**Term Deposit Telemarketing!**

Elijah Sartin, esartin@bellarmine.edu

Richard Osborn, rosborn@bellarmine.edu

**ABSTRACT**

Banks around the world make money by lending out their customers’ money. Term deposit accounts contribute significantly to the pool of money that they can lend out and therefore are crucial to the success of a bank. Telemarketing has become one of the major avenues that banks use to get customers signed up for term deposit accounts. Figuring out what leads to a successful term deposit telemarketing campaign helps the banks to continue making money. These key factors can be examined to determine the success of future telemarketing campaigns and give insight into what kinds of groups should be targeted in the future. These a range of factors include age, occupation, marital status, education level, and certain economic indicators, among other things. Through an exploratory data analysis and the use of machine learning methods, we were able to come to various conclusions about the contributing factors to success in bank telemarketing campaigns.

1. **INTRODUCTION**

Our data set consists of data collected through a Portuguese banking institution. The data was collected to determine the success of a telemarketer campaign that was *trying* to convince clients to subscribe with a term deposit. We will be trying to test whether the client has subscribed to the term deposit with a variety of other factors, such as age and the consumer price index. We will be working with the bank-additional.csv file which contains 10% of the random samples and will provide enough information to draw conclusions about the population parameters.

There are some factors that we have chosen to exclude from our data analysis since they did not provide relevant information. For example, we removed duration because it is a remarkably high predictor that the longer your duration in the campaign, the more likely you are to make a term deposit.

1. **BACKGROUND**

The goal of the data set was to determine the success of a telemarketing campaign. There are three separate classes of attributes. The first is the bank client data which includes the clients' demographics as well as the type of loans they already have and if they have defaulted on any previous credit. The second class is campaign-related attributes. This includes the methods in which the clients were contacted and the amount of overall time that a client was campaigned towards. The final class is social and economic attributes. This includes statistics such as the consumer price index and the consumer confidence impact. The consumer price index compares the average price of a typical basket of goods and the Euribor 3 Month Rate is the interest rate that certain banks loan at. All these classes were sampled from the Portuguese bank clients to determine if they would put down a term deposit.

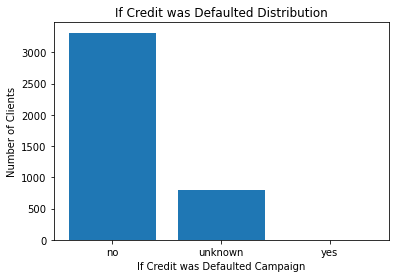
1. **EXPLORATORY ANALYSIS**

This data set contains 4119 samples with 21 columns of numerical and categorical data types.

**Table 1: Data Types**

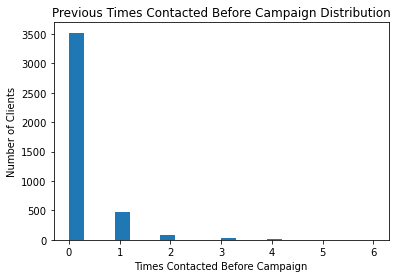
|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Age | Ratio/int64 |
| Job | Nominal/object |
| Marital | Nominal/object |
| Education | Nominal/object |
| Default | Ordinal/object |
| Housing | Ordinal/object |
| Loan | Ordinal/object |
| Contact | Nominal/object |
| Month | Interval/object |
| Day of Week | Interval/object |
| Duration | Ratio/int64 |
| Campaign | Ratio/int64 |
| Previous Days | Ratio/int64 |
| Previous Campaigns | Ratio/int64 |
| Previous Outcome | Ordinal/object |
| Employment Variation Rate | Ratio/float64 |
| Consumer Price Index | Ratio/float64 |
| Consumer Confidence Index | Ratio/float64 |
| Euribor 3 Month Rate | Ratio/float64 |
| Number of Employees | Ratio/int64 |
| Y (Output) | Ordinal/object |

The default distribution is interesting since there is a substantial proportion reporting that they did not default on their credit. However, it was difficult to show but there was one single person who did default. There could be an issue with people not wanting to admit that they defaulted over the phone to a telemarketer or that they did not know. Since it was so skewed, we decided to drop this from our data set.



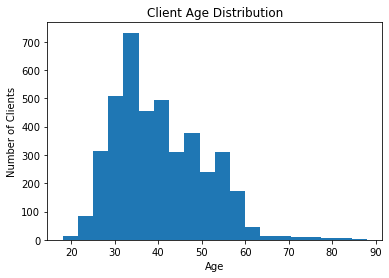
**Figure 1: Credit Distribution**

The previous number of times contacted before the campaign was 0 for most contacts. This can be good since there is a new pool of people to choose from. If people are being contacted again but didn’t subscribe to a term deposit the first time, then they will not subscribe for a term deposit.



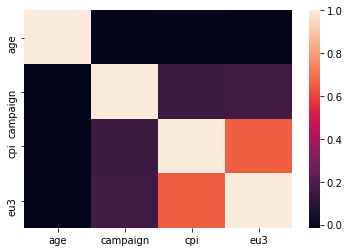
**Figure 2: Previous Contacts Distribution**

Age is normally distributed, but there are a few outliers that are extremely elderly. Most of the banks’ customers are between the ages of 30-40.



