Project Report

William Chan (25961039), Yung Chyi Siah (27717518), Connor Lickliter (27794628), Mihir Bhatt (25175319), Alec Kajewski (26895765)

May 20, 2018

Introduction

This project presented a task that is regularly faced by data analysts; it challenged us to produce the best model to predict the probability that a prospective client will subscribe to a bank term deposit on the basis of several predictors, with the end goal being that this model could be used to direct and target the bank's future marketing campaigns. The data was based on multiple phone calls to the prospective clients, with seven predictors and seven variables being included. The seven predictors were a variety of different attributes of the prospective client, with multiple different types of data including categorical, binary and ratio data. The variables related to the last contact of the current marketing campaign, with the data again coming in several different forms. For our training set, we were provided with both the predictors and the variables, however, for our test set, off which we based our models, we only had access to the predictors. Therefore, the aim of this analysis was to create the most accurate classifier, as measured against the test set, which could be used to direct future marketing campaigns.

A brief summary of some of what was thought to be related variables including jobs, marital status and whether someone already has credit in default can be seen below:

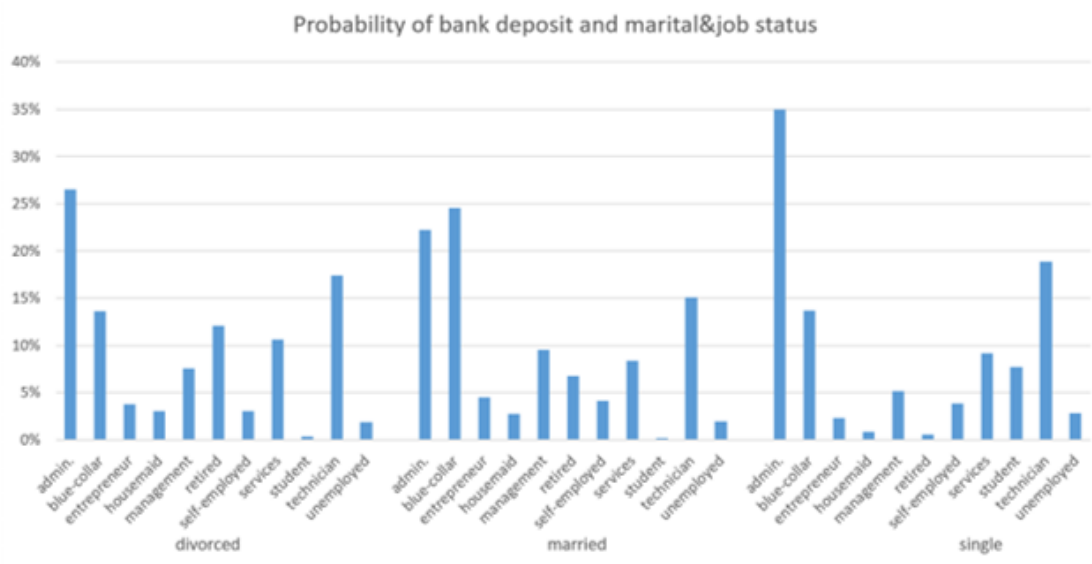

Figure 1: Sumary of job and marital status and their effect on the probability of someone subscribing to a bank deposit.

From these two plots alone it can be seen that the probability that someone will subscribe to a bank deposit is highly dependent on their field of occupation. In Figure 2, only participants with no credit defaults were plotted due to the small amount of people with credit in default. Once again it is logical that people with no loan is more likely to put more money into the bank. Based on other similar summaries, relevant and non-relevant variables can be identified and methods to predict this probability can be developed.

