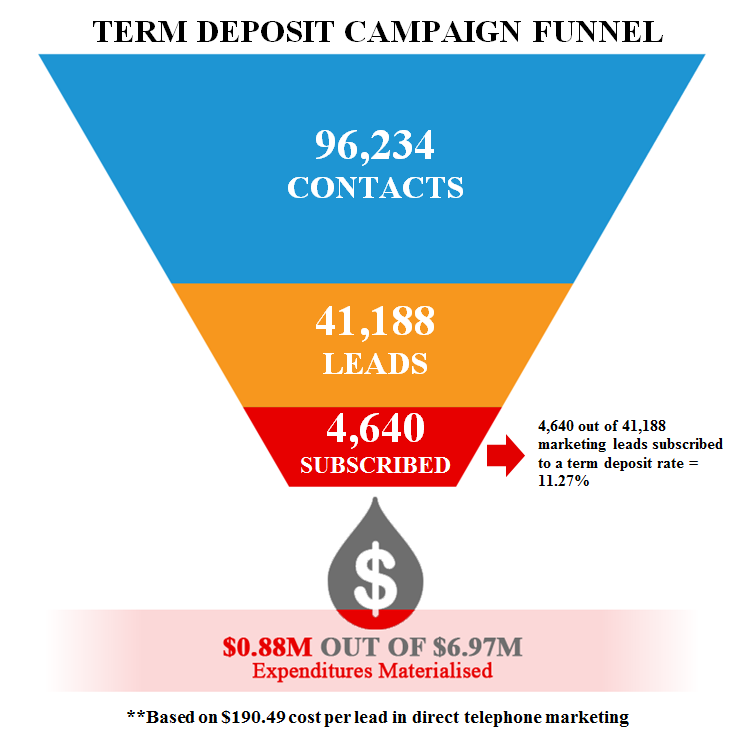
**COMP5310: Binary Classification Model to Predict Term Deposit Campaign Subscription**

Emil Laurence Pastor

epas6415@uni.sydney.edu.au

The University of Sydney

**Executive Summary**



According to the *Gartner 2016-2017 CMO Spend Survey,* marketing budgets continued their steady ascent in 2016, climbing to 12% of company revenue(Pemberton, 2017). Consequently, this portion of the budget translates to millions of dollars spent on advertisements and promotions to improve overall branding, customer engagement and to drive more sales. Do the millions of dollars in marketing expenditure can really bring the value to the company and their customers? A marketing data suggest that the average lead to sales conversion rate in upselling a product or service to an existing customer is just 11.27% in a typical direct marketing campaign. Furthermore, this means that around 88.73% of the expenditures didn’t translate to actual sales. Moreover, the measly lead to sales conversion rate and unnecessary operational expenditures is the result of (1) poor market segmentation, (2) channel selection, (3) customers receiving several campaigns over a period and (4) not taking advantage of the customer and transactional data. In line with this, companies should re-evaluate their current marketing strategy on how they can best use their existing assets to improve their KPIs.

Figure 1 Term Deposit Campaign Funnel

To overcome this challenge, companies like banks should monetise on the immense amount of customer and transactional data available in their systems to improve their marketing strategy to a data driven and highly targeted analytical campaign. Having said that, this project will take advantage years’ worth of dataset alongside its attributes to build a binary classification model that will predict if the lead will subscribe to a product or service. Furthermore, this project includes evaluation of the most appropriate classification algorithm and data attributes to get the right blend of algorithm and features (data attributes). Thus, this approach will translate to an increase in the lead to sales conversion rate, increase operational efficiencies, improved customer segmentation and better customer satisfaction through the reduction of unnecessary contact while learning more about your customer.

**I. Problem Statement**

C

ompanies like banks spend millions of dollars on marketing which includes advertisements, promotions, and campaigns. Consequently, companies should re-evaluate their expenditure and strategy alongside with the result marketing KPIs like lead to sales conversion rate. In line with this, the data gathered between May 2008 to November 2010 from a European bank show that telemarketers contacted the leads 2.57 times to have 4,640 out of 41,188 marketing leads subscribed to a term deposit direct marketing campaign. Furthermore, this result in a low lead to sales conversion rate of 11.27% and $0.88M out of $6.97M expenditure translated into sales using $190.49 cost per lead in direct telephone marketing (Lohrey, 2013). Thus, this scenario creates a long-term and adverse problem resulting to poor customer engagement and satisfaction, decrease in operational efficiencies and low marketing KPIs like lead to sales conversion rate. A successful solution would be to build a binary classification model that will label customers who will be campaign leads and be invited by the telemarketers to subscribe to a term deposit. To be able to build the model, this project aims to answer the following challenges:

1. Determine the suitable features (attributes) to build the classification model.
2. Evaluate the most appropriate machine learning algorithm to classify whether the lead will subscribe to the term deposit.

