**PREDICTIVE ANALYTICS PROJECT REPORT**

(Project Semester August-December 2023)

***Customer Conversion Prediction***

Submitted by

Raju Devnath

Registration No 12014332

Programme and Section BTECH(CSE) K20DP

Course Code INT 234

Under the Guidance of

**Baljinder kaur UID 27952**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**



**DECLARATION**

I, Raju Devnath, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 2 Nov 2023 Name of the student

Reg. No- 12014332 Raju Devnath

**Acknowledgement**

In preparation of my assignment, I had to take the help and guidance of some respected persons, who deserve my deepest gratitude. As the completion of this assignment gave me much pleasure, I would like to show my gratitude Baljinder kaur, Course Instructor, for giving me a good guideline for assignment throughout numerous consultations. I would also like to expand my gratitude to all those who have directly and indirectly guided me in writing

Many people, especially my classmates have made valuable comment suggestions on my paper which gave me an inspiration to improve the quality of the assignment.

**Table of Content**

1.Introduction

2.Objectives/Scope of the Analysis

3.Source of dataset

4.ETL process

5.Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization
6. List of Analysis with results
7. Training and selecting ML model
8. References
9. Bibliography

**Introduction**

In the age of data-driven decision-making, this report documents our predictive analysis project's objectives, scope, and significance. We aim to harness data for accurate predictions, guiding proactive decision-making. The report covers data collection, , model selection, training, evaluation, and deployment. This project seeks to improve efficiency, reduce costs, and enhance decision-making across various domains, demonstrating the power of predictive analysis in our organization.

This project aims to explore the vast and exciting field of Machine Learning, with a specific focus on leveraging the power of the R programming language. R, known for its robustness and versatility in statistical analysis, provides an excellent platform for conducting sophisticated machine learning tasks.

The primary objective of this project is to develop machine learning models that can analyze and interpret complex datasets. By utilizing various preprocessing R packages and tools, we aim to build models that can learn from data, identify patterns, and make decisions with minimal human intervention.

In addition to R, we will also be using other tools and technologies to aid in our analysis. These may include data visualization libraries to help understand our data better, and various machine learning libraries to implement different algorithms.

Through this project, we hope to gain a deeper understanding of machine learning concepts and how they can be applied in real-world scenarios using R. We believe that the knowledge and skills acquired from this project will be invaluable in our future endeavors in the field of data science and machine learning.

**Objectives/Scope of the Analysis**

This project aims to predict customer conversion for an insurance company. Using historical data, we will develop a machine learning model that can identify which customers are most likely to purchase a policy. By understanding the patterns and behaviors of previous customers, the model can make predictions and provide valuable insight to the company. The objective is to increase the rate of successful conversions and ultimately improve the company's bottom line.

Some of the objectives we are trying to analyse in this project are-

1. Analysing the Duration of call for the targeted and subscribed customers for insurance.
2. The number of customers targeted and subscribed for insurance on the basis of their age.
3. The number of customers targeted and subscribed for insurance on the basis of their type of job.
4. The month of the year in which most customer targeted and subscribed for insurance.
5. The days of the month in which most customer targeted and subscribed for insurance.

**Source of the Dataset**

We are working for a new-age insurance company and employ multiple outreach plans to sell term insurance to your customers. Telephonic marketing campaigns still remain one of the most effective way to reach out to people however they incur a lot of cost. Hence, it is important to identify the customers that are most likely to convert beforehand so that they can be specifically targeted via call. We are given the historical marketing data of the insurance company and are required to build a ML model that will predict if a client will subscribe to the insurance.

Link to download the dataset: https://drive.google.com/file/d/1LQpnpzH3NNaNMsPy6RBdocmChbsUOc0s/view?usp=drive\_link

Columns: There are 11 columns each having a title which is self-explanatory.

Rows: There are 52860 rows each having a mobile with at least a distinct feature.

