University of Cape Town

Msc. Data Science

Dissertation

Hospital Readmission Risk

Author:
Amos Mugova

Student Number:
MGVAMO001

Supervisors:

Mr Sulaiman Salau

Dr Sebnem Er

Thursday 13th June, 2024

Declaration

1. I declare that this minor dissertation is my own work and has not been presented for any academic qualification or assessment at any other University.
2. I assert that I have neither permitted nor will permit the replication of my work by others for the purpose of presenting it as their own.
3. I acknowledge and have appropriately cited all major contributions and quotations from others' work within this minor dissertation.

I hereby grant the University of Cape Town permission to copy and disseminate this work, or any part thereof, for study and research purposes.

Signed by candidate

Amos Mugova

Thursday 13th June, 2024

Abstract

Hospital readmissions are a significant challenge in healthcare, as they lead to increased costs, higher risk of mortality, treatment complications, and patient distress. This minor dissertation, set within the South African healthcare framework, investigates the potential of both traditional clinical screening tools and advanced statistical learning methods for predicting hospital readmission risk. The methods considered include the LACE score, decision trees, logistic regression, random forests, gradient-boosting methods, and neural networks.

The study uses data from South Africa's privately insured demographic, provided by a private insurer. It includes a comprehensive array of patient information such as demographics, prescribed medications, medical procedures undergone, and historical hospital usage. Feature selection methods were used to identify relevant variables for model training, and the effectiveness of these variables was assessed based on their ability to differentiate between patients at risk of hospital readmission within 30 days after discharge.

The statistical learning methods' efficacy was measured using several performance indicators, such as prediction accuracy, F1 score, Area Under the Receiver Operating Characteristics Curve (AUC), Area Under the Precision-Recall Curve (AUC-PR), and the Matthews Correlation Coefficient (MCC). The study found that the neural network model outperformed the other statistical learning methods evaluated across various metrics.

Moreover, the research extends the range of variables used to predict hospital readmissions beyond the traditional LACE score, incorporating critical factors such as the frequency and costs of previous hospital visits, expenses related to specialist services, patient age, and the primary diagnosis category.

Acknowledgements

The journey towards the completion and subsequent publication of this dissertation has been nothing short of an invigorating challenge. It demanded the investment of unwavering dedication, inexhaustible patience, and significant sacrifices. The road was arduous and steep, but a steadfast commitment to this intellectual pursuit underpinned each step.

My heartfelt gratitude extends to my son, Tinotenda, who has been my bedrock of inspiration, a wellspring of joy, and a ceaseless motivator since his birth. His presence and spirit have been the beacon that illuminated the path during the darkest hours of this journey. His infectious enthusiasm has continually rekindled my vigour and lent an even more profound meaning to my endeavours.

I sincerely appreciate my esteemed supervisors, Mr Sulaiman Salau and Dr Sebnem Er. Their seasoned expertise and guidance have been instrumental in the maturation of this dissertation. They nurtured my research skills, continually challenged me to broaden my horizons, and walked with me every step of the way. Their unfaltering faith in my capabilities has been a monumental force that empowered me to traverse this scholarly journey.

As I close this chapter, I feel humbled and immensely enriched by the experiences, the knowledge gained, and the growth this journey has fostered within me. My gratitude for all those who have been a part of this journey knows no bounds.

Contents

List of Figures

List of Tables

List of abbreviations

AUC	Area Under the Receiver Operating Characteristics Curve
AUC-PR	Area Under the Precision-Recall Curve
CART	Classification and Regression Tree
CPT	Current Procedural Terminology
DRG	Diagnosis-Related Group
GBM	Gradient Boosting Machines
GDP	Gross Domestic Product
ICD-10	International Classification of Diseases, Tenth Revision
Lasso	Least absolute shrinkage and selection operator
MCC	Matthews Correlation Coefficient
MDC	Major Diagnostic Categories
ROC	Receiver operating characteristic
SVM	Support vector machine
USD	United States Dollars

Chapter 1

Introduction

1.1 Context and background to the study

Healthcare is indispensable to any contemporary culture, and every country invests substantially in their healthcare systems. According to the World Health Organisation (2021), global spending on healthcare more than doubled in real terms over the past two decades, reaching 8.5 trillion United States Dollars (USD) in 2019, equivalent to 9.8% of the global Gross Domestic Product (GDP). Despite this high expenditure, healthcare costs are increasing at a prohibitive rate, primarily driven by factors such as rising readmission rates, the effects of ageing populations, population growth, increasing numbers of people living with long-term conditions, long-term co-morbidities, and cumulative increases in treatment and technology costs, to name a few (Lewis et al., 2011; Hosseinzadeh et al., 2013; Stone et al., 2010; Strandberg et al., 2011).

Hospital readmissions account for a sizable portion of healthcare spending in many countries (Hosseinzadeh et al., 2013). In clinical terms, readmission refers to a patient's return to a hospital or medical facility within a set period, typically 30 days, after being discharged from an initial stay (Dreyer and Viljoen, 2019; Stone et al., 2010). These subsequent admissions may occur at the same hospital or a different one and could be for scheduled or emergency interventions, encompassing

both surgical and medical treatments (Stone et al., 2010).

Hospital readmissions substantially burden the healthcare system and carry warning signs of poor-quality care (Jencks et al., 2009). The World Health Organisation (2005) highlighted hospital readmissions as a significant undesirable outcome of healthcare systems, emphasising the reduction of such occurrences as a critical strategic goal. Many countries have since implemented policies and incentives to mitigate the severity of high hospital readmissions on costs and healthcare outcomes. For example, the United States of America mandated multiple initiatives through the Patient Protection and Affordable Care Act of 2010 (Kocher and Adashi, 2011) and the implementation of the Hospital Readmissions Reduction Program (HRRP) in 2012 (McIlvennan et al., 2015), all aiming to reduce hospital readmissions.

The growth of accessible health-related data presents unprecedented opportunities to conduct analyses to improve a patient's health and reduce related costs. Today, patient records and administrative data, such as surgeon and room availability, are kept in electronic formats that allow healthcare professionals and administrative staff to utilise decision-support systems to optimise the allocation of the limited healthcare resources to enhance patient outcomes and minimise costs (Wasylewicz and Scheepers-Hoeks, 2019). These decision-support systems are evolving, and there is great interest among healthcare stakeholders in applying statistical learning methods in healthcare and developing more advanced support systems to aid in the prediction of hospital readmission (Huang et al., 2021; Hosseinzadeh et al., 2013).

In the 2020/21 financial year, South Africa's total private healthcare expenditure amounted to approximately R178.04 billion (Council for Medical Schemes, 2022). During the same period, R58.4 billion was spent on public healthcare (Department of Health, 2022). This brings the total healthcare spending to R236.4 billion South African Rand (R), accounting for approximately 6.76% of the country's GDP (Statistics South Africa, 2022). Despite this significant investment, South Africa continues to face substantial healthcare challenges. The country is grappling with a high prevalence of Human Immunodeficiency Virus (HIV) and Acquired Immune

Deficiency Syndrome (AIDS), along with tuberculosis and an increasing burden of non-communicable diseases. These health issues are compounded by weak economic growth and a scarcity of critical healthcare resources, including physicians, hospital beds, and surgeons. Moreover, disparities between the public and private healthcare sectors contribute to South Africa's lagging healthcare outcomes compared to other middle-income countries (Informa, 2019).

While applying statistical learning methods to predict and investigate significant factors contributing to hospital readmissions is common worldwide, such research is notably scarce in South Africa (Dreyer and Viljoen, 2019). This gap is significant because international studies and methods are not directly transferable to the South African context, given its unique healthcare ecosystem characterised by a high disease burden, prevalent comorbidities, and pronounced social inequalities, as discussed earlier. This minor dissertation presents an extensive exploration of how statistical learning methods can be adapted to predict and identify key factors associated with the risk of hospital readmission among the privately insured population in South Africa.

1.2 Aims and objectives

This minor dissertation aims to:

- identify the most effective statistical learning method for predicting all-cause 30-day hospital readmissions.
- Determine the critical factors associated with hospital readmissions within the privately insured South African population.

Objectives:

The goals of this minor dissertation are outlined as follows:

- To explore the application of conventional clinical screening and statistical learning techniques in predicting the likelihood of hospital readmission.

- To compare the efficacy of these methods using various model performance metrics, including the Area Under the Receiver Operating Characteristic Curve (AUC), the Area Under the Precision-Recall Curve (AUC-PR), the F1-score, Precision, and Recall.
- To utilise statistical learning methods to identify critical factors associated with hospital readmissions in the privately insured population of South Africa.

1.2.1 Layout of the minor dissertation

This minor dissertation has six chapters, organised as follows:

- **Chapter 2:** This section analyses various predictive models for hospital readmissions, focusing on the evolution and application of these methodologies over time.
- **Chapter 3:** Discusses the theoretical underpinnings of the statistical learning methods for predicting hospital readmissions. This includes an exploration of their evolution and the optimisation strategies applied to these methods.
- **Chapter 4:** Details the data used and summarises the data wrangling and feature engineering processes used in this minor dissertation.
- **Chapter 5:** Explores applying statistical learning methods and the LACE score to training and unseen datasets. It also identifies and summarises the critical factors associated with hospital readmissions, as revealed by this minor dissertation.
- **Chapter 6:** This section wraps up the minor dissertation by summarising the key findings, exploring their practical implications, and proposing possible future research in the field.

Chapter 2

Literature review

The management of increasing readmission rates is a significant challenge healthcare systems face worldwide (World Health Organisation, 2021). Often, hospitals discharge patients as a strategy to free up beds in overburdened systems. However, this approach can inadvertently lead to higher hospital readmission rates, especially if the discharge is premature and not all factors are adequately considered (World Health Organisation, 2021). While it might seem intuitive that healthcare practitioners like physicians and nurses, who have personal familiarity with their patients, would be best suited to assess the risk of readmission, there is a growing shift towards relying on statistical learning methods (Lewis et al., 2011). This shift is grounded in three primary theoretical reasons, suggesting that statistical learning methods may offer a more informative alternative over traditional clinician-based methods (Van Walraven et al., 2010; Lewis et al., 2011).

Firstly, statistical learning methods have the capacity for regular and widespread screenings across entire patient populations, a task that is not feasible for individual healthcare professionals. Secondly, while clinicians are limited in their ability to predict outcomes for unfamiliar patients, statistical learning methods can incorporate a wide array of factors. This encompasses patient engagements with various facets of the healthcare system, socio-economic factors like deprivation, and the diverse approaches of different hospitals, all contributing to enhanced predictions of

readmissions (Van Walraven et al., 2010; Lewis et al., 2011). Thirdly, clinicians are naturally prone to cognitive biases, which can hinder their ability to transform individual observations into reliable estimates for the broader population.

Various statistical learning approaches have been innovated and implemented to precisely pinpoint patients with a high risk of readmission, underscoring the crucial need for an advanced and adaptive support framework in healthcare administration (Corrigan and Martin, 1992).

This chapter describes the concept of hospital readmission explored in this minor dissertation and subsequently reviews the scientific literature on clinical and statistical learning techniques for predicting hospital readmission.

2.1 Definition of hospital readmission

The concept of hospital readmissions first appeared in medical literature in Woodside (1953). Since then, a vast corpus of literature on hospital readmissions has emerged, discussing their prevalence, causes, associated patient and hospital characteristics, and prevention strategies (Sheingold et al., 2016; Wiseman et al., 2019; Kansagara et al., 2011). Readmission denotes a patient being admitted to a hospital or healthcare facility within a stipulated period after the initial admission (Hackbarth et al., 2007). Such readmissions can take place in the same hospital or a different one and may involve scheduled or emergency surgical or medical treatments (Stone et al., 2010).

The literature presents varying timeframes for analysing readmissions (Heggestad and Lilleeng, 2003). Shorter spans, such as two weeks, are also employed, and certain readmission studies investigate considerably more extended periods, like six or twelve months (Benbassat and Taragin, 2000). Nonetheless, a one-month (28-31-day) interval post-discharge is commonly adopted (Ashton and Wray, 1996; Heggestad and Lilleeng, 2003). The choice of shorter periods (less than a month) is preferred when hospital readmissions serve as a predictor of the outcome of a previous hospital stay

or disease episode, necessitating the consideration of the link between care processes and outcome measures (Ashton et al., 1997; Hammermeister et al., 1995; Benbassat and Taragin, 2000). Conversely, selecting a more extended timeframe is associated with an increased emphasis on the disease's natural progression and community factors (Heggestad, 2002).

Hospital readmissions are categorised as either all-cause or disease-specific. All-cause readmissions are defined as any hospital readmission occurring within a specified timeframe, irrespective of the underlying disease. In contrast, disease-specific readmissions are directly linked to the initial hospitalisation and pertain to complications, exacerbations, or sequelae related to the original illness or its treatment(Ruff et al., 2021). Much of the research dedicated to predicting 30-day readmission risks has historically focused on specific diseases, such as congestive heart failure (Balla et al., 2008), cancer (Francis et al., 2015), or cases requiring emergency readmissions (Sushmita et al., 2016). While these disease-specific models are undeniably crucial, there is a growing consensus among researchers that models predicting all-cause readmission risks, which are not limited to any particular disease condition, are equally crucial (Kansagara et al., 2011; Wiseman et al., 2019). This minor dissertation is dedicated to examining 30-day all-cause hospital readmissions.

2.2 Predicting hospital readmission risk

The methods for predicting hospital readmissions are diverse, ranging from straightforward clinical rule-based approaches to advanced statistical learning techniques. To facilitate a structured literature review, these methods have been categorised based on the type of model they employ:

- Clinical rule-based methods: These are straightforward, easy-to-use approaches where predefined clinical rules guide decision-making.
- Regression-based methods: These methods use statistical regression techniques to predict readmissions based on various factors and variables.

- Decision tree methods: These methods employ a hierarchical, tree-like decision-making structure, where readmission predictions are made through a series of branching choices based on specific criteria.
- Others: This category includes more complex techniques, such as neural networks and support vector machines, which are part of advanced statistical learning methodologies.

The subsequent subsections will delve into the literature in greater detail, exploring how each method has been applied and developed for predicting hospital readmissions.

2.2.1 Clinical rule-based methods

The historical landscape of clinical rule-based readmission risk assessment is characterised by the use of various screening techniques to predict hospital readmissions. These techniques, primarily rely on a limited set of readily available clinical data and facilitate straightforward computations (Morgan et al., 2019). Key among these is the **LACE** score, an acronym for **L**ength of stay, **A**cuity level of admission, **C**omorbidity condition, and **E**mergency room use (Van Walraven et al., 2010). The **mLACE** score, introduced by Morgan et al. (2019), is a modification of the LACE score that refines this assessment by including factors such as emergency department visits in the six months preceding hospital admission rather than focusing solely on emergency room use during the current admission. Another significant tool is the **HOSPITAL score**, developed by Donzé et al. (2013), which evaluates seven variables: hemoglobin level, discharge from an oncology service, sodium level, procedures during the index admission, type of admission, the number of admissions in the past year, and length of stay. While various tools have been employed, the LACE score has gained widespread acceptance and usage in predicting hospital readmissions, as noted by Ben-Chetrit et al. (2012) and Van Walraven et al. (2010). This score has been particularly valued for its efficacy and ease of application in diverse clinical settings.

The LACE score, developed using data from 4812 patients, was further validated

on a substantial cohort of 1000000 patients, utilising records from 2004 to 2008 in Ontario, as documented by (Van Walraven et al., 2010). To assess its predictive accuracy, the C-statistic, or concordance statistic, was used. In its derivation set, the LACE score achieved a C-statistic of 0.71; in the validation set, it recorded 0.69 (Van Walraven et al., 2010).

In a separate study, Dreyer and Viljoen (2019) applied the LACE score at Tygerberg Hospital's Department of Internal Medicine in Cape Town. Their analysis encompassed 11826 admissions from 1 January 2014 to 31 March 2015, during which 1242 patients were readmitted within 30 days, indicating a 10.5% readmission rate. The study reported a C-statistic of 0.63, highlighting moderate predictive ability.

They identified the number of comorbidities as the most significant risk factor for readmission. Other factors contributing to 30-day readmissions included hospital-acquired infections, negative drug reactions (with a particular emphasis on warfarin toxicity), sub-optimal discharge planning, and mistakes made by doctors.

The LACE score had to undergo practical, unavoidable changes to address its flaws. A significant issue for healthcare providers using the LACE score is the uncertainty of the patient's length of stay at the time of admission, as this detail typically becomes clear only at discharge. Calculating the length of stay at the end of the hospital stay postpones the planning for discharge, rendering the initial method of applying the LACE score impractical (El Morr et al., 2017). Accordingly, El Morr et al. (2017) proposed the LACE-rt, a modified LACE score that can be used in a real-time setting that bases the length of stay (L) on the patient's past (rather than current) acute care admission within the last 30 days. However, the LACE-rt score performed worse than the original LACE score (C-statistic 0.684, 95% CI 0.679-0.691) (El Morr et al., 2017). The authors argued that this disparity in performance was to be expected and is attributable to population cohort differences (El Morr et al., 2017). The mean age in their sample was 74.29 years, compared to 61.3 years in the Van Walraven et al. (2010) population investigated, and the LACE score is known to perform worse in older people (Cotter et al., 2012).

The modified (mLACE) score is another modification aimed at improving the LACE score's clinical utility and international generalisability, as noted by Sarah Rimar MD (2014). Unlike the traditional LACE score that measures severity based on whether a patient is admitted through the emergency department, the mLACE score evaluates the severity based on whether the patient is on observation or admitted as an inpatient (Sarah Rimar MD, 2014). However, the mLACE score's ability to predict readmissions was found to be moderately effective, indicated by a C- statistic of 0.67.

The LACE score and its variants are easy-to-use scoring tools with high transparency and interpretability. However, the LACE score has variable results in the literature for predicting 30-day readmission (Cotter et al., 2012), especially in different patient populations outside Ontario where it was initially validated (McNaughton et al., 2013; Wong et al., 2011). In particular, Cotter et al. (2012) showed that the LACE score is relatively poor at predicting hospital readmission in an older UK population. The method also has low accuracy and discriminative power as noted by Wang et al. (2014). Some researchers also believe that more complex statistical learning methods that incorporate a broader range of predictors, including medical comorbidities and basic demographic information, can generate more precise risk assessments compared to the LACE score (Hammill et al., 2011; Cotter et al., 2012; Kansagara et al., 2011). An example is the regression-based methods discussed in the next section.

2.2.2 Regression based methods

Regression methods are commonly applied in medical studies for their ability to forecast outcomes, adjust for confounders, and assess relationships (Stoltzfus, 2011). Logistic regression, in particular, stands out as a powerful and reliable method in statistical analysis. It is employed to investigate how multiple independent variables influence a binary outcome, offering insights into the distinct impact of each variable. It was developed by Cox (1958) as a refinement of linear regression. The technique uses a logistic function to depict the association between a single dependent binary variable and one or more independent variables, which may be nominal, ordinal,

interval, or ratio-level.

Logistic regression offers a significant benefit over other methods for predicting hospital readmissions, such as the LACE score, by incorporating a broader range of variables while maintaining ease of interpretation (James et al., 2013). One key feature, the Odds Ratio (OR) derived from the model's coefficients, provides a straightforward way to understand the probability of readmission in comparison to the probability of not being readmitted. This makes logistic regression not only a versatile tool for analysis but also accessible for interpreting the impact of various factors on readmission risks (Schober and Vetter, 2021). Moreover, unlike the LACE score, logistic regression can readily incorporate additional independent predictors. In the literature, it has been observed that incorporating additional variables such as the characteristics of individual patients, the attributes of healthcare providers including the speciality, age, experience, practice environment, and patient volume of physicians, as well as the availability of community healthcare resources, enhances the accuracy of models predicting hospital readmissions (Kind et al., 2014; Corrigan and Martin, 1992; Cotter et al., 2012; Kansagara et al., 2011).

To this end, logistic regression is a widely used technique in predicting hospital readmission (Boulding et al., 2011; Artetxe et al., 2018). Demir et al. (2009) used logistic regression to predict the readmission risk of Chronic Obstructive Pulmonary Disease (COPD) patients admitted to England's hospitals. Their study used patient-level predictors measured over seven years, including demographics, admission events, diagnosis codes, and length of stay (LOS). Solely using patients' history of readmissions, their model had a C-statistic of 0.71, demonstrating that a simple logistic model with no covariates has the potential to estimate the risk of readmission. Their research found an association between the number of previous readmissions and an elevated risk of future readmission. However, their model was designed for a particular cohort and cannot be easily generalised to other patient groups.

