Predictive Model on Campaign Impacts

Final Report for DSC148-WI24

Donald Taggart

dtaggart@ucsd.edu

Andrew Peng

apeng@ucsd.edu

**Introduction**

Direct marketing campaigns allow you to target the idea customer by understanding their needs and wants in order to gain customers in the most cost-effective way. However, while this method does result in a much higher customer-response rate, direct marketing and sale campaigns come with very high costs regarding management and administrative overhead. These costs can be reduced if the marketing campaign targets individuals and groups that are much more likely to purchase said product. In this project, we will build a model that will predict an individual’s response to a direct marketing campaign.

# Dataset

**1.1 Identify Dataset**

The dataset used will be data collected from a Portuguese banking institution based through phone calls (bank-full.csv). This **bank**[[1]](#footnote-1) dataset was found on UCI’s machine learning repository and was obtained as a .csv file. The dataset contains 45,211 instances with 16 features as well as a corresponding variable for each instance to denote if the marketing campaign was successful in gaining a new term deposit subscriber. This data ranges from May 2008 to November 2010 and has no missing values.

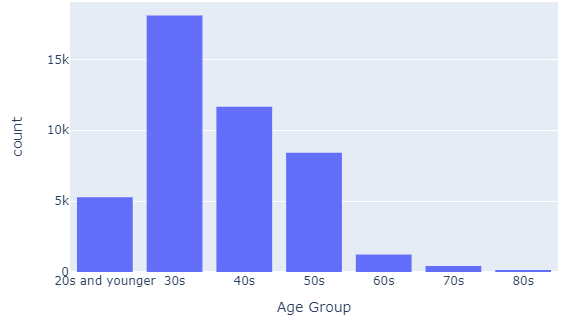
Columns within the dataset focus on the basic features of each individual contacted (age, job category, marital status, education). It also included a few more in-depth measurements of each individual’s financial status (loan, housing, balance) as well as data summarizing if the individual had been contacted before for a similar campaign and the outcome of said campaign (last contacted, number of contacts, previous outcome).

**1.2 Basic Exploratory Data Statistics**

The most basic thing to acknowledge is that, after searching through this dataset, there are no NaN or missing values in the dataset. However, for columns where the data is quantitative, values that do not fit into any existing category are saved as “unknown” instead.

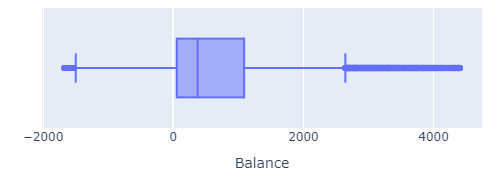
To gain a better understanding of the individuals who were contacted during the marketing campaign, we explored most of the columns within the dataset.

**age** records the numeric ages of each of the individuals contacted during the direct marketing campaign. However, a majority of the data is centered around the ages of 30 to 50, with a very small number of responses coming from those in the 60+ age ranges (Figure 1).



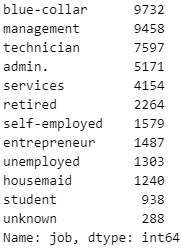
**Figure 1: Frequency Distribution of Age Groups**

**balance** describes the numeric average yearly balance in euros. However, here we see that some people have negative yearly balances (Figure 2). This is very likely due to overdrafts where individuals spend more than what they have in their account. This can also be attributed to having overdraft fees and other late payments[[2]](#footnote-2). This will prove very useful and interesting to see if those with lower average account balances are less likely to subscribe a term deposit.



**Figure 2: Box plot of balance accounting for a single standard deviation**

**job** has 12 predetermined categories for jobs. These include the following (Figure 3).

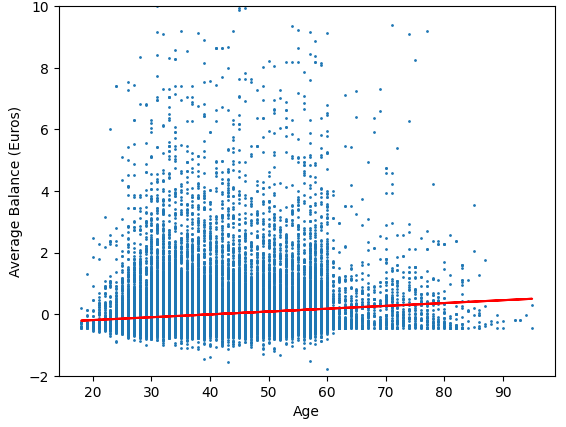


**Figure 3: Frequency of Job Categories**

From this we can gleam a few basic statistics. Firstly, the most common job categories are “blue-collar”, “management”, and “technician”, suggesting that the dataset does indeed cover a diverse range of professional backgrounds, preventing too much skewness from having a predominant job category available. Additionally, the wide spread of jobs is positive as different professions have varying financial needs, influencing their potential decision for a subscription to a term deposit. Thus, by including **job** in the classification prediction model, we can develop a more effective marketing strategy to tailor the marketing campaign to specific needs and preferences of different job categories.

