A Machine Learning Approach to Predicting the Employability of a Graduate

Masego Modibane

A dissertation submitted to the Faculty of Commerce, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Philosophy.

January 3, 2019

MPhil in Data Science specialising in Financial Technology,
University of Cape Town.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Philosophy to the University of Cape Town. It has not before been submitted for any degree or examination.

Signed by candidate

Masego Modibane

January 3, 2019

Abstract

For many credit-offering institutions, such as banks and retailers, credit scores play an important role in the decision-making process of credit applications. It becomes difficult to source the traditional information required to calculate these scores for applicants that do not have a credit history, such as recently graduated students. Thus, alternative credit scoring models are sought after to generate a score for these applicants. The aim for the dissertation is to build a machine learning classification model that can predict a students likelihood to become employed, based on their student data (for example, their GPA, degree/s held etc). The resulting model should be a feature that these institutions should use in their decision to approve a credit application from a recently graduated student.

Acknowledgements

I would like to thank my parents, Rapula and Sarah, and the rest of family, especially my grandaunt Granny, for their continued support in the furthering of my education.
I would like to thank Co-Pierre Georg for his belief in - and supervision over this project.
I would like to thank Allan Davids for the help with shaping this dissertation to what it is today.
I would like to thank Sanlam Investments for funding my Masters degree.
I would like to thank DataFirst for access to the data used in the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Credit scores play a vital role in the decision-making process of granting consumer credit because they assist in evaluating the financial risk of lending to a particular client. Thomas *et al.* (2017) define credit scores as a set of decision models which determines who will get credit and how much credit they should get. There is an abundance of research available on credit scoring models built on machine learning algorithms. Baesens *et al.* (2015) brilliantly summarise the literature available. The article also provides an extensive comparison of these machine learning algorithms.

Alternative methods of credit scoring should be considered where there is multiple sources in which individual data can be collected and used to get a better view of the behaviours of the individual. Information about a candidate is now more easily accessible and needs to be included in models that can provide more access to financial products. *Transunion*, a consumer credit reporting agency, have recently introduced an alternative credit scoring product[1]. The product aims to provide a more holistic view of the South African consumer to a potential lender. This can lead to more South Africans being considered for loans that they can afford.

The scope of this dissertation is to build an index[2] that will use university student data to predict employability after graduation. The intention of this index is to act as a supplement to an existent credit application score model, as opposed to a behavioural scoring model. The resulting model seeks to improve a recent graduates probability of obtaining credit and access to other products and services that need a credit score. This aim is in support of the vision set out by the South African Treasury on financial inclusion in South Africa(*Achieving Effective Financial Inclusion in South Africa: A Payments Perspective*, 2014).

The ideal data set for this kind of classification model would include informa-

[1] The type of data used is not disclosed but is said to benefit the consumer in the eyes of the lender (*TransUnions New CreditVision Model Uses Alternative and Trended Data to Better Predict Credit Risk, Providing Millions of South Africans with More Opportunities to Gain Access to Credit*, 2018)

[2] We will be using *index* and *classification prediction model* interchangeably throughout the discussion

tion about the candidates grade point average (GPA), age, highest degree obtained, type of high school education as well as the candidates assessment marks and their university societal participation. This kind of data can be tracked by the higher education institutions at different levels of interaction with the candidate. The *South Africa - Graduation Destination Survey 2012* (2015) survey data provides us with most of these aformentioned variables in the form of questions asked in the survey. The data also provides an indication of whether the candidate is employed or not. The data is collected from a survey done on the 2010 cohort of graduates from four of the Western Cape Universities. The aim of the survey was to gauge graduate employment (and unemployment). It focussed on various aspects of the graduates path to their employment status as well as the future steps of the graduate. Further discussion of this survey data is in the *Data description section* of Chapter 3.

According to the South African National Credit Regulations of 2006, approved credit institutions that need the student data mentioned can get the information by requesting for the data from the candidates educational institution/s(Mpahlwa, 2006)[3]. However, only once consent has been given by the candidate.

Our research approach is to investigate numerous machine learning algorithms across different domains that could be applicable to a case of binary classification. Credit scoring literature deals mostly with regression problems but could easily be extended to classification problems given the outcome of the original regressed score. The literature for alternative credit scoring is a relatively new field and it poses more difficult to find relevant articles. The target with this dissertation is to add to this literature. After investigating a number of machine learning models and performance measures used in both general classification prediction and in credit scoring, we selected 9 models that will be trained and tested on 5 performance measures through a two-stage testing procedure. The first stage of testing used a data set that only has the most important variables for *Employability* prediction according to the backward selection process, a feature selection method discussed in more detail in Chapter 3. The data set is then divided into a training and test set. All models are then trained with the training set from the abridged data set and tested using the test data set. The top 3 best performing models are then selected to enter the second stage of testing. The second stage applies a stress-testing environment. A new data set consisting of all the other variables that were not seen as important for *Employability* prediction by the feature selection method is now used. The data set is then also divided into a training and test set. The models are re-trained on this new data set and tested. The best prediction model is found to be a *Bagged decision tree* according to the performance results of both stages. How-

[3] Chapter 3, Section 18.1 of *National Credit Regulations of 2006* (Mpahlwa, 2006)

ever, the *Adaboost* model is a very close second. Surprisingly, all models performed extremely well on all measures chosen at the first stage of testing. However, the performance declined at the second stage of testing. This should be expected as the variables used where not originally seen as important in predicting *Employability*. The differences between the model performances were tested for statistical significance. It is found that the models still were able to predict an *Employed* test data point correctly. Yet, they did not fair as well in predicting an *Unemployed* test data point. Thus, all other measures that involve the (mis)classifying of *Unemployed* data points observed significant differences in their performances

The rest of the dissertation is setup as follows;

- Chapter 2 provides a discussion of the literature review that focussed on popular and alternative credit scoring models, graduate employability prediction models as well as general machine learning algorithms used for classification problems.
- Chapter 3 breaks down the method behind building the resulting best prediction model. This includes the data description, as well as the explanation of the models and performance measures used.
- Chapter 4 gives a summary of the results obtained through the two-stage testing procedure used where ultimately the best model is chosen. The chapter also includes some data exploration of the most important variables in the data set.
- Chapter 5 provides a conclusion from the results as well as future recommendations based on the outcomes of the dissertation.

Chapter 2

Literature Review

The aim of this dissertation is to merge two domains of literature, i.e. *Employability prediction* and *Application credit scoring*, to create an alternative approach to the consumer credit score. The contribution to literature is the consideration of a graduate employability index as a supplement to an existing application credit scoring model.
There are three main branches of literature being reviewed, namely;

- Alternative credit scoring
- Employability prediction and,
- Credit scoring using Machine Learning techniques

Alternative credit scoring literature is looked at with two approaches in mind;

1. Alternative models using non-traditional technologies
2. Alternative models using alternative data but the same technologies.

The two approaches could overlap like in the case of Berg *et al.* (2018) where they predict consumer default by augmenting a credit bureau score with the information content of the user's digital footprint i.e. the digital information left online when accessing or registering on a website.

Keeping in mind the first approach mentioned, there are also alternative credit lending in the form of peer-to-peer (P2P) lending[1]. The paper by Lin *et al.* (2013) suggests that there exists a relationship between a borrower and their social network friendships, as well as their credit quality i.e. interest rates on loans and default rates. A concept like this could be combined with a traditional lending platform and the candidate's social network presence. However, ethical considerations need to also be considered before implementing such an alternative model.

With regards to the second approach mentioned, Šušteršič *et al.* (2009) decided

[1] Online lending where lenders are connected with borrowers, eliminating the need for financial institutions

to use a traditional credit scoring approach on accounting data[2], including transactions data and account balances. They believed that there was correlation between the accounting data and credit history data which was proven by the results and the alternative model performed better than they had expected.

Employability prediction literature is generally concerned with identifying the factors that educational institutions need to consider to evaluate the employability of their students (Mishra *et al.*, 2016; Piad *et al.*, 2016; García-Peñalvo *et al.*, 2018).

Alternatively, papers such as Jantawan and Tsai (2013) aim to assist human resource directors in identifying factors that could affect the employee's performance in their first year of work after graduation.

A comparison of Naïve Bayes models and decision tree algorithms[3] appear frequently in these types of papers (Jantawan and Tsai, 2013; Mishra *et al.*, 2016; Piad *et al.*, 2016). Additionally, Piad *et al.* (2016) also looked into the use of logistic regression which ultimately performed the best in their paper, whilst, Mishra *et al.* (2016) also considered a neural network.

Extensive research is available on the different methods applied to and developed for credit scoring. Baesens *et al.* (2015) provide a comprehensive comparison of classifiers as well as explanations on some of these classifiers that are investigated. They have managed to use 41 classification methods on 8 credit scoring data sets with 6 performance measures. Baesens *et al.* (2015) also emphasise the importance of using different types of performance measures to assess discriminatory power, accuracy and correctness of a scorecard. They also elaborate on the use of statistical hypothesis testing to compare difference in performance measures and argued that previous literature did not consider this enough.

In the literature, there are many papers that focus on a novel ensemble algorithm such as those discussed in Antonini *et al.* (2010); Tsai (2014); Marqués *et al.* (2012).

Antonini *et al.* (2010) introduces a derivation of the bagging algorithm[4], where the subset of the data set instead of the whole data set is randomly drawn at each base model. It can be used specifically for data that is imbalanced or prone to missing data, like that of credit scoring data (Antonini *et al.*, 2010).

Tsai (2014) introduces the notion of integrating unsupervised and supervised[5] techniques. The unsupervised learning part of the algorithm would act as a data

[2] as opposed to credit history data
[3] More explanation on this in Appendix A
[4] like the one discussed in 3.4.3
[5] in the form of ensemble methods

reduction tool whilst the supervised section would train the newly clustered and reduced data set. Motivation for this method came from the authors of Tsai (2014) arguing that advanced machine learning techniques are not fully assessed for cases of financial distress.

Another advanced technique for credit scoring includes that of Marqués *et al.* (2012). They explore composite ensembles where a combination of a data resampling technique (such as bagging) and an attribute subset selection method (such as rotation forests) is done to form a new model with the aim of improving prediction performance.

The advanced methods are worth noting to give an idea of where classification prediction modeling is headed in literature. However, they are not used for this dissertation due to time constraints.

In the following chapter (3), there is an explanation of the methodology followed to select the best model. This also includes a more in-depth explanation of the models trained[6] and performance measures used[7], as per the suggestions of the aforementioned papers.

[6] refer to 3.4
[7] refer to 3.5

Chapter 3

Methodology

This chapter gives an elaboration on the methodology used to select the best model for predicting *Employability*. A two-stage selection process is adopted. In stage one, all the models discussed in section 3.4 were trained using an abridged data set and tested using the performance measures mentioned in section 3.5. In stage two, the three best performing models, from stage one, will be retested on a new set of variables that do not include the variables deemed vital to predicting *Employability* according to the feature elimination process[1]. The best performing model overall is then determined. The second stage is put in place to observe which model will still be able to predict *Employability* correctly even though vital variable points are missing.

The rest of the chapter includes a description of the data, the data preprocessing techniques applied and a basic explanation of the models and performance measures used.

3.1 Data description

The classification models will be evaluated on the *South Africa - Graduation Destination Survey 2012* (2015) survey data set. The survey as well as the reporting of its results forms part of the *Cape Higher Education Consortium (CHEC)'s* ongoing work on graduate attributes (*Pathways from university to work*, 2013). The data is from a survey done in 2012, on the 2010 cohort of graduates from four of the Western Cape Universities[2]. The focus of the survey was to gauge graduate employment (and unemployment) as well as get the future steps of the graduate. About 8190 graduates across different disciplines and qualification types were approached to take part in

[1] The feature elimination process used is the backward selection process and the method is discussed in more detail under *Data reduction* in section 3.2

[2] These include University of Cape Town, Stellenbosch University, University of the Western Cape and Cape Peninsula University of Technology.

the survey.
The survey was broken up into 5 sections;

1. High school information
2. Life at university prior to 2010 qualification
3. Employment status and relevance of qualification
4. Further studies done after 2010 qualification
5. Future plans of candidate involving studying and migration

The data set also included the candidates age, their matric year math and physical science academic performance as well as their undergraduate grade point average (GPA).

The original data set has 4864 observation points with 178 variables including the dependent variable of employment status at the time of the survey. Redundant variables were omitted. Through the data preprocessing techniques, discussed in section 3.2, a data set of 49 variables (1 dependent, 48 independent) is used for the first stage of model selection. The probability distribution of the classes in the dependent variable has about 75% of the data points classified as *Employed* and about 25% classified *Unemployed*. This imbalance is kept constant, through stratified sampling, for the division of the data into a test and training set.

3.2 Data preprocessing

Data preprocessing can help with improving the quality of the data that will be fed to the machine learning algorithms (Han *et al.*, 2011). In this section, we introduce the different data preprocessing techniques suggested by Han *et al.* (2011) and Scheule *et al.* (2017) that are applied to the original data set. The resulting transformed data set is used to train and test the different machine learning models that are discussed in section 3.4. Figure 3.1 gives a high level view of what is to follow in this section.

