Bank Marketing Project

By: Jack Molesworth
With thanks to Springboard mentor: Benjamin Bell

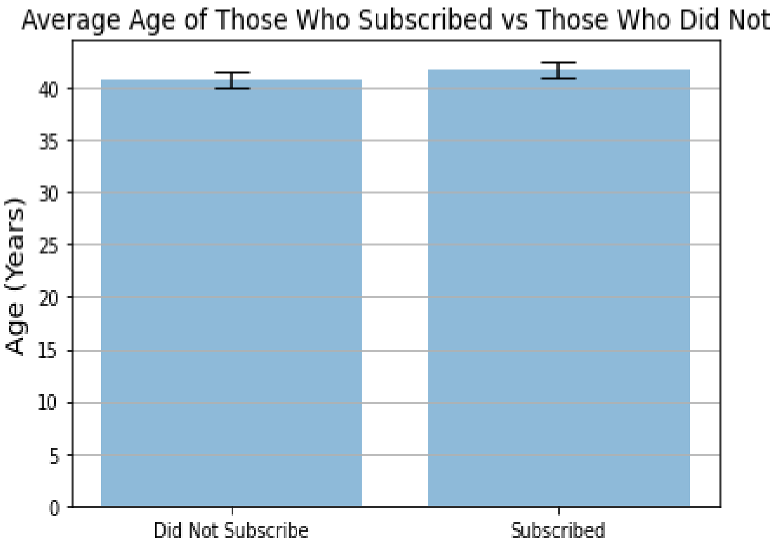
Introduction

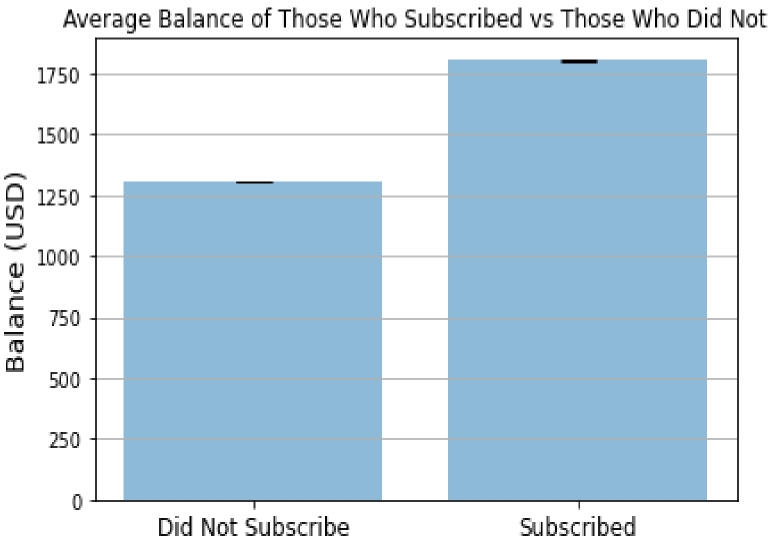
- Telemarketing plays a vital role in the banking industry.
- With telemarketing, banks can reach out to potential customers for long term deposit subscriptions.
- If banks can predict which potential customers are most likely to subscribe, they can specifically target them and avoid wasting time and money on those who are unlikely.
- This project aims to create a model that can predict whether a potential customer will subscribe based on data collected from a Portruguese banking institution.
- After this model is created, I will threshold it for profitability.

