IST 707 Final Report

Mark Nash, Mike DeMaria, Noah Goldie, Joey Eovaldi

Introduction

Our assignment was to apply machine learning techniques to a novel problem to create a positive result for a business. For this exercise we chose to use the Bank Marketing Data Set from the UCI Machine Learning Repository. This data set uses data, from what can be assumed to be a Portuguese bank, detailing call records of sales agents attempting to sell term deposits to customers (the European equivalent of a CD). We analyzed the data to find which clientele to target first when making sales calls, in order to maximize profit for the call agents and the bank.

Assumptions

The data set used in this exercise did not indicate anything about the size or scope of the company wherein the data was derived. As such, we made some reasonable assumptions to provide business context, justification and ROI on this exercise. We positioned the company as holding 500,000 clients. This would be about double the size of Empower Federal Credit Union, a regional bank chain. (Cummings, 2021). The metric of 400 calls per agent with a 20% connection rate was derived from an anonymous interview with a manager of the Preferred Client Partners Group (PCPG), a division of Equitable Financial Life Insurance Company. Industry analysis has confirmed this connection metric; a study performed in 2012 by the Keller Center found a 28% connection rate (Bettencourt, 2012). The figure of $100,000 to replace an employee is based on a recent PCPG job posting, which shows a base salary of $50,000, a training period of two years and working with up to 20 Financial Professionals (Equitable, 2023). As per the job requirements, FINRA Series 7 and Series 66 exams cost $477 (FINRA, 2023). Life and Health Licenses are issued on a state by state basis. Given the licensing costs and training period, we believe the quoted replacement cost of $100,000 may be an underestimate. The successful lead conversion rate of approximately 12% comes from the provided call center data. We also assumed that all calls successfully connected to a customer; we derived this from the fact that every call in the data set had a duration that was greater than 0. Lastly, upon examination of the economic indicators featured in the data set we came to the conclusion that this data was most likely collected during the 2008-2009 global financial crisis.

Business Process

The business has 500,000 clients with 20 Financial Professionals. The desire is to attempt annual touchpoints, which requires 25,000 calls per agent per year. However, each agent can only place at most 400 calls in a week (20,000/year). This means that 20% of clients must be excluded from the call lists. Each week, the agents receive a list of 400 leads. Given the 20% connection rate, agents only hold about 80 conversations per week. The agents are free to choose to call all 400 leads just once, or attempt to follow up and recall the same lead multiple times per week, resulting in more customers not receiving a touchpoint. Within these 80 conversations, about 10 conversations result in a sale. In the original list of 400 leads, there should be about 48 (12%) who are likely to result in a sale. The agents are trying to find these 48 “golden clients” in the lead list, but currently have no basis for choosing one over the other.

The marketing department is unsure of their campaign's effectiveness. They want to know how effective prior campaigns were, and if there is a seasonality component. Marketing would like to know what are the highest and lowest selling periods, economic conditions that lead towards sales, the results of prior campaigns and use customer demographics to craft targeted content.

Data Set

We obtained the Bank Marketing Data Set from the UCI Machine Learning Repository (Moro et al., 2014). The data set has 20 input variables, 1 output variable and 41,188 observations, broken up into 3 groupings. The first is client data, which includes age, job classification, marital status, education, credit default, mortgage loan and personal loan. The group of data relates to the campaigns: call type (cell/land line), last contacted month and day, duration of last call, contact points within the current campaign, days since last contact, prior contacts and prior contact outcome. Finally, the last data group is economic attributes at the time of contact, which includes employment, consumer price index, consumer confidence index, Euribor rate and employment numbers. Finally, the data set includes a binary output variable representing if the customer purchased a term deposit. The data does not include any form of customer ID or agent ID, so we cannot determine if a product was sold to the same customer multiple times, nor could we determine if the data was skewed due to a few agents performing exceptionally. For the data study, we are assuming each success was with a unique customer, and the win/loss rate was evenly distributed amongst the agents.

Models

The predictive models were trained on a subset of the predictors in the data. The economic predictors were excluded because they are not relevant to customer prioritization, and month was also excluded because it correlated too highly with these economic factors. Call duration was also excluded since it cannot be known before the call is made. Lastly, campaign was excluded because the model would not be able to account for new campaigns in the future.

Since the constraints of the problem demand a specific proportion of positive classifications (80%), the model selection was based on precision. We wanted as many sales as possible in the 80% of clients that get called. To maintain this rigid proportion, each of the classification models were made to produce classification probabilities rather than strict classifications. The top 80% of probabilities were then classified as positive and the bottom 20% as negative. It is also important to note that – again, due to the constraints of the problem – the highest achievable precision is 14.1% (which would be achieved if all 11.3% of the sales were contained within the 80% predicted positive; $\frac{0.113}{0.800} = 0.141$).