**II. Data Set**

**1. General Information**

In this project, the Bank Marketing Data from *kaggle.com* and the *UCI Machine Learning repository* was used as the data sets for data exploration and analysis. The zip file consists of two *csv* files: (1) *bank-additional-full.csv* contains 41,188 records will be used as training data set for the next stage of the project and (2) *bank-additional.csv* contains 4,119 records will be used as the test data. Furthermore, the data set consists of 1 output variable (campaign outcome *y*) and 20 categorical and numeric features can be grouped into 4 categories: Customer Demographics, Customer Bank Information, Current and Historical Marketing Activity and Socio-Economic Measures. The mix of different data categories provides a broader perspective in understanding the marketing bank and how does each feature contributes on the overall campaign outcome.

Table 1 Features and Output Data

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Data Type** |
| **Customer Demographics** | | |
| Age | Customer Age | Numeric |
| Job | Job type (Administration, Blue Collar, Management, etc.) | Categorical |
| Marital Status | Marital Status (Single, Divorced, Married) | Categorical |
| Educational Attainment | Highest Educational Attainment (High School, University Degree) | Categorical |
| **Customer Bank Information** | | |
| Credit Default Flag | Flag if the customer has credit in default | Categorical |
| Housing Loan Flag | Flag if the customer has housing loan | Categorical |
| Personal Loan Flag | Flag if the customer has personal loan | Categorical |
| **Historical Marketing Activity** | | |
| Contact Channel Type | Marketing contact channel type | Categorical |
| Contact Month | Last contact month | Categorical |
| Contact Day of Week | Last contact day | Categorical |
| Last Contact Duration | Last contact duration in seconds | Numeric |
| Campaign Contacts | Number of contacts performed during this campaign and for this customer | Numeric |
| Previous Campaign Last Contact | Number of days that passed by after the customer was last contacted from a previous campaign | Numeric |
| Previous Campaign Contact | Number of contacts performed before this campaign | Numeric |
| Previous Campaign Outcome | Outcome of the previous marketing campaign | Categorical |
| **Socio-Economic Measures** | | |
| Quarterly Employment Variation Rate | Employment variation rate | Numeric |
| Monthly Consumer Price Index | Consumer price index | Numeric |
| Monthly Consumer Confidence Index | Consumer confidence index | Numeric |
| Daily Euribor 3Month Rat | Euribor 3-month rate | Numeric |
| Quarterly Number of Employees | Number of employees | Numeric |
| **Current Campaign Result** | | |
| Campaign Outcome | Term Deposit campaign outcome | Categorical |

In data management perspective, the data has been modelled into Star Schema to conform and organise the data set before performing further data deep dives and analysis as shown in *Figure 2*. The categorical variables are converted into a dimension which consists of source key, numerical target for labelling and target categorical value.



Figure 2 Logical Data Model

**2. Data Ingestion, Cleansing and Transformation**

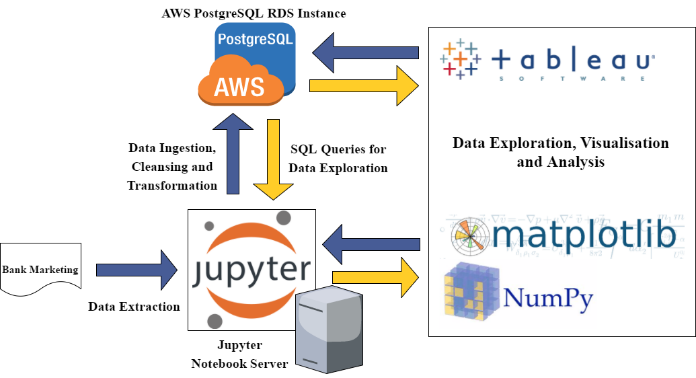


Figure 3 Data Architecture

The Data Architecture as shown in *Figure 3* provides a high-level strategy and tools used for this stage of the project. Having said that, *Jupyter Notebook* is used for compiling and execution of *Python 3.6* codes with the use of *matplotlib* for visualisation, and *numpy* for statistical and numerical calculations, the *AWS PostgreSQL* instance was created to stage, cleanse and transform data to fit into the data model and Tableau was used for most of the data visualisation activities.

The extracted training data set was loaded into a staging table *STG\_BANK\_TRAINING* before performing any cleansing and transformation activities. Consequently, the staging table data were cleansed and transformed to be loaded into the *FACT\_BANK\_TRAINING* table and dimension tables. In addition, *FACT\_BANK\_TRAINING\_KEY* was loaded with categorical data that was labelled and transformed into its numerical equivalent in preparation as input data in the chosen Machine Learning algorithm. Furthermore, the mapping sheet found in *Appendix A* of this paper provides the cleansing and transformation specification used prior to loading into the fact table.



Figure 4 Data Flow from Staging to Star Schema Tables

As part of the cleansing and the transformation activity as shown in *Figure 5*, two columns were created, namely: Record Key to give a unique id for each record and Age Bucket to group the Age data into buckets.



Figure 5 Cleansing and Transformation Process

**3. Data Exploration**

The next step is to explore the data set after performing data cleansing and transformation process. The exploration activity gives insight and refinement of understanding of the data that will be used in the succeeding process towards building the classification model. The succeeding part of this section provides different insights through data visualisation.

The descriptive statistical measures were calculated as part of data calculation to determine the centrality and dispersion of data. The individual plots can be found in the *IPython Notebook*.

Table 2 Descriptive Statistical Measures for some features

|  |  |
| --- | --- |
| **Attribute Name** | **Statistical Measures** |
| Age | Mode: 31  Mean: 40.02  Median: 38.00  Standard deviation: 10.42 |
| Job | Mode: Administration |
| Marital Status | Mode: Married |
| Educational Attainment | Mode: University Degree |
| Credit Default Flag | Mode: No |
| Housing Loan Flag | Mode: Yes |
| Personal Loan Flag | Mode: No |
| Contact Channel Type | Mode: Cellular |
| Contact Month | Mode: May |
| Contact Day of Week | Mode: Thursday |
| Last Contact Duration | Mode: 85  Mean: 258.29  Median: 180.00  Standard deviation: 259.28 |
| Campaign Contacts | Mode: 1  Mean: 2.57  Median: 2.00  Standard deviation: 2.77 |
| Previous Campaign Last Contact | Mode: 0  Mean: 0.22  Median: 0.00  Standard deviation: 1.35 |
| Previous Campaign Contact | Mode: 0  Mean: 0.17  Median: 0.00  Standard deviation: 0.49 |
| Previous Campaign Outcome | Mode: Unknown |

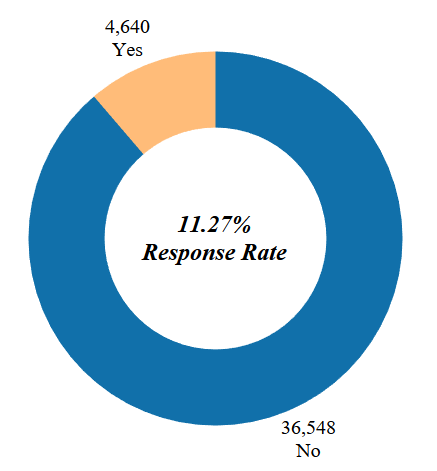


Figure 6 Campaign Outcome Distribution

*Figure 6* shows that the campaign outcome of the training data set provides 11.27% response rate, which means that this project will be dealing with imbalanced classification because 88.73% belongs to a single class. Thus, this implies that there is a necessity to use methods of treating imbalanced data set prior to building the classification model.

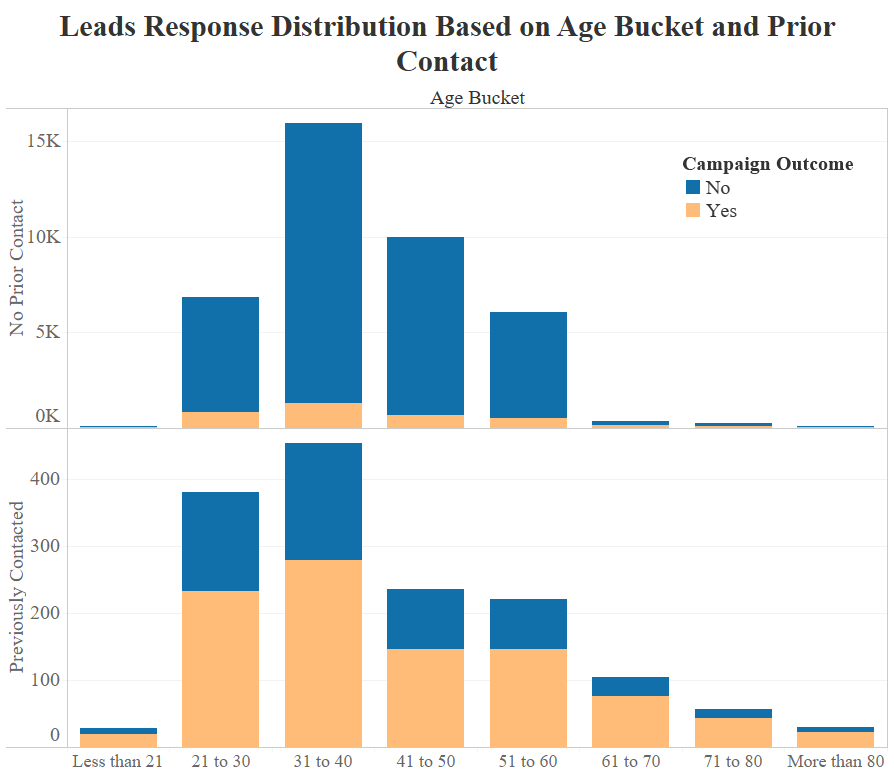


Figure 7

The data also suggest that there has been a considerable improvement in response rate for those customers who has been previously contacted compare to customers with no prior marketing contact. In addition, most of the customers contacted are relatively young compare the general population of the data. Furthermore, *Figure 8* shows that the campaign outcome resulted favourably for those customers who were contacted fewer times than those who had spammed through several calls from telemarketers.

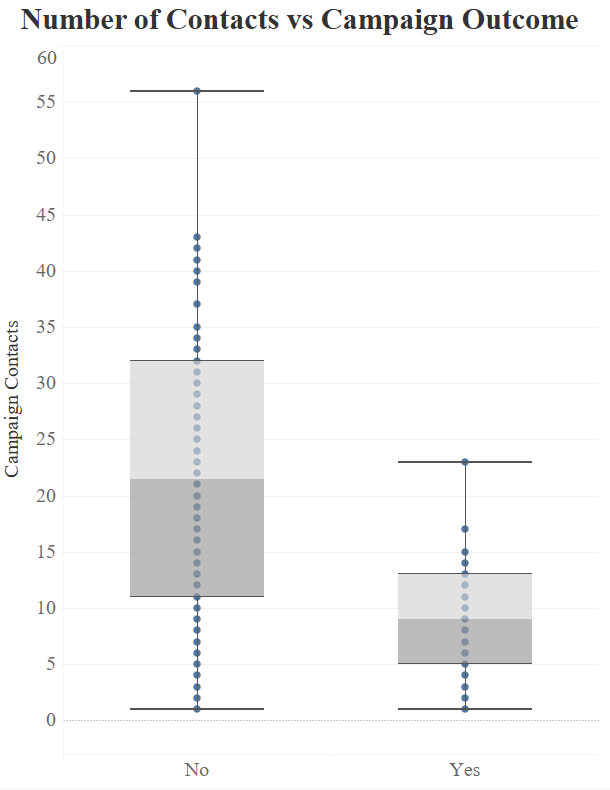


Figure 8

The response distribution across month of contact as shown in *Figure 9* was also explored and found that the number of leads subscribed to the term deposit had been less than 1,000 regardless of the amount of leads generated during that month. Having said that, the lead to sales conversion rate between May and August including November was less than the average compared to the rest of the year.

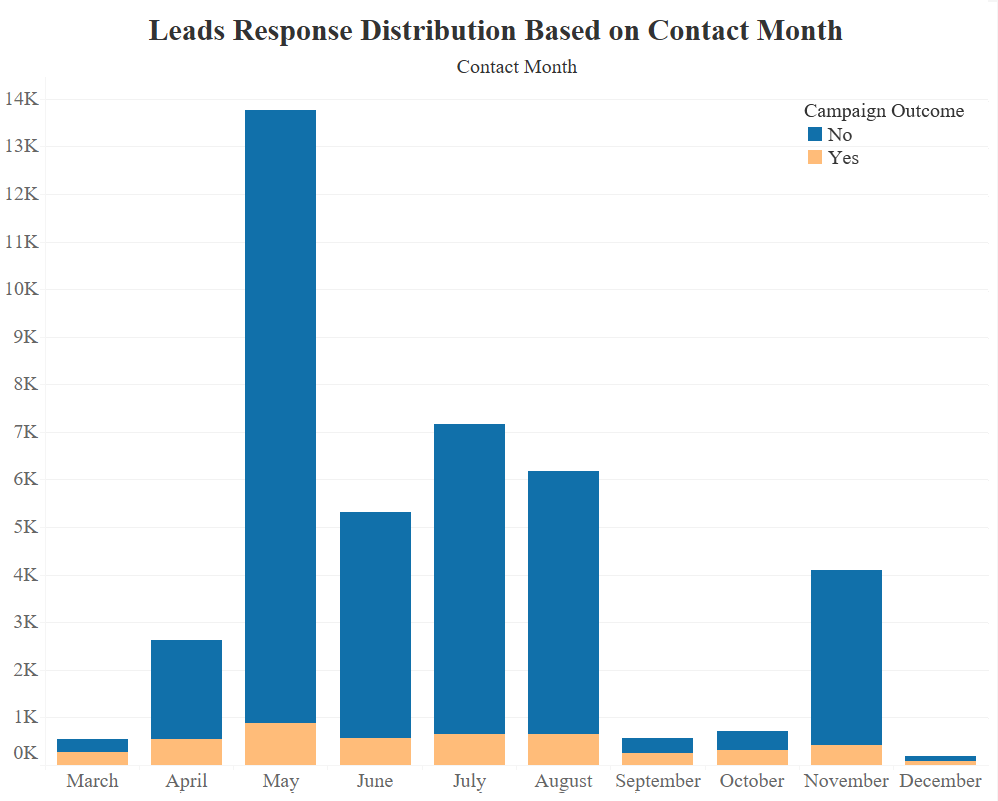


Figure 9

In addition, it was observed that better response rate was achieved between Tuesday to Thursday compare to Monday and Friday, though there is no considerable difference in terms of response rate as shown in *Figure 10*.

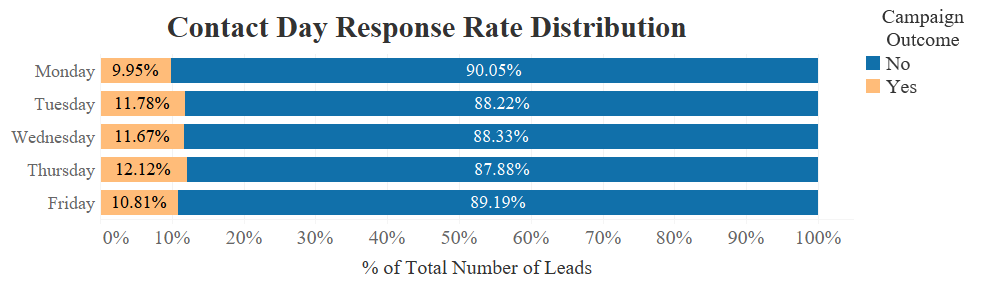


Figure 10

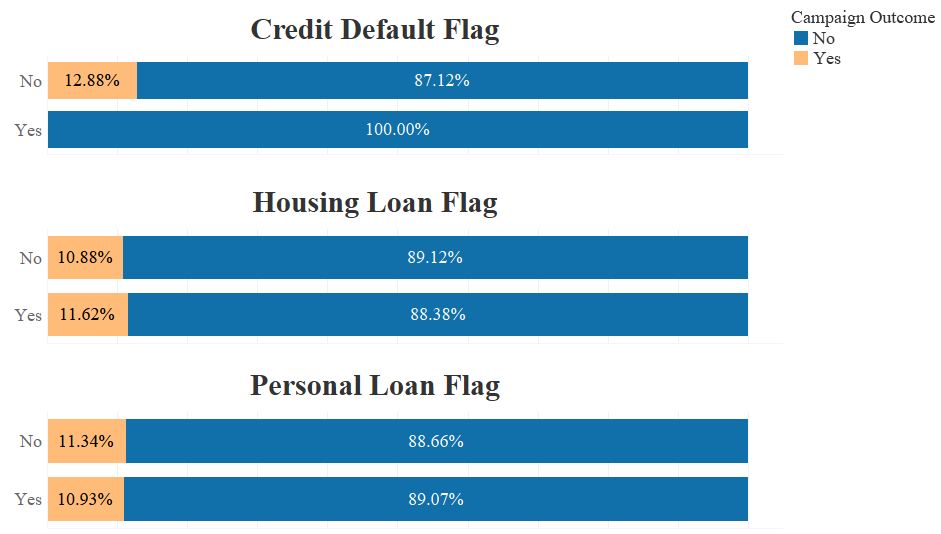


Figure 11 Response Distribution with Customer Bank Information Flags

It was observed that the response distribution is similar regardless if the customer has a personal or a housing loan. In contrast, customers having credit default did not subscribe to term deposit. Thus, it is suggested to exclude customer with credit default from the pool of leads.



Figure 12 Contact Channel and Campaign Outcome

The mobile phone was the primary contact channel used to reach the leads. In addition, the bulk of the leads who agreed to subscribe to the term deposit was contacted through a mobile phone. Thus, it is recommended to pick customers with mobile phone over those with only a telephone available.

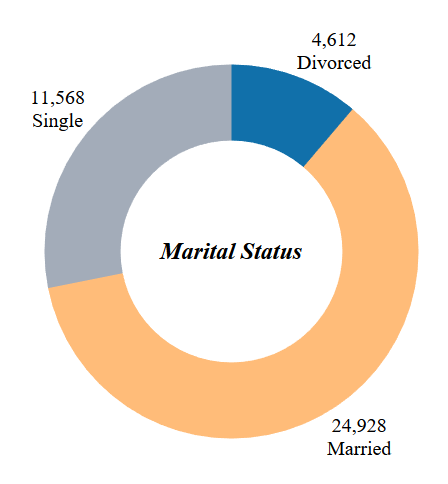


Figure 13 Marital Status

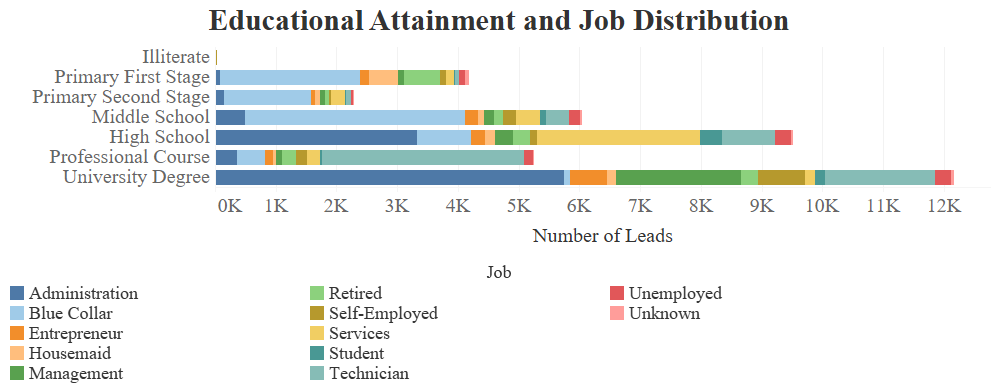


Figure 14

Moving to the customer demographics, it is observed that a significant portion of the customers contacted in the training set is married (*Figure 13*). Meanwhile, those with at least High School degree dominated the pool of leads where vast portion is working in Administration jobs (*Figure 14*).

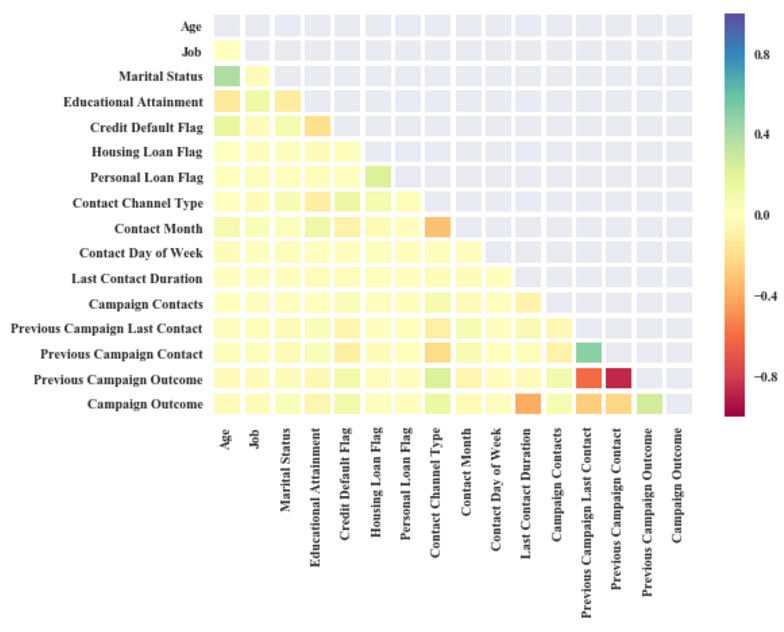


Figure 15 Correlation of features and output variable

Part of data exploration is to determine the correlation coefficients amongst the features. Having said that, the Pearson correlation method to determine correlation coefficients and visualise using the heatmap in *Figure 15*. In addition, it is found that the last contact duration has negative correlation with the campaign outcome whilst previous campaign outcome has positive correlation with campaign outcome. The next phase of the project is to utilise the insights gathered from this phase to build and evaluate the performance of the classification model.

**III. Proposal and Evaluation Framework**



Figure 16 CRISP-DM from *ipredictt.com/products.html*

The Cross Industry Standard Process for Data Mining (CRISP-DM) will be used in providing solution delivery for this project. In additional, the process model consists of six phases as follows (CRISP-DM Consortium, 2000):

* 1. Business Understanding
  2. Data Understanding
  3. Data Preparation
  4. Modelling
  5. Evaluation
  6. Deployment

Furthermore, the first three phase of model has been performed in this project as describe in section 2 of this report. The succeeding phase of the project will be described as below:

**1. Modelling**

This phase of project involves feature selection of the 20 available attributes that will be used as input parameters to ensure optimal classification model. The two-feature selection method that may be used in this project is as follows:

* + 1. Filter Method – this method is used in selection features based on various statistical correlation score with the campaign outcome.
    2. Wrapper Method – this method considers the selection of a set of features as a search problem, where different combinations are prepared, evaluated, and compared to other combinations. (Brownlee, 2014)

The next step in modelling is to determine the list of possible classification modelling algorithms to be used. For this project, the following algorithms will be used to build the classification model:

* 1. Logistic Regression
  2. Naïve Bayes
  3. Support Vector Machine (SVM)

Hence, these algorithms will be subject to performance evaluation to select the best classifier to support the business problem. In addition, this phase of the project will involve review and calibration of the cleansed and transformed data set to ensure that it fits the parameters needed for the algorithms mentioned.

**2. Evaluation**

This stage of the project involves thorough evaluation of the different model built to ensure that the classification met the business requirement. To select the best classifier, the following metrics will be used to evaluate the model:

* 1. Receiver Operating Characteristic (ROC) – shows performance of a binary classifier system as its discrimination threshold is varied (Wikipedia).
  2. Confusion Matrix – is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known (Markham, 2014).

Table 3 Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Class | | |
| Actual Class |  | Class = Yes | Class = No |
| Class = Yes | a | b |
| Class = No | c | d |

*a: TP (true positive)*

*b: FN (false negative)*

*c: FP (false positive)*

*d: TN (true negative)*

Below are the measures in evaluating the performance of the classification model:

**3. Deployment**

The last stage of the project involves summarisation and analysis of result, documentation and writing of the learning experience and future recommendation to further improve the performance of the classification model.

**IV. References**

Brownlee, J (2014, October 4) An Introduction to Feature Selection. Retrieved from http://machinelearningmastery.com/an-introduction-to-feature-selection/.

CRISP-DM Consortium (2000, August). CRISP-DM 1.0 Step-by-step Data Mining Guide. Retrieved from http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf.

Lohrey, J. (2013). The Average Success Rate of Direct Marketing. Retrieved from http://smallbusiness.chron.com/average-success-rate-direct-marketing-73648.html

Markham, Kevin (2014, March 25) Simple guide to confusion matrix terminology. Retrieved from http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

Moorman, C. and Finch, T.A. (2017, January 24). Marketing Budgets Vary by Industry. Retrieved from http://deloitte.wsj.com/cmo/2017/01/24/who-has-the-biggest-marketing-budgets/

Moro, S. (2014) Bank Marketing Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Pemberton, C. (2016, December 12). Gartner CMO Spend Survey 2016-2017 Shows Marketing Budgets Continue to Climb. Retrieved from http://www.gartner.com/smarterwithgartner/gartner-cmo-spend-survey-2016-2017-shows-marketing-budgets-continue-to-climb/

**Appendix A: Source to Target Mapping Sheet for the Fact Table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source File** | **Source Table** | **Source Column** | **Source Data Type** | **Target Table** | **Target Column** | **Target Data Type** | **Transformation Rule** | **Lookup Table** |
|  |  |  |  | FACT\_BANK\_TRAINING | RECORD\_KEY | INTEGER | Generate Record Sequence Number |  |
| bank-additional-full.csv | stg\_bank\_training | age | VARCHAR(100) | FACT\_BANK\_TRAINING | AGE | INTEGER | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | age | VARCHAR(100) | FACT\_BANK\_TRAINING | AGE\_BUCKET | VARCHAR(50) | Group the Age in buckets of 10: WHEN age < 20 THEN 'LESS THAN 20' WHEN age BETWEEN 21 AND 30 THEN |  |
| bank-additional-full.csv | stg\_bank\_training | job | VARCHAR(100) | FACT\_BANK\_TRAINING | JOB | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_JOB |
| bank-additional-full.csv | stg\_bank\_training | marital | VARCHAR(100) | FACT\_BANK\_TRAINING | MARITAL\_STATUS | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_MARITAL\_STATUS |
| bank-additional-full.csv | stg\_bank\_training | education | VARCHAR(100) | FACT\_BANK\_TRAINING | EDUCATIONAL\_ATTAINMENT | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_EDUCATIONAL\_ATTAINMENT |
| bank-additional-full.csv | stg\_bank\_training | default\_ | VARCHAR(100) | FACT\_BANK\_TRAINING | CREDIT\_DEFAULT\_FLAG | VARCHAR(10) | Convert the Source Column (Source Value) into Target Value | DIMN\_FLAGS |
| bank-additional-full.csv | stg\_bank\_training | housing | VARCHAR(100) | FACT\_BANK\_TRAINING | HOUSING\_LOAN\_FLAG | VARCHAR(10) | Convert the Source Column (Source Value) into Target Value | DIMN\_FLAGS |
| bank-additional-full.csv | stg\_bank\_training | loan | VARCHAR(100) | FACT\_BANK\_TRAINING | PERSONAL\_LOAD\_FLAG | VARCHAR(10) | Convert the Source Column (Source Value) into Target Value | DIMN\_FLAGS |
| bank-additional-full.csv | stg\_bank\_training | contact | VARCHAR(100) | FACT\_BANK\_TRAINING | CONTACT\_CHANNEL\_TYPE | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_CHANNEL\_TYPE |
| bank-additional-full.csv | stg\_bank\_training | month\_ | VARCHAR(100) | FACT\_BANK\_TRAINING | CONTACT\_MONTH | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_MONTH |
| bank-additional-full.csv | stg\_bank\_training | day\_of\_week | VARCHAR(100) | FACT\_BANK\_TRAINING | CONTACT\_DAY\_OF\_WEEK | VARCHAR(50) | Convert the Source Column (Source Value) into Target Value | DIMN\_DAY\_WEEK |
| bank-additional-full.csv | stg\_bank\_training | duration | VARCHAR(100) | FACT\_BANK\_TRAINING | LAST\_CONTACT\_DURATION | INTEGER | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | campaign | VARCHAR(100) | FACT\_BANK\_TRAINING | CAMPAIGN\_CONTACTS | INTEGER | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | pdays | VARCHAR(100) | FACT\_BANK\_TRAINING | PREVIOUS\_CAMPAIGN\_LAST\_CONTACT | INTEGER | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | previous | VARCHAR(100) | FACT\_BANK\_TRAINING | PREVIOUS\_CAMPAIGN\_CONTACT | INTEGER | IF pdays = 999 THEN 0 ELSE previous END |  |
| bank-additional-full.csv | stg\_bank\_training | poutcome | VARCHAR(100) | FACT\_BANK\_TRAINING | PREVIOUS\_CAMPAIGN\_OUTCOME | VARCHAR(20) | Convert the Source Column (Source Value) into Target Value | DIMN\_CAMPAIGN\_OUTCOME |
| bank-additional-full.csv | stg\_bank\_training | emp\_var\_rate | VARCHAR(100) | FACT\_BANK\_TRAINING | QUARTERLY\_EMPLOYMENT\_VARIATION\_RATE | DECIMAL(12,4) | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | cons\_price\_idx | VARCHAR(100) | FACT\_BANK\_TRAINING | MONTHLY\_CONSUMER\_PRICE\_INDEX | DECIMAL(12,4) | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | cons\_conf\_idx | VARCHAR(100) | FACT\_BANK\_TRAINING | MONTHLY\_CONSUMER\_CONFIDENCE\_INDEX | DECIMAL(12,4) | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | euribor3m | VARCHAR(100) | FACT\_BANK\_TRAINING | DAILY\_EURIBOR\_3MONTH\_RATE | DECIMAL(12,4) | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | nr\_employed | VARCHAR(100) | FACT\_BANK\_TRAINING | QUARTERLY\_NUMBER\_OF\_EMPLOYEES | DECIMAL(12,4) | Change Data Type |  |
| bank-additional-full.csv | stg\_bank\_training | y | VARCHAR(100) | FACT\_BANK\_TRAINING | CURRENT\_OUTCOME | VARCHAR(10) | Convert the Source Column (Source Value) into Target Value | DIMN\_FLAGS |