**Capstone Report**

**Title: Prediction of Client Subscriptions to Term Deposits**

**By: Bhavika Dobanaboina**

**KSU ID: 811323122**

**Project Overview:**

This project aims to predict client subscriptions to term deposits through analysis of direct marketing campaign data from a Portuguese bank. The objective is to develop a predictive model by exploring feature-subscription relationships. By optimizing marketing strategies, the goal is to boost term deposit subscriptions, enhancing campaign effectiveness and profitability.

**Purpose:**

The purpose is to develop a predictive model to determine if customers will subscribe to a term deposit using data from direct marketing campaigns conducted by a Portuguese banking institution. The focus is on analyzing campaign data to forecast client subscriptions to term deposits.

**Objective:**

Preprocess and analyze direct marketing campaign data to develop a predictive model for term deposit subscriptions. Explore feature-subscription relationships to optimize marketing strategies, enhancing the bank's success in attracting subscriptions and improving campaign effectiveness and profitability.

**Methods used in this project:**

For the classification problem, I would like to use Logistic Regression and Decision Trees.

confusion Matrix, ROC curve was used to evaluate model’s performance.

**Overview of Dataset:**

This dataset is collected from UCI Machine Learning Repository, it records phone-based direct marketing campaigns of a Portuguese bank, often requiring multiple contacts to determine if clients would subscribe to a bank term deposit ('yes') or not ('no').

It consists of 45212 rows and 17 columns, and variables like

|  |  |  |
| --- | --- | --- |
| **Variable names** | **Description** | **Numerical/**  **Categorical?** |
| Age | How old the person is. | Numerical |
| Job | What kind of work they do. | Categorical |
| Marital | Whether they're married, single, etc. | Categorical |
| Education | How much schooling they have. | Categorical |
| Default | Whether they have unpaid debts. | Categorical |
| Balance | How much money they typically have? | Numerical |
| Housing | Whether they have a loan for a house. | Categorical |
| Loan | Whether they have a personal loan. | Numerical |
| Contact | How the bank reached out to them (e.g., phone, email). | Categorical |
| Day of Week | What day of the week they were last contacted? | Numerical |
| Month | What month they were last contacted? | Categorical |
| Duration | How long did the last conversation last? | Numerical |
| Campaign | How many times they've been contacted in this marketing campaign? | Numerical |
| Pdays | How many days ago they were last contacted from a previous campaign? | Numerical |
| Previous | How many times they were contacted in previous campaigns? | Numerical |
| Poutcome | What happened in the previous marketing campaign (e.g., success, failure). | Categorical |
| Y | Whether they've agreed to a term deposit. | Numerical |

**Data cleaning:**

The data was examined carefully to identify missing values, inconsistencies, spelling errors and duplicates.

* **Missing Values:**

Percentage of missing values in each column was calculated:

A black and white text

Description automatically generatedNo missing value has been found.

**Finding no of rows and columns**



**Data Transformation:** Converting ‘y’ column to factor with 2 levels having values ‘yes’ and ‘no’. These 2 values indicate whether a customer would subscribe to term deposit or not.

A screen shot of a computer

Description automatically generated

**Converting categorical variables to Numerical variables:**

* Since 9 variables out of 17 variables are categorical, converting all these variables into individual numeri
* cal variables.
* Each class of every categorical variable has been transformed into a new variable with values 0 and 1.

A screenshot of a graph

Description automatically generated

**Data Exploration:**

* ***Descriptive statistics of all columns:***

A screenshot of a computer

Description automatically generated

* ***Finding the number of customers who have subscribed and not subscribed to the term deposit:***

A graph with a bar and a number of text

Description automatically generated with medium confidence

* We could observe that only 5000 customers have subscribed to the term deposit, out of 45000 customers. Clearly, the data is imbalanced, and a sampling technique will be used to sample an equal number of rows from both classes.
* ***Finding the age group of customers who would subscribe to term deposit:***

A screenshot of a diagram

Description automatically generated

* We could see that the customers with Median age of 30 have subscribed to the term deposit and customers between the age of 35 and 45 have not subscribed.
* ***Analyzing if the customers would subscribe to the term deposit or not based on their jobs:***