Data cleaning

Quality data can be defined as data that is complete, accurate and consistent (Han *et al.*, 2011). The aim of data cleaning is to transform raw data into quality data. Given that we are dealing with survey data, we are bound to find missing values. Missing values refers to missing data entries that are denoted with *NAs* in the data set. This contributes to the data being incomplete (Han *et al.*, 2011). One way

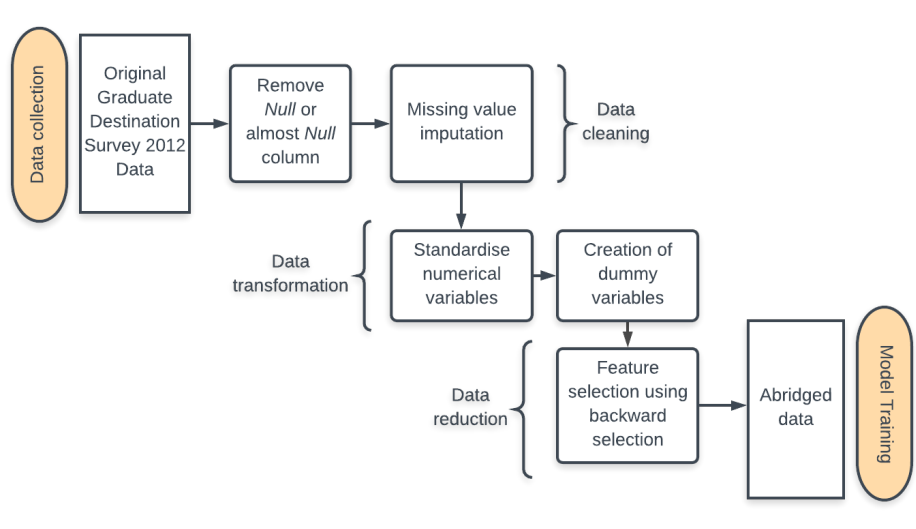


Fig. 3.1: Data preprocessing procedure performed on the *South Africa - Graduation Destination Survey 2012* (2015) data before model testing

to eliminate this problem is to impute the missing values. Various value imputation techniques are available but only the *imputation using decision trees* technique, presented by Rahman and Islam (2013), is used on the missing values in the data[3]. *Imputation using decision trees* is a non-parametric approach to value imputation that uses all other variables to construct a decision tree to predict the missing value of a particular variable.

Prior to imputing of the values, there has to be sufficient valid entries for that specific variable to be able to construct a decision tree. A variable is omitted from the data set if 95% or more of the data points in that variable are classified as missing values.

Data transformation

Data transformation is being considered because it is important to realise the nature of the data that will be presented to particular machine learning algorithms. For example, neural networks are sensitive to non-standardized numerical variables which can therefore influence the output of that model (Han *et al.*, 2011). The aim of data transformation is to ensure that a suitable data set is presented to an algorithm.

The min-max standardization is a transformation of the values of a variable to numbers between a new minimum value (normally 0) and a new maximum value (normally 1). These new minimum and maximum values are based on the minimum and maximum values of the original set of values for that variable. A min-

[3] refer to Appendix A for more explanation on decision trees

max standardization is applied to all numeric variables in the data set except that of the *GPA*[4] variable which is transformed into a percentage value. Numerous other standardization techniques exist but are beyond the scope of this dissertation[5].

Another form of data transformation is the creation of dummy variables. Dummy variables are created when categorical variables need to be converted to a numeric value normally to be able to compute difference between data points. This is done by creating new variables to represent each level of the original categorical variable. A value of 1 indicates that the data point does belong to that specific category-level variable and a value of 0 will reflect at every other category-level variable. This transformation will increase the number of variables in the data set and should only be considered when needed in machine learning algorithms that involve distance measures like the *k*-Nearest Neighbours (see 3.4.1) and Rotation Forests (see 3.4.3) to name a few.

Data reduction

Data reduction techniques can be used when dealing with a large data set. *Large* refers to the volume of the data for purposes of this dissertation and what large means is different from user to user. The objective of the data reduction techniques is to reduce the size of the data without significantly changing the integrity of the original data set. Thus, the abridged data should be more efficient than the original data when applied to the models without notably skewing the results of those models.

The data is reduced through a feature selection technique known as the backward selection process. The backward selection process of feature elimination is a stepwise selection process that begins with fitting a full model i.e. a model that includes all the variables in the original data set. A *Bagged decision tree* [6] is the chosen model because of its non-parametric properties. At each step of the backward selection process, the number of variables in the model decreases by one variable. The best model at each step is chosen according to the model's prediction accuracy (James *et al.*, 2013). The optimal number of variables is the best model with the highest prediction accuracy overall.

[4] Grade point Average
[5] refer to Han *et al.* (2011) for more techniques
[6] see section 3.4.3 for more information on what this is

3.3 Training and Test data

The abridged data[7] is divided into a training and test data set. The training set will be used for training of the machine learning models. The test set is used to determine whether the trained model is a good prediction model for graduate employability.

The percentage of the data that will be used for training and testing is generally subjective. There exists extensive research that suggests how to divide the data such that we keep a balance of the variance between parameter estimates (result from too little training data) and the variance of the performance statistic (result from too little test data). The data set is divided such that 95% of the abridged data is used for training and the remaining 5% is used as test data from the methodology suggested in Amari *et al.* (1997).

Stratified sampling is used to allocate the data points in either the training or test data set. This means that the division of the data is done at random without replacement whilst preserving the original dependent class variable imbalance (Han *et al.*, 2011).

3.4 Models

This section provides simple explanations of all the models used in the first stage of model testing and have been broken up the models into 3 groups, namely *Individual Classifiers, Neural Networks* and *Ensemble methods*, based on the complication of the algorithm.

3.4.1 Individual Classifiers

Models that only require one iteration of the data being fed into the algorithm are classified as *Individual Classifiers*.

Logistic Regression

The Logistic Regression method is derived from statistical theory and has a relation to linear regression (Trevor *et al.*, 2009). The mathematical derivation will not be discussed, however, it is important to note that the classification problem will be modelled using a multiple logistic regression (*mlr*) model.

For binary classification, the *mlr* model outputs a probability of a data point belonging to a specific level of the categorical dependent variable i.e. the probability

[7] the data that has gone through data prepocessing as demonstrated in figure 3.2

of the data point being classified as *Employed*. The classification is determined by a probability threshold. This threshold is especially vital when dealing with imbalanced data. One method to determine the optimal probability threshold is to use the Receiver Operating Characteristic (ROC) curve which is explained in more detail in section 3.5.2.

This method has the disadvantage of having distributional assumptions of the variables that involve means and covariance matrices (Thomas *et al.*, 2017; Han *et al.*, 2011). However, due to its popularity in binary classification, as suggested by (James *et al.*, 2013), it is considered here.

k-Nearest Neighbours

The *k* - Nearest Neighbour (KNN) classifier can be categorised as a lazy learner (Han *et al.*, 2011). A lazy learner stores information of the training data and will only predict a classification based on similarity once a test data point is introduced. The test data point is classified to the class of which the majority of the *k*-Nearest Neighbours belongs to. The form of the data is important because the distance is taken between numeric values.

The method is distance-based and it is important that data is normalised so that the difference in measurement scales of the variables does not skew the resulting model. The euclidean distance measure is used to calculate this distance but many others can be used but are not discussed here (Han *et al.*, 2011).

The selection of *k* can be user-specified. For purposes of this dissertation, we tested the accuracy[8] of the model at every possible integer value of *k* of the KNN classifier using the training data set. The KNN model with the best performing rate over all is thus the optimal model[9]. This value of *k* can change if more training data points are added to the model and therefore should be monitored continuously. Thomas *et al.* (2017) suggest using a distribution of *k* instead of a single value of *k*. However, this approach will not be used here but could be a future consideration.

The KNN model is easy to interpret and is dynamic since the training data can easily be changed or updated. However, it can be computationally expensive (especially when *k* is large) and requires efficient storage. *Partial distance* and *pruning* methods could have been considered to alleviate the computational inefficiency, as suggested by Han *et al.* (2011), but are omitted for this discussion.

[8] as defined in section 3.5.1

[9] Refer to figure B.8 to view the plot of KNN model at the first stage of testing.

Support Vector Machines

Support Vector Machines (SVMs) are a semi-parametric method that support different functional forms and require the user to select that form a priori (Baesens *et al.*, 2015). It applies a non-linear transformation, also known as a kernel trick, on the original data set such that the data is in a higher dimension. In the higher dimension, a decision boundary (hyperplane) is derived through optimization routines which divides the observations between the two classes (i.e. *Employed* and *Unemployed*). The classification of the data point is then determined by which side of the decision boundary the data point will fall on through the non-linear transformation of the independent variables. The explanation of the mathematics of the non-linear transformation as well as how the decision boundary is created is not discussed here[10].

SVMs are more effective and accurate when classification between the classes is not linearly divisible given the original set of independent variables. However, the method is computationally inefficient (Han *et al.*, 2011). SVMs are also less prone to overfitting since the complexity of the classifier is not defined by the dimensionality of the data but rather the number of training data points that are used in constructing the decision boundary (which are also known as support vectors (Han *et al.*, 2011)).

Popular choices for the non-linear kernels include *Polynomial*, *Sigmoid* and *Radial basis*. We will omit the *Sigmoid* kernel as, according to Han *et al.* (2011), it has relations to the multi-layer perceptron which is discussed later in section 3.4.2. *Radial basis* kernel has more local behaviour, which means that the decision boundary takes shape to the behaviour of the neighbours of a specific data point (James *et al.*, 2013). The *Polynomial* kernel transforms the decision boundary from linear to a more flexible one.

A disadvantage of the considered kernel transformations is that they require parameters to be set. The parameters were chosen in a grid search of the combinations of parameters. The best performing combination of parameters was selected through the best AUC[11] value and results as shown in table B.2.

3.4.2 Multi-layer feed-forward perceptron Neural Networks

Neural networks are modelled after the manner in which the neurons in the human brain communicate and process information (Thomas *et al.*, 2017). The multi-layer feed-forward perceptron neural network structure consists of an input layer, one or

[10] refer to (James *et al.*, 2013; Han *et al.*, 2011)
[11] discussed in more detail in section 3.5.2

more hidden layers and an output layer. All the independent variables are fed into the input layer. The variables are then transformed through the hidden layers and eventually a value is calculated for the output layer. At each layer, a weighted sum of the inputs is calculated for each neuron in the layer. Each neuron, in the current layer, then transforms the weighted sum through an activation function where the result will be used as an input for the next layer of the network (see figure 3.2).
The neural network is feed-forward and this ensures that the weights do not feed into a previous layer of the network.

An activation function is chosen by the user and for a classification problem, popular choices include *threshold, logistic* or *tanh* functions (Thomas *et al.*, 2017).
The number of hidden layers is user-specified. Hand and Henley (1997) argue that two hidden layers are enough for any type of prediction problem. A perceptron with no hidden layer is better suited for linearly separable data whilst a multi-layer perceptron is suited for non-linearly separable data[12] (Rumelhart *et al.*, 1986).

The weights are calculated through training under back propagation algorithm with the aim of minimizing the *entropy* error function (Thomas *et al.*, 2017; Han *et al.*, 2011; Trevor *et al.*, 2009).

The number of neurons per hidden layer is user-specified, however, Thomas *et al.* (2017) suggest that an optimal number of neurons can be determined after the initial training stage.

For binary classification, the output layer has one output value with a value that could be seen as a probability[13] of being classified as *Employed*.
The derivation behind a neural network is difficult to interpret, however, it does have a higher tolerance to deal with noisy data and is not effected by correlation between variables unlike a logistic regression model (Han *et al.*, 2011).

3.4.3 Ensemble Methods

An *Ensemble Method* is a technique of creating a composite model. Therefore, a model based on a number of types of models (also known as base learners/classifier). Each base learner can have a sampled version of the original data set, depending on the process of the chosen ensemble method being used. A test data point is then classified into the class through a type of voting procedure.

For Homogeneous ensemble methods, we discuss *Bagging, Boosting* and *Rotation Forests*. The methodologies are similar in that they have the same type of model as a base learner but differ in the data set sampling, base classifier dependence at

[12] where it is not possible to distinguish between classes with a straight boundary line
[13] the value is not necessarily a number between 0 and 1

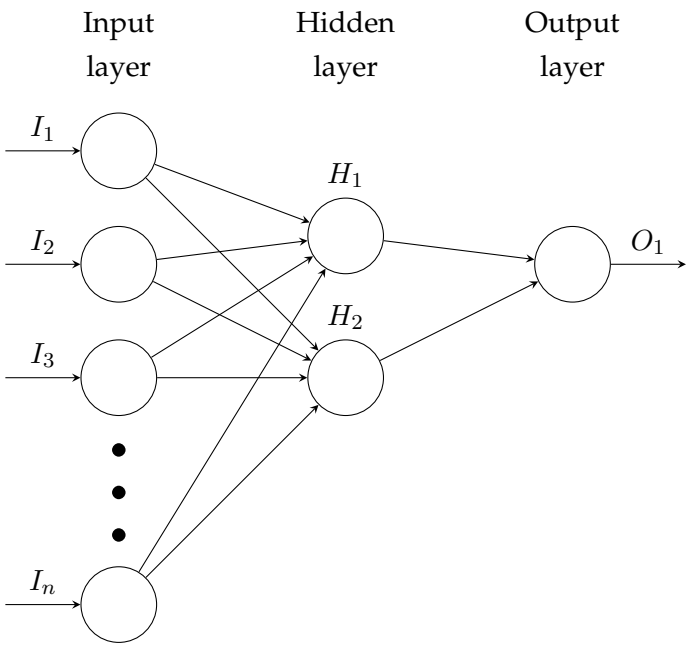


Fig. 3.2: Skeleton of the multi-layer feedforward neural network model where n is the number of variables used in the abridged data set

each iteration and weights of the voting process. For all ensemble methods we will be using decision trees as base learners for ease of interpretability[14].

