**Coursework**

**ST3189 Machine learning**

**Tristan Leanca**

**Candidate No. A13375**

Using unsupervised learning techniques such as Principal Component Analysis or Cluster Analysis to summarise the information in the data

The job of a data scientist is to first and foremost, understand, decipher, describe, visualise and perhaps communicate vast amounts of data with the aim of gathering as much useful information from them as possible in order to understand the patterns that bring about such information and henceforth communicate their findings in a way that positively affects business decisions. In summary, a data scientist’s role generally speaking involves making sense of disorganized, unstructured data, from various sources such as online databases or datasets, or perhaps in the past decade, social media feeds. It must be stressed that machine learning is a tool in a data scientist’s repository, and in this project, I will focus mostly machine learning, which can simply be described as a group of techniques used by data scientists to learn, and predict data.

Machine learning algorithms can be separated into two categories: supervised and unsupervised learning. In the case of supervised learning, there are dependent or response variables

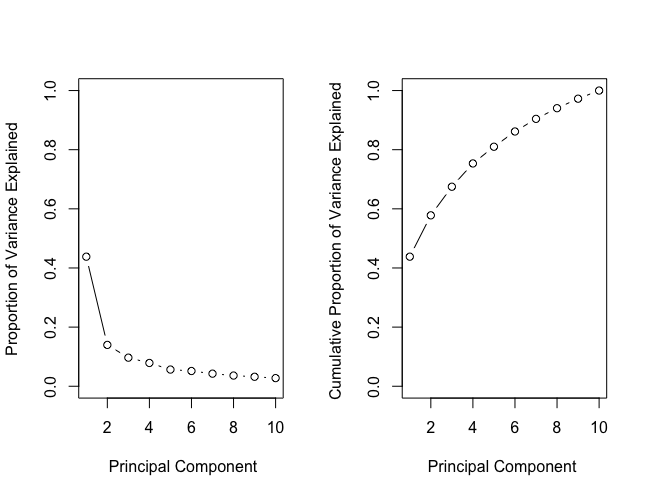
that we are interested in and our goal is to understand the relationship between these dependent variables and the other variables called predictor variables, i.e., variables with which we estimate and predict the dependent variable. Furthermore, supervised learning can be divided into two additional subcategories: regression and classification. In the case of regression, our dependent variable, Y is quantitative. Examples include prices of stock market indices, cholesterol levels in adults or even perhaps the age of a given population. In classification however, Y takes predetermined values in an unordered set such as: digit 0-3, survived/died, divorced/married, etc. Overall, the purpose of supervised learning is to accurately predict new data, understand which of the inputs or dependent variables affect the outcome (dependent variable) and finally to assess the quality of our predictions and inferences.

The other category of machine learning, unsupervised learning is perhaps a bit trickier in that you do not have a dependent variable to estimate. Analyses implemented under unsupervised learning can give us an idea of what the underlying pattern behind the data may be, but it cannot enable us to make accurate predictions and it therefore difficult to know how well you are performing, us you are unable to measure the quality of predictions. However, the data gathered can be of great use. It can be used to segment markets for consumers and hence create different strategies to accommodate the different characteristics as of those segmentations. Lastly, unsupervised learning can be used as a pre-processing step for supervised learning, for instance, if you have a data set with a large number of predictor variables, you employ a technique such as Principal Component Analysis in order to reduce the number of variables and hence much more easily interpret and apply machine learning models to the data.

In the 2016 European Working Conditions Survey, we are given data relating to various features and variables which were collected from individuals such as gender, age and ratings of the general wellbeing of the individual, such as happiness, and relaxation. The dataset has ten columns or variables and 7183 rows or observations. The objective of such a survey may be to understand how much the qualitative aspects of a person’s life may have an impact on the

more quantitative professional aspects of his/her life such as performance.

We will use one unsupervised method techniques to try to understand and describe the data, one of which is Principal Component Analysis (PCA). PCA is a tool that is helpful for visualizing data before it is processed as described above and thereafter apply supervised method methods. PCA represents a low-dimensional representation of the data and thus perhaps be able to visualise and describe the data. According to James et al. “It finds a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated”. And each linear combination of the features is represented in descending order of variability in new variables called ‘principal components’. Hence, the first principal component will have the largest variance, the second principal component the second largest variance, etc. A simple example to understand PCA is to imagine a group of people you have to photograph, and it requires you to take the photo in the angle that shows most clearly all the people in the group. This is akin to finding the principal component that maximises variability.