Data Wrangling

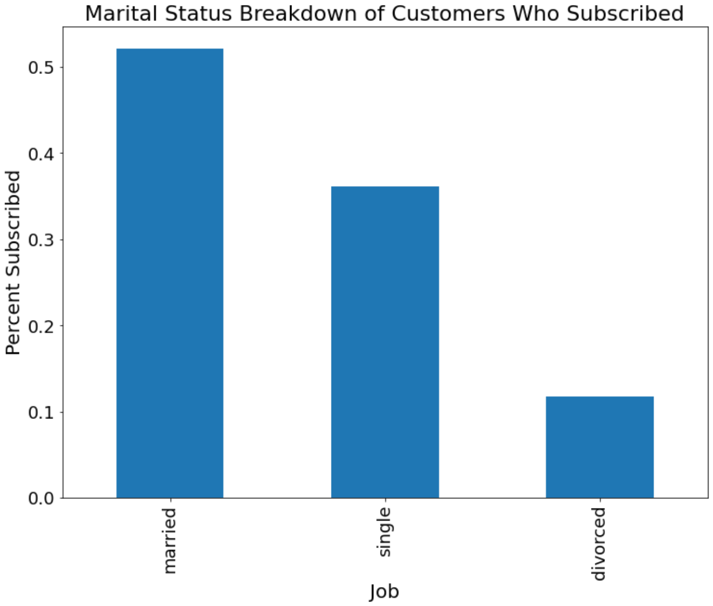
- For this project, I used the Bank Marketing Dataset, available on the UCI Machine Learning Repository
- The dataset contains 21 columns, including the outcome variable which tells whether or not the customer subscribed.’
- Cleaning and preparation steps:
    - Dropping “duration” column because the duration of each call is not known before the call is performed and by the end of the call, the outcome is known.
    - Dropping duplicate values