Wennberg et al. (2006) also developed a logistic regression model aimed at identifying patients with a high risk of hospital readmission in England. Their study used

demographic, medical, pharmaceutical, and psychological predictors from data collected over five years and established that the key predictors for hospital readmission included age, sex, ethnicity, number of previous admissions, and clinical condition. Their model had a ROC statistic of 0.685 (Wennberg et al., 2006).

Complementing the findings of Wennberg et al. (2006) is the work of Billings et al. (2013), who used multivariate logistic regression models to identify patients at risk of hospital admission. Additionally, they explored the effects of incorporating data from various sources, including hospital inpatient and outpatient services, as well as electronic medical records from general practitioners (GPs). The results indicate that including more predictors enhanced the predictive capability of the model. Specifically, the ROC statistic rose from 0.731 in the basic model to 0.780 in the comprehensive model after the introduction of additional variables. Despite this improvement, the model's true positive rate (sensitivity) remained quite low. Billings et al. (2013) also pointed out the shortcomings of the data utilised in their research, emphasising that high-risk patients frequently possess critical attributes related to care requirements and capacity that administrative data fail to capture (Zhu et al., 2015).

Howell et al. (2009) investigated the use of a multivariate logistic regression model to forecast hospital readmissions among patients with chronic conditions like congestive heart failure, chronic obstructive pulmonary disease, diabetes, or dementia, drawing on data from Queensland Hospital in Australia. The research identified multiple variables such as age, comorbidities, socio-economic challenges, and the count of prior hospitalisations as significant predictors of readmission Howell et al. (2009). The model demonstrated moderate predictive accuracy, with the area under the receiver operating characteristic curve reported at 0.65 (Howell et al., 2009).

Logistic regression-based models are ineffective for modelling the nonlinear relationships found in medical data (Schober and Vetter, 2021). They also rely on stringent underlying assumptions (Stoltzfus, 2011) such as:

- Logistic regression presupposes a linear connection in the logit for continuous independent variables, such as age, implying that there should be a straightforward relationship between these variables and the logit-transformed results.
- Logistic regression requires absence of multicollinearity, or redundancy, among independent variables. This means that the independent variables should not be too highly correlated with each other.
- Lastly, logistic regression usually demands a substantial sample size. A common rule of thumb is to have at least 10 instances of the least common outcome per independent variable included in the model.

As a consequence, when compared to other statistical learning methods such as decision trees and random forests, they tend to have low prediction accuracy when applied to hospital administrative data (Zhu et al., 2015). Several statistical learning methods have been used to improve our understanding of the predictors of readmission. These include Random forest, Gradient boosting machine (GBM), Extreme gradient boosting (XGBoost), and support vector machine, which are discussed in the sections that follow (Darabi et al., 2021).

2.2.3 Decision tree

Decision trees have gained popularity as a statistical learning method for predicting hospital readmissions (Demir et al., 2009). These models, which are non-parametric, predict an outcome variable based on independent variables, regardless of whether these variables are continuous, discrete, or categorical (Cox, 1958). Introduced in the early 1980s, decision trees offered an advantage over logistic regression by enabling the learning of non-linear decision boundaries effectively (Cox, 1958). Furthermore, decision trees eliminate the need for a predefined parametric relationship between predictor variables and the predicted outcome, a requirement often necessary in logistic regression methods (James et al., 2013; Austin, 2007).

To this end, decision trees are one of the most utilised techniques in medical and healthcare applications. Demir et al. (2009) used decision trees to predict emer-

gency readmissions of 963 patients with chronic obstructive pulmonary disease and asthma. In their study, the predictor variables included medical comorbidities, previous utilisation of medical services, patient demographics, and socio-demographic and social determinants. They divided their data into derivation and validation samples. Classification trees and logistic regression models were fitted using data in the derivation sample, and predictive accuracy was evaluated using the validation sample by means of the area under the receiver operating characteristic (AUROC) curve (Harrell et al., 2001). In this context, AUCROC is defined as the proportion of times the model accurately discriminates between readmitted and non-readmitted patients.

Table 2.1 below presents the mean area under the ROC curve for each model in both the derivation and validation samples from the study conducted by Demir et al. (2009). The results demonstrate that both the decision tree model and logistic regression possess good discriminative abilities. Specifically, the decision tree model achieved a mean ROC curve area of 0.948 in the derivation samples, which slightly decreased to 0.924 in the validation samples, representing a minimal decrease of 0.024. Meanwhile, the logistic regression model attained a mean ROC curve area of 0.928 in the validation sample, as reported by Demir et al. (2009).

Receiver Operating Characteristic (ROC) Area:		
Model	**Training sample (%)**	**Validation sample (%)**
Decision tree	94.8	92.4
Logistic regression	97.7	92.8

Table 2.1: Findings from the study carried out by Demir et al. (2009)

Their research also reveals that factors associated with previous medical service usage, including the length of hospital stay and past readmission history, are significant predictors of hospital readmission. The findings indicate a 90% likelihood of readmission for patients whose prior hospital stays (both emergency and non-emergency) exceeded half a day and who had experienced two or more emergency admissions within the preceding 90 days.

James et al. (2013) argued that decision trees more closely mirror human decision-making processes compared to logistic regression. The outcomes of decision trees can be visually represented as decision rules. This graphical representation makes them more intuitive and easier to comprehend in a clinical setting than the outcomes derived from many other statistical methods (Garzotto et al., 2005; Demir et al., 2009).

However, one significant limitation of decision trees is their dependency on the sample; a slight variation in sample size can notably affect the model's outcome, as noted by James et al. (2013). Additionally, decision trees often do not achieve the same level of predictive accuracy as some other classification methods, such as random forest. Despite this, their predictive performance can be substantially improved through the use of ensemble techniques like bagging, random forests, and boosting, which aggregate multiple decision trees. These ensemble methods are discussed in the subsequent section.

2.3 Ensemble methods

Ensemble methods aim to improve the accuracy of results by combining multiple models instead of using a single model (James et al., 2013; Harrell et al., 2001). Ensemble methods can be divided into two types based on their structure: parallel and sequential, as displayed in Figure 2.1 below (Xia et al., 2017).

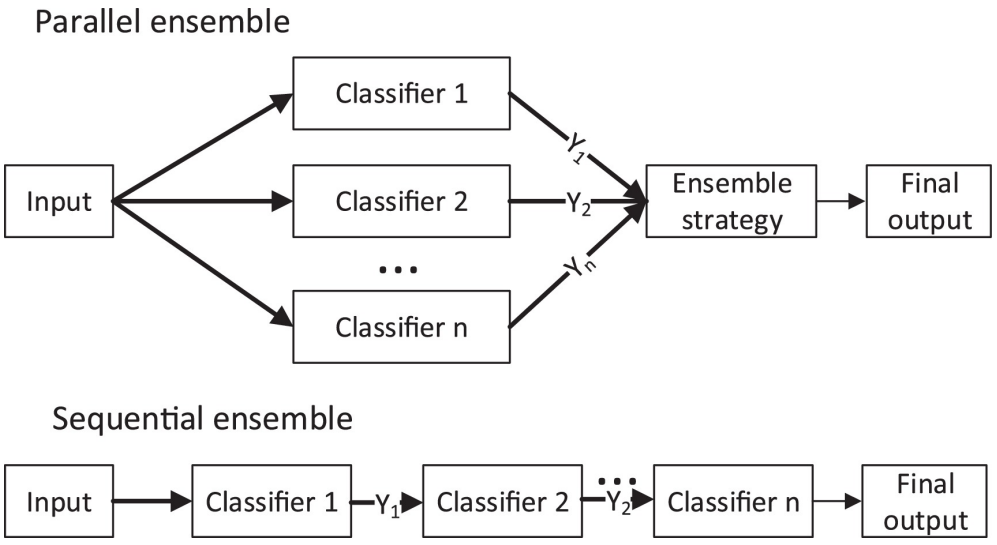


Figure 2.1: Parallel ensemble and sequential ensemble (Xia et al., 2017)

The parallel ensemble structure integrates various learning algorithms, each independently generating a model in parallel. This approach allows for combining diverse predictive strengths and weaknesses of different models. In contrast, the sequential ensemble method involves sequentially feeding the output of one model into the next, refining predictions at each step until an outcome is achieved. Both the Random Forest and Gradient Boosting Machine discussed below, employ these ensemble techniques. Random Forest is an example of a parallel ensemble method while Gradient Boosting Machine is a sequential ensemble method.

2.3.1 Random Forest

As noted earlier, one issue with decision trees is their relative instability (James et al., 2013). This means that even a slight change in the data can lead to a significantly different decision tree. A solution to this problem is to use ensemble learning methods such as Random Forests. Random Forests mitigate instability by constructing multiple decision trees and aggregating their results. Random forest was first introduced in 2001 by Breiman (2001) and is based on the CART algorithm (Breiman and Friedman, 1984) and the bagging ensemble method (Leo, 1996). The random forest algorithm constructs each tree using a random sample of the observation and feature space from the original dataset. This random sampling means

that each tree in the forest is different because it is trained using a different portion of the data and has the effect of correcting the tendency of individual regression trees to overfit the training data (Breiman, 2001). Unlike conventional decision tree models, a random subset of all the predictors identifies the best classifier at each node (Liaw and Wiener, 2002).

Random forests have been widely used in the healthcare literature to predict hospital readmissions. Zhu et al. (2015) compared the performance of a random forest algorithm, neural network model, support vector machine (SVM), non-linear SVM, Cox regression and LACE score in predicting 30-day hospital readmission. Their model used patient characteristics, such as age and drug risk, from the general U.S. population sample data. The results of Zhu et al. (2015)'s research are provided in Table 2.2 and show that the random forest model outperformed logistic regression and ranked second with high accuracy (74.4%) and sensitivity (87.4%).

Model	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
LACE scores	43.5	51.8	21.8
RBFNN	54.6	56.1	49.3
Logistic regression	57.9	60.5	49.3
Random forest	74.4	87.4	30.7
PSO–SVM with RBF	78.4	97.3	8.6

Table 2.2: Findings from Zhu et al. (2015)'s study

Random forests are one of the most accurate learning algorithms available and can handle missing observations efficiently (Hastie et al., 2009). However, the main disadvantage of random forests is that they tend to overfit some datasets with noisy classification problems. Furthermore, random forests are likely to perform poorly when the number of variables is large, but the fraction of relevant variables is small (Hastie et al., 2009). At each split, the chance may be low for the relevant variables to be selected, making the variable importance scores from the random forest unreliable (Hastie et al., 2009).

2.3.2 Gradient Boosting Machine

The Gradient Boosting Machine (GBM) is a powerful method for enhancing decision tree predictions, applicable to a broad spectrum of statistical learning tasks, including regression and classification (James et al., 2013). In contrast to the parallel model construction of random forests, GBM adopts a sequential strategy, where each subsequent tree is developed based on the residuals or errors of the preceding trees (James et al., 2013; Hastie et al., 2009). This method, known as boosting, does not adjust sample weights but instead fits each new tree to the residual errors made by the previous trees (James et al., 2013). Specifically, samples that were incorrectly predicted by prior trees have their errors used to train the next tree, effectively giving those samples more focus in subsequent iterations. The final model aggregates the predictions of all individual trees by summing their outputs, rather than using a weighted voting mechanism, thereby improving prediction accuracy and model robustness (Freund et al., 1996).

Sushmita et al. (2016) evaluated five statistical learning methods for predicting 30-day hospital readmission using data from the Nationwide Readmission Database. Their study compared the Gradient Boosting Machine (GBM), LACE, Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression. Table 2.3 shows that the results from their study indicate that the Gradient Boosting Machine is one of the best methods, with a sensitivity of 90.43%, specificity of 18.24%, and precision of 29%.

The benefits of GBMs include improved predictive accuracy, the ability to work with categorical and numerical values, and their ability to manage missing data without requiring imputation (James et al., 2013). However, GBMs may overfit the training data, necessitating additional methods such as cross-validation. GBM results are also less interpretable, although this is easily addressed with various tools such as variable importance and partial dependence plots, among others (James et al., 2013; Hastie et al., 2009).

Algorithm	Sensitivity (%)	Specificity (%)	Precision (%)
LACE	76.42	38.95	31.63
SVM	98.11	1.84	26.98
Decision Trees	94.07	9.04	27.65
Random Forest	84.76	25.60	29.63
Logistic Regression	92.47	13.24	28.26
GBM	90.43	18.24	29.02

Table 2.3: Performance comparison of different statistical learning methods for predicting 30-day hospital readmission. The results are based on a study conducted by Sushmita et al. (2016).

2.3.3 Other statistical learning methods used for predicting hospital readmission

A review conducted by Huang et al. (2021) focused on predictive models of hospital readmission that specifically use statistical learning methods and are based on all types of databases across different healthcare settings in the United States of America. The objective of Huang et al. (2021)'s scoping review was to synthesise the current literature on the types of statistical learning methods used in predicting hospital readmissions in the United States of America. Furthermore, their review provides a comparative analysis of predictive performance, specifically in terms of the Area Under the Curve (AUC), across all statistical learning methods used for hospital readmission prediction, and the findings of their study are concisely illustrated in Figure 2.2 below.

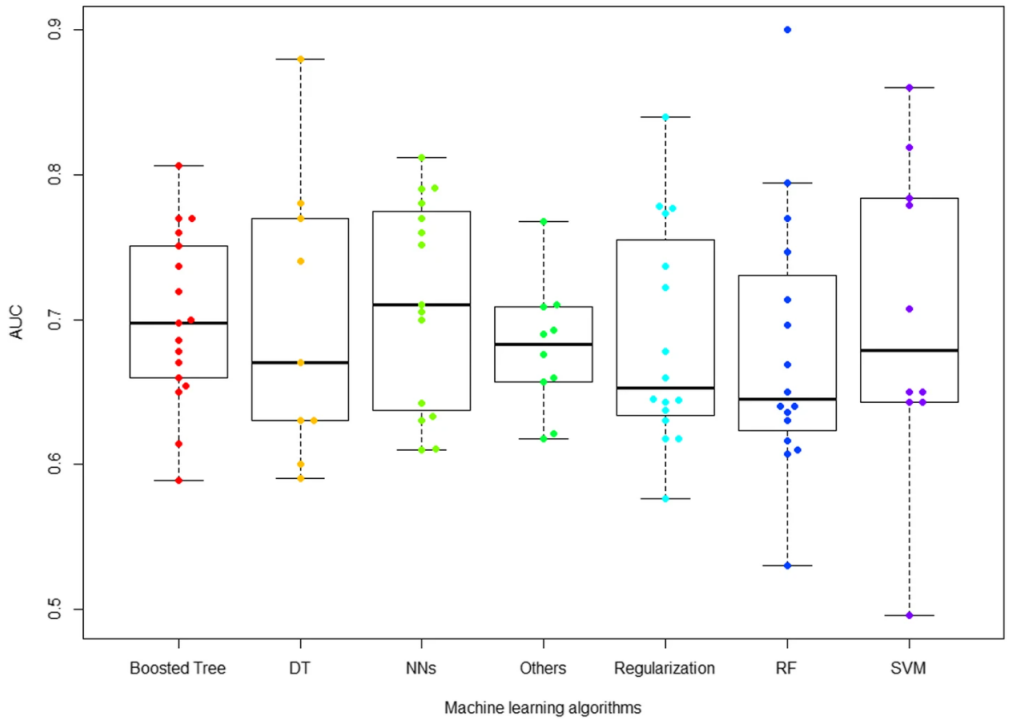


Figure 2.2: The results of Huang et al. (2021)'s review

Their research showed that statistical learning methods like support vector machines, regularised logistic regression, neural networks, and tree-based algorithms are often employed to forecast hospital readmissions in the United States of America. Their analysis, which encompassed 43 studies, found that regularised logistic regression (28%), neural networks (33%), tree-based methods (53%), and support vector machines (23%) were the algorithms most frequently implemented. Notably, a significant proportion of these studies (65%) reported AUC values exceeding 0.70. However, the studies also demonstrated considerable variation in AUC values, with a median of 0.68, an interquartile range (IQR) of 0.64 to 0.76, and an overall range of 0.50 to 0.90 (Huang et al., 2021).

2.3.4 Summary of statistical learning methods

The review of the aforementioned published studies indicates that statistical learning models outperform both threshold models and clinician knowledge in identifying patients at risk of future readmission. However, the literature does not establish

a consensus on the superiority of any specific statistical learning technique. Demir et al. (2009) suggested that this absence of consensus could be attributed to the diverse definitions of hospital readmission present in the literature. Furthermore, the model's intended purpose is an essential factor to consider when assessing the relative accuracy of these techniques. The majority of studies that aim to accurately predict the 30-day risk of readmission mainly focus on specific disease cohorts, such as patients with congestive heart failure (Balla et al., 2008), cancer (Francis et al., 2015), or emergency readmissions (Sushmita et al., 2016). While these disease-specific models are beneficial, there is considerable value in developing all-cause readmission risk models that are not limited to specific conditions.

The literature also indicates that the accuracy of readmission prediction heavily depends on the predictors used. The following section will explore these predictors in more detail.

2.4 Factors associated with hospital readmission

Extensive research has been conducted on the factors influencing hospital readmission, with studies highlighting the importance of the chosen factors in the accuracy of readmission prediction models (Anderson and Steinberg, 1984; Holloway et al., 1988; Kansagara et al., 2011; Cotter et al., 2012; Kind et al., 2014). These factors range from patient-specific clinical and resource utilisation characteristics to sociodemographic data such as age, gender, marital status, and educational attainment (Curry et al., 2005). Additionally, the variables related to healthcare providers, including variables like physician specialisation, experience, practice setting, and the availability of healthcare resources, have been examined (Holloway et al., 1988). However, there remains a debate in the literature regarding which factors are most predictive of hospital readmission. The following subsections delve into the literature review of these factors.

2.4.1 Socio-demographic data

Most studies concur that demographic characteristics in isolation do not exert a significant predictive impact on hospital readmissions (Curry et al., 2005). Nevertheless, several studies have identified that specific demographic attributes exhibit a more significant predictive capacity than others (Holloway et al., 1988). For example, the research conducted by Holloway et al. (1988) revealed a positive correlation between age and the probability of hospital readmission. This relationship persists until the age of sixty-five. Beyond this age, the association between age and readmission rates diminishes. Gender differences in readmission rates have been observed, particularly in studies with shorter follow-up periods. These studies indicate that males have higher readmission rates, although this discrepancy diminishes and becomes statistically insignificant in studies with longer follow-up durations (Anderson and Steinberg, 1984; Graham and Livesley, 1983). The role of marital status, especially living alone, in predicting readmission risk has yielded inconsistent findings across various studies. Some research suggests a link between living alone and higher readmission risk, while other studies do not support this association (Curry et al., 2005; Holloway et al., 1988). This variability underscores the complexity of predicting hospital readmissions and the multifactorial nature of the underlying risk factors.

2.4.2 Prior utilisation data

Including the prior cost and medical service utilisation data has been shown to significantly enhance the predictive accuracy of readmission models (Curry et al., 2005). A study conducted by van Barneveld et al. (1997) demonstrate that incorporating a year's prior utilisation data can increase the explained variance from as little as 3% with solely socio-demographic data to approximately 26%. This underlines the value of considering inpatient and outpatient utilisation data in predicting readmission.

2.4.3 Diagnostic data

Diagnostic data also plays a critical role in readmission rates, with distinct patterns observed in medical versus surgical admissions (Corrigan and Martin, 1992; Curry et al., 2005; Tsai et al., 2014). Surgical site infections, for instance, are a significant predictor of readmission in surgical cases (Snyders et al., 2020). The presence of chronic diseases and the need for multiple surgical procedures also elevate readmission risks, alongside factors like poor health status and physical impairments (Kansagara et al., 2011; Holloway et al., 1988).

2.4.4 Clinical and Pharmacy data

Despite the acknowledged predictive significance of pharmaceutical data in statistical learning models for forecasting hospital readmissions, this type of data is often underutilised in current research and clinical practice (Corrigan and Martin, 1992). The underrepresentation of pharmaceutical data in these models is noteworthy, especially considering the substantial evidence linking certain medications to heightened readmission rates. Specifically, medications such as steroids, narcotics, anticholinergics, and antibiotics have been identified as strong predictors of increased readmission risks (Kansagara et al., 2011).

The underutilisation of pharmaceutical data in predictive models may be due to various factors, including the complexity of integrating medication data into existing models, data privacy concerns, and the dynamic nature of patients' medication regimens (Corrigan and Martin, 1992). Additionally, the potential interactions between multiple medications, known as polypharmacy, further complicate the predictive analysis. This complexity suggests that more comprehensive and sophisticated models are needed to incorporate pharmaceutical data accurately (Corrigan and Martin, 1992).

