Course Number: CSDA 1050

Course Name: Advanced Analytics Capstone

Professor Name: Hemant Sangwan

Assignment Title: A data-driven approach to predict term deposit subscriptions

Group Number: 1

Team Members: Fan Zhang, Surin Singh, Maryam Zulfiqar, Mohammadreza Arabkhani

TABLE OF CONTENTS

EXECUTIVE SUMMARY

KEY BUSINESS OBJECTIVES & FINAL RECOMMENDATIONS

Marketers will agree that direct mail and telephone methods pack a powerful personal punch than more indirect, general brand creation and promotion methods. Our mission is to empower businesses with valuable insights and analytics from their current marketing strategies' processes and technologies. Our approach uses a balanced assortment of techniques that include; filling gaps in useful data by information gathering and data structuring, finding overlooked and valuable insights from historical data, generating live insights on current campaigns, and influencing future initiatives by using advanced and holistic predictive models. Our product, specifically for financial institutions requiring an increase in customer term deposit subscriptions, is an analytical tool focused on improving direct marketing campaigns by telephone. This client's main objective was to address the need for understanding which potential customer best responds to direct marketing efforts. Machine learning models were built to identify whether a customer would subscribe to a term deposit based on a Portuguese banking dataset's parameters. The important parameters were job, education, duration of the last call and number of contacts performed before this campaign. Using logistic regression and random forest classifier, we could predict 90% and 91% accuracy, respectively. Our final recommendations to the client are to improve data acquisition on potential customers for direct contact methods, improve incentives to purchase on the first contact, introduce campaigns to highlight advantages of term deposits to individuals who have a job in administration or have a university degree.

PROJECT PROPOSAL

One of the most personal decisions people make involves their finances. Direct-marketing methods via telephone give consumers time to digest the information and make educated choices. Once an advanced understanding of their current and potential customers is obtained, a personalized direct-telephone campaign can introduce new products or encourage subscription to products previously denied by the customer. This project's scope is set to make predictions on customer subscriptions. Keeping in mind the limitations of economic stability at the time of data collection and other observations regarding the Portuguese bank dataset. By undergoing this challenge, we aim to:

- Pinpoint important factors that affect a customer's decision to subscribe to a term deposit

- Identify which group of customers is more likely to subscribe to a term deposit
- Reduce phone calls to our customers by our prediction

This will allow us to implement marketing strategies for our client.

METHODOLOGY

PRELIMINARY ANALYSIS, DATA MANIPULATION, DESCRIPTIVE ANALYSIS

The Bank Marketing dataset was obtained from the UCI Machine Learning Repository. This secondary dataset is a repository of 41,188 entries. It is related to direct marketing campaigns of a Portuguese banking institution based on phone calls. Often, more than one contact to the same client was required to assess if the bank term deposit would be ('yes') or not ('no') subscribed (Bank Marketing Data Set, n.d.). There are 21 attributes consisting of eleven categorical and ten numeric variables.

Average Customer:

- After duplicate rows were removed from the dataset, Table 1 in the appendix shows that age ranges from 17 to 98 years old. The duration of a call in the sample is anywhere from 0 to 4,918 seconds. An average individual in our sample is a 40-year-old and has a last contact duration of 258 seconds. This customer has been contacted three times during this campaign. The number of contacts performed before this campaign is zero, with a consumer price index of $94 monthly.

y:

- Now, we will look at the dataset closely. Figure 1 in the appendix shows a distribution of 89% entries labelled with 'no' labelled (0) who did not subscribe, and 11% entries labelled with 'yes' labelled (1) who did subscribe.

job:

- In Table 2, a cross-tabulation of the target variable 'y' with job shows that out of the 11% of the people who subscribed to a term deposit, 29% have an administration job, and 16% are employed as technicians, while 14% have a blue-collar job.

marital:

- A cross-tabulation between 'y' and marital yielded that out of the 11% who subscribed to a term deposit, 55% were married, and 35% of them were single, as seen in Table 3.

education:

- We did a cross-tabulation of 'y' with education to understand education's influence on the target variable (Table 4). Out of 11% of the people who did subscribe to a term deposit, 36% had a university degree, and 22% had a high school diploma.