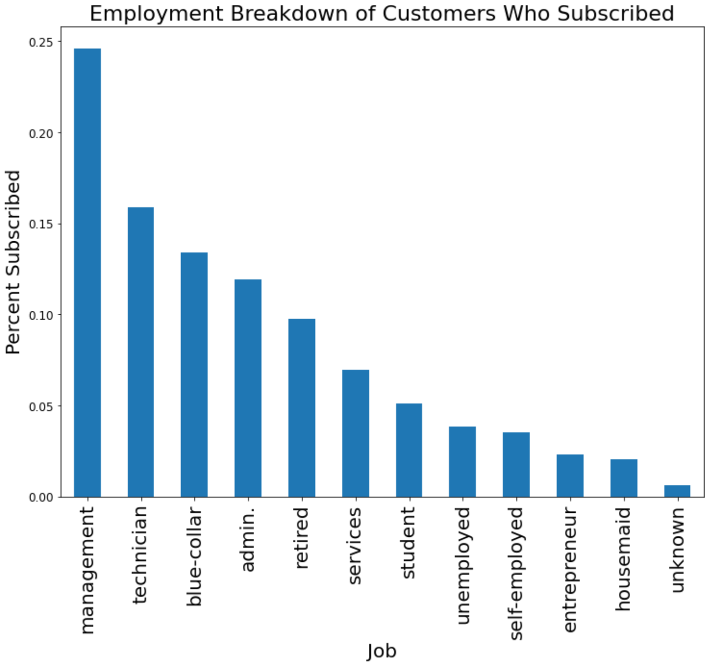
Exploratory Data Analysis



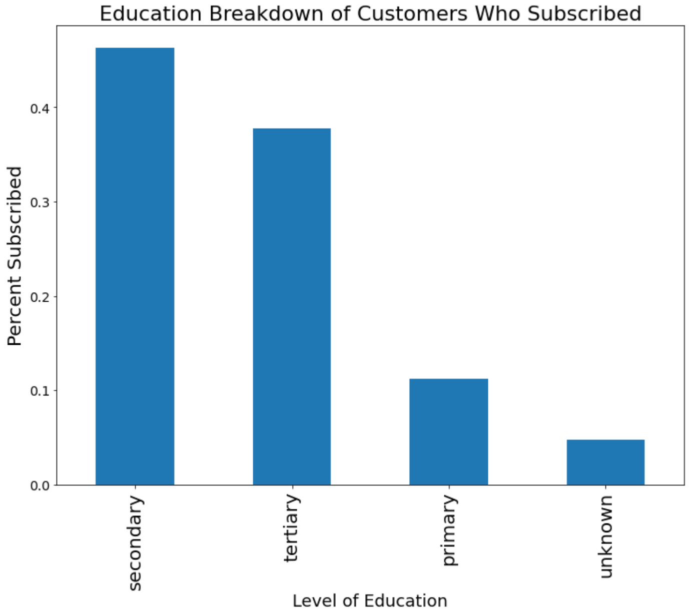


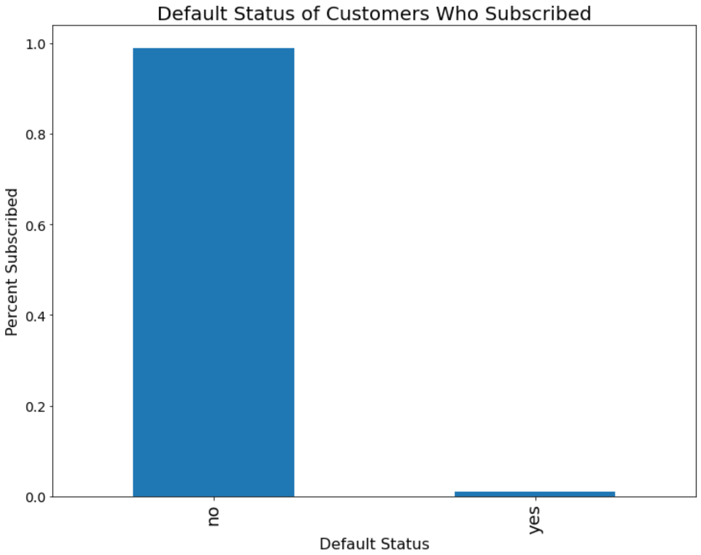
Exploratory Data Analysis (Continued)



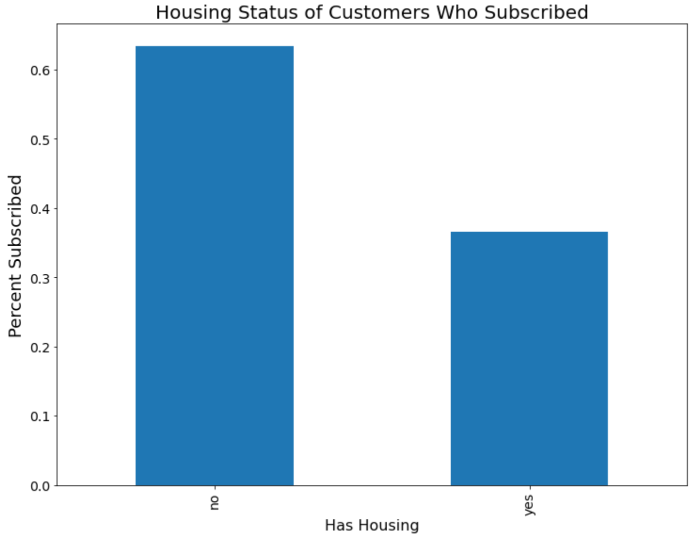


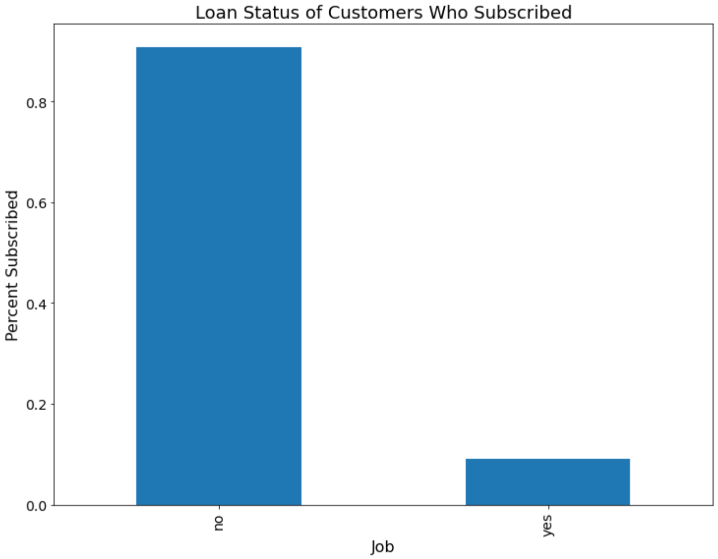
Exploratory Data Analysis (Continued)