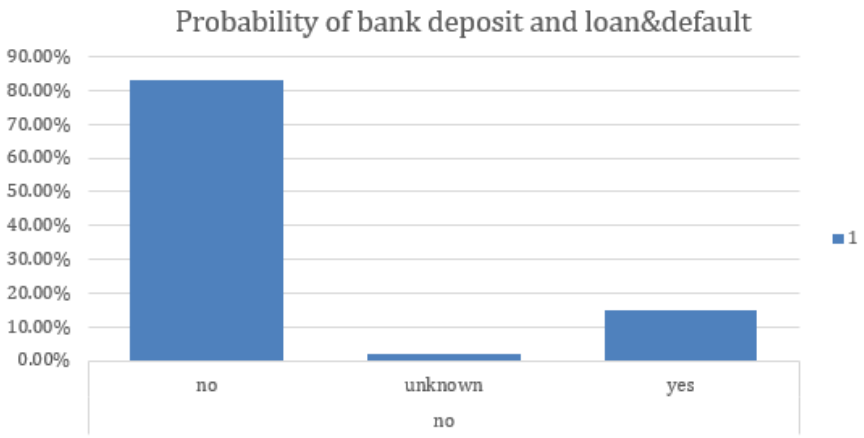

Figure 2: Probability of bank deposit against existing credit status.

Methodology

To get a better understanding of the various predictors in the data set, various plots and tables were initially created to show the range and concentration of the values and to identify any potential outliers. For example, a histogram was created to determine the distribution of the predictor age.

From here, a logistic regression model was constructed as many of the predictors were qualitative and binary. By utilising a logistic regression, the probability that Y belongs to certain predictors can be modeled rather than modelling the response Y directly (Nordhausen, K., 2014).

Initially, a model with many variables was predicted as seen below.

```
fit1 = glm(y ~ age + job + marital + edu + housing + loan, data = my_data, family=binomial)
```

From this initial model, the other predictions were tried where less independent variables were used to limit the chances of overfitting. Furthermore, to gauge the final scores of the model when submitted to Kaggle, the predictions were also evaluated using the function of in sample log loss as seen below. The predictions from the model are first calculated and then written as a data frame.

```
fit1.pred = predict(fit1, test_set, type="response")
fit1.train.pred = predict(fit1, type="response")
prediction = data.frame(1:10182, fit1.pred)
log_loss = function(prediction){
  log_loss = -mean(my_data$y*log(prediction) + (1 - my_data$y)*log(1-prediction))
  return(log_loss)
}
```

By utilsing the in sample log loss different models were able to be compared. Furthermore, relevant and potentially significant variables were able to be identified as these fits would result in a lower in sample log loss. Furthermore, various other combinations of the predictors were tried, including using age^2 and categorising the jobs into those that had more than 1000 responses (admin, blue-collar, entrepreneur, self-employed, services and technician) and those with less than 1000 responses (housemaid, retired, student, unemployed and unknown).

From here, other methods such as probit regression and decision trees were tried. Whilst logistic and probit regression are similar in that they are both generalised linear models that can be used to model categorical predictor variables and are generally written as $\hat{Y} = f(\alpha + \beta x)$, they differ in that logit models use the cumulative distribution function of the logistic distribution to define f(), whereas probit models use the cumulative distribution function of the standard normal distribution to define f().

The following probit model tried was:

```
fit4 = glm(y ~ single + default + loan + pdays + poutcome_binary,
           family = binomial(link = "probit"), data = my_data)
```

Decision tree models divide the data into smaller subsets and were tried for their potential to capture any nonlinearity in the predictors. However, these models did not give better predictions than the logistic regressions.

Results and Discussion

We began with a logit model using a wide range of predictor variables, then attempted to reduce the number of variables used through the use of significant levels and by manipulating the data set. One example of this is with the jobs variable, rather than having all of the job types included, the variable was reduced to a 1 only if the participant is retired or a student otherwise it was 0. This was done as only the student and retired observations were significant in the logit regressions. This approach was done with other variables that only had some observations of significance such as the outcome of the previous campaign and marital status. However, the models that resulted, while all variables were significant, did not have the predictive power as measured by the log loss function as one of the models that was attempted when surveying the data.