default:

- In Table 5, a cross-tabulation of the target variable 'y' with default shows that out of the 11% who subscribed for a term deposit, 90% did not have a credit in default.

loan:

- A cross-tabulation between 'y' and loan in Table 6 showed that 83% out of the 11% who subscribed to a term deposit did not have a personal loan.

contact:

- As seen in Table 7, 83% of the people who subscribed for a term deposit used a cellular phone.

duration:

- A correlation analysis was performed on the dataset, as seen in Table 8, and a positive correlation of 0.4 was obtained with the target variable 'y' and the duration of the call with the client in seconds.
- The mean time of a call for people who subscribed to a term deposit is 553s. In contrast, the mean time for a call for people who did not subscribe is 221s, indicating the longer the call, the likelier they were to subscribe.

campaign:

- Figure 2 in the appendix shows that about half of the people who subscribed for a term deposit did so after the first contact was made with them.

previous:

- A positive correlation of 0.2 can be seen in Table 8 between the 'y' variable and the variable 'previous' (the number of contacts performed before this campaign and for this client).
- The mean of the variable 'previous' of those who did not subscribe for a term deposit is 0.132. The mean number for those who subscribed to a term deposit is 0.493. This leads us to believe the higher the number of contacts performed before this campaign, and with a client, the likelier they will subscribe for a term deposit.

poutcome:

- It can be seen in Table 9 out of the 11% who subscribed to a term deposit, 68% are new clients, and 19% were successful in the previous marketing campaign.

The variables emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m and nr.employed had strong correlations with our target variable 'y' as seen in Table 8, and we deem them important for analysis. The variables housing, month, day_of_week and pdays were removed from the analysis. The remaining variables, age, job, marital, education, default, loan, contact, duration, campaign, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m and nr.employed will be used in our Logistic Regression and Random Forest Classifier models to predict if people will subscribe to a term deposit.

MODEL BUILDING AND EVALUATION

Logistic Regression:

- To understand model performance, we divided the dataset into a 90% training set and a 10% test set while maintaining the ratio indicated in Figure 1 if the client subscribed or did not subscribe to a term deposit.
- The bank marketing dataset contains attributes such as age, job, marital (inputs), which are independent variables in our modelling and y (the output) is the dependent variable that is separated into a binary class – 'no' or 'yes' if the client subscribed for a term deposit. In logistic regression, we predict a category. If a client purchased a term deposit, 'no' or 'yes.' To implement the regression model, we ensure that the dataset is divided into 'no' or 'yes,' setting the target variable in this dichotomous manner.
- To interpret the metrics using our prediction case on y, for example, when our logistic regression model predicted 'y is going to be 'no' (0)', it is accurate 93% of the time, 'yes' (1) is predicted with 69% precision. In Recall, if the client didn't subscribe to a term deposit - 'no' (0) in the test set, our logistic regression model can identify it 98% of the time; if the client did subscribe - 'yes' (1) is predicted 40% of the time. The final percentage (weighted average) of the right prediction for if a client subscribed or did not subscribe, as seen in Table 10, is 90%.

Random Forest Classifier:

- We next implement the Random Forest Algorithm, which is excellent with a large dataset with higher dimensionality and handles unbalanced data. Random forest tries to minimize the overall error rate, so when we have an unbalanced dataset, the larger class will get a low error rate while the smaller class will have a larger error rate (Kho, 2019). We again split the dataset into 90% training and 10% testing in our random forest classifier. For greater accuracy, the random_state parameter controls the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node was set to 0 (3.2.4.3.1. sklearn.ensemble.RandomForestClassifier, n.d.). The number of trees (n_estimators) was set to 200 since if there are more trees, it will not allow overfitting trees in the model.
- The Random Forest Classifier generates a weighted average value, as seen in Table 11, that tells us how the model performed. In our prediction case for 'y,' for example, when our Random Forest model predicted 'y is going to be 'no' (0)', it is accurate 94% of the time, 'yes' (1) is predicted with 63% precision. In Recall, if there is 'y' that is 'no' in our test set, our Random Forest model can identify it 96% of the time; 'yes' is predicted 52% of the time. The final percentage (weighted average) of the right prediction for 'y' is 91%.

