FINAL PROJECT: BANK MARKETING MODEL & DATA ANALYSIS REPORT

Presented To:

Vitaly Druker
Professor of Data Analysis (STAT 716)

Prepared By:

Peter Salamon
Bryan Ronga
Halid Idrissou

December 9th, 2018

Introduction

Our final project revolves around the bank marketing dataset. The scenario that we adopt, along with this dataset, is that a Portuguese bank seeks to increase its subscriber base. In pursuit of this goal, the institution conducts campaigns in which potential clients are contacted and their information collected. The bank's aim is to predict whether an individual will subscribe to their bank based on the person's demographic information and data from previous contacts with the potential client. The goal of this analysis is to produce a statistical classification model that predicts whether or not a client will subscribe a term deposit.

The data used in this analysis is hosted on the University of California Irvine's Machine Learning Repository website which can be found at https://archive.ics.uci.edu/. The dataset was originally collected by a Portuguese banking institution during their marketing campaigns, conducted via telephone, between May of 2008 and November of 2010. The dataset was also originally use in the research paper entitled *Data-Driven Approach to Predict the Success of Bank Telemarketing* [Moro et al., 2014] in which researchers propose a "...data mining approach to predict the success of telemarketing calls for selling bank long-term deposits" (Moro et al.). Each observation of the dataset consists of several features that describe the client's demographic information and financial status. The dependent variable of the set is a simple binary 'yes' or 'no' which represents whether or not the client had successfully purchased a subscription with the bank.

Description of Data

The entire data set was divided into a train-test split of two seperate files. Taking a closer look at the training data, there are a total of 31,647 observations and 17 features, including the dependent variable, in all. Each feature, it's corresponding meaning, and data type are as follows:

'*age'* - The age of client. (Numeric)

'*job'* - Type of job held by client. (Factor)

'*marital'* - The marital status of the client. (Factor)

'*education'* - The highest level of education completed by the client. (Factor)

'*default*' - Has the client ever defaulted on a previous debt? (Factor)

'*balance'* - The client's average yearly balance, in euros. (Numeric)

'*housing'* - Does the client currently possess a housing loan? (Factor)

'*loan'* - Does the client currently possess a personal loan? (Factor)

'*contact'* - The type of communication over which the client was reached. (Factor)

'*month'* - The month of year that the client was contacted. (Numeric)

'*day'* - The day, of the month, that the client was contacted. (Numeric)

'*duration'* - The length of the last communication with the client, in seconds. (Numeric)

'*campaign'* -The number of times, during the current campaign, that the client was reached. (Numeric)

'*pdays'* - The number of days ago that the client was last reached. (Numeric)

'*previous'* - The total number of times that the client was contacted. (Numeric)

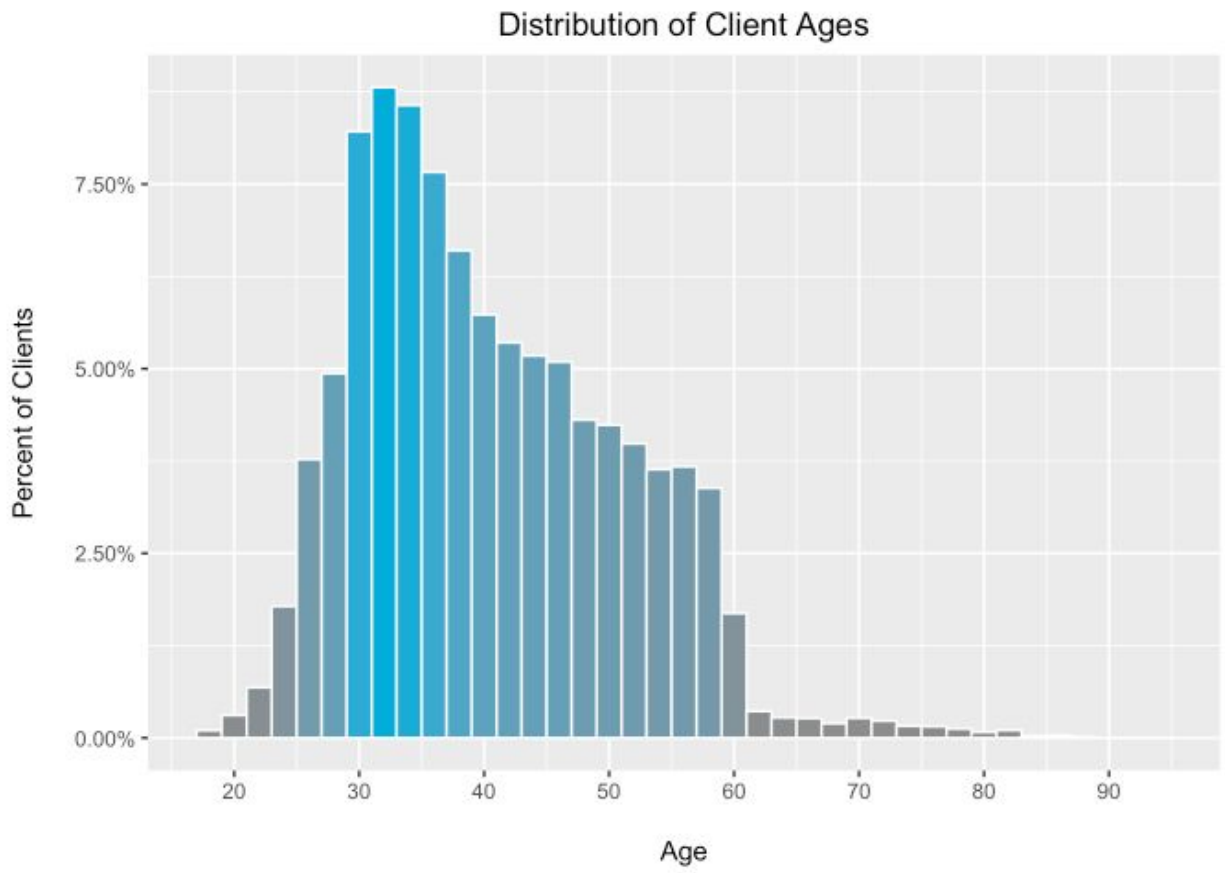
'*poutcome'* - The outcome of the previous marketing campaign on the client. (Factor)

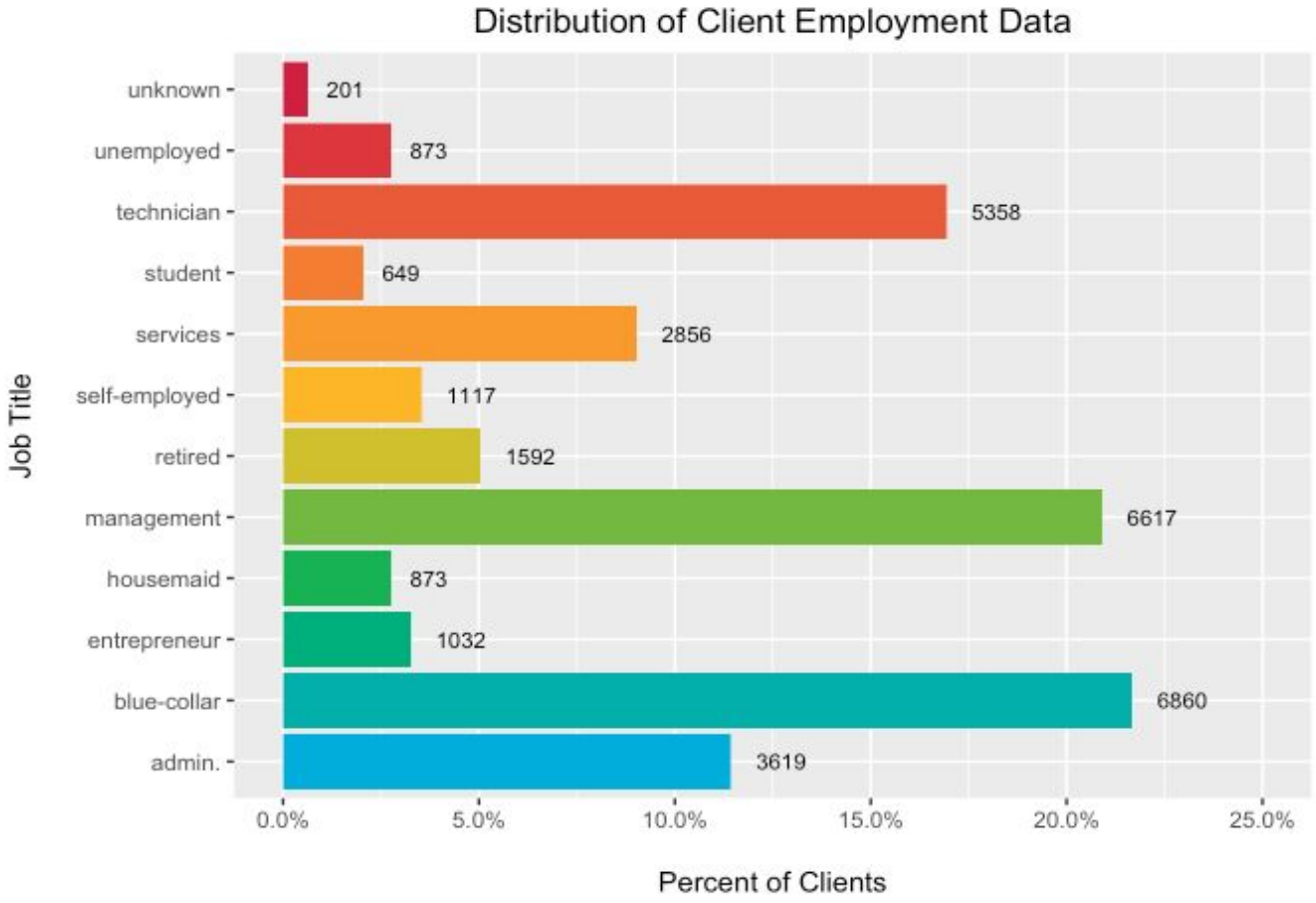
'*y*' - Whether or not the client has subscribed a term deposit. (Factor; our dependent variable)

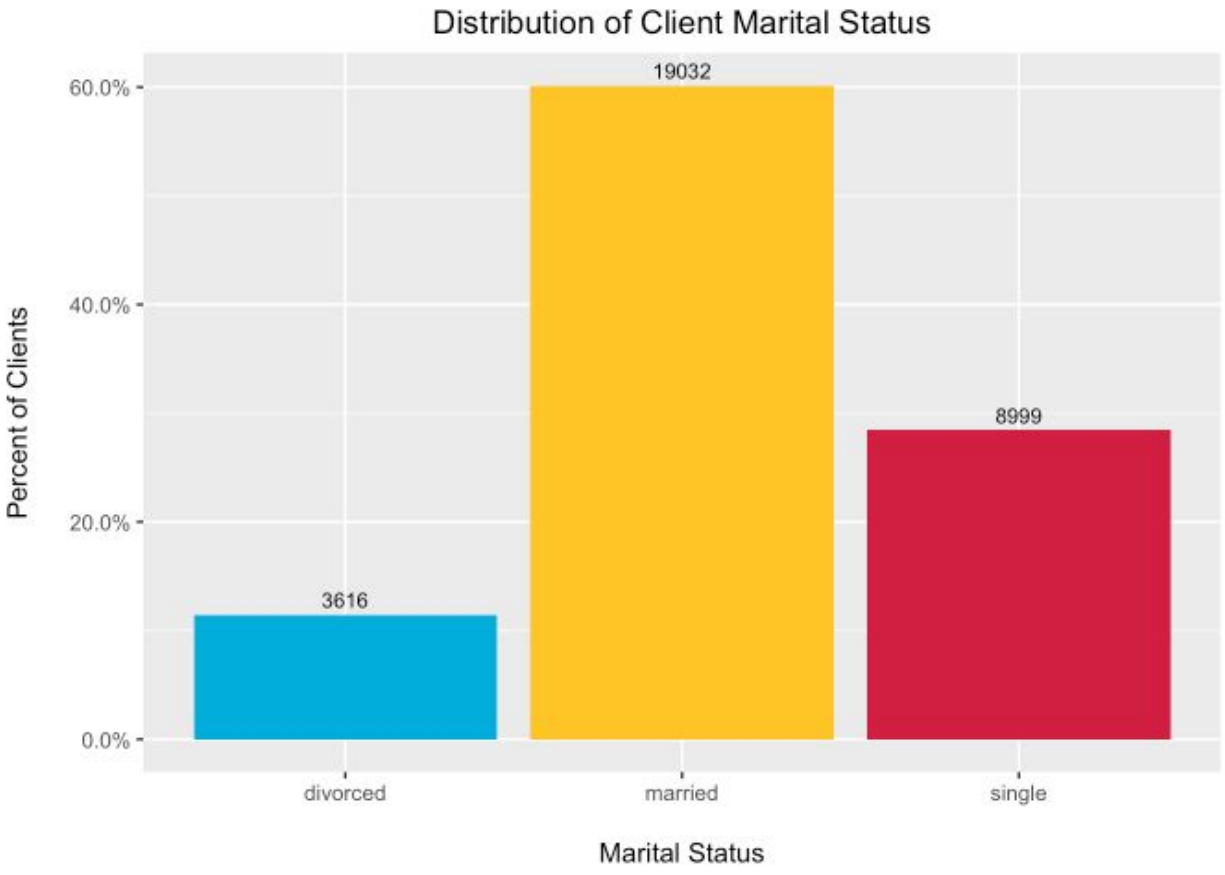
A summary report was obtained for the training data which specified the min, median, mean, max, and quartile values for each feature, if it was a numeric variable, and the unique values with their corresponding counts, if it was categorical. The corresponding output can be seen below.

