**CAMPAIGN ANALYSIS**

**BANK MARKETING DATA**

**Table of Contents**

[**Executive Summary** 1](#_Toc73980011)

[**Introduction** 1](#_Toc73980012)

[**Description of the dataset** 2](#_Toc73980013)

[Original Data 2](#_Toc73980014)

[Data Pre-processing 3](#_Toc73980015)

[Account holders and Campaign Profile 3](#_Toc73980016)

[**Methodology** 4](#_Toc73980017)

[Market Basket Analysis(MBA) 4](#_Toc73980018)

[Clustering 5](#_Toc73980019)

[**Results** 6](#_Toc73980020)

[Market Basket Analysis(MBA) 6](#_Toc73980021)

[Clustering 8](#_Toc73980022)

[**Conclusion** 10](#_Toc73980023)

[**Future Analysis** 11](#_Toc73980024)

[**Appendix** 11](#_Toc73980025)

[Appendix I – Transactional variables 11](#_Toc73980026)

**Table of Tables**

[Table 1 Variable Description of Original Data 2](#_Toc60493356)

[Table 2 Subsets of the data 3](#_Toc60493357)

# **Executive Summary**

Finance and banking are one of the most extensive and extremely competitive markets. Any organisation going face-to-face with the big players need to fortify that they understand precisely how buyers like to interact with their sales and marketing processes. Customers today lean on both online as well as offline means to shop before deciding. Almost on every occasion, an offline phenomenon such as making a phone call or visiting a branch is a positive indicator of a possible conversion. In this report, a bank marketing dataset of a Portuguese bank is selected where the marketing campaigns were based on phone calls. The report is based on the Bank Marketing Data from [**UCI ML Repository**](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing).

Initially, the dataset is split into training data, testing data and evaluation data. The data is imbalanced in terms of outcome category. SMOTE(Synthetic Minority Over-sampling TEchnique) was performed to handle the imbalance in the dataset. Various pre-processing techniques such as scaling, creating dummy variables were performed to make data suitable for modelling. The outcome of the campaign is predicted using various classification and prediction models. < ADD HERE >

# **Introduction**

Banks and financial institutions exist to offer financial services to people and to make huge profits. Having said that, banks also devote remarkable resources and business intellect to gain capital. One of the most common ways for banks to do this is to engage in direct marketing campaigns like phone calls and face-to-face meetings to promote and provide services. Phone calls, i.e. Telemarketing is a conventional marketing technique that helps to soar profits for any given business. Moreover, it also offers a more interactive and personal medium of sale service which can initiate an instant rapport with the prospective customers. Furthermore, telemarketing can help an organization to reach out more customers than with in-person or by going door-to-door and it can benefit a company to sell a product to both existing and new customers. For banking industry, telemarketing can be useful to communicate with large number of customers and offer them with all the services that they have for them. This may include information about loans, term deposits, mortgages, Overdraft facility, Credit cards etc.

For this project, a data set of a Portuguese Bank direct marketing campaign is used. This dataset is obtained from the [**UCI ML Repository**](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The primary objective of this project is to find the best model to predict whether a customer will subscribe for a term deposit or not using various classification techniques. Our secondary objective is to determine what factors in this data set would contribute the most for the sale of term deposits to the potential customers. The target users for this project are the marketing team of a banking institution who are looking to increase their inflow of cash deposits. The following sections of this report includes the description of the dataset in detail, all the methods (classification techniques) that has been applied on the dataset to get the results and eventually the best one is described thoroughly. Moreover, the result for the best model is presented followed by conclusion which summarises the most important findings and the scope of future research is suggested.

# **Description of the dataset**

## Original Data

The original dataset used in this analysis was provided in the Comma Separated Values(.csv) format. The spreadsheet consists of:

* 45211 observations(rows)
* 17 variables(columns)
  + 4 Demographic Variables – age, job, marital, education.
  + 4 Variables representing Economical and Socio-Economical standing of the customers – balance, default, housing, loan.
  + 8 Variables representing campaign information – contact, day, month, duration, campaign, pdays, previous, poutcome.
  + 1 Target variable – y.