INSIGHTS AND SUMMARIZING RESULTS

Our data exploration has shown that job, marital, education, default and loan are the main factors influencing someone subscribing to a term deposit. The group of customers who are married and have administrative or technician job with a university degree or a high school diploma with no credit in default and no personal loan are likely to subscribe to a term deposit. Additionally, about 50% of the customers did not purchase the deposit in the first contact. People have a higher chance of subscribing if the call lasts longer, so it might be challenging to reduce phone calls. Obtaining customers' demographic information in advance before the phone call would help target customers and improve subscriptions. These are the market strategies that should be of high priority.

Regarding the metric of accuracy, the two models are similar with 90-91% accuracy in identifying people who will subscribe no or yes to a term deposit. Since we are not interested in identifying individuals who do not subscribe to a term deposit, this metric is not as valuable. The precision and especially recall metrics are most useful since we could focus on the ‘yes’ subscription predictions. Given that the Random Forest Classifier has the highest recall value of 52% and 63% precision, it is the favoured model to employ going forward. Also, we will improve our Random Forest model to predict 'yes' subscriptions with even more precision than 63%, and if there are 'yes' subscriptions, we hope to identify them more than 52% of the time.

Our focus on this project was on people who subscribed to a term deposit, but useful information can also be gathered from those who did not subscribe to a term deposit. Marketers can use this to find reasons for the decline and other products they would be interested in. A limitation of this project is the amount of deposit each customer subscribed to is not known. This could result in bias in our conclusion. For example, the recommended population deposit purchased could be lower than the amount of deposit purchased by the remaining group of customers. Acquiring relevant data would be a challenge for us.

CONCLUSION

The ability to predict individuals who will subscribe to a term deposit would be of interest to those companies which require this criterion to identify target recipients. This model can be employed by the financial institution bank, our primary stakeholder, to increase their net revenue. Market Executives may wish to utilize this model to craft the marketing campaign to meet bank targets carefully. The Customer Service Team will play an important part in activating and delivering the marketing campaign. This model has proven successful in its ability to predict potential customers. The choice model will be deployed with the usage of a user-friendly application for the analysts and marketing team to stream static data into the app, conduct standard analysis on the client demographics, and make educated predictions of whether a future campaign will be successful on that individual. Maintenance of the model would be undertaken monthly to validate the performance of the model. After a new campaign has been completed, the attributes' data will change; therefore, we will need to update our model.

REFERENCES

Bank Marketing Data Set. (n.d.). Retrieved from https://archive.ics.uci.edu/ml/datasets/Bank Marketing

Kho, J. (2019). Why Random Forest is My Favorite Machine Learning Model. Retrieved from https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

3.2.4.3.1. sklearn.ensemble.RandomForestClassifier. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

APPENDIX

	age	duration	campaign	pdays	previous	emp. var. rate	cons. price. idx	cons. conf. idx	euribor 3m	nr. employ ed
count	41164	41164	41164	41164	41164	41164	41164	41164	41164	41164
mean	40	258	3	962	0	0	94	-41	4	5167
std	10	259	3	187	0	2	1	5	2	72
min	17	0	1	0	0	-3	92	-51	1	4964
25%	32	102	1	999	0	-2	93	-43	1	5099
50%	38	180	2	999	0	1	94	-42	5	5191
75%	47	319	3	999	0	1	94	-36	5	5228
max	98	4918	56	999	7	1	95	-27	5	5228

Table 1: Statistical information on numerical attributes after duplicate entries were removed

job	admin.	blue-collar	entrep reneur	house maid	manag ement	retir ed	self-em ployed	services	stud ent	techn ician	unemp loyed	unkn own
y												
no	24.82	23.58	3.65	2.61	7.11	3.51	3.48	9.97	1.64	16.44	2.38	0.8
yes	29.11	13.76	2.67	2.29	7.07	9.36	3.21	6.96	5.93	15.74	3.10	0.8

Table 2: Cross-tabulation of the 'y' target variable with job

marital	divorced	married	single	unknown
y				
no	11.32	61.28	27.21	0.19
yes	10.26	54.55	34.93	0.26