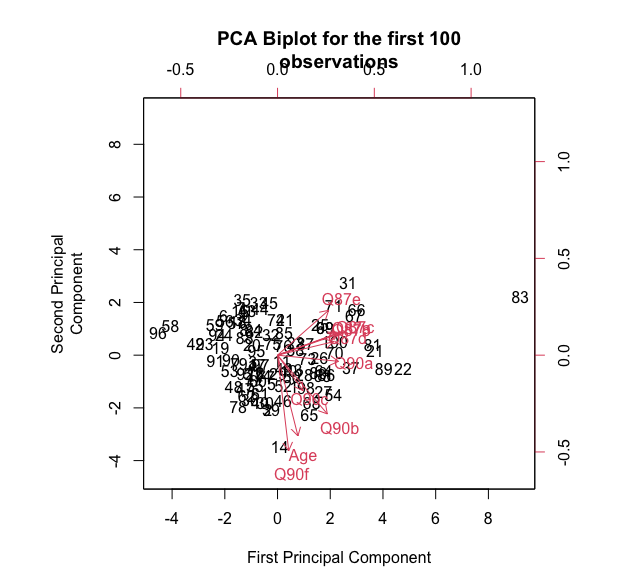
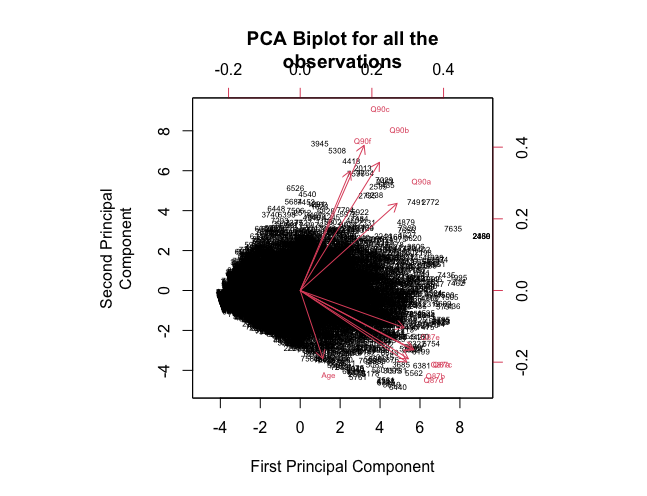


The scree plot above on the left plots the proportion of total proportion of variance explained.

The right plot above shows the cumulative proportion of total

variance explained by the principal components.

Dimension reduction simply means taking the ten variables (we discount gender as it as an unordered categorical variable) and by using PCA, acquire new variables that explain best the dataset. The question to ask next, is how many of these new variables, or principal components do we need to best explain the data. Unfortunately, there is no clear-cut answer. We construct a diagram called a scree plot which shows the proportion of variance explained by the principal components, and the rule a thumb is, you choose the number of components where there is a kink, or an elbow in the plot. Hence, according to the above plot, we choose the first two components, which account for about 60% of the variability in the data.



These two diagrams plot the first two principal components for the EWCS data. The numbers

correspond to the score for the first two principal components. The red arrows indicate the first two eigenvectors or loading vectors. The left diagram represents a biplot for all the observations whereas the right one for the first 100 observations.

It is important to note that the variables in the EWCS dataset (except for Age) are categorical variables, also known as factors, which as described above is a variable that can take a limited number of values. Unsupervised learning methods generally speaking should not be applied to factors, however, in this case, the variables take on an ordered set, therefore it is appropriate to apply PCA or clustering methods.

After we apply PCA, we construct two diagrams called biplots of the whole dataset, and because there are a large number of observations, of the first hundred observations. We can see that the observations are distributed quite uniformly among the two plots and the data is quite dispersed, however there are still some observations to be made. On the second diagram, which is a plot of the first hundred observations, Q90a(“At work I feel full of energy”) is shown to be entirely dominated by the first principal component, whereas in the full biplot, Q90a is divided amongst the principal components, suggesting a different contribution to the data. Likewise, the loading vectors Age and Q90f(“In my opinion, I am good at my job”) are near each other suggesting a correlation between an individual’s age and their self-described job proficiency. The plot representing the whole dataset though has the two eigenvectors further apart suggesting a disparity between the two plots, and we cannot draw conclusions on the whole dataset based on the first observations only.