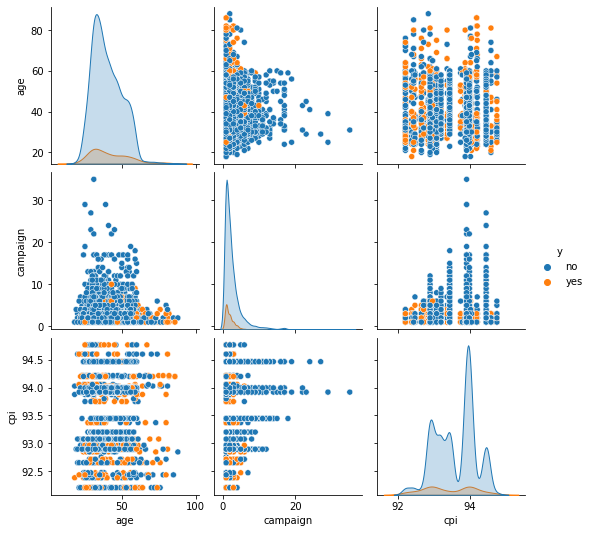
**Figure 3: Age**

We decided to use these variables due to the low correlation between them and variables with high correlation often not improving training.



**Figure 4: Heatmap**

We noticed through the pair plot that age tends to sway the outcome of a customer signing up for a term deposit account.



**Figure 5: Pairs plot comparing age, campaign, and the consumer price index**

1. **METHODS**
   1. *Data Preparation*

We started by importing all the necessary packages, such as Seaborn, Pandas, NumPy, and other packages for our data. Then we decided to drop the columns number of employees, the consumer confidence index, and the employment variation rate. We decided to drop these immediately because they were not what we wanted to test for the output. We wanted more demographic measurements and data that we were more confident with. We also decided to drop days of the week since it was almost uniform among all the days of the week, with Friday having slightly less. The parameters we found best suited the accuracy score were using age, campaign, consumer price index, and the Euribor 3 Month Rate. The goal was to test different parameters with each other until we got a sufficient accuracy score. There was not a need to normalize our data since the large sample size overshadowed the few outliers. Outliers were not a significant problem for the testing that we did. There was significant testing with bin widths to see what different shapes the graphs would make the most useful.

* 1. *Experimental Design*

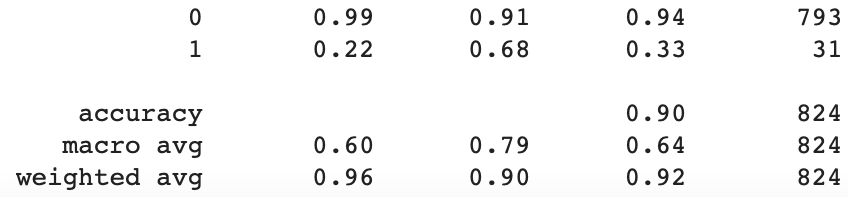
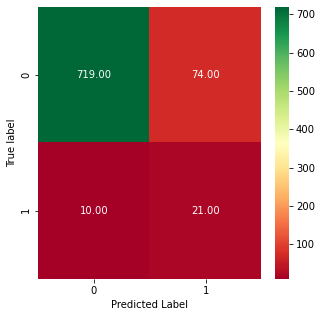
Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All four (4) raw features with 80/20 split for train and test- Decision Tree |
| 2 | All four (4) raw features with 70/30 split for train and test- Decision Tree |
| 3 | All four (4) raw features with 80/20 split for train and test- Random Forest |
| 4 | All four (4) raw features with 70/30 split for train and test- Random Forest |

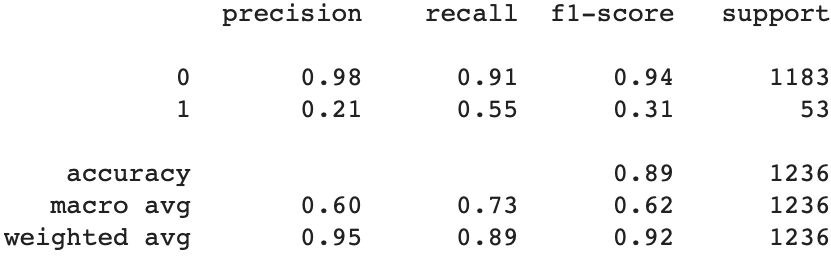
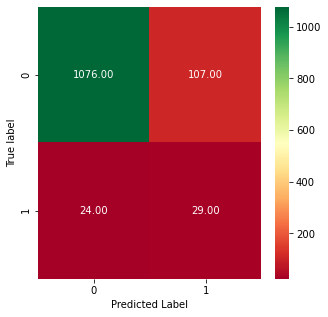
* 1. *Tools Used*

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, and SKLearn 0.18.1.

