Machine Learning Project

Table of Contents

Introduction

Machine Learning consists of techniques that use algorithms on data to mimic the way humans learn, while improving its accuracy over time. It implements supervised and unsupervised learning methods to perform several tasks such as: analyses to visualise datasets; regressions to make predictions; and classifications of observations.

This project consists of three sections. These sections analyse the outputs of the datasets given to execute the three tasks outlined above to analyse and draw conclusions. The first section uses the unsupervised learning method, namely the Principal Component Analysis (PCA), to visualise and describe the European Working Conditions Survey 2016 dataset (EWCS). The second section is a Logistic Regression model built to predict the final grade of students in two subjects (Mathematics and Portuguese), from two Portuguese schools. A Regression Tree model was also constructed seeking to improve predictive performance. In the third section a classification model is built using the Classification Tree method to predict if a client will subscribe to a term deposit of a bank telemarketing dataset. Another classification method, the K-Nearest Neighbours, is also used in search of improved predictability.

This project was completed using the R programming language.

Part One

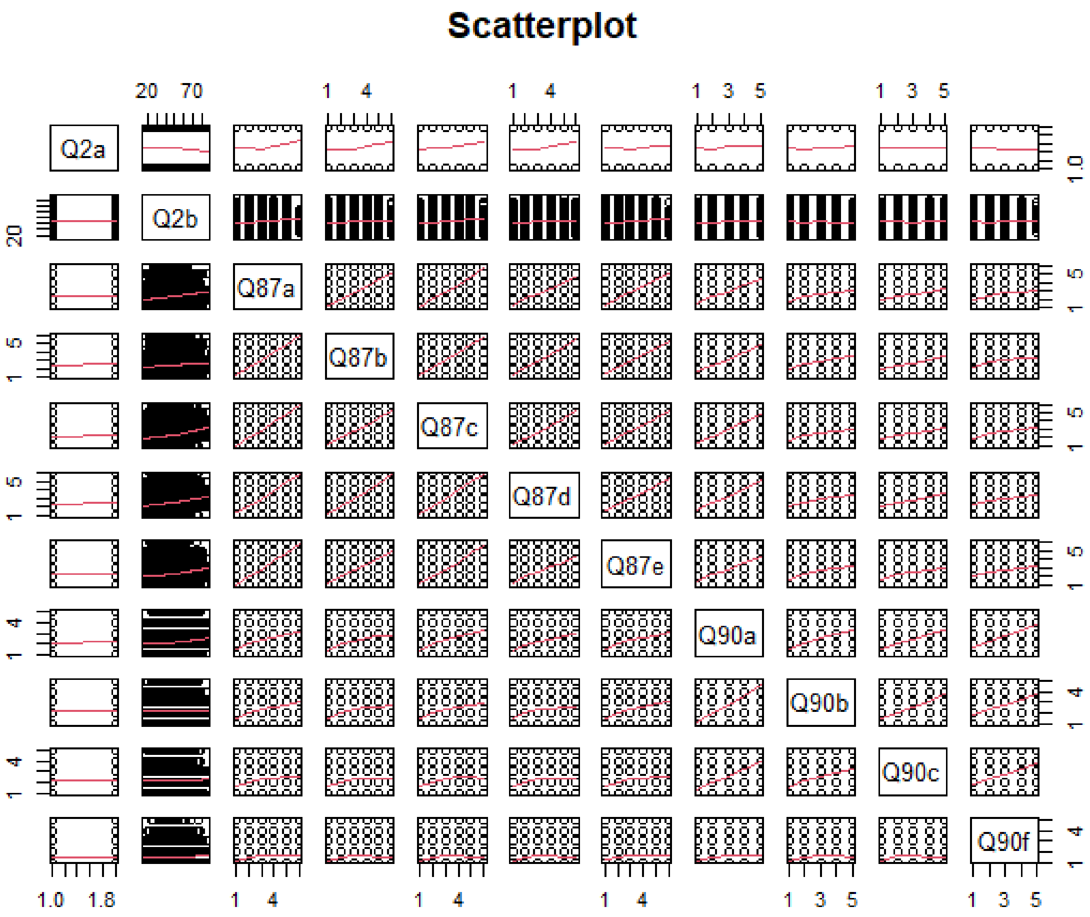
In this section, the EWCS dataset was provided to work with. The objective is to implement an unsupervised learning method to visualise and describe this dataset. The method chosen to perform this task is the PCA.