DESCRIPTION OF ATTRIBUTES

1. age (numeric)
2. job : type of job
3. marital : marital status
4. educational\_qual : education status
5. call\_type : contact communication type
6. day: last contact day of the month (numeric)
7. mon: last contact month of year
8. dur: last contact duration, in seconds
9. num\_calls: number of contacts performed during this campaign and for this client
10. prev\_outcome: outcome of the previous marketing campaign
11. Output variable (desired target): y - has the client subscribed to the insurance?

**ETL process**

The Dataset was download from the link attached above.

The dataset was in .csv format.

The dataset read by in RStudio.

Basic Analysis of Dataset From Problem Statement And Features:

* It is a supervised learning problem - We are predicting target variable .
* From target variable we can clearly understand it is a classification problem .
* From target variable we can tell it is a binary classification problem. target = (y/n) .

Loading and preprocessing the dataset in rStudio:

df=as.data.frame(read.csv("train.csv"))

View(df)

summary(df)

str(df)

prop.table(table(df$y))

#From the above result we can clearly understand that the dataset is imbalanced.

#Checking for Na and duplicate values

sum(is.na(df))

sum(duplicated(df))

df <- df[!duplicated(df), ]

unique\_values <- sapply(df, unique)

#From all the above results all values are unique which means there is no incorrect or wrong df that is spelling mistake, upper case and lower case mismatch of each values.

#Exploring the dataset and replacing the unknown values

#changing datatype of target column

df$target <- ifelse(df$y == "yes", 1, ifelse(df$y == "no", 0, NA))

#checking and removing outliers

q1 <- quantile(df$age, 0.25)

q3 <- quantile(df$age, 0.75)

IQR <- q3 - q1

upper <- q3 + 1.5 \* IQR

lower <- q1 - 1.5 \* IQR

df$age <- ifelse(df$age < lower, lower, ifelse(df$age > upper,upper, df$age))

**Analysis on dataset**

1. Introduction

1.Analysing the Duration of call for the targeted and subscribed customers for insurance.

1. General Description

The analysis of call duration for targeted and subscribed customers for insurance involves examining the length of calls made to these customers. This analysis can provide valuable insights into customer behavior and the effectiveness of the calls.

1. Specific Requirements, functions and formulas and prediction model

R packages: dplyr and ggplot2

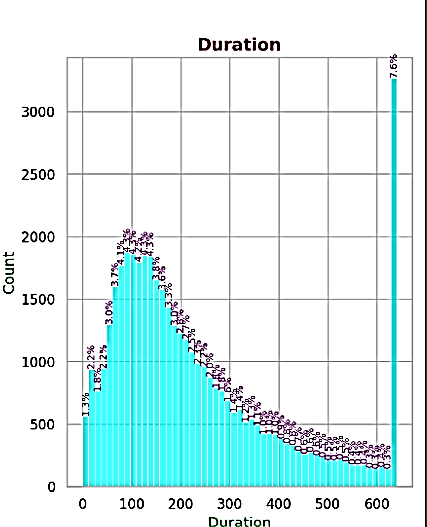
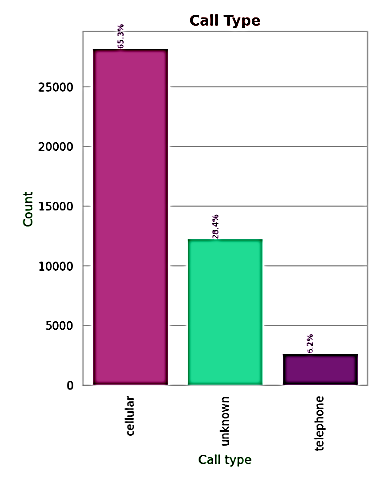
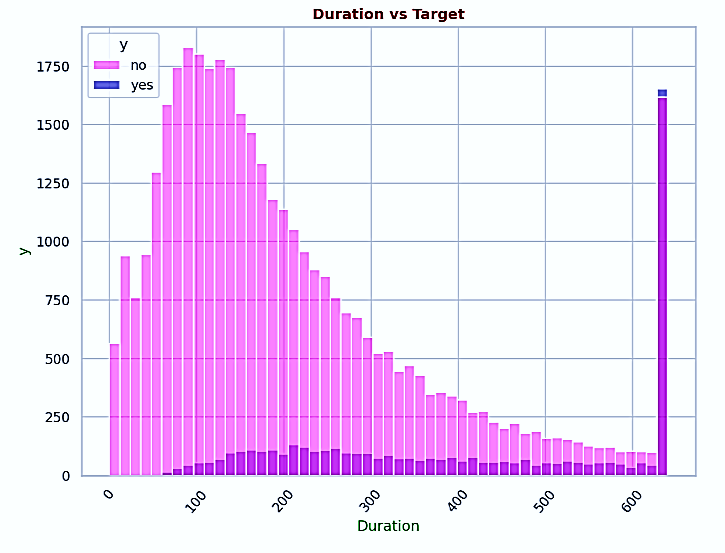
Statistical tool: Distribution graph