Moreover, the incorporation of pharmaceutical data into predictive models not only enhances the accuracy of readmission predictions but also provides a more holistic view of patient health and risk factors (Kansagara et al., 2011). Understanding

the impact of specific medications on readmission rates can inform clinical decision-making and targeted interventions to reduce these risks. For instance, closer monitoring and follow-up for patients prescribed high-risk medications could be implemented to mitigate the likelihood of readmission.

Overall, integrating pharmaceutical data into statistical learning models represents a significant opportunity for improving the accuracy and utility of hospital readmission predictions. This approach necessitates a multidisciplinary effort involving collaboration between clinicians, pharmacists, data scientists, and healthcare administrators to effectively harness the potential of pharmaceutical data in reducing hospital readmissions.

2.4.5 Summary of factors used in predicting readmissions

The literature on hospital readmissions presents a complex and diverse set of factors influencing patient readmissions, highlighting the challenges in accurately predicting these events. A critical insight from this extensive literature review is that there is a need for prediction models to incorporate a wide range of factors, from patient characteristics to healthcare data, so as to identify individuals at high risk of readmission better and effectively.

Studies consistently show that a patient's previous readmissions significantly impact their likelihood of future readmissions (Curry et al., 2005). The interplay of age and gender with readmission rates has also been a focal point, as underscored by the literature review. Notably, Soeken et al. (1991) emphasised age, length of stay, and prior hospitalisation as primary predictors of readmission. Similarly, Wennberg et al. (2006) highlighted the importance of factors like age, gender, ethnicity, the number of previous hospitalisations, and the patient's clinical condition in predicting readmissions. Further, other researchers have expanded on these findings by identifying additional critical variables, including the performance of surgery, disease complications, depression, and comorbidity (Snyders et al., 2020; Demir et al., 2009; Howell et al., 2009).

Although the predictors used for predicting hospital readmissions are numerous and diverse, they can be efficiently categorised into four primary groups for practical implementation in prediction models:

- Socio-demographic predictors such as age and gender
- Prior utilisation and cost predictors: These predictors relate to a patient's interactions with healthcare services.
- Diagnostic, health status, and functionality predictors: This category encompasses medical diagnoses, overall health conditions, and functional status such as the ability to sit and walk.
- Clinical and pharmacy predictors: This involves clinical treatments and medication usage information.

A comprehensive understanding of these categories is beneficial and practically crucial and was adopted in this minor dissertation. The next chapter gives theoretical underpinnings of the statistical learning methods used within the scope of this minor dissertation.

Chapter 3

Theory of model employed

While numerous statistical learning methods for predicting the risk of hospital readmission have been explored in the literature, as outlined in the preceding chapter, this chapter delves into the theoretical underpinnings of the statistical learning techniques employed in this minor dissertation. Understanding the theory behind these techniques is crucial for fully grasping their strengths, limitations, and application contexts in modelling readmission risk. The evaluation metrics used to assess the performance of these models are also discussed. These metrics provide quantifiable means of comparing the efficacy of different models and understanding the trade-offs among various statistical learning techniques in predicting hospital readmissions.

3.1 LACE score

Numerous clinical scoring tools and statistical learning methods have been developed and are continually evolving, playing a significant role in enhancing our understanding of hospital readmissions. Among these, the LACE score is a particularly notable tool, gaining widespread recognition in academic literature for its effectiveness (Ben-Chetrit et al., 2012; Van Walraven et al., 2010). The choice to select the LACE score as a baseline method for comparative analysis in this minor dissertation is rooted in its well-documented strengths, as extensively detailed by Van Walraven et al. (2010):

- **Simplicity**: The LACE score is characterised by its unambiguous and transparent structure, rendering it highly interpretable and user-friendly.
- **Minimal complexity**: Comprising only four elementary factors that are widely accessible to clinicians and can be determined with reliability, the LACE score epitomises parsimony.
- **Moderate discrimination and accuracy**: The LACE score demonstrates a balanced blend of discrimination and accuracy in predicting the risk of early mortality or hospital readmission.
- **Validity and transparency of derivation methods**: The methodologies underlying the creation of the LACE score have been conducted with due diligence and intellectual rigour, enhancing its credibility and acceptance in the clinical community.

The framework of the LACE score comprises four variables, specifically engineered to pinpoint patients susceptible to readmission within 30 days post-discharge (Van Walraven et al., 2010). These are:

- Length of stay (L)
- Acuity of the admission (A)
- Comorbidity of the patient (assessed via the Charlson comorbidity index score) (C)
- Emergency department utilisation spanning six months before admission (E)

Table 3.1 details the scoring framework for each variable. The cumulative LACE score, which emerges from the summation of the pertinent attribute points, serves as the final risk assessment score for the patient. This score operates on a spectrum ranging from 0 (corresponding to a modest 2.0% projected risk of readmission within 30 days) to 19 (translating to a significant 43.7% expected risk), as detailed in Table 3.2 below.

In summary, the LACE score's salient features of simplicity, accessibility, balanced

performance, and methodological integrity render it an invaluable benchmark in the landscape of readmission prediction. Its ability to distil complex patient information into a concise and actionable metric demonstrates its relevance and applicability in contemporary healthcare settings.

Attribute	**Value**	**Points**
Length of stay, ("L")	≤ 1	0
	1	1
	2	2
	3	3
	4 - 6	4
	7 - 13	5
	≥ 14	7
Acute (emergent) admission ("A")	No	0
	Yes	3
Comorbidity (Charlson comorbidity index score) ("C")	0	0
	1	1
	2	2
	3	3
	≥ 4	5
Visits to emergency department 6 months prior ("E")	0	0
	1	1
	2	2
	3	3
	≥ 4	4

Table 3.1: LACE score for the quantification of the risk of readmission within 30 days after discharge Van Walraven et al. (2010)

The application of the LACE score, as discussed in the preceding chapter, comes with intrinsic limitations. The methodology behind the LACE score, particularly, has been critiqued for its relatively limited accuracy and discriminative power (Cotter et al., 2012). Although its simple design has certain benefits, it might also

LACE score	Expected probability(%)	Derivation group	Validation group
0	2.0	0.0 (0.0 - 61.5)	0.0 (0.0 - 46.1)
1	2.5	1.4 (0.2 - 5.1)	3.0 (0.8 - 7.6)
2	3.0	2.6 (0.5 - 7.5)	2.7 (0.5 - 7.8)
3	3.5	5.6 (2.2 - 11.4)	2.5 (0.5 - 7.2)
4	4.3	3.9 (2.0 – 6.9)	2.3 (0.9 – 4.8)
5	5.1	4.4 (2.2 – 7.9)	6.7 (3.9 – 10.8)
6	6.1	4.7 (2.3 – 8.7)	4.5 (2.0 – 8.5)
7	7.3	7.6 (4.9 – 11.4)	8.5 (5.8 – 12.0)
8	8.7	6.3 (3.8 – 9.8)	8.0 (4.9 – 12.2))
9	10.3	11.7 (6.8 – 18.8)	8.7 (5.0 – 14.2)
10	12.2	14.5 (9.4 – 21.3)	13.6 (8.7 – 20.2)
11	14.4	18.6 (11.5 – 28.4)	18.1 (10.9 – 28.3)
12	17.0	20.8 (11.7 – 34.4)	10.4 (4.5 – 20.5)
13	19.8	17.3 (7.9 – 32.9)	17.4 (7.5 – 34.3)
14	23.0	28.6 (12.3 – 56.3)	36.4 (15.7 – 71.7)
15	26.6	8.3 (0.2 – 46.4)	18.8 (3.9 – 54.8)
16	30.4	50.0 (18.3 – 100)	29.4 (9.6 – 68.6)
17	34.6	33.3 (6.9 – 97.4)	42.9 (8.8 – 100)
18	39.1	100.0 (12.1–100)	–
19	43.7	0.0	–

Table 3.2: Expected and observed probability of death or unplanned readmission within 30 days after discharge, by LACE score (Van Walraven et al., 2010)

inadvertently limit its predictive effectiveness. Some scholars in the field advocate for the use of more complex statistical learning models, such as Random Forest and Gradient-Boosted Machines (Hammill et al., 2011; Cotter et al., 2012). These models include a broader array of variables, encompassing medical comorbidities and primary demographic data, and are believed to achieve a higher accuracy (Kansagara et al., 2011). Owing to their comprehensive nature, these statistical learning models are considered more adept at capturing the subtle interactions among various patient

factors. This capability potentially leads to more precise predictions in assessing the risk of hospital readmission.

This criticism of the LACE score accentuates the perennial tension between simplicity and complexity in predictive modelling. While the LACE score's streamlined design facilitates quick and transparent risk assessment, it may simultaneously limit the model's ability to capture the multifarious nature of patient readmission risks fully. Consequently, the search for a balanced model that harmoniously integrates both ease of use and predictive accuracy remains an ongoing challenge and a vibrant area of research within the healthcare analytics community. The following subsection details Logistic regression, which has been applied in literature as an attempt to overcome some of the shortcomings of the LACE score.

3.2 Logistic regression

Logistic regression, first conceptualised by Cox (1958), emerged as a specialised linear regression adaptation designed to address classification problems. Unlike its linear counterpart, logistic regression does not adhere to the same stringent assumptions concerning the distribution and relationship of independent variables. This renders it a more flexible modelling tool suitable for handling binary outcomes.

Recognised as a cornerstone in the pantheon of supervised learning methods for classification, logistic regression warrants inclusion in any comprehensive comparison of modelling techniques (Hastie et al., 2009). The theoretical framework of logistic regression outlines a mathematical relationship between a singular dependent binary variable and one or multiple independent variables, which may encompass various measurement levels such as nominal, ordinal, interval, or ratio (Cox, 1958).

Mathematically represented in Equation 3.1, the logistic regression model characterises the probability of the dependent event occurring as a linear combination of the independent variables.

$$Pr(Y = 1) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}} \tag{3.1}$$

where in readmission prediction :

- $\Pr(Y = 1)$ is the probability of patient being readmitted
- $X_1, X_2, ...X_p$ are p independent predictors
- $\beta_0, \beta_1, ..., \beta_p$ are the estimated coefficients

To interpret the meaning of the coefficient, one can express the Equation 3.1 above in terms of the odds ratio. The odds of an event refer to the probability of the event occurring relative to the probability of it not occurring, essentially comparing the chance of occurrence to non-occurrence (Hastie et al., 2009). Odds close to zero and infinity represent very low and very high readmission probabilities, respectively. The odds are expressed in Equation 3.2:

$$\begin{aligned} \frac{Pr(Y = 1)}{Pr(Y = 0)} &= \frac{Pr(Y = 1)}{1 - Pr(Y = 1)} \\ &= \exp^{(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)} \end{aligned} \tag{3.2}$$

By taking the logarithm of both sides of Equation 3.2 we arrive at the following:

$$\log \left(\frac{Pr(Y = 1)}{Pr(Y = 0)} \right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{3.3}$$

The left-hand side of Equation 3.3 is called the log odds or logit.

3.2.1 Fitting logistic regression

The model coefficients $\beta_0, \beta_1, ..., \beta_p$ in Equation 3.1 are unknown and are estimated based on the available training data. Although there is a wide range of coefficient estimation and optimisation algorithms, such as the Gradient Descent and Bayesian methods, our logistic regression models are trained using the **maximum likelihood estimation** method, a commonly used method in the literature (Hastie

et al., 2009). Maximum likelihood estimation is a method of estimating the parameters $\beta_0, \beta_1, ..., \beta_p$, such that the predicted probability of readmission for each patient using Equation 3.1 corresponds as closely as possible to the patient's observed readmission status. The likelihood function is defined as the probability of observing the set of outcomes in the dataset, given the explanatory variables. More generally, the likelihood function can be written as:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} Pr(Y = 1|X = x_i)^{y_i} Pr(Y = 0|X = x_i)^{1-y_i} \\ &= \prod_{i=1}^{n} f(x_i)^{y_i}[1 - f(x_i)]^{1-y_i} \end{aligned} \tag{3.4}$$

Where $f(x_i) = e^{\beta_0+\beta_1 X_1+....+\beta_p X_p}/(1 + e^{\beta_0+\beta_1 X_1+....+\beta_p X_p})$.

We want to find the $\beta_0, \beta_1, ..., \beta_p$ estimates that maximise the likelihood of our observed outcomes. This is equivalent to minimising the cross-entropy loss function, which is obtained by taking the negative log-likelihood of Equation 3.4.

$$CE(\beta) = -\sum_{i=1}^{n}[y_i log f(x_i) + (1 - y_i) log(1 - f(x_i))] \tag{3.5}$$

The gradient descent algorithm is used to determine the optimal set of parameters that minimise the cost function, as outlined in Equation 3.5. The impact of each independent variable on the predictions can be directly measured by its coefficient, which indicates the expected change in the dependent variable for a one-unit change in the independent variable, assuming all other variables remain constant (Hastie et al., 2009).

Logistic regression models, when developed using the maximum likelihood estimation technique, are prone to issues like over-fitting and sensitivity to sparse data (Hastie et al., 2009). Over-fitting happens when a model captures not just the underlying patterns but also the noise in the training data, which can degrade its performance on unseen data. To address over-fitting, regularisation techniques are employed (Hastie et al., 2009). These techniques involve the use of all predictors

in the model but adjust the estimated coefficients towards zero compared to those obtained by least squares estimation. This approach effectively reduces variance, helping to mitigate the risk of over-fitting and improve model generalisation (Halevy et al., 2009; Hastie et al., 2009).

3.2.2 Ridge-logistic regression

Imposing a penalty on the logistic likelihood function reduces both over-fitting and the complexity of the logistic models (James et al., 2013). Ridge-logistic regression achieves this by adding a ridge penalty to the logistic likelihood function, as shown below:

$$L(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.6}$$

where :

- L(β) is the logistic likelihood function described in Equation 3.4 above
- $\lambda \geq 0$ is the ridge penalty parameter which controls the amount of penalty imposed on the $\beta_0, \beta_1, ..., \beta_p$ coefficient parameters.

When $\lambda = 0$, there is no penalty, and the loss function in Equation 3.6 is the same as in logistic regression. However, as λ gets larger, the logistic regression coefficients $(\beta_0, \beta_1, ..., \beta_p)$ are shrunk toward zero in order to minimise the penalised loss function. Shrinking the logistic regression coefficients towards 0 reduces their variance and reduces over-fitting. However, all available variables are incorporated in the model, and hence, the ridge penalty can not be used if variable selection is preferred (Hastie et al., 2009).

3.2.3 Lasso-logistic regression

A disadvantage of ridge regularisation, as noted above, is that all predictors are included in the final model. To overcome this, the Least Absolute Shrinkage and Selection Operator (Lasso) can be used to regularise the logistic regression. In a

Lasso-logistic regression, the parameters are obtained by minimising the logistic log-likelihood function subject to a penalty, specified in Equation 3.7.

$$L(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.7}$$

where :

- L(β) is the logistic likelihood function described in Equation 3.4
- $\lambda \geq 0$ is the lasso tuning parameter which controls the amount of penalty imposed on the $\beta_0, \beta_1, ..., \beta_p$ coefficient parameters.

As with ridge regression, $\lambda = 0$ returns logistic regression, while larger values of λ cause the estimates to be shrunk toward zero. Lasso regression, therefore, helps to reduce the model complexity and multi-collinearity by shrinking some of the coefficients to zero. While Lasso regression is popular, it has certain limitations. A notable drawback is its lack of robustness when handling features that are highly correlated. In such cases, Lasso tends to arbitrarily select one variable from the correlated group and disregard the others, which can impact the model's effectiveness (Zhou, 2013). A solution to this is to use Elastic nets which are discussed in the next section.

3.2.4 Elastic net logistic regression

Elastic nets were introduced by Zou and Hastie (2005) to overcome the disadvantages of Lasso and ridge logistic regression. In an elastic net-logistic regression, the parameters are obtained by minimising the logistic log-likelihood function subject to a penalty, which is a combination of both the ridge and lasso penalties, specified in Equation 3.8.

$$L(\beta) + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{3.8}$$

where :

- $L(\beta)$ is the logistic likelihood function described in Equation 3.4
- $\lambda_1 \geq 0$ is the ridge tuning parameter
- $\lambda_2 \geq 0$ is the lasso tuning parameter

This method enhances feature selection by incorporating the ridge regression penalty to address issues of high correlation among variables, while simultaneously leveraging the Lasso penalty to ensure that only relevant variables are included in the final model. Nevertheless, logistic regression-based models are ineffective for modelling the nonlinear relationships often found in medical data (Schober and Vetter, 2021). Decision trees, discussed in the following section, address this limitation.

3.3 Decision trees

A decision tree is a classification method that divides the input space into disjoint regions, assigning a class to each region based on the majority target value of the training instances within that region, as described by (James et al., 2013; Hastie et al., 2009). James et al. (2013) further argued that decision trees more accurately reflect human decision-making compared to other statistical learning techniques like logistic regression, highlighting their intuitive alignment with human reasoning processes Demir et al. (2009). The following subsections detail the three techniques employed for conducting recursive binary splitting at each node during the construction of the decision tree.

3.3.1 Classification error

In a decision tree model, the classification error rate is determined by the proportion of training samples within a certain region that do not match the most frequently occurring class. The construction of a decision tree involves selecting splits at each stage that minimise this classification error rate, effectively reducing the rate of incorrect predictions among the observations. Classification error is defined as:

$$E = 1 - max_k(\hat{p}_{mk}) \tag{3.9}$$

where: $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class. The classification error is not sufficiently sensitive for tree-growing (Hastie et al., 2009), and different measures of node impurity, that is, how 'pure' the leaf nodes are, are often used.

3.3.2 Gini index

The Gini index, defined in Equation 3.10, is the metric used to assess the quality of splits inside each tree. It is a measure of node impurity or variability within the leaf nodes and is defined as:

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.10}$$

where: $\hat{p}_{mk}$ is the proportion of observations in outcome category $k = 1, 2 \cdots, K$ within leaf node $m = 1, 2, \cdots, J$.

From Equation 3.10, if all of the $\hat{p}_{mk}$ are close to zero or one, the Gini index takes on a small value. It is, for this reason, the Gini index is regarded as an indicator of node purity, meaning a lower value suggests that a node primarily consists of observations from one class. At each step during tree growth, we, therefore, choose the split that produces the greatest reduction in the Gini index.

3.3.3 Cross-entropy

Entropy serves as another option to the Gini index and it is presented as:

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}) \tag{3.11}$$

Since $0 \leq \hat{p}_{mk} \leq 1$, it follows that $0 \leq \hat{p}_{mk}$ log $\hat{p}_{mk}$, therefore, like the Gini index, the entropy will take on a value near zero if the $\hat{p}_{mk}$'s are all near zero or near one.

The Gini index and entropy are metrics often used to assess the effectiveness of a split in a decision tree, given their higher sensitivity to the purity of nodes compared to the classification error rate (James et al., 2013). These metrics help in identifying splits that best segregate the data into homogeneous groups or classes. While either the Gini index, entropy, or classification error rate can be applied during the tree pruning process, the classification error rate is generally favoured for optimising the predictive accuracy of the final pruned tree Hastie et al. (2009). This optimisation process results in a set of rules that can be applied to classify new data points effectively.

A primary drawback of decision trees is that minor alterations in the distribution of key features can substantially affect the model's performance (Hastie et al., 2009). The predictive performance of decision trees can thus be enhanced considerably by aggregating numerous decision trees using ensemble techniques, such as random forests, which is discussed in the next section.

3.4 Random Forest

Introduced by Breiman (2001) in 2001, the Random Forest algorithm builds upon the principles of the Classification and Regression Trees (CART) algorithm (Breiman and Friedman, 1984) and the bootstrap aggregating (bagging) technique (Leo, 1996). It generates each tree by selecting a random subset of observations from the initial dataset. An outline of a random forest algorithm for classification is outlined in Algorithm 1.

Algorithm 1 Random Forest for Classification (James et al., 2013).

For $b = 1$ to B :

(a) Draw a bootstrap sample Z^* of size n^* from the training data.

(b) Grow a random-forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum node size n_{min} is reached.

i. Select m variables at random from the p variables.

ii. Pick the best variable/split-point among the m.

iii. Split the node into two daughter nodes.

Output the ensemble of trees $\{T_b\}_1^B$

To predict at a new point x:

Classification: Let $\hat{C}_b(x)$ be the class prediction of the *bth* random-forest tree.

Then $\hat{Cf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$

- Where B : is the number of independent and identically distributed trees.