A graph showing different colors of stripes

Description automatically generated with medium confidence

* Among the customers who have not subscribed to the term deposits, most of them are unemployed, student, services, self-employed and housemaids. Interestingly, customers who have blue-collar and management qualifications have also shown no interest in subscribing to term deposits.
* ***Analyzing if the customers who have unpaid debts(default) have subscribed or not to term deposits:***

A blue rectangular object with a bar

Description automatically generated

A graph with a red rectangle

Description automatically generated

* It is obvious that the customers with unpaid debts have not subscribed to the term deposits.
* ***Analyzing if the customers with personal loans have subscribed to the term deposits or not:***

A graph with a bar and a number of columns

Description automatically generated with medium confidence

* Clearly, most of the customers who have personal loans have not subscribed to the term deposits.
* ***Analyzing the customers who have subscribed to the term deposits using data of previous campaigns:***

A graph with blue bars

Description automatically generated with medium confidence

* During the previous campaigns, the bank reached out to more customers between April and August.

A graph with a bar

Description automatically generated

* As we can see, the success rate of previous campaigns is very low.

A graph with a number of bars

Description automatically generated with medium confidence

* To most of the customers, the bank has reached out only 2 times.
* ***Finding the Correlation between all numerical variables:***

A diagram of a graph

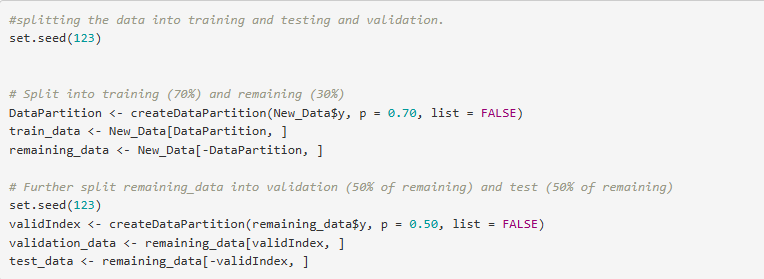
Description automatically generated with medium confidence

* From the correlation matrix, there is a positive correlation between balance, age and previous campaign to the number if days customers were reached out before.

On the other hand, there is a negative correlation between the campaign and the duration of the phone call.

**Data partition**

Our final dataset has 52 variables and 45212 observations. This dataset was partitioned into Training, validation and Test datasets with 70% of data being in the Training set, 15% in Validation and the remaining 50% in Test Dataset.



Training datasets will be used to train the models and Validation set will be used to tune the hyperparameters. After tuning hyperparameters, the model with the best evaluation metrics will be considered as the final model and that final model will be deployed on to the Test dataset (Unseen data).

**Model selection**

* As the target variable ‘y’ has 2 levels “yes” or “No”- this problem will be classified as a categorical type of problem with 2 classes. So, I chose the following algorithms:

**K-Nearest Neighbors (k-NN)** is a classification algorithm that predicts the class of a data point based on the majority class of its k nearest neighbors in the training data. In simpler terms, similar data points are likely to have the same class.

**Logistic Regression:** For binary classification problems, logistic regression is used to estimate the probability that a given instance belongs to a specific class. After predicting these probabilities, we set a classification threshold to categorize each instance into either class 0 or class 1. We then compare the predicted classifications to the original data to evaluate the model's performance.

Metrics such as accuracy, precision, specificity, and sensitivity are influenced by the chosen classification threshold. Therefore, selecting the appropriate threshold is crucial and should be based on the specific business problem.

Advantages of Logistic Regression:

* Simplicity and Interpretability: Logistic regression is easy to use and interpret. Coefficients directly show how attributes influence class likelihood.
* Probabilistic Predictions: Provides probabilities that can calibrate decisions and assess prediction confidence.
* Few Hyperparameters: Requires adjustment of only a few parameters like classification threshold or regularization strength, simplifying model selection.

**Decision Trees:**

Decision trees work by breaking down data into smaller groups based on input attributes, forming a tree-like structure to predict outcomes. This approach categorizes instances into different classes depending on their feature values.

