**Project report**

**By Alden Maiyor**

**Process overview**

the following diagram shows the overall end-to-end process for defining, designing and delivering the capstone project.

Diagram

Description automatically generated

**Problem statement**

This project is looking the data obtained from a Portuguese banking institution which was running a telemarketing campaign to convince people to make a term deposit with them. A term deposit is a cash investment held at a financial institution where the money is held for a fixed amount of time for a agreed upon interest rate. The issue that the bank was facing was that running a telemarketing campaign costs a lot of money and resources and as such they want to ensure that they are able to maximize the effectiveness of their telemarketing campaign.

**Industry**

This project lies within the domain of marketing in the banking industry. As it stands right now, a major part of a banks revenue comes from term deposits. The way that a bank earns money from term deposits is when a customer makes a term deposit, their money is given to the bank for a fixed amount of time. The bank then uses this money as capital to provide loans to other people and businesses at a higher interest rate or to invest in various assets. One of the key concepts of this banking cycle is that the bank uses needs to have a sufficient sum of money to keep this cycle rolling and earn profits. This is why term deposits are so important to banks.

**Stakeholders**

The main stakeholders for this project are the management and decision makers for the Portuguese banking institution. The reason that they care about the results of this project is that they have invested a substantial amount of resources to initiate a telemarketing campaign. They did this because they believe that the increases profits from the increased amount of term deposits made would outweigh the costs of running the campaign. As such they would want to make the market targeting as efficient as possible to minimize wasted marketing money while maximizing the effectiveness of the campaign. The shareholders would expect that this project would improve their marketing campaign beyond just random calling targets.

**Business question**

The business question that needs to be answers is how can the bank increase the effectiveness of their telemarketing campaign to increase the amount of people that make a term deposit. By answering the question, the bank will be able to increase their profit margins.

The required level of accuracy of this model is be better than assuming everyone would make a term deposit, which is 11.6%. The implications of a false positive is that the bank wastes marketing money targeting the wrong person while the implication of a false negative is that the bank misses out on a potential customer.

**Data Question**

The business can be turns into two data questions:

1. Based on previous data, how can we identify and target the people who are most receptive towards this marketing campaign?
2. Can we segment our population base to more effective tailor marketing campaigns towards different customer segments?

The data required to answer this question is previously marketing data on customer characteristics as well as whether they decided to make a term deposit or not.

**Data**

I got this dataset from Kaggle, it was originally submitted to the UCI machine learning repository by the Portuguese banking institution themselves. The dataset contains 45211 entries and 15 variables. The data itself is relatively reliable as it was provided the banking institution itself. The quality of the data was quite good with not many missing values. The data was collected from the telemarketing already done by the banking from the years 2008-2010.

**Data Science process**

**Data analysis**

*Data wrangling*

Several columns were dropped due to not being relevant.   
several rows were dropped due to missing data.   
days since last contacted column was turned into a binary column ‘previously contacted?’

*Data pipeline*

This data wrangling step can easily be turned into a pipeline to take in new data and make it usable for the machine learning models.

***Exploratory Data Analysis***

***% of people that subscribed to a term deposit after the marketing call***

Chart, bar chart

Description automatically generated

Above we can see the break down of people who subscribed to a term deposit and those that did not. 88.3% of people did not while 11.6 % of people did.

***Education***

Chart, bar chart, waterfall chart

Description automatically generated

we can see a breakdown of people who subscribed to a term deposit broken down by their education. Here we can see that as levels of education rise, so does the likelihood of subscribing to a term deposit. This shows us that there is a positive correlation between levels of education and likelihood of subscribing to a term deposit.

***Previously contacted***

Chart, bar chart

Description automatically generated

We can see that those who were previously contacted were much more receptive to the marketing campaign and subscribed to a term deposit. Where if they were previously contacted, 23% of people subscribed to a term deposit while if they were not previously contacted this number drops to 9%

***Age***

Chart, line chart

Description automatically generated

We can see that the tail ends of the age distribution(so the younger and older people) have much higher term deposit rates than middle aged people

***Loan Status***

Chart, bar chart

Description automatically generated

Above we can see a graph showing the breakdown of whether subscribed to a term deposit based on their loan status, on the top we can see that only 8% of people who had a housing loan subscribed to a term deposit compared with 17% of people who did not have a housing loan.

On the bottom graph we can a similar trend with personal loans, whereby only 7% of people who had a personal loan subscribed to a term deposit compared with the 13% of people without a personal loan. This suggests that people without debt are more likely to make a term deposit.

***Jobs***

Chart, bar chart

Description automatically generated

Above , we can see the breakdown of jobs by whether they subscribed to a term deposit. We can see that retired people and students have a much higher likelihood of making a term deposit with 23 and 29% respectively, compared with jobs such as entrepreneur with 8% and blue collar jobs with 7%

***Duration of Phone call***

A picture containing table

Description automatically generated

Above we can see the duration of phone call for those that did make a term deposit and those that did not make a term deposit. We can see that the average phone call duration is much longer for those that did make a term deposit compared with those that did not.

**Data Modelling**

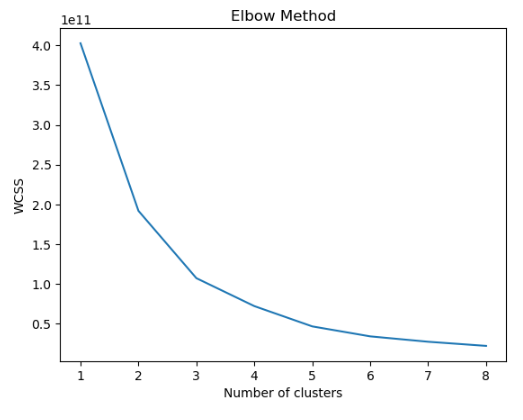
For this project I wanted to make two machine learning models, one model for clustering the data and one predictive model.