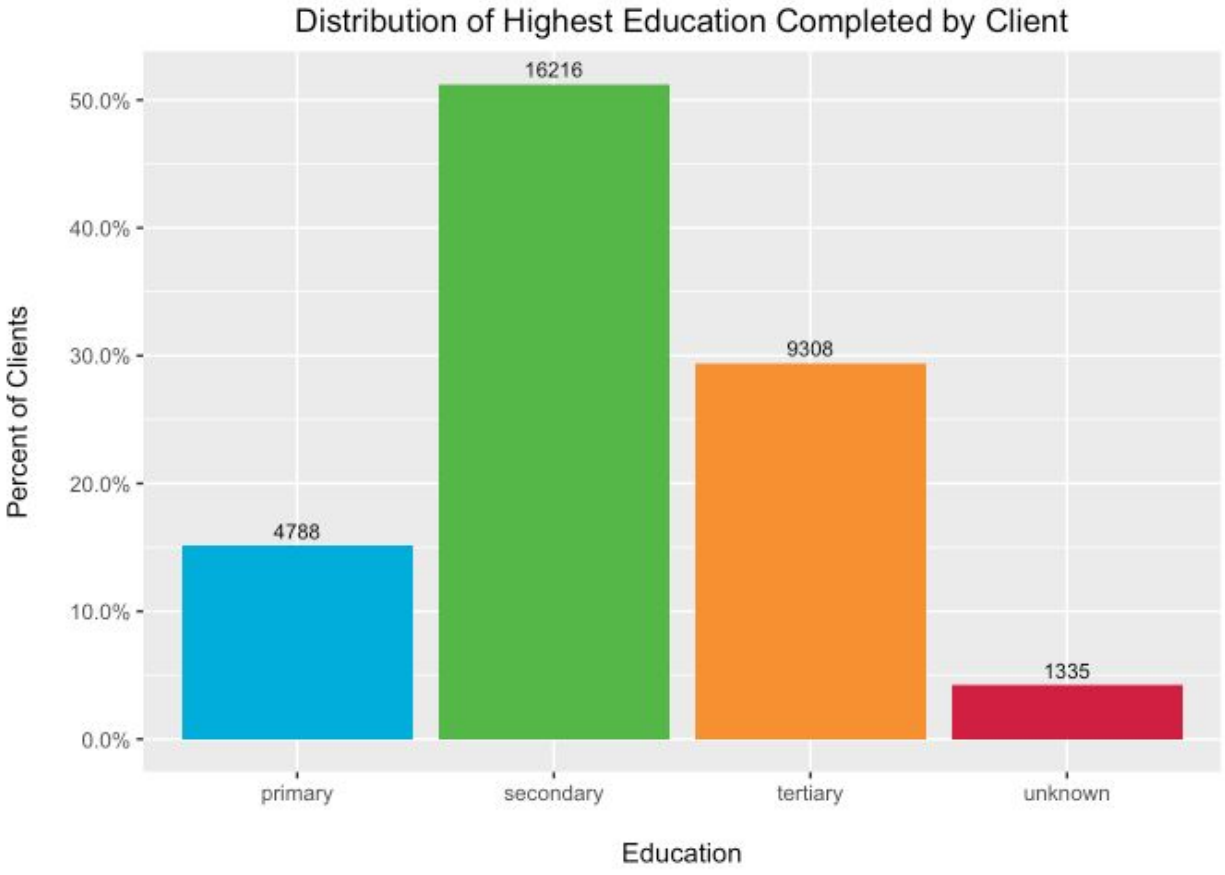
```
##       age                job          marital          education
##  Min.   :18.00   blue-collar:6860   divorced: 3616   primary  : 4788
##  1st Qu.:33.00   management :6617   married :19032   secondary:16216
##  Median :39.00   technician :5358   single  : 8999   tertiary : 9308
##  Mean   :41.01   admin.     :3619                    unknown  : 1335
##  3rd Qu.:49.00   services   :2856
##  Max.   :94.00   retired    :1592
##                  (Other)    :4745
##  default        balance       housing      loan            contact
##  no :31079   Min.   : -6847   no :14079   no :26629   cellular :20444
##  yes:  568   1st Qu.:    72   yes:17568   yes: 5018   telephone: 2035
##              Median :   450                           unknown  : 9168
##              Mean   :  1351
##              3rd Qu.:  1420
##              Max.   :102127
##
##       day            month         duration         campaign
##  Min.   : 1.00   may    :9678   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    :4827   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    :4379   Median : 179.0   Median : 2.000
##  Mean   :15.81   jun    :3782   Mean   : 258.3   Mean   : 2.776
##  3rd Qu.:21.00   nov    :2745   3rd Qu.: 317.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    :2047   Max.   :4918.0   Max.   :63.000
##                  (Other):4189
##      pdays           previous          poutcome       y
##  Min.   : -1.00   Min.   : 0.0000   failure: 3443   no :27961
##  1st Qu.: -1.00   1st Qu.: 0.0000   other  : 1281   yes: 3686
##  Median : -1.00   Median : 0.0000   success: 1039
##  Mean   : 40.06   Mean   : 0.5809   unknown:25884
##  3rd Qu.: -1.00   3rd Qu.: 0.0000
##  Max.   :871.00   Max.   :58.0000
```

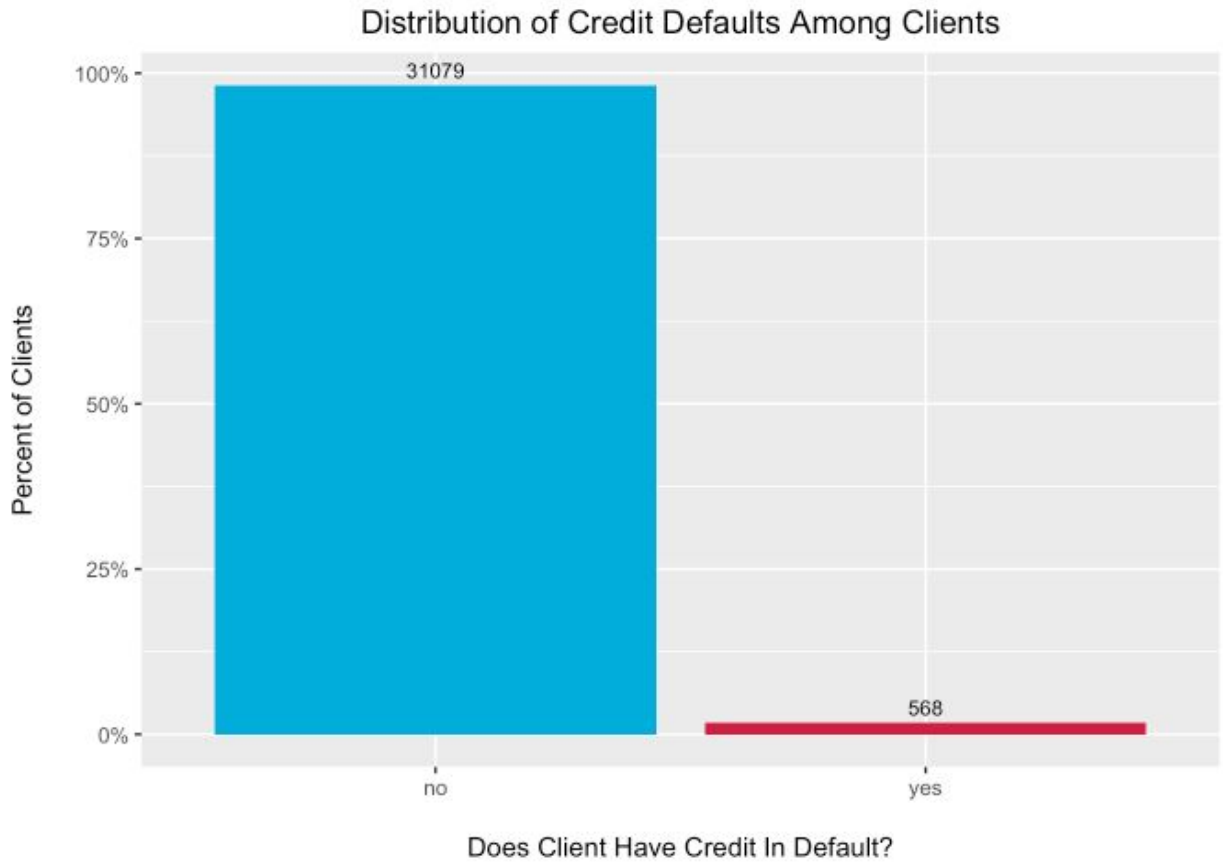
This report served as a holistic representation of each individual feature. What followed was a visual analysis of each feature in graphical form.

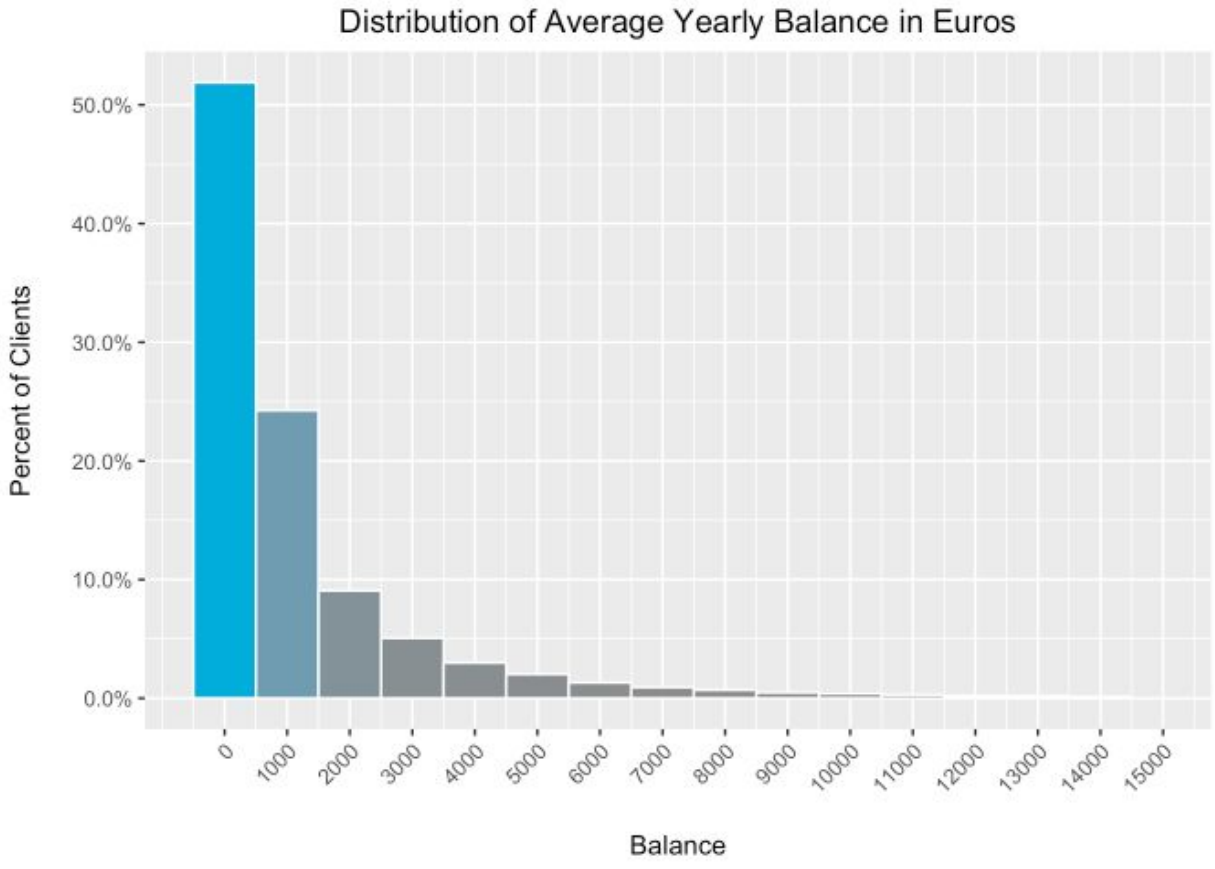

Distribution of Client Ages
Percent of Clients
7.50%
5.00%
2.50%
0.00%
20
30
40
50
60
70
80
90
Age