For the Heterogeneous ensemble approach, we have created a model that uses an equally-weighted majority vote of the classifications given by the 5 individual classifier models discussed earlier (i.e. *KNN, logistic regression, polynomial* and *radial SVM*, as well as a *Multi-layer Perceptron Neural Network*). This model aims to increase accuracy which could also lead to overfitting of a model on the training data.

Selective ensembles could have also been considered. These types of ensembles chose a subset of the original base models instead of using all the models. Baesens *et al.* (2015) elaborate on these, however, we will not review them here.

Bagging

Bagging is a simple ensemble method that is also known as *bootstrap aggregation* (Han *et al.*, 2011; James *et al.*, 2013). For each base classifier, the original data set is resampled with replacement and therefore has repeated data points in the resulting data set (also known as bootstrap sampling). The bootstrapped data set is fed into the base classifier algorithm to produce a model. After all base classifiers are

[14] see appendix A

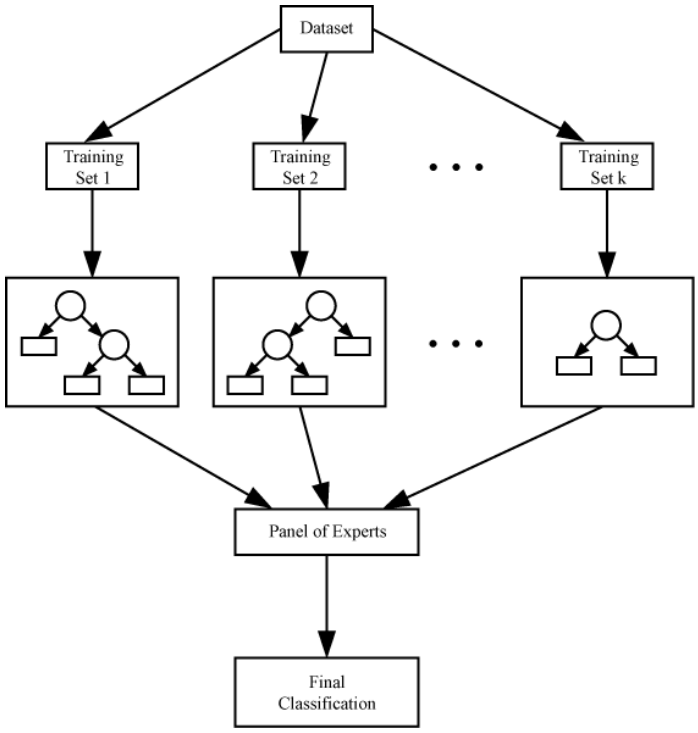


Fig. 3.3: A general view of a bagged decision tree model where each training set is a bootstrapped version of the original data set that is fed into a decision tree base classifier. The *Panel of Experts* is the classification outcome of each base classifier and the *Final Classification* is a result of the majority of the classification outcomes of the base classifications

(*Decision tree ensemble using Bagging algorithm*, 2009)

created, the test data point is fed into all the models simultaneously and a classification is obtained for each base model. The data point is then classified into the class that the majority of the models have classified the data point into. Figure 3.3 shows a general view of a *bagged decision tree* model.

The method has the advantage of reducing the variance of individual classifiers. It is also hardly affected by issues of outliers (Han *et al.*, 2011). *Bagging*, however, can suffer from bias due to dominant independent variables in the data set (James *et al.*, 2013), thus methods such as *Random* and *Rotation forests* (see 3.4.3) were developed. The method is not recommended for real time classification due to it being computationally expensive especially when there is a large number of base classifiers.

Adaptive Boosting

Adaptive boosting is an ensemble method originating from the boosting method that aims to improve the performance of the algorithm (Freund and Schapire, 1997). It is also known as the *adaboost* method. Boosting is different from bagging (see section 3.4.3), in that each base classifier is assigned its own vote, depending on how well it has performed as opposed to an equal vote (Han *et al.*, 2011).
The *adaptive boosting* method, introduced by Freund and Schapire (1997) focuses on improving what is considered a "weak learner". Thus, the method puts more emphasise on trying to improve classifiers that have a high misclassification/error rate. Freund *et al.* (1999) and Han *et al.* (2011) provide a more detailed explanation of how the algorithm works.
Adaboost demonstrates the following advantages;

- Fast, simple and easily programmable. (Freund *et al.*, 1999)
- it only needs the user to specify how many base learners/iterations should be constructed.

However, if the data has many outliers, the performance of *adaboost* will suffer since the method focuses on trying to improve the classification of "difficult-to-classify" data points (Freund *et al.*, 1999). The method also has the potential to overfit due to its focus on improving the misclassification rate and consequently will have better accuracy over a *bagging* model (Han *et al.*, 2011).

The base learner chosen is a decision tree with a tree depth of 1 as suggested by Antonini *et al.* (2010). They state that an *adaboost* with stumps handles unbalanced data sets well as the balanced error is minimised.

Rotation Forest

Rotation forest is a classifier ensemble method that is presented by Rodriguez *et al.* (2006). It is derived from a focus on feature extraction and can be broken down into two phases, training and classification.

In the training phase, a rotation matrix is prepared by applying Prinicipal Components Analysis (PCA)[15] on a randomly selected subset of the original data set. The data used needs to be completely numeric data and the categorical variables need to be transferred into dummy variables in order for the PCA to work. A classification decision tree is then built on the rotated data set by applying the rotation matrix on the original data set. This decision tree is referred to as a base learner. The number of base learners is specified by the user. Each base learner will be different from the previous because of the random selection of a subset of the variables.

[15] More explanation on how PCA works can be found in Chatfield and Collins (1980)

In the classification phase, the test data point is fed to each base learner and the average of each class prediction probability is calculated. The data point will be assigned to the class with the highest average. A more detailed derivation of this method can be found in the article by Rodriguez *et al.* (2006).

This ensemble method follows a procedure similar to that of *bagging* and *random forests* in decision tree theory. The *rotation forests* method promotes individual accuracy and diversity within the ensemble. Accuracy is achieved by retaining all principal components when creating the rotation matrix. Diversity is encouraged through the feature extraction of each base learner. The method, however, looses interpretability of variable importance once the PCA is applied to the base learner.

3.5 Performance measures

In this section, we give a simple explanation of the performance measures used in the two stages of model testing. A comparison of the performance measures is also done and is discussed in the Chapter 4.

3.5.1 Confusion matrix and measures for binary classification

A confusion matrix is a contingency table of predicted versus actual class classification. The values in the rows represent the number of data points in the actual classification whilst the columns represent the predicted classes of the data points or vice versa. The values on the diagonal are correct classifications and all other values are misclassifications. The aim with all classification models is to reduce the number of the misclassifications. Table 3.1 gives the structure of the confusion matrix for the binary classification done for this dissertation. Positive classification refers to the *Employed* class and negative classification is *Unemployed* class.

Numerous measures can be derived from the confusion matrix (Sokolova and Lapalme, 2009). However, the measures chosen to compare the different models are based on their property of invariance as described in the article (Ballabio *et al.*, 2018) as well as the financial cost of having a bad prediction. Therefore, we chose measures that are sensitive to particular changes in the confusion matrix. It is important to see difference in measures when there is a change to a positive and negative class prediction, therefore *recall* and *specificity* were selected.

Recall, also known as *sensitivity* or as the True Positive rate (TPR), is calculated as follows, using the definitions in table 3.1;

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

It is the rate of those test data points predicted as *Employed* when they are actually *Employed*. The goal with this measure is to show the model's ability to identify *Employed* data points. If this measure is low, i.e. approaching 0, this implies that the data points that should be identified as *Employed* are being heavily misclassified and defeating the purpose of the model.

Specificity, also known as the True Negative rate (TNR), is calculated as follows, using the definitions in table 3.1;

$$TNR = \frac{TN}{TN + FP} \tag{3.2}$$

It can be seen as the rate of those test data points predicted as *Unemployed* when they are actually *Unemployed*. The goal with this measure is to identify the model's ability to reject data points of other classes (Ballabio *et al.*, 2018). The False Positive rate (FPR) can be derived from the *specificity* where;

$$\begin{aligned} FPR &= \frac{FP}{FP + TN} \\ &= 1 - specificity \end{aligned} \tag{3.3}$$

The lower the FPR the better, as a high FPR indicates that the model is misclassifying *Unemployed* data points as *Employed*, and this could ultimately have negative financial implications.
FPR and sensitivity are important in deriving the Receiver Operating Characteristic (ROC) curve[16].

Accuracy is measure of overall effectiveness of a classifier. It is a commonly used measure for assessing classification accuracy. It is calculated, using the definitions in table 3.1, as;

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.4}$$

The measure experiences bias against imbalanced classes and although it is usually the first to be considered it can not serve as the only measure of performance.

3.5.2 Receiver Operating Characteristic (ROC) Curves and Area under the curve (AUC)

Receiver Operating Characteristic (ROC) Curve is a figure that visualises the trade-off between the True Positive rate (TPR) and the False Positive rate (FPR) along the range of all possible prediction probability thresholds that would classify the data points into classes. A good model, when analysing the ROC, will generally have a high TPR and low FPR, much like *ROC1* of figure 3.4. This is an indication that the

[16] more elaboration in section 3.5.2

Predicted→ Actual↓	Employed	Unemployed
Employed	TP	FN
Unemployed	FP	TN

Tab. 3.1: Confusion matrix with a binary classification for the case of predicting Employment where TP = True Positive, FN = False Negative, FP = False Positive and TN = True Negative

positive events (i.e. predicting employment) will be predicted as positive events. If the ROC curve demonstrates a diagonal line, the TPR and FPR of positive events to negative events is the same for all prediction thresholds and the model is seen as no better than a random classifier (Han *et al.*, 2011). The curve can also be used to find the optimal prediction probability threshold (Thomas *et al.*, 2017). There are a number of measures that can be calculated using the ROC curve (Ballabio *et al.*, 2018). However, only the area under the ROC curve will be discussed here.

The area under the ROC curve (AUC) is a measure of the model's ability to avoid false classification i.e. its discriminatory power (Sokolova and Lapalme, 2009). Therefore, the AUC measures how well is the model classifying *Employed (Unemployed)* data points as *Employed (Unemployed)* predictions. A value approaching 1 indicates that the model is very likely to be classifying correctly and a value approaching 0.5 means the model has a 50/50 chance of being correct.

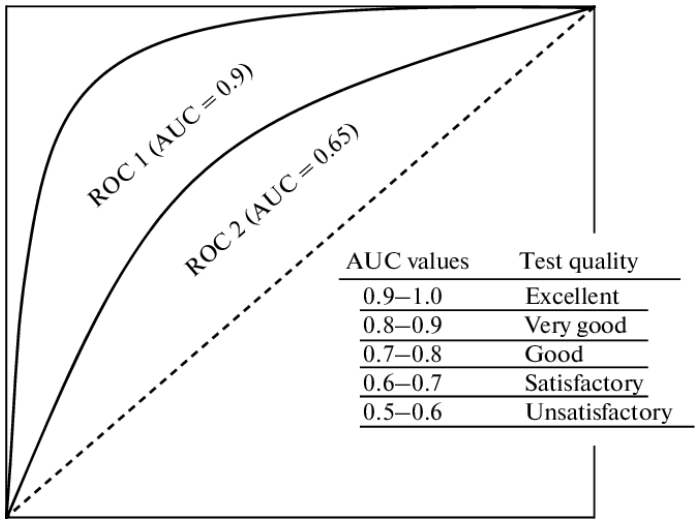
Sokolova and Lapalme (2009) suggest that the value of the AUC, apart from calculating the area, can be derived using information from the confusion matrix (see 3.5.1) ;

$$\begin{aligned} &\frac{1}{2} \times \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \\ &= \frac{1}{2} \times (specificity + sensitivity) \end{aligned} \tag{3.5}$$

3.5.3 Gini Index

The Gini index or Gini coefficient, not to be confused with the definition used in Economics, is a measure of model performance over all possible values of the percentage threshold that determines whether a data point is classified as *Employed* or *Unemployed* (Thomas *et al.*, 2017). It is also referred to as Youden's J Index (Ballabio *et al.*, 2018; Youden, 1950). It can be derived from the AUC (see 3.5.2) and is calculated as;

$$2 \times AUC - 1 \tag{3.6}$$



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Fig. 3.4: General structures of Receiver Operating Characteristic (ROC) Curves
This figure demonstrates a breakdown of what the different values of the Area under the ROC curve could describe about the model's overall discriminatory power (*ROC Analysis*, 2015)

The result is a value between 0 and 1 where 0 indicates that the model performs no better than a random classifier and 1 indicates that the model is a perfect classifier. The Gini index linearly transforms the AUC into a percentage value that calculates the performance of the classifier.

It has the disadvantage of considering all possible threshold percentage values as opposed to a selected range of threshold values.