Each observation has a score on the principal components, and two understand and attempt to interpret these scores, we look at the loadings. We see from our plot that the first principal component is mostly dominated by the Q87a-Q87e variables, which describe the general wellbeing of a worker, whereas the second principal component is mostly dominated by the Q90a-Q90f variables, which describes the worker’s job satisfaction. This means is that the Q87 variables are the features that contribute to the most variability in the data, hence they are more important to consider when you analyse this dataset, for example if you’d apply some regression method.

Overall, the biplot with which we interpret the data and attempt to describe it is quite difficult to interpret as there are no particular clusters, or separations in the data. Rather, it is shown to be as a concentrated black cloud in the middle of the plot with conclusions being able to drawn on feature selection rather than describing perhaps certain patterns among the data.

Regression model for predicting final grade of a student in a Portuguese schools

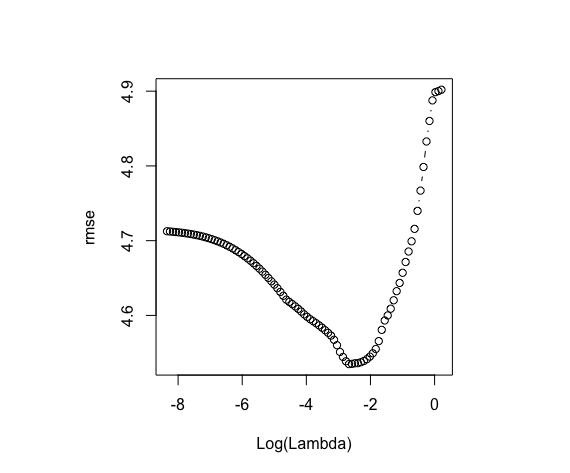
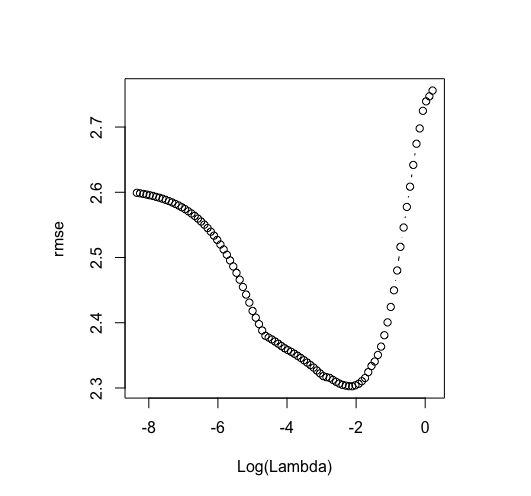
Our next task will involve building a regression model to for the variable G3 (final grade) in two different subjects (mathematics and Portuguese) given by two datasets respectively. The attributes or variables included are student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. The data also includes previous grade results, however for our analysis, it is imperative we predict the final grade excluding the previous grades (G1 and G2) as this will allow us to predict the grade with higher accuracy and we will be able to identify other factors that influence school performance, which gives us an invaluable insight into the social or demographic factors that affect grade results. Individually, our datasets have both 33 variables, 649 and 395 observations (students) and after we merge the data to obtain students that have both taken Portuguese and mathematics, we end up with 53 variables and 382 rows. As a result of the merge of the datasets, we obtain some duplicate variables, so we will compare two numbers between the machine learning models, as we have two separate final grades.

Before I begin the analysis, it is pivotal I explain a statistical concept called the ‘bias-variance tradeoff’ as we will discuss it through the next two parts. First of all, bias is the inability of a machine learning method to capture the true relationship between predictor and response variable. Variance is the difference in fits between datasets. If a model fits a set of observations very well and then fits next set of observation poorly, this is called over-fitting in machine learning parlance. Ideally, you’d want a model that has low bias and low variance, however, this is where the term ‘bias-variance tradeoff’ comes in, as it is quite difficult to minimize both. Changing one almost invariably comes at the expense of the other and it plagues the whole of machine learning.