EWCS 2016 Dataset: Principal Component Analysis

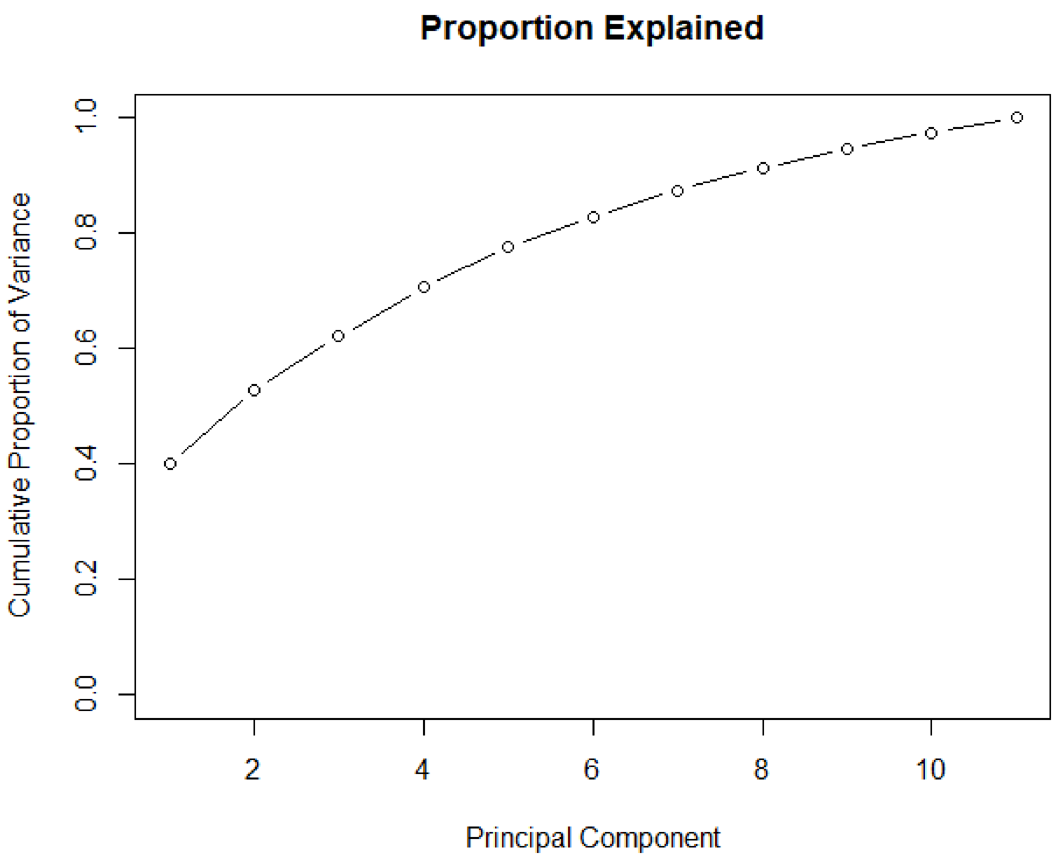
This survey contains lots of questions which can make it problematic to work with. PCA is a technique used to reduce the dimension of the dataset in question so that there will be fewer relationships between the variables and a reduction in the likelihood of overfitting the model.

The EWCS survey contains eleven (11) questions asked and seven thousand, six hundred and forty-seven (7,647) people completing the survey. PCA was performed and a few interesting findings were noticed.