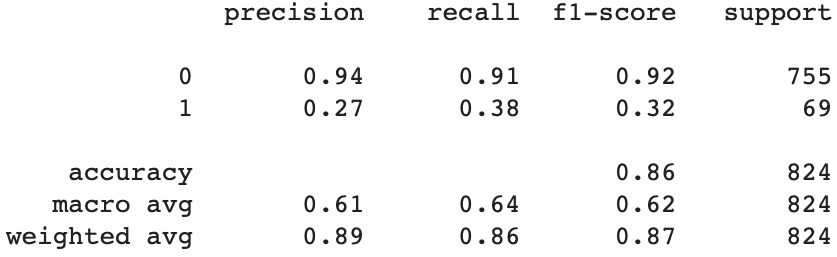
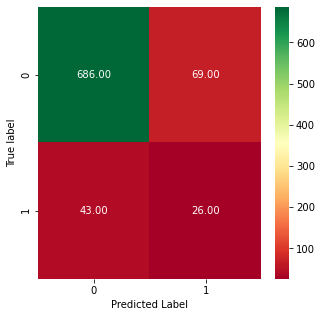
1. **RESULTS**
   1. *Classification Measures*



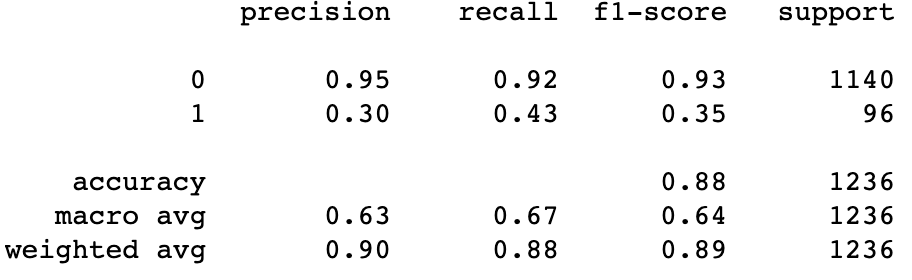
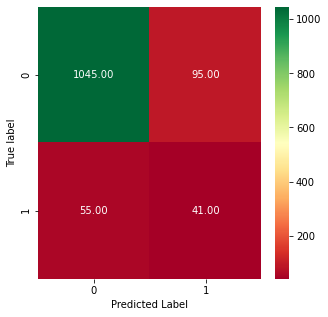
**Figure 6: 80/20 Split Using Decision Tree**



**Figure 7: 70/30 Split Using Decision Tree**



**Figure 8: 80/20 Split Using Random Forest**



**Figure 9: 70/30 Split Using Random Forest**

* 1. *Discussion of Results*

The main determiner of what a good model was for us was the accuracy score. The 80/20 split using the Decision Tree machine learning gave us the highest accuracy and precision. The classification reports indicated that the decision tree was the better predictor. One reason for this is that there was increased support for the random forest models, thus changing the results slightly. We still felt that the precision and accuracy of the decision tree was better.

* 1. *Comparison of Models*

Overall, we decided that the decision tree model was better than the random forest model due to the accuracy results.

* 1. *Problems Encountered*

The first problem that we encountered was trying to get a firm understanding of the data set. We were not overly familiar with bank telemarketing prior to this assignment. During the exploratory data analysis, we had difficulty finding too many errors, as the data the bank collected was clean already. There were a decent number of unknowns throughout the data that did cause slight woes when it came to prediction.

* 1. *Limitations of Implementation*

This might not be the best predictor for our data since there is low support for the number of positives so there was not a whole lot to predict if they subscribed to a term deposit or not. There could have been better systems, but these worked for our data set.

Discuss the limitations of your model(s). Are there reasons your models might not be the best way to predict the target data? What other models might work better?

* 1. *Improvements/Future Work*

For future improvements, we could have tried different machine learning models such as Naïve Bayes to see if we got different results. We would have also tried to use other variables, but we did test a few different combinations. We would like to use the same tests on a different data set to see how it compares.

What would you like to do to improve your model in future work? Some items you might consider discussing are performing more experiments, using different models, adding or removing variables, finding a different data set, etc.

1. **CONCLUSION**

Overall, there were a lot of problems that were overcome but this led us to useful conclusions about the data set. The problems were tough to deal with because we could not get certain parts of the model to work. However, we dealt with it and felt that this was a decent model for our data since we got a maximum accuracy of .9. There is potential for other models to work better without data, but we feel that this is an accurate model, and that the decision tree model gave us slightly better results with smaller supports. I would say that once we got past the initial challenges of coding, machine learning was smooth with problems that were easier to fix, Overall, it was a successful analysis, and applying machine learning to a data set was successful in showing the potential of machine learning.

**REFERENCES**

<https://realpython.com/logistic-regression-python/>

<https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>

<https://github.com/r2klabs/MachineLearningExamples>

*Division of Labor*

* We were both together during the entirety of the time we worked on it and bounced ideas off each other. Below are some of the things we worked on more specifically.
* Richard
  + Bar Graphs
  + Markdowns
  + Placing/Citing Figures
  + Conclusion
* Elijah
  + Abstract
  + Random Forest
  + Decision Tree
  + Pair plot