Overview

In this paper I propose using the Data Science approach to predict the probability that bank clientele will open new term deposit accounts (certificate of deposit accounts). Using the data collected from a Portuguese retail bank between the years 2008 and 2010, we analyzed 20 different variable features in predicting successful outcomes. Through random forest classifier (AUC = 0.783) and logistic regression (AUC = 0.758) models I was able to reach about 90% accuracy for both models. The results of this study can be helpful for bank managers in determining how to market new products efficiently increasing successful output results. Further research in improving the models would involve gathering more variables, such as, marketing agent effectiveness and specific client background information.

Introduction Project Proposal

The success of banks is reliant on the support of consumers participating in their products in order to earn a surplus profit. Consumers opening a new term deposit account, or a certificate of deposit in the United States, allow banks to hold funds for a specific amount of time without the possibility of consumers unexpectedly withdrawing their money. Without the risk of consumers withdrawing, banks can either invest these funds in a high rate of return investments or loan to other borrowers collecting the net interest margins. However, not all consumers are educated on investing their money and need to be sold on opening a new account. Banks use various marketing strategies varying between telemarketing and in branch advertisements in order to inform consumers term deposits that they are offering.

The issue with blindly calling consumers is it can be taxiing on bank resources without the guarantee of success. It would be more cost effective if we could maximize our outcome by predicting the probability of individuals opening accounts based on variables collected from previous term deposit campaign telemarketing attempts. Using classification machine learning models, we can determine how multiple variables affect the outcome of whether or not consumers opened new term deposit accounts.

There are many classification models available which can help determine the importance of variables in determining the results. Ideally using various models and then determining which model provides optimal results is the best method.

Data Collection/Wrangling

The data is gathered from a Portuguese bank, from May 2008 to November 2010, there are a total of 41,188 entries and 20 different variable inputs available. Of the entries available those that were successful in opening term accounts totaled 4,640 entries. Although successful entries

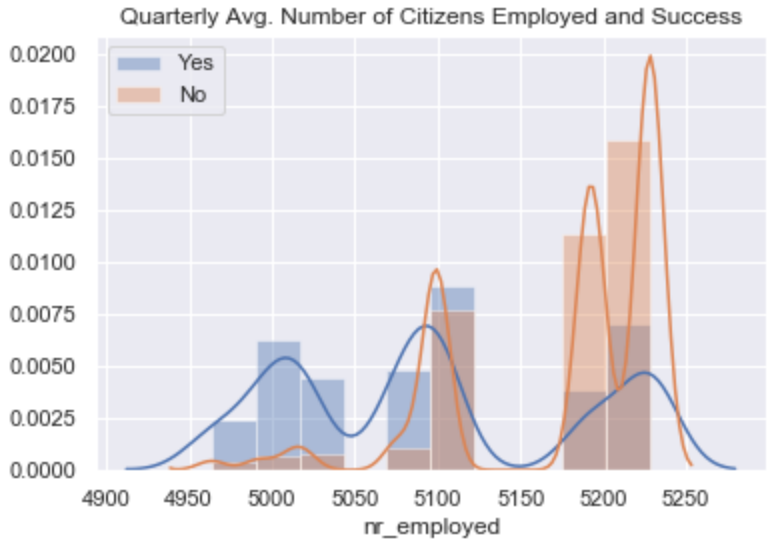
make up less than 12% of our data, it does not violate the independence condition and through machine learning we should still achieve reliable results.

None of our entries contain any null values. From the 20 available inputs, the variable of duration of call is eliminated because it is unknown prior to a call being placed and therefore is not valuable for our model. For the variables which were noted in text they must be encoded into integers in order to be used in our models. Lastly, some variables had to be renamed because they contained periods which will interfere with coding.

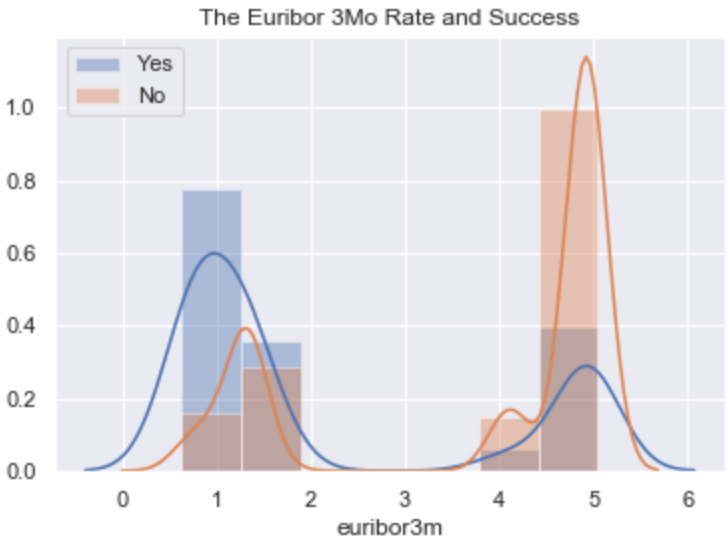
Exploratory Data Analysis

As previously mentioned, the number of successful calls make up less than 12% of our total data. Viewing each variable separately, it is possible to see some trends from the data.