An example model using these new variables is found below:

```
fit = glm(y ~ retiredstudent + single + poutcome_binary,
            data = my_data, family = binomial)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.61321    0.02654 -98.473  < 2e-16 ***
retiredstudent   0.76599    0.08278   9.253  < 2e-16 ***
single           0.25854    0.04731   5.465 4.63e-08 ***
poutcome_binary  1.98677    0.20530   9.678  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: Summary of example model

The three coefficients correspond to binary variables of whether a participant is retired, student, single or whether they have subscribed to the bank term deposit in the past. All three variables are ones that should have a high impact on a person's financial stability and thus correlate to how likely they will sign up. However this model indicates that being single, student or retired will increase the probability that someone will join the deposit program. The in sample log loss of this model performed very well but it did not achieve a good test set score on Kaggle.

The model with the lowest out of sample log loss is as follows as well as the coefficients of the model:

```
fit = glm(y ~ marital + default + loan + pdays + poutcome, data=my_data, family=binomial)
```

As can be seen, all else being held equal most of the variables have the effect of decreasing the probability a customer will subscribe to a deposit. In particular a person who has credit in default results in a much higher chance of them not joining. This is in line with real world expectations. Unexpectedly, a married person as well as someone who has previously joined also has a lower probability of joining according to this model.

A probit version with the same variables used was also submitted but did not achieve as low a log loss as the logit model.

A number of alternative methods were trialed but were found to be ineffective in comparison to the linear regression model used in the final prediction. One method that was tested was a tree based method, which used recursive division of data to form a regression tree. However, several problems were encountered when using this method.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.013088   1.015446   0.013   0.9897
maritalmarried       -0.015704   0.070173  -0.224   0.8229
maritalsingle         0.211766   0.075447   2.807   0.0050 **
maritalunknown        0.317973   0.475606   0.669   0.5038
defaultunknown       -0.332609   0.055167  -6.029 1.65e-09 ***
defaultyes           -9.098839 113.712158  -0.080   0.9362
loanunknown          -0.102615   0.147489  -0.696   0.4866
loanyes              -0.028102   0.060748  -0.463   0.6436
pdays                -0.002428   0.001020  -2.380   0.0173 *
poutcomenonexistent  -0.059189   0.103086  -0.574   0.5659
poutcomesuccess      -0.518722   1.030481  -0.503   0.6147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Summary of chosen model