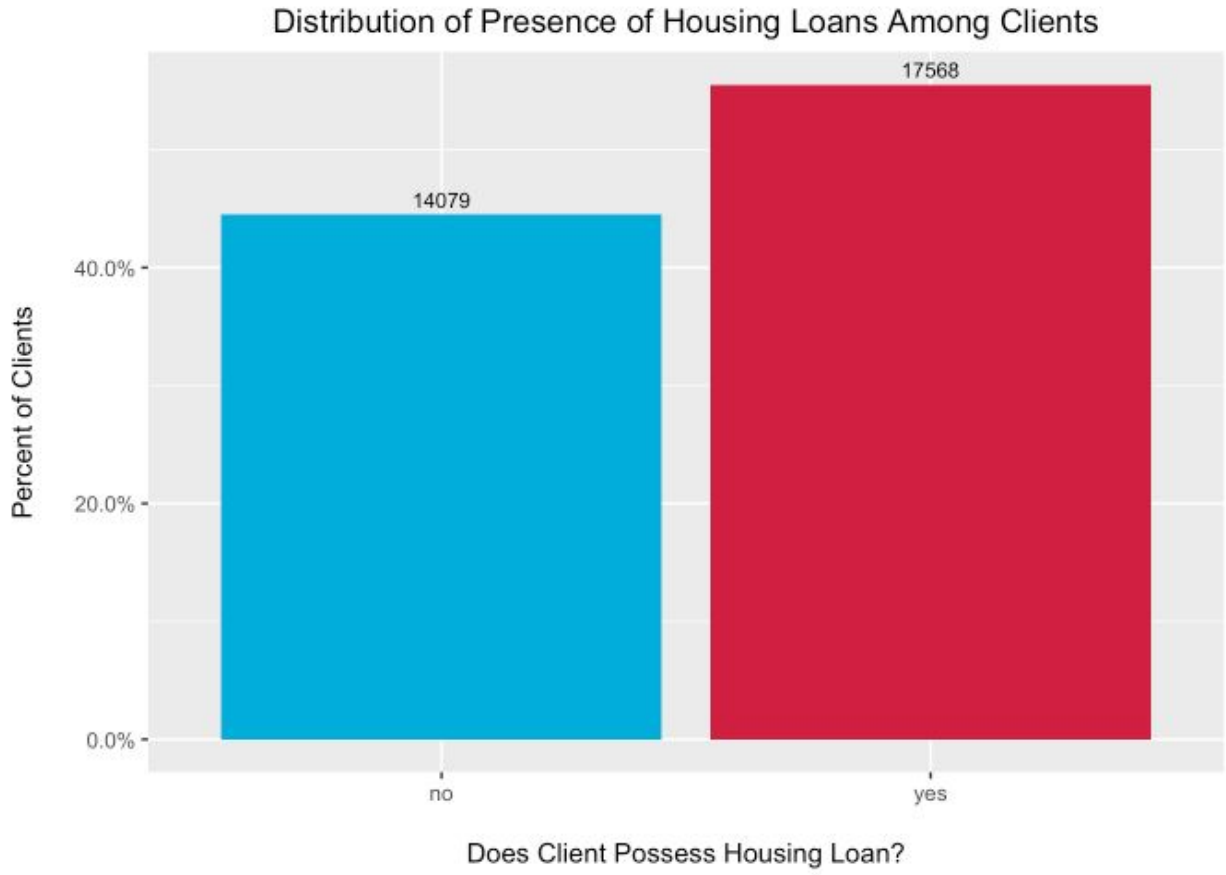

Distribution of Client Employment Data
Job Title
unknown
unemployed
technician
student
services
self-employed
retired
management
housemaid
entrepreneur
blue-collar
admin.
201
873
5358
649
2856
1117
1592
6617
873
1032
6860
3619
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
Percent of Clients

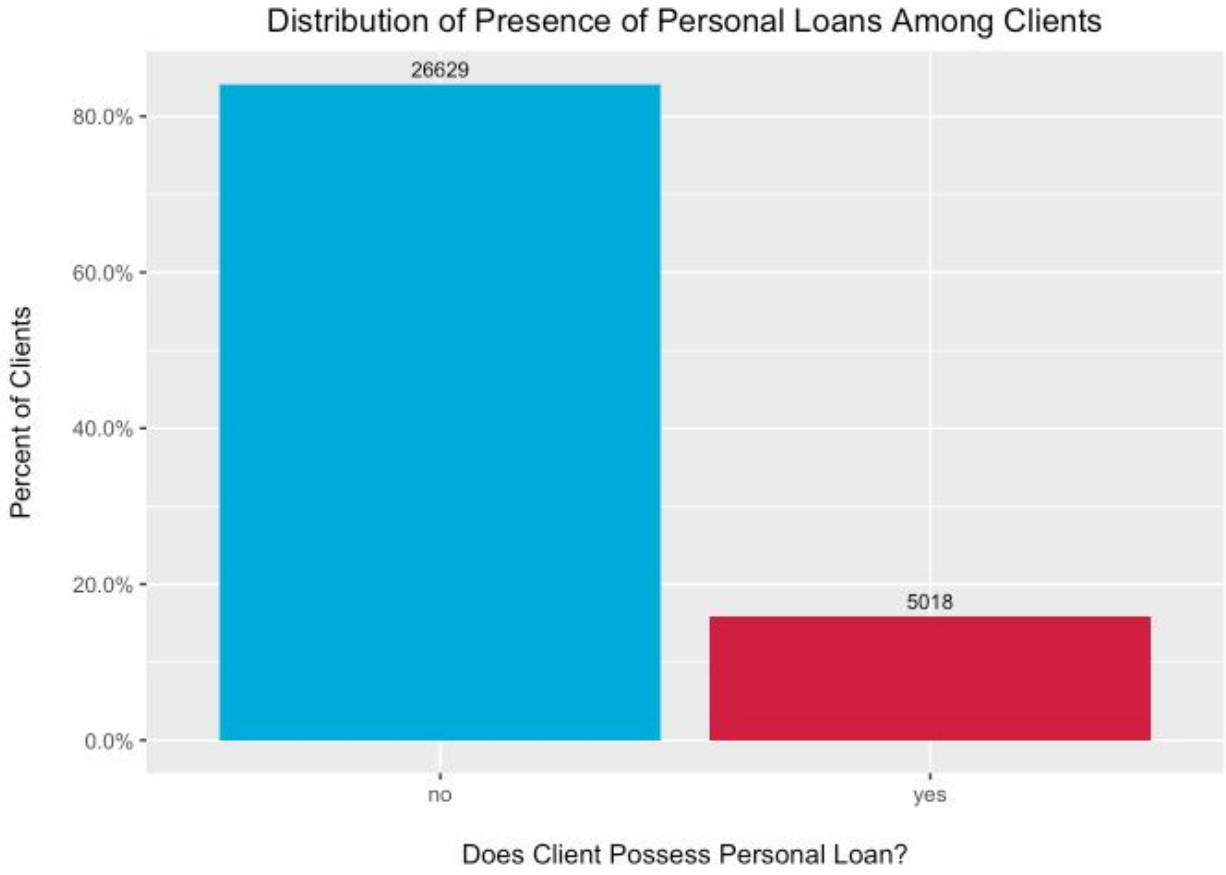

Distribution of Client Marital Status
60.0%
40.0%
20.0%
0.0%
Percent of Clients
19032
8999
3616
divorced
married
single
Marital Status

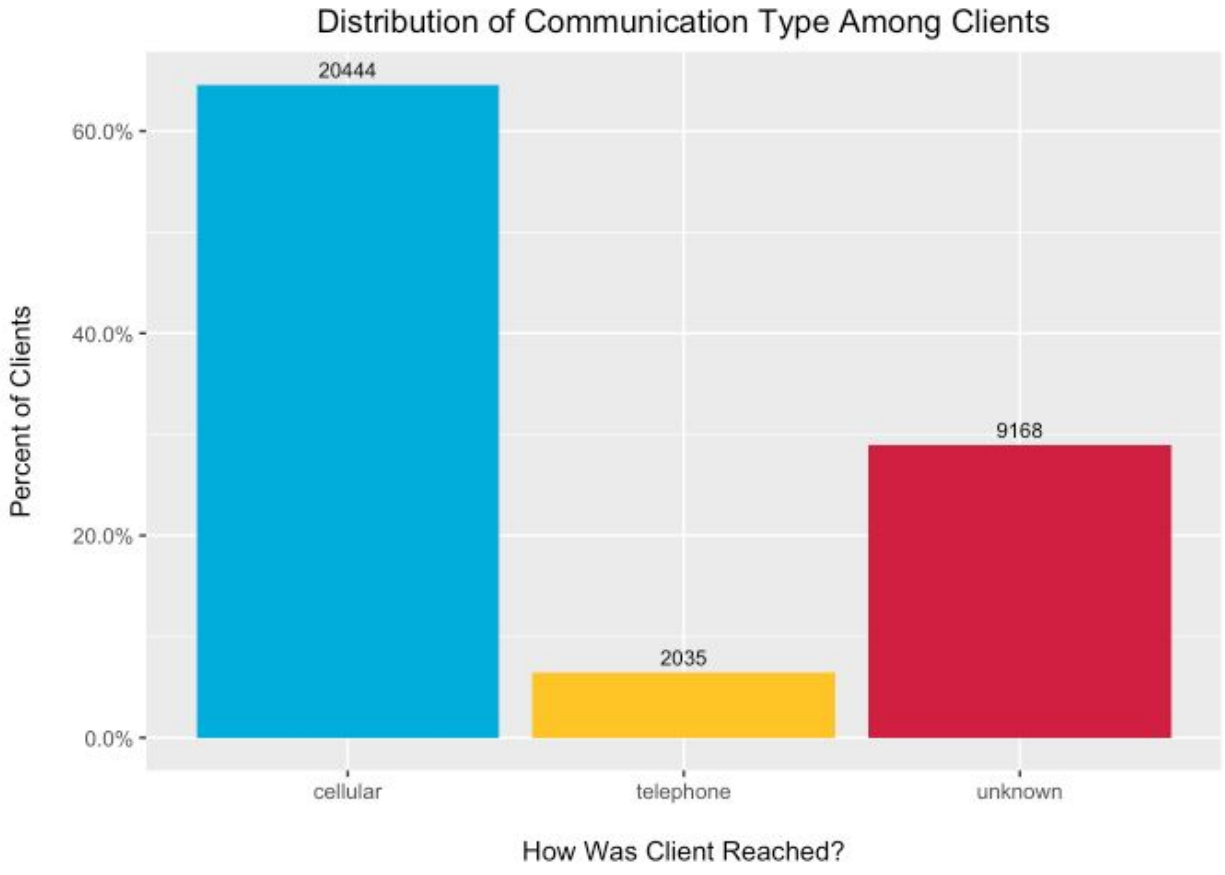

Distribution of Highest Education Completed by Client
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Percent of Clients
16216
9308
4788
1335
primary
secondary
tertiary
unknown
Education