Table 3: Cross-tabulation of the 'y' target variable with marital

education	basic. 4y	basic. 6y	basic. 9y	high. school	illiterate	professional. course	university. degree	unknown
y								
no	10.26	5.75	15.25	23.21	0.04	12.71	28.72	4.05
yes	9.23	4.05	10.20	22.23	0.09	12.83	35.96	5.41

Table 4: Cross-tabulation of the 'y' target variable with education

default	no	unknown	yes
y			
no	77.67	22.32	0.01
yes	90.45	9.55	0.00

Table 5: Cross-tabulation of the 'y' target variable with default

loan	no	unknown	yes
y			
no	82.35	2.42	15.24
yes	82.97	2.31	14.73

Table 6: Cross-tabulation of the 'y' target variable with loan

contact	cellular	telephone
y		
no	60.98	39.02
yes	83.03	16.97

Table 7: Cross-tabulation of the 'y' target variable with contact

	y
age	0.03
job	0.03
marital	0.05
education	0.06
default	-0.1
loan	-0.005
contact	-0.1
duration	0.4
campaign	-0.07
previous	0.2
poutcome	0.1
emp.var.rate	-0.3
cons.price.idx	-0.1
cons.conf.idx	0.05
euribor3m	-0.3
nr.employed	-0.4
y	1

Table 8: Correlation of all the variables with the target variable

poutcome	failure	nonexistent	success
y			
no	9.98	88.70	1.31
yes	13.04	67.68	19.28

Table 9: Cross-tabulation of the 'y' target variable with poutcome

	precision	recall	f1-score	support
0	0.93	0.98	0.95	3644
1	0.69	0.40	0.50	473
accuracy			0.91	4117
macro avg	0.81	0.69	0.73	4117
weighted avg	0.90	0.91	0.90	4117

Table 10: Prediction table for if a client subscribed or did not subscribe to a term deposit using logistic regression

	precision	recall	f1-score	support
0	0.94	0.96	0.95	3665
1	0.63	0.52	0.57	452
accuracy			0.91	4117
macro avg	0.78	0.74	0.76	4117
weighted avg	0.91	0.91	0.91	4117

Table 11: Prediction table for if a client subscribed or did not subscribe to a term deposit using random forest classifier

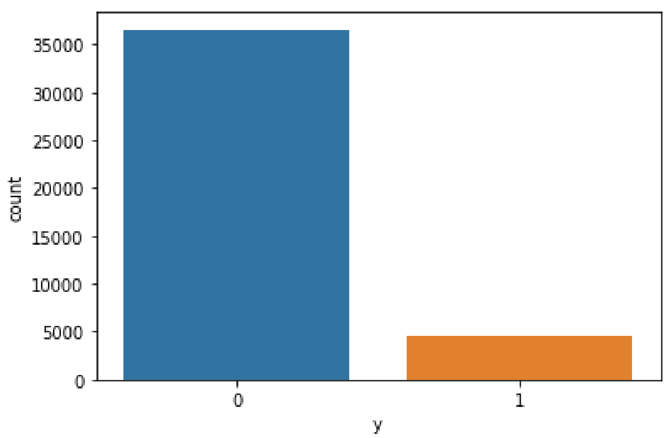

Figure 1: y categories, 0 & 1 for 'no' or 'yes' respectively

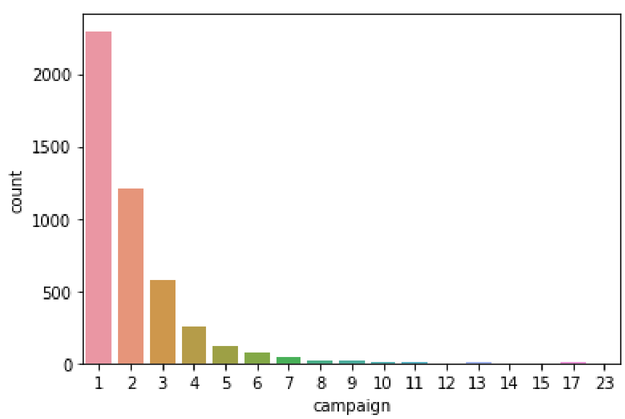

Figure 2: Count for the target variable 'y' when the outcome is 'yes' with campaign