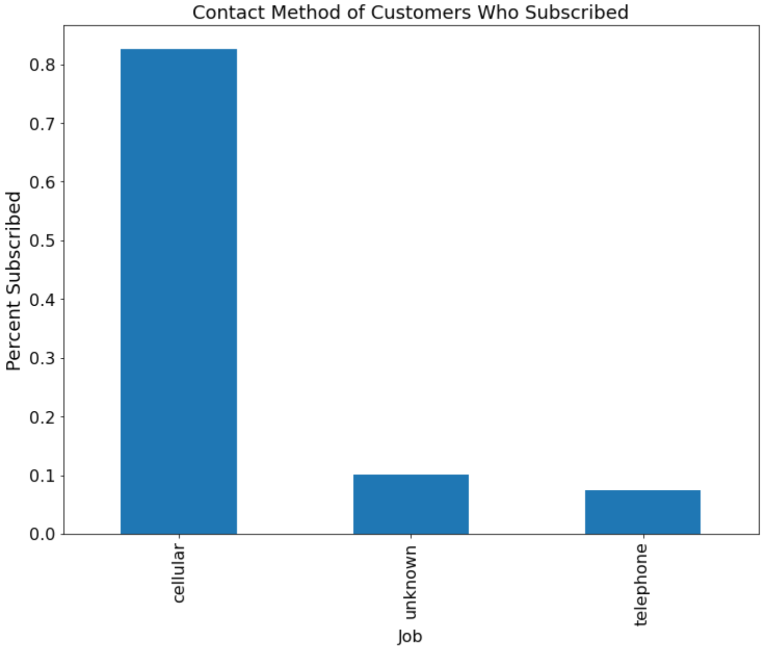


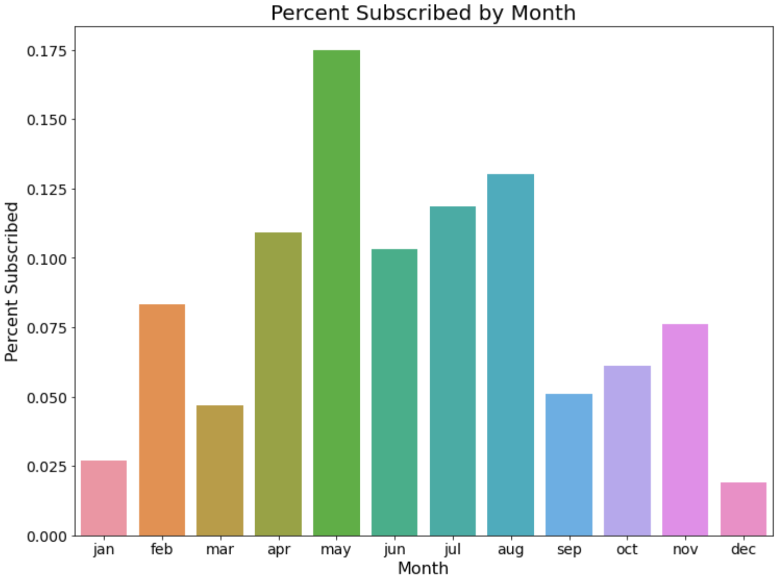
Exploratory Data Analysis (Continued)