The first machine learning model that we will use for this prediction is multiple linear regression. Linear regression is a simple but powerful tool for assessing the relationship between predictor and response variable. The issue that arises with applying linear regression with many variables, is that due to the many predictors (explanatory variables) it may be difficult to interpret the model due to possible collinearity between the predictors and possibility of outliers, and because we have a few categorical predictor variables, linear regression model will automatically create so-called dummy variables depending on the number of categories in each variable, so the number of features may increase from 33 to more than 70, and it may hence be difficult and tough to accurately assess the model and its predictors. However, I still believe it can be a powerful tool to predict the accuracy of the model, and we will later compare its predictive performance when we compare it to other regression models. Applying the model to predict both final grades, we deduce that the most statistically significant variables in predicting final grades are ‘failures’ (past class failures), ‘Dalc’ (consumption of alcohol during workday) and ‘goout’ (going out with friends), all which seem to be quite intuitive. The mean squared errors are quite different, remarkably enough. I will summarise them in a table.

For most of our machine learning models we will test the prediction accuracy of the model by dividing our data into training and testing data set. We use our training set to, unsurprisingly train our model on and our test data set to test or validate our model on and hence evaluate the model performance. A common metric used to compare regression model performances is the mean squared error (MSE), and what it does is measures the average of the squares of the differences between the observed and predicted value of the data points.

This slightly different method we will be applying is called regularization. It is a slightly altered method of linear regression in which the variables are reduced or eliminated, and the resulting model is much simpler to interpret. In mathematical parlance, it does so by minimizing the residual sum of squares and in addition a shrinkage penalty. These methods are called ridge and lasso regression, or a combination of both, elastic-net regression. I will compare all three of them and use cross-validation (a resampling method) to determine the optimal value of the tuning parameter. After we compared the three methods, the optimal alpha parameter with lowest MSE is 0.8, i.e. elastic-net regression is best method to use. Also, lambda parameter that minimizes MSE is 0.2. For the math prediction, ridge regression is optimal method.



*A plot showing the minimum of lambda for the Portuguese A plot showing the minimum of lambda for math final*

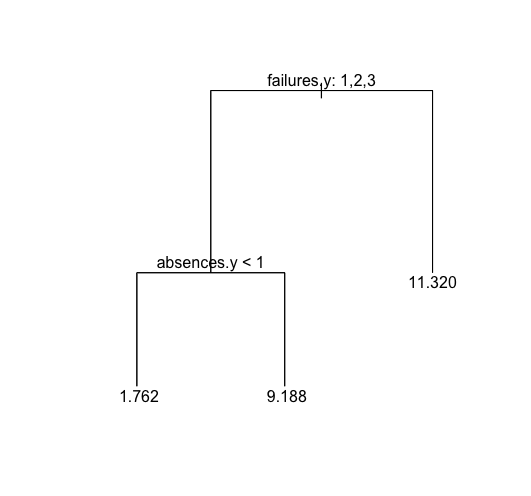
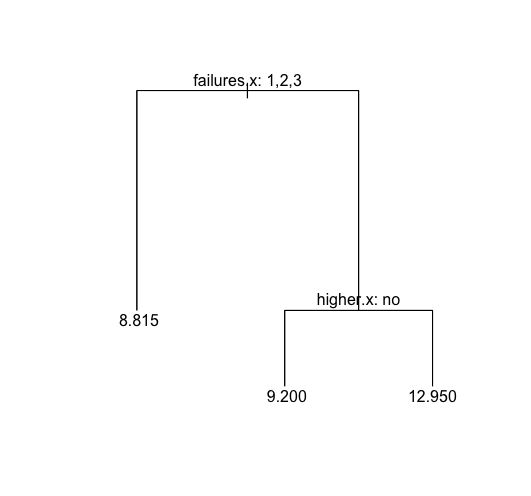
*final grade prediction with a lambda value of 0.2 for alpha = grade prediction with a minimum lambda value of 0.07*

*0.8. for alpha = 0.8*

|  |  |  |
| --- | --- | --- |
| Model | MSE Portuguese | MSE  Math |
| Linear regression | 6.860816 | 27.50616 |
| Regularisation | 6.084707 | 17.18129 |
| Random forests | 5.107372 | 13.18615 |
| Whole-tree | 6.777447 | 18.16165 |
| Boosting | 5.361108 | 15.42135 |
| Pruned-tree | 6.131727 | 16.88778 |

The last methods we will utilize are called tree-based methods which involve segmenting the predictor space into simple regions. The set of decisions used to stratify the predictor space are used to build trees, and since our dataset implicates a regression problem, these are called regression trees.