**Clustering**

For the clustering model I chose to use k-prototype, which is a machine learning algorithm which takes in both categorical and numerical data. I did this so that it would be easier to interpret the characteristics of each cluster.

For this model I chose to use all of the features from the dataset. I did this as it would allow me to get the most accurate picture of each cluster.

I ran this model with the cluster numbers ranging from 2-9. I then plotted the cost value for each model against their cluster number. Using the elbow method I saw that the optimal number of clusters was 2. This plot can be seen below.



**Results**

***Cluster 1***

* more likely to make a term deposit compared to cluster 2 (15.6%)
* slightly older (average age = 43)
* on average better educated
* much higher average bank balance ($ 12020)
* more likely to be in management or retired than cluster 2
* less likely to have an active personal loan (6%)
* less likely to have an active housing loan (45.3%)

***Cluster 2***

* less likely to make a term deposit(11.5%)
* slightly younger (average age = 40)
* less educated on average
* much lower average bank balance ($ 896)
* more likely to be working in blue-collar, service or admin jobs than cluster 1
* more likely to have an active personal loan (16.9%)
* more likely to have an active housing loan (56.7%)

**Feature engineering**

Before putting the data into models, I performed SMOTE to deal with the imbalanced nature of the data. SMOTE is a technique used to balanced imbalanced dataset by oversampling the minority group.

**Predictive modelling**

I used all of the present in the features present in the dataset to create the model.

*Feature importance (taken from Random Forests feature importance)*

The features that were most important in this model was Duration of phone call, housing loan, balance and age.

Chart, funnel chart

Description automatically generated

*Models tested*

For this project I used and tested 5 models, Random forest, Stacking with logistic regression, support vector classifier and decision trees, Stacking with KNN, Random Forests, and Naïve bayes, Bagging and Adaboost.

*Model training length*

Random Forests **–** 41s

Stacking (Logistic Regression, SVC, DT) **–** 3m 3s

Stacking (KNN, RF, Naïve Bayes) **–** 35s

Bagging **–** 1.6s

Adaboosting**–** 1.6s

*Model Performance metrics*

For predictive modelling I chose f1 score as my evaluation metric. The f1 score is a mix of both precision and recall. Precision is the ratio of true positive predictions vs the total predicted positives while recall is the true positive predictions compared to the true total positives. I chose f1 score as a balance between precision and recall to align with the goal of minimizing wasted marketing money (precision) while not missing out on potential customers (recall).

*Model selection*

The best model was stacking(KNN, RF, Naïve bayes)

Table

Description automatically generated

**Outcomes**

*Exploratory data analysis insights*

* Customers are more likely to make a term deposit if they had been contacted before
* Younger people and older people are more receptive to marketing than middle aged people
* The longer the duration of the phone call, the more likely a person is to make a term deposit
* If a person no housing loans or personal loans, they are more likely to make a term deposit
* Students and retired people are the most likely occupations to make a term deposit

*Clustering*

* two clusters
* Main defining characteristic are their bank balance, their job and their loan status

*Predictive modelling*

* We chose f1 score as our metric to get a balance of precision and recall
* Our best model was a stacking model using KNN, RF and GNB
* We managed to obtain a 89% accuracy with a 0.53 F1 score

**Implementation**

Both the predictive model and the clustering model are likely to improve the effectiveness of the marketing campaign.

**Data Answer**

**Question 1:**

Based on previously data, how can we identify and target the people who are most receptive towards this marketing campaign?

**Answer:**

Yes, we were able to identify people who are more likely to be receptive towards this marketing campaign with an accuracy rate of 89% and a F1 score of 0.53 for our best model.

**Question 2:**

Can we segment our population base to more effective tailor marketing campaigns towards different customer segments?

**Answer:**

Yes, we were able to segment our population base into two segments separated mainly based on their bank balance, job and loan status.

**Business Answer**

**Question:**

How can we increase the effectiveness of our telemarketing campaign to increase the amount of people that make a term deposit.

**Answer:**

Yes we are able to increase the effectiveness of our telemarketing campaign based on our predictive modelling. In addition to this, we are also able to use our clustering model increase our likelihood of success by tailoring various term deposit plans for different customer segments.

**Confidence in the models**

We can be relatively sure that using the predictive model targeting would improve our base rate of calling random people.

**Response to stakeholders**

Overall, this project shows the importance of using data when making business decisions. One recommendation for the stakeholders is to create different loan plans to target the different cluster, for example people in cluster 2 on average have a lower bank balance than cluster 1. By creating a term deposit plan that has a shorter commitment term, it might increase the chances that people in cluster 2 make a term deposit as their money is not locked away for as long.

**End – to – end solution**

The overall end to end solution is to provide this model to the telemarketing team. The team then can then input the data about potential customers and see if they are likely to make a term deposit or not. They can then prioritize calling those who have a high likelihood of making a term deposit.

**References**

Data cleaning and EDA –Project EDA.ipynb

Data Modelling –project modelling.ipynb

Data source - <https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets>

**Resources used in project:**

matplotlib   
pandas  
seaborn  
plotly

**Modelling resources**

Kmodes - https://kprototypes.readthedocs.io/en/latest/  
scikit-learn - https://scikit-learn.org/stable/about.html#citing-scikit-learn  
Imblearn - https://scikit-learn.org/stable/about.html#citing-scikit-learn  
mlxtend - http://rasbt.github.io/mlxtend/