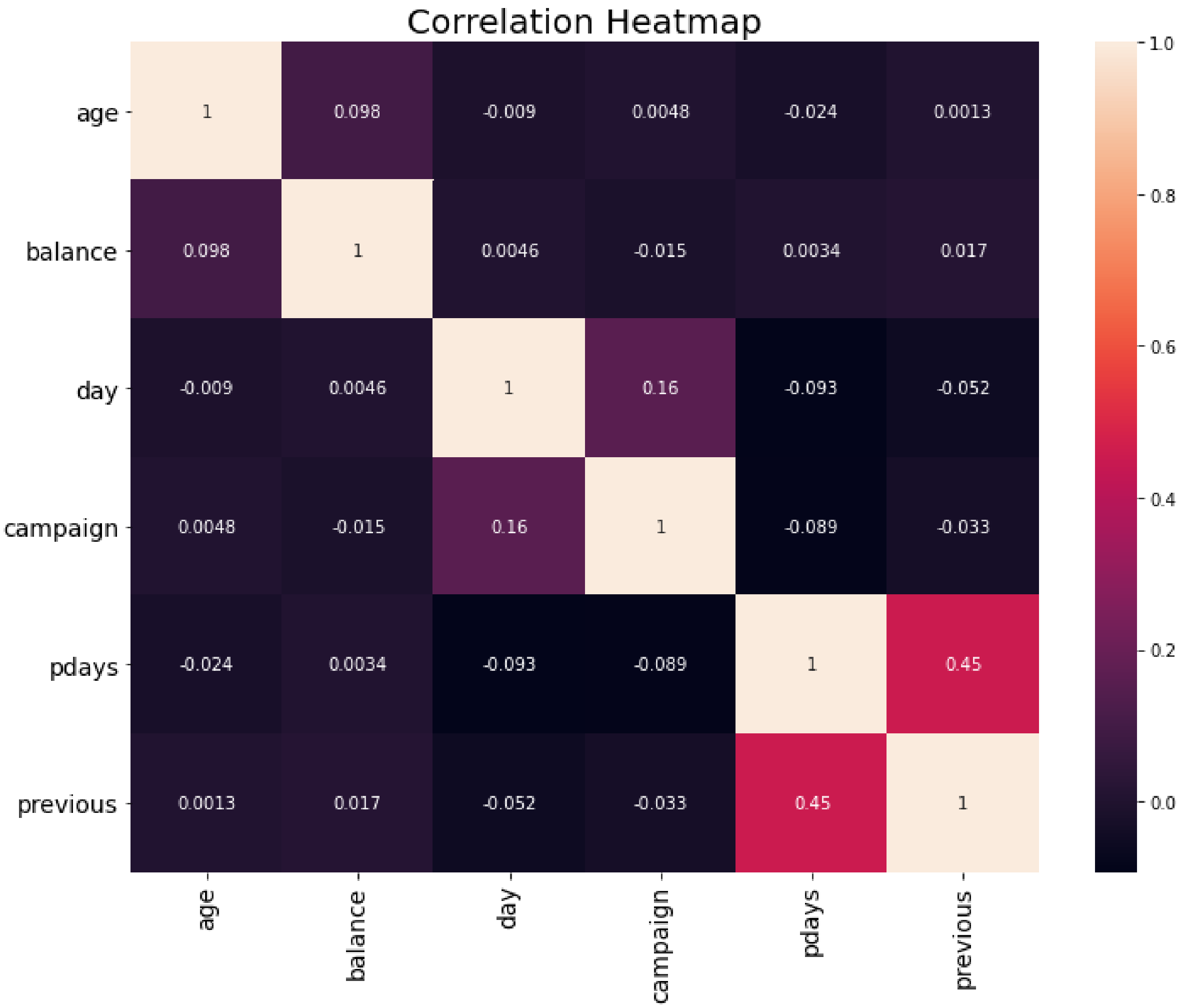


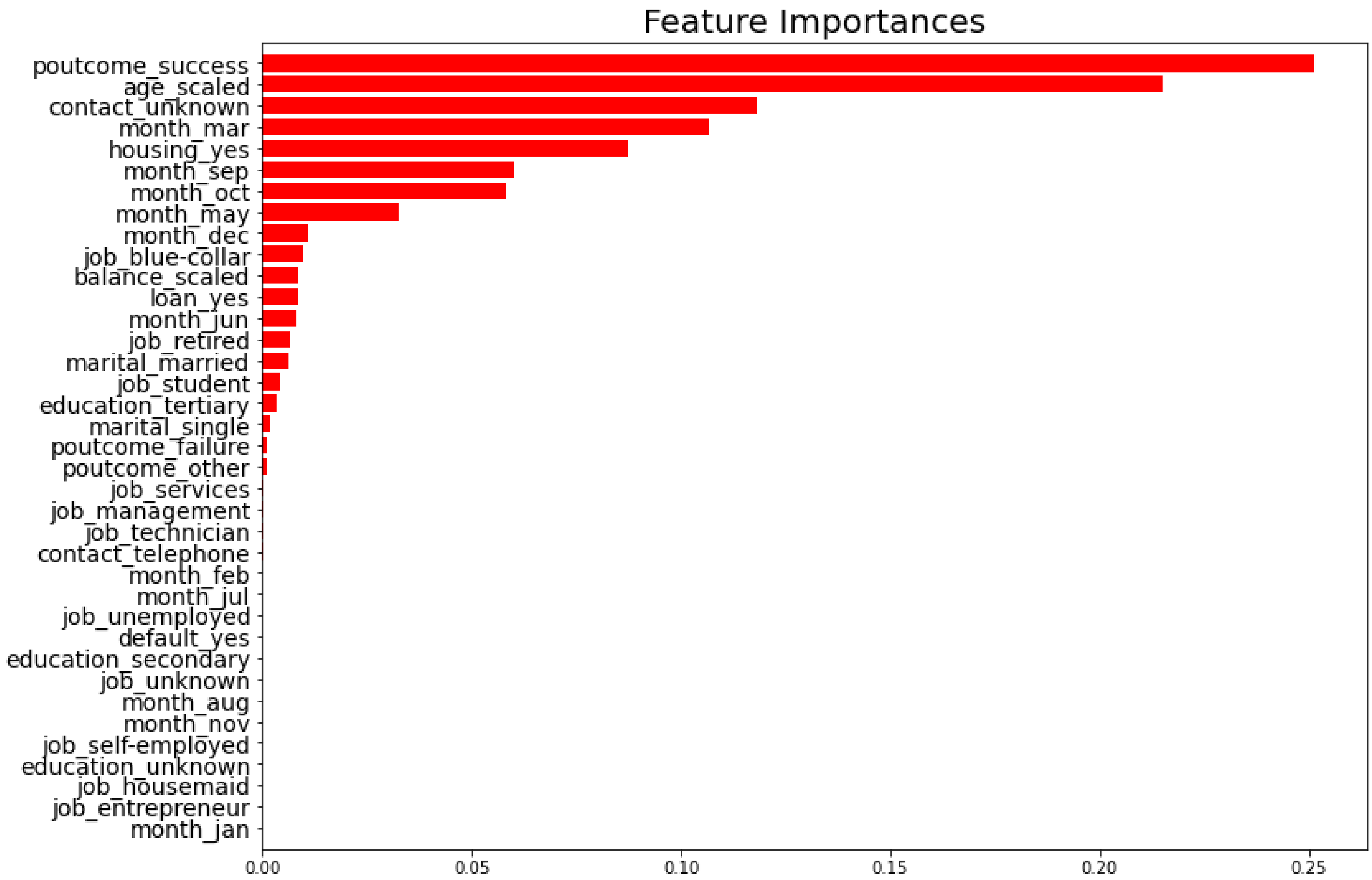


Exploratory Data Analysis (Continued)