This chapter has provided an explanation of the components in the model testing procedure. It started off with the data set description to provide more understanding of the kind of data that is used. The section that follows that, is the data preprocessing section that explained how the data is transformed as well as the reasoning of why the transformation is required. There is also some elaboration of how the two-stage testing procedure is going to work with the different models and versions of the data set. A further explanation of the type of models considered for testing as well as the performance measures used in evaluating the models is given. The final model parameters, for the models that require parameters, can be seen in table B.2. In the following chapter (4) we discuss the results of the two-stage testing. It also includes some conclusions made on the data based on the data exploration that is performed.

Chapter 4

Results

This chapter is broken up into two sections. The first section is a discussion on the behaviours of the top 48 variables as per the data reduction procedure mentioned in section 3.2. The second section is a review of the model performances in the two-stage testing procedure[1].

Data Exploration

In stage one of the model testing process, the variables of the data set used included the top 48 variables for predicting *Employability* according to the feature selection process [2]. The backward selection process discussed under *Data reduction* of section 3.2 is the feature selection process used. Figure B.1 shows the prediction accuracy of the model against the number of variables used in the model. The model with the best overall accuracy is the one with 48 variables. Table B.1 provides the question description as well as the options of the most important variables for predicting *Employability*, according to the backward selection process. It is important to note that the variable names used are in the form of the original question number of the survey. This was done for the ease of reference and it would have been complicated to rename the original 178 questions.

The most dominant independent variable was the candidates employment status being *Employed in the informal sector*[3], between graduating and current formal employment. The employment status of the candidate, between graduating and current employment, played a major role in *Employability* prediction by accounting for 7 of the 48 most vital variables.

[1] Two-stage testing as explained in chapter 3

[2] Note that the some of the 48 variables include the dummy variables that were created and there is repetition of an original categorical variable at different category levels

[3] This means that one works for an unregistered, informal trader, maker or seller of goods and services *South Africa - Graduation Destination Survey 2012* (2015)

Other noteworthy variables were whether the candidate looked for work by *responding to job advertisements through employment websites*[4], as well as if the candidates primary method of finding current employment was *through the help of a lecturer*[5]. Thus, the combination of the students efforts of looking for a job through online job advertisements and the assistance received from their lecturers are key features to finding employment after graduating.

It is also found that whether the candidate is studying *full-time* or *part-time* is important[6]. One of many assumptions as to why the variable is important, could be that a part time student is more likely to get employed because they already getting work experience.

The degree level also plays an important role in *Employability* prediction especially if the degree is a *Masters degree by coursework and research*[7].

The age range to which the candidate belonged is also important, but surprisingly the age range of between 22 and 30 was not important. We were more interested in this age range, as it includes a lot more recent undergraduates. This could imply that this age range had such diversity between the *Employed* and *Unemployed* data points that it would be more difficult to predict *Employability*.

The *Neighbour distance plot*[8] is meant to help us identify potential clusters within the data. The clusters are formed by firstly matching the data points to the nodes that display properties similar to that of the data point and then identify how similar the properties of the nodes are from each other by using a distance measure. The procedure of determining the number of nodes as well as the formula for dividing the data according to certain properties is outside the scope of this dissertation [9]. Heatmaps of the *neighbour distance* plot as well as the individual variable plots from table B.1 are constructed using Self-Organising Maps as developed by Kohonen (1990). Looking at Figure 4.1, the darker blue (red) the node is the more (dis)similar the node is to its neighbour. We would like to see two distinct clusters, where the one will represent an *Employed* cluster and the other an *Unemployed* cluster. Yet, from figure 4.1, we observe one oddly shaped cluster in the bottom right corner of the plot as many dark blue nodes are seen in that region. Other clusters are difficult to identify. There are a few dark red nodes which suggest significant difference to neighbours but these nodes could also be housing outliers.

[4] Question 3.1.13 Answer:f of the *South Africa - Graduation Destination Survey 2012* (2015) survey
[5] Question 3.4.6 Answer: 2 of the *South Africa - Graduation Destination Survey 2012* (2015) survey
[6] Question 2.1 Answer:1 or 2 of the *South Africa - Graduation Destination Survey 2012* (2015) data
[7] Question 2.4.1 Answer:1 of the *South Africa - Graduation Destination Survey 2012* (2015) survey
[8] also known as U-matrix
[9] However, better elaboration and explanation of methods used are in Kohonen (1990); Hua *et al.* (2009)

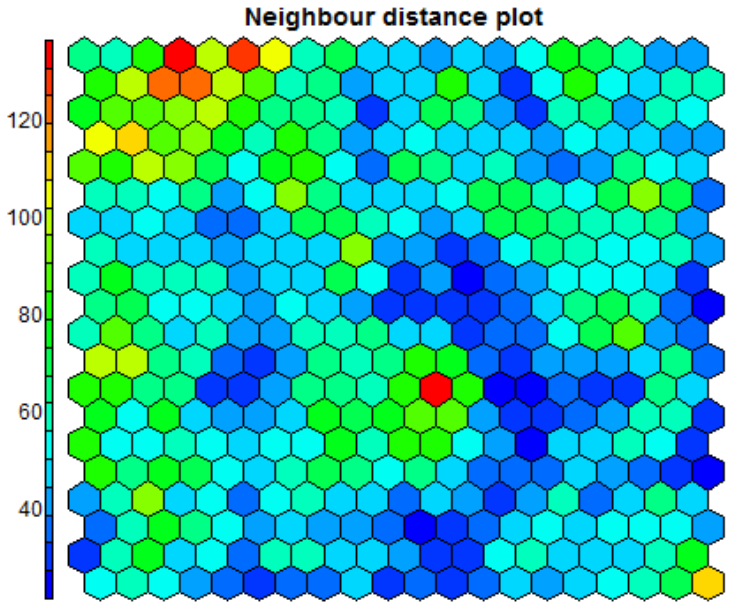


Fig. 4.1: Neighbour distance plot of the *South Africa - Graduation Destination Survey 2012* (2015) survey data
This figure shows the distance between the nodes where the darker blue (red) the node is the more (dis)similar the node is to its neighbour. An oddly shaped cluster is seen in the bottom right corner of the plot otherwise the plot seems disjoint and clusters are hard to identify

After looking at the *Neighbour distance plot*, we look at the individual variable *property plots* to examine which variables contribute more to certain clusters, i.e. what are the properties that cause the separation of the clusters. Thus, the aim is observe the *property plots* that behave similarly or show contrast to that of the *Neighbour distance plot*. Surprisingly, the *property plots* of the top 48 variables[10] do not display patterns similar to that of the *Neighbour distance plot*. This discovery needs to be investigated further as it is unusual and makes it difficult to determine which variables acted as cluster dividers.

However, while observing these individual property plots we can also identify which properties compliment or contrast each other. The property plot of *q2_1* and *q3_3*, fig 4.2, behave eerily similar where the former shows information about whether the candidate did their degree part-time or full-time and the latter variable indicates the employment status of the candidate before pursuing their degree. This relationship would make sense because it is more likely that a candidate that was unemployed before pursuing their degree is more likely to be doing their degree full-time[11].

Variables *q3_3* and *q3_3_2* seem to show inverse relationships, refer to figure 4.3. Variable *q3_3* seeks to identify whether the candidate worked before their 2010 qualification whilst *q3_3_2* indicates whether the candidate who had employment before their degree is still working at the same place after obtaining their degree. The difficulty with comparing this is that *q3_3* contains the responses of those candidates that did not work prior to the degree whilst *q3_3_2* does not. This could explain the distribution as possible empty nodes could be in the data. The dark red nodes in the *q3_3_2* plot could contain the empty nodes which corresponds to dark blue nodes in the *q3_3* plot that could be the responses of those candidates that were unemployed before pursuing their 2010 qualification. This way we could deduce that the candidates that were unemployed before their 2010 qualification are represented by the dark blue nodes in the *q3_3* plot.

[10] Table B.1

[11] This is not the only case, it is just an example of where this case seems relevant

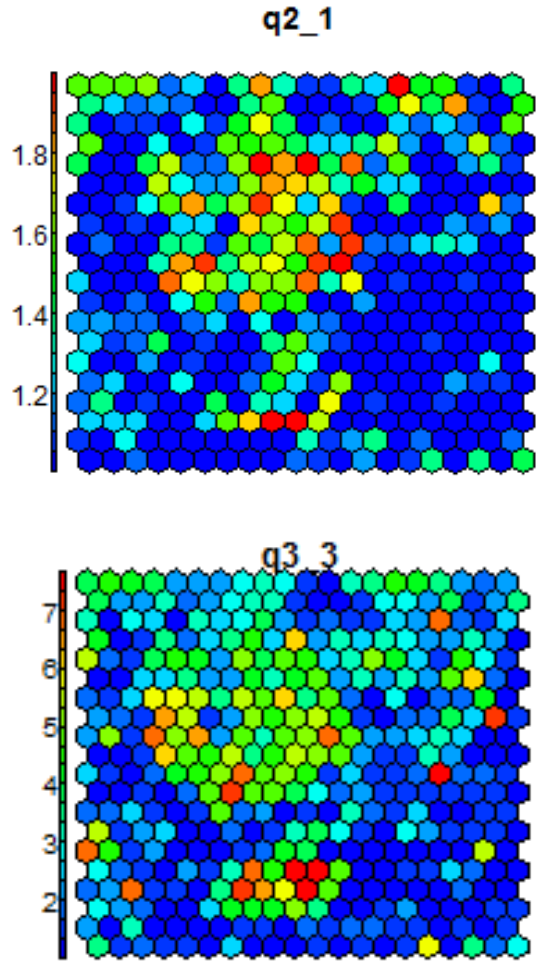


Fig. 4.2: Property plot comparison of question 2.1 and 3.3 of *South Africa - Graduation Destination Survey 2012* (2015) data

This figure shows the distance between the nodes where the darker blue (red) the node is the more (dis)similar the node's response is to its neighbour. The property plot of *q2_1* and *q3_3* behave eerily similar where the *q2_1* plot shows information distribution about whether the candidate did their degree part-time or full-time and the *q3_3* plot indicates the employment status of the candidate before pursuing their degree. The figures tend to have darker blue nodes at the same regions of their respective maps which shows that similar properties are seen at the data points in the nodes of these plots. We suggest that one explanation could be that a candidate that was unemployed before pursuing their degree is more likely to be doing their degree full-time. So if we assume the dark blue nodes in the *q2_1* plot represent the unemployed before 2010 qualification nodes and the dark blue nodes in the *q3_3* plot represent the candidates doing the degree full-time*.

* this is just one interpretation

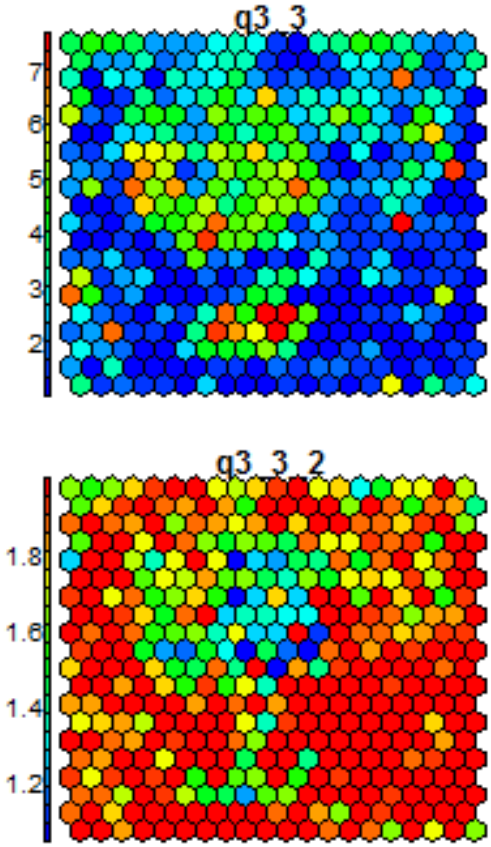


Fig. 4.3: Property plot comparison of question 3.3 and 3.3.2 of *South Africa - Graduation Destination Survey 2012* (2015) data

This figure shows the distance between the nodes where the darker blue (red) the node is the more (dis)similar the node's response is to its neighbour. Variable *q3_3* seeks to identify whether the candidate worked before their 2010 qualification whilst *q3_3_2* indicates whether the candidate who had employment before their degree is still working at the same place after obtaining their degree. The dark red nodes could contain the empty nodes which corresponds to dark blue nodes in the *q3_3* plot that could be the responses of those candidates that were unemployed before pursuing their 2010 qualification.

There are individual property plots that display patterns worth noting, refer to figures 4.4 and 4.5.
Observing the heatmap of the variable *q3_4_13f*, refer to figure 4.4 which translates to the candidate *looking for work by responding to the job ads on employment websites*, the distribution is almost completely similar i.e. not much difference in the responses to this variable between the nodes. This suggests that majority of the population behaved similarly in the response to this particular answer in question 3.4.13 of the *South Africa - Graduation Destination Survey 2012* (2015) data. This could imply that either most candidates either did or did not look for work by responding to the job ads on employment websites.

Question *q3_4_6* involved identifying the primary method used by the candidate to find their current job. Figure 4.5 displays an interesting pattern for this question that seems to have 2 clear groupings that are not necessarily clusters because the one group is more or less dark red nodes and that implies that the nodes in that region are very different in their behaviour for that particular question, however, a very distinct cluster in the bottom right corner of the plot is observed. This cluster could represent the nodes that had *through the help of a lecturer* as their primary method of finding their current employment.

