Introduction

All businesses use data to determine marketing decisions. Financial institutions are no different and gather information on their clients and potential clients in order to sell them different products like term deposits. Term deposits are a fixed-term investment that includes depositing money into an account at a financial institution.[1] These fixed terms often range in maturity length from one month to multiple years. This offers the client a lower risk investment to save money while gaining a small amount of interest while the financial institution loans that investment to other clients seeking to take out loans. Longer fixed-term investments will generate more interest for both the bank and the client so it is in the interest of the bank to solicit more clients to subscribe to these investments and clients to deposit large amounts into these investments.

This project is designed to take anonymized banking data from a Portuguese bank and build a classification model to predict if a client is likely to subscribe to a term deposit based on a model built from client data. This project is aimed to develop and identify characteristics of previous and current clients that have subscribed to term deposits and use this information going forward to reach new and returning clients. Additionally, elucidation of some characteristics unable to be used to develop the model may still help gain insight into improved marketing results of future campaigns.

The dataset used for this analysis is a subset of an initial dataset gathered by Moro et. al. of 79354 client entries gathered between May 2008 to November 2010, ordered by date.[2,3] The original success rate was 8% (6499 clients accepting the subscription). In their data preparation phase, they rejected variables that were irrelevant (e.g. gender of client, banking agent making contact, etc.) and dropped client entries that contained missing values. This reduced the dataset to 45,211 client entries with a success rate of 11.7% (5,289 clients accepting the subscription, see figure 1).[4]

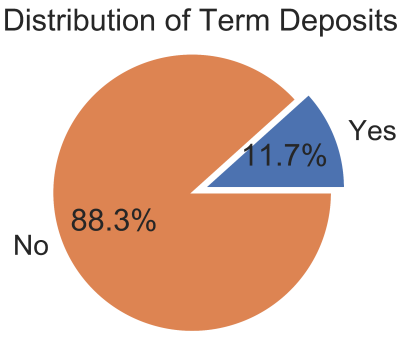

Figure 1: Distribution of Term Deposits for the Reduced Dataset

[1] https://www.investopedia.com/terms/t/termdeposit.asp
[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modeling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.
[3] A link to the dataset: https://archive.ics.uci.edu/ml/datasets/bank+marketing
[4] 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

The dataset contains 17 unique attributes broken down into four general categories: 1) bank client data, 2) data related to last contact of the current campaign, 3) other attributes related to previous campaigns, and 4) subscription rate, 'y', the desired target. Note that there is a second data set that contains extra feature columns to total 21 unique attributes. This data set was not used for this analysis as it contained the extra variables that were rejected in the data preparation by Moro et. al. in their analysis.

Bank client data includes common demographic data for past and present clients. These features include age, job type, marital status, and terminal education. Specific financial information about the client having various debts would also factor into the decision so having credit in default, a personal loan, or a housing loan was also reported for each client.

Data relating to the last contact of the current campaign includes the contact communication type, the date of the last contact broken down by month and day, and duration of the call (in s). Similarly, the number of times each client was contacted during this campaign, including the last contact, was also shared in the dataset.

It should be noted that the duration of the call can highly affect the output target (e.g. if the duration is zero, then the subscription rate will be 'no'). This duration information is unavailable as we do not know how long we will have the customers on the call before we make the call. Hence, we will need to disregard this category for the modeling step. To quote the source of this data[5]: "... the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model." While this information can't be directly used in the modeling process, it may be useful for marketing methods on future solicitations for these term deposits.

Data relating to previous campaigns includes the number of times each client was contacted before the current campaign, the outcome of the previous marketing campaign, and the number of days that passed since the client was last contacted. Some clients were not contacted before the current campaign and thus they may have filler data in these features. For example, the 'pdays' feature has entries with "-1" corresponding to clients not previously contacted and the 'poutcome' feature has entries with "unknown" that may also include these clients.

The outcome feature, 'y', asks if the client has subscribed to a term deposit. In this data set, we're looking to accurately predict this value for new clients based on the other features.

Data Wrangling

A few steps were taken during the data wrangling process to make the subsequent steps a little bit easier. One such change was to convert the 'y' outcome column from string choices "yes" or "no" into numeric values of "1" or "0", respectively. This new column was named 'y_bool'. This

[5] https://archive.ics.uci.edu/ml/datasets/bank+marketing

column contained the identical information as the column “y”, but it was able to be processed to make graph work-up easier.