The random forest algorithm introduces two key aspects of randomness in the development of decision tree-based models (Hastie et al., 2009). These are:

- **Bootstrap sampling**: This method involves selecting data points from the training dataset randomly with replacement to form a bootstrap sample. Random forests employ numerous such samples to develop a sequence of decision tree models, each derived from a distinct subset of the training data.
- **Random feature selection**: Additionally, the random forest algorithm chooses a random subset of features for constructing each decision tree, thereby omitting certain attributes from the bootstrap sample's training data.

The use of random sampling ensures that each tree in the forest is unique, as it is trained on a distinct segment of the data. This approach effectively addresses the tendency of individual regression trees to overfit the training data. (Breiman, 2001).

The random forest algorithm benefits greatly from averaging the predicted value

of each tree. To illustrate this, consider a set of B independent and identically distributed trees $T_1, T_2, \cdots, T_B$ each with variance σ^2. The variance of their average is given by:

$$\begin{aligned} Var[\bar{T}] &= Var[\frac{T_1 + T_2 + \cdots + T_B}{B}] \\ &= \frac{\sigma^2}{B} \end{aligned} \tag{3.12}$$

Averaging the trees in a random forest mitigates the variance inherent in individual trees. Although the trees generated via the bagging process share identical distribution, they do not necessarily operate independently (James et al., 2013). The practice of selecting different features randomly for each tree further diminishes their variance, contributing to the decorrelation among the trees within the forest, enhancing the overall model's robustness (Hastie et al., 2009). Mathematically, Equation 3.13 illustrates the variance of B identically distributed random variables, each with a variance of σ^2 and pairwise correlation of ρ.

$$Var[\bar{T}] = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3.13}$$

It can be noted from Equation 3.13 that reducing the pairwise correlation between the variables and increasing the number of B reduces overall variance. This is accomplished during the tree development phase by randomly choosing the input variables (James et al., 2013).

3.4.1 Hyper-parameter tuning

The random forest algorithm has several parameters that must be tuned during training. These include:

- Minimum node size, which controls the structure of each individual tree in the forest.
- The number of trees grown in the forest.

- Number of variables to be considered at each split or sample size for training each tree.

Precise and generalisable random forest models may be created by fine-tuning the aforementioned parameters for each model until a set of parameter values that give the best model for a given data in a reasonable amount of time is found. This can be done by using grid search, that is, where a set of values for each parameter for which a search is required is specified then a model is trained for every combination of the parameter values.

As mentioned in the preceding chapter, the primary drawback of random forests is their propensity to overfit datasets with noisy classification problems (Hastie et al., 2009). Furthermore, random forests may perform suboptimally when the number of variables is large, but the fraction of pertinent variables is small (James et al., 2013). Boosting represents an alternative approach for enhancing the predictions resulting from a decision tree and can be applied to various regression or classification problems.

3.5 Boosting

Boosting is a method that involves constructing a composite model from an ensemble of weaker predictive models, typically employing decision trees (Hastie et al., 2009). These trees are developed sequentially, each building on the information gleaned from its predecessors. Training a boosted model occurs in stages, gradually refining the model by minimising a specific loss function throughout the training process (Friedman, 2001). A schematic representation of a boosting algorithm for classification is given in Algorithm 2.

Algorithm 2 Boosting for classification trees (Hastie et al., 2009).

1: Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.

2: For $b = 1, 2, ..., B$ repeat:

(a) Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r).

(b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{3.14}$$

(c) Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) \tag{3.15}$$

3: Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{3.16}$$

3.5.1 Tuning parameters

Boosting involves three key parameters that require tuning. The first is the number of trees, denoted as B. The optimal value for B is determined through cross-validation. The second parameter is the shrinkage parameter, also known as the learning rate, represented by λ. This parameter governs the pace at which the boosting algorithm learns, with common values being 0.01 or 0.001. The choice of λ is dependent on the specific problem at hand, and a very small λ may necessitate a significantly large B to achieve satisfactory results. The third parameter, represented by d, refers to the number of splits in each tree, which dictates the complexity of the boosted ensemble (Hastie et al., 2009). Often, values of $d = 1$ or 2 work well.

3.6 Model performance

In this minor dissertation, the prediction models are assessed using carefully selected evaluation metrics, especially considering the challenges of imbalanced datasets.

An imbalanced dataset is characterised by a significant discrepancy in the number of instances between classes, a common issue in hospital readmission data where one class, for example, readmitted patients, is much less frequent than non-readmitted patients. To address this, the models are evaluated using metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), Area Under the Precision-Recall Curve (AUC-PR), F1 score, Matthews Correlation Coefficient (MCC), and misclassification error derived from the confusion matrix.

Binary classification results in four possible outcomes:

Classified as Positive:

- **True Positive** (TP): Correctly identified positive cases
- **False Positive** (FP): Incorrectly identified positive cases from actual negatives

Classified as Negative:

- **True Negative** (TN): Correctly identified negative cases
- **False Negative** (FN): Incorrectly identified negative cases from actual positives

The model metrics discussed in detail in the subsection that follows summarise these outputs of a binary classifier with single numbers.

3.6.1 Accuracy

In a binary classification context, accuracy represents the proportion of correct predictions out of all predictions made. It measures the model's effectiveness in accurately classifying both true positives and true negatives. This is quantified as shown in Equation 3.17.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.17}$$

3.6.2 Precision and recall

- **Precision** is the proportion of cases correctly classified as positive. This is given by:

$$precision = \frac{TP}{TP + FP} \tag{3.18}$$

- **Recall**, also known as the true positive rate, is the fraction of actual positive cases that the model correctly identifies. It is calculated by dividing the number of true positive predictions by the total number of actual positives.

$$recall = \frac{TP}{TP + FN} \tag{3.19}$$

3.6.3 Matthews Correlation Coefficient (Absolute MCC)

The Matthews Correlation Coefficient 3.20, as defined in Equation 2, offers a metric for evaluating the effectiveness of a binary classifier in identifying true positives, false positives, true negatives, and false negatives.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.20}$$

It is termed a correlation coefficient because it reflects the degree of correlation between the actual outcomes and the predictions made by the classifier. A coefficient of 1 signifies perfect prediction accuracy, -1 denotes a classifier that consistently predicts the contrary class, and a score of 0 suggests that the classifier's performance is equivalent to making random guesses.

3.6.4 F1 score

The F1 Score is an additional measure of classification performance, calculated as the harmonic mean of precision and recall.

$$F1score = \frac{2 \times precision \times recall}{precision + recall} \tag{3.21}$$

An F1 score of 1 signifies flawless precision and recall, meaning every positive instance is identified correctly without any negative instances being incorrectly labelled as positive. If precision or recall is significantly lacking, the F1 score approaches zero, indicating diminished classification accuracy.

3.6.5 Receiver operating characteristics (ROC) curve

The Receiver Operating Characteristic (ROC) Curve graphically showcases a binary classifier's effectiveness, illustrating its capacity to differentiate between classes (Hastie et al., 2009). This curve plots sensitivity, representing the true positive rate (TPR), against 1 - specificity, also termed the false positive rate (FPR), across various classification thresholds. The area under the ROC curve (AUC) serves as a comprehensive metric of model performance, ranging from 0 to 1. A higher AUC indicates superior ability of the model to distinguish between the two classes in a binary classification problem.

3.6.6 Area under the precision-recall curve (AUC-PR)

The AUC-PR metric assesses the effectiveness of a binary classification model by measuring the area under the curve of Precision versus Recall. This curve omits the consideration of true negatives, which is significant in scenarios with imbalanced data where true negatives can significantly outnumber other outcomes, diminishing the visibility of variations in metrics such as false positives. The AUC-PR is particularly responsive to changes in True Positives, False Positives, and False Negatives, more so than the AUC. Therefore, for datasets with a high degree of imbalance, the AUC-PR is preferred over the AUC.

3.7 Summary

This chapter provides a summary of the modelling techniques used in this minor dissertation, including a detailed description of each method and the hyperparameters that were tuned during the training phase of each model. Additionally, the chapter outlines the performance metrics employed to evaluate the models. The subsequent chapter, among other topics, details the data sources utilised to compile the dataset for analysis, the process of feature selection, and the preprocessing steps applied to the dataset.

Chapter 4

Data and pre-processing

This chapter presents a comprehensive examination of the data that forms the backbone of this minor dissertation. The initial section offers an in-depth look at the dataset, including its sources, the nature of the data, its scope, and other relevant characteristics, such as the types of variables included, the period covered, and the sample size.

The chapter then progresses to a detailed exploration of the various data preprocessing techniques used in the study, highlighting their pivotal role in enhancing the effectiveness and accuracy of data analysis and model building. These techniques include:

- Data cleaning: Tackling missing values, duplicates, and errors to ensure data accuracy and reliability.
- Data transformation: Standardising or normalising data for consistency which is especially crucial for models sensitive to variable scales.
- Feature engineering: Generating new features from existing data, potentially through combining variables, creating interaction terms, or employing methods like PCA.
- Feature selection: Selecting the most relevant variables for the model to reduce complexity and improve performance using methods like correlation analysis,

wrapper methods, and filter methods.

- Handling imbalanced data: Applying techniques like oversampling, undersampling, or specialised algorithms to address imbalanced data.
- Data partitioning: Dividing the dataset into training, validation, and test sets for model evaluation and to prevent overfitting.

The chapter underscores the importance of meticulous data preparation and processing to ensure robust, interpretable, and reliable analysis.

4.1 Use of personal data declaration

It is important to clarify that this minor dissertation did not involve the use of any personal data, meaning no information that could identify individuals or is related to individual identities was employed in developing features or training the models presented.

4.2 Data used

The statistical learning models in this minor dissertation were trained using the hospital admission data from a privately insured medical aid member base in South Africa. The data covered the period from January 2021 to December 2021 and included 682381 admissions records. There were 77733 all-cause readmissions within 30 days in the dataset, representing a readmission rate of 11.4%. This is consistent with the observed readmission rates globally, which range from 10% to 25% depending on the area of study and the readmission definition used (Dreyer and Viljoen, 2019). In total, 64 distinct features (Table 4.1) were recorded from each admission.

Feature	Feature subtypes	Feature information example
Demographic features	Basic demographic information	Age
		Gender
		Race
		Marital status
		Province
	General health information	Smoking status
		Alcohol intake status
Admission and discharge information	Admission information	Reason for admission
		Hospital admission date
		Planned or unplanned admission
		Admitting doctor practice type
		Total number of procedures performed
		Hospital type (eg mental health or acute facility)
		Month of admission
		Length of stay
		Days spend in general ward
		Days spend in high care ward
		Days spend in intensive care unit
	Discharge information	Hospital discharge date
		Total cost of the admission
Clinical features	Diagnosis information	Principle diagnosis
		Major diagnostic category (MDC)
		Diagnostic related group (DRG)
		DRG severity
		ICD-10 codes
		CPT codes
Chronic information	Chronic conditions registered	Chronic indicator
		Number of chronic conditions
		Chronic condition
Prior costs and utilisation features	Amount paid	Costs
	Visits count	Utilisation

Table 4.1: Summary of features used in predicting 30-day all-cause hospital readmission risk

In accordance with the methodology outlined in Chapter 2 of the literature review, the features were chosen based on their highlighted importance in literature and organised into these categories:

- demographic information
- diagnostic, health status, and functionality data.
- prior utilisation/cost data.
- clinical data.

The following subsections describe these features and how they relate to readmission. Additionally, Table 6.3 in Appendix 3 provides comparative descriptive statistics between patients who were readmitted to the hospital within 30 days and those who were not. This comparison specifically includes variables measuring the number of visits and the amount paid.

4.2.1 Demographic features

Demographic features, including age, gender, and marital status, were integrated into our hospital readmission prediction models, aligning with existing research highlighted by Kansagara et al. (2011). Table 4.2, highlights a slight gender disparity in readmission rates. Specifically, of the 387728 female patients, 43144 (11.1%) were readmitted within 30 days. In comparison, 34589 out of 294653 male patients, representing 11.7%, were readmitted during the same period.

Variable	Variable subtype	Total number of patients	Number of patients not readmitted	Number of patients readmitted	Readmission rate (%)
Gender	Female	387728	344584	43144	11.1
	Male	294653	260064	34589	11.7

Table 4.2: Summary of the demographic features used in predicting all-cause 30-days readmission risk

Additionally, Table 4.3 below illustrates that while the overall median age of all patients is 49 years, those readmitted had a higher median age of 53 years. The

interquartile range also differed such that readmitted patients had lower and upper quartile ages of 37 and 66 years, respectively, compared to 29 and 62 years for non-readmitted patients.

Measure	Total patients	Readmitted patients	Not readmitted patients
Median (age in years)	49	49	53
Interquartile rage (IQR) (age in years)	(30 , 63)	(29 , 63)	(37 , 66)

Table 4.3: Summary statistics for age in years of readmitted and non-readmitted patients

Figure 4.1 depicts a graph comparing the number of readmitted and non-readmitted patients within 30 days of discharge across different age groups. The readmission rate was 7.5% (10640 out of 140995) for patients under 20 years and 15.3% (20370 out of 133566) for those aged 65 and above.

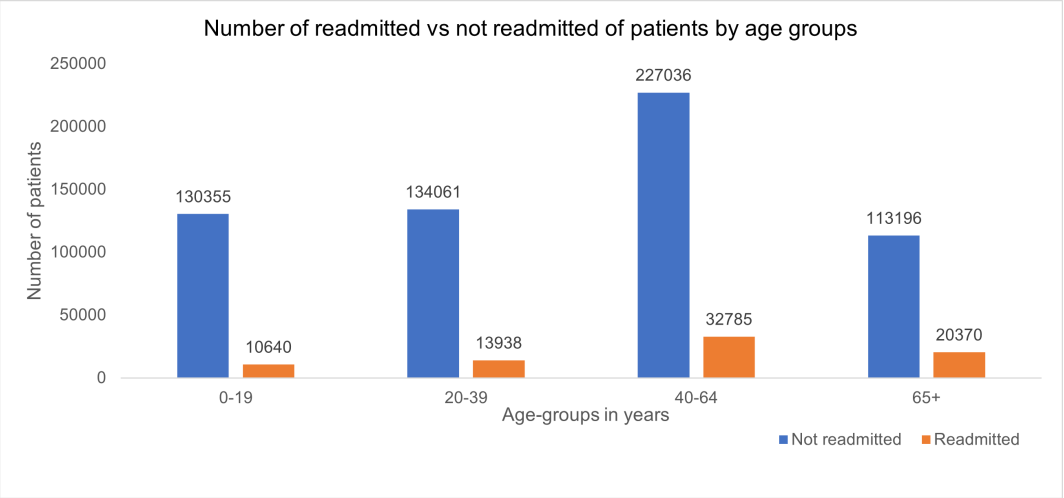


Figure 4.1: Number of patients readmitted and not readmitted within 30 days by age groups.

The next subsection provides a breakdown of readmitted patients and those not readmitted based on the initial diagnosis and the severity of the initial hospital admission.

4.2.2 Diagnosis class and severity

Table 4.4 presents a detailed admissions breakdown based on whether surgery was performed. To classify these cases, Diagnosis-Related Groups (DRGs), a system designed to group hospital cases with expected similarities in resource use and costs, were employed. DRGs consider various factors such as the principal and secondary diagnoses, surgical procedures, patient age, sex, and discharge status.

As indicated in Table 4.4, the majority of admissions in this study were medical, comprising 56%, while surgical admissions constituted 44%. The readmission rate for surgical admissions stood at 10.2%, aligning with other research findings (Tsai et al., 2014), while the medical admissions displayed a slightly higher readmission rate of 12.3%.

Variable	Variable subtype	Total number of patients	Number of patients not readmitted	Number of patients readmitted	Readmission rate (%)
Admission class	Medical	381881	334788	47093	12.3
	Surgical	300500	269860	30640	10.2
Diagnosis related group severity levels	Without co-morbidity or complications	377250	344401	32849	8.7
	With co-morbidity or complications	202227	175203	27024	13.4
	With major co-morbidity or complications	102904	85043	17861	17.4

Table 4.4: Admission class and diagnosis-related group severity levels summary for patients

Admissions were stratified into three severity levels based on clinically significant secondary diagnoses, assessing the presence and extent of complications or comorbidities. A key observation from Table 4.4 is the variation in readmission rates among these DRG severity levels. Admissions with major comorbidities or complications experienced the highest readmission rate at 17.4%, compared to 13.4% and 8.7% for admissions with minor comorbidities or complications and those without any, respectively. Consistent with the literature, such as Lefèvre et al. (2017), factors like comorbidities, complications, and the nature of the admission (medical or surgical) were all important in the modelling readmission risk.

The subsection that follows delves into the role of prior utilisation data. These factors are crucial in developing comprehensive predictive models for assessing hospital readmission risk, as evidenced by the literature reviewed in Chapter 2 of this minor dissertation.

4.2.3 Prior utilisation

The analysis dataset included comprehensive information on health service utilisation six months before readmission, an aspect proven to be a significant predictor of readmission risk in numerous studies, including studies by van Barneveld et al. (1997) and Curry et al. (2005).

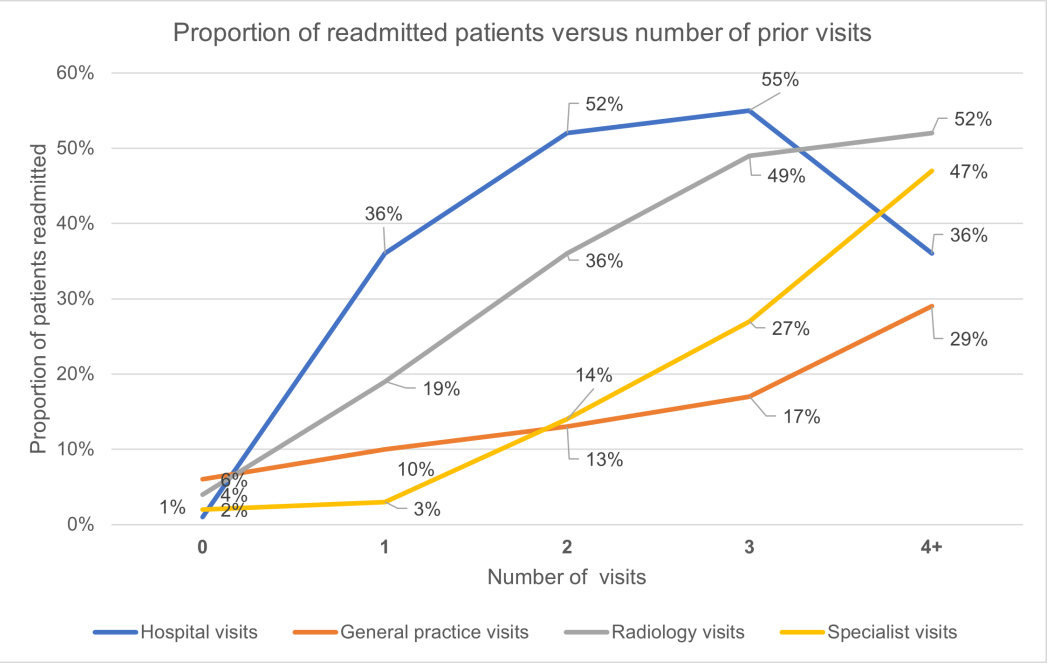


Figure 4.2: Prior utilisation history six months before the admission and their relationship with hospital readmission

Figure 4.2 presents an insightful relationship between readmission rates and the number of hospital visits prior to readmission. A clear trend is observed: the proportion of readmitted patients increases with the number of hospital visits, reaching a peak of 55% for those with three previous visits. This trend reverses for patients with four or more visits, where the readmission rate drops to 36%.

In addition, Figure 4.2 suggests a positive correlation between readmission rates and the frequency of radiology visits within the previous six months. The readmission rate hits a high of 52% for patients with at least four radiology visits. Additionally, the data show a relationship between readmission rates and prior consultations with specialists and general practitioners, with readmission rates reaching 47% and 29%, respectively, for patients with four or more visits to these healthcare providers.

The following section focuses on Major Diagnostic Categories (MDCs) as possible predictors for hospital readmissions. MDCs, which group diseases and conditions based on the affected organ system, are crucial in predicting readmissions (Snyders et al., 2020). This is due to their ability to reflect the overall nature of a patient's health condition. The section will provide a comprehensive overview of MDCs through descriptive statistics as a means of aiding the understanding of their role in readmission prediction.

4.2.4 Clinical data

Clinical data have been integrated into numerous models that predict hospitalisation (Corrigan and Martin, 1992; Curry et al., 2005). In this minor dissertation, such data were selected to encompass diagnostic codes and various medical codes utilised by physicians, medical professionals, non-physician practitioners, hospitals, ambulatory facilities, and laboratories to delineate procedures and services conducted. These diagnostic codes were employed to classify admissions into 23 mutually exclusive Major Diagnostic Categories (MDCs), each reflecting a specific diagnosis area in line with literature (Curry et al., 2005). The diagnoses within each MDC correspond to a singular organ system or aetiology and are generally associated with a particular medical speciality.