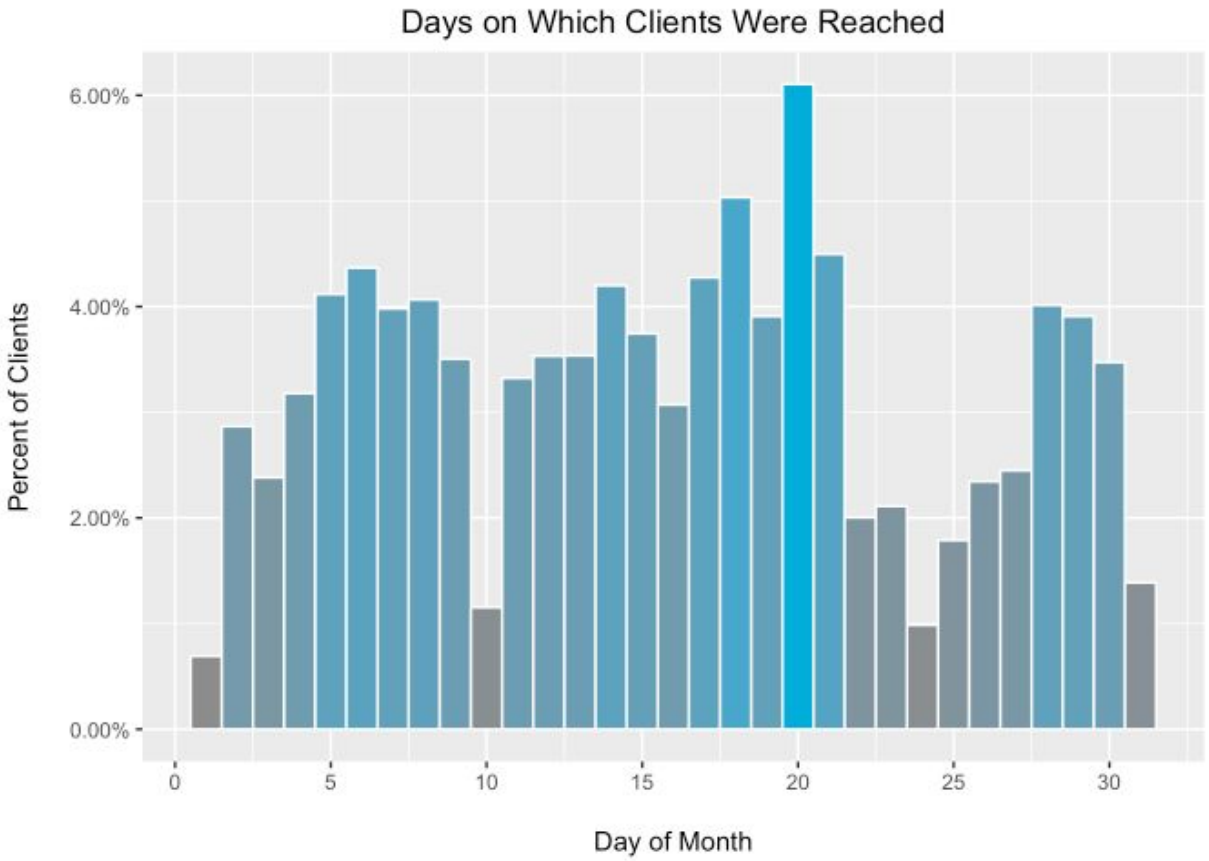

Distribution of Credit Defaults Among Clients
100%
75%
50%
25%
0%
Percent of Clients
31079
568
no
yes
Does Client Have Credit In Default?

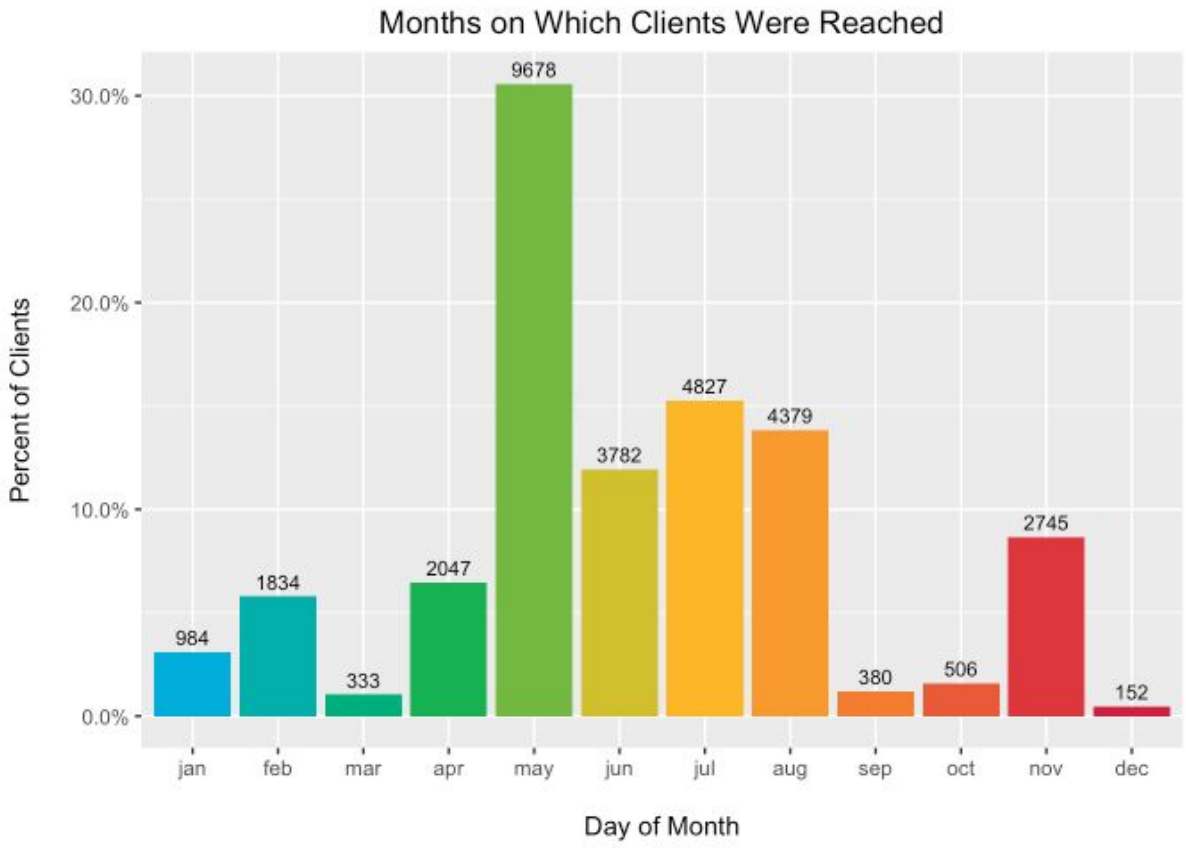

Distribution of Average Yearly Balance in Euros
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Percent of Clients
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
Balance

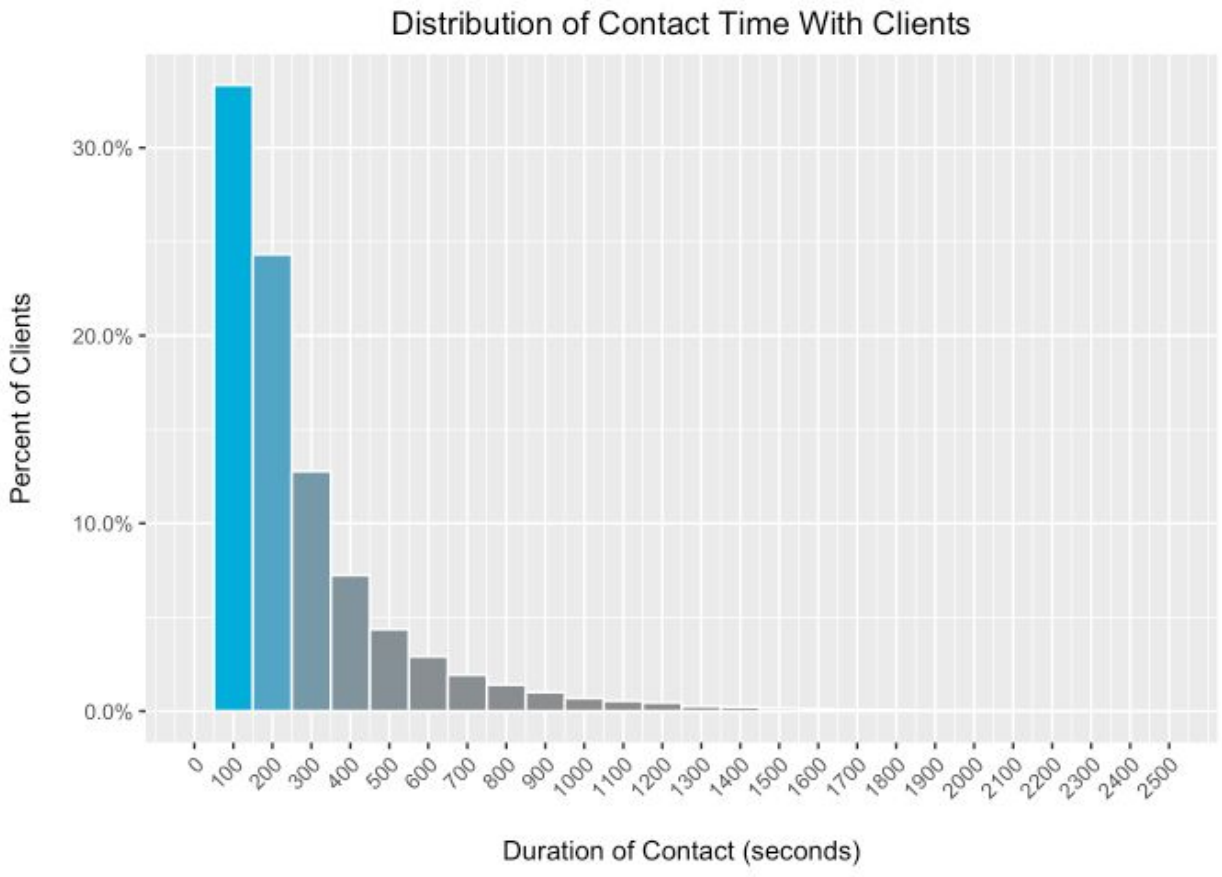

Distribution of Presence of Housing Loans Among Clients
17568
14079
40.0%
20.0%
0.0%
Percent of Clients
no
yes
Does Client Possess Housing Loan?