1. Analysis results

Based on the analysis of insurance sales calls:

* Most targeted customers had longer calls around 1750 seconds.
* Least targeted customers had shorter calls between 100 to 200 seconds.
* Cellular calls were the most common for both targeted and subscribed customers.
* The highest subscription rate was observed for customers contacted via cellular networks.
* The lowest subscription rate was for customers with an unknown call type.

1. Visualization

1. Introduction

2. The number of customers targeted and subscribed for insurance on the basis of their age.

1. General Description

Analyzing the number of customers targeted and subscribed for insurance based on their age involves examining the age distribution of these customers. This analysis can provide valuable insights into customer behavior and the effectiveness of the insurance sales strategy.

1. Specific Requirements, functions and formulas and prediction models

R packages: dplyr and ggplot2

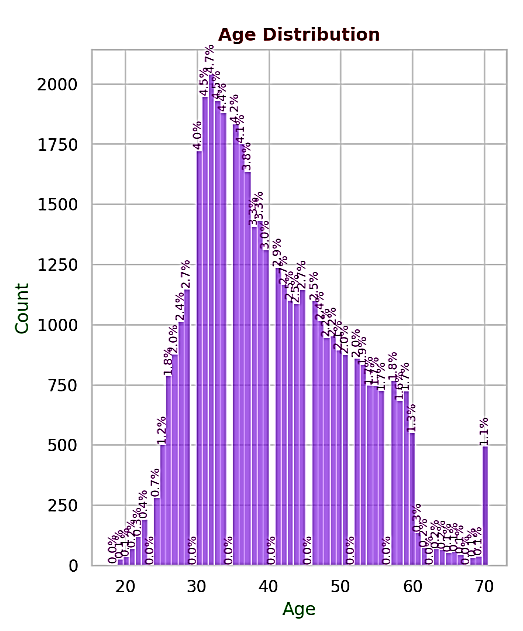
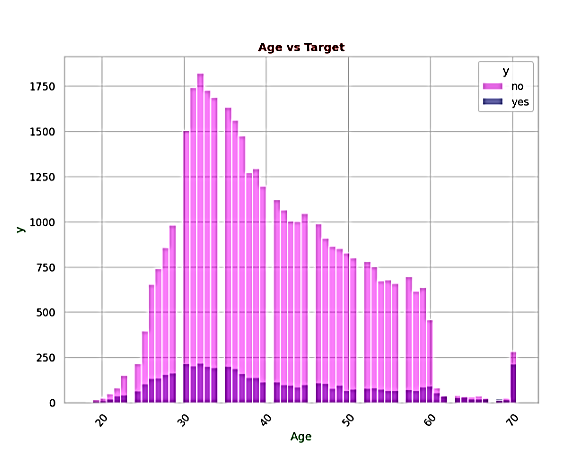
Statistical tool: Distribution graph

1. Analysis results

Based on the analysis of insurance sales calls:

* The primary target for insurance sales calls are middle-aged people, specifically those aged between 30 to 40 years.
* The least targeted age groups are those below 20 years and above 60 years.
* Interestingly, the age group with the highest subscription rate is also the middle-aged people.

