Portuguese Bank Marketing

Analysis and Recommendations

By: Richard Broyles

Capstone 2

Project Objective

- To classify whether or not a customer will subscribe to deposit program in Portugal.
- Bank was using telemarketing campaign to get potential customers to open a new savings account.
- Data set is publicly available from the UCI data repository (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing/)

Data Characteristics

- Dataset has 16 features, with one outcome variable.
- Dataset has over 45,000 data points available.

Data Cleaning

- Several changes were made to prepare the data for analysis:
  - All ambiguous values, such “other” or “unknown” were removed.
  - All outliers more than 3 standard deviations away from the mean were dropped.
  - Change the ‘y’ variable to ‘response’ and convert it to binary values (1/0).

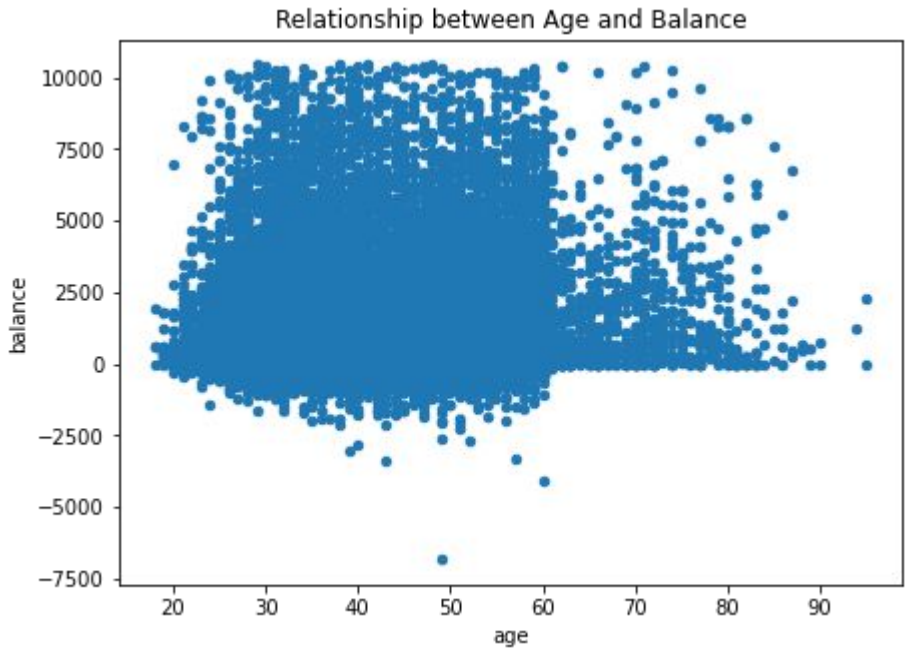
Exploratory Data Analysis

- Key variables in this analysis were:
  - Age – most customers were in their 30s – 40s
  - Balance – large variability; over 2200 records removed due to outliers
  - Duration – length of a call in seconds
  - Campaign – times a customer was called.

Age vs. Balance

No direct relationship exists between these two variables, however, there are some interesting points that were worth investigating.

There were several people who were in their 60's who had a low balance. This is explained by the fact that most people in their 60s are retired and have no steady income.