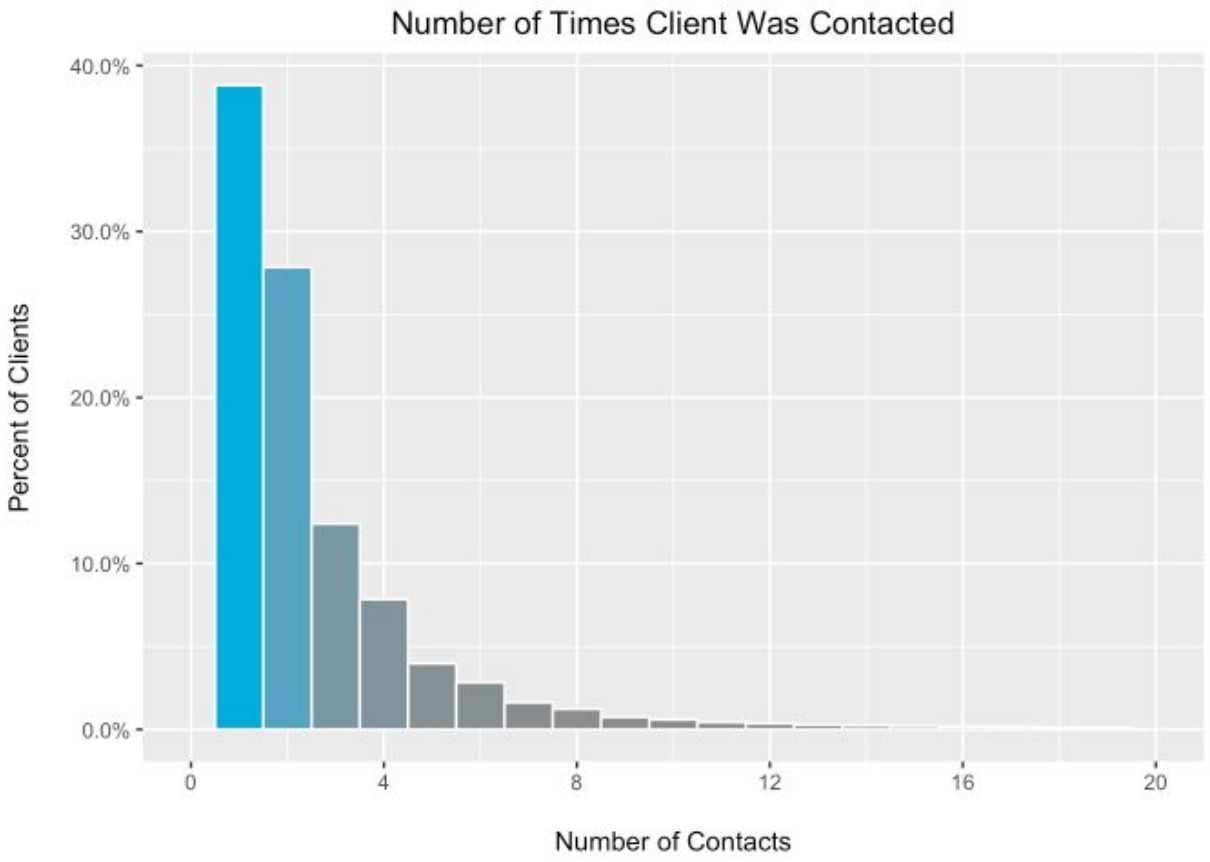

Distribution of Presence of Personal Loans Among Clients
26629
80.0%
60.0%
40.0%
20.0%
0.0%
5018
Percent of Clients
no
yes
Does Client Possess Personal Loan?

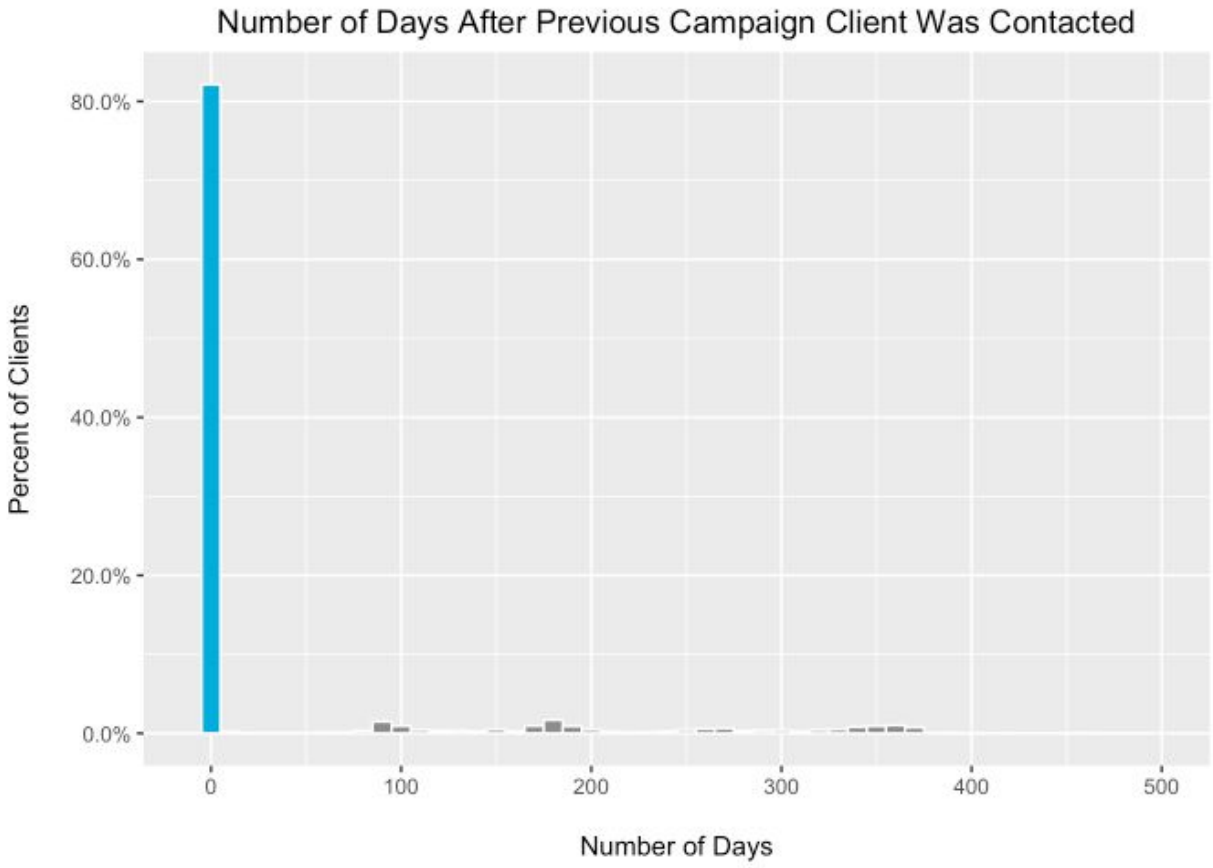

Distribution of Communication Type Among Clients
20444
9168
2035
60.0%
40.0%
20.0%
0.0%
Percent of Clients
cellular
telephone
unknown
How Was Client Reached?

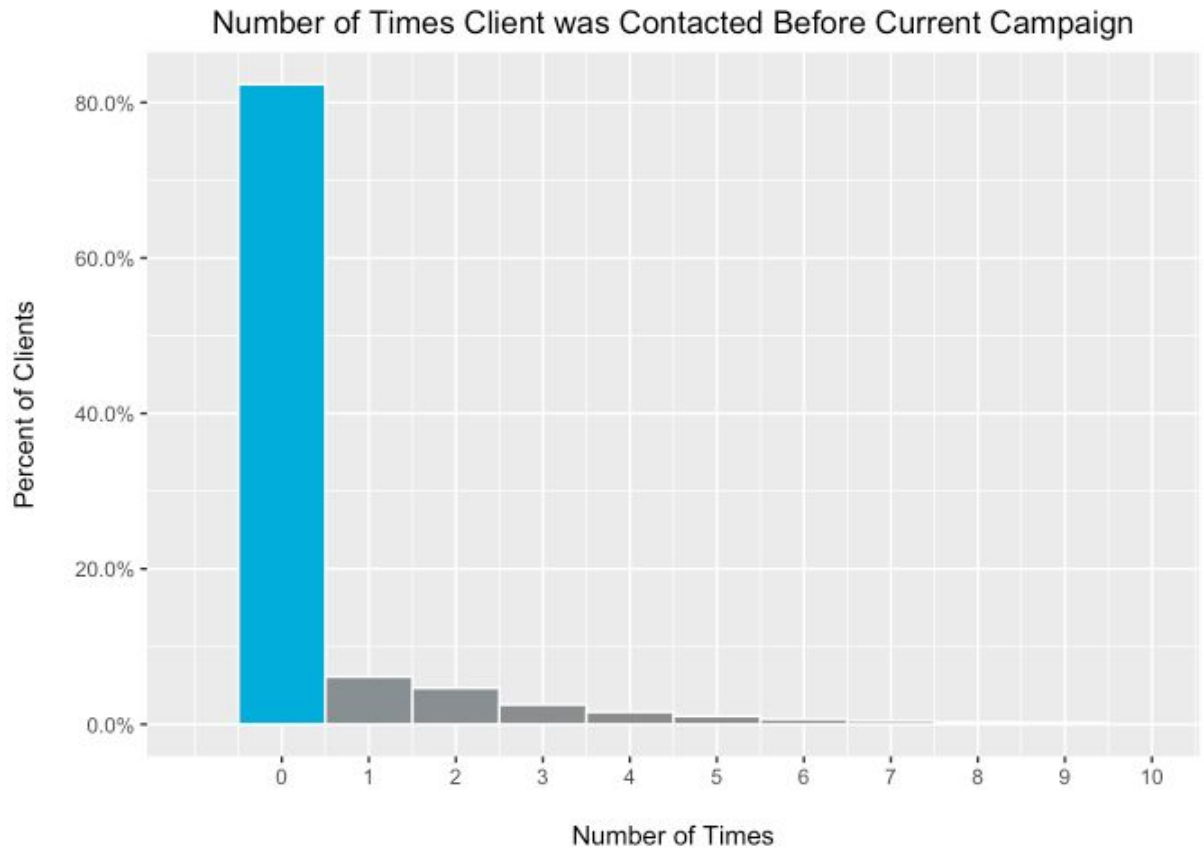

Days on Which Clients Were Reached
6.00%
4.00%
2.00%
0.00%
Percent of Clients
0
5
10
15
20
25
30
Day of Month

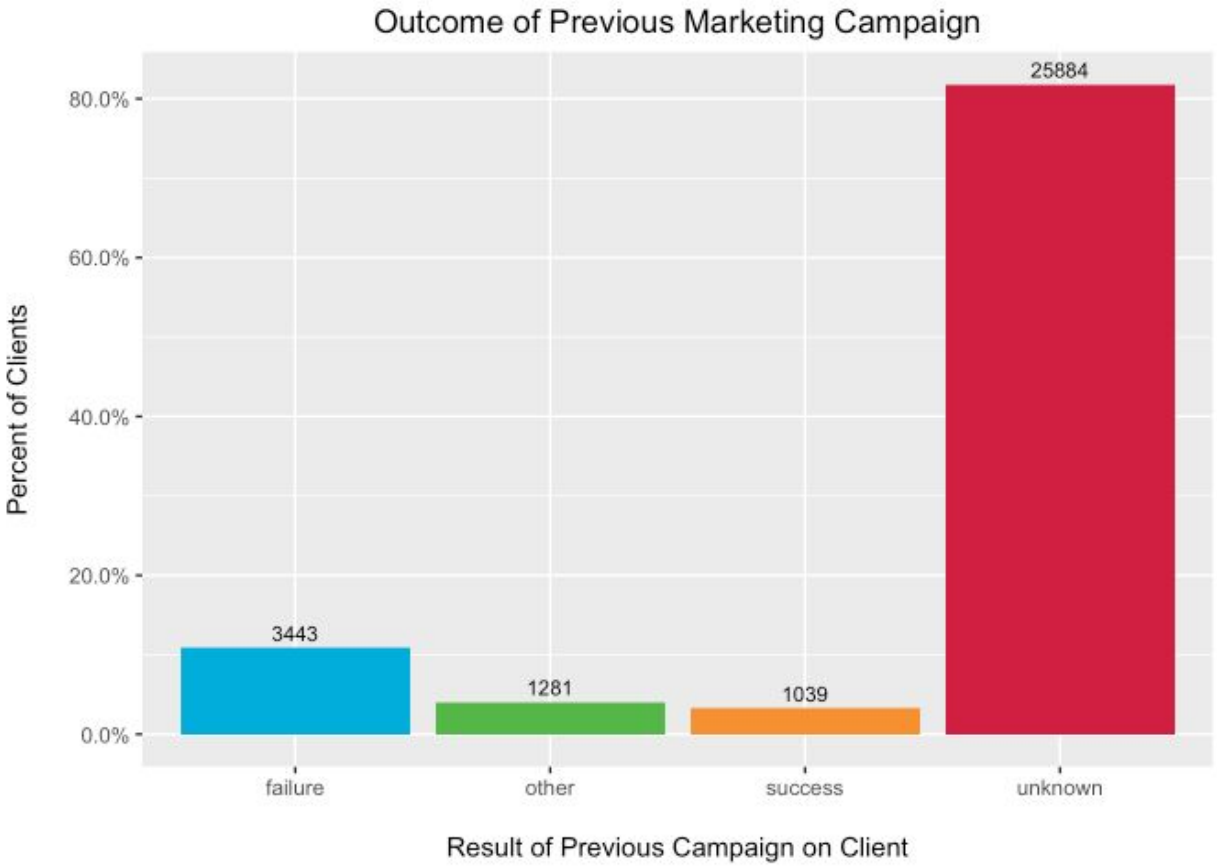

Months on Which Clients Were Reached
Percent of Clients
30.0%
20.0%
10.0%
0.0%
984
1834
333
2047
9678
3782
4827
4379
380
506
2745
152
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
Day of Month