For the feature columns, only the “age” and “pdays” columns were wrangled significantly. Age data was grouped into five year blocks between 20 and 59 years old (20-24, 25-29, etc…), with two bookend groups of 18-19 year olds and 60-95 year olds. This data was used primarily for graph formation, not modeling purposes. Other age range groupings appear in the Jupyter notebook in additional supplemental graphs.

The time since the client was last contacted in days (pdays) was changed into categorical data (pdays_cat). As listed in the original data set, a value of “-1” for “pdays” corresponds to the client not previously contacted. For the remaining clients, a whole number was given. These integer values were converted into four categories: no contact, contacted within the last three months, contacted within the last year, and more than a year since last contacted.

Exploratory Data Analysis

Using the data gleaned from the clients’ age, the subscription rate was determined for each age category range. As seen in Figure 2, the largest group of clients is the group of 30-34 year olds, yet they only have a 10.8% subscription rate within the group. This subscription rate is similar to that seen for a majority of the population. The average subscription rate for ages 25-59 is 10.6%. The largest groups of subscribers by percentage appear to be the two bookend ranges. The group under 20 years old has an almost 40% rate of subscriptions while the 60+ year old group has about ⅓ subscription rate. Each of these bookend groups could be different from the rest of the age groups as they have a much smaller sample size within their respective categories.

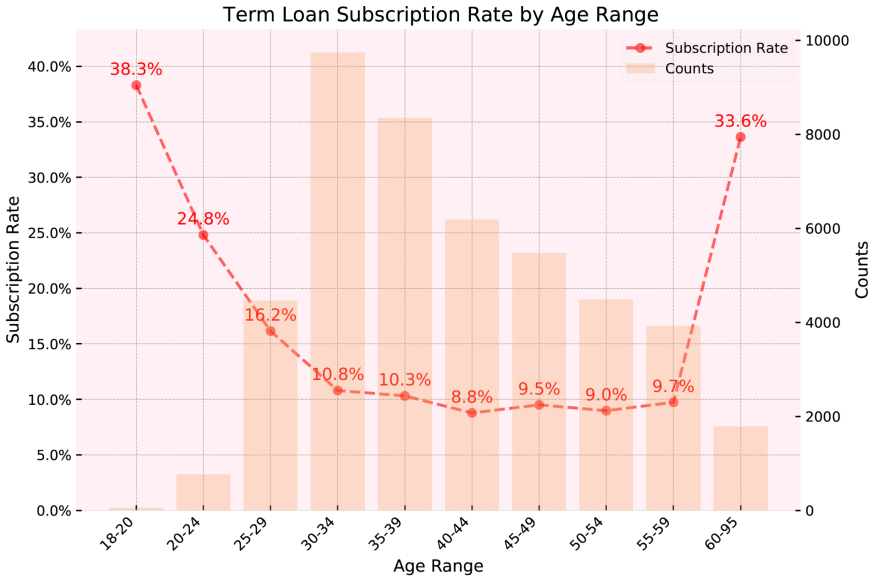

Figure 2: Term Loan Subscription Rate by Age Range

Residential population distribution of Portugal broken down by age gathered by Instituto Nacional de Estatística for the years 2008-2010 reports a different demographic breakdown to age compared to the client data.[6] The general population data was averaged between all three years and then reported as a percentage versus the total population, reported as the red squares in Figure 3. While not reported on this figure, the age group of 0-14 year olds were included in the total population and influenced the percent population values. As reported, the population of Portugal was evenly spread of each age range with each group roughly between 10% to 30% of the total population. By comparison, the ages of most clients within the banking data set were largely younger, as reported with blue circles. Half of all clients within the data set are within the age range of 15 to 24 years old and about 85% of clients are under 40 years old. This disparity very clearly shows that a majority of the client population is younger than 40 and that's the age range that should be strongly marketed towards for these subscriptions.