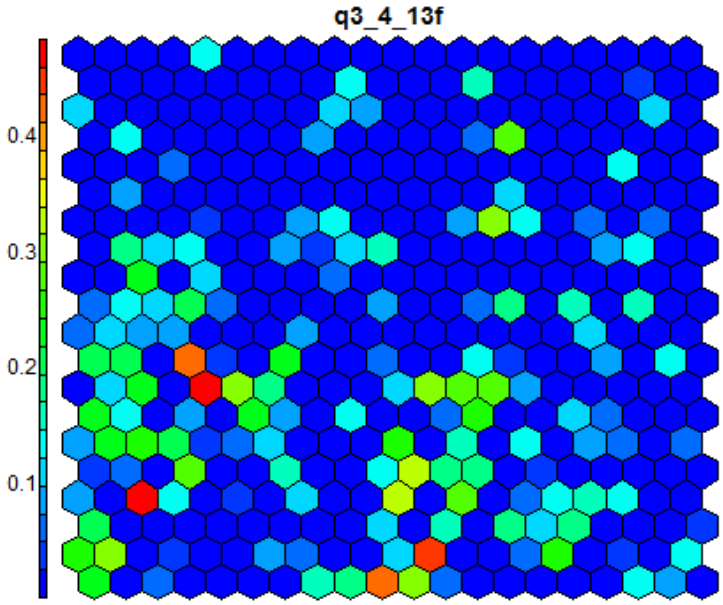


Fig. 4.4: Property plot of question 3.4.13 answer:f of *South Africa - Graduation Destination Survey 2012* (2015) data
This figure shows the distance between the nodes where the darker blue (red) the node is the more (dis)similar the node's response is to its neighbour. The property plot of the distribution is almost completely similar (not much difference between the information contained in the nodes) suggesting that the majority of the population responded similarly to this particularly question and answer in the survey.

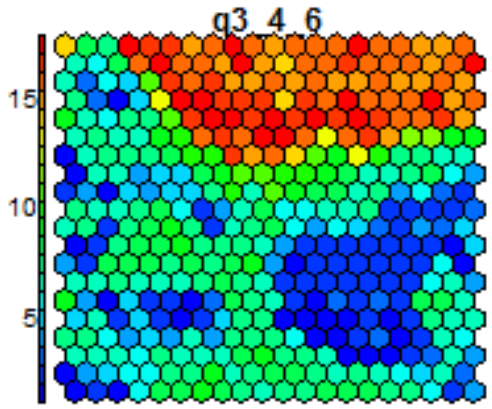


Fig. 4.5: Property plot of question 3.4.6 of *South Africa - Graduation Destination Survey 2012* (2015) data

This figure shows the distance between the nodes where the darker blue (red) the node is the more (dis)similar the node's response is to its neighbour. The question involved identifying the primary method used by the candidate to find their current job. The property plot of the question displays an interesting pattern that seems to have 3 groupings that are not necessarily clusters. The one group is more or less dark red and that implies that the nodes in that region of very different in their behaviour for that particular question. There is also a very distinct cluster in the bottom right corner of the plot.

Model Performance Results

The best model for *Employability* prediction is selected through a two-stage testing process. In stage one, all models discussed in section 3.4 are trained and tested using the abridged *South Africa - Graduation Destination Survey 2012* (2015) data set[12]. Table 4.1 summarises the performance results of stage-one testing and provides the rank of the model per performance measure as well as overall average rank of the model. The top 3 models are selected using the top 3 highest average ranks, i.e. *Adaboost*, *Polynomial Support Vector Machine* and *Bagged decision trees*. The parameters for these models can be found in table B.2. Generally, all the models performed extremely well with the test set. Many factors could have contributed to this phenomenon including;

- the training and test set data set being extremely similar in nature. Similar with regards to the class imbalance of the dependent variable (*Employability*) or in the data points makeup[13]. The class imbalance was fixed through the stratified sampling, however, the data point similarity was not tested but could possibly account for the test set doing so well.
- that the variables selected are all vital in predicting *Employability* despite the model used.
- that there were inclusion of too many variables to begin with and maybe more variable reduction could have taken place.

All the models had great *Accuracy* and *Recall* where all of them are displaying values of above 90% . A value of 100% for *Recall* is observed for the *Bagged decision trees* and *Adaboost* models. Thus, these models always predicted the *Employed* data points correctly.

The *Specificity* did not go as well but still good nonetheless where the worst performer is the *KNN*[14]. This suggests that the *KNN* model does not predict the *Unemployed* data points as well as the other models.

The *Adaboost* model has the best AUC and Gini index performance values (value of 0.98361 and 0.96721 respectively). Therefore, of all the models the *Adaboost*[15] has the best ability to avoid false classification at all possible percentage threshold values.

The *individual classifiers*[16] collectively performed worse overall. This should be expected given that the ensemble methods were created to improve the perfor-

[12] refer to chapter 3 and figure 3.1 for details about what is meant by the abridged data set
[13] i.e. the data between candidates are similar
[14] *k*-Nearest Neighbour
[15] Refer to table B.3 to view the results of the variable importance of the *Adaboost* model
[16] this is including the neural network

mance of a individual classifiers Han *et al.* (2011).

[17] Refer to figure B.9 to view the results of the variable importance of the *Bagged decision tree* model

	Accuracy	Recall	Specificity	AUC	Gini index	Average rank
Logistic Regression	0.98347 (4)	0.98895 (7)	0.95082 (3)	0.96989 (7)	0.93977 (7)	5.6
k-Nearest Neigh-bours	0.92975 (9)	0.95580 (9)	0.85246 (9)	0.90413 (9)	0.80826 (9)	9
Radial SVM	0.98347 (4)	0.99448 (3)	0.95082 (3)	0.97265 (4)	0.94529 (5)	3.8
Polynomial SVM	0.98760 (2)	0.99448 (3)	0.96721 (1)	0.98084 (2)	0.96168 (2)	2
Multi-layer NN	0.96694 (8)	0.98343 (8)	0.91803 (8)	0.95073 (8)	0.90146 (8)	8
Bagged decision trees[17]	0.98760 (2)	1.00000 (1)	0.95082 (3)	0.97541 (3)	0.95082 (3)	2.4
Adaboost	0.99174 (1)	1.00000 (1)	0.96721 (1)	0.98361 (1)	0.96721 (1)	1
Rotation Forest	0.98347 (4)	0.99448 (3)	0.95082 (3)	0.97265 (4)	0.94529 (5)	3.8
Simple average model	0.98347 (4)	0.99448 (3)	0.95082 (3)	0.97265 (4)	0.94530 (4)	3.6

Tab. 4.1: Performance Results of the test data on the differently trained machine learning models where SVM stands for Support Vector Machine, NN stands for Neural Network and Adaboost refers to the Adaptive Boosting model using decision trees as base classifiers. The number in parenthesis is the rank number of that model for that performance measure.

	Accuracy	Recall	Specificity	AUC	Gini index	Average rank
Polynomial SVM	0.87597 (2)	0.93782 (2)	0.69204 (3)	0.81507 (3)	0.63013 (3)	2.6
Bagged decision trees	0.89147 (1)	0.95337 (1)	0.70769 (2)	0.83053 (2)	0.66106 (2)	1.6
Adaboost	0.87209 (3)	0.91192 (3)	0.75385 (1)	0.83288 (1)	0.66576 (1)	1.8

Tab. 4.2: Performance Results of the test data on the top 3 machine learning algorithms as per the results in table 4.1 where SVM stands for Support Vector Machine and Adaboost refers to the Adaptive Boosting model using decision trees as base classifiers. The number in parenthesis is the rank number of that model for that performance measure.

The next stage of the model testing is to remove the most important variables[18] from the abridged data set at the point before the data reduction was performed on the data in the data preprocessing step[19], refer to figure 3.2. We now test the top 3 models with the 352 independent variables that were initially rejected by the data reduction technique. Table 4.2 shows the performance results of the models using the abridged data. The *Bagged decision trees* has a slight edge over the *Adaboost* model. The overall performance values have indeed decreased, as expected given the new data set.

Table 4.3 gives a better view of the differences between the performance values of the models from the the different testing stages. A *t*-test is performed to identify whether the differences were significant, and the *p*-value for each performance measure is also included in table 4.3[20].

The average difference change for *Specificity* and *Gini index* i.e. -25% and -32% respectively, show the models seem to have become worse at correctly classifying *Unemployed* data points as well as showing a decrease in its overall ability to predict correctly regardless of probability threshold. There is also a significant change in the overall *Accuracy* of the models, as a result of the poor performance of predicting *Unemployed* data points . When analysing the *p*-value, the performance difference

[18] as per the data reduction process discussed in section 3.2 and resulting most important variables displayed in Table B.1

[19] Note that the class imbalance was kept constant

[20] More explanation of the t-test performed, the p-value as well as significance level is in section B.4 of the Appendix

	Accuracy Δ(st1, st2)	Recall Δ(st1, st2)	Specificity Δ(st1, st2)	AUC Δ(st1, st2)	Gini index Δ(st1, st2)
Polynomial SVM	-11%	-6%	-28%	-17%	-34%
Bagged decision trees	-10%	-5%	26%	-15%	-30%
Adaboost	-12%	-9%	-22%	-15%	-31%
Average Change	-11%	-6%	-25%	-16%	-32%
p-value	0.0342*	0.2307	0.045*	0.0144*	0.0144*

Tab. 4.3: Summary of the changes in the value of the performance measures when a stress test is performed on the top 3 machine learning models as well as the resulting p-values of the *t*-test performed on the differences where * indicates significant at a 5% level of significance

can be seen as statistically significant depending on the level of significance. Looking at a 5% level of significance, the only measure that does not have a significant difference is *Specificity*. Thus, the models still predicts the *Employed* data points well. This outcome supports the argument that all the variables in this data set are important to predict if one is *Employed* regardless of the model used. Yet, the variables that were discarded by the data reduction technique proved to hardly contribute to predicting *Unemployed* data points .

The aim of the drastic variable change is to see which model would remain reliable when certain variables were no longer available[21]. The best model overall for predicting *Employability* is the *Bagged decision tree*. This decision is based on its consistence through both stages of testing and having the lowest difference in performance between the two testing stages. However, the *Adaboost* is a close second and it also happens to be homogeneous ensemble. Surprisingly, the heterogeneous ensemble *(Simple average model)* did not even pass the first round of testing. This could be because of the overall poor performance of the individual classifiers.

This chapter summarises the results of the two-stage testing process as well as

[21] The important variables are seen in table B.1 in the Appendix which have a further description of the variable.

included some findings made from the data exploration. There was unfavourable findings in the data exploration that have made it more difficult to understand why the top 48 variables, as per the feature selection process were actually chosen. This has encouraged further investigation and possibly reevaluating the feature selection process used. The performance results of the models are, however, as expected. The ensemble methods dominated the performance results where the top 2 models were homogenous ensemble methods. In the following chapter, we do provide a more comprehensive conclusion of these results. There is also a discussion of future recommendations for our findings.

Chapter 5

Conclusion

The aim of this dissertation is to build an *Employability* index. We achieve this by investigating and testing 9 machine learning algorithms[1]. The best model is chosen by using a two-stage testing process and 5 prediction measures[2]. The two-stage testing process involves two versions of the abridged *South Africa - Graduation Destination Survey 2012* (2015) data set. In stage one, the most important variables are included in the data set. The top 3 models are selected according to their performance values . In stage two, we include all other unimportant variables in the data set. The purpose for doing this is to create a stress test on the models to test whether the models will still predict as well as before. The top 3 models are then re-trained and tested. The best model for predicting *Employability* is found to be a *Bagged decision trees*.

The results of the feature selection and data exploration are generally unexpected. They support the notion of why it is important to do these processes before evaluating a model. According to the findings of the data exploration, the following conclusions are made;

- Omit more of the variables from the onset because some of the variables are not relevant to a student who has not been employed before/yet.
- The correlation/association between the data points as well as between variables needs to be tested as part of the data preprocessing procedure.
- The individual property heatmaps[3] of the most important variables[4] did not show behaviours like that of the entire data set, as seen by the *Neighbour distance* plot. A discovery that requires further investigation to understand as to why this is the case. Naturally, we would expect these plots to demonstrate similar behaviours. Given this outcome we could look at a different feature selection method. Alternatively, we could even create heatmaps of all the

[1] Refer to section 3.4
[2] More details in section 3.5
[3] created using Self-organising maps
[4] refer to B.1

variables in the data set and select the variables that has a relationship with that of the *Neighbour distance plot*.

- More involved data exploration could have been done by looking at the behaviour of the data for *Employed* or *Unemployed* data points.
- Variables such as *GPA* and the degree field were not deemed important. This is an unexpected result as naturally one would think that this would be related to one getting employed. This could also be a result of inclusion of many variables that are not relevant to a graduate that has not had employment before their qualification. These variables dominate the variables that should be relevant to all candidates.

A mixture of machine learning algorithms are trained and tested. This mixture includes;

- individual classifiers such as Logistic Regression, *k*-nearest neighbours and support vector machines (polynomial and radial),
- a neural network,
- homogeneous ensembles such as Rotation Forest, Adaboost and a Bagged decision tree, as well as
- a heterogeneous ensemble in the form of a *simple weighted average model* that contains all the individual classifiers as base classifiers.