To visualise the relationship between the questions asked in this survey, a scatterplot was created relating answers for each question to each other. Standard answers were given to choose from for the questions asked, and for those questions that were not about personal details (gender and age) the values were given a rank for which the higher the value selected the less likely what was asked in the question occurred. Also, the questions were all painted in a positive perspective. In this light, it is considered all questions asked are of a standard nature and the connections between them can be assessed.



For the questions that are not personal details, there are positive relationships between them. Persons who answered “All the time” or “Always” for one question tend to answer similarly for the other questions and vice versa. This means that persons who are in good spirits, calm, active, well rested and have a daily life full of things that interest them share similar traits when it comes to their jobs. However, it is observed that the relationship is a bit weaker, which may be caused by a slightly negative perception to working.



After performing the PCA on the dataset, it is found that some 62.1% of the variation in the data can be explained using three (3) groupings, while 70.7% can be explained using four (4).

Part Two

Two datasets were provided to build a regression model with the aim of interpreting the data and assessing the model's predictive performance. The data were surveys taken at two secondary schools in relation to the performance in two subjects: Mathematics and Portuguese.

The objective is to design a model that can predict the final grade of the students in the two subjects stated above. The two datasets were merged to remove duplicates (students that took both surveys). This merger resulted in having two variables for final grades, one for the final grades in Mathematics and the other for final grades in Portuguese. It also resulted in an increase in the number of variables from thirty-three (33) in the individual groups to fifty-three (53) in the merged group.

The two datasets used were also adjusted to remove all the quotation marks around the variables prior to being uploaded and merged so that each variable would be read into the model separately.

The first period grades and second period grades were removed from the merged dataset used. This is because the final grade which is being predicted is highly correlated with the first period and second period grades.

The merged dataset, which has three hundred and eighty-two (382) observations, was divided into two groups. The first group consisting of two hundred (200) observations was used to build the model, while the second group with the remaining observations was used to test its accuracy.

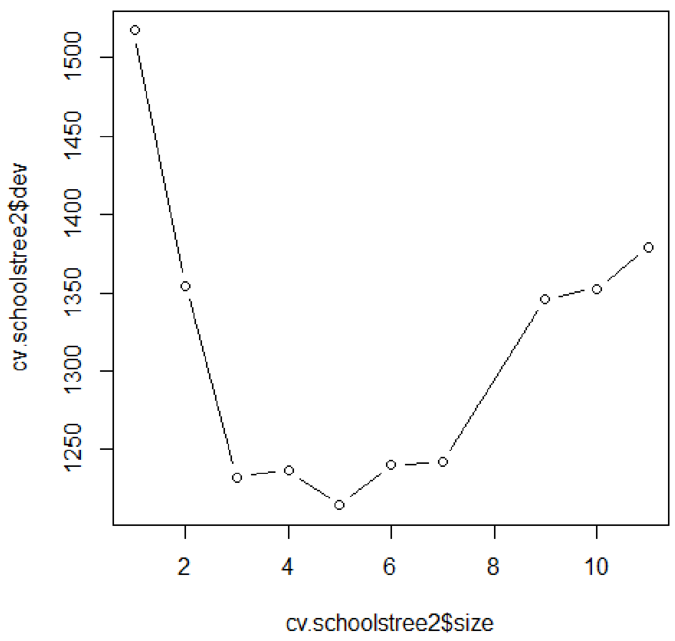
A regression tree model was first created to predict the final grades for the subjects and a review of the model's predictivity was completed. A random forest model was then built to predict the final grades to determine if this model has better predictivity than the regression tree.