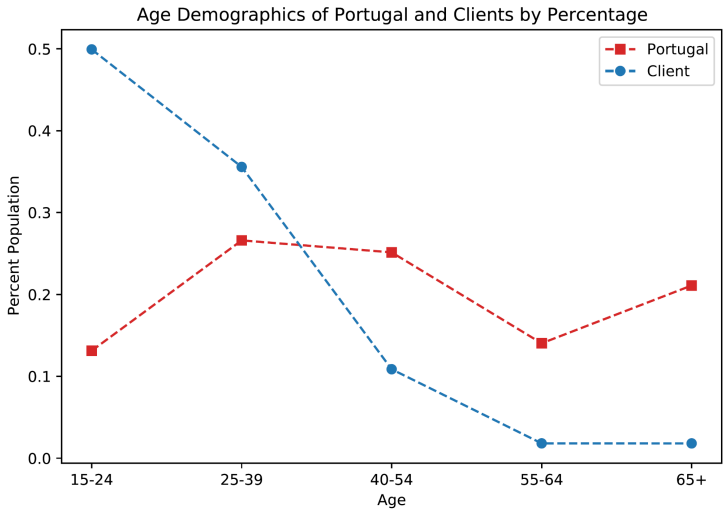

Figure 3: Percent Population of Portugal by Age Range (red squares) and Percent Population of Clients by Age Range (blue circles)

Another interesting feature that was discovered was the relationship between the last time a client was contacted (pdays_cat) and the subscription rate, seen in Figure 4 (left). As noted above, while most of the clients were not previously contacted in a past campaign, those that were contacted before the current campaign did have an increased rate of subscribing to the term deposit. This relationship saw roughly a double or more increase in subscriptions for all three time ranges: up to "3 months", up to "1 year", and "more than 1 year" since last contacted. While the specific subscription rates for these previously contacted clients are most likely not robust due to smaller samples compared to the clients not previously contacted, there seems to be evidence to support following up with previously contacted clients independent of their current subscription status. Figure 4 (right) also depicts that contacting a client previously at least once roughly doubles the subscription rate from 9.1% to 20.8%; however, contacting more

[6] Instituto Nacional de Estatística, IP, ESTATÍSTICAS DEMOGRÁFICAS 2010, pp. 35, Lisboa, Portugal, 2012. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=102686059&PUBLICACOESmodo=2

than three times doesn't increase the chances for a successful subscription much more than that.

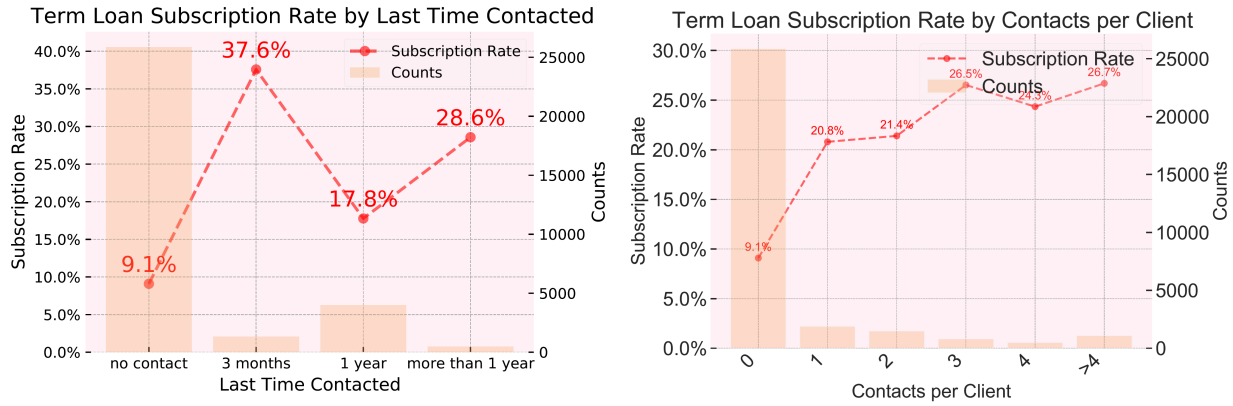

Figure 4: Term Loan Subscription Rate by Last Time Contacted (Left) and by Contacts per Client from Previous Campaigns (Right)

After looking at the specific data distribution of the duration column with the subscription rate, a general pattern emerges. As shown in Figure 5, clients who interacted with the sales representative for longer lengths of time were more likely to subscribe to the term limit. Specifically, the percent success for subscribing rises as time on the call continues to increase to 60% success around the time point 16 minutes (960 seconds). After this point, the success rate mostly stays constant. However, most of the clients have contact durations closer to 5 to 6 minutes (336 s) which has a much smaller success rate around 10% or lower. Trying to keep clients engaged in the call or by explaining details of the product may lead to increased subscription rate in the future. Alternatively, reaching out to clients that are actually interested in the products offered will also most likely increase duration as well and increase probability of successes.