1. Visualization

1. Introduction

3. The number of customers targeted and subscribed for insurance on the basis of their type of job.

1. General Description

Analyzing the number of customers targeted and subscribed for insurance based on their job type can provide valuable insights into which professions are more likely to need and subscribe to insurance. This information can help insurance companies tailor their marketing strategies and product offerings to meet the needs of different professional groups.

1. Specific Requirements, functions and formulas and prediction models

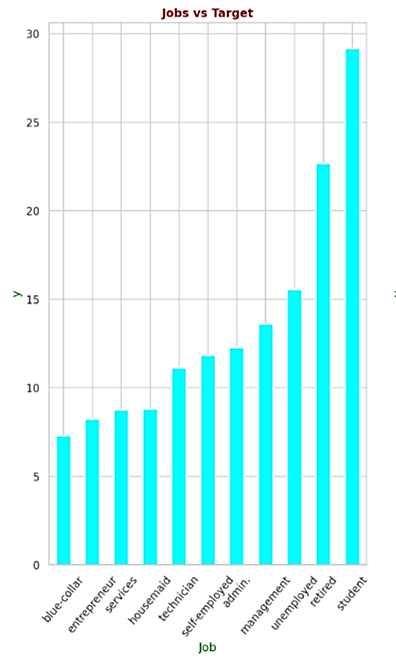
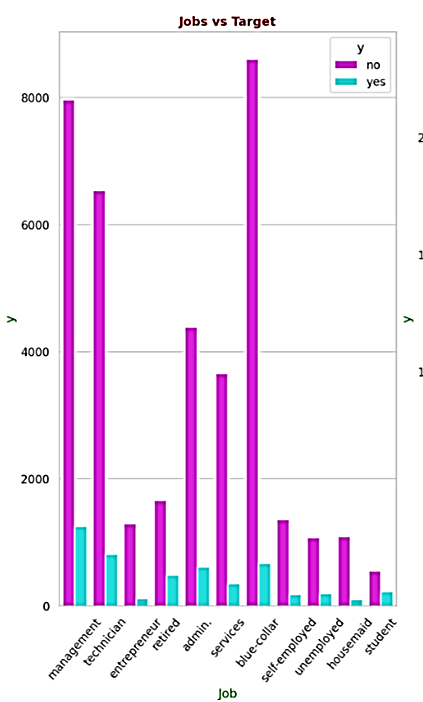
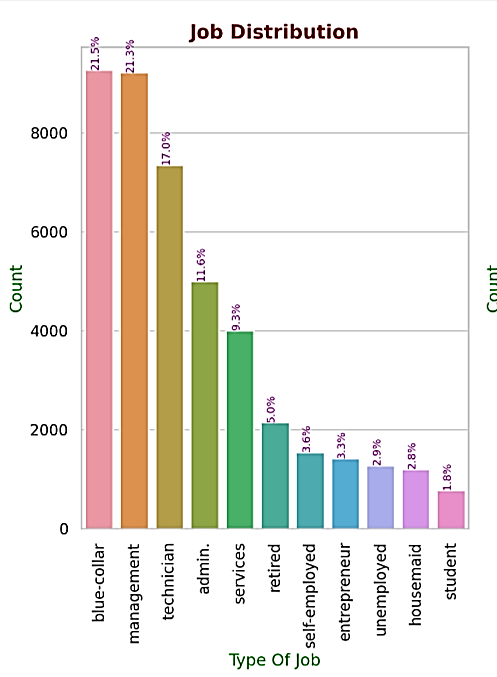
R packages: dplyr and ggplot2

Statistical tool: Distribution graph

1. Analysis results

* The most targeted job categories for insurance sales calls are blue-collar workers and management professionals.
* The least targeted job categories are students and housemaids.
* Among those targeted, blue-collar workers are less likely to subscribe to insurance.
* On the other hand, management professionals have the highest subscription rate.