Table 4.5 illustrates the distribution of readmission rates across each MDC. A considerable proportion of patients (12.7%) are admitted due to Diseases and Disorders of the Digestive System (MDC 06), followed by those admitted for Diseases and Disorders of the Musculoskeletal System and Connective Tissue (MDC 16), which

includes conditions such as arthritis, osteomyelitis, tuberculosis of other bones, and meningococcal arthritis. The lowest readmission rate of 0.3% was recorded for burns (MDC 22).

MDC	Major Diagnostic Category (MDC) Description	Number of patients, n (%)
01	Diseases & Disorders of the Nervous System	39372 (5.8)
02	Diseases & Disorders of the Eye	43788 (6.4)
03	Diseases & Disorders of the Ear	28921 (4.2)
04	Diseases & Disorders of the Respiratory System	53896 (7.9)
05	Diseases & Disorders of the Circulatory System	52724 (7.7)
06	Diseases & Disorders of the Digestive System	86415 (12.7)
07	Diseases & Disorders of the Hepatobiliary System & Pancreas	11772 (1.7)
08	Diseases & Disorders of the Musculoskeletal System & Connective Tissue	68859 (10.1)
09	Diseases & Disorders of the Skin	33956 (5.0)
10	Endocrine	21470 (3.1)
11	Diseases & Disorders of the Kidney & Urinary Tract	36237 (5.3)
12	Diseases & Disorders of the Male Reproductive System	12903 (1.9)
13	Diseases & Disorders of the Female Reproductive System	31935 (4.7)
14	Pregnancy	8684 (1.3)
15	Newborns & Other Neonates	9734 (1.4)
16	Diseases & Disorders of Blood	6489 (1.0)
17	Neoplastic Disorders (Haematological & Solid Neoplasms)	9783 (1.4)
18	Infectious & Parasitic Diseases	58495 (8.6)
19	Mental Diseases & Disorders	32495 (4.8)
20	Alcohol/Drug Use & Alcohol/Drug Induced Organic Mental Disorders	3384 (0.5)
21	Injuries	9377 (1.4)
22	Burns	2020 (0.3)
23	Factors Influencing Health Status & Other Contacts with Health Services	19672 (2.9)

Table 4.5: Major Diagnostic Category (MDC) list and number and proportion of patients in each category

Now that a solid understanding of the data utilised in this minor dissertation has been established, the subsequent section provides a detailed exploration of the various data preprocessing techniques employed. This section outlines the transformation of raw data into a refined format suitable for practical research and modelling. It emphasises the considerable impact of these preprocessing steps on the overall

effectiveness and accuracy of the statistical learning models fitted in this minor dissertation.

4.3 Data extraction and preprocessing

As explained in Section 4.2, the initial dataset curated for this minor dissertation comprised 64 explanatory variables, a subset of which is germane to the prediction of hospital readmissions, as the integration of irrelevant features can undermine the performance of the employed learning algorithm, potentially leading to overfitting. The concept of feature selection adeptly addresses this challenge by weeding out superfluous and redundant data, leading to a decrease in computation time, an enhancement in learning accuracy, and a more nuanced understanding of the learning model (Hastie et al., 2009). The forthcoming subsections detail the feature selection techniques and data preprocessing approaches deployed in this minor dissertation.

4.3.1 Constant variance features

Firstly, highly stationary variables, that is, predictors with zero or near zero variance, were excluded to reduce sparseness and invalid features (Kuhn et al., 2013). This was done by comparing the frequency ratio of the most prevalent value to a 95% threshold as noted by Kuhn et al. (2013).

4.3.2 Correlated features

Pearson's correlation coefficients (Equation 4.1) were calculated for all pairs of numerical variables to identify multicollinearity.

$$P(X_j, X_i) = \frac{\text{cov}(X_j, X_i)}{\sigma_{x_j}\sigma_{x_i}} \tag{4.1}$$

where :

- $cov(X_j, X_i)$ is covariance between feature X_j and X_i
- σ_{x_i} is the standard deviation of feature X_i

- σ_{x_j} is the standard deviation of feature X_j.

The correlation coefficients for all numerical variables used in this minor dissertation are presented in Table 6.2 of Appendix 2. While there is no widely accepted benchmark in the literature for defining a **significant** correlation, this minor dissertation considers a coefficient above 0.70 (or below -0.70) to be significant (Chennamaneni et al., 2008). Only a few variables, specifically **ICU approved amount** and **ICU days**, slightly exceed this threshold, indicating that the majority of variables exhibit weak collinearity. For the purposes of this minor dissertation, all variables were retained; however, the interpretation of the results should carefully consider these correlations.

4.3.3 Data normalisation

In this minor dissertation, numerical variables underwent a standardisation process before being used in the modelling stages to enhance the performance and interpretability of the models. This process ensures that each feature contributes equally and meaningfully to the predictions. Standardisation is particularly important when features are measured on different scales, with some features having a much larger range of values than others. The standardised values for each variable were calculated as shown in Equation 4.2 below.

$$Z_j = \frac{X_j - \bar{X}_j}{S_{Xj}} \tag{4.2}$$

where :

- Z_j is the standardised score
- X_j is the value of the j^{th} feature
- $\bar{X}_j$ is the mean of the j^{th} feature
- S_j is the standard deviation of the j^{th} feature.

4.3.4 Categorical Variables

In this minor dissertation, categorical variables within the dataset were transformed using one-hot encoding, a widely used method for processing categorical data (Hastie et al., 2009). This technique converts a categorical variable with 'n' unique categories into 'n' separate binary variables, each representing one category. For each record, these binary variables will have a value of 1 if the category is present and 0 if it is not, thereby clearly delineating the presence or absence of each category in the dataset (Lantz, 2019).

4.4 Imbalanced data

An imbalanced data set arises when the classification categories are not approximately equally represented, often characterised by a significant disparity in the number of observations across the classes (Chawla et al., 2002). Within the context of this minor dissertation set, only 11.4% of patients were readmitted within 30 days, illustrating such an imbalance. The literature emphasises the utility of under-sampling the majority class to augment a classifier's sensitivity towards the minority class (Chawla et al., 2002). In pursuit of enhanced classification performance, this minor dissertation utilises the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). This innovative method amalgamates the over-sampling of the minority class with the under-sampling of the majority class. Specifically, SMOTE leverages the nearest neighbours algorithm to synthesise new examples of minority class data, facilitating an over-sampling strategy that accurately reflects the minority class (Chawla et al., 2002).

4.5 Data partitioning

The data utilised in this research were carefully partitioned into distinct train (ntrain=540024) and test (ntest=102357) dataset, a division fundamental to rigorous model validation. The training dataset, which comprises an expansive 540024 (84%) of the overall data, functioned as the foundational basis upon which potential

predictive associations were trained. To fortify this process, 10-fold cross-validation was employed on the training dataset, a statistical validation technique known for its efficacy in providing an unbiased assessment of a model's performance (James et al., 2013). While cross-validation is commonly done on the entire dataset, separating the test set and performing cross-validation on the training data is a strategic approach that ensures model robustness and prevents information leakage (Brownlee, 2020). This method enhances the model's ability to generalise to new data, leading to more reliable and accurate predictive performance.

To implement 10-fold cross-validation, the training data is segmented into ten equal fractions or folds. The model is then trained and validated ten times, each utilising nine folds for training and the remaining one for validation (James et al., 2013). The iterative nature of this process allows for the calculation of performance measures such as accuracy and error rate across multiple partitions. The mean performance across all iterations provides a robust indication of the model's predictive aptitude, thereby facilitating the fine-tuning of hyperparameters.

The residual 15% of the data, delineated as the test set, was harnessed as an independent and previously unseen set for evaluating the performance of each algorithm examined in this minor dissertation. The primary function of this test set is to gauge the models' generalisability, ensuring that their fit to the training data translates into competent performance on new, unseen data (James et al., 2013). It is an impartial benchmark, appraising how effectively the models are expected to perform in real-world applications where the outcomes are unknown. This training structure, validation (achieved through cross-validation), and testing constitute a sophisticated and widely acknowledged methodology conducive to the realisation of reliable and robust model development (Hastie et al., 2009).

4.6 Conclusion

Upon the successful completion of data extraction and preprocessing stages, the progression of this minor dissertation led to the training of diverse models targeted

at predicting hospital readmissions. The next chapter delineates the findings of these modelling techniques, encapsulating not only the obtained results but also the model performance.

Chapter 5

Results and discussion

This chapter presents the comprehensive outcomes and insights from this minor dissertation's analytical process. It commences by addressing the initial objective of this minor dissertation through an examination of the predictability of 30-day hospital readmission using various statistical learning methods. Eight statistical learning methods were developed, their respective performances were compared, and the best-performing method was selected. After this, the latter section of the chapter pivots its focus towards profiling important features integral to predicting hospital readmissions, specifically within the context of the privately insured population in South Africa.

5.1 Baseline model: LACE score results

The LACE score was computed for each admission in the training and testing datasets. The findings are summarised in Table 5.1 and Table 5.2, respectively, and they reveal that the overall proportion of readmitted patients remains consistent between the two datasets, at 11.4% and 11.3%, respectively. Moreover, the readmission rates observed escalate markedly with higher LACE scores, starting from 7.3% for patients with a LACE score of zero and reaching as high as 70.2% for those with the maximum LACE score of 19 in the training dataset. A similar pattern is evident in the testing dataset as well.

LACE score	Number of patients	Number of patients with no 30-day readmission	Number of patients with a 30-day readmission	Observed readmission proportion (%)
0	56361	52226	4135	7.3%
1	47248	43260	3988	8.4%
2	68384	62860	5524	8.1%
3	75060	69586	5474	7.3%
4	78345	70535	7810	10.0%
5	59245	52486	6759	11.4%
6	36093	32035	4058	11.2%
7	47303	40821	6482	13.7%
8	21864	19141	2723	12.5%
9	21522	18030	3492	16.2%
10	15017	12592	2425	16.1%
11	12224	10027	2197	18.0%
12	9936	7900	2036	20.5%
13	7931	6331	1600	20.2%
14	7017	5311	1706	24.3%
15	5086	3721	1365	26.8%
16	3640	2655	985	27.1%
17	3055	2171	884	28.9%
18	1943	1318	625	32.2%
19	2750	819	1931	70.2%
Total	**580024**	**513825**	**66199**	**11.4%**

Table 5.1: LACE scores and number of patients readmitted to the hospital within 30 days in the training data set (ntrain= 580024)

LACE score	Number of patients	Number of patients with no 30-day readmission	Number of patients with a 30-day readmission	Observed readmission proportion (%)
0	10146	9407	739	7.3%
1	8376	7653	723	8.6%
2	12150	11210	940	7.7%
3	13150	12191	959	7.3%
4	13534	12223	1311	9.7%
5	10508	9277	1231	11.7%
6	6328	5652	676	10.7%
7	8632	7479	1153	13.4%
8	3784	3301	483	12.8%
9	3824	3205	619	16.2%
10	2664	2238	426	16.0%
11	2109	1723	386	18.3%
12	1715	1379	336	19.6%
13	1376	1093	283	20.6%
14	1202	907	295	24.5%
15	897	658	239	26.6%
16	621	459	162	26.1%
17	529	377	152	28.7%
18	355	257	98	27.6%
19	457	134	323	70.7%
Total	**102357**	**90823**	**11534**	**11.3%**

Table 5.2: LACE scores and number of patients readmitted to the hospital within 30-days in the testing data set (n= 102357)

Consistent with existing literature (Donzé et al., 2013), the LACE scores were further segregated into three categories: LACE scores of 0-4 points were deemed as low risk for readmission, 5-6 points as intermediate risk, and seven or more points as high risk. In the training dataset, encompassing 580024 patients, 325398 (56.1%) were classified as low risk, 95338 (16.4%) as intermediate risk, and 159288 (27.5%) as high risk for 30-day readmission, as depicted in Table 5.3.

LACE score	Risk category	Patients in each category	Number of patients with no 30-day readmission	Number of patients with a 30-day readmission	Observed readmission rate in each category (%)
0-4	Low	366731	336324	30407	8.3%
5-6	Intermediate	107385	95192	12193	11.4%
>=7	High	179433	147421	32012	17.8%
Total		**653549**	**578937**	**74612**	**11.4%**

Table 5.3: LACE score summary in each risk category and observed number of patients readmitted within 30 days of hospital discharge in the training dataset (n=653549)

Similarly, the testing dataset, which also includes 102357 patients, categorised 57356 (56.0%) as low risk, 16836 (16.4%) as intermediate risk, and 28165 (27.5%) as high risk for 30-day readmission as shown in Table 5.4.

LACE score	Risk category	Patients in each category	Number of patients with no 30-day readmission	Number of patients with a 30-day readmission	Observed readmission rate in each category (%)
0-4	Low	57356	52684	4672	8.1%
5-6	Intermediate	16836	14929	1907	11.3%
>=7	High	28165	23210	4955	17.6%
Total		**102357**	**90823**	**11534**	**11.3%**

Table 5.4: LACE score summary in each risk category and observed number of patients readmitted within 30 days of hospital discharge in the test dataset (n=102357)

Table 5.3 also reveals a pattern of escalating readmission rates across different categories. It shows an 8.3% risk of readmission for those classified as low risk, an 11.3% risk for the intermediate group, and the highest risk of 17.9% for those categorised as high risk. A similar pattern can be seen in the test dataset in Table 5.4.

The results above can be represented via a confusion matrix, as illustrated in Table 5.5 for the testing dataset. A confusion matrix is a unique form of a contingency table that enables the visualisation and summarisation of a binary classifier's outcomes (Hastie et al., 2009). The confusion matrix presented in Table 5.5 shows that, in the test dataset, the LACE score identified 28165 patients as being at high risk of readmission within 30 days of hospital discharge. Among these individuals, 4955 were readmitted, while 23210 were not. Furthermore, the LACE score indicated that

74192 patients had a low to intermediate risk of readmission; out of these, 67613 were not readmitted, but 6579 were. While the overall misclassification error rate for the LACE score is relatively low at 29.1%, the class-specific error rate among patients who were readmitted is considerably high at 57.0%.

		Actual outcome		**Total number of patients**
		Number of patients with a 30-day readmission	Number of patients with no 30-day readmission	
LACE score outcomes	High risk	4955	23210	**28165**
	Low and intermediate high risk	6579	67613	**74192**
Total number of patients		**11534**	**90823**	**102357**
Misclassification error		**57.0%**	**25.6%**	**29.1%**

Table 5.5: Confusion matrix based on LACE score risk categories of patients in the test data set for predicting readmission within 30 days after discharge (ntest=102357

Additionally, performance metrics, including Recall, Precision, Specificity, F1-score, and MCC, which are crucial for the comparative analysis of the statistical learning models in this minor dissertation, were calculated as detailed in Chapter 4. The results are summarised in Table 5.6. These results align with those observed in the literature (Van Walraven et al., 2010; Donzé et al., 2013; Morgan et al., 2019).

Model	**Dataset**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**MCC (%)**
LACE score	Testing	70.9	43.0	17.6	25.0	12.3

Table 5.6: The performance metrics of the LACE score for predicting 30-day all-cause hospital readmissions in privately insured South African patients in the test datasets.

The LACE score, which uses four critical factors, namely, length of stay, acuity of admission, patient comorbidity, and emergency department utilisation (measured by visits in the previous six months) (Van Walraven et al., 2010), demonstrated moderate effectiveness in predicting hospital readmission risk. As detailed in Table 5.6, the LACE score achieved an accuracy rate of 70.9% in the test dataset, accompanied by a precision rate of 43.0%, a recall rate of 17.6%, an F1-score of 25.0%, and

a Matthews Correlation Coefficient (MCC) of 12.3%. Building on this foundational analysis, the following sections explore additional predictors introduced in Chapter 4 and assess their significance in predicting hospital readmissions within 30 days.

5.2 Logistic regression models results

Logistic regression models were implemented in R (Ripley et al., 2001) using the H2O package (Malohlava et al., 2016), a powerful and efficient Java-based interface that supports both local and cluster-based deployments. For each hyperparameter, a grid search was employed to fine-tune the models by exploring a range of values.

A logistic regression was initially fitted by maximising the log-likelihood function in Equation 3.4. Subsequently, by adding a penalty term to the loss function, penalised logistic regression models were also fitted, namely Ridge, LASSO, and Elastic net logistic regressions. The following three subsections illustrate the cross-validation results of the degree of shrinkage (λ) imposed on the regression coefficients for these penalised regression models. Lastly, the model performance of all these logistic models on the testing dataset is discussed.

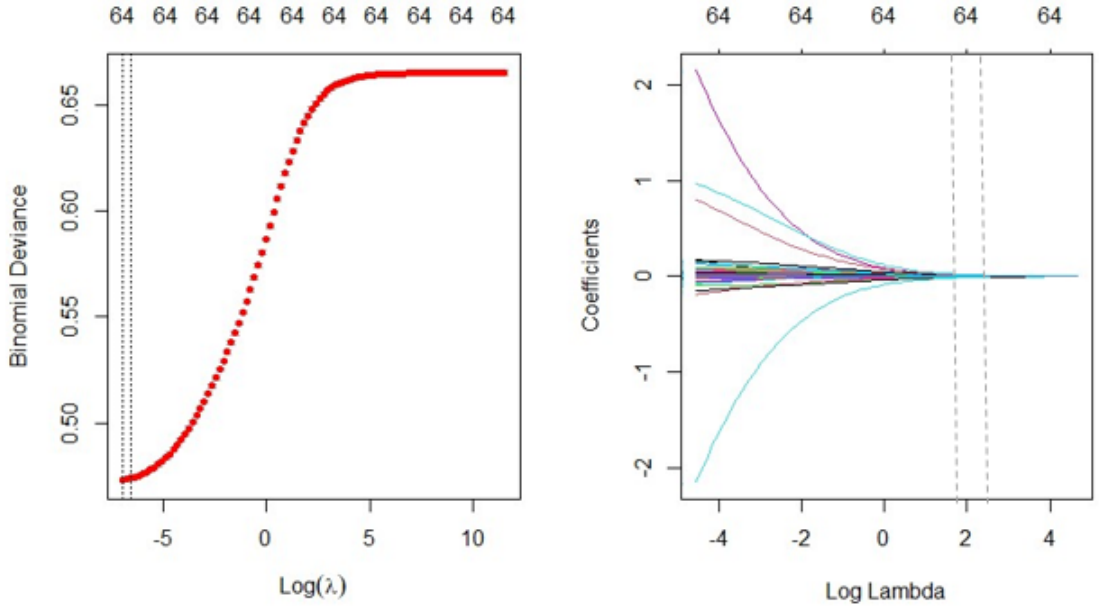
5.2.1 Ridge regression

The ridge logistic regression penalty parameter λ in Equation 3.6 was determined through 10-fold cross-validation to ensure good out-of-sample performance. Figure 5.1 shows the binomial deviance errors of the 10-fold cross-validation results for different values of $\log(\lambda)$ and the corresponding ridge logistic regression coefficient paths.

The plot on the left depicted in Figure 5.1 illustrates the 10-fold cross-validated binomial deviance as a function of $\log(\lambda)$ for the ridge-regularised model. The numbers at the top of the plot denote the number of predictors used by the model, which, in this case, are 64 variables. The two dotted vertical lines on the left plot represent the value of $\log(\lambda) = -3$ ($\lambda = 0.001$), leading to the minimum binomial deviance and

the optimal value of log(λ) = -2.838 ($\lambda = 0.001451$), which results in the minimum cross-validated error.

