**Table of Contents: Page No.**

1. Task 1

Introduction 2

Dataset Merging 2

Data Pre-processing 2

Exploratory Data Analysis 3

Binning 9

* 1. Bar Graphs 5
  2. Line Graph 6
  3. List 6
  4. Heat Map 7

1.10Data Balancing 7

2. Task 2

2.1 Neural Network 8

2.2 Decision Tree 11

2.3 Conclusion 13

3. Task 3

3.1 Concepts Node 15

3.2 Text parsing Node 16

3.3 Sentiment Node 19

3.4 Topics Node 20

3.5 Categories Node 21

3.6 Conclusion 22

4. Task 4 23

5.References 25

**Task 1:**

**Introduction**:

This analysis contains 2 datasets - all\_campaign.sav and all\_personal.sav - which contain the details of a Campaign contact information and response outcome and the customers’ personal characteristics respectively. The campaign was conducted by an International Bank for the promotion of a fixed term saving account. The campaign dataset has information on the customer ID, contact type, duration and response of customers and personal dataset has the customer ID, age, region, job, marital status, education, default, balance, housing, and loan details of the customers. The target variable for the analysis is ‘response’, which the the customer response of the campaign.

**Dataset Merging:**

Since the data were in 2 different datasets with customer ID column common, they were merged together in order to get a detailed dataset (Fig 1.). Merging has been done via Data->Merge Files->Add Variables option.

A screenshot of a computer

Description automatically generated

*Fig 1. all\_campaign and all\_personal dataset merged into a single dataset.*

Proper meaningful labels are added to all the variables to easily understand the variables (Fig 2.).

**Data pre-processing:**

A summary of the missing values and outliers has been generated in Fig 3. There are no missing values for any of the variables and a total count of 3390 rows are there in the dataset. There are 4 numerical variables- customer ID, age, balance, duration and 9 categorical variables - region, job, marital status, education, default, balance, housing, loan and response. Age is having a lower extreme value of 32 and higher extreme of 720, while balance and duration has 1159 and 1568.

A screenshot of a computer

Description automatically generated

*Fig 2. Labelling the dataset.*

A screenshot of a data sheet

Description automatically generated

*Fig 3. Statistics of all variables in the dataset.*

**Exploratory Data Analysis:**

To understand the relationship between customer age and their yearly balance, a scatter plot is created (fig 4.). Customers of all age groups have an yearly balance of Rs 0- Rs 25,000 and very few customers have a good yearly balance of more than Rs 25,000.

A graph of a number of blue and red dots

Description automatically generated

*Fig 4. Scatter plot of customer age and yearly balance of customer*

A graph of different colored columns

Description automatically generated

*Fig 5. Bar chart of different contact types and response*

From fig 5, it is evident that most of the customers who were contacted by any mode – mobile, telephone, unknown, were not interested. Majority of the customers who responded were contacted for a duration od around 200 secs, by any mode of contact.

*Fig 6. Pie chart representation of credit default/not for contacted customers.*

98% of customers who were contacted were not having any default while less than 2% customers had default (Fig 6.).

Fig 8. is a bar graph representing that 45% of secondary educated customers did not respond to the campaign, while 5% responded.

55.77% of contacted customers have housing load and 44.23% does not have housing loan, which is given by pie chart (Fig 9.).

*Fig 7. Missing values and mode of categorical variables*

It is evident from fig 7. That there are no missing values in categorical variables. So no need of any missing values handling methods in this dataset.

A graph of a number of blue and red bars

Description automatically generated

*Fig 8. Education of contacted customers and their response*

*Fig 9. Pie chart of distribution of customers with and without housing loan*

*Fig 10. Job of contacted customers and their response*

15-20% of customers from management, technician and other jobs did not respond to the campaign. Less than 5% of customers from different jobs responded.

A blue pie chart with a red triangle and a triangle

Description automatically generated

*Fig 11.* *Pie chart of distribution of customers with and without personal loan*

Out of the total contacted customers, 83.9% have personal loan while 16% don’t have (Fig 11.).

*Fig 12.* *Bar chart of contacted customers from various regions*

20%-30% of customers contacted were from London, South-East and North-West and small number were contacted from North-East and East Midlands (Fig 12.).

A pie chart of response of customer

Description automatically generated

*Fig 13. Pie chart distribution of customer response*

Only 11% of contacted customers actually responded to the campaign while 88% did not respond (Fig 13.).

**Binning:**

Binning is the process of grouping data into intervals or "bins" to make it easier to analyse or visualize. Here, customer age and contact duration are numerical data which can be grouped into different intervals for better understanding. Binning to equal interval gives better knowledge on which age groups were contacted more and for how much duration (in secs).

A screenshot of a statistics

Description automatically generated

*Fig 14. Summary statistics of contacted customer age*

From fig 14, the mean age of customers is 40 years, minimum is 18 years and maximum is 95 years. So intervals can be taken as 20 years, starting from 1-20 years. The binned graph is given below (fig 15.).

