Predicting the Success of Bank Marketing Campaigns

By: Jack Molesworth

1. Introduction:

Telemarketing plays a vital role in the banking industry. Banks rely on their telemarketing departments to reach out to customers and secure them for long term deposits. Despite the usefulness of telemarketing, banks lose revenue on calls that do not successfully yield a subscriber. If banks could know the likelihood that a customer will subscribe beforehand, they could focus their efforts solely on those with a high likelihood, thus maximizing profitability.

Using marketing data collected from a Portroguese banking institution, I aim to identify what factors influence the likelihood of whether a potential customer will subscribe. After these factors are identified, I will build a predictive model and threshold it for profitability.

---

2. Client Profile:

My client for this project will be any bank that is looking to maximize the profitability of its marketing efforts. The data I will be using was collected from a Portrugese banking Institution, but can be adjusted if given data from other banks.

---

3. Data Wrangling and Exploratory Data Analysis:

The dataset I used was acquired from the UC Irvine Machine Learning Repository. It contained no duplicates or null values.

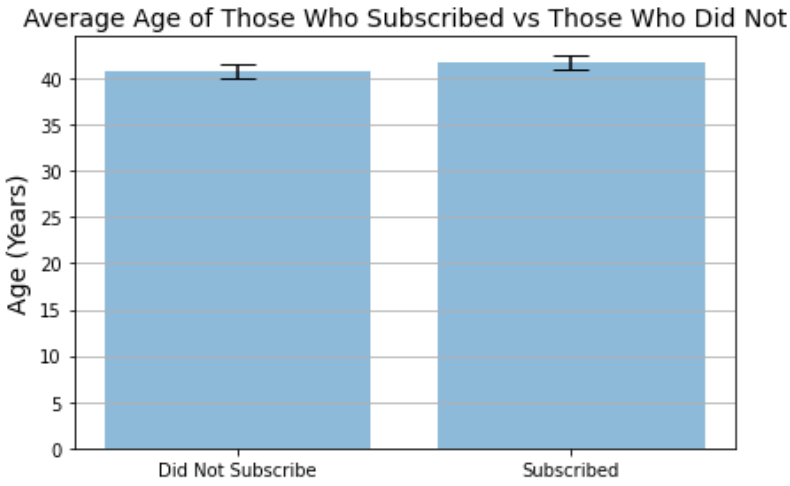

Figure 3.1: Average Age of Those Who Subscribed vs Those Who Did Not

Figure 3.1 shows a comparison of the average age of those who subscribed vs those who did not. Based on this plot, the average age of those who did not subscribe appears to be slightly lower than that of those who did.

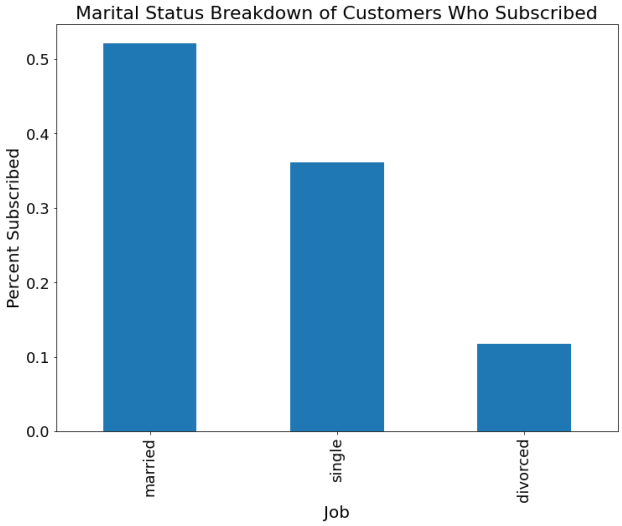

Figure 3.2: Marital Status Breakdown of Customers Who Subscribed

Figure 3.2 displays the marital status breakdown of customers who subscribed. Married customers appear to represent the largest percentage of customers who subscribed, followed by single, then divorced.