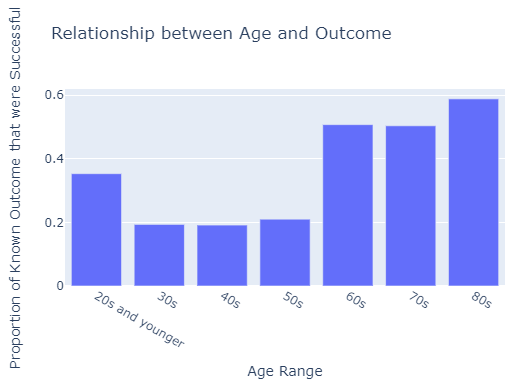
**1.3 Interesting Exploratory Data Analysis**

When doing a simple regression of age on average balance (standardized with z-scores) in euros, a very interesting relationship comes to light (Figure 4). While this could be conditional to the lack of individuals contacted for ages 60 and up, the spread of data increasingly narrows. Interestingly enough, after the age of 50, the number of individuals with negative bank account balances tapers off. However, a little after the age of 60, that number falls nearly to zero. A likely source of this comes from the OCED database regarding retirement ages[[3]](#footnote-3). In the years 2008 to 2010, the average age of retirement (and thusly the age most people accessed their pension and retirement funds) was 62 years old. This graph displays that feature as many do not overspend their retirement pensions once they receive them. In context of this project.

****

**Figure 4: Relationship between Age and Balance (Standardized)**

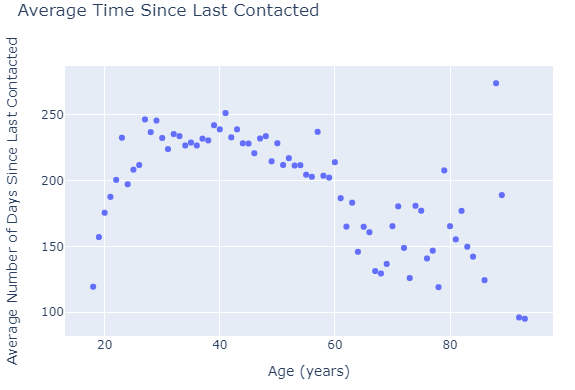
To further ask if people that have retired may be more likely to accept a subscription to a term deposit, we took a look at the relationship between age and the proportion of past interactions (from previous bank marketing campaigns) that were both recorded and successful. Here, we can see a clear jump in the outcome proportion for those ages 60 and up (Figure 5).

****

**Figure 5: Relationship between Age and known outcome of those previously contacted**

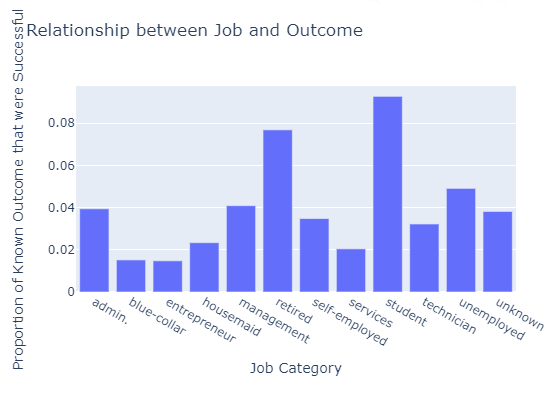
It’s easy to see the large jump of success rates of prior contact records. Thus, it’s fair to say that there will be drastic correlation between age and the predicted value of a successful conversion of contact to subscription.

Unsurprisingly, this trend is noted by the banks themselves. Using data they collected, we can see the average time since last contacted based on age (Figure 6).



**Figure 6: Relationship between Age and average last known contact**

The disparity and frequency in ages contacted along with each age group’s propensity to accept a subscription will be a key point of the machine learning process. To quantify this into yet another way, we plotted the relationship between those who were contacted and the outcome along with the category of their job (Figure 7).

****

**Figure 7: Relationship between outcome and job category**

Here, we see further evidence along with a surprising fact. The two categories of people that had the highest proportion of those with the propensity to accept a subscription were students and those that were retired. Unsurprisingly, retirees once again had some of the highest proportions regarding outcome of prior contact: this makes it yet again a large point of contention within our final model and something that we will focus on. However, surprisingly, students also had a tendency to accept bank marketing campaigns for term deposits based on prior data. As seen in figure 6, the age range of a student was also one of the most contacted groups as well. This is likely because these are the two groups that are the most financially free (or starting out in life) and are the ones most capable of pursing a term deposit for basic interest.