A graph of a number of people

Description automatically generated with medium confidence

*Fig 15. Equal interval binned bar graph of customer age*

After binning age, it is clear that 48% from 21-39 years and 44% from 40-58 years makes the majority of customers. Less than 1% customers from 0-20 years and 77+ years age group.

The mean duration of contact is 257.61 secs while maximum is 4918 secs (fig 16). The intervals are divided into bins of interval of 600, starting from 0-600 secs. From fig 17, it is clear that most customers were contacted for a duration of less than 600 secs (10 mins).

A screenshot of a statistics

Description automatically generated

*Fig 16. Summary statistics of contact duration*

*Fig 17. Equal interval binned bar graph of contact duration*

**Data Splitting:**

Data splitting is necessary for training the model and then evaluating the model performance on test data. This helps assess how well the model generalizes to new, unseen data. The whole dataset is split into 70% for training and 30% for testing using the option Data->Split into files. After splitting, both train and test datasets are saved in separate files. So, 23,736 data points out of 33909 has been retained in training set and 10173 is for testing (fig 18.). The random seed value used is 8888 to get a consistent output value very time the model is run.

*Fig 18. Number of data points in training dataset*

**Task 2**

**Logistic Regression**

Binary logistic regression is a statistical method used to model the relationship between one or more independent variables and a binary output variable. It estimates the probability of the output variable taking a particular value (usually 0 or 1) based on the values of independent variables.

Logistic regression is done using the steps Analyse->Regression->Binary Logistic. In first model, select all variables and choose response as dependant variable. For categorical variables, choose ‘last’ as reference category.

The significance level chosen for the analysis is 0.1. Therefore, any variables with p-value greater than 0.1 will be either rejected or merged to base class, accordingly.

Training Model 1:

Chosen independent variables –

1. Contact
2. Duration
3. Age
4. Region
5. Job
6. Marital
7. Education
8. Default
9. Balance
10. Housing
11. Loan

Dependent variable-Response

The categorical variable coding is given in fig 19., which shows the reference class (last category given by .000 in all columns) and other categories.

Fig 20. shows the significance of each category. Categories with p-value less than 0.1 are significant and others need to be removed.

Since Region and Job has p-values greater than 0.1, they are removed for model 2 (appendix).

Training Final Model:

Chosen independent variables –

1. Contact
2. Duration
3. Age
4. Marital
5. Education
6. Balance
7. Housing
8. Loan

Dependent variable-Response

In the final model, Region and Job variables are removed and tertiary education is merged with reference class (fig 21.). Currently, education has only 3 categories- primary, secondary and others (tertiary+unknown). The p-value of all variables are now less than 0.1, indicating that the variables are statistically significant. Age, housing, personal, contact and duration have positive relation with response. So, for each one-unit increase in variable, the log odds of response increase by co-efficient value. Marital and education have negation relation which states that log odds of response decrease by co-efficient value for each variable (fig 22.).

A screenshot of a table

Description automatically generated

*Fig 21. Categorical variable codings for final model*

A table with numbers and a few black text

Description automatically generated with medium confidence

*Fig 22. Variables in the equation: p-value of variables for final model*

Model Evaluation Using Testdata:

The ROC curve and the area under the ROC curve (AUC) are used to evaluate the performance of the logistic regression model. The ROC curve (fig 23.) plots the true positive rate against the false positive rate, illustrating the model's ability to discriminate between positive and negative outcomes. The curve in our analysis shows a strong performance, bending towards the upper left corner. The AUC (fig 24.) value of 0.860 further confirms this, indicating that the model has a high level of accuracy, with an 86% probability of correctly distinguishing between positive and negative instances. This demonstrates that the model is effective and reliable in predicting the outcome variable.

A graph with a line

Description automatically generated

*Fig 23. ROC of test results*

A screenshot of a test

Description automatically generated

*Fig 24. Area under ROC curve*

**Decision Tree Analysis**

Decision tree analysis is used to classify data into distinct categories by recursively splitting the dataset based on feature values. This method creates a tree-like model of decisions, where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.

Training Model:

A graph of a curve

Description automatically generated with medium confidence

A screenshot of a test

Description automatically generated

Model Evaluation Using Testdata:

A graph with a line and a red line

Description automatically generated

A screenshot of a test

Description automatically generated

**Appendix:**

Logistic Regression

Model 2:

Chosen independent variables –

1. Contact
2. Duration
3. Age
4. Marital
5. Education
6. Default
7. Balance
8. Housing
9. Loan

Dependent variable-Response

The categorical variable codings for model 2 are given in fig 21. From frig 22, it shows that default has p-value 0.163 (>0.1) and Education (3) has a value 0.664 (>0.1). Since the reference category of education has p-value <0.001, education (3) need to be merged with reference category in model 3.

A screenshot of a table

Description automatically generated

*Fig 21. Categorical variable codings for model 2*

*A table with numbers and text

Description automatically generated*

*Fig 22. Variables in the equation: p-value of variables for model 2*