The tree function found that only two of the variables in the data set were significant: age and month which resulted in a set of extremely limited predictions as seen in Figure 5. While the software has chosen to select the month of the last contact as an important variable, it does not seem likely that whether a person will join a term deposit is based on when they were contacted. The tree predicts that if a person is contacted in March then the probability of them subscribing is higher at 0.481. Depending on the location the data was gathered from, March could be the end of a financial year or when tax returns occur and thus a person is more likely to deposit into the bank. However, the log loss of these models performed badly when compared to the regression models performed previously. When the random forest function was tested, complications arose due to the binary (classification) dependent variable in the training data set being used to form a probability regression for the test set prediction.

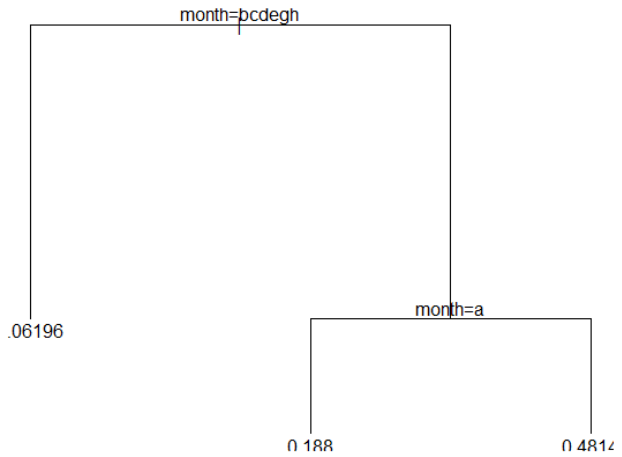

Figure 5: Decision Tree from R based on all variables in the dataset.

Other restrictions around the models that could be used resulted from several of the significant variables being categorical rather than continuous. This meant that linear discriminant analysis was also unable to be utilised effectively in the prediction.

The evaluation method that was chosen to decide which models to submit was the in-sample log loss. However it was discovered that a low in-sample log loss is not indicative of a low test set log loss also. The following plot shows that while the in-sample log loss remained fairly consistent with only small changes the test set log loss is affected dramatically. It can even be seen that while the in-sample log loss is decreasing, the out-of-sample log loss is going up. This meant there was no concrete method by which to decide which models to submit.

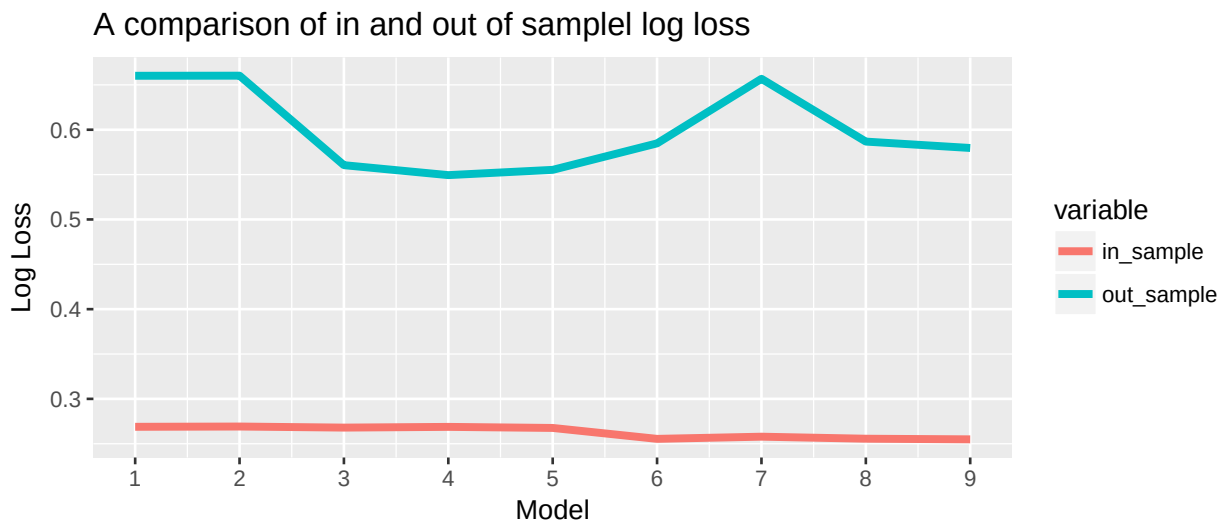

Figure 6: A comparison of in and out of sample log loss using submitted models.

As can be seen in the plot again, the second peak of the out of sample log loss occurred when the prediction method was altered from regression methods such as logit and probit to decision trees. However, due to the limited number of trees that were possible, the score obtained was not as low as the logistic models.

Conclusion

Through thorough investigation into the data set obtained through a marketing campaign of a banking institution, a series of models were tested and independently assessed in order to find the most effective model for predicting the likelihood of client subscription. The main findings from this analysis were that the logistic regression model was the most efficient method to form predictions due to the complications arising from the other various methods, as well as the logistic regression model's simplicity. The following logistic regression model was found to have the lowest in sample log loss:

```
fit = glm(y ~ marital + default + loan + pdays + poutcome, data=my_data, family=binomial)
```

With a test log loss of approximately 0.55, it suggests that the final model has a reasonable accuracy in finding the probability that a client will subscribe to a long term deposit.

To further increase the accuracy, and hence lower the log loss of future models, it is suggested that various other methods of classification are investigated, such as clustering and tree based methods. Additionally, for the logistic regression model, further investigation into transformations of each significant predictor could take place.

References

Nordhausen, K., 2014. An Introduction to Statistical Learning-with Applications in R by Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani. International Statistical Review, 82(1), pp.156-157.