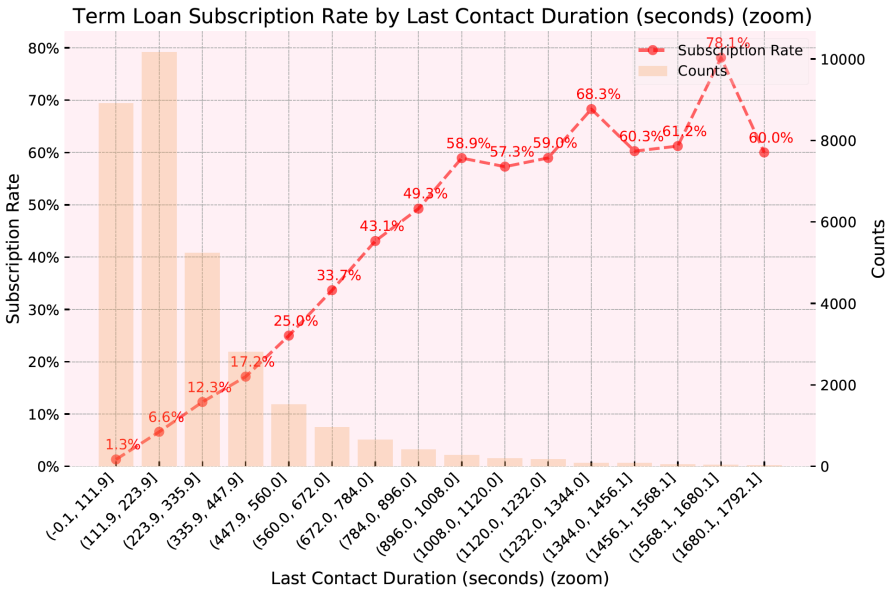

Figure 5: Term Deposit Subscriptions vs Duration of Client Call

Heat Map

The heat map of features was generated using 16 variables: 15 features and the outcome variable, the boolean version “y_bool”. The “duration” feature was omitted due to the discussion above in the EDA. The heat map was generated using a combination of methodologies to interact with the mixture of categorical data and numerical data. To generate the values in each square of the heat map, either Cramer’s V (for categorical vs categorical data), Pearson’s R (for numerical vs numerical data), or a correlation ratio (for categorical vs numerical data) was used. The functions used to determine the Cramer’s V and correlation ratio were adapted from an article by Shaked Zychlinski.[7] The Pearson’s R function was pulled from the scipy.stats library.