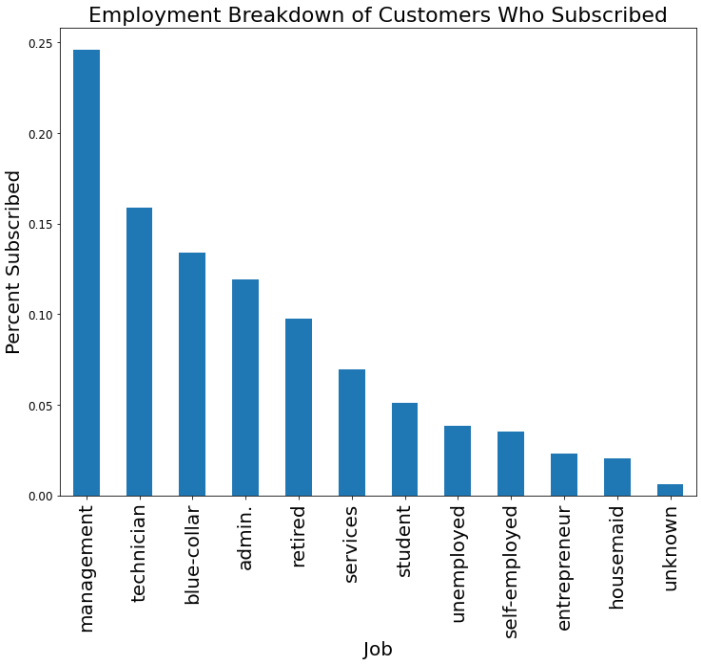

Figure 3.3: Employment Breakdown of Customers Who Subscribed

Figure 3.3 displays the employment breakdown of customers who subscribed. Customers who work in management appear to represent the largest percentage of customers who subscribed while those whose job is unknown appear to represent the lowest percentage.

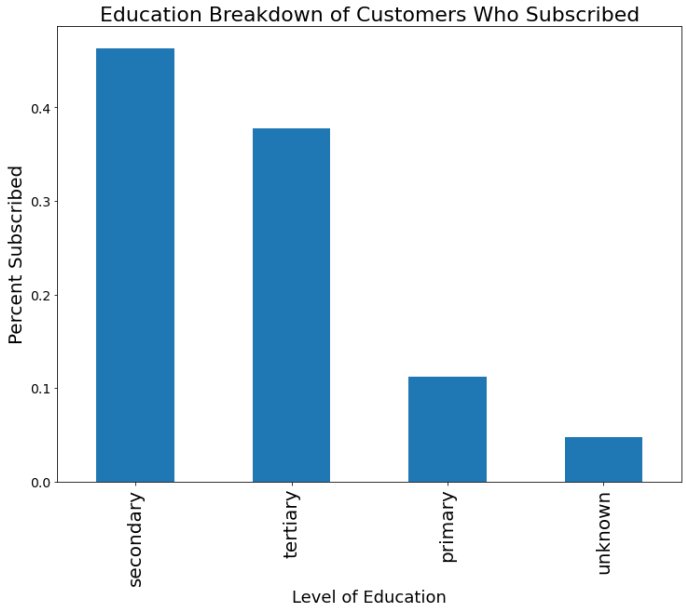

Figure 3.4: Level of Education Breakdown of Customers Who Subscribed

Figure 3.4 displays the breakdown of customers who subscribed by level of education. Secondary education appears to make up the highest percentage, followed by tertiary, primary, and unknown.

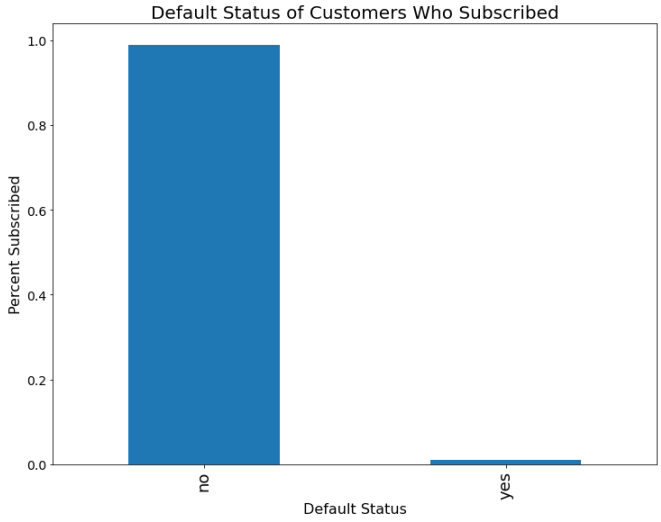

Figure 3.5: Default Status of Customers Who Subscribed

Figure 3.5 displays the percentage of customers who defaulted vs those who did not out of the customers who subscribed. Customers who did not default appear to make up a significantly larger percentage than that of those who did.