Student Dataset: Regression Tree

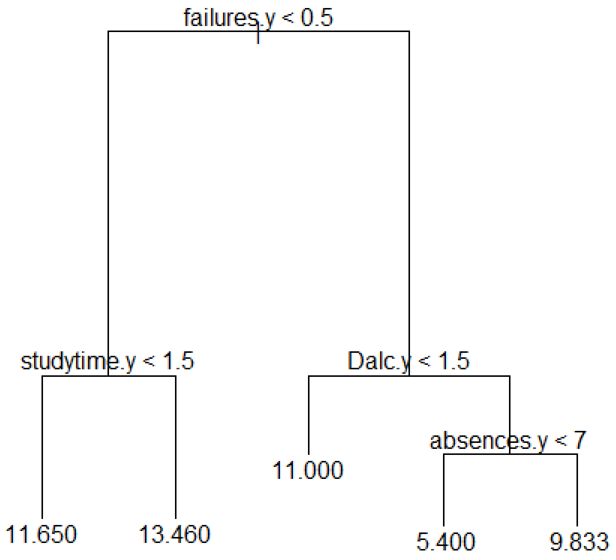
This model was used on the dataset to predict both Mathematics and Portuguese final grades using the other factors we have information for about the students, which has little to no correlation with the final grade.

Based on the result, this model was able to predict the grades of students within 4.33 marks of the true grade for Mathematics, and within 2.98 marks for Portuguese.

Using the Portuguese subject, it can be seen in the cross-validation graph below that a tree of size five (5) will produce the best results.

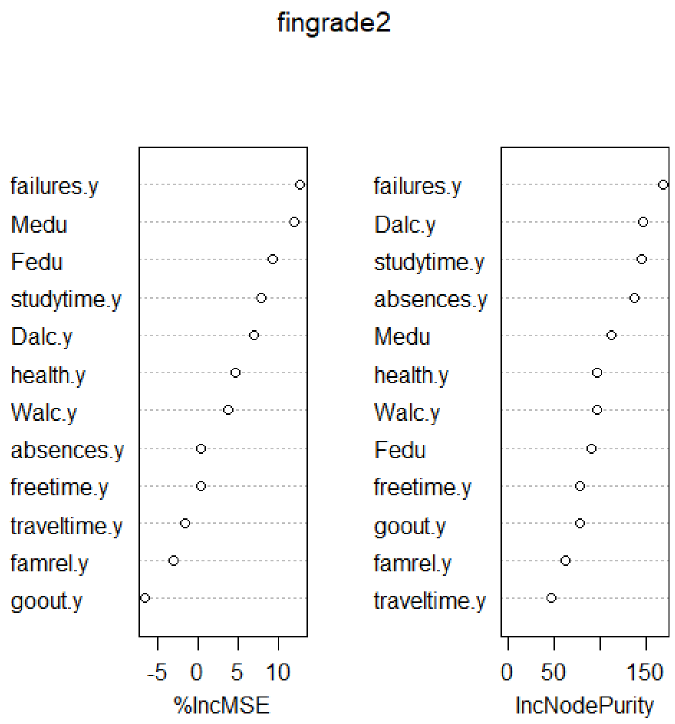


The graph below shows the regression tree using the recommended size from the graph above. It uses the: number of times a student failed previous classes; length of time they spend studying; amount of alcohol consumption during the week; and number of absences. These qualities may have an impact on a student's performance in school.



Student Dataset: Random Forest

The random forest model was applied to the same data for both Mathematics and Portuguese. There were very slight improvements in the ability to predict the final grades. For Mathematics, this model was able to predict the final grades of students within 4.16 marks, while for Portuguese it was able to do so within 2.83 marks.



In the display above, the importance of the different attributes towards the final grade prediction is shown for the Portuguese subject. The same attributes used in the regression tree can be seen as the most important contributors in this method.

It may be difficult to predict a student's grade. There are many factors that need to be considered which can affect their performance: Factors outside of those used in the survey, such as possible tragedies experienced, perception of the subject or mental abilities, and factors within the attributes, such as whether they actually studied, or scope of the syllabus studied during study time.

Part Three

In an attempt to predict if a client would subscribe to a term deposit using the marketing campaigns data of a Portuguese bank, a classification model was constructed using a classification tree. Another model was built using the K-Nearest Neighbours method seeking for an improvement in the accuracy.

The dataset used was also adjusted to remove all the quotation marks around the variables prior to being uploaded so that each variable would be read into the models separately.