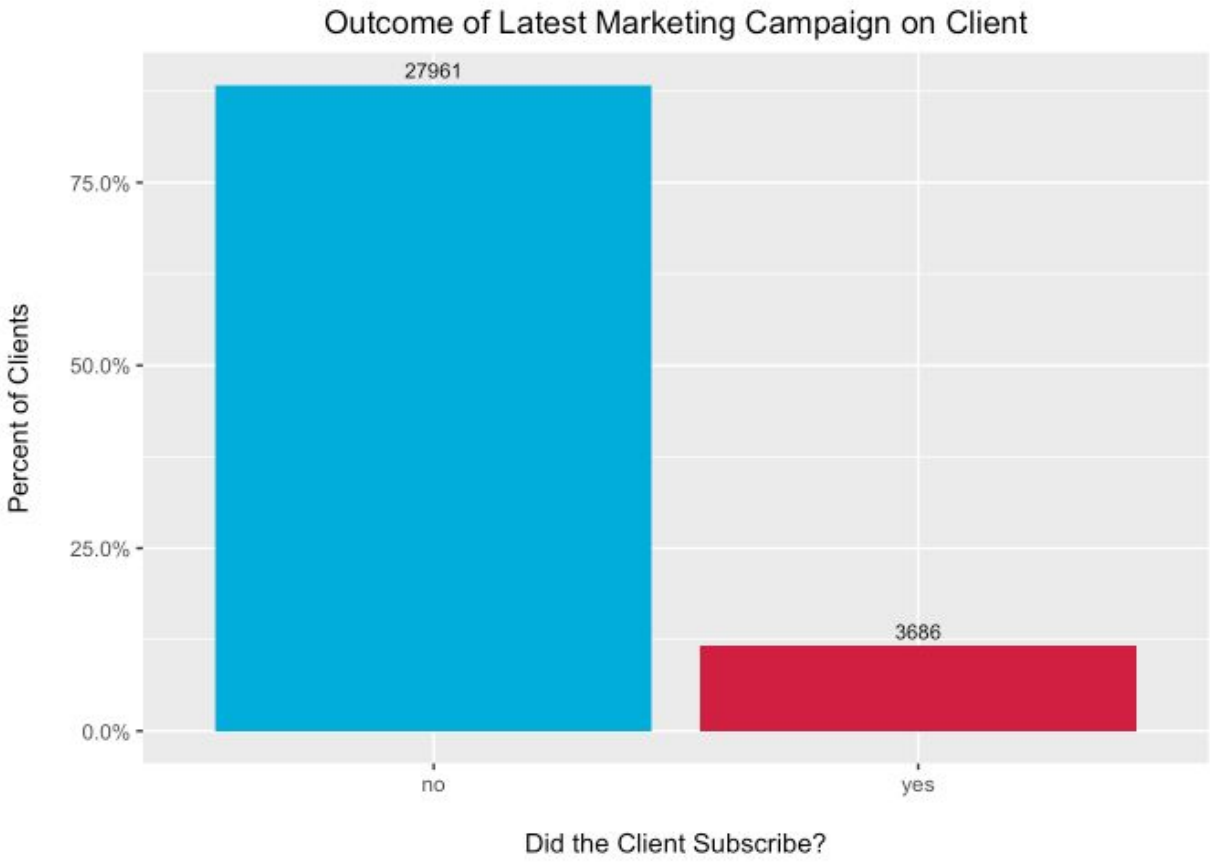

Distribution of Contact Time With Clients
Percent of Clients
30.0%
20.0%
10.0%
0.0%
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
Duration of Contact (seconds)


Number of Times Client Was Contacted
Percent of Clients
40.0%
30.0%
20.0%
10.0%
0.0%
0
4
8
12
16
20
Number of Contacts


Number of Days After Previous Campaign Client Was Contacted
Percent of Clients
80.0%
60.0%
40.0%
20.0%
0.0%
0
100
200
300
400
500
Number of Days


Number of Times Client was Contacted Before Current Campaign
Percent of Clients
80.0%
60.0%
40.0%
20.0%
0.0%
0
1
2
3
4
5
6
7
8
9
10
Number of Times


Outcome of Previous Marketing Campaign
Percent of Clients
80.0%
60.0%
40.0%
20.0%
0.0%
3443
1281
1039
25884
failure
other
success
unknown
Result of Previous Campaign on Client



Of all the plots, the final plot, above, is the most striking. The graph exhibits the number of 'yes' and 'no' responses we have, for our dependent variable, in the training set. We will refer to the 'yes' responses are positive cases and the 'no' responses as the negative cases. Over 80% of the cases are negative. This will need to be taken into consideration during model building and performance analysis as this may cause our models to be biased towards the prediction of the negative cases making it harder for the positive cases to be predicted.

Other considerations that will need to be made include whether or not there are numerical variables that would be better suited as factor variables. We must make sure that each feature is of an appropriate data type in order to prevent our model from learning any false relationships between the independent variables and the dependent variable. It may be best for discrete numerical variables, such as age, month, or day, to be turned into factor variables. Such features

do not possess values in between them and would not make sense to treat them as continuous, numerical values. In addition, we should consider which variables, in any, should be left out of the model. Features with large amounts of missing data would provide little predictability to a model and may even introduce false relationships or cause overfitting as well. From the graphs of 'Outcome of Previous Marketing Campaign' and 'Number of Days After Previous Campaign Client Was Contacted', we see that the feature 'poutcome' is missing concrete values for over 80% of the observations in the data. Likewise, 'pdays', although a numerical variable, holds the values '-1' which makes up ~80% of its observations. According to documentation accompanying the datasets for this analysis, the value '-1' is used to encode a missing value. These features may be removed as they provide little additional information and hold the potential of introducing false relationships to the models. A final consideration that needs to be made is the statistical model that will be used to predict our dependent variable. Because our dependent variable is a binary category, the model we choose will be some sort of classifier. Popular classification algorithms include Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). In order to decide on which of these models to continue with for the duration of the analysis, base models should be fit using each algorithms and their corresponding performances evaluated.

Methods & Results

During model selection, a model was fit using all four of the algorithms previously mentioned. Below is a table with each model type, fit and evaluated on a train-test split on the original training data, along with their accuracies, specificities, and sensitivities on the test set.

Model\Metric	Accuracy	Sensitivity	Specificity
Logistic Regression	0.8927	0.9835	0.2035
SVM	0.8862	0.9814	0.1642
Random Forest	0.8964	0.9485	0.5007
KNN	0.8832	0.9862	0.1018

During the analysis, the R Markdown environment delegated our positive cases for the dependent variable as the 'negative class'; and so, focus was put on maximizing the amount of positive cases captured by or model. In other words, focus was directed towards maximizing specificity. Due to its comparative accuracy and high specificity, relative to the other models, the Random Forest was chosen for further analysis and tuning.

In order to address the imbalance of positive and negative cases in the dependent variable, an attempt was made to improve upon the Random Forest by boosting the model. It was

thought that, as a result of the weights applied by a boosted model during training, the model would experience a decrease in bias and capture a greater proportion of positive cases of the dependent variable. Below are the performance results of the XGBoost model, a particular type of boosted Random Forest, after parameter tuning, on the test set.

Model\Metric	Accuracy	Sensitivity	Specificity
XGBoost	0.8935	0.9751	0.2741

Based on the above results, it was concluded that the XGBoost model does not perform any better than the base Random Forest and, in fact, performs worse in terms of specificity capturing ~20% less of the positive test cases than the original Random Forest model.

Throughout the duration of this analysis, Random Forest models were created using the base R 'randomForest()' function. During training of any single decision tree in the forest, the algorithm samples data points from the train set pool. In this way, the model creates many trees using many, different samples of the data set and, in effect, simulating the process of cross-validation. It is for this reason that all random forests were fit without the use of cross-validation. Although, the same train-test split is maintained to ensure equivalent comparisons are made between successive Random Forest models.

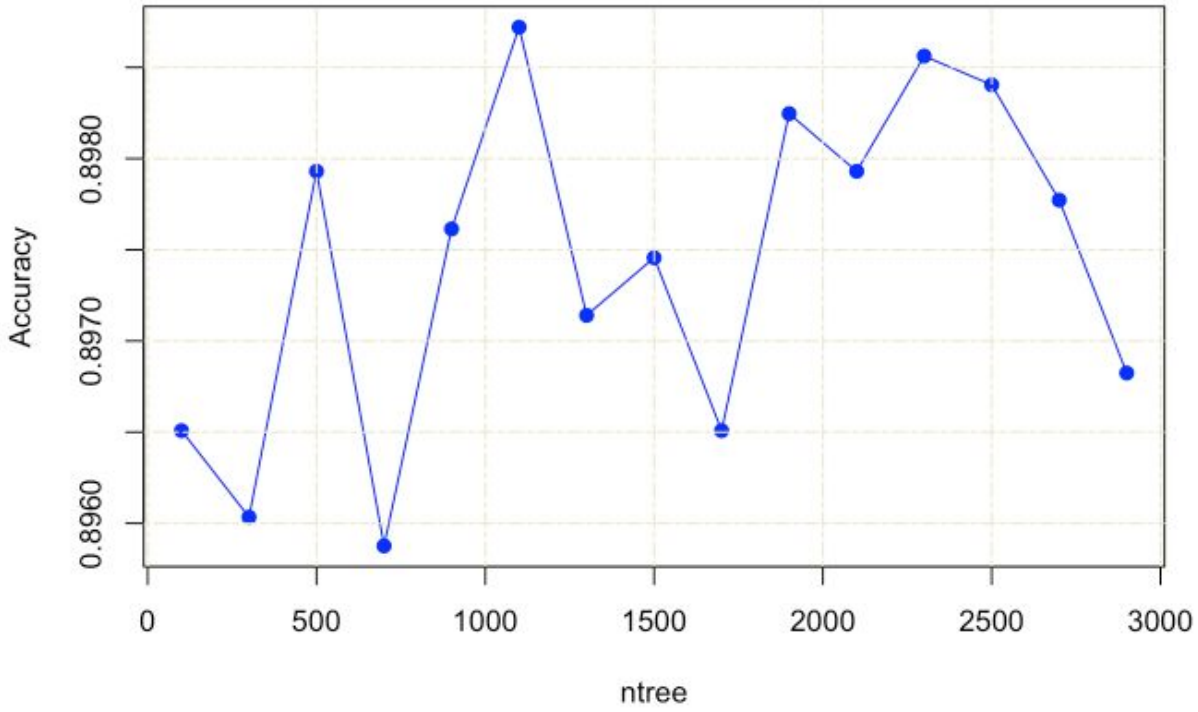
In an attempt to further improve the base Random Forest, additional economic indicators were added to the original feature set. Economic indicators provide general prognoses of the economic status of a country. It was thought that macroeconomic influences may have an affect on whether or not an individual decides to purchase a subscription with our Portuguese bank. These additional features include net disposable income, euribor three month rate, consumer