Their biggest advantage is how simple is to interpret them, however they may suffer in prediction accuracy. So, we will use two more methods that are based on aggregating many trees to improve the accuracy. This may come at the cost of some loss of interpretation. We will first build a tree, then prune it, in order to avoid overfitting.



*A pruned tree using cost-complexity pruning for A pruned tree using cost-complexity pruning for predicting final grade in*

*predicting the final grade in Portuguese. We interpret mathematics. Like in the left diagram, ‘failures’ the most important*

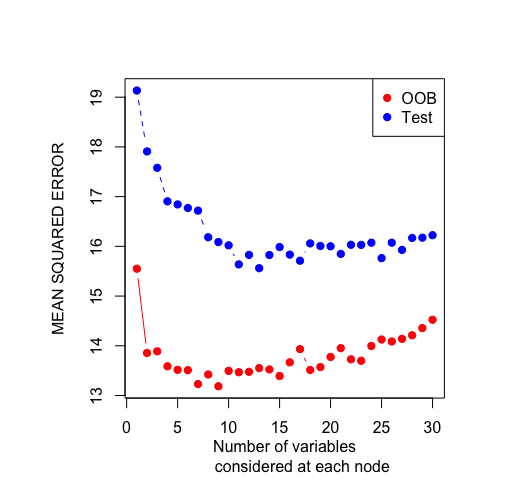
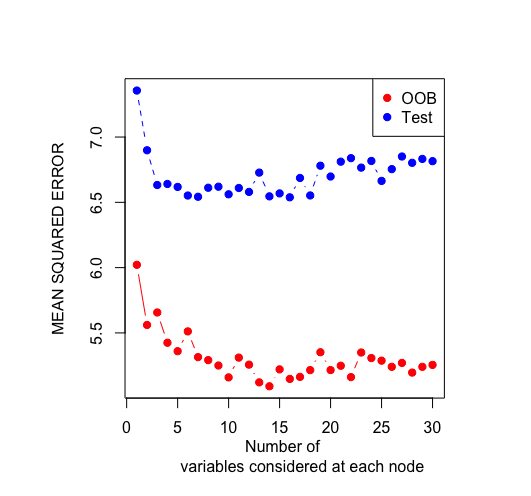
*this tree in the following: ‘failures’ is the most important feature in predicting final grade. However, unlike the left diagram,*

*feature in determining final grade and then higher (students absences are second most import in prediction.*

*wants to take further education)*

Finally, in order to (potentially as it is not guaranteed) improve our accuracy, we employ random forests and boosting. Both methods involve building a very large number of trees and averaging their results. The former involves building a number of decision trees on bootstrapped (resampled) training samples, and when the tree is built, at each node only a random number of predictors is considered. The latter works in a comparable way except that trees are grown sequentially to improve on the performance on the previous collection of trees.

For random forests, we will vary the features considered at each node, and there is no need to divide into train and test sets as the model has an in-built feature called OOB’s (Out of bag errors). OOB errors simply put are methods to estimate prediction error using bootstrapping ( again, a method of resampling).



*A plot that shows the error for test and OOB A plot that shows error for test and OOB considering the number*

*of variables at each node at tree building for Portuguese of variables at each node for predicting mathematics*

*final grade final grade*

For boosting, we will vary a few parameters to obtain the most accurate model. The number of trees, lambda and interaction depth are varied using cross validation. For the Portuguese final grade prediction, the most important variables, as in the previous models are ‘Dalc’ (daily alcohol consumption) and ‘failures’. For the mathematics prediction, the most important variables are ‘absences’ and ‘mjob’ (mother’s job).

Overall, comparing the mean squared errors of all the machine learning models we employed, random forests give the best overall predictions, both for the Portuguese and math final grade G, and intriguingly, features such as the number of school absences, the number of past class failures, extra educational support, sex of the student, reason for choosing the school, going out with friends, mother’s job and health as significant factors in the student’s final performance with past failures and absences accounting for most of the impact in our model. We also can see that linear regression for predicting the Portuguese final grade is not far off the accuracy of the other models, however, for Mathematics, it is a bit off. It shows linear regression can still be a handy and powerful tool in machine learning.