Advantages of Decision Trees:

* Easy Understanding: Decision trees visually represent how classification decisions are made, making it straightforward to explain these decisions to non-experts.
* Handling Various Data Types: They handle both numerical and categorical data effectively, determining optimal thresholds for numerical features and creating subsets for categorical variables without needing extensive preprocessing.
* Scalability**:** Decision trees efficiently handle large datasets by splitting data into subsets based on feature values. This flexibility enables them to perform well across different data sizes and complexities.

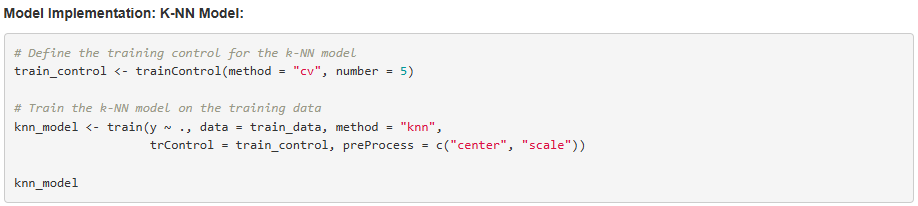
**Model building**

During this phase, we created and tested all three models to see how well they work with our data.

We'll choose the model that fits best based on how well it predicts outcomes. Once each model is ready, we'll use the Validation data to see how accurate they are.

For KNN, logistic regression and Decision Trees, this step will also help us adjust settings to make the models perform even better with new data.

**Implementation of KNN Model:**

****

KNN Model was run on the Training data and cross validation was used to train the model on every fold. Cross Validation with 9 folds gave the highest accuracy, so the final model used k=9.

A screenshot of a computer

Description automatically generated

**Evaluating the performance of the KNN Model:**

After deploying the model on the Validation data, Confusion matrix was used to test the accuracy of the predictions and the results are as below:

A screenshot of a computer

Description automatically generated

* Accuracy of the Model is 89.41%
* Sensitivity is 97%
* Specificity is 27%

As our problem is predicting if a customer will subscribe to the term deposit or not, it is important for the bank to identify all customers who would subscribe so that they can target their marketing campaigns towards those customers. It would be a loss for the bank if they don’t identify all the customers who might subscribe. Hence, instead of considering Accuracy as an evaluation metrics (which does not account for Type 1 and Type 2 errors), we would consider **Recall** as an evaluation metrics.

Recall = TP/TP+FN

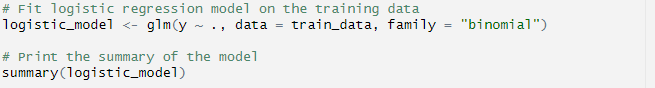
Recall = 216/ (216+ 577) =27%

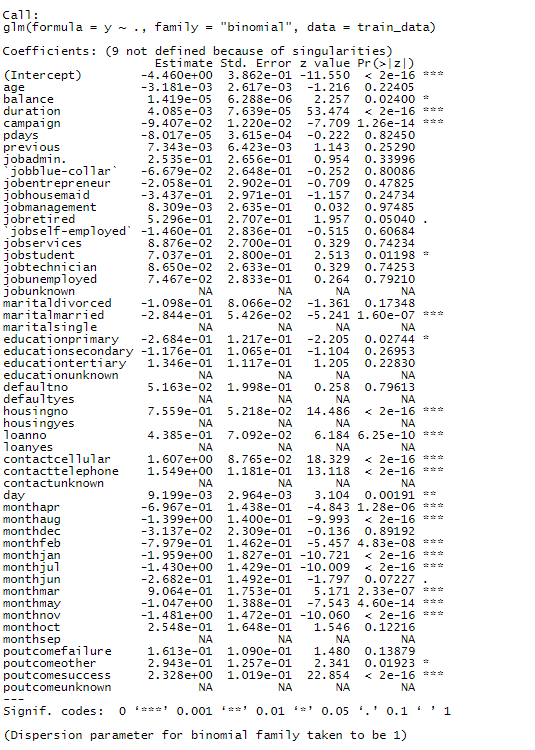
A screenshot of a computer

Description automatically generated

**Implementation of Logistic regression:**

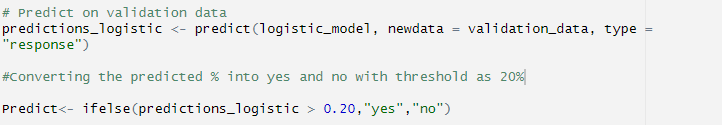
* Developed a logistic regression model to predict client subscription to term deposits.
* The model predicts the probability of a client subscribing based on various features.