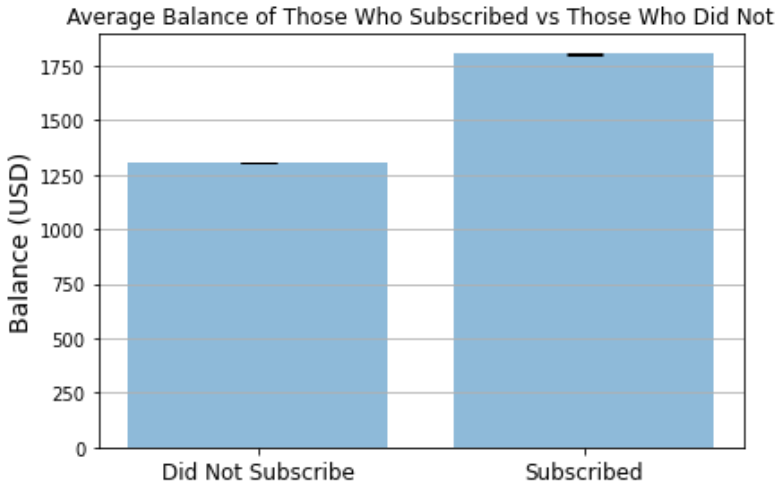

Figure 3.6: Average Balance of Customers Who Subscribed vs Those Who Did Not

Figure 3.6 displays the average account balance of those who subscribed vs those who did not. The average balance of those subscribed appears to be significantly higher than that of those who did not by roughly $500.

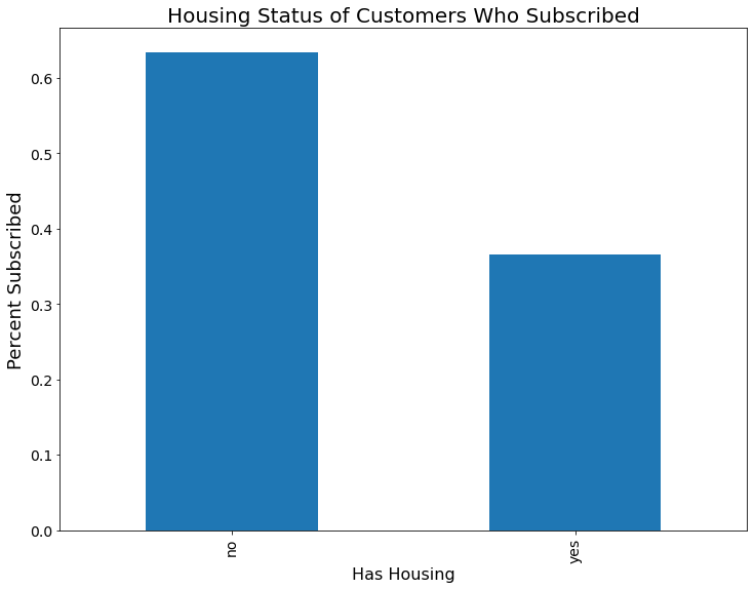

Figure 3.7: Housing Status of Customers Who Subscribed

Based on figure 3.7, customers who subscribed appeared to not have housing at a higher rate than that of those who did have housing.