Figure 5.1: Explanatory plots for cross-validated errors and Ridge coefficients paths



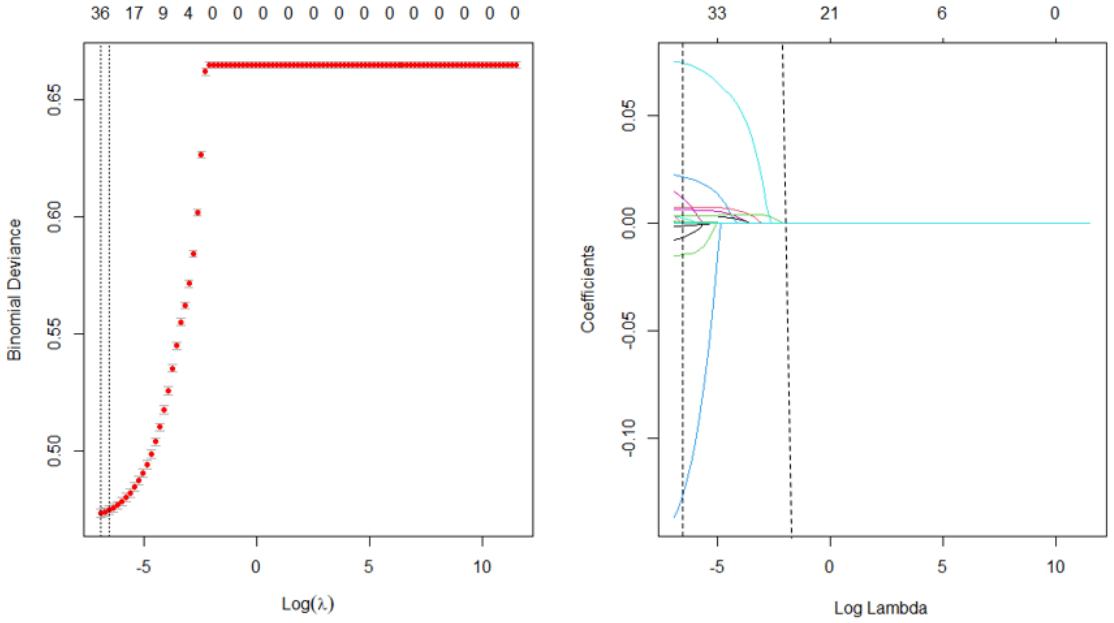
The plot on the right in Figure 5.1 visually displays how increasing lambda shrinks the coefficients for each variable. It can also be seen from Figure 5.1 that the higher the lambda value, the more the coefficients are shrunk toward zero. However, they will never be precisely zero. This particular characteristic might not be desirable if we want the model to select important variables. In the subsequent sections, LASSO regression is introduced and explored as a methodological approach to address and potentially overcome this limitation.

5.2.2 LASSO regression

LASSO regression was also implemented, and the results presented in Figure 5.2 show the 10-fold cross-validation results for the optimum value of lambda and the coefficient paths. As with ridge regression, the predictors were standardised before optimising the lasso loss function in Equation 3.7. The plot on the left in Figure 5.2 indicates an optimal log(λ) value of -2.434 ($\lambda = 0.00368$). The plot on the right in Figure 5.2 illustrates that as the lambda value increases, more coefficients will

be shrunk towards or, in some instances, reach a zero value. Moreover, when the lambda was set to its optimal value, the model used roughly 36 predictors from a total pool of 64.

Figure 5.2: Explanatory plots for cross-validated errors and Lasso coefficients paths



The coefficients of the following features were shrunk to precisely zero and were removed as potential features for lasso regression.

- The neonatal ICU claimed amount
- The number of radiology visits in the last 60 days
- The number of hospital visits in the last 60 days
- The number of times the patient visits a clinic in the last 60 days
- The number of times the patient consulted a nurse in the last 60 days
- Whether the patient is the main member or a dependent

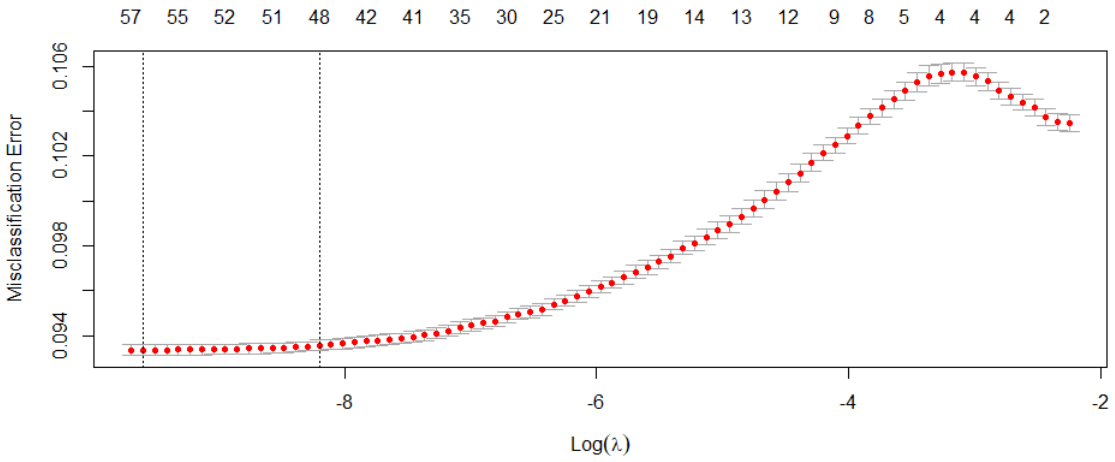
When interpreting the results above, it is crucial to remember that the lasso penalty tends to pick one variable at random when predictor variables are correlated. Consequently, the abovementioned variables might still possess predictive power but

are correlated with other variables. The Elastic net, discussed in the subsequent subsection, amalgamates the benefits of LASSO and Ridge regression methods.

5.2.3 Elastic net logistic regression

As indicated in Equation 3.8, Elastic net regularisation combines LASSO and Ridge penalties. The first penalty, denoted by α, controls the elastic net distribution between the $l1$ and $l2$ norms.

Figure 5.3: Explanatory plots for Elastic net cross-validated errors



The 10-fold cross-validation results are shown in Figure 5.3 and indicate that the value of log λ is -8.345, resulting in the minimum cross-validated misclassification error. At this optimal value of λ, the model will use approximately 48 out of the 64 predictors. The following subsection describes and discusses the model performance for all four logistic regression models trained in this minor dissertation.

5.2.4 Logistic regression models results

In light of the imbalance dataset, a comprehensive evaluation of both logistic regression and penalised models was undertaken utilising a variety of metrics. The estimated values of unstandardised and standardised $\beta_0, \beta_1, ..., \beta_p$ coefficients are for a logistic regression are shown in Table 6.1 in the Appendix. Performance metrics such as the AUC, AUC-PR, F1 score, MCC, and the misclassification error were

also calculated for the training and test datasets and results are presented in Tables 5.7 and 5.8, respectively.

Model	AUC (%)	AUC-PR (%)	F1 score (%)	MCC (%)	Misclassification error (%)
Logistic	90.34	44.03	51.07	49.17	17.55
Ridge	90.52	44.24	51.36	49.37	17.28
Lasso	90.63	44.26	51.93	50.09	16.96
Elastic-net	91.03	45.91	51.9	50.51	16.66

Table 5.7: The performance results of various logistic regression models for predicting readmission within 30 days after discharge on the training data set (n=580024)

The performance outcomes of the logistic models on the training dataset are notably similar, with minor variances contingent on the specific metric employed, as depicted in Table 5.7. Nevertheless, the Elastic Net model exhibited marginally superior performance, evidenced by its AUC (91.03%), AUC-PR (45.71%), F1-score (51.9%), MCC (50.51%), and misclassification errors at 16.66%.

Similarly, in the test data set, the performance outcomes of the logistic models were consistent. The Elastic net model had the best overall misclassification error of 16.45%. In comparison, both the Lasso and Ridge regression had a slightly higher error of 17.33% and 17.96%, respectively, and lastly, unpenalised logistic regression had the highest overall misclassification error of 17.96%.

Model	AUC (%)	AUC-PR (%)	F1 score (%)	MCC (%)	Misclassification error (%)
Logistic	90.26	44.02	51.01	49.08	17.96
Ridge	90.38	44.24	51.11	49.09	17.95
Lasso	90.67	45.81	50.77	49.14	17.33
Elastic-net	91.07	45.86	52.04	50.58	16.45

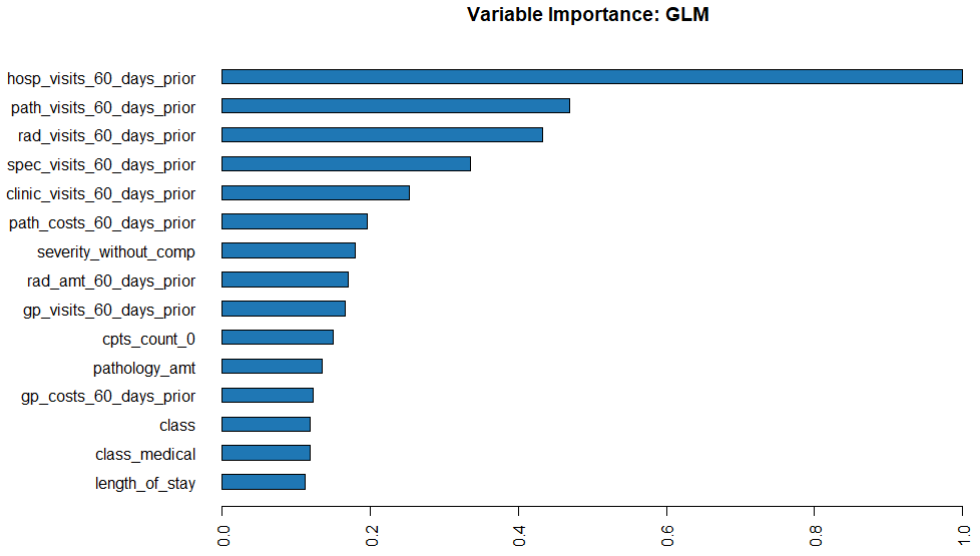
Table 5.8: The performance results of various logistic regression models for predicting readmission within 30 days after discharge on the test data set (ntest=10235)

5.2.5 Variable importance

The identification of significant predictors within a statistical learning model holds paramount importance. Such identification aids in evaluating the model's validity based on domain knowledge, uncovering new insights, and potentially improving data collection methodologies, among other benefits. Breiman (2001) introduced variable-importance measures for random forests, facilitating the assessment of covariates' contribution to a prediction model's accuracy. The fundamental concept measures the change in a model's performance when the influence of a selected explanatory variable or a group of variables is removed. If a variable is significant, the model's performance is expected to degrade upon permutation of that variable's values—the more substantial the alteration in performance, the greater the variable's influence.

The Elastic net model emerged as the top-performing model within the logistic family. This model identified the predictors displayed in Figure 5.4 as the most important factors in predicting the risk of hospital readmission within 30 days. These predictors encompass the patient's prior utilisation history, such as the frequency of hospital visits and engagements with pathology laboratories, radiology, specialist doctors, and clinics in the preceding 60 days.

Figure 5.4: Elastic Net variable importance plot



Mirroring the LACE score, the length of stay and the number of hospital visits in the six months before admission emerged as important predictors. The following subsection focuses on tree-based machine-learning techniques.

5.3 Tree-based models results

The statistical learning models for predicting hospital readmission within 30 days were developed based on various factors, including clinical diagnoses and drug prescriptions prior to patient discharge, as detailed in Chapter 4. Tree-based statistical learning models, specifically random forest, decision tree, and gradient boosting machine (GBM), were trained, and their performance was evaluated on both the training and test datasets.

A comprehensive grid search, encapsulating all potential hyper-parameter combinations, was performed using cross-validation on the training set. Ten-fold cross-validation was employed to select the optimal hyper-parameter and facilitate internal validation.

5.3.1 Hyper-parameters

In this minor dissertation, the critical parameters for the tree-based algorithms were categorised into two distinct groups: tree-specific parameters and ensemble parameters. Tree-specific parameters directly influence the behaviour and performance of the individual trees being built. The following are essential tree-specific parameters utilised in this study:

- **Tree depth** (max_depth): This parameter determines the maximum depth of each tree. Deeper trees can capture more complex patterns but also heighten the risk of overfitting.
- **Minimum samples split** (min_samples_split): This parameter specifies the minimum number of samples required in a node for it to be considered for splitting. Higher values help prevent the model from learning overly specific

relationships from the training data, thereby controlling overfitting.

- **Minimum samples leaf**(min_samples_leaf): Sets the minimum number of samples that must be present at a leaf node. Smaller leaves may cause the model to capture noise in the training data, while larger leaves generally lead to a more generalised model.

Ensemble parameters, on the other hand, impact the overall training process and the performance of the models. These parameters manage aspects like the number of trees in the model, the rate at which the algorithm learns, and the method for combining the results from individual trees. Key ensemble parameters include:

- **Learning rate**(learning_rate): This parameter governs the rate at which the model learns. Lower values lead to a more gradual learning process, potentially enhancing performance but requiring a more significant number of trees to capture data complexities. Conversely, higher values accelerate learning at the risk of overfitting
- **Number of trees**(n_estimators): This represents the total number of trees to be built, equating to the number of boosting stages. While more trees can increase accuracy, it also raises the risk of overfitting and the computational burden.
- **Subsample**: Denotes the sample fraction used for fitting each base learner. Values below 1.0 can reduce variance and increase bias, introducing a method known as "Stochastic Gradient Boosting".
- **Max features**: This parameter determines the number of features to consider for the best split. Appropriate tuning can aid in reducing overfitting and enhancing model performance, especially in high-dimensional data scenarios.

These parameters were tuned using cross-validation to derive the optimal model configuration, with the optimal values presented in Table 5.9.

	Statistical learning method		
Parameter	Random forest	Decision tree	Gradient boosting machine (GBM)
max_depth	5	5	5
min_samples_split	10	10	10
min _samples_leaf	5	5	5
min split improvement	0.001	0.00001	0.00001
learning rate	0.22	0.1	0.52361
n_estimators	500	1	500
Subsample	n/a	n/a	0.56125
max features	20	20	20
balance classes	FALSE	FALSE	FALSE

Table 5.9: he optimal parameters for the decision tree, random forest and GBM model obtained using a 10-fold cross-validation

The hyperparameters leading to the best cross-validation performance were chosen for each tree-based model. Tables 5.10 and 5.11 display the performance metrics of these models on the training and test datasets, respectively. The Random Forest model exhibited slightly superior performance on the training dataset, as demonstrated by its metrics in Tables 5.10. The GBM model achieved the lowest overall misclassification error of 9.87% in the training set.

Model	**AUC (%)**	**AUC -PR(%)**	**F1 score(%)**	**MCC (%)**	**Misclassification error (%)**
Decision tree	94.34	59.58	62.65	59.28	12.40
Random forest	95.81	68.50	65.42	62.84	10.51
GBM	95.13	67.03	65.37	62.51	9.87

Table 5.10: The performance results of tree-based models for predicting readmission within 30 days after discharge on the training data set (ntrain=580024)

Similarly, on the test dataset, the tree-based models showed consistent performance, with minor deviations in specific metrics, as shown in Table 5.11.

Model	AUC (%)	AUC -PR(%)	F1 score(%)	MCC (%)	Misclassification error (%)
Decision tree	93.22	58.12	60.77	59.26	12.49
Random forest	95.21	67.93	64.49	64.66	9.66
GBM	95.01	65.72	65.28	63.46	9.81

Table 5.11: The performance results of tree-based models for predicting readmission within 30 days after discharge on the test data set (ntest=102357)

Figure 5.5 presents the ROC curves for the tree-based methods evaluated in this minor dissertation using the test dataset.

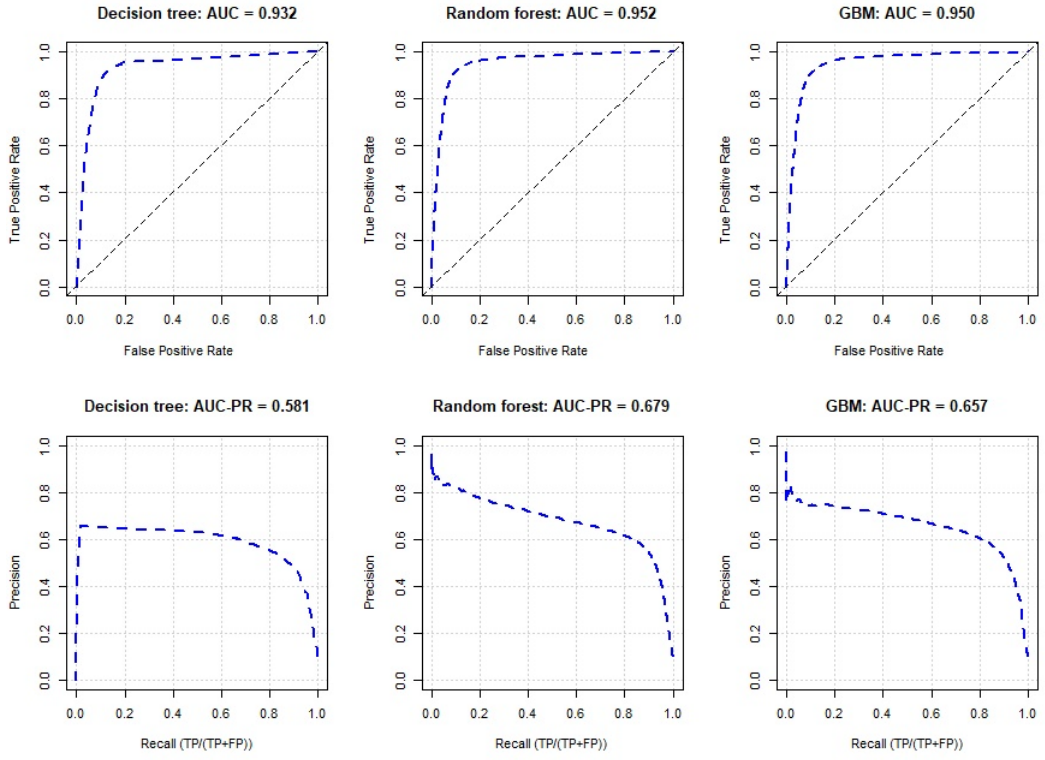


Figure 5.5: AUC and AUC-PR curves for the tree-based methods on test data set

Notably, the Random Forest model achieved the highest AUC and AUC-PR values, substantiating its ability to discriminate between readmission and non-readmission patients. The marginal discrepancy observed between the AUC and AUC-PR metrics for the GBM and Random Forest models highlights that, although the Random

Forest model exhibits a marginal edge in these metrics, the GBM model similarly possesses a significant discriminative capacity in the predictive modelling for hospital readmissions.

5.3.2 Variable importance for random forest model

Figure 5.6 shows the top 20 important predictors from the random forest model for predicting readmission. The predictors are arranged in descending order of importance, with percentage values representing each predictor's proportional importance scaled to 100%. The plot suggests that the model's most important predictor is the expenditure on hospital expenses in the preceding 60 days. Most prior utilisation features are critical for predicting hospital readmissions, a finding that aligns with existing literature (Soeken et al., 1991; Wennberg et al., 2006). Other important variables include the base DRG, the patient's age, length of stay, number of diagnoses, and the admitting doctor.