The interesting trends from the exploratory data analysis are tables of the social-economic variables, the three month Euribor rate and the employment rate. It would make more sense for individuals to invest their money when the economy is strong, however, the plotted trends show opposite results.



When the employment number is high, the proportion of successful calls to failures is also the opposite of what is expected. When the economy is doing well there should be more individuals investing their money, however the graph shows that when employment is low more individuals are likely to open accounts.

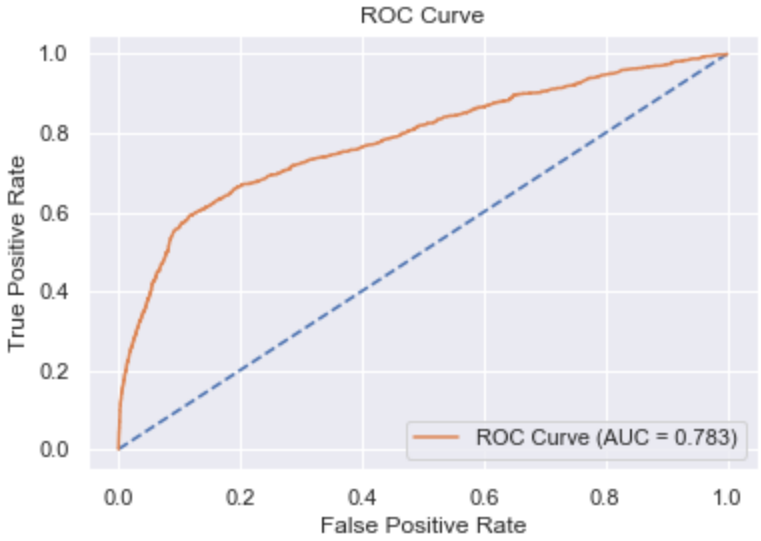


Likewise when viewing the table for the three month Euribor rate, the interest rates at which banks lend to one another. Individuals should want to invest when the market has high interest rates which can provide a higher returns, however, the data shows that when the market rate is below two percent more individuals opened accounts than when the market rate was high.

Experiment, Results and Analysis

The problem hinges on whether or not a new account is opened by a client. Therefore a classification model will be needed to predict the outcome given the independent variables. From the models available we opted to use a random forest classification and logistic regression models.

In order to find optimal hyperparameters for our random forest model we used GridSearchCV to find the best estimator values. Grid search provides results for each combination of the hyperparameters while cross validation ensures that all the data is represented within the training of the estimators. After completing the GridSearchCv the optimal parameters were n_estimators of 100 and max_depth of 3. Inserting the hyperparameters into our model we were able to achieve an 89.9% prediction accuracy on the test data.



The ROC curve plots the true positive rate against the false positive rate and the area under the curve, or the AUC, shows the models sensitivity differentiating between 0 and 1, or not successful and successful marketing calls. Our model has a 78.3% probability of determining correctly whether or not a call is a 0 or 1.

```
euribor3m          0.229407
nr_employed        0.198787
pdays              0.141971
previouscamp       0.132129
emp_var_rate       0.086354
cons_conf_idx      0.071493
cons_price_idx     0.042659
monthenc           0.023152
previous           0.022377
contact_method     0.014926
dtype: float64
```

Random forest model also ranks our variable features in order of influence on our model. Our model shows that the Euribor and the employment rate are two of the most important features.

Suggestions/Conclusion/Continuing Studies

The results of our models provided significant results in determining the probability of whether or not clients will open new term deposit accounts. Bank managers will be able to utilize this information in order to better cater to individuals more likely to open accounts. Having a more precise list of clients to contact will help market to our target clientele instead of trial and error in selecting who to contact. The model can also help sort individuals who physically visit a bank

location, if the system points out that a client has a high probability of opening an account then time should be spent marketing new products towards them.

The possibilities are endless in researching actions of clients. Multiple variables regarding clients behaviors will help banks better predict future revenues based on the probability of how many individuals are likely to participate in products offered by the bank. Instead of just focusing on whether or not an individual will default on a loan or commit fraud, we are able to provide more information towards a customer's value. Knowing a customer's value helps banks know which clients generate present and future revenue, and therefore can allocate resources towards retaining and marketing for high value clients. The banking industry is competitive and keeping clients satisfied is key to retaining business. The best way to retain clients is providing services best tailored to each customer based on their profiles. Therefore, using machine learning models will help banks allocate resources in an optimized manner maximizing revenue gains.