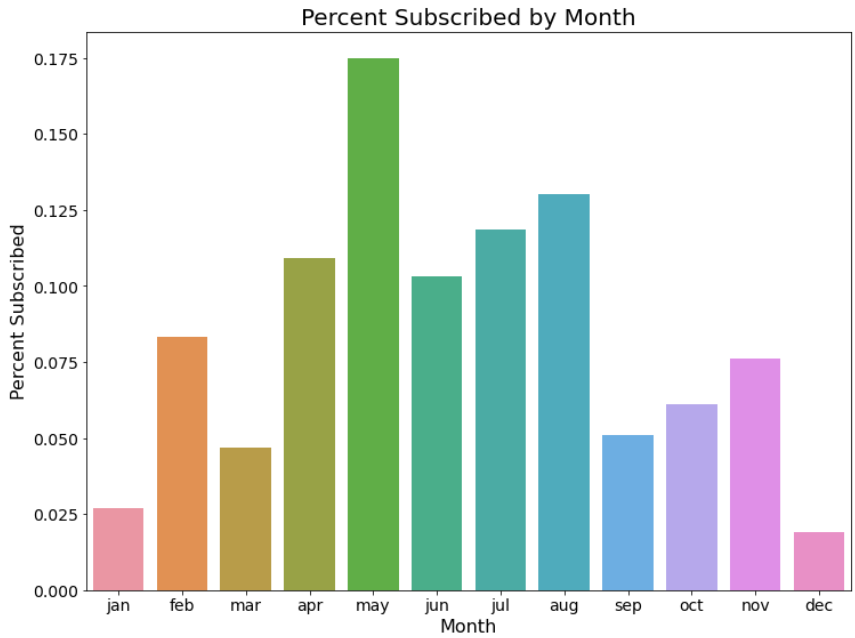

Figure 3.8: Subscription Breakdown by Month

Figure 3.8 shows the rate of new customer subscriptions by month in descending order. May appears to have the highest rate of new subscriptions while December appears to have the lowest.

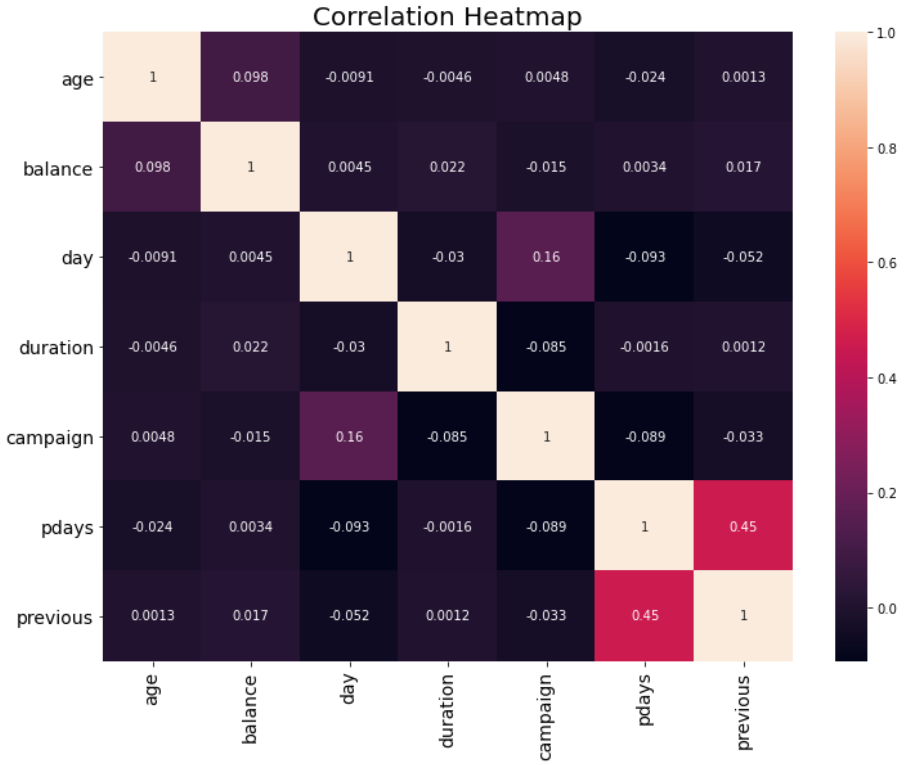

Figure 3.9: Correlation Heatmap

Based on the correlation heatmap shown in figure 3.9, previous and pdays appear to have the strongest positive correlation while age and day have the strongest negative correlation.