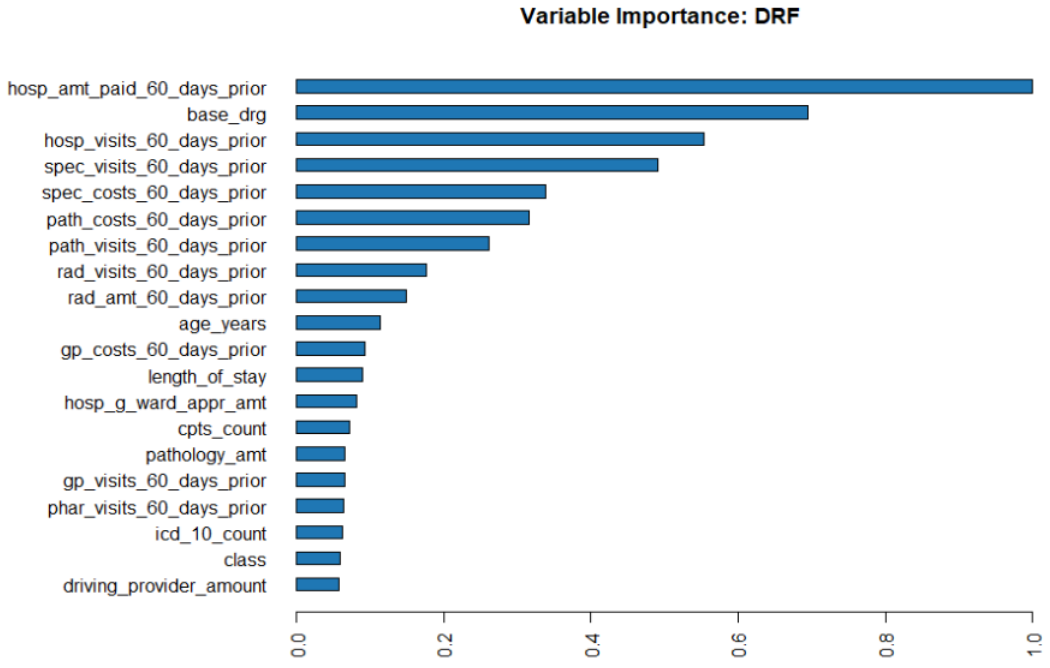


Figure 5.6: Variable importance of Random forest model

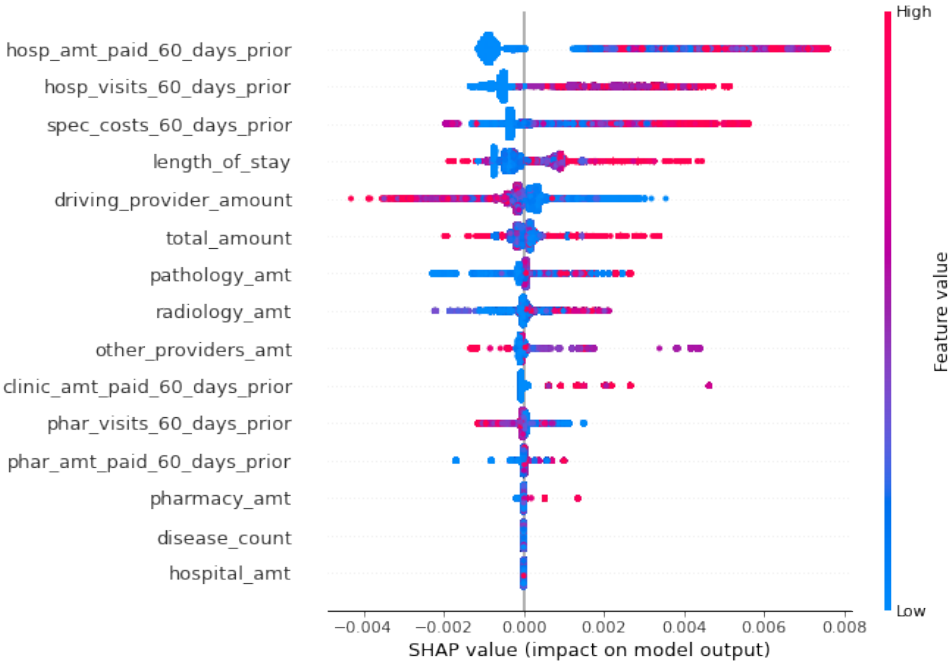
The following subsection aims to investigate these highly important predictors through

a novel technique called Shapley Additive explanation.

5.3.3 Feature impact

Figure 5.7 showcases the SHapley Additive exPlanation (SHAP) values for key predictive factors in the top-performing random forest model, blending feature significance with their effects on the model's outcomes. In the SHAP summary plot, each dot represents a data point from the training set, with the colour indicating the feature's value from low to high. This visualisation elucidates how each attribute contributes positively or negatively to the model's predictions and illustrates the distribution of these feature values(Lundberg and Lee, 2017).

Figure 5.7: SHAP summary plot



Similar to the results in Figure 5.6, Figure 5.7 illustrates that the total amount spent on previous hospital admissions within 60 days is the most important feature used in the model. Furthermore, it can be observed that the lower the spending on prior hospital admissions in the last 60 days, the lower the risk of readmission. Based on the distribution of the feature representing spending on hospital admissions in the previous six months - as shown in Figure 5.7, most patients have negative SHAP

values.

5.3.4 Explaining individual predictions

To demonstrate the impact of explanatory variables on readmission predictions, this section examines two arbitrary patients from the unseen test dataset, shown in Figures 5.8 and 5.9. Each case employs SHAP values, represented as vectors that either increase (positive value) or decrease (negative value) the likelihood of readmission. These vectors counterbalance each other to settle at the model's prediction for each patient.

Figure 5.8: SHAP explanation force plots for Patient A

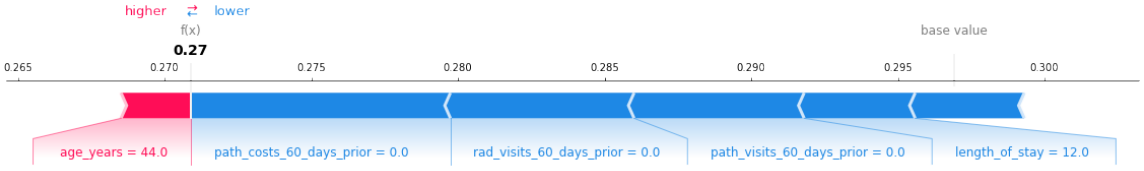
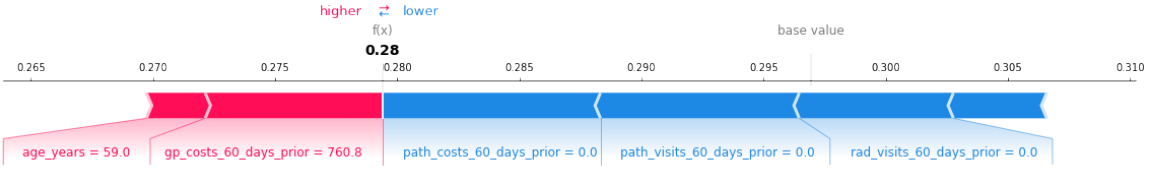


Figure 5.9: SHAP explanation force plots for Patient B



The baseline prediction shows that the dataset's average predicted probability of readmission is approximately 0.297. In contrast, the final predicted readmission risks for Patients A and B are computed to be 0.27 and 0.28, respectively. Notably, the absence of visits and costs associated with pathology and radiology services contributes to a reduction in the predicted risk of readmission for both patients relative to the baseline probability. Nevertheless, Patient B's prediction is marginally higher than that of Patient A, which can be attributed to factors such as advanced age and significant healthcare expenditure on general practitioner services in the

sixty days preceding the prediction. This examination highlights how individual patient characteristics and healthcare interactions affect their final prediction of hospital readmission risk.

5.4 Neural network performance results

To predict the risk of all-cause hospital readmission within 30 days, a multi-layer feed-forward artificial neural network was trained using H2O (Malohlava et al., 2016). This minor dissertation examined key parameters integral to the functioning and performance of neural networks, such as:

- **Activation function**: Determines neuron activation, introducing non-linearity into the model. This allows the network to learn and represent complex data patterns.
- **Hidden layers**: Located between input and output layers, these layers are pivotal in feature extraction and processing, enabling the network's learning and predictive capabilities.
- **Epochs**: Each epoch represents a full pass of the training dataset through the network. The number of epochs affects learning depth, with more epochs potentially improving learning at the risk of overfitting.
- **Rho**: A hyperparameter governing the decay rate of the moving average of squared gradients crucial for adjusting the learning rate and ensuring training convergence.
- **Epsilon**: A small constant ensuring numerical stability in network algorithms.
- **Rate**: Refers to the learning rate, dictating the step size in optimisation. Higher rates can accelerate convergence but risk overshooting the optimal point, while lower rates might slow down the learning process.
- **Rate annealing**: This technique gradually decreases the learning rate, starting higher for rapid initial progress and reducing it to fine-tune model weights, aiding in stable and optimal convergence.

These parameters were fine-tuned for optimal model performance, with their optimal values presented in Table 5.12.

activation	hidden	epochs	rho	epsilon	rate	rate_annealing
Rectifier	128,128,128	100	0.99	1.00E-08	0.01	0.000002

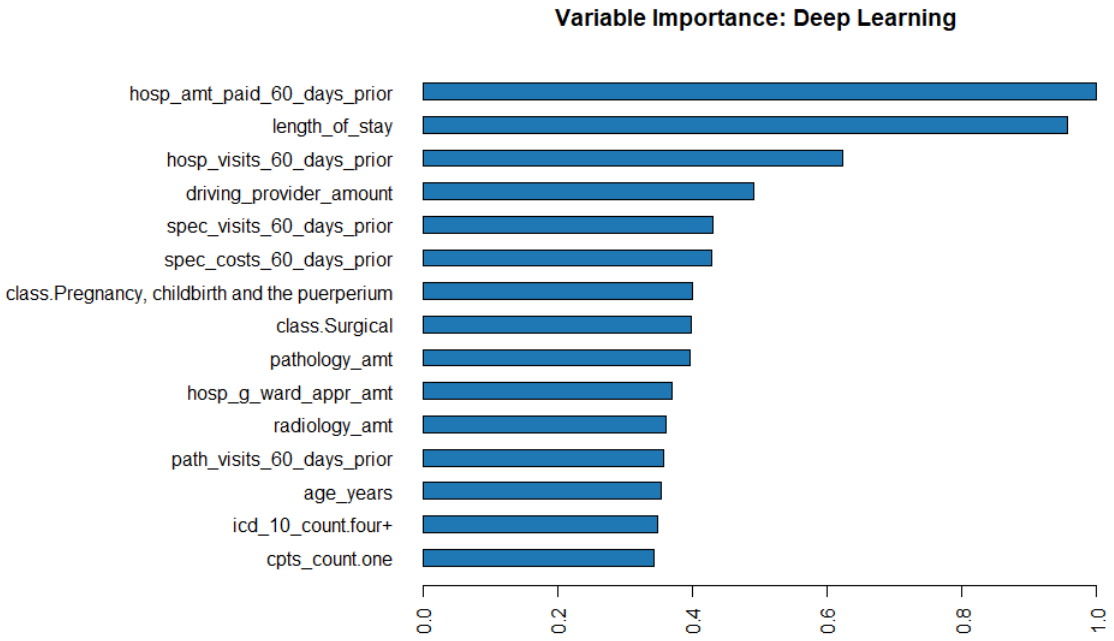
Table 5.12: hypermeters for the best-performing neural network model

Based on these parameters, the final model's performance is detailed in Table 5.13. The results demonstrate a remarkable ability to discriminate between patients at high and low risk of 30-day post-discharge hospital readmission. The ROC curves exhibited an AUC of 95.17% on the unseen (test) dataset. This model outperformed tree-based methods in AUC-PR, F1-score, and MCC metrics on the test dataset. However, it is noteworthy that while the neural network achieved the misclassification error of 10.5% on the test dataset, Random Forest and Gradient Boosting Machine (GBM) methods reported slightly lower errors, with 9.66% and 9.81%, respectively on the same dataset.

Dataset	AUC (%)	AUC-PR (%)	F1 score (%)	MCC (%)	Misclassification error (%)
Training data	96.07	66.88	68.26	66.65	8.93
Unseen test data	95.17	64.51	65.22	63.72	10.5

Table 5.13: The performance results of the deep neural network model for predicting readmission within 30 days after discharge on the training and unseen test data set

The neural network model identified key variables for hospital readmission, as shown in Figure 5.4. These include length of stay, previous utilisation metrics, and admission type, focusing on surgical and pregnancy-related cases. Additionally, patient-specific characteristics such as age and diagnosis, categorised by ICD-10 codes, were important factors that aligned with existing literature findings Holloway et al. (1988); Curry et al. (2005); Corrigan and Martin (1992).



5.4.1 Conclusion of results

It is important to note that all the statistical learning methods examined in this minor dissertation performed better than the standard LACE score method, as shown in Table 5.14.

Model	AUC (%)	AUC -PR(%)	F1 score(%)	MCC (%)	Misclassification error (%)
LACE score	70.90	30.22	25.00	12.30	29.10
Logistic	90.26	44.02	51.01	49.08	17.96
Ridge	90.38	44.24	51.11	48.09	17.95
LASSO	90.67	45.81	50.77	49.14	17.33
Elastic-net	91.07	45.86	52.04	50.48	16.45
Decision tree	93.22	58.12	60.77	59.26	12.49
Random forest	95.21	67.93	64.49	64.66	9.66
GBM	95.01	65.72	65.28	63.46	9.81
Neural network	95.17	64.51	65.22	63.73	10.50

Table 5.14: The performance results of all the statistical learning methods and LACE score for predicting readmission within 30 days after discharge on the test data set (ntest=102357)

While the neural network achieved minimally better results than GBM on metrics such as AUC and MCC, the random forest and gradient boosting machine (GBM) methods reported lower misclassification errors on the unseen data, registering 9.66% and 9.81%, respectively.

The important variables for hospital readmission prediction identified across all statistical learning methods explored in this minor dissertation predominantly include the length of stay, the number and costs of prior hospital visits, and the primary diagnosis class. It is noteworthy that the ranking of these variables in a variable importance plot varies across different statistical learning models. This variation is attributed to many underlying factors, necessitating an understanding of the distinct characteristics and methodologies inherent to various predictive modelling techniques. The literature highlights several key factors that contribute to these discrepancies:

- Model-specific feature handling: Different models have unique ways of evaluating and handling features. For example, tree-based methods like Random Forests assess variable importance based on metrics like mean decrease in impurity or mean decrease in accuracy, which depend on how and where the variable is used in the trees (Breiman, 2001). Linear models, in contrast, might consider the size and significance of the coefficients associated with each variable (Hastie et al., 2009).

- Interactions and correlations: Some models can capture complex interactions between variables more effectively than others. Models that account for interactions may assign greater importance to variables that are part of significant interaction terms, even if those variables are not as important individually (Friedman, 2001).

- Regularization and feature selection: Techniques such as LASSO (Least Absolute Shrinkage and Selection Operator) apply penalties to the coefficients of regression models, which can lead to some variables being entirely excluded from the model. This influences their perceived importance (Hastie et al.,

2009).

- Randomness: Ensemble statistical learning methods, in particular, introduce randomness in variable selection and combination, leading to variability in variable importance across different runs or data subsets (Leo, 1996).

Given these considerations, the observed discrepancies in the ranking of variable importance across different statistical learning methods, as noted in this minor dissertation, are to be expected. This variability emphasises the necessity for a holistic approach to model evaluation, incorporating multiple models and domain knowledge of healthcare professionals to comprehensively understand the influence of various variables on the prediction of hospital readmissions.

The next chapter presents the outcomes and insights gained from this minor dissertation, offering a summary of the significant findings. It further discusses the implications of these findings and suggests areas for future research that could build upon the foundation established by this work.

Chapter 6

Conclusions and future work

This section encapsulates the key insights derived from the master's minor dissertation, building upon the goals and objectives introduced in Section 1.2. It concludes with a series of recommendations, laying the foundation for future research endeavours that could naturally follow from the groundwork laid out in this minor dissertation.

6.1 Conclusion

This minor dissertation's goals are outlined in Section 1.2, however for convenience; they are restated below:

Aims:

- identify the most effective statistical learning method for predicting all-cause 30-day hospital readmissions.
- Determine the critical factors associated with hospital readmissions within the privately insured South African population.

This minor dissertation enhances the prediction of hospital readmission risks within the privately insured South African population, moving beyond the conventional LACE score, which relies on Length of stay, Acuity of admission, Comorbidities,

and Emergency department visits. The investigation into eight statistical learning models demonstrated a notable improvement in prediction performance metrics over the LACE score, asserting their superiority in key performance indicators such as AUC, AUC-PR, F1 score, MCC, and misclassification error. These metrics are essential for evaluating the efficacy of predictive models in healthcare, where precision and dependability are critical.

Among the assessed models, the neural network emerged as particularly effective, highlighting the value of sophisticated statistical learning methods for managing the complex, non-linear data relationships typical in medical datasets. The Random Forest and GBM models, known for their robustness and ability to process extensive datasets with numerous variables, also yielded encouraging results, especially in minimising misclassification errors. This is crucial in medical predictions, where the cost of a false negative (not identifying a patient at risk) can be exceptionally high. The results suggest a balance must be struck between model complexity and the desired level of accuracy. Although the neural network slightly surpassed the GBM and Random Forest models in terms of performance metrics, its complexity might not warrant preference over the more interpretable GBM and Random Forest models for hospital readmission predictions.

Furthermore, this minor dissertation broadens the scope of variables beyond those included in the LACE score, incorporating factors such as the number and costs of previous hospital visits, length of stay, specialist professionals' costs, age, and primary diagnosis class. This expanded approach mirrors the complex nature of hospital readmissions by encompassing a more comprehensive array of patient-related factors and recognises the importance of diagnosis-related groups and major diagnostic categories in predicting readmissions. The consideration of disease-specific models also underscores the applicability of statistical learning methods.

Although centered on South Africa's private healthcare system, the implications of this study are far-reaching, offering valuable insights for a broad spectrum of healthcare practitioners. These insights can guide the creation of targeted, multi-

disciplinary strategies for patients at elevated risk of readmission, potentially improving patient outcomes, curtailing healthcare costs, and enhancing the quality of care. However, approximately 16% of South Africa's population is covered by private healthcare insurance in 2022 (Statistics South Africa, 2022; Council for Medical Schemes, 2022). To this end, while findings from the private healthcare sector can inform improvements in the public sector, they must be adapted to address the unique challenges and contexts of public healthcare. Careful consideration of resource constraints, patient demographics, and operational differences is essential to ensure that any applied strategies are effective and sustainable.

In summary, this minor dissertation underscores the intricate dynamics of hospital readmissions and establishes the advantages of advanced statistical learning models over traditional approaches. Its findings have the potential to transform how healthcare providers predict and manage hospital readmissions, leading to better patient care and more efficient healthcare systems.

Nevertheless, it is imperative to recognise the limitations of this minor dissertation and the avenues for further research it opens. The reliance on data from private hospitals calls for an investigation into the generalisability of these findings to public healthcare settings and a broader demographic.The advent of health-related big data and the integration of vast datasets into predictive models represent significant areas for future exploration. Additionally, examining various readmission time frames and the possibility of multiple readmissions could deepen the understanding and predictive accuracy regarding hospital readmission risks. Future research should also assess the effectiveness of statistical leaning methods in forecasting disease-specific readmission outcomes across different patient groups

Appendix 1: Logistic regression coefficients

Variable Name	Coefficients	Standardised Coefficients	Variable Name	Coefficients	Standardised Coefficients
Intercept	-10.56932686	-9.90625109	length_of_stay	-0.03473994	-0.24750252
Pregnancy, childbirth, and the puerperium	-0.36393235	-0.36393235	high_care_days	0.00947541	0.02289731
Surgical	-0.07440585	-0.07440585	icu_days	0.00996200	0.02899056
cpts_count.one	-0.10426911	-0.10426911	beneficiary_number	-0.00356158	-0.00559056
cpts_count.three	-0.03189776	-0.03189776	hosp_high_care_appr_amt	0.00000118	0.01906442
cpts_count.two	-0.09775724	-0.09775724	hosp_n_natal_icu_clm_amt	-0.00000155	-0.02328172
cpts_count.zero	0.61465264	0.61465264	hosp_icu_appr_amt	-0.00000219	-0.08887175
icd_10_count.one	-0.13711100	-0.13711100	hosp_g_ward_appr_amt	-0.00000096	-0.04226220
icd_10_count.three	-0.04617978	-0.04617978	hypertension	-0.08119020	-0.03745124
icd_10_count.two	-0.03467816	-0.03467816	hyperlipidemia	-0.05136239	-0.01959237
severity.W CC	0.59716718	0.59716718	diabetes_mellitus_type_ii	0.02315121	0.00822699
severity.W MCC	0.65842054	0.65842054	diabetes_mellitus_type_i	0.04596314	0.00461158
severity.W/O CC	0.46659311	0.46659311	ischemic_heart_disease	0.06396887	0.01244393
gender.M	-0.00639666	-0.00639666	cardiomyopathy	0.12486908	0.02531116
age_years	-0.00060925	-0.01405046	mental_illness	0.10572307	0.03739669
low_cost_option_flag	-0.00593821	-0.00135183	disease_count	0.00548800	0.01320972
hospital_amt	0.00000289	0.24694427	rad_visits_60_days_prior	0.14308141	0.16956610
total_amount	-0.00000101	-0.12814729	hosp_visits_60_days_prior	1.08176891	1.10237812
driving_provider_amount	-0.00000339	-0.04202021	gp_visits_60_days_prior	0.03358129	0.06444349
anaesthetist_amt	0.00000625	0.03307014	nurse_visits_60_days_prior	-0.00390982	-0.00307562
radiology_amt	0.00000048	0.00274966	clinic_visits_60_days_prior	1.48097762	0.21931763
pathology_amt	0.00001067	0.09806472	phar_visits_60_days_prior	0.03627012	0.08294961
pharmacy_amt	0.00001023	0.02030527	path_visits_60_days_prior	0.03839680	0.16055887
other_providers_amt	-0.00000026	-0.00326998	spec_visits_60_days_prior	0.01706885	0.13509421
driving_provider.51	-1.61432312	-1.61432312	rad_amt_60_days_prior	0.00001080	0.06553407
driving_provider.54	-0.72256893	-0.72256893	hosp_amt_paid_60_days_prior	-0.00000032	-0.02161366
driving_provider.62	-0.66524836	-0.66524836	gp_costs_60_days_prior	0.00003760	0.08146629
driving_provider.64	0.10010157	0.10010157	nurse_paid_60_days_prior	0.00000169	0.00233930
driving_provider.92	0.03420307	0.03420307	clinic_amt_paid_60_days_prior	0.00001850	0.04576416
driving_provider.95	0.04902118	0.04902118	phar_amt_paid_60_days_prior	0.00000007	0.00082306
class.Medical	-0.56958769	-0.56958769	path_costs_60_days_prior	-0.00001310	-0.09755965
class.Newborns and other neonates	-0.21584509	-0.21584509	spec_costs_60_days_prior	0.00000168	0.03435567
class.Pre MDC	0.16778048	0.16778048	co_payment_amt	0.00000400	0.00387172