The classification models tested were Decision Trees, k-Nearest Neighbors, Linear Kernel SVM, Naive Bayes, and Random Forest. Model precisions were estimated using 5-fold cross validation.

Model	**Precision**	**Model Details**
Decision Tree	8.89%	Complexity = 0.05
k-Nearest Neighbors	12.72%	k = 35

Linear SVM	11.96%	Cost = 2
Naive Bayes	13.07%	Laplace smoother = 1; kernel density estimation with 1.25x bandwidth adjustment
Random Forest	12.39%	Predictors per tree (mtry) = 7

The strongest model was Naive Bayes using kernel density estimates for the numerical predictors and a Laplace smoothing constant of 1. With a precision of 13.07%, we expect that the implementation of this model would increase sales by 16% without any other changes to business practices. Additionally, the precision would get better as the total number of clients increases. If the total clients were to increase from 500,000 to 550,000, we expect that this model would increase sales by 23% above the baseline rate.

Predictive Measures

We ran four linear models to find which variables best predicted the success of a sales call. We decided to run four different models, so that each model could specifically focus on the factors that relate to each other and not be skewed by noise of unrelated factors. The first model we ran was our 'basic' model that analyzed factors such as age, living situations, and marital status. We found that calling single clients was more likely to lead to a successful sales call. In our second model, we looked at variables relating to time and found that sales calls made in the months of March, September, October, and December are more likely to lead to a successful sale. These findings are most likely related to the different marketing campaigns run during these months, but nonetheless are interesting predictive findings. We also found that calls made during the months of May, June, July, August, and November predict an unsuccessful sales call. Our third model focused on the different jobs that our clients have and students and retirees are the two factors that could predict a successful sales call. Other jobs, such as technicians, service workers, blue collar workers, and entrepreneurs are less likely to be on the other end of a successful sales call. The fourth linear model we ran focused on the education that our clients have received. Those with a four-year University degree predicted successful sales calls, while those with PhD's predicted unsuccessful sales calls. All of these models simply gave us insight as to which variables influence the result of a sales call, so we could take closer looks at these variables and better predict which clients to call first.

Findings and Conclusions

Overall, we found that there are very actionable insights that can be drawn out of this dataset that can be used to optimize a call center.

We believe that the best application of machine learning with this dataset would be to run the model at the start of every week to create a prioritized call list (similar to a priority queue data structure in computer science) and run the call center according to that list.

Furthermore, we found, and shared in our presentation, some actionable insights about customer qualities and sales processes that lead to higher success rates. Some of these insights include:

- Conversion rates were above average with those:
    - Who do not work
    - Who are single
    - Who have previously purchased the product

    - Who have previously been contacted but did not buy the product
    - Who have a university degree
    - Who were on the phone for longer
        - There is a direct positive correlation between call duration and success rate
- Conversion rates were below average with those:
    - Who had never been contacted before
    - Who work in blue collar jobs
    - Who have PhDs

Based on these trends and insights, we are making a few recommendations to the marketing department. They are as follows:

1. Shift marketing efforts away from the summer months with a heavy lean on the end of the year.
2. Ensure repeated contacts are made.
    a. We should not write off a customer who has declined on the first round, as subsequent rounds may have better results. Current customers are an excellent source to repeatedly market towards.
3. Direct marketing efforts towards local colleges and universities, perhaps going so far as to negotiate a branding partnership.
    a. Single individuals, university degree holders and current students are all more likely to buy.

An important point to note with the data set and model is that it is up to the call center managers to determine how to assign calls to sales reps. For example, they could give the low percentage sales to their better sales reps to boost the chances of a sale; or they could give their better sales reps the high percentage sales as a reward for good work. Due to the fact that the data set had no information regarding which agent was placing the calls we cannot come to any conclusions regarding the relationship between sales rep skill and success rate. While we have confidence in our predictions of which customers are most likely to make the purchase, the onus is still on the financial professional to seal the deal.

Works Cited

B. Cummings & B. Carhart. Empower Federal Credit Union. 2021. 2021 Annual Report. Retrieved on 2023-04-23 from https://www.empowerfcu.com/Empower/media/PDFs/EmpowerAnnualReport2021web.pdf

L.A. Bettencourt. Baylor University. 2012. Achieving Service Excellence in Real Estate: The Fundamental Tenents. Retrieved on 2023-04-23 from https://www.baylor.edu/business/kellercenter/doc.php/194525.pdf

Equitable. 2023. Entry Level Salaried Financial Professional (PCPG). Retrieved on 2023-04-23 from https://equitable.taleo.net/careersection/eqh_1/jobdetail.ftl?job=23000014&tz=GMT-04%3A00&tzname=America%2FNew_York

FINRA. 2023. Qualification Exams. Retrieved on 2023-04-23 from https://www.finra.org/registration-exams-ce/qualification-exams

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. Retrieved on 2023-04-10 from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#