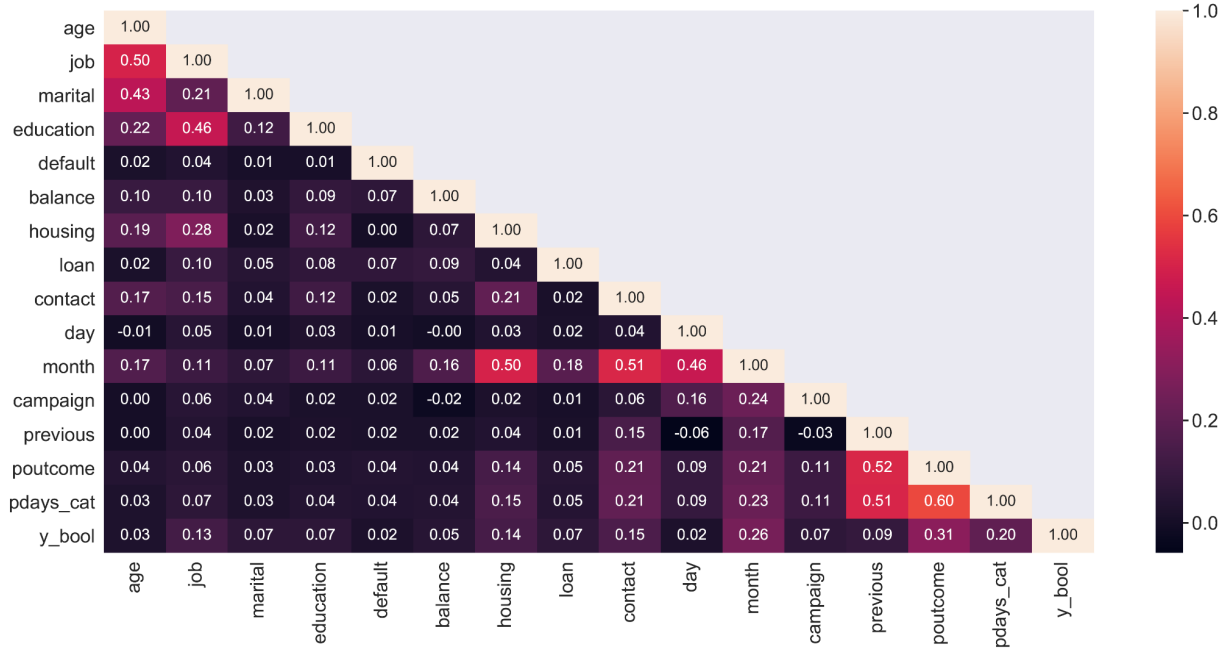

Figure 6: Heat Map of Features Generated from Cramer's V, Pearson’s R, and Correlation Ratio

Based on the heat map, there are a few features that should be noted for collinearity. Values that typically have correlation numbers higher than 0.7 or lower than -0.7 can be omitted as there is a risk for multicollinearity. In general, most of the heatmap in Figure 6 is not at a high risk of correlation and we can assume these features are mostly independent.

The client’s age has a larger correlation with the client’s job and marital status and a lesser correlation with the client’s education and status on a housing loan. These correlations generally are understandable as people within the age range most popular with the term deposit (25 - 39) will have similar demographic information based on education, jobs, and marital status. The correlation between the job and the age can also be seen in Figure 7, where the two highest

[7] https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

subscription rates by job are student (28.0%) and retired (22.4%). These jobs generally match with younger and older clients, respectively.

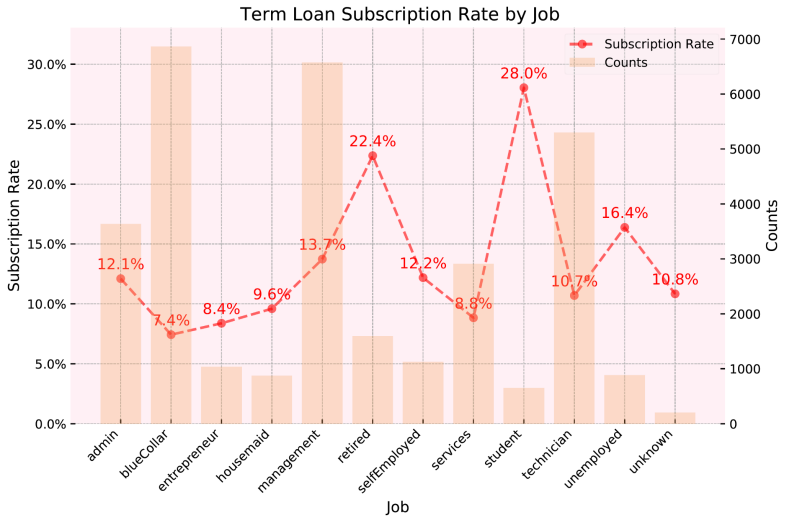

Figure 7: Term Deposit Subscriptions vs Job of Client

The client's education seems to be correlated with their job, which again would not be surprising. Interestingly, there are also correlations between the month of the call in the current campaign with the status of a housing loan, communication contact type, and the day of the week for the call in the current campaign. Generally, these correlations are large yet they are not expected to influence the model generation for this particular project.

Feature Importance

A standard methodology using the random forest classifier was selected to determine the feature importance for the data set. To help further clarify which features had an influence in the outcome, a "random" feature was added to each client entry. This random feature was a number between 0 and 1 and was to provide a noise baseline.

Based on the results of the Random Forest Classifier shown in Figure 8, 15 features were deemed to be more important than the baseline "random" feature. The top five features that have the most influence are: if the client previously subscribed to a term deposit in a previous campaign (poutcome_success), age of client, if the client does not have a housing loan, and if the number of days that passed by after the client was contacted from a previous campaign was 3 months or less.

Unfortunately, some of the features in this list are not helpful. One such example is not knowing which method a client was contacted (contact_unknown) having a larger importance. Similarly, a client both having and not having a housing loan is similarly important, which doesn't give any useful information. Lastly, the difference between the 9th through 15th important features are not

much more than the random baseline and probably don't have a strong influence on the final outcome.