1. Visualization



1. Introduction

4. The month and day of the year in which most customer targeted and subscribed for insurance.

1. General Description

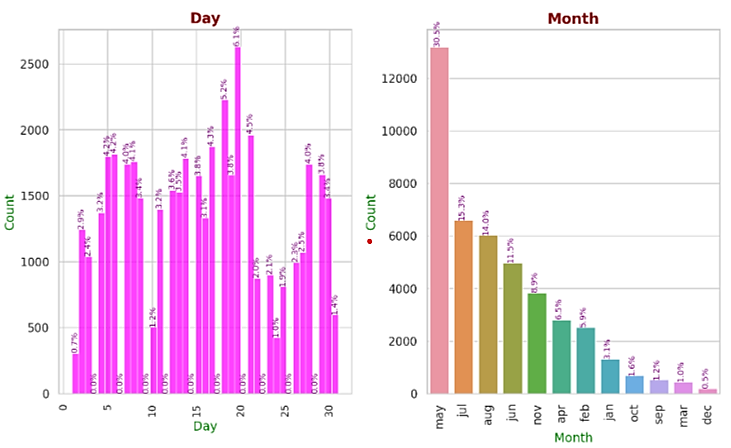
This could indicate the presence of certain seasonal trends or customer behaviors during these months and days that make them more conducive to subscribing to insurance.

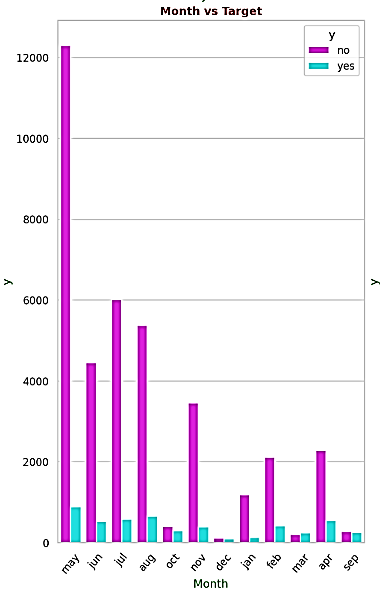
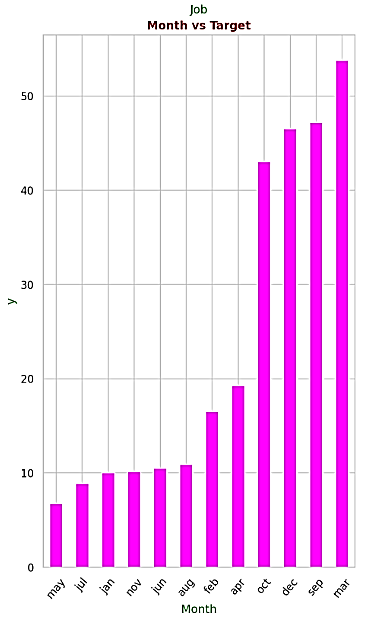
1. Specific Requirements, functions and formulas and prediction models

R packages: dplyr and ggplot2 Statistical tool: Distribution graph

1. Analysis results

* The month with the most targeted customers was May and least in December.Despite being the most targeted, May also had the lowest subscription rate.
* The highest subscription rates were observed in March and September.
* The mid days of the month has most subscribed customer.