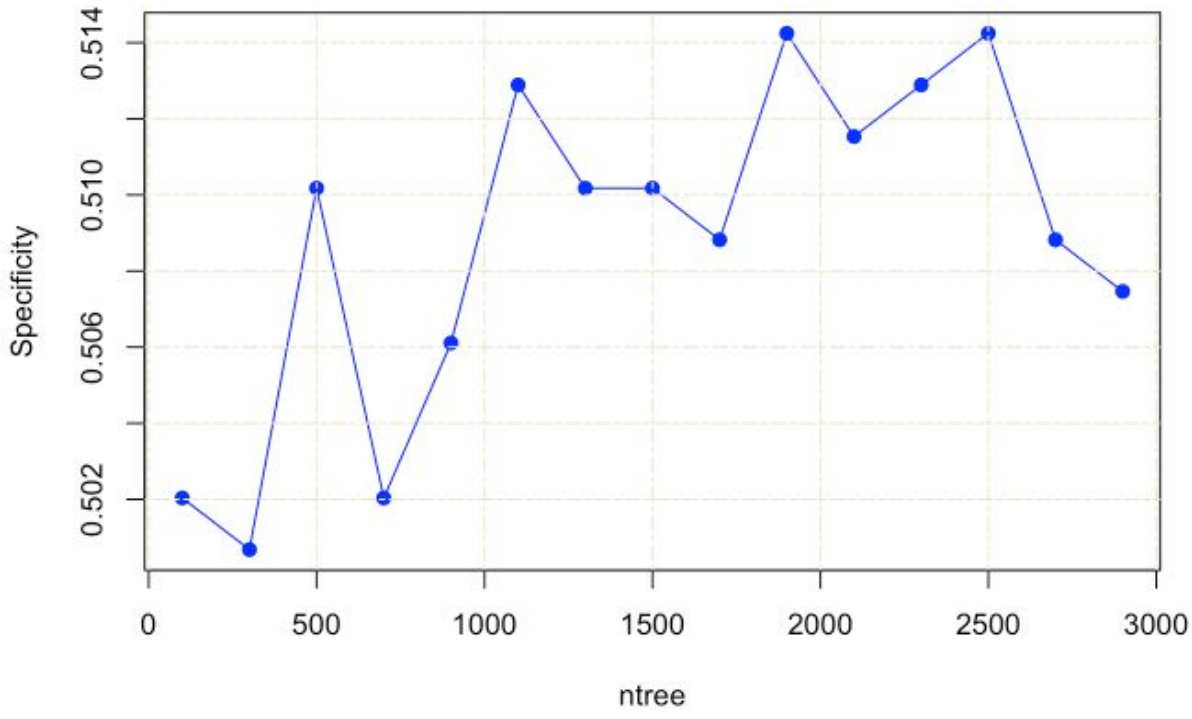
spending, loan growth, consumer confidence, inflation rates, employment variation rate, employment rate, and the unemployment rate all specific to the country of Portugal and the time frame, May 2008 to November 2010, in which each observation in the dataset was recorded. All economic data was gathered from https://tradingeconomics.com and conjoined with the original data in Microsoft Excel.

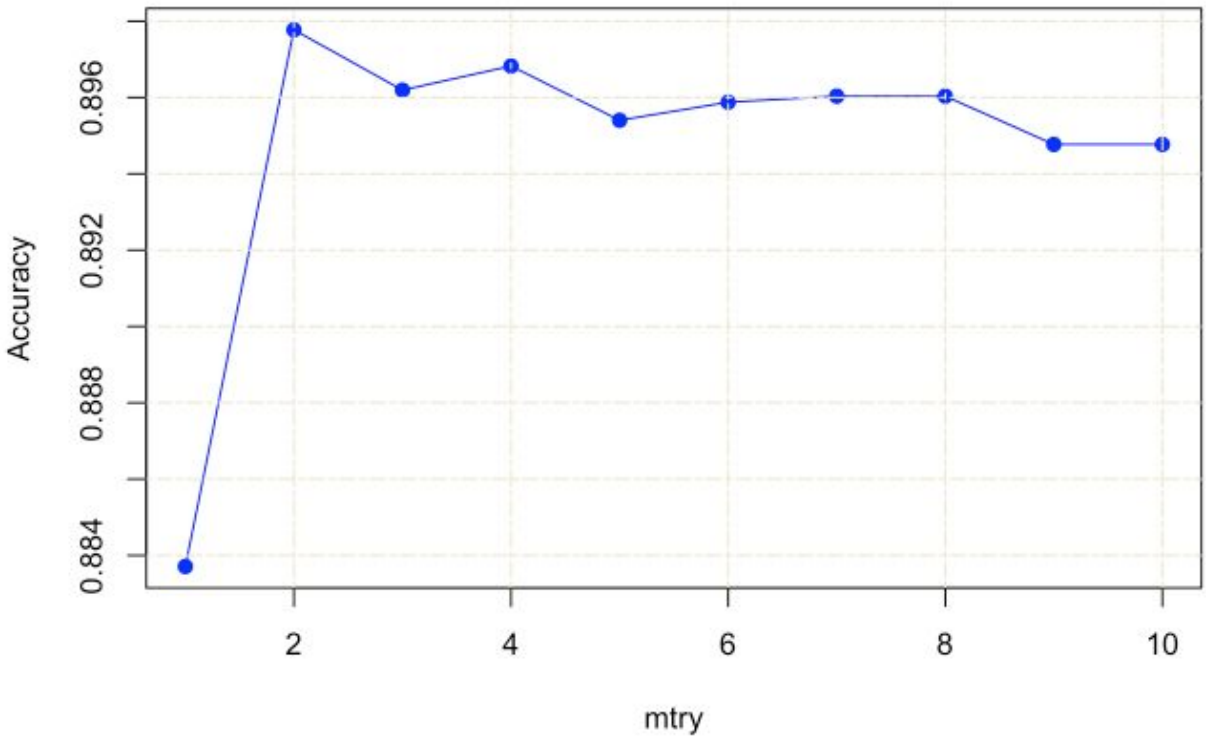
Creation of the model continued using two different datasets; a training set with the original features and a dataset including the new economic features. To remain true to the scenario, it was decided that no economic features would be appended to the final test set provided to us. This forces us to utilize only those features present in the test set for training of our final model. As a result, the final test set predictions would be made with a model trained on only the original set of features. A second model would still be built and evaluated using the new features. Although the new features will not be utilized within the final model, comparison of the two models would be able to highlight the differences in performance that the addition of the new features would have on the Random Forest Model. No additional feature transformations were done on the original features of the dataset.

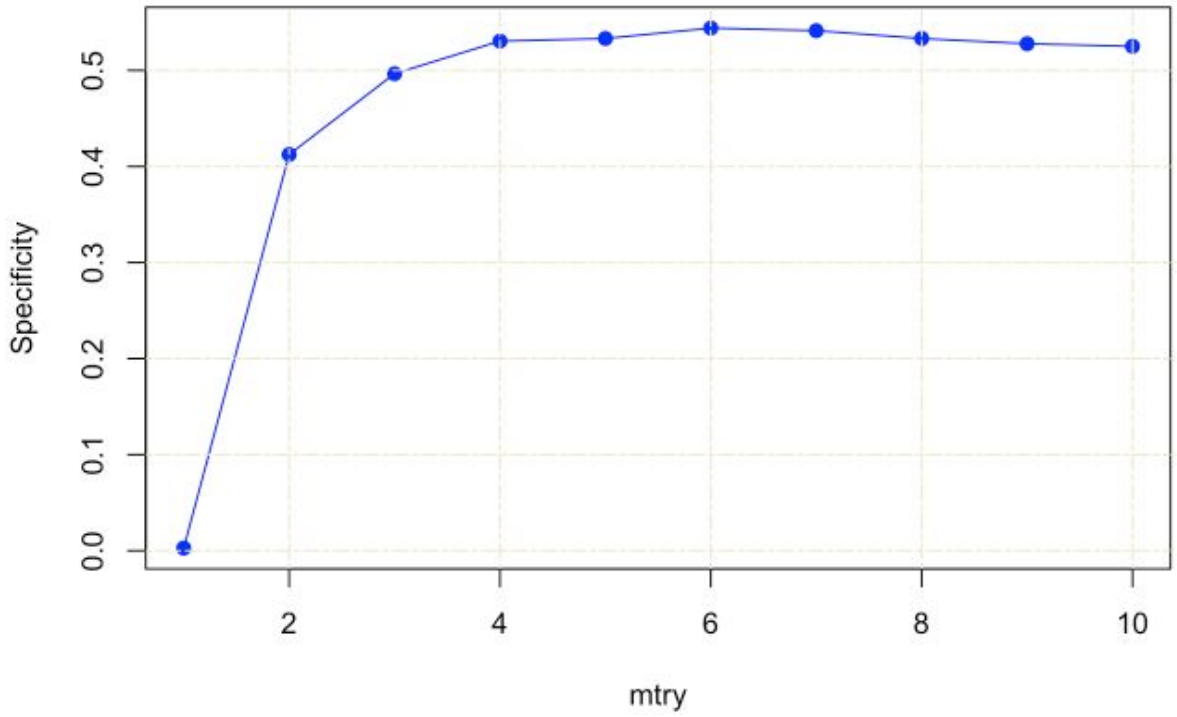
After the Random Forests were fit with both original and new features, model improvement continued with the tuning of the model parameters. Multiple random forests were fit and evaluated using varying values of a single parameter while all other parameters were held constant at their default values. The model's accuracy on the test set was then recorded. Finally, plots of the model's accuracy and specificity, on the test set, as well as the varying parameter values were made. The parameter value that maximized both the accuracy and specificity was selected as the ideal value for the parameter. The parameters selected for tuning in the Random

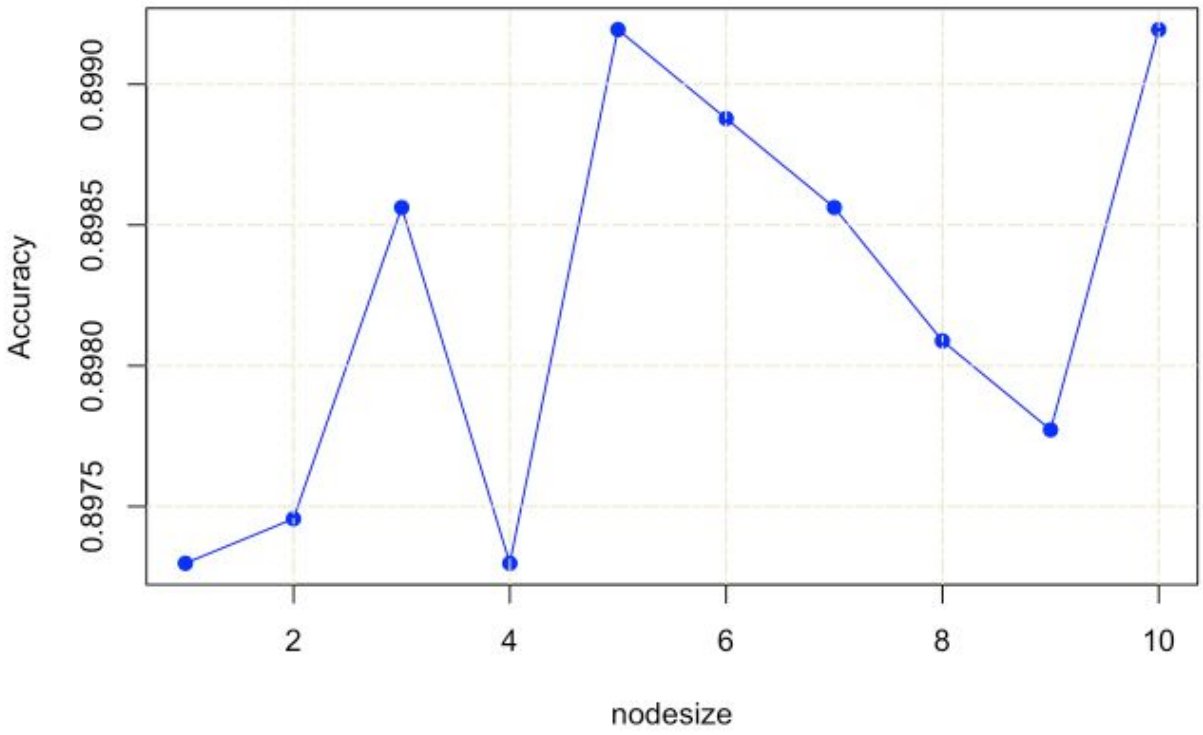
Forest model were ‘ntrees’, ‘mtry’, ‘nodesize’, and ‘cutoff’. The following are parameter optimization graphs for the model trained on original features only.