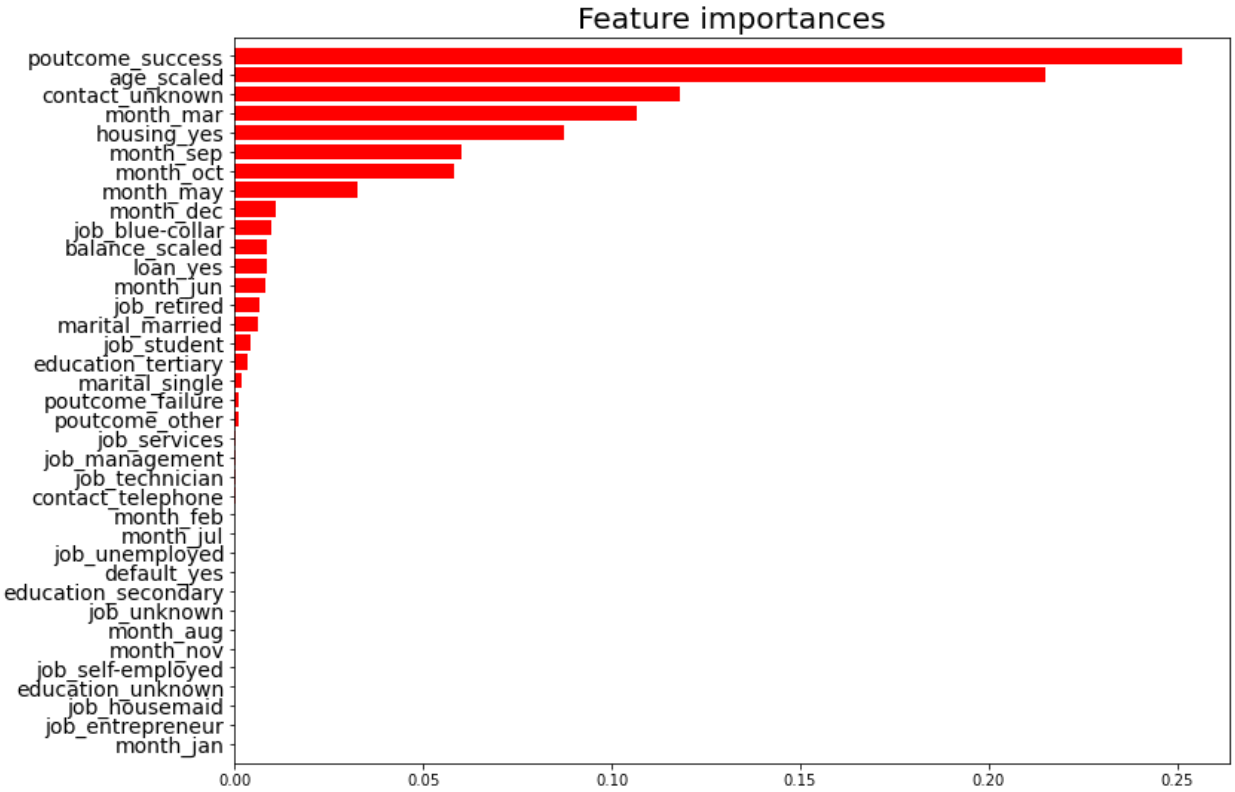

Figure 3.10: Feature Importance Bar Chart in Descending Order

Based on the Random Forest feature importance bar chart shown in figure 3.10, only 20 out of 37 variables appear to have any influence over the performance of the model. Of these 20 variables, poutcome_success had the highest influence while poutcome_other had the lowest. This makes sense that customers who had previously subscribed deposit would be likely to do so again.