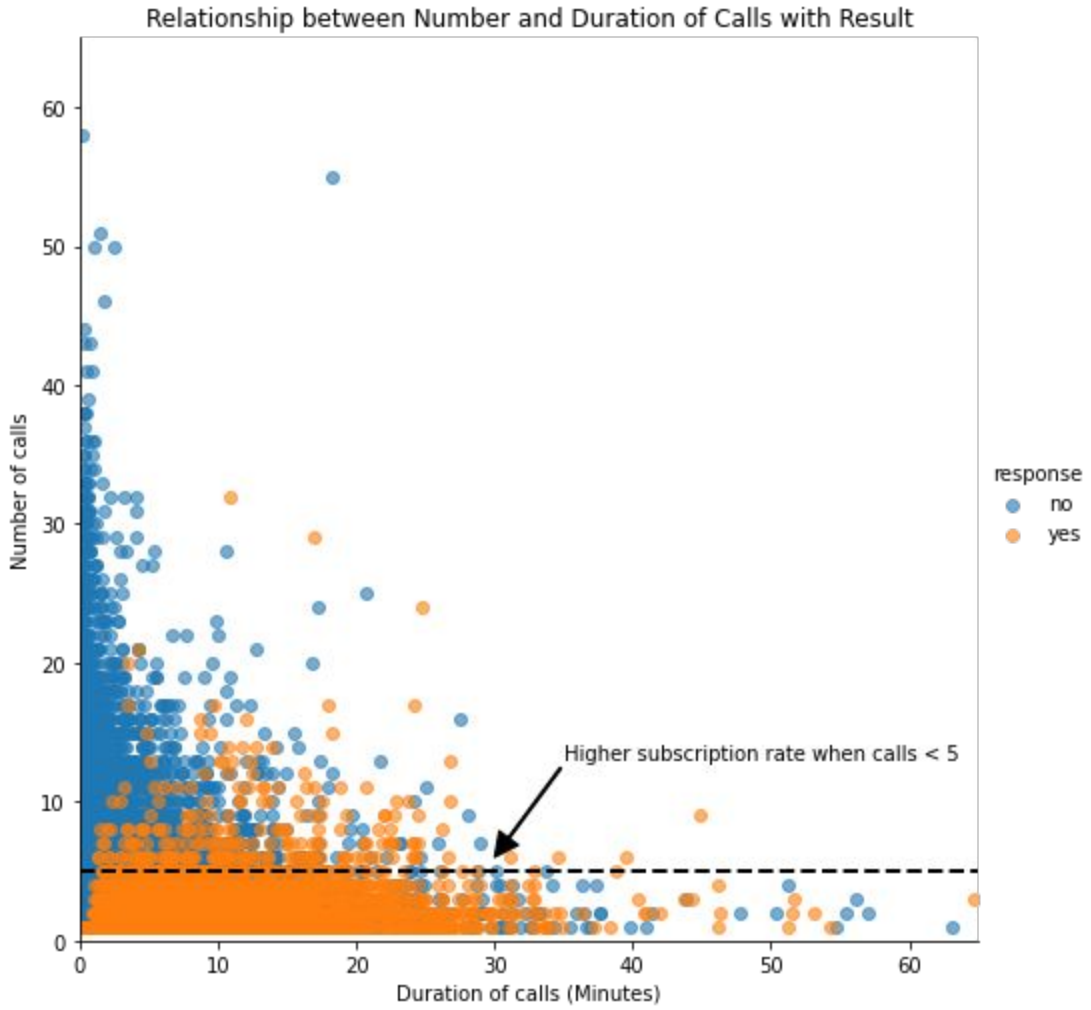


Duration of calls vs. Number of calls (Campaign)

Two distinct groups appear in this chart.

One group are the number of customers who said ‘yes’ to opening an account. Most of these customers were contacted a maximum of 5 times and each call lasted under 30 minutes.

The group in blue are the clients who said ‘no’ to opening an account. The more times a customer was contacted about opening an account, the less likely they are to open one.



Data Visualization

- After some exploratory analysis was complete, the next step was to see if any relationships can be found among the columns.
- This analysis focused on the relationship between the subscription rate and main variables.