Using these features (while heavily relying on those who are retired as well as their ages), we will determine and improve the dataset and thusly the model mostly through focusing on age, occupation, as well as other factors where people might be financially free enough to pursue a subscription to a term deposit.

# Predictive Task

All material on each page should fit within a rectangle.

# Model

will be added when the publications are assembled.

# Literature

This dataset has been used by past studies, but mainly in the context of improving algorithmic fairness in machine learning, specifically in clustering algorithms. The following studies collectively highlight the versatility of the dataset and its ease of use by applying it to a range of challenges in machine learning.

However, the main difference between articles citing this dataset and the way we used this dataset lay in the way the dataset was used. Rather than a primary feature, the UCI bank dataset was more of a subsidiary to prove fairness in clustering algorithms. Thus, the state-of-the-art method was more akin to clustering over classification. This is likely due to its clean nature and lack of NaN values, leading to minimal work needed to prepare it for use. Thus, the main novelty of our work is more so regarding the fact that we are using recall as a evaluation metric rather than proving a hypothesis based on a new form of clustering or fairness.

**Clustering with Fairness Constraints: A Flexible and Scalable Approach[[4]](#footnote-4)**

The authors (Ziko, Granger, Yuan, and Ayed) do not intend to predict the outcome variable (term deposit subscription), but instead use the bank dataset as a control to contrast a synthetic dataset in order to demonstrate the effectiveness of their algorithm in handling real-world data with various demographic proportions. The main focus of the dataset is far different from our investigation. Instead of maximizing recall to correctly classify instances of data, the authors were more focused on comparing their method with existing fair-clustering methods to show the advantages of the method they had theorized regarding balancing fairness and clustering objectives.

While we focused on a more general overview (also focusing on past interactions with prior marketing campaigns, these authors constrained their dataset to create two initial clusters based on marital status to test fairness by dividing the data into three equal-sized groups, which was very different from the subset of data we used as we used every instance available.

Similar to one of our baselines, Ziko, Granger, Yuan, and Ayed proposed a clustering algorithm that involved K-means clustering.

**Fair Algorithms for Clustering[[5]](#footnote-5)**

The authors (Bera, Chakrabarty, Flores, and Negahbani) in this article also do not intend to predict the outcome variable. Similar to the previous article, these authors used the UCI bank dataset as a way to focus on metrics to solve the problem of finding low-cost fair clustering methods. This dataset was one of several used to evaluate their algorithm. They focused on fair clustering, ensuring that each cluster had fair representation of different groups in the data. For clustering, the authors employed algorithms based on the lp-norm objective, which includes k-means, k-median, and k-center objectives.

While somewhat similar to the XGBoosting and k-nearest neighbors we tuned parameters for, these authors were less concerned about the performance of the model based on predicting the outcome variable and more concerned about how the model they had created stacked up against additional, state-of-the-art models that were normally used. While these authors ended up focusing more on the k-means objective, our project looked more in-depth into XGBoost. Additionally, we also focused on tuning hyper-parameters to increase the recall of our model while the authors of this article were more concerned with allowing groups to lie in multiple protected groups.

**ToPs: Ensemble Learning with Trees of Predictors[[6]](#footnote-6)**

This article (Jinsung Yoon, William R. Zame, and Mihaela van der Schaar, Fellow, IEEE) was the most similar to the model that we created. Here, the authors test the difference in two instantiations of ToPs (trees of predictors) that included adaboost, linear regression, logistic regression, logitboost, and random forests. Many of which we used in our baseline and final model. These instantiations allowed them to explore improvement of models, and also allowed them to compare the performance of their models to create a new approach to ensemble learning and prove that their method was better than other casual uses of state-of-the-art-methods in terms.

These authors employed a similar prediction model in predicting for the same y-variable (which allowed for a comprehensive evaluation of the ToPs algorithm’s effectiveness in comparison to other machine learning algorithms in predicting the acceptance of bank marketing offers based on clientele features. However, the main difference regarded on how they used their new approach. Unlike our project (which used recall), the authors here utilized area under the ROC curve as the loss function due to the unbalanced nature of the dataset that we also noticed.

# Result

## Comparison

## Significance

## Hyperparameter

## Conclusions

# References

1. Bowman, M.,

1. https://archive.ics.uci.edu/dataset/222/bank+marketing [↑](#footnote-ref-1)
2. https://www.self.inc/blog/bank-balance-negative [↑](#footnote-ref-2)
3. https://stats.oecd.org/Index.aspx?DataSetCode=PAG# [↑](#footnote-ref-3)
4. https://arxiv.org/pdf/1906.08207v1.pdf [↑](#footnote-ref-4)
5. https://proceedings.neurips.cc/paper\_files/paper/2019/file/fc192 b0 c0d270dbf41870a63a8c76c2f-Paper.pdf [↑](#footnote-ref-5)
6. https://arxiv.org/pdf/1706.01396.pdf [↑](#footnote-ref-6)