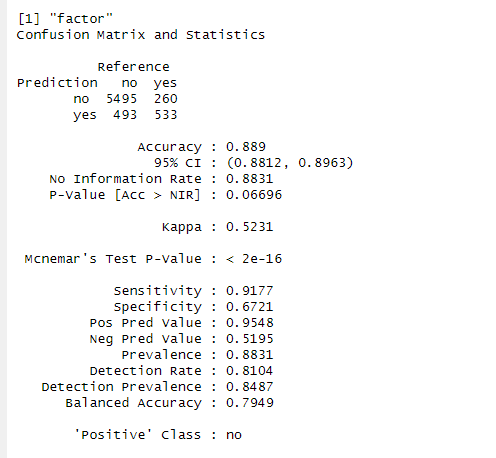
 Significant Predictors: Predictors with a P-value less than 5% are considered as significant predictors.

* Positive Influences: balance, duration, jobstudent, housingno, loanno, contactcellular, contacttelephone, monthmar, poutcomeother, poutcomesuccess.
* Negative Influences: campaign, maritalmarried, educationprimary, monthapr, monthaug, monthfeb, monthjan, monthjul, monthmay, monthnov.
* **Evaluating the performance of Logistic Regression Model:**



After training the model, it was deployed on to the Validation data. Predictions were % of a customer belonging to a specific class.

A classification threshold of 20% was considered appropriate as we want to predict every customer who would subscribe to a term deposit. Below are the results:



Like KNN Model, we are considering Recall as our evaluation metrics.

Recall = TP/TP+FN = 533/533+260 = 67.2%.

However, as Accuracy, Recall and Sensitivity can be changed based on Classification threshold, we are also considering ROC curve.

A graph of a curve

Description automatically generated

**Implementation of Decision Trees:**

**A close-up of a computer code

Description automatically generated**

Decision Trees model was implemented on the Training dataset and a method ‘Anova’ was used to identify the importance of variables.

Hyperparameters minsplit and maxdepth was also tuned to improve the performance of the model.

A computer code with numbers and symbols

Description automatically generated

* **Important Predictors**:
  + duration of the last contact
  + poutcome indicating the outcome of the previous marketing campaign
  + housingno
  + housingyes
  + monthmay
  + monthaug
  + age

A diagram of a tree

Description automatically generated

Top of Form

* **Evaluating the performance of Decision Trees Model:**Bottom of Form

This model was also deployed on to the Validation dataset to tune the hyperparameters so that the model generalizes well.

**A screenshot of a computer

Description automatically generated**

**Comparing the results of 3 Models:**

As we used Recall to be the evaluation metrics, we could see that the logistic regression model gave a Recall of 67% with a classification threshold of 20%.

Hence, we chose **Logistic Regression** model as the model which generalizes well on unseen data and deploys this final model on to the Test Dataset.

**Model Insights:**

* The duration of the last contact is the most significant predictor.
* The previous marketing campaign outcome (poutcomesuccess) significantly influences the prediction.
* The model provides probabilistic outcomes, aiding in decision-making based on confidence levels.

**Insights from Exploratory Data Analysis:**

* 5000 customers have subscribed to the term deposit, out of 45000 customers.
* Customers with a Median age of 30 have subscribed to the term deposit and customers between the age of 35 and 45 have not subscribed.
* Among the customers who have not subscribed to the term deposits, most of them are unemployed, student, services, self-employed and housemaids. Interestingly, customers who have blue-collar and management qualifications have also shown no interest in subscribing to term deposits.
* Most customers with unpaid debts and personal loans have not subscribed to the term deposits.
* During the previous campaigns, the bank reached out to more customers between April and August. However, the success rate of previous campaigns is very low.