Correlation Heatmap
age
balance
day
campaign
pdays
previous
1
0.098
-0.009
0.0048
-0.024
0.0013
0.098
1
0.0046
-0.015
0.0034
0.017
-0.009
0.0046
1
0.16
-0.093
-0.052
0.0048
-0.015
0.16
1
-0.089
-0.033
-0.024
0.0034
-0.093
-0.089
1
0.45
0.0013
0.017
-0.052
-0.033
0.45
1
age
balance
day
campaign
pdays
previous
1.0
0.8
0.6
0.4
0.2
0.0


Feature Importances
poutcome_success
age_scaled
contact_unknown
month_mar
housing_yes
month_sep
month_oct
month_may
month_dec
job_blue-collar
balance_scaled
loan_yes
month_jun
job_retired
marital_married
job_student
education_tertiary
marital_single
poutcome_failure
poutcome_other
job_services
job_management
job_technician
contact_telephone
month_feb
month_jul
job_unemployed
default_yes
education_secondary
job_unknown
month_aug
month_nov
job_self-employed
education_unknown
job_housemaid
job_entrepreneur
month_jan
0.00
0.05
0.10
0.15
0.20
0.25

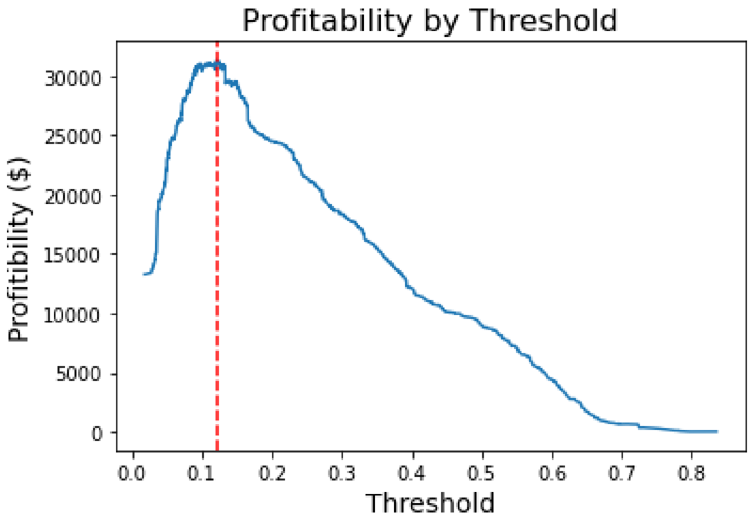
Modeling

- Modeling steps included:
  - Sorting the features of the dataframe in order of importance.
  - Creating separate data frames for different feature sets (top 5, top 10, top 20, top 30, all)
  - Testing each feature set on a random forest model to see which one performed the best
- With an ROC-AUC score of .7702, using all features performed the best.
- After establishing the all features set as the best one, I tested other algorithms (Logistic Regression and KNN)
- Of the three algorithms I tested, Random Forest maintained the highest ROC-AUC score

Thresholding for Profitability

- To maximize the usefulness of my model, I thresholded it for profitability.
- In order to do this, I needed to do this, I needed to estimate some numbers that would be provided in a business setting, such as revenue per subscription and cost per call.
- To estimate cost per call, I assumed that a bank employee would be making $50 and 10 calls in an hour. This would make the cost per call $5.
- For revenue per subscription, I assumed a value of $50.
- Next, I needed to use these numbers to calculate revenue and cost.
- To calculate revenue, I took the number of true positives and multiplied it by revenue per subscription.
- To calculate cost, I multiplied cost per call by the number of predicted positives.
- I was finally able to calculate profit by subtracting cost from revenue.

Thresholding Results



I achieved a max profitability of $31,350 at threshold .1208. At the lowest threshold, .0180, profitability is $13,280. Thresholding gives us a 136% increase in profitability.

Conclusion

With the model I created, 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 .1209, 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 of 0.0179, and $31,350 at threshold .1209 where profitability was maxed, the model was able to increase profits by 136%.

Future Work/Improvements

- **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.

Source

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing