**Data-Mining-Solutions-for-Direct-Marketing-Campaign**

# 1.INTRODUCTION

Data mining is a technique which extracts and analyses the large set of data and predict all possible outcomes through machine learning. It has the ability to predict the future by using the historical data. Which helps us as a guide to make decisions at present. It is a most effective marketing strategy which is in practice by many organisations in the recent years.

Banks follows set of marketing strategies to increase the sales of their products. The most common method is direct marketing campaign which is tele marketing. These banks will make phone calls to increase their customers and profits. To minimize the costs of the marketing campaign these banks follow data mining techniques to predict and reduce the costs. These data mining techniques will predict the interests of the customers and the success rate by doing the data mining classification algorithms.

This data mining technique will allow the bank to predict the customers interest of subscribing to the product. A data mining classification algorithm is used to evaluate the interest of the customer. In this, the given data set is initially divided into two parts. One is train set and the other one is test set. Train set will predict the future values by using the data. Where as test set is set of data which is not seen by the model.

Banks uses clients information such as age, job, education and marital etc to predict the success rates. They classify a model which helps to identify the customers interest on buying their product. By doing this, banks will minimize the cost which will help them in efficient use of their resources.

# 2.DATA

There is a set of data given from a bank which contains the personal information of the customers which is used in a marketing campaign. The given data set contains 41188 instances and 21 attributes. These 21 attributes are categorised into 5 types.

1. Demographic information : Age, Job, Marital, Education
2. Financial information : Default, Housing, Loan, y
3. Current campaign : Contact, Month, Day\_of\_week, Duration, Campaign
4. Previous campaign : pdays, previous, poutcome
5. Economic situation : emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

Data features are classified into two types, one is categorical features and the other one is numerical features.

Categorical features are job, marital, education, default, housing, loan, contact, month, day\_of\_week, poutcome, y.

Numeric features are age, duration, campaign, pdays, previous, emp.var.rate, cons.price.index, cons.conf.idx, euribor3m and nr.employed.

# 3.Class Distribution

In order to find out the given data is balanced or imbalanced, class distribution is applied.

![Chart, bar chart

Description automatically generated]()

The difference is very high as the given data is imbalanced.

# 4.Univariate analysis

Univariate analysis will analyse the data in a simple form. As the name suggests it applies to one variable. It is a basic analysis applied to the given data. The given data contains 21 attributes, both categorical and numeric variables. Univariate analysis is applied to each of the variable separately. It looks on the both range and central tendency of the values. Its an easy way to analyse the given data. It will distinguish between the classes. We know the given data contains 21 attributes, each variable again sub categorised, univariate analysis will help us to find out which specific category is more likely to show interest.

Applying univariate analysis is applied to both the categorical features and numeric features.

## 4.1 Univariate analysis on categorical features:

### 

### 4.1.1 Job

**![Chart, bar chart

Description automatically generated]()**

Job: Housemaid, services, admin, blue-collar, technician, retired, management, unemployed, self-employed, unknown, entrepreneur and student.­­

The above plot indicates that the different types of jobs that customers do. From the plot it is clear that the maximum customers who are doing admin job have the highest chances of buying the product, also it also shows they are the highest in not showing any interest on the product.

### 4.1.2 Marital

![Chart, bar chart

Description automatically generated]()

Marital: Married, single, divorced and unknown.

From the plot, most of the customers having married as marital status.

## 4.1.3 Education:

![Chart, bar chart

Description automatically generated]()

Education: Basic 4y, high school, basic 6y, professional course, university degree and illiterate.

### 4.1.4 Default

Default indicates whether the customer is having credit or not.

![Chart

Description automatically generated]()

Default: Yes, no and unknown.

### 4.1.5 Housing

It implies the status of the customer having house loan. From the plot majority of the customers having a house loan.

Chart, bar chart

Description automatically generatedHousing: Yes, no and unknown.

4.1.6 Loan

Chart, waterfall chart

Description automatically generated

Loan: Yes, no and unknown.

### 4.1.7 Contact

Chart, bar chart

Description automatically generated

Contact: Telephone and cellular.

### 4.1.8 Month

![Chart, bar chart

Description automatically generated]()

Month: March, April, May, June, July, August, September, October, November and December. From the plot the distribution is varying between each month so it will be helpful to predict the customers decision.

### 4.1.9 Day\_of\_week

![Chart, bar chart

Description automatically generated]()

Day of week: Monday, Tuesday, Wednesday, Thursday and Friday.

In the above plot the distribution among all the days is similar. That is, on average 17% of the customers rejected and 2.1% of the customers agreed.

### 4.1.10 Poutcome

The plot indicates majority fall under not exist category. That is, most of the customers are not contacted earlier. They are all new customers. Therefore this feature is important in predicting the output.

**![Chart, waterfall chart

Description automatically generated]()**

Previous out come: Success, Failure and notexist.

## 4.2 Univariate analysis on Numeric features:

### 4.2.1 Age

The mean age of the customers is 40.

**![Chart, box and whisker chart

Description automatically generated]()**

### 4.2.2 Duration

This attribute is important to consider, the duration here is about duration of the call last contacted represented in seconds. It is a known feature as if the duration=0, then the ‘y’ becomes no (i.e., y = no). After the end of the call, the variable ‘y’ is known. Therefore it becomes one of the important attribute in predicting. Therefore a graph is plotted

**Chart, box and whisker chart

Description automatically generated**

![Chart

Description automatically generated]()

### 4.2.3 Campaign

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

### 

### 4.2.4 Pdays

**![Chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

### 4.2.4 Emp.var.rate

Employment variation rate is indicated in quarterly manner.

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

### 

### 4.2.5 Cons.price.idx

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

4.2.6 Cons.conf.idx

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

### 4.2.7 Euribor3m

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

### 

The graph shows the difference in the distribution, hence this is will be useful in predicting the outcome.

### 4.2.8 Nr.employed

**![Chart, box and whisker chart

Description automatically generated]()**

**![Chart, histogram

Description automatically generated]()**

It is also a key variable as it also has the difference in the distribution.

By doing univariate analysis, its clear that

1. The mean age of the customers is 40
2. Majority of the customers are married.
3. Most of them are doing admin jobs.
4. Most of them are having university degree as education status
5. Majority of them don’t have loans.
6. Most of them are new customers who haven’t contacted previously.

# 5. Correlation of features

**![Chart, treemap chart

Description automatically generated]()**

A correlation graph is plotted among the variables. It is observed that, emp.var.rate, cons.price.idx euribor3m and nr.employed has the maximum correlation where euribor3m and nr.employed is having a highest correlation of ‘0.95’. Therefore these are the important variable in predicting the results.

# 6.Encoding Technique

## 6.1 Response Coding

Whenever solving a machine learning problem, response coding technique is used to represent the categorical data. It represents, the chances of data belonging to a particular class.

Mathematical expression is given as

**P(class=X | category=A) = P(category=A ∩ class=X) / P(category=A)**

## 6.2 Laplace Smoothing

It handles the probability of zero values in naïve bias and represented as

![Graphical user interface

Description automatically generated with medium confidence]()

Where,

Alpha = Smoothing parameter

K = No.of dimensions in the data,

N = No.of reviews.

## 6.3 Pseudo Code

Code for response coding with Laplace smoothing

Alpha : Used for Laplace smoothing

Feature : Categorical features

Df : [‘train\_df’,’test\_df’,’cv\_df’]

Algorithm

Consdier all unique values and the number of occurrences of given feature in train data frame

Build a vector (1\*2), the first element = (number of times it occurred in class1 + 10\*alpha / number of times it occurred in total data+20\*alpha)

Feat\_dict is like a look up table, for every categorical data it store (1\*2) representation of it

For a value of feature in df

If it is in train data

We add the vector that was stored in ‘feat\_dict’ look up table ‘res\_fea’

If it is not there,

We add [1/2 , 1/2]

Return ‘res\_fea’

# 7. Random forest

Random forest is the one of the top classifier techniques which are in use now a days in machine learning techniques. It is a collection of set of decision trees. Decision trees usually have high variance and low bias. It gives outputs based on the conditions present in the data given. This algorithm creates trees which are dependent to each other. This technique works on the numeric data. It is flexible for both classification and regression problems and it improves the accuracy by decreasing the overfitting problems in decision trees. As it combines the outputs of the various trees, it requires more computation. The time taking will be more as it makes lot of decision trees into one form in order to determine the class. When compared to other algorithms, Random forest works on the data that contains categorical variable. This method is more likely to be used in banking.

**![Table

Description automatically generated]()**

Random Forest output Fig

By applying random forest algorithm we have got 0.942 accuracy

# 8. K - Nearest Neighbour

KNN is a simplest machine learning algorithm. It is used to solve classification problems and regression problems. KNN is a supervised learning algorithm. In order to predict the future, it uses labelled set of data which is given. That is, we have given a data from a bank which is known, KNN uses the data and predicts the data whether it is positive or negative. It stores all the data which is available and categorize a new data point on distance functions when a new data appears. The value of the K is chosen at the beginning. It is very fast in terms of calculation time. It makes simple when there is a training data set is provided. But the output accuracy will completely depends on the data that is provided.

**![Chart, line chart

Description automatically generated]()**

KNN Output Fig

By applying KNN algorithm we have got 0.924 accuracy.

# 9. Summary

Classification of the success rate of a customer to subscribe is predicted through a given bank telemarketing campaign data using two different machine learning algorithms (i.e., Random forest and KNN). The data set consists of 21 attributes. Class distribution of the variables is imbalanced as the there is lot of difference between the positive class and negative class. By doing univariate analysis on both numerical and categorical variables, analysed the most important features which are useful to predict the success rate. They are emp.var.rate, cons.price.idx, cons.price.idx, euribor3m and nr.employed. These variable also have the highest correlation among them. Among the variables, both euribor3m and nr.employed has the highest correlation value of ‘0.95’. Then implemented the machine learning algorithm “RANDOM FOREST”. It is a collection of decision trees and also it can handle the categorical values. By applying Random Forest algorithm on the given data, got an accuracy of “0.942”. Then implemented “K – NEAREST NEIGHBOURING” technique. It takes a bit more time for execution compared to Random Forest method. It handles the numeric variables and it works on the distance method, where K is an integer. By applying KNN algorithm on the given data, got an accuracy of “0.924”. So it is clear that Random Forest is the more accurate when compared to KNN.

# 10. References

1. An Implementation and Explanation of the Random Forest in Python by Will Koehrsen.
2. Introduction to K – Nearest Neighbours: A powerful Machine Learning Algorithm ( with implementation in Python and R) by Tavish Srivastava.
3. Comparison of Random forest, K – Nearest Neighbouring and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery.
4. Predicting Demographic and Financial Attributes in a Bank Marketing Dataset by Samiraj Ejaz.
5. Github Machine learning techniques by Sukanta Roy.
6. Resopanse Coding for Categorical Data by Dharmik Thakkar.
7. Laplace smoothing in Naïve Bayes algorithm by Vaibhav Jayaswal.
8. Naïve Bayes Classification using Python by James McCaffrey.
9. The 7 most important data mining techniques by Larry Alton.
10. A review of data mining techniques by Sang Jun Lee (University of Nebraska-Lincoln, Lincoln, Nebraska, USA), Keng Siau (University if Nebraska-Lincoln, Lincoln, Nebraska, USA).