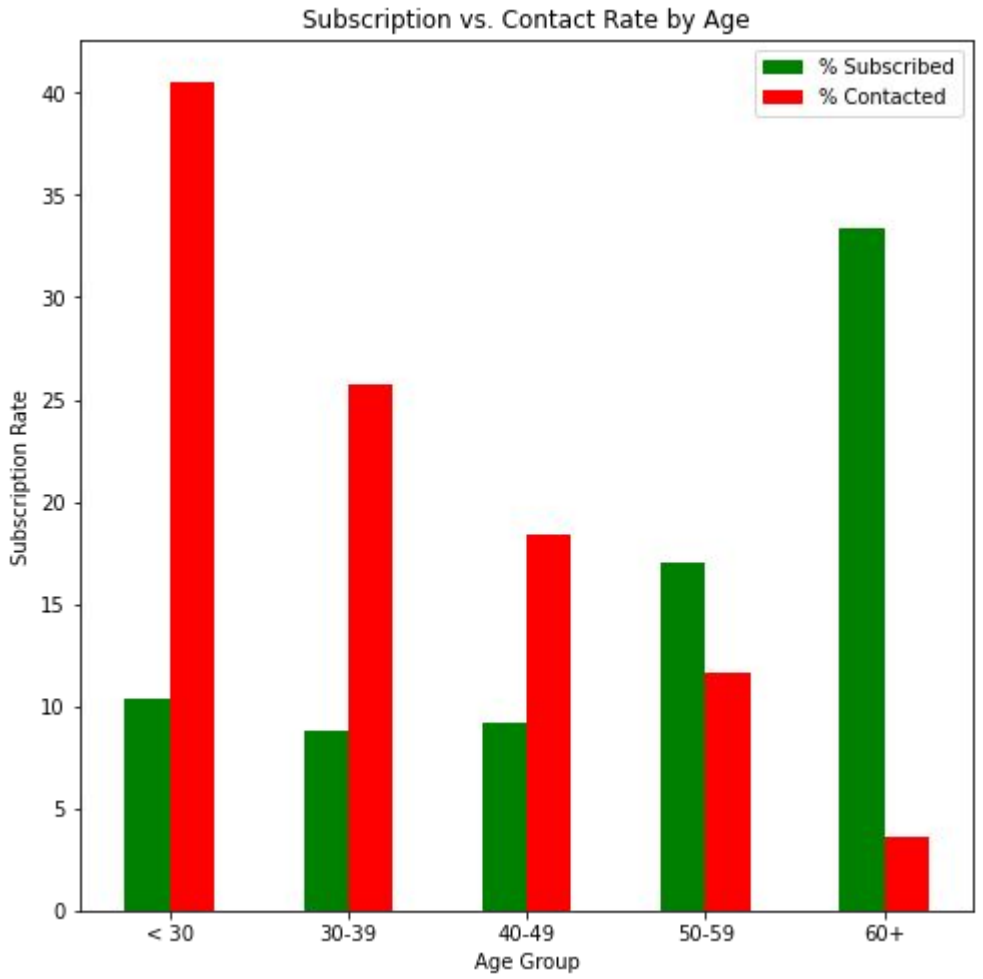
Age vs. Subscription Rate

The bank here is focusing on pushing their deposit to people in several distinct groups.

However, the bank is having more success in only two age groups:

1) People who are between the age of 50-59.
2) People who were over the age of 60.