Building a predictive classification model to assess client subscription

for term deposits

Our next dataset to analyse is related with a marketing campaign presented by a Portuguese banking institution. The marketing campaign was based on phone calls. The aim of the campaign was to determine if the client will subscribe or not to a term deposit. The machine learning models we will use are part of the mentioned above supervised learning, that we have a dependent variable which we must predict, infer and perhaps analyse. As explained, the Y (dependent) – variable in this data set is categorical, i.e. it has to possible outcomes: ‘Yes’ or ‘No’. Therefore, we must use classification techniques to estimate the two(binary) outcomes.

We are given a set of 4516 observations with 17 variables. We will however discard the variable

“pdays” as it contains a very large number of empty slots, in mathematical jargon, null values, thus it will not help us accurately predict our subscription decision. Another variable we will discard is “duration”, which records the length of time on the telephone. However, when we have new data (observations) to predict, we evidently do not have this information. Moreover, if the variable takes on the value 0, the obviously the response variable will take on the value ‘no’, as the campaign is based on a phone call. So, we remain with 15 variables and 4516 observations.

The machine learning techniques we will utilise in our analysis and build the classification models are logistic regression, a model similar to linear regression, however unlike linear regression, with logistic regression we estimate probabilities, and based on the threshold (usually 0.5) we choose, we categorise the predicted observation into their respective categories, which in our case are ‘Yes’ or ‘No’. Other models we use are tree-based methods, which include building and pruning trees, random forests and support vector machines, which is classification method of separating the observation using a hyperplane, which, again, in mathematical jargon is a line in 1-dimension to a plane in 3-dimensions and a hyperplane in n-dimensions. Lastly, we will use linear discriminant analysis.

We will once again compare our result from our various machine learning methods by dividing our dataset into training and testing data, train our models on the training data and then test, or validate our trained models on the testing dataset. The validation metric we will use will be the accuracy of the test, i.e., the percentage of the observations correctly predicted by our model. Another metric for comparison to be used are ROC curves and AUC. I mentioned above that classification methods classify or separate observations based on a threshold that you choose. What the receiver operating characteristics (ROC) does is compile all these thresholds into a single plot due to the fact that different threshold may yield different results, and in some cases, depending on the needs of the outcome of the model, certain threshold may be used. This is where I introduce the terms ‘specificity’ and ‘sensitivity’, which are statistical terms for ‘True Negative Rates’ and ‘True Positive Rates’ respectively. Depending on the type of dataset analysed, for instance let’s say need to classify patients if they have a virus or not. In this case, it is absolutely imperative to classify each patient correctly in order to minimise risk of an outbreak, i.e., we need as little false negatives as possible, so we lower the threshold even if it means more false positives, and hence get higher sensitivity. ROC precisely does that; it plots True Positive Rate against False Negative Rate. For our analysis, we do not have a specific threshold in mind, hence we will use the AUC (area under the curve of the ROC) to compare the models and all their thresholds.

Before I introduce the models, it is important to note that the class label we are trying to predict is imbalanced. There are approximately 90% of ‘no’s’ and 10% of ‘yes’s’. Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This may hence impact the results in the models and lessen the predictive performance. There are techniques for reducing class imbalances, however such methods are out of the scope of this project.

As mentioned above, logistic regression is a method for modelling the probability of a categorical or in our case binary outcome. Applying the model, the most significant variable in prediction is ‘poutcome’ (outcome of previous marketing campaign) which is as expected. Clients that have subscribed previously are much more likely to subscribe again due to the campaign. Then, strangely enough the month of the campaign, maybe explained by some specific event in that month that would stimulate people to subscribe more. This result is based on fitting the model on the whole data, and also on fitting it on training data. It is interesting to observe that we do not overfit, and we will see throughout this data that differences in prediction accuracies are minimal between full data fitting and