Table 6.1: Standardised and unstandardised logistic coefficients

Appendix 2: Correlation matrix of all numerical variables.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27
V 1	1.000																										
V 2	0.532	1.000																									
V 3	0.472	0.524	1.000																								
V 4	0.395	0.230	0.199	1.000																							
V 5	0.726	0.375	0.326	0.446	1.000																						
V 6	0.093	0.065	0.006	0.064	0.109	1.000																					
V 7	0.712	0.341	0.356	0.397	0.686	0.085	1.000																				
V 8	0.684	0.355	0.184	0.303	0.573	0.158	0.493	1.000																			
V 9	0.521	0.272	0.213	0.300	0.479	0.083	0.408	0.419	1.000																		
V 10	0.716	0.335	0.294	0.274	0.715	0.034	0.606	0.430	0.233	1.000																	
V 11	0.530	0.269	0.216	0.302	0.477	0.082	0.407	0.418	0.733	0.235	1.000																
V 12	0.731	0.333	0.300	0.271	0.719	0.034	0.603	0.427	0.224	0.738	0.235	1.000															
V 13	0.718	0.575	0.515	0.365	0.685	0.107	0.607	0.709	0.371	0.529	0.372	0.538	1.000														
V 14	0.105	0.023	-0.009	0.048	0.102	0.045	0.079	0.192	0.064	0.059	0.065	0.057	0.116	1.000													
V 15	-0.037	-0.041	-0.042	-0.024	-0.024	0.022	-0.015	-0.033	-0.009	-0.017	-0.009	-0.017	-0.049	0.134	1.000												
V 16	0.034	-0.009	-0.021	0.020	0.044	-0.002	0.028	0.076	0.019	0.022	0.020	0.022	0.035	0.145	0.107	1.000											
V 17	0.000	-0.007	-0.006	0.000	0.000	0.000	0.006	0.003	-0.001	0.005	-0.001	0.004	-0.004	0.004	0.007	-0.001	1.000										
V 18	-0.015	0.003	0.001	-0.017	-0.018	0.000	-0.012	-0.026	-0.015	-0.008	-0.015	-0.009	-0.015	-0.020	-0.019	0.003	-0.004	1.000									
V 19	-0.011	0.000	-0.001	0.014	0.012	0.062	0.007	-0.041	-0.007	-0.005	-0.006	-0.004	-0.015	-0.018	0.071	0.035	0.013	0.039	1.000								
V 20	0.135	0.019	-0.025	0.030	0.149	0.076	0.098	0.244	0.088	0.086	0.087	0.084	0.139	0.626	0.169	0.201	0.012	-0.018	0.006	1.000							
V 21	0.119	0.020	-0.027	0.025	0.115	0.068	0.105	0.267	0.092	0.061	0.090	0.057	0.133	0.609	0.153	0.107	0.014	-0.019	-0.028	0.790	1.000						
V 22	0.063	0.047	0.026	0.057	0.051	0.059	0.047	0.131	0.035	0.005	0.037	0.005	0.097	0.488	0.105	0.042	0.008	-0.008	0.019	0.296	0.338	1.000					
V 23	0.140	0.034	-0.004	0.020	0.113	0.036	0.083	0.221	0.072	0.096	0.075	0.097	0.141	0.552	0.097	0.125	0.006	-0.023	-0.059	0.673	0.625	0.232	1.000				
V 24	0.041	0.005	-0.003	0.010	0.039	0.007	0.033	0.081	0.018	0.021	0.017	0.020	0.050	0.200	0.084	0.478	-0.001	0.010	-0.018	0.254	0.183	0.081	0.248	1.000			
V 25	-0.009	-0.006	-0.028	-0.005	0.019	0.184	0.004	-0.011	0.010	-0.007	0.010	-0.007	-0.014	0.024	0.152	-0.017	0.011	0.020	0.251	0.115	0.093	0.052	0.038	-0.014	1.000		
V 26	0.140	0.031	-0.019	0.051	0.175	0.102	0.104	0.220	0.090	0.097	0.089	0.095	0.138	0.609	0.152	0.164	0.005	-0.018	-0.012	0.645	0.713	0.315	0.650	0.231	0.116	1.000	
V 27	0.081	0.052	0.005	0.006	0.063	0.053	0.078	0.173	0.053	0.040	0.052	0.038	0.095	0.503	0.122	0.068	0.012	0.018	0.009	0.555	0.648	0.272	0.636	0.189	0.073	0.527	1.000

Table 6.2: Correlation matrix of all numerical variables.

Where:

- Variable 1 : Hospital amount
- Variable 2 : Driving provider amount
- Variable 3 : Anaesthetist amount
- Variable 4 : Radiology amount
- Variable 5 : Pathology amount
- Variable 6 : Pharmacy amount
- Variable 7 : Other provider's amount
- Variable 8 : Length of stay
- Variable 9 : High care days
- Variable 10 : ICU days
- Variable 11 : High care approved amount
- Variable 12 : ICU approved amount

- Variable 13 : General ward approved amount
- Variable 14 : Radiology visits 60 days prior
- Variable 15 : Hospital visits 60 days prior
- Variable 16 : General practice visits 60 days prior
- Variable 17 : Nurse visits 60 days prior
- Variable 18 : Clinic visits 60 days prior
- Variable 19 : Pharmacy visits 60 days prior
- Variable 20 : Pathology visits 60 days prior
- Variable 21 : Specialists visits 60 days prior
- Variable 22 : Radiology amount 60 days prior
- Variable 23 : Hospital amount 60 days prior
- Variable 24 : General practice amount 60 days prior
- Variable 25 : Pharmacy amount 60 days prior
- Variable 26 : Pathology amount 60 days prior
- Variable 27 : Specialists amount 60 days prior

Appendix 3: Summary statistics: The average number of visits and the amounts paid by patients who were readmitted to the hospital within 30 days compared to those who were not.

	Number of patients with no 30-day readmission	**Number of patients with 30-day readmission**
Average of total_amount	60374.02	83512.93
Average of length_of_stay	4.09	5.76
Average of driving_provider_amount	7257.32	7733.46
Average of hosp_visits_60_days_prior	0.25	1.41
Average of hosp_amt_paid_60_days_prior	6748.35	60335.65
Average of spec_visits_60_days_prior	1.73	10.19
Average of spec_costs_60_days_prior	3397.76	21425.75
Average of gp_visits_60_days_prior	1.09	2.21
Average of gp_costs_60_days_prior	621.10	1902.44
Average of nurse_visits_60_days_prior	0.01	0.04
Average of nurse_paid_60_days_prior	20.42	52.50
Average of clinic_visits_60_days_prior	0.02	0.05
Average of clinic_amt_paid_60_days_prior	195.79	715.50
Average of phar_visits_60_days_prior	2.23	3.12
Average of phar_amt_paid_60_days_prior	1948.93	5147.13
Average of path_visits_60_days_prior	1.38	6.03
Average of path_costs_60_days_prior	1654.05	9013.35
Average of rad_visits_60_days_prior	0.33	1.60
Average of rad_amt_60_days_prior	1209.84	5965.49
Average of anaesthetist_amt	2136.90	2415.95
Average of radiology_amt	2300.21	2976.80
Average of pathology_amt	3772.09	5568.15
Average of pharmacy_amt	194.57	483.49
Average of other_providers_amt	2891.59	4645.20
Average of co_payment_amt	133.15	100.10
Average of high_care_days	0.46	0.75
Average of icu_days	0.41	0.63
Average of hosp_high_care_appr_amt	2965.41	5884.34
Average of hosp_icu_appr_amt	5257.52	8869.55
Average of hosp_g_ward_appr_amt	28301.93	36618.93

Table 6.3: The average number of visits and the amounts paid by patients who were readmitted to the hospital within 30 days compared to those who were not.

Bibliography

Anderson, G. F. and Steinberg, E. P. (1984). Hospital readmissions in the medicare population. *New England Journal of Medicine*, 311(21):1349–1353.

Artetxe, A., Beristain, A., and Grana, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, 164:49–64.

Ashton, C. M., Del Junco, D. J., Souchek, J., Wray, N. P., and Mansyur, C. L. (1997). The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. *Medical care*, pages 1044–1059.

Ashton, C. M. and Wray, N. P. (1996). A conceptual framework for the study of early readmission as an indicator of quality of care. *Social science & medicine*, 43(11):1533–1541.

Austin, P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality. *Statistics in medicine*, 26(15):2937–2957.

Balla, U., Malnick, s., and Schattner, A. (2008). Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. *PMIP*, 87(5):294–300.

Ben-Chetrit, E., chen shuali, C., Zimran, E., Munter, G., and Nesher, G. (2012). A simplified scoring tool for prediction of readmission in elderly patients hospitalized in internal medicine departments. *The Israel Medical Association journal : IMAJ*, 14:752–6.

Benbassat, J. and Taragin, M. (2000). Hospital readmissions as a measure of quality of health care: advantages and limitations. *Archives of internal medicine*, 160(8):1074–1081.

Billings, J., Georghiou, T., Blunt, I., and Bardsley, M. (2013). Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. *BMJ open*, 3(8):e003352.

Boulding, W., Glickman, S., Manary, M., Schulman, K., and Staelin, R. (2011). Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. *The American journal of managed care*, 17:41–8.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. and Friedman, J. (1984). Charles j stone, and richard a olshen. *Classification and regression trees*.

Brownlee, J. (2020). How to avoid data leakage when performing data preparation.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chennamaneni, P., Echambadi, R., Hess, J. D., and Syam, N. (2008). How do you properly diagnose harmful collinearity in moderated regressions? *Retrieved May*, 25:2011.

Corrigan, J. M. and Martin, J. B. (1992). Identification of factors associated with hospital readmission and development of a predictive model. *Health services research*, 27(1):81.

Cotter, P. E., Bhalla, V. K., Wallis, S. J., and Biram, R. W. S. (2012). Predicting readmissions: poor performance of the LACE index in an older UK population. *Age and Ageing*, 41(6):784–789.

Council for Medical Schemes (2022). 2020/21 industry report. Technical report.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Curry, N., Billings, J., Darin, B., Dixon, J., Williams, M., and Wennberg, D. (2005). Predictive risk project. *Literature review. Londres: The King's Fund*.

Darabi, N., Hosseinichimeh, N., Noto, A., Zand, R., and Abedi, V. (2021). Machine learning-enabled 30-day readmission model for stroke patients. *Frontiers in Neurology*, 12.

Demir, E., Chaussalet, T., Xie, H., and Millard, P. H. (2009). Modelling risk of readmission with phase-type distribution and transition models. *IMA Journal of Management Mathematics*, 20(4):357–367.

Department of Health (2022). Annual report 2020/21. Technical report.

Donzé, J., Lipsitz, S., Bates, D. W., and Schnipper, J. L. (2013). Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. *BMJ*, 347.

Dreyer, R. and Viljoen, A. (2019). Evaluation of factors and patterns influencing the 30-day readmission rate at a tertiary-level hospital in a resource-constrained setting in cape town, south africa. *South African Medical Journal*, 109(3):164–168.

El Morr, C., Ginsburg, L., Nam, S., Woollard, S., et al. (2017). Assessing the performance of a modified lace index (lace-rt) to predict unplanned readmission after discharge in a community teaching hospital. *Interactive Journal of Medical Research*, 6(1):e7183.

Francis, N. U., Mason, J., Salib, E., Allanby, A., Messenger, D., Allison, A. s., Smart, N. J. ., and Ockrim, J. B. (2015). Factors predicting 30-day readmission after laparoscopic colorectal cancer surgery within an enhanced recovery programme. *PMIP*, 17(7):294–300.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Garzotto, M., Beer, T. M., Hudson, R. G., Peters, L., Hsieh, Y.-C., Barrera, E., Klein, T., and Mori, M. (2005). Improved detection of prostate cancer using classification and regression tree analysis. *J Clin Oncol*, 23(19):4322–9.

Graham, H. and Livesley, B. (1983). Can readmissions to a geriatric medical unit be prevented? *The Lancet*, 321(8321):404–406.

Hackbarth, G., Reischauer, R., and Miller, M. (2007). Report to the congress: promoting greater efficiency in medicare. *Washington, DC: MedPAC*.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12.

Hammermeister, K. E., Shroyer, A. L., Sethi, G. K., and Grover, F. L. (1995). Why it is important to demonstrate linkages between outcomes of care and processes and structures of care. *Medical care*, pages OS5–OS16.

Hammill, B. G., Curtis, L. H., Fonarow, G. C., Heidenreich, P. A., Yancy, C. W., Peterson, E. D., and Hernandez, A. F. (2011). Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):60–67.

Harrell, F. E. et al. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Heggestad, T. (2002). Do hospital length of stay and staffing ratio affect elderly patients' risk of readmission? a nation-wide study of norwegian hospitals. *Health services research*, 37(3):647–665.

Heggestad, T. and Lilleeng, S. E. (2003). Measuring readmissions: focus on the time factor. *International Journal for Quality in Health Care*, 15(2):147–154.

Holloway, J. J., Thomas, J. W., and Shapiro, L. (1988). Clinical and sociodemographic risk factors for readmission of medicare beneficiaries. *Health Care Financing Review*, 10(1):27.

Hosseinzadeh, A., Izadi, M., Verma, A., Precup, D., and Buckeridge, D. (2013). Assessing the predictability of hospital readmission using machine learning. *Proceedings of the twenty-seventh AAAI conference on artificial intelligence*, pages 1532–1538.

Howell, S., Coory, M., Martin, J., and Duckett, S. (2009). Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research*, 9(1):1–9.

Huang, Y., Talwar, A., Chatterjee, S., and Aparasu, R. R. (2021). Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC medical research methodology*, 21(1):1–14.

Informa (2019). Industry insights: South africa healthcare market overview. Technical report.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Jencks, S. F., Williams, M. V., and Coleman, E. A. (2009). Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428. PMID: 19339721.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., and Kripalani, S. (2011). Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA*, 306(15):1688–1698.

Kind, A. J., Jencks, S., Brock, J., Yu, M., Bartels, C., Ehlenbach, W., Greenberg, C., and Smith, M. (2014). Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Annals of internal medicine*, 161(11):765–774.

Kocher, R. P. and Adashi, E. Y. (2011). Hospital Readmissions and the Affordable Care Act: Paying for Coordinated Quality Care. *JAMA*, 306(16):1794–1795.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Lefèvre, J. H., Reboul-Marty, J., de Vaugrigneuse, S., and Zeitoun, J.-D. (2017). Readmissions after surgery: a french nationwide cross-sectional study of 1,686,602 procedures performed in 2010. *World journal of surgery*, 41(1):31–38.

Leo, B. (1996). Bagging predictors in machine learning.

Lewis, G., Curry, N., and Bardsley, M. (2011). Choosing a predictive risk model: a guide for commissioners in england. *London: Nuffield Trust*, 20.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Malohlava, M., Hava, J., and Mehta, N. (2016). Machine learning with sparkling water: H2o+ spark. *H2O. ai Inc.*

McIlvennan, C. K., Eapen, Z. J., and Allen, L. A. (2015). Hospital readmissions reduction program. *Circulation*, 131(20):1796–1803.

McNaughton, C. D., Collins, S. P., Kripalani, S., Rothman, R., Self, W. H., Jenkins, C., Miller, K., Arbogast, P., Naftilan, A., Dittus, R. S., and Storrow, A. B. (2013). Low numeracy is associated with increased odds of 30-day emergency department or hospital recidivism for patients with acute heart failure. *Circulation: Heart Failure*, 6(1):40–46.

Morgan, D. J., Bame, B., Zimand, P., Dooley, P., Thom, K. A., Harris, A. D., Bentzen, S., Ettinger, W., Garrett-Ray, S. D., Tracy, J. K., and Liang, Y. (2019). Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA Network Open*, 2(3):e190348–e190348.

Ripley, B. D. et al. (2001). The r project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, 1(1):23–25.

Ruff, C., Gerharz, A., Groll, A., Stoll, F., Wirbka, L., Haefeli, W. E., and Meid, A. D. (2021). Disease-dependent variations in the timing and causes of readmissions in germany: A claims data analysis for six different conditions. *Plos one*, 16(4):e0250298.

Sarah Rimar MD, Joseph Musto MD, J. K. M. M. M. B. M. S. S. M. P. R. K. M. (2014). Application of the lace and modified lace indices at rush university medical center: A retrospective study. *Annals of Internal Medicine*, 161(11):765–774.

Schober, P. and Vetter, T. (2021). Logistic regression in medical research. *Anesthesia Analgesia*, 132:365–366.

Sheingold, S. H., Zuckerman, R., and Shartzer, A. (2016). Understanding medicare hospital readmission rates and differing penalties between safety-net and other hospitals. *Health Affairs*, 35(1):124–131. PMID: 26733710.

Snyders, P., Swart, O., and Duvenage, R. (2020). Thirty-day readmission rate: A predictor of initial surgical severity or quality of surgical care? a regional hospital analysis. *South African Medical Journal*, 110(6):537–539.

Soeken, K. L., Prescott, P. A., Herron, D. G., and Creasia, J. (1991). Predictors of hospital readmission: a meta-analysis. *Evaluation & the health professions*, 14(3):262–281.

Statistics South Africa (2022). Government spending breaches r2 trillion. Technical report.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, 18(10):1099–1104.

Stone, J. L., Hoffman, G., et al. (2010). *Medicare hospital readmissions: Issues, policy options and PPACA*. Congressional Research Service Washington, DC.

Strandberg, T., Pitkälä, K., and Tilvis, R. (2011). Frailty in older people. *European geriatric medicine*, 2(6):344–355.

Sushmita, S., Khulbe, G., Hasan, A., Newman, S., Ravindra, P., Roy, S. B., De Cock, M., and Teredesai, A. (2016). Predicting 30-day risk and cost of" all-cause" hospital readmissions. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Tsai, T. C., Orav, E. J., and Joynt, K. E. (2014). Disparities in surgical 30-day readmission rates for medicare beneficiaries by race and site of care. *Annals of surgery*, 259(6):1086.

van Barneveld, E. M., van Vliet, R. J., and van de Ven, W. P. (1997). Risk-adjusted capitation payments for catastrophic risks based on multi-year prior costs. *Health policy*, 39(2):123–135.

Van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557.

Wang, H., Robinson, R. D., Johnson, C., Zenarosa, N. R., Jayswal, R. D., Keithley, J., and Delaney, K. A. (2014). Using the lace index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders*, 97.

Wasylewicz, A. and Scheepers-Hoeks, A.-M. (2019). *Clinical Decision Support Systems*, pages 153–169.

Wennberg, D., Dixon, J., Billings, J., et al. (2006). Combined predictive model–final report.

Wiseman, J. T., Guzman, A. M., Fernandes-Taylor, S., Engelbert, T. L., Saunders, R. S., and Kent, K. C. (2019). General and vascular surgery readmissions: a systematic review. *Journal of the American College of Surgeons*, 35(1):124–131. PMID: 25067801.

Wong, E. L., Cheung, A. W., Leung, M., Yam, C. H., Chan, F. W., Wong, F. Y., and Yeoh, E.-K. (2011). Unplanned readmission rates, length of hospital stay, mortality, and medical costs of ten common medical conditions: a retrospective analysis of hong kong hospital data. *BMC health services research*, 11(1):1–8.

Woodside, M. (1953). A follow-up of psychiatric patients; one year's survey of patients discharged from the york clinic. *Guy's Hospital reports*, 102(1):70–75.

World Health Organisation (2005). Do current discharge arrangements from inpatient hospital care for the elderly reduce readmission rates, the length of inpatient stay or mortality, or improve health status? Technical report, Genf, Schweiz.

World Health Organisation (2021). Global expenditure on health:public spending on the rise? Technical report, Genf, Schweiz.

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*, 78:225–241.

Zhou, D.-X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters*, 83(9):2108–2112.

Zhu, K., Lou, Z., Zhou, J., Ballester, N., Kong, N., and Parikh, P. (2015). Predicting 30-day hospital readmission with publicly available administrative database. a conditional logistic regression modeling approach. *Methods of information in medicine*, 54.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.