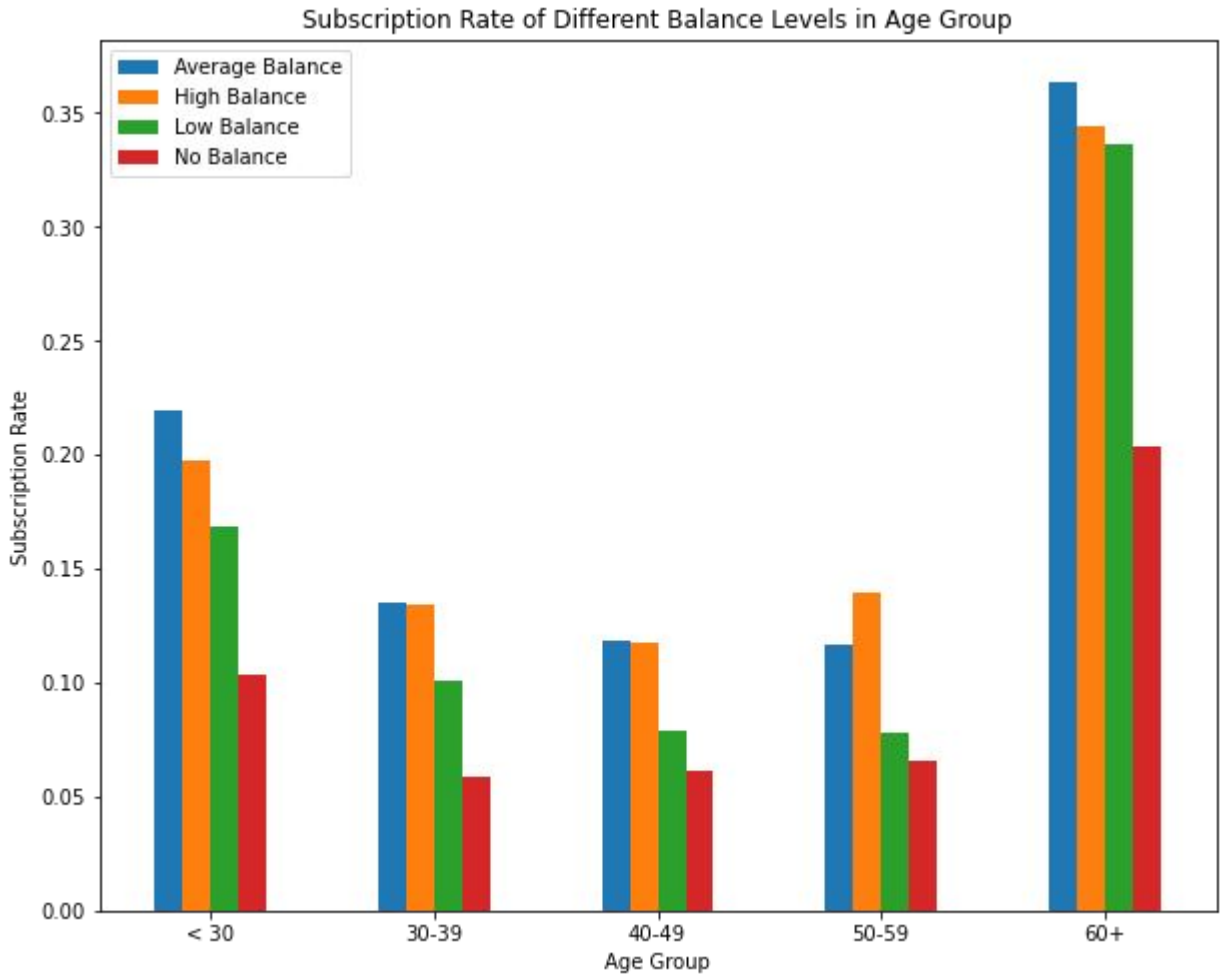
The bank had numerous contacts with people under the age of 30, despite the low number of subscriptions.



Subscription Rate for Balance Levels

People over the age of 60 had a high subscription rate because they saw a term deposit as a worthwhile investment. People over the age of 60 who had at least some money in the bank subscribed more often.

Another group the bank should focus on is the under 30 group, the majority of which are students. They saw the deposit as a way to have a steady income stream while they were in school.



Modeling

- Classification algorithms were used to analyze the customer's statistics.
- The algorithms used were:
  - Logistic Regression
  - K-Neighbors
  - Random Forest
  - Gaussian Naïve Bayes
- Data was prepared by:
  - Selection of the most relevant customer information.
  - Categorical variables were converted to dummy variables.
  - Dataset was split into 80/20 training/test.

Modeling (cont’d)

- The model that had the best performance was Logistic Regression, which had an accuracy score of 89.08%, but this number is misleading.
- The dataset is heavily unbalanced, with most of the customers replying ‘no’ to opening a deposit.
- The accuracy score was biased, and further evaluation was required.

Regression Analysis

- Duration of phone call is correlated to the outcome of the campaign; can be used as another indicator of success.
- Six regression algorithms were used:
  - Linear Regression
  - Lasso
  - Ridge
  - ElasticNet
  - K-Neighbors
  - DecisionTree
- Best performing regression algorithm was Ridge, with a MSE of 17.78.
- Indicates that this model is a sound model for predicting the target variable.

Conclusions

- Target customer profile was established with the following features:
  - Age (Age < 30 or Age > 60)
  - The job type (Students and Retired people)
  - Balance of at least 5000 euros.
- By utilizing both a classification and regression model, bank will be able to predict the customer's response.
- Predicting the duration of a call and adjusting the marketing plan will increase the efficiency of the campaign.

Recommendations

- Improve the timing of the campaign.
- Smarter marketing design.
- Create a better services provision.