Looking at the summary of the data, the average age of the participants in the campaigns were 41 years, with the youngest being 19 years old and the oldest being 87 years old. Most of them have management level jobs followed by blue-collar jobs. The least number of persons were retired.

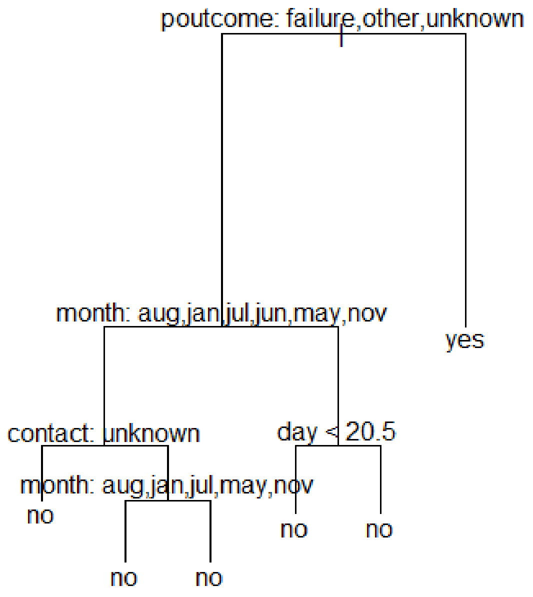
More than half of the participants are married, while a little more than quarter are single. Most of them are educated up to a secondary school level and the second most have tertiary level education.

Bank Dataset: Classification Tree

The classification tree used four areas in its prediction: outcome of the previous marketing campaign, last month contacted of the year, communication type used to contact participants, and the last contact day of the week.

The duration of the last contact is left out of the model due to its significant impact on the output target, e.g., once someone is not contacted, then the result would be that they will not subscribe. When someone is contacted, whether he/she will subscribe would be known. Also, once someone is not contacted, then they will not subscribe.

The diagram below depicts the classification tree determined by the model.


poutcome: failure,other,unknown
month: aug,jan,jul,jun,may,nov
yes
contact: unknown
day < 20.5
month: aug,jan,jul,may,nov
no
no
no
no
no

From the diagram, the model considered that the client would subscribe only based on the previous marketing campaign outcome.

This model provided an ***89.8%*** accuracy in its predictions which is a good result.

Some 1,521 clients were used to test the model's predictivity. Some 1,341 clients were accurately predicted to not subscribe and 25 to subscribe. The model incorrectly predicted that 16 would subscribe and 139 would not subscribe.

In an attempt to refine the model by using less branches, the outcome was the same since the tree shows that the decision to subscribe is solely based on previous marketing campaign outcomes, and not based on the months contacted, days contacted and method of contact. These other categories do not have an impact on a person's decision.

Bank Dataset: K-Nearest Neighbours

Another method used to identify whether a client would subscribe to a term deposit was the K-Nearest Neighbours method. This method was built in an attempt to increase the accuracy of prediction.

Again, with the duration of the last contact highly affecting the output, it was left out of the model.

The accuracy of predicting whether a client would subscribe to a term deposit was good but decreased slightly from 89.8% using the classification tree method to 89.1% using K-Nearest Neighbours.

```
=======================================
                 predicted default
actual default       no     yes   Total
---------------------------------------
no                 1329      28    1357
                  0.874   0.018
---------------------------------------
yes                 138      26     164
                  0.091   0.017
---------------------------------------
Total              1467      54    1521
=======================================
```

Out of the 1,521 clients used to test the predictivity, the model predicted accurately that 1,329 clients would not subscribe and 26 would subscribe. It also incorrectly predicted that 28 clients would subscribe and 138 would not subscribe.

Conclusion

The three (3) tasks required were completed: the first task being to visualise and describe the European Working Conditions Survey 2016, the second was to perform regression to predict a student's final grade, and third being to classify a bank's marketing campaigns to predict if a client will subscribe a term deposit.