The models all performed fairly well on the test set in both stages. In stage one, the individual classifiers as well as the neural network did not perform as well as the other types of machine learning algorithms overall. However, the polynomial support vector machine is one of the top 3 models based on performance measures. This result suggests that individual classifiers can still be seen as relevant in machine learning applications. The homogeneous ensembles performed really well were 3 of them appear in the top 5 models based on performance results. Another unexpected result included that of the disappointing performance of the neural network. It placed second to last in terms of performance. However, a different structure of the neural network could have produce better results. The worst performer overall was the *k*-nearest neighbour. Looking at the figure 4.1 this result makes more sense. There was only one possible cluster seen and even that cluster was oddly shaped. However, the behaviours of the neighbours did not seem similar enough to suggest consistency in prediction using the neighbour information.
In stage two, there is an expected decrease in performance results. Some performance measures are showing some drastic percentage differences, refer to table 4.3. The *Gini index* measure displayed the largest decrease of about -32% on average between the models. This implies that in this stage of testing the models had some

difficulty performing well overall probability thresholds of these models. Through all the changes, the *bagged decision tree* prevailed. An investigation of whether the use of bagged decision trees in the feature selection as well as the missing value imputation discussed in Chapter 3.2 could have influenced this result.

Besides the main aim of using the resulting index to assist a credit scoring application process, other uses were identified. We suggest that the *Employability* index can also be used for targeted marketing by a bank or a retailer etc. It could also be used in conjunction with a student loan securitisation model or student loan application process. Bursary-granting institutions should consider using the index when evaluating a potential student bursar. The index can also contribute to the *Graduate Employability prediction* literature. It can aid in determining an effective way for different kinds of students to approach their job hunting experience.

Future Recommendations

For future studies based on this discussion, one could look at;

- combining student data set with the National Income Dynamic Study (NIDS) data to predict *employability* as well as the potential salary range of a candidate;
- the use of different performance measures including the H-measure and the Kolmogorov-Smirnov statistic as well as provide more elaboration on the type I and II errors of the data;
- the use of more data sets and different kind of data sets as done by Baesens *et al.* (2015). One of the limitations of this project is that we did not create our own data set. There is better control of the data entered into the model, if the variables initially required are actually obtained by the researchers as opposed to using data that was intended for another purpose.
- more extensive data exploration using more unsupervised learning techniques such as *association rule mining*;
- the use of different and more complicated models like the ones mentioned in Tsai (2014); Marqués *et al.* (2012); Antonini *et al.* (2010). Additionally, one can do a comparison of the simple models versus complicated models;
- attempting to prove that there is a correlation between employability and the paying off of a loan in similar fashion as that suggested by Ballabio *et al.* (2018);
- attempting to prove that the probability of obtaining credit is higher when a graduate discloses their student data for the use of a *employability* index; and

- introduce the use of regularized models to reduce the potential overfitting of models to the training data.

This dissertation has achieved its aim but with unexpected findings. The data exploration was disappointing but the performance results of the models was not surprising. Many future recommendations are discussed as there are many other approaches that could have been used to achieve the target of this dissertation. This work hopes to pioneer the literature that combines *graduate employability* with credit application scoring.

Bibliography

Achieving Effective Financial Inclusion in South Africa: A Payments Perspective (2014).

Amari, S., Murata, N., Muller, K. ., Finke, M. and Yang, H. H. (1997). Asymptotic statistical theory of overtraining and cross-validation, *IEEE Transactions on Neural Networks* **8**(5): 985–996.

Antonini, G., Elisseeff, A. and Paleologo, G. (2010). Subagging for credit scoring models, *European Journal of Operational Research* **201**(2): 490 – 499.
URL: *http://www.sciencedirect.com/science/article/pii/S0377221709001532*

Baesens, B., Lessmann, S., Seow, H.-V. and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research* **247**(1): 124 – 136.
URL: *http://www.sciencedirect.com/science/article/pii/S0377221715004208*

Ballabio, D., Grisoni, F. and Todeschini, R. (2018). Multivariate comparison of classification performance measures, *Chemometrics and Intelligent Laboratory Systems* **174**: 33–44.

Berg, T., Burg, V., Gombović, A. and Puri, M. (2018). On the rise of fintechs–credit scoring using digital footprints, *Technical report*, National Bureau of Economic Research.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and regression trees.

Brownlee, J. (2016). Classification and regression trees for machine learning, https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

Chatfield, C. and Collins, A. J. (1980). *Principal component analysis*, Springer US, Boston, MA, pp. 57–81.
URL: *https://doi.org/10.1007/978-1-4899-3184-9_4*

Decision tree ensemble using Bagging algorithm (2009). https://commons.wikimedia.org/wiki/File:DTE_Bagging.png.

Freund, Y., Schapire, R. and Abe, N. (1999). A short introduction to boosting, *Journal-Japanese Society For Artificial Intelligence* **14**(771-780): 1612.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of computer and system sciences* **55**(1): 119–139.

García-Peñalvo, F. J., Cruz-Benito, J., Martín-González, M., Vázquez-Ingelmo, A., Sánchez-Prieto, J. C. and Therón, R. (2018). Proposing a machine learning approach to analyze and predict employment and its factors, *International Journal of Interactive Multimedia and Artificial Intelligence* **5**(2): 39–45.
URL: *http://www.ijimai.org/journal/sites/default/files/files/2018/02/ijimai_5_2_5_pdf_12552.pdf*

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques*, Elsevier.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **160**(3): 523–541.

Hua, G., Skaletsky, M. and Westermann, K. (2009). Exploratory analysis of cia factbook data using kohonen self-organizing maps, *Case Studies In Business, Industry And Government Statistics* **3**(1): 48–59.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. 112, Springer.

Jantawan, B. and Tsai, C. (2013). The application of data mining to build classification model for predicting graduate employment, *CoRR* **abs/1312.7123**.
URL: *http://arxiv.org/abs/1312.7123*

Kohonen, T. (1990). The self-organizing map, *Proceedings of the IEEE* **78**(9): 1464–1480.

Lin, M., Prabhala, N. R. and Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending, *Management Science* **59**(1): 17–35.

Marqués, A., García, V. and Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment, *Expert Systems with Applications* **39**(12): 10916–10922.

Mayo, M. (2016). Decision tree classifiers: A concise technical overview, https://www.kdnuggets.com/2016/10/decision-trees-concise-technical-overview.html.

Mishra, T., Kumar, D. and Gupta, S. (2016). Students' employability prediction model through data mining, *International Journal of Applied Engineering Research* **11**(4): 2275–2282.

Mpahlwa, M. (2006). National credit regulations, 2006, (3): 18.1.

Pathways from university to work (2013). *resreport*, Cape Higher Education Consortium (CHEC).

Piad, K. C., Dumlao, M., Ballera, M. A. and Ambat, S. C. (2016). Predicting it employability using data mining techniques, *2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC)*, pp. 26–30.

Rahman, M. G. and Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques, *Knowledge-Based Systems* **53**: 51 – 65.
URL: *http://www.sciencedirect.com/science/article/pii/S0950705113002591*

ROC Analysis (2015). http://mlwiki.org/index.php/ROC_Analysis.

Rodriguez, J. J., Kuncheva, L. I. and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method, *IEEE transactions on pattern analysis and machine intelligence* **28**(10): 1619–1630.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors, *nature* **323**(6088): 533.

Scheule, H., Rösch, D. and Baesens, B. (2017). *Credit Risk Analytics: The R Companion*, Create Space Independent Publishing Platform.
URL: *https://epub.uni-regensburg.de/36395/*

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing & Management* **45**(4): 427 – 437.
URL: *http://www.sciencedirect.com/science/article/pii/S0306457309000259*

South Africa - Graduation Destination Survey 2012 (2015). Cape Higher Education Consortium [producer].
URL: *http://www.datafirst.uct.ac.za/dataportal/index.php*

Šušteršič, M., Mramor, D. and Zupan, J. (2009). Consumer credit scoring models with limited data, *Expert Systems with Applications* **36**(3, Part 1): 4736 – 4744.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417408002996*

Thomas, L., Crook, J. and Edelman, D. (2017). *Credit scoring and its applications*, Vol. 2, Siam.

TransUnions New CreditVision Model Uses Alternative and Trended Data to Better Predict Credit Risk, Providing Millions of South Africans with More Opportunities to Gain Access to Credit (2018).
URL: *https://newsroom.transunion.co.za/transunions-new-creditvision-model-uses-alternative-and-trended-data-to-better-predict-credit-risk-providing-millions-of-south-africans-with-more-opportunities-to-gain-access-to-credit/*

Trevor, H., Robert, T. and JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction.

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress, *Information Fusion* **16**: 46–58.

Youden, W. J. (1950). Index for rating diagnostic tests, *Cancer* **3**(1): 32–35.

Appendix A

Decision trees

The classification decision tree is a flowchart-like tree structure, as seen in the figure A.1. The tree begins at the root (the topmost node) and works its way through the test criteria presented by internal nodes to ultimately reach a leaf node that contains a class prediction (i.e. classify the data point as *Employed* or *Unemployed*) (Han *et al.*, 2011). At every step of an internal node creation, all of the independent variables are considered and the resulting criteria is based on the best performing variable under a specific error rate. Figure A.2 gives a high-level look at the algorithm used to derive the tree and the test criteria at the root and internal nodes. It is discussed more extensively in the article by Breiman *et al.* (1984).
Decision trees are used in many instances including the following;

- feature selection (Han *et al.*, 2011)
- rule extraction (Han *et al.*, 2011)
- missing value imputation (Rahman and Islam, 2013)
- classification or regression prediction (Breiman *et al.*, 1984)

just to name a few.

Table A.1 summarises the advantages and disadvantages of decision trees as suggested by James *et al.* (2013).

Advantages	Disadvantages
Easy model to explain and interpret	One tree does not provide the same kind of prediction accuracy and thus methods like *Bagging* and *Random forests* were developed.
Do not need data preprocessing of categorical variables into dummy variables	Not a robust model and so is very sensitive to change in the input data

Tab. A.1: Advantages and Disadvantages of Classification Decision Trees

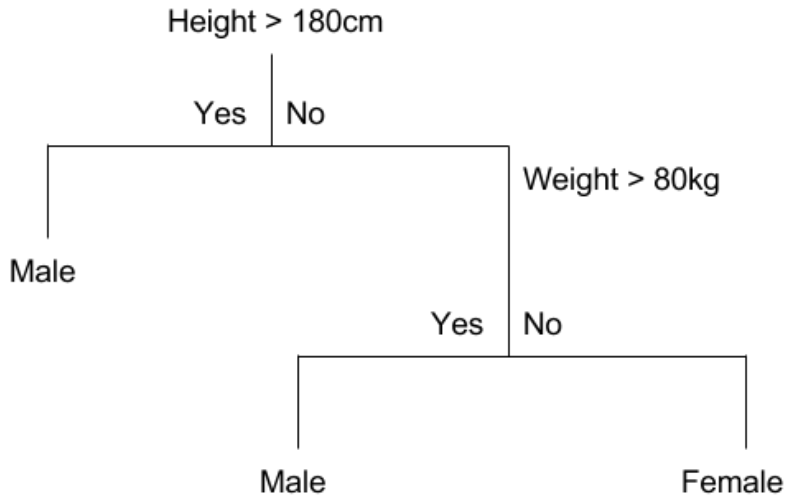


Fig. A.1: Basic example of a decision tree classifying between the classes *Male* or *Female*. *Height>180cm* is seen as the root node and *Weight>80kg* is an internal node. The nodes containing *Male* and *Female* are known as terminal nodes (Brownlee, 2016)

INPUT: S, **where** S = *set of classified instances*
OUTPUT: *Decision Tree*
Require: $S \neq \varnothing$, $num\_attributes > 0$

```
1:  procedure BuildTree
2:      repeat
3:          maxGain ← 0
4:          splitA ← null
5:          e ← Entropy(Attributes)
6:          for all Attributes a in S do
7:              gain ← InformationGain(a, e)
8:              if gain > maxGain then
9:                  maxGain ← gain
10:                 splitA ← a
11:             end if
12:         end for
13:         Partition(S, splitA)
14:     until all partitions processed
15: end procedure
```

Fig. A.2: Basic structure of a decision tree algorithm (Mayo, 2016)

Appendix B

Results

This section includes explanations, tables and graphs that may supplement that of what is discussed in the *Results* in chapter 4.

B.1 Feature Selection process

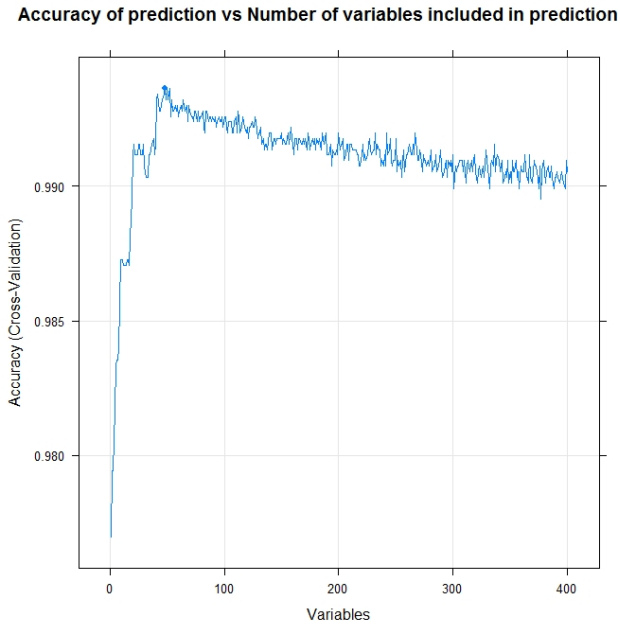


Fig. B.1: Graph of the Accuracy of prediction against the number of variables included in the model
This figure shows the highest prediction accuracy of the model at the inclusion of different number of variables. The highest prediction value over all was seen at 48 variables.