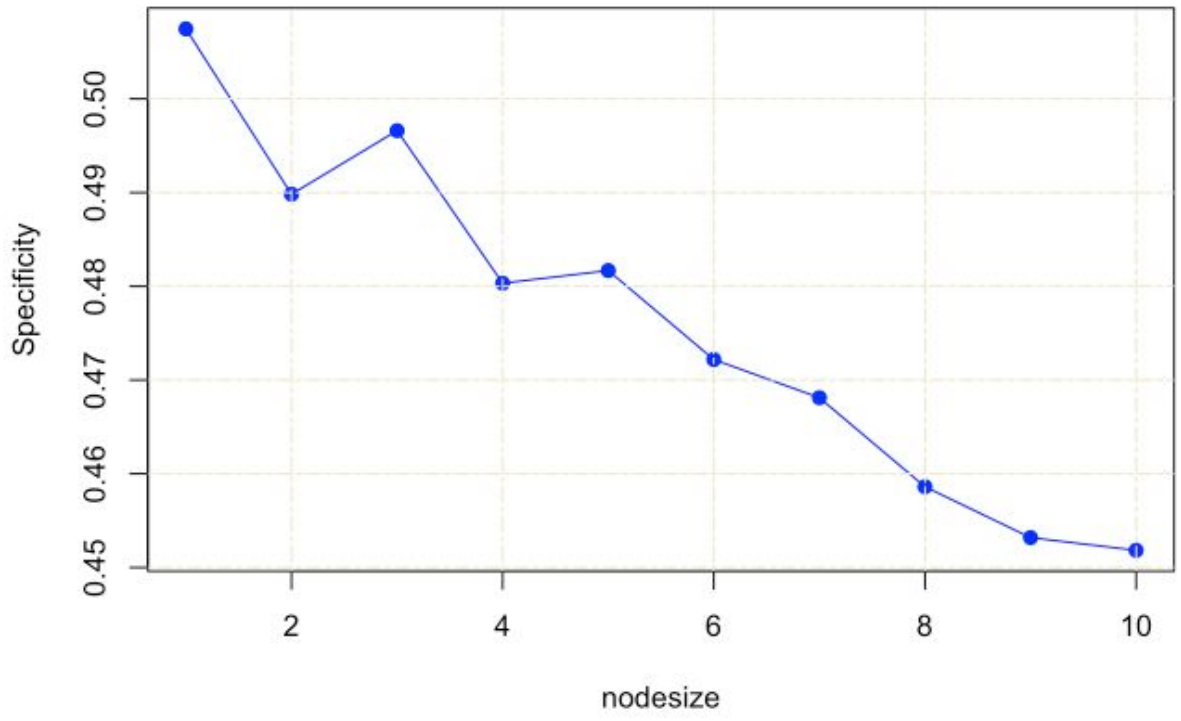


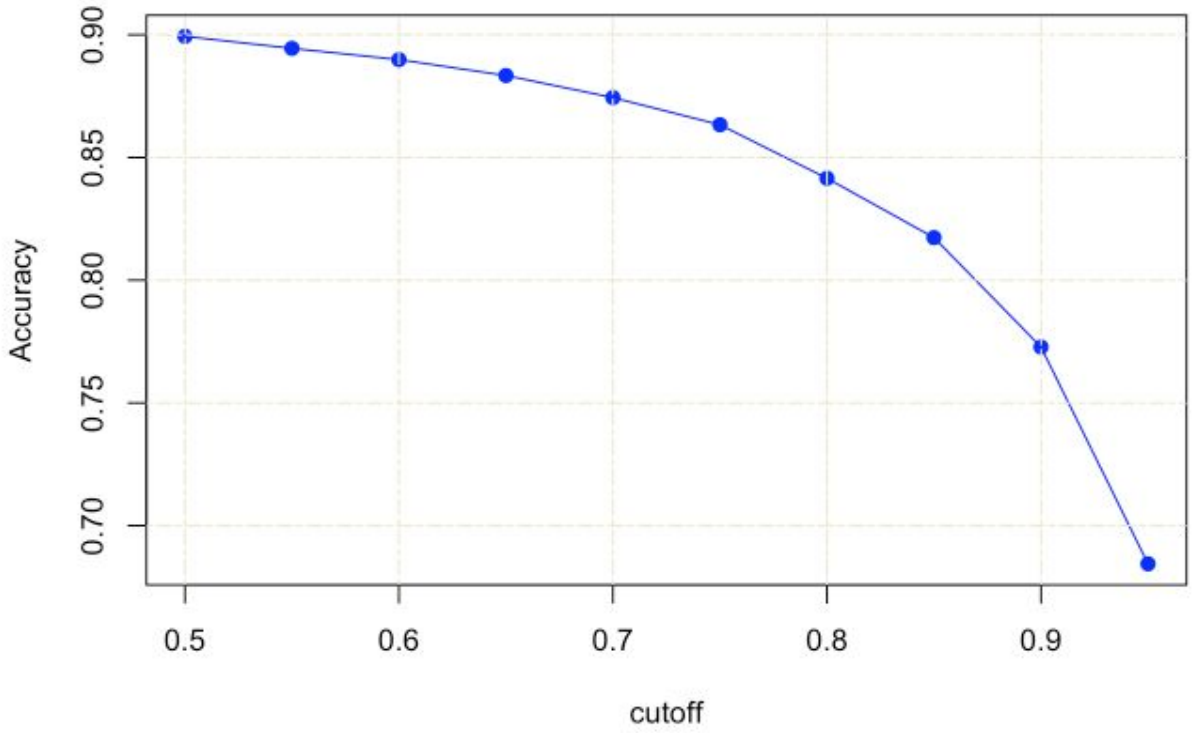


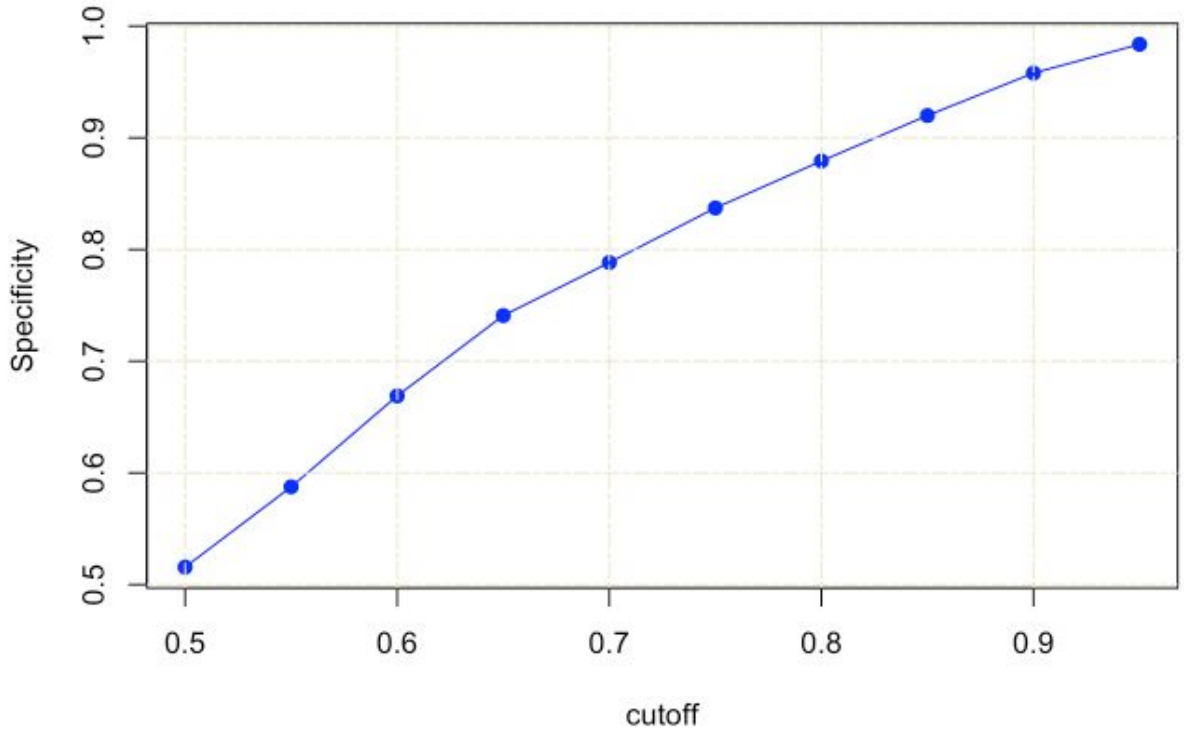

Accuracy
0.884
0.888
0.892
0.896
2
4
6
8
10
mtry


Specificity
0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10
mtry


Accuracy
0.8975
0.8980
0.8985
0.8990
2
4
6
8
10
nodesize


Specificity
0.45
0.46
0.47
0.48
0.49
0.50
2
4
6
8
10
nodesize


Accuracy
0.70
0.75
0.80
0.85
0.90
0.5
0.6
0.7
0.8
0.9
cutoff


Specificity
0.5
0.6
0.7
0.8
0.9
1.0
0.5
0.6
0.7
0.8
0.9
cutoff

Similar optimization charts were made for the model trained on the new economic features.

Below is a chart summarizing the optimal values for each parameter and each model.

Model\Optimized Parameter	ntrees	mtry	nodesize	cutoff
Random Forest (Original Features)	2100	4	1	(0.6,0.4)
Random Forest (New Features)	700	6	1	(0.6,0.4)

The final Random Forest models were refit using the optimized parameters and their performance evaluated. Below is a summary of the models with and without new features as well as with and without optimized parameters. Note, 'RF' is the abbreviation for 'Random Forest'.

Model\Metric	OOB Error Rate	Accuracy	Sensitivity	Specificity
RF (Original Features; Not Optimized)	9.48%	0.8964	0.9485	0.5007
RF (New Features; Not Optimized)	10.03%	0.9044	0.9515	0.5473
RF (Original Features; Optimized)	10.29%	0.8896	0.9174	0.6784
RF (New Features; Optimized)	11.12%	0.8953	0.9150	0.7460

Based on the results above, the model with the the additional economic features achieved a 5% greater specificity and ~1% greater accuracy than the model with only the original variables. Unfortunately, tuning of the parameters does not have seemed to increase the accuracy of the models drastically although specificity has increased by 17-20% for both models. The model highlighted in red is the model that was used to produce the final test set predictions.

Conclusion/Discussions:

After the analysis of the data with different models, we came to the conclusion that the best predictive model is the Random Forest. Below are the variable importances for both models with the original (left) and the new (right) features.

```
##           MeanDecreaseGini
## age             426.597391
## job             356.083777
## marital          96.357650
## education       123.639489
## default           9.106537
## balance         441.185966
## housing         116.316587
## loan             40.282468
## contact         100.504144
## day             872.522448
## month           650.579481
## duration       1420.207423
## campaign        235.558298
## previous        285.717061
```

.

```
##                          MeanDecreaseGini
## age                             70.4965053
## job                             86.8887197
## marital                         19.9489891
## education                       28.2678913
## default                          0.8611439
## balance                         79.2303054
## housing                         16.1235653
## loan                             7.0068761
## contact                         11.5051037
## day                            198.2215533
## month                           47.6837266
## duration                       331.7490115
## campaign                        49.9374555
## previous                        39.9971212
## net.disposable.income            8.1294671
## euribor3m                       77.8347091
## consumer.spending                9.5586725
## loan.growth                     64.5713316
## consumer.confidence             27.4475114
## nr.employed                     54.7133588
## unemployment.rate               27.8590140
## inflation.rate                  16.8823865
## emp.var.rate                    18.2707556
```

According to the Mean Decrease in Gini values, the top three most important features are ‘duration’, ‘age’, and ‘job’ for the original features and ‘duration’, ‘day’, and ‘balance’ for the new features. As far as limitations are concerned, missing information in the form of unknown categories proved to be a hindrance to model performance. In addition, tuning the parameters for the Random Forests proved costly in terms of run time and computational load forcing parameters to be tuned manual resulting in less precise estimates for optimal parameter values. Above all, the lack of positive cases in the dependent variable introduced a large amount of bias to our model. Further future analysis should be conducted that includes the collection of a greater variety of inferential features on clients, more data points, especially those in which the client is observed to have purchased a subscription, and research into more aggressive marketing tactics that utilize the insights gained from this analysis.

Work Cited

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014