1. Visualization



**List of Analysis and results**

1. Most targeted customers had longer calls around 1750 seconds. Least targeted customers had shorter calls between 100 to 200 seconds.
2. The primary target for insurance sales calls are middle-aged people, specifically those aged between 30 to 40 years. The least targeted age groups are those below 20 years and above 60 years. Interestingly, the age group with the highest subscription rate is also the middle-aged people.
3. The most targeted job categories for insurance sales calls are blue-collar workers and management professionals. The least targeted job categories are students and housemaids. Among those targeted, blue-collar workers are less likely to subscribe to insurance. On the other hand, management professionals have the highest subscription rate.
4. The month with the most targeted customers was May and least in December. Despite being the most targeted, May also had the lowest subscription rate. The highest subscription rates were observed in March and September. The mid days of the month has most subscribed customer.

**Training and selecting ML model**

# Set the random seed for reproducibility

set.seed(3)

# Split the dataset into training and testing sets

# SplitRatio is set to 0.8 for an 80/20 split

split <- sample.split(df, SplitRatio = 0.8)

x\_train <- x[split, ]

x\_test <- x[!split, ]

y\_train\_labels <- y[split]

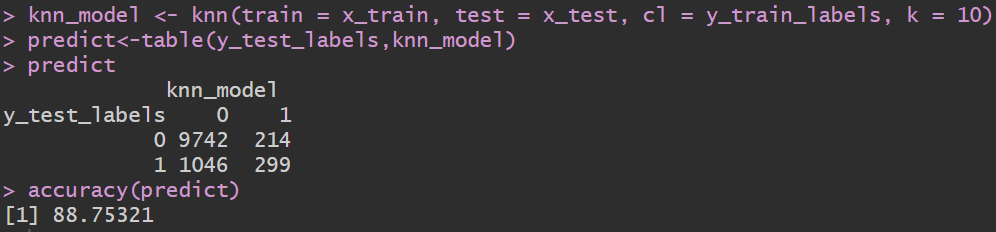
y\_test\_labels <- y[!split]

str(df)

*KNN model training*

knn\_model <- knn(train = x\_train, test = x\_test, cl = y\_train\_labels, k = 10)

predict<-table(y\_test\_labels,knn\_model)



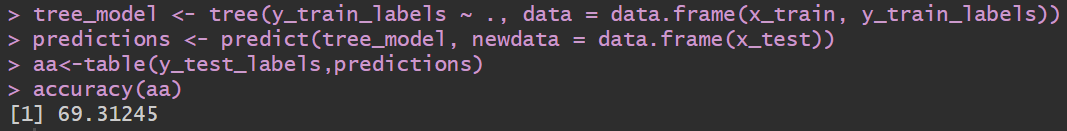
*Decision Tree model training*

tree\_model <- tree(y\_train\_labels ~ ., data = data.frame(x\_train, y\_train\_labels))

predictions <- predict(tree\_model, newdata = data.frame(x\_test))

predict<-table(y\_test\_labels,predictions)

accuracy(predict)



*Random Forest model training*

install.packages("randomForest")

library(randomForest)

classi=randomForest(x=x\_train, y=y\_train\_labels,ntree=500)

y\_pred=predict(classi,newdata=x\_test)

aa<-table(y\_test\_labels,predictions)



So the best model for our dataset would be KNN as it has the highest accuracy in among all the other ML algorithm.

**References**

[**www.kaggle.com**](http://www.kaggle.com)

[**www.google.com**](http://www.google.com)

[**www.youtube.com**](http://www.youtube.com)

[**www.airbnb.com**](http://www.airbnb.com)

[**http://insideairbnb.com/get-the-data/**](http://insideairbnb.com/get-the-data/)

[**https://cran.r-project.org/**](https://cran.r-project.org/)

**Bibliography**

**Shubham Bathwal**

**https://www.kaggle.com/datasets/shubhambathwal/flipkart-mobile-dataset?resource=download**

**DGOMONOV**

**https://www.kaggle.com/code/dgomonov/data-exploration-on-nyc-airbnb**