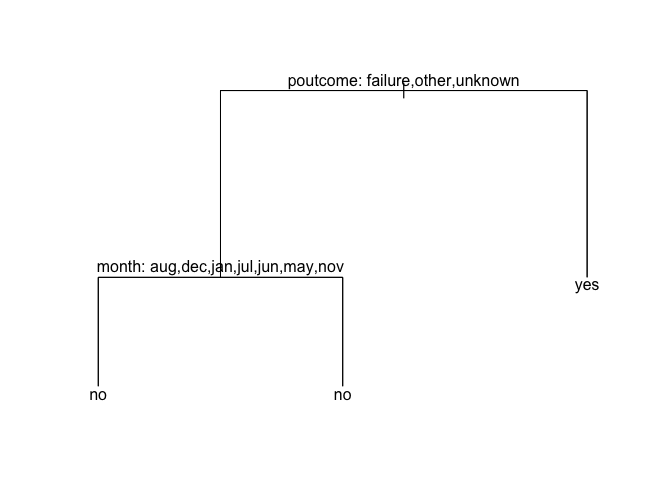
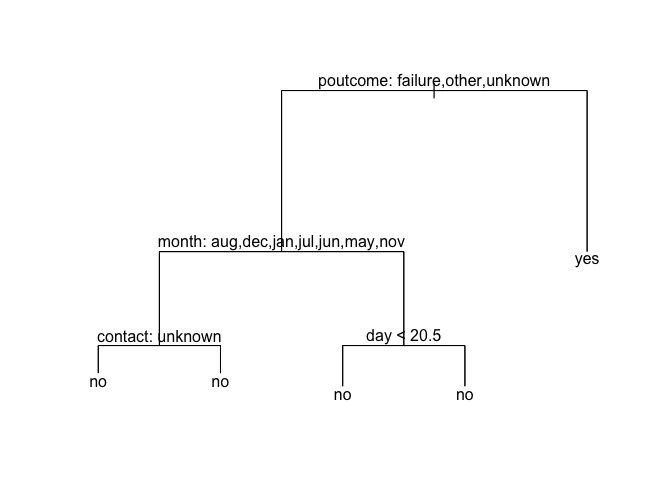
by using training and testing sets.

As we used trees in our previous dataset, we will use them now, only difference being is we use them for classification. They are called decision trees and are a very simple but powerful tool to classify and more importantly interpret and visualize data.

Decision trees for bank dataset Decision trees for bank dataset

Decision trees for bank dataset

Dde Decision trees for bank dataset



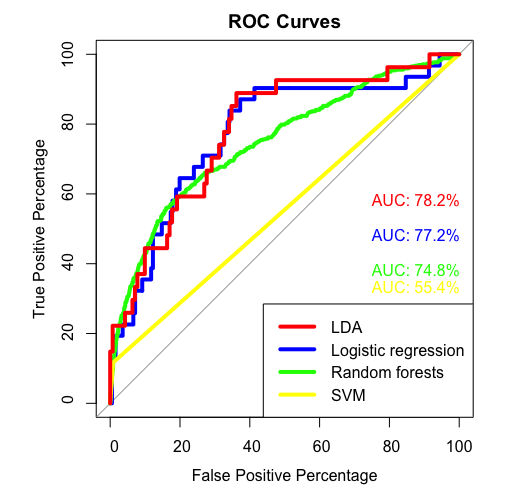
The left diagram is a tree fit on the training set of the bank data. As you can tell, it is very simple to interpret. It shows that ‘poutcome’ which is the outcome of the previous marketing campaign is most important variable in prediction.

The right diagram however is a plot of a pruned tree after cross-validation. Likewise, ‘poutcome’ remains the variable of utmost importance, and then month.

As previously explained, despite a decision tree being an easy tool to interpret the data with, it can be at times inaccurate. We will employ random forests, the more powerful tool based itself on building many decision trees. However, when applying random forests, it is important to vary parameters of model in order to maximize prediction accuracy and unlike other models, we do not require to split the data into train and test set, as it automatically does that by calculating the so-called Out Of Bag Error Rate.

I will briefly explain what linear discriminant analysis (LDA) is. It is method similar to PCA, that finds the linear combination of variables that separates classes, however unlike PCA’s who’s objective is to maximise variance, LDA finds the combinations that maximises separability between classes.

|  |  |  |
| --- | --- | --- |
| Model | Prediction Accuracy | AUC |
| Logistic regression | 0.876652 | 77.2% |
| Trees | 0.8858839 | - |
| Pruned-trees | 0.8878628 | - |
| Random forests | 0.891724 | 74.8% |
| SVM | 0.8957784 | 55.4% |
| LDA | 0.8452381 | 78.2% |



Even though all the models display a very high prediction accuracy, our ROC curves that for different probability thresholds, support vector machines accomplishes the worse, whereas the other methods are comparable both in terms of AUC values and prediction [[1]](#endnote-1)power.

1. References: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications

   in R. New York :Springer, 2013 [↑](#endnote-ref-1)