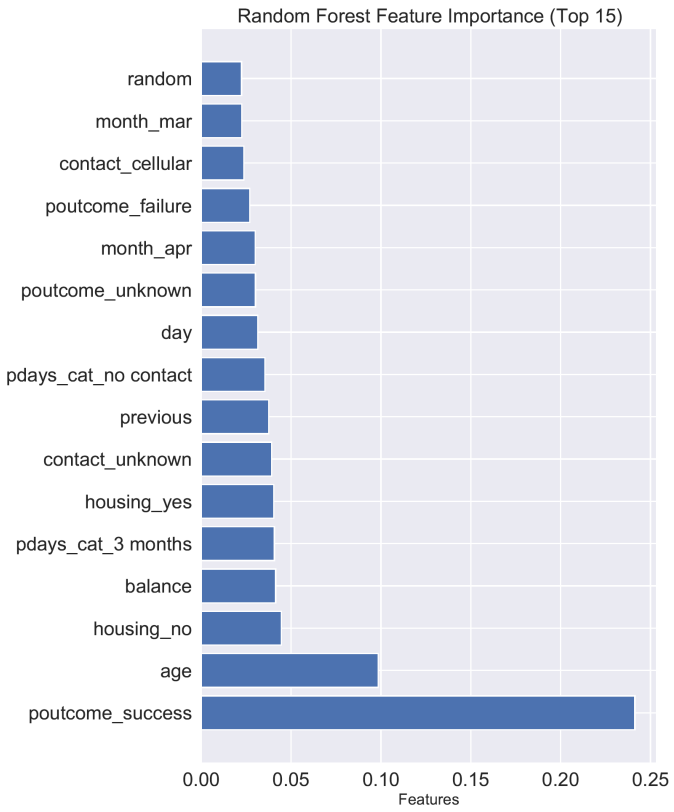

Figure 8: Top 15 Feature Importance of Data Set using Random Forest Classifier

Results

Model Selection - Choosing a Classifier

Three different classifier models were chosen for further analysis: Random Forest, Logistic Regression, and XGBoost. Each of these underwent a grid search to tune hyperparameters based on time allotted and previous data. Each set of hyper parameters were tested with cv=3 to boost replicability. The training of the models took about one hour with the Random Forest taking the longest and the Logistic Regression taking the shortest. The results of the grid search on the hyperparameters for the three models are listed in Table 1. From the results of the three classifier models, both the optimized Random Forest and XGBoost classifiers give roughly equal ROC-AUC scores of 0.797 and 0.804, respectively, and should be relatively equal in efficacy, however the Logistic Regression model was not far behind. For the remaining analysis, the XGBoost model was chosen.

Classifier	ROC-AUC Score	Best Hyperparameter Values
Random Forest	0.797	{'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 275}
Logistic Regression	0.782	{'C': 75.43, 'penalty': 'L2''}
XGBoost	0.804	{'objective': 'binary:logistic', 'learning_rate': 0.10, 'max_depth': 4}

Table 1: Performance for the Random Forest, Logistic Regression, and XGBoost Classifiers[8]
Note: ROC-AUC Scores were generated on test data.

Profit Curve

According to the world bank, Portugal's lending interest rate in 2019 was 6.12%.[9] Based on this lending interest rate, it can be assumed for this analysis the revenue from each term deposit could be roughly 5% of the deposit amount, with up to 1% being paid to the client. Other factors can lead to deprecation of the return including inflation (typically 2% each year) and defaults on loans.

An article by Schnabl et. al. on banking deposit risk states that the net interest margin[10] for a term deposit was insulated from interest rate changes between 1955-2015, staying at 2-3%, as those costs are passed on to clients.[11] They continue to say that up to 2% is used to pay for operation costs and the remaining up to 1% is the return on investment. Specifically, the World Bank reported the net interest margin for Portugal in 2009 to be 1.5715%.[12]

A prominent bank in Portugal, CGD, is listing a term deposit of one year with a minimum deposit amount of €10,000 and no maximum deposit amount.[13] For the profit analysis, a deposit amount of €10,000 for 1 year will be used in the calculation. The interest on term deposits and resulting loans were calculated by standard compound interest formulas compounded daily. Loans taken out using the capital provided by these term deposits generated the total revenue. Of this total revenue, 1% was assumed to go back to the client as payment for the term deposit and 1.57% was assumed to be net profit for the bank as indicated by the net interest margin, €158.36.