Base question number	Elaboration of question [1]
q3.4.1.1	Question: *What was your employment status between graduating and starting the job you had on the 1st of September 2012* [2]*?* Answer: • *5 - Employed in the informal sector (you are an unregistered, informal trader, maker or seller of goods and services* ***(1)*** • *2 - Employed part/full time in the private sector (e.g, in a registered tax paying business, company or institution* ***(10)*** • *4 - Employed part/full time in public sector (e.g. in a government department, university, science council, public school or public health centre* ***(11)*** • *1 - N/A → I was studying fulltime; not working and not looking for work at all* ***(15)*** • *6 - Unemployed and looking for work* ***(23)*** • *8 - N/A The job I had on the 1st of September, I started soon after studying* ***(31)*** • *3 - Self-employed in the private sector (you are registered for tax purposes)* ***(35)***
q3.4.11	Question: *Did your 2010 qualification lead to any of the following*? Answer [3]: • *f - none of the above* → (0 - ***(2)*** and 1 - ***(3)***) • *d - Increased tasks and responsibilities* → (0 - ***(4)*** and 1 - ***(5)***) • *a - A promotion to a higher rank, position or level* → (0 - ***(7)*** and 1 - ***(8)***) • *b - A pay increase* → (0 - ***(12)*** and 1 - ***(13)***) • *c - Increased benefits* → (0 - ***(26)*** and 1 - ***(27)***)
q3.4.13	Question: *How did you look for work?*: • *f - I responded to job ads on employment websites* → (0 - ***(6)*** and 1 - ***(9)***)

q3.4.10	Question: *On a scale of 1 to 5 with "1" being "not at all" and "5" being "to a large extent", to what extent...*: • *b - were you able to apply what you learned in your 2010 qualification in the job you had on the 1st of September 2012?* – Answer: *5* ***(14)*** • *a - was the job that you did on the 1st of September 2012 related to the field in which you did your 2010 qualification* – Answer: *5* ***(17)*** • *c - were you satisfied with your 2010 qualification in relation to the job you had on the 1st of September 2012?* – Answer: *5* ***(18)*** – Answer: *3* ***(48)***
q4.1.6	Question: *On a scale of 1 to 5 with "1" being "not at all" and "5" being "to a large extent", to what extent did your 2010 qualification prepare you for further studies?* Answer: • *5* ***(16)*** • *4* ***(25)***
q4.1	Question: *Were you registered for and studying towards another qualification at a university on the 1st of September 2012*: Answer: • *2 - No* ***(19)*** • *1 - Yes* ***(20)***
q3.3.2	Question: *On the 1st of September 2012 did you still have the same job that you had just before you started studying towards the qualification you obtained in 2010?* Answer: • *2 - No* ***(21)*** • *1 -Yes* ***(22)***
q3.4.6	Question: *What was your primary method of finding the job you had on the 1st of September 2012?* Answer: • *2 - Through help of a lecturer* ***(24)***

q3.3	Question: *What was your employment status before you started studying towards the qualification you obtained in 2010?* Answer: • *r1_2 - Employed in public/private sector or self employed* ***(28)*** • *5 - Employed (part- or full-time in the public sector (e.g., in a government department, university, science council, public school or public health centre)* ***(41)*** • *r1_1 - Unemployed and not looking for work* ***(42)***
q2.1	Question: *While studying towards the qualification in 2010 were you mostly full-time or part-time?* Answer: • *1 - Full-time* ***(29)*** • *2 - Part-time* ***(30)***
q4.1.5	Question: *Why did you chose to study further after graduating in 2010*[4]*?* Answer: • *b - To improve my chances of getting a job as I have yet to find one* $\rightarrow$ (0 - ***(32)*** and 1 - ***(37)***)
q3.3.3	Question: *On what basis were you employed in the job you had just before you started studying towards the qualification you obtained in 2010?* Answer: • *2 - Temporary/contractual* ***(33)*** • *1 - Permanent* ***(39)***
q4.1.4.4	Question: *Did you complete your current qualification by the 1st of September 2012?* Answer: • *2 - No (I deregistered or discontinued this qualification)* ***(34)*** • *1 - Yes* ***(36)***
q2.4.1	Question: *What type of qualification did you obtain in 2010?* Answer: • *1 - Masters degree by coursework and research* ***(38)***

age_r1	Question: *Age range?* Answer: • *1 - 21 or younger* ***(40)*** • *3 - 31 or older* ***(44)***
q3.4.3	Question: *What was your occupation in the job you had on the 1st of September 2012?* Answer: • *6 - Clerk* ***(43)***
q4.1.3	Question: *In which field were you studying on the 1st of September 2012?* Answer: • *2 - Business & Commerce* ***(45)***
q3.4.4	Question: *What basis were you employed in the job you had on the 1st of September 2012?* Answer: • *1 - Permanent* ***(46)***
q3.3.4	Question: *Were you full-time or part-time in the job you had just before you started studying towards the 2010 qualification?* Answer: • *1 - Full-time (40 hours per week)* ***(47)***

Tab. B.1: Table of the most important variables, in order of importance, as per the backward selection process. A bagged decision tree is used as the base model for the process. The rank of the described variable is in bold brackets. The footnotes from this table are seen after table B.2

B.2 Variable Property plots

This section has the graphs of the top 48 variables[5]. Only the original variable is plotted and not the dummy variable form. Hence there aren't 48 graphs seen but 23. The graphs have been grouped according to the original questions asked which can be seen in table B.1. It is easier to see contrasts/similarity when the property plots are side by side. Generally it is difficult to see similarities and contrasts within the plots presented here.

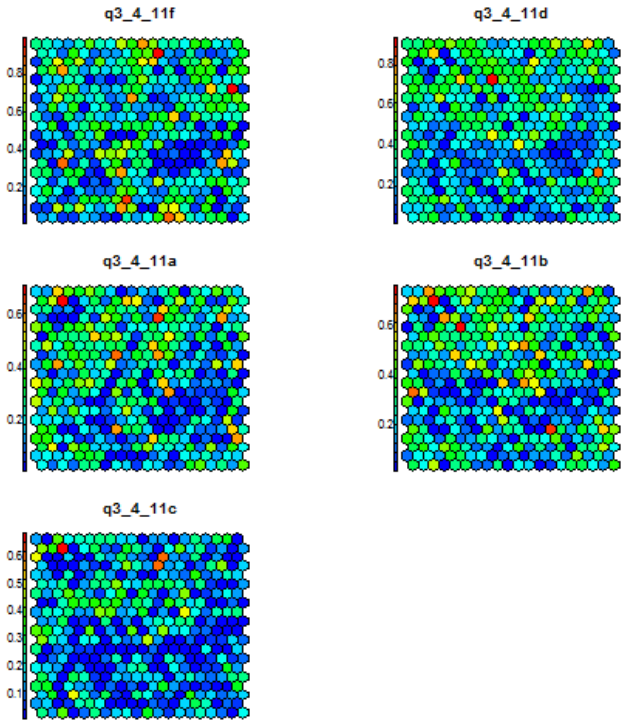


Fig. B.2: Property plot comparison of question 3.4.11 and options a,b,c,d and f of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar is the behaviour/reactions of the data points in that node to its neighbour. The main question involved indicating what lead the candidate to pursue their 2010 qualification. The plot for option c, where the candidate received *increased benefits*, seems to have more similarity between nodes. Generally, the plots do not show any relation between each other. Yet, the plots all seem to show more similarity between the nodes.

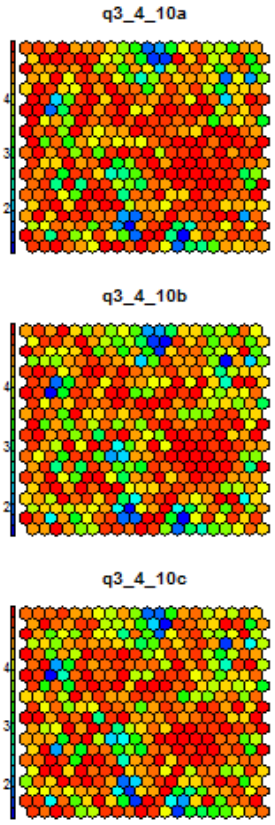


Fig. B.3: Property plot comparison of question 3.4.10 and options a,b, and c of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar is the behaviour/reactions of the data points in that node to its neighbour. The dissimilarity within the plots makes sense as these plots are option specific for a particular section of questions. These questions involve the candidates qualification in relation to their employment where the responses range in value from 1 to 5. This complication can explain the dissimilarity as more options were given to respondents. Interestingly, if one looks closer at the plots the behaviour appears similar.

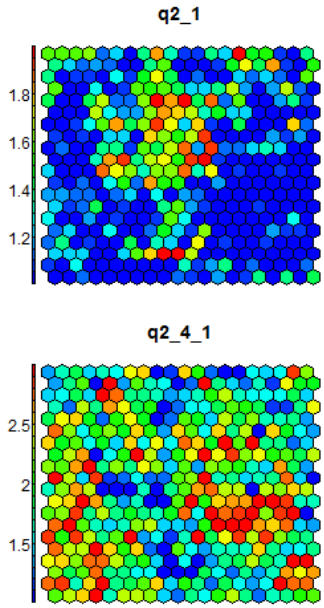


Fig. B.4: Property plot comparison of question 2.1 and question 2.4.1 of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar is the behaviour/reactions of the data points in that node to its neighbour. Plot of question 2.1 seem to be completely similar whereas question 2.4.1 has no clear pattern. The question/variable is seen in table B.1 and given the nature of these questions it is understandable about why there is no relation seen. The one question deals with whether the degree was done part-time or full-time, whilst the other question deals with indicating the type of degree that the candidate completed in 2010.

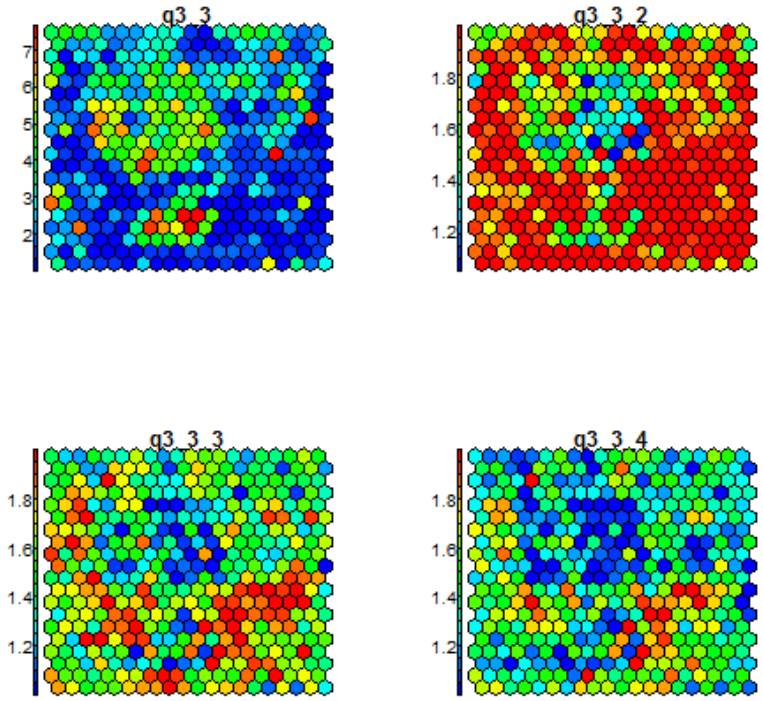


Fig. B.5: Property plot comparison of question 3.3 , 3.3.2, 3.3.3 and question 3.3.4 of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar is the behaviour/reactions of the data points in that node to its neighbour. Question 3.3 and 3.3.2 show contrasting behaviours. Question 3.3.2 is only relevant to *Employed* candidates. Thus, there are bound to be empty nodes that represent the *Unemployed* population that would have not answered this question.