---

4. Modeling

Before I began modeling, I used the function SelectKBest and sorted each feature of the dataframe in descending order by k score. The variable poutcome_success had the highest k score, while job_unknown had the lowest. After sorting each feature by k score, I created separate dataframes with different feature sets (5, 10, 20, 30, all) to see which one performed the best in a random forest model. Running separate random forest models for each feature set produced the following results:

Feature Set	ROC AUC Score
Top 5 k-score features	0.5806
Top 10 k-score features	0.6040
Top 20 k-score features	0.6812
Top 30 k-score features	0.7364
All features	0.7727

Based on the results, using all features produced the highest ROC AUC score. Once I identified which feature set produced the highest ROC AUC score, I wanted to see if the random forest model performed better than other algorithms (KNN, Logistic Regression). Running all three algorithms produced the following results:

Algorithm	ROC AUC Score	Best Params
Random Forest	0.7727	criterion='entropy', max_depth=9, max_features='log2', min_samples_leaf=2, n_estimators=10
KNN	0.7762	n_neighbors=9, p=1
Logistic Regression	0.7628	c=0.1

Of the three algorithms I tested, Random Forest performed the best.

Thresholding for Profitability:

After establishing that Random Forest was the best performing model, I wanted to threshold it for profitability. To do this, I created a function that would calculate profitability at a range of thresholds and return a dataframe with the profit made ($) at each threshold. In order to calculate profit, I first needed to make an estimate for two key variables: cost and revenue. Before I could calculate these variables, other information needed to be estimated such as revenue per subscription and cost per call. To estimate these numbers, some estimations needed to be made. For example, I estimated that the cost per call would be $5 if an employee is paid $50 an hour

and makes 10 phone calls an hour. As for revenue per subscription, I assumed a value of . In a real work setting, I would have the actual data, but for the sake of testing this model, I used $50.

Revenue was calculated by multiplying the number of true positives by revenue per subscription. Cost was calculated by multiplying the cost per call by the number of times the probability was over each threshold and taking the sum of every product. Finally, profit was calculated by subtracting cost from revenue. Plotting the resulting dataframe returned the following curve:

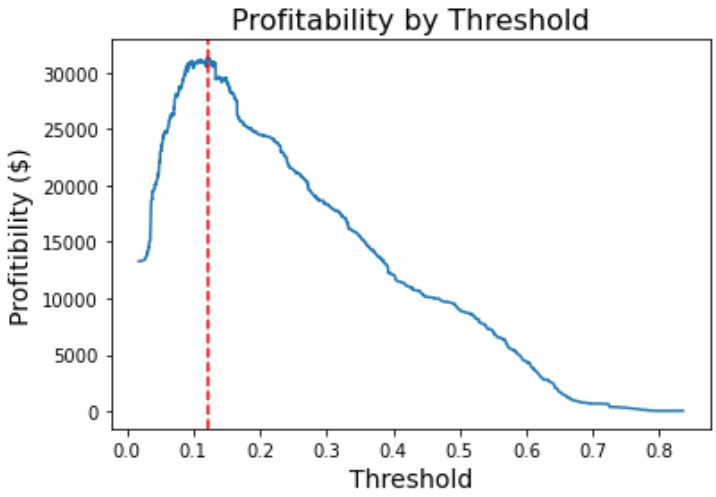

Figure 4.1: Profitability curve

Based on the profitability curve shown in figure 4.1, the optimal threshold was 1.2 with a profit of about $31,350.

---

5. Conclusion

With this model, banks will be able to pick and choose which customers to call based on whether or not they meet the optimal threshold. With the current data, if a customer is not likely to subscribe at threshold .12, then the bank would be better off not calling them. Of course, this could change when real workplace data is used, but for now, this is the result. Given that profitability was $13,280 at the lowest threshold, 0.0179, and $31,350 at threshold .1209 where profitability was maxed, the model was able to increase profits by 136% over the baseline.

---

6. Future Improvement

Ideas for improvement would include:

- **Use real numbers for profitability function:** In the profitability function, I used made up numbers for cost per call and revenue per subscription. I tried to make these numbers as realistic as possible, but if this model were to be used in a real world setting, it would need the actual company numbers for these variables in order for the model to be useful.

- **Gather additional customer data:** The dataset provided was fairly limited with only 12 usable variables not including the outcome variable. Gathering additional customer information could improve the predictive strength of the model.

- **Use banking data collected from other countries:** There could be cultural differences which influence the success of marketing campaigns in Portugal. Because of this, the trends in this dataset may not appear for US banks, for instance. If the model is to be used in another country, then it should be built based on data from that country.