[8] Please see the Jupyter notebook for a full workup of the data.
[9] https://tradingeconomics.com/portugal/lending-interest-rate-percent-wb-data.html
[10] "In finance, net interest margin is a measure of the difference between interest paid and interest received, adjusted for the total amount of interest-generating assets held by the bank."
https://www.investopedia.com/ask/answers/061715/what-net-interest-margin-typical-bank.asp
[11] https://www.frbsf.org/economic-research/wp-content/uploads/sites/4/Session-1-Paper-2-Savov.pdf
[12] https://fred.stlouisfed.org/series/DDEI01PTA156NWDB
[13] https://www.cgd.pt/Particulares/Poupanca-Investimento/Depositos-a-Prazo-e-Poupanca/Pages/Deposito-Prazo-3-Anos.aspx

The cost for each call was calculated by estimating the cost per call center employee. An estimated value for the monthly salary for an employee is €3000. Breaking this down to 4 weeks of 5 workdays with 40 hours, we have a cost per hour of €18.75. Estimates for number of calls per hour can vary. Based on our client data, the average duration of a call lasts 4.29 minutes, but this does not include other tasks like finding the call and subsequent reporting of the call and other tasks part of the job. To be safe, the estimation on the amount of time used per call is 15 minutes. Therefore the cost per call is €4.69. However, there is additional time setting up, processing, and closing the term deposit so the time for each process will be considered to be one full hour of bank worker time that costs €18.75.

Using these numbers, the cost matrix was generated below in Table 2. The value for a true positive (TP) corresponds to the net profit after the cost of making the call is factored in (e.g. €158.36 - €18.75). The value for the false positive (FP) is the cost of the call (e.g. - €18.75). The cost values for both the false negative (FN) and true negative (TN) are zero because no resources were used for these categories. False negative classifications do represent a potential loss in future profits as these could be additional customers that could generate revenue, but this is not included within this discussion as determination of the exact monetary value assigned to this loss is outside the scope of this modeling discussion.

	Predicted Success	Predicted Failure
Actual Success	**TP** €139.61	**FN** €0
Actual Failure	**FP** -€18.75	**TN** €0

Table 2: Cost Matrix for the Profit Curves

The coding of the profit curve was largely adapted by the work of E. Saslow.[14] The choice of using each of the three trained, optimized models was made because no other metric was used besides ROC-AUC and a judgment couldn't be made without additional evidence. The results from the profit curves can be seen in Figure 9. Each model had a similar trend increasing steadily. The Logistic Regression model predicted the lowest maximum profit compared to the Random Forest and XGBoost models. However, the Random Forest model predicted a lower probability threshold where the net profit would be equal to zero, compared to the higher probability threshold predicted by the Logistic Regression and the XGBoost. Since the Random Forest and the Logistic Regression both mimic sections of the XGBoosted model and the XGBoost model gives the highest maximum, this is the curve that will be considered in the discussion. The maximum profit predicted by the XGBoost profit curve was at a probability threshold value of 0.142. This corresponds with the maximum profit for the client data set would be boosting the subscription rate to 14.2%.

[14] https://github.com/Esaslow/profit_curves Their work is also highlighted on their website in a blog post: http://www.elliottsaslow.com/galvanize/profit-curves-in-python-for-turnover-rate-of-wireless-contracts