The table below offers a brief description of each variable(For the detailed table go to Appendix)

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| ***age*** | Continuous | Age of the customer in years |
| ***job*** | Nominal | Type of job: “blue-collar”, “management”, “technician”, “admin”, “services”, “retired”, “self-employed”, “entrepreneur”, “unemployed”, “housemaid”, “student”, and “unknown” |
| ***marital*** | Nominal | Marital Status: “Married”, “Divorced”, “Single” |
| ***education*** | Nominal | Level of Education: “Primary”, “Secondary”, “Tertiary”, “Unknown”. |
| ***default*** | Binary | Has credit in default? : “No”, “Yes” |
| ***balance*** | Continuous | Average yearly average, in euros. |
| ***housing*** | Binary | Has housing loan? : “No, “Yes” |
| ***loan*** | Binary | Has personal loan? : “No”, “yes” |
| ***contact*** | Nominal | Communication type : “telephone”, “cellular”, “unknown” |
| **day** | Discrete | Last contact day of the month |
| **month** | Nominal | Last contact month of the year |
| **duration** | Continuous | Last contact duration, in seconds |
| **campaign** | Continuous | Number of contacts performed during this campaign and for this client(Includes last contact) |
| **pdays** | Continuous | Number of days that passed by after the client was last contacted from a previous campaign(-1 means client was not previously contacted.) |
| **previous** | continuous | Number of contacts performed before this campaign and for this client |
| **poutcome** | Nominal | Outcome of the previous marketing campaign : “unknown”, “other”, “failure”, “success” |

Table 1 Variable Description of Original Data

## Data Pre-processing

The owner of the data mentioned that data is cleaned on the [**UCI ML Repository**](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing).And after careful analysis the same conclusion was reached. One observation was removed from the data for having extreme value for “previous” variable.

Later in the analysis the dummy variables were created for categorical variables having more than two categories. Also, binary variables were transformed using label encoders. The numerical variables were not on equal footing and the scaling was performed to adjust the variance of the variables in order to nullify the effect of the larger variables on the models. The number of features/predictors increased from 17 to 53 after initial data pre-processing. The data was split into three different subsets for training(80%), validation(10%) and testing(10%) randomly.

The outcome variable in the training data was imbalanced: 87% of the observation has “No” value for the outcome variable. This imbalance could affect the model training as model would be able to learn more about the “No” observation compared to “Yes” observations. This problem can be targeted by using over-sampling or under-sampling. The over-sampling was performed using SMOTE to have more data for the training of the model. The following table has the details of all three subsets after performing the SMOTE.

|  |  |  |
| --- | --- | --- |
| **Subset** | **Observations, Features // % of Yes // % of No** | |
| **Before Pre-processing** | **After Pre-processing** |
| ***Training*** | (36168, 17) // 13 // 87 | (63758, 42) // 50 // 50 |
| ***Validation*** | (4521, 17) // 11 // 89 | (4521, 42) // 11 // 89 |
| ***Testing*** | (4522, 17) // 11 // 89 | (4522, 42) // 11 // 89 |

Table 2 Subsets of the data

## Account holders and Campaign Profile

The age of the account holders of Portuguese Bank is between 18 to 95 years with an average age of 40 years. Most of the account holders are middle-aged ranging from 33 to 48 years. Only 2.63% of the total account holders are senior citizens(age > 60 years) with the highest average annual balance of nearly 2600$. On the other hand the young adults(18 <= age <= 32 ) have the lowest average annual balance of 1000$.

Most of the account holders have blue-collar jobs or jobs in management. Moreover, very few housemaids and students have bank accounts in the bank. Furthermore, most of the account holders are married and have secondary education followed by account holders who have tertiary education.

Nearly 2% of the account holders failed to repay a debt back to the bank. Most of the high balance account holders are in management jobs or retired and had tertiary education. Given the high balance, most of the account holders do not have default credit, personal loans, and home loans.

Almost 9% of the account holders who have negative balance are working middle-aged adults. In contrast to the high rollers, most of the negative balance holders have completed the secondary level of education. Moreover, compared to high rollers 72% of these bank holders have house loans. Also, 87% of these account holders were never previously contacted regarding the campaign which could be explained by the negative balance.

Around 64% of the account holders were contacted on their cellular devices regarding the term deposit which makes sense as time when data was collected mobile phones were more prominent. The data also shows that the account holders who had cellular devices are younger(Median age of 38 years) compared to the account holders who had telephones(Median age of 47 years).

Contacts regarding the term deposits were generally made in the third week of the month. The least amount of call was made on the first day of the month. There are low number of calls on 31st day of the month which is understandable since not all months have 31 days. Also, Majority of the contacts were made in the middle of the year(May - Aug). December has the least number of the contacts that can be explained by the festive season(employees not working/customers don't have time to talk to the bank employees/last month of the fiscal year). In addition to that, Most common duration for the contact is 3 minutes(180 seconds). The contacts with longer duration were usually made with middle-aged and old people.