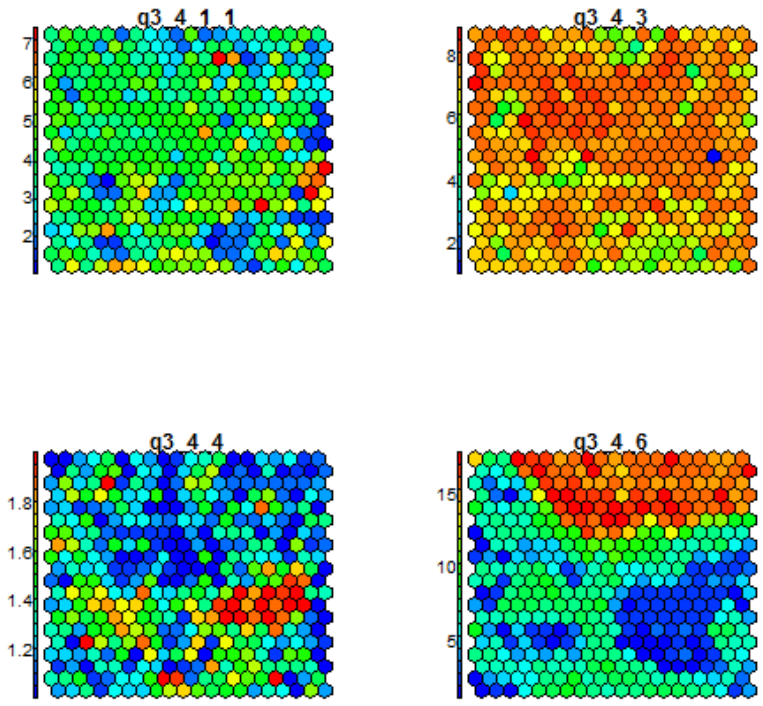


Fig. B.6: Property plot comparison of question 3.4.1.1, 3.4.3, 3.4.4 and question 3.4.6 of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar is the behaviour of the data points in that node to its neighbour. None of the figures here show similar or contrasting behaviours. Yet, the plot of question 3.4.6 displays an interesting pattern which suggests at least one cluster of nodes behaving similarly. Whilst question 3.4.3 seems to have most of the nodes displaying dissimilarity with their neighbour, the implication of this is that the spread of the kinds of job that the *employed* candidates had were generally not the same. This is understandable because the data was collected from candidates of different disciplines.

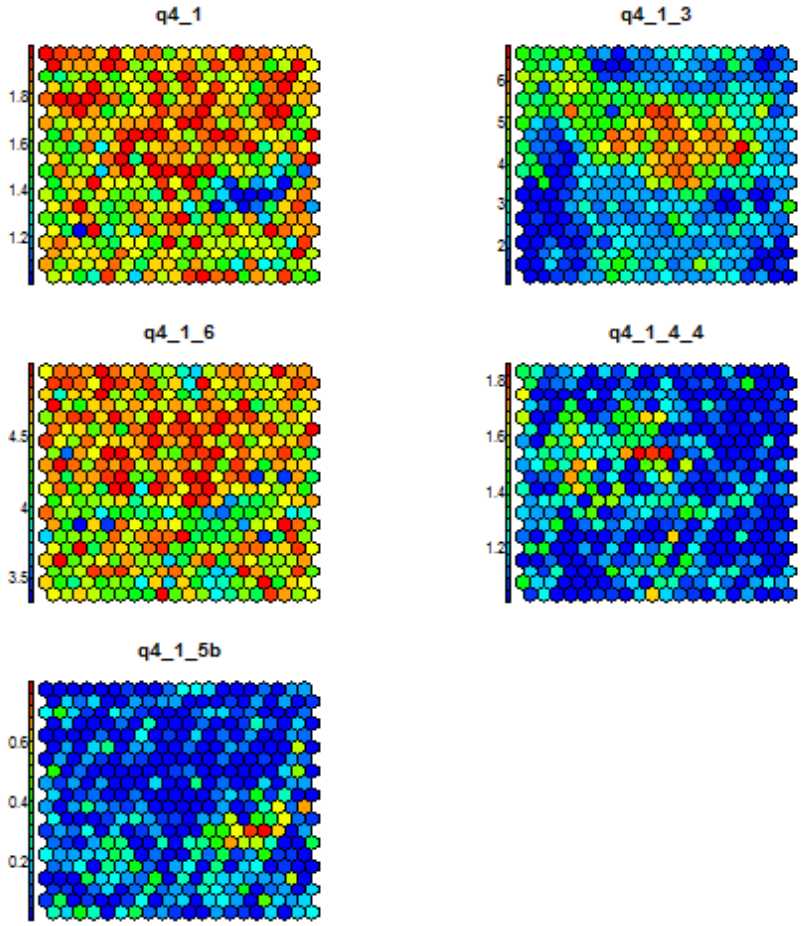


Fig. B.7: Property plot comparison of question 4.1 , 4.1.3, 4.1.6, 4.1.4.4 and question 4.1.5b of the *South Africa - Graduation Destination Survey 2012* (2015) data where the darker the blue (red) the more (dis)similar the behaviour of the node is to their neighbour nodes. None of the figures here show similar or contrasting behaviours.

B.3 Results of Models

This section provides supplementary information for the models included in the first stage of testing. Note that not all models are being discussed here. It starts of with a table of the parameters used for certain models and then further information is included from the different models.

Machine learning algorithm	Parameters
k-Nearest Neighbours	$k = 46$
Radial SVM	$\Gamma = 0.005$ $cost = 10$
Polynomial SVM	$\Gamma = 0.005$ $cost = 10$ $degree = 4$
Multi-layer NN	Hidden layers: 1 Number of neurons: 2
Bagged decision trees	Number of base classifiers (decision trees) : 500
Adaboost	Tree depth: 1 Number of iterations: 20
Rotation Forest	Number of random variables selected per subset (K): 40 Number of base classifiers (L): 10

Tab. B.2: Table of parameters used in machine learning algorithms for both stages of testing[6]

[1] with most important answers in order of importance

[2] This is the date of reference for the survey

[3] Note that all these answers refer to whether option was selected or not, for example if the candidate selected *f* or did not select *f* was seen as vital

[4] Note that the answer refers to whether option was selected or not, for example if the candidate selected *b* or did not select *b* was seen as vital

[5] Note that the variables were broken up into dummy variables as per the data preprocessing process and thus repetitions of a question were seen

[6] The *cost* mentioned in the SVM models is calculated according to the cost stated in the R's *e1071* package (and is not the same as the cost calculated for example in the explanation by James *et al.* (2013))

k-Nearest Neighbours

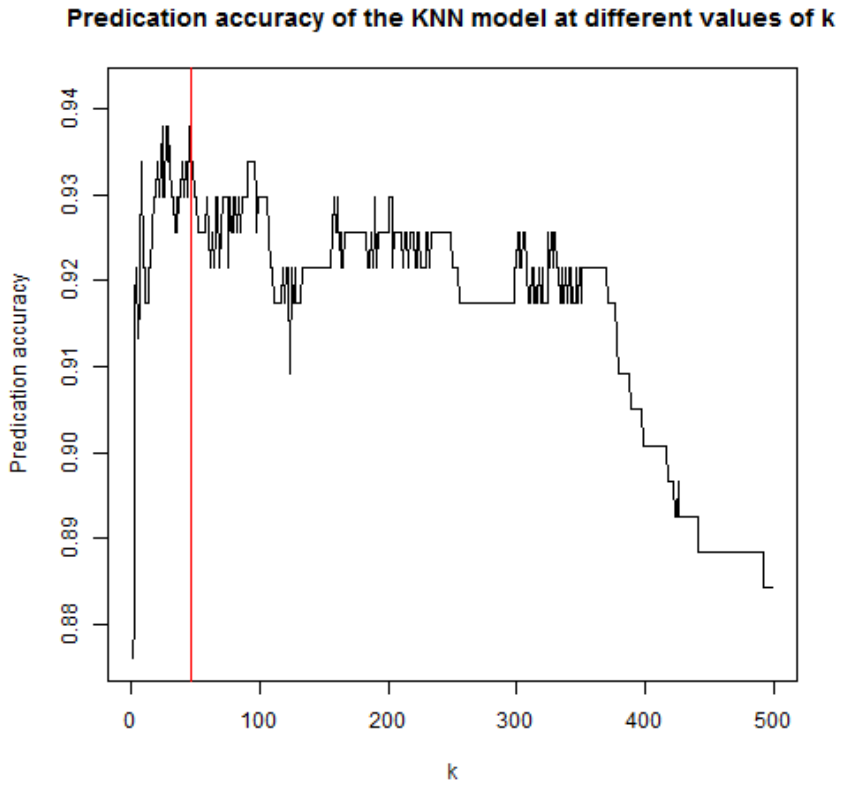


Fig. B.8: Prediction accuracy of the *k*-Nearest Neighbours model at different values of nearest neighbours
This figure shows the level of prediction accuracy at different numbers of nearest neighbours included in the model. The model with the best overall prediction is the one with 46 nearest neighbours included.

Adaptive boosting

Variable name	Information gain (%)
q3.4.1.1 option 5	14.41536389
q3.4.11. option f (0)[7]	5.159997613
q3.4.11. option b (0)	4.194041303
q3.4.3 option 6	4.16542532
q3.4.11 option a (0)	4.135867373
q3.4.11 option d (0)	3.890052623
gpa	3.111631965
q3.4.13 option f (0)	3.111631965
q3.4.10 option b (5)[8]	2.624942276
q4.1.6 (5)[9]	2.448216838

Tab. B.3: Variable Importance Table for Adaptive boosting model[10]

[7] 0 indicates that the absence of this option is seen as important

[8] A rating of 5 is given here where 5 meant as *to a large extent*. Refer to table B.1 for specifics of the question/variable

[9] A rating of 5 is given here where 5 meant *to a large extent*. Refer to table B.1 for specifics of the question/variable

[10] The *cost* mentioned in the SVM models is calculated according to the cost stated in the R's *e1071* package (and is not the same as the cost calculated for example in the explanation by James *et al.* (2013))

Bagged decision trees

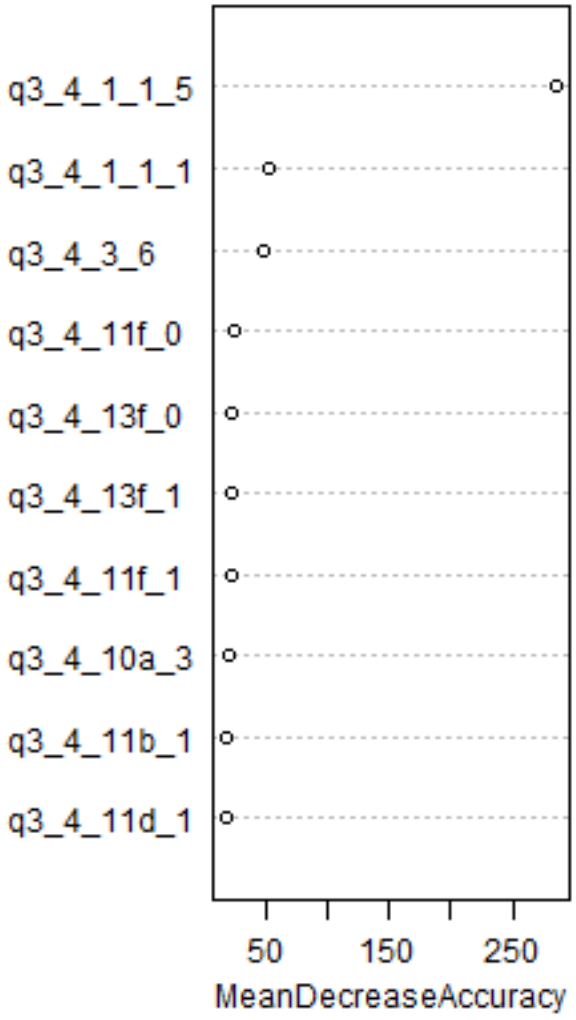


Fig. B.9: Variable Importance plot of the top 10 variables used for the Bagged decision tree model in descending order of importance
The figure gives an sign of how much on average does the accuracy decrease when the specific variable is removed from the model. For bagged decision trees question 3.4.1.1 and option 5 is important which is the same as the result of the feature selection process. The question involved indicating what the candidate's employment status was before pursing their 2010 degree where option 5 is *Employed in the informal sector*. Looking at the Mean Decrease Accuracy for this variable it was an extremely vital variable as it is almost 5 times more important than the 2nd most important variable.

B.4 Paired sample t-test explanation

The t-test performed here is known as a paired sample t-test used for determining whether the average difference between two sets of observations is zero. We use this type of test when there is a dependence between the two sets of observations. We base our performance measures on the same models that were trained before but using different versions of a data set. This t-test can help us determine whether there is a statistical significance in the difference between the performance measures.

The test has two hypotheses. The null hypotheses which always states that the average difference is 0. as well as the alternate hypothesis is dependent on what the expected outcome is. We are interested in finding a difference, no matter if it positive or a negative difference. The hypothesis can be defined as follows;

$$H_0 : \mu_d = 0$$
$$H_A : \mu_d \neq 0$$

where $\mu_d =$ average difference between the performance values of the top 3 models[11].

There are various assumptions that need to be satisfied to use this t-test[12]. The one assumption that could be violated is the assumption that the differences are normally distributed. This is hard to confirm, given that there are only 3 observations per t-test. Yet, we will still assume normality.

Once we have decided on the alternative hypothesis, we can proceed with calculating a test statistic using the sample average and standard deviation of the differences. A *p*-value from that test statistic is then calculated.

The *p*-value (also known the probability of observing the test statistic under the H_0) assists in determining statistical significance. For a two-tailed test the *p*-value is calculated as;

$$p - value = 2 \times p(T >\mid t \mid) = 2 \times (1 - p(T <\mid t \mid))$$

where $p(T <\mid t \mid)$ is the probability of observing a value less than the test statistic value t[13] of a *t*-distribution with 2[14] degrees of freedom.
The *p*-value can then be compared to a significance level (usually 5% or 0.05). For example, a *p*-value higher(lower) than 5% indicates that there is a more(less) than 5% chance that average difference satisfies the null hypothesis i.e. there (are) no significant differences between the performance measures.

[11] as per the stage one testing
[12] These assumptions can be read on **?**
[13] Note that *t* could be a negative value that is why we using the absolute value of the test
[14] n-1 degrees of freedom where n is the number of observations, i.e. 3