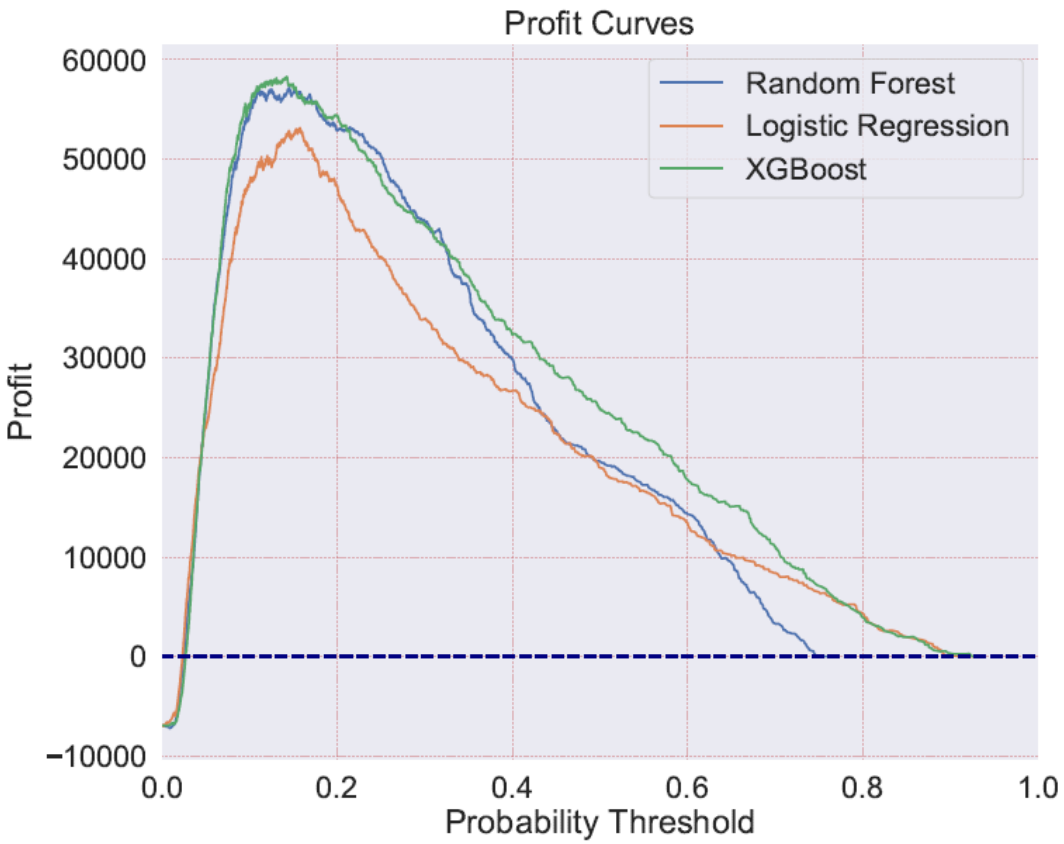

Figure 9: Profit Curve in Profit (Euros) vs Probability Threshold

Conclusions

Overall, each of the classification models used in this analysis were roughly equal in efficacy, with an ROC-AUC score of around 0.80. Based on this conclusion, each of these models would be able to reasonably predict the desired outcome of whether a client could subscribe to a term deposit. Using this data, a profit curve was generated looking at the goal percentage of subscriptions to maximize profits. The XGBoost optimized classification model performed the best out of the three models for the profitability curve. It was seen that the maximum probability threshold value predicted was at 14.2% successful subscriptions of the client population.

This data and the subsequent model posed an interesting question about trying to predict who would be likely to subscribe to a term deposit. However, this data was from over a decade ago at the time of writing. According to an article from Wharton Business School,[15] Portugal had been suffering from a prolonged economic recession that started in 2003 and worsened in 2008 from the global financial crisis caused by the US banking crisis in 2007-2008. To recover from this recession, Portugal had been subjected by the International Monetary Fund and European Union economic austerity measures that may have influenced the ability and attitude of the Portuguese population to make financial decisions, including subscribing to term deposits.

[15] https://knowledge.wharton.upenn.edu/article/portugals-economic-recovery-how-much-came-from-ditching-austerity/

Comparing the data listed by Moro et. al. to other time windows both before 2008 and after 2010 could tease out if this subscription behavior is partially influenced by external economic factors within the Portuguese economy or the global economy at large. Additionally, considering inflation over periods of time for the net income for the term deposits could alter the profitability but increase complexity in the data analysis.

In the exploratory data analysis above, external data about Portugal's age demographics were included in the discussion comparing the client data and the general population at large. Alternatively, if more data about the job, education, marital status, etc. distribution for the general population of Portugal compared to the client data featured in this data set, there may be overarching differences in this population that chooses to bank with the source(s) of the data compared to the greater population of Portugal. This information may be useful to a broader analysis increasing the probability of successful subscriptions.

Lastly, additional profit loss could be determined by better anticipating the unrealized losses of the false negative client labels, customers who desired to subscribe to a term deposit but were not contacted. This loss in profit does not directly cost the bank money as no resources were used to pursue these clients. Yet these customers could offer a group that could be better marketed towards to increase subscriptions and future profits.