Moreover, on average account holders were contacted 2 times. The high number of contacts were made to the account holder working a blue-collar job, which could be explained by the nature of their job. Lastly, the success rate for the previous campaign is quite low(less than half of failure rate).

# **Data Modeling**

The Primary objective of our project is to be able to predict whether an account holder will subscribe to the term deposit or not. The data modeling was done by using two different languages: Python and R.

## Python

The analysis done in Python consists of different supervised learning models capable of handling binary outcome variable such as Logistic Regression, Decision Tree, Random Forest, Extra Trees, KNN, SVM, MLP and QDA. The search of the best hyperparameters was done using grid search algorithm which was performed by manually selecting values for hyperparameters and then looping through the combinations of these values.

## R

# Since the response variable subscribed(whether user subscribed to a term plan or not) is dichotomous, binomial response distribution can be used to model it. Also, few of the covariates are skewed in distribution since GLM doesn’t require the covariates to be normal so the skewness in covariates is acceptable. Collectively, the decision was made in the favour of Binomial GLM.

# ***[Note – Data Pre-processing was carried out again to accommodate data in R which can be found*** [here](https://github.com/NandishRPatel/DATA-SCIENCE-PROJECTS/blob/master/BANK%20MARKETING%20DATA/analysis_in_R_Notebook.Rmd)***]***

# **Results**

## Python

The best set of hyperparameters values were chosen for each modeling technique based on the performance metrics such as accuracy of the model on training and validation set, Individual class accuracies. The performance of these models can be seen in the *table 3*.

## R

While conducting analysis using Binomial GLM, variables “balance” and “age” were found to be insignificant so they were dropped from the further modelling. The final model in R had a great predictive power as it can be seen in *table 3*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Class 0 Accuracy** | **Class 1 Accuracy** | **Overall Accuracy**  **(Train)** | **Overall Accuracy**  **(Validation)** |
| *Logistic Regression* | 0.847 | 0.787 | 0.857 |  |
| *Decision Tree* |  |  |  |  |
| *Random Forest* |  |  |  |  |
| *Extra Trees* |  |  |  |  |
| *KNN – Uniform Weights* |  |  |  |  |
| *SVM – Poly* |  |  |  |  |
| *MLP* |  |  |  |  |
| *QDA* |  |  |  |  |
| *GLM in R* |  |  |  |  |

# **Conclusion**

* Data Quality – Coles data set raised concerns about data quality – the age, Postcode variable and the sequential invalid entries in other variables like pmethod, nchildren, homeown.
* Despite the data quality issues, it has provided great insights. The key results and recommendations are:
* Coles should put newly marketed products near the shelf of bread, milk, cereal, and banana So that customers are exposed to the new products. Coles should put these products to the end of the aisle So that customers have to walk more in the stores, in turn, they will be tempted to buy other products they see on their way to the most purchased product aisle.
* Baby Food should be displayed near the shelves of Napies, Fruit, Chocolate and Olive Oil.
* Coles could probably increase sales by advertising/marketing householCleaners, coffee, frozenmeal, TomatoSauce with Vegetables(if possible stock them near).
* Customers who bought Vegetables & Fish and Fruit and Fish are 2.2 and 2.0 times more likely to buy householCleaner respectively. Coles should spatially separate fish, vegetables, fruit, and householCleaners for greater travel distance so that customers will be encouraged to purchase other products.
* Coles should start marketing campaigns for family or house related products towards young customers(from cluster 3) as the majority of them are home owners and parents. These groups of customers are on a tighter budget and they will respond better to deals and discounts on the branded products that coles has to offer.
* On the other hand, luxury products should be targeted to customers(from cluster 1) who do not own a house and are females independent of their income and age.

# **Future Analysis**

* Recording customer’s postcode adequately would help identify patterns by states or suburbs.
* Recording data of as many customers and as many transactions as coles can by using various methods like using a points card like Flybuys. More data and correct data would help coles get meaningful insights from the data.
* Quantity of items purchased alongside the value would be more helpful in assessing customer’s preferences and spending power.
* MBA can be carried out on more generalized product categories for less-frequently purchased products, for example, put all the frozen items in one category that would increase the support of the products.

# **Appendix**

## Appendix I – Transactional variables