SunbaseData Machine Learning Assignment

Customer Churn Prediction

Pushkara A

06-09-2023

Summary

Objective of this project is to predict whether customer will leave churn out or not, and to show case machine learning skill.

I have developed a machine learning model to predict which customer will churn out and which customer will remain. For this I have built several models like Logistic Regression algorithm, Random Forest, Decision Tree, K-Nearest Neighbor. I optimized the model Logistic Regression model as it has better accuracy than other models. After model optimization same model has been saved and used to deployed it to predict the customer will churn out or not in real time.

Introduction

This is a machine learning project, which was built to predict will customer churn out or not. Machine Learning model was built on Logistic Regression model, because after comparing accuracy of the Logistic Regression model with other model like Random Forest, Decision Tree, K-Nearest Neighbor, accuracy of model built using Logistic Regression was bit higher than other models.

Steps involved in building the predictive model:

- Importing of excel file.
- Data Preprocessing
- Feature Handling.
- Model Building.
- Model Optimization.

To import the dataset from the local machine pandas has been used.

After importing the dataset (excel file), Data preprocessing was adopted to explore the dataset like are null value present in the dataset, how many columns are there and their data types. It is important to explore the dataset before training the model so a developer can know does dataset has any null or nan values, any outliers in the dataset, because those values can hinder or decrease the accuracy value and it can lead to wrong prediction. Once the dataset has been explored next step is adopted that is one hot encoding and splitting the dataset into 2 parts namely training set and testing set.

Next adopted is feature handing. Once the dataset has been explored, and no null values or no outliers exists it is safe to do feature handling. In feature handling columns which has yes or no, male and female will be converted to Boolean values, because to train the machine learning model, values in the dataset or variable values should be int, float. Using the get_dummies variable we can convert values like “Yes” or “No” and “Male” and “Female” vales can be turned to Boolean values, by passing drop_first = True as the parameter to the get_dummies function, we can remove the column which could be just extra column. After converting the values to Boolean values, we can adopt StandardScaler() function to convert the integer value of different column to same unit, so machine learning model can get trained better and give higher accuracy.

After feature scaling, next step is “Model Building”. In this step I have built several models namely “Logistic Regression”, “Random Forest”, “Decision Tree”, “K-Nearest Neighbor” and their accuracies were 50.43, 49.56, 49.94, 49.77 before model optimization.

After “Model Building” I adopted the “Model Optimization” because accuracies were too low, with the help of the GridSearchCV I have improved the accuracy of the model from 50.43 to. I have adopted 2 ideas to improve the accuracy of the logistic regression model, they are “Cross-Validation” and “Hyperparameter tuning”.

After building a model I saved the model in model.sav format for feature use. I had built a basic website so I can check a customer in future will churn out or not in real time.

Snapshots

Overview

Overview | Alerts 5 | Reproduction

Dataset statistics

Number of variables	9
Number of observations	100000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.9 MiB
Average record size in memory	72.0 B

Variable types

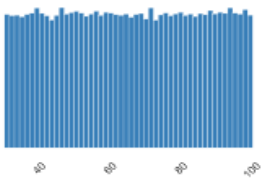
Numeric	5
Categorical	4

Figure 1.0: Dataset Statistics and Variable types

Monthly_Bill

Real number (ℝ)

Distinct	7001	Minimum	30
Distinct (%)	7.0%	Maximum	100
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	65.053197	Memory size	781.4 KiB



More details

Figure 2.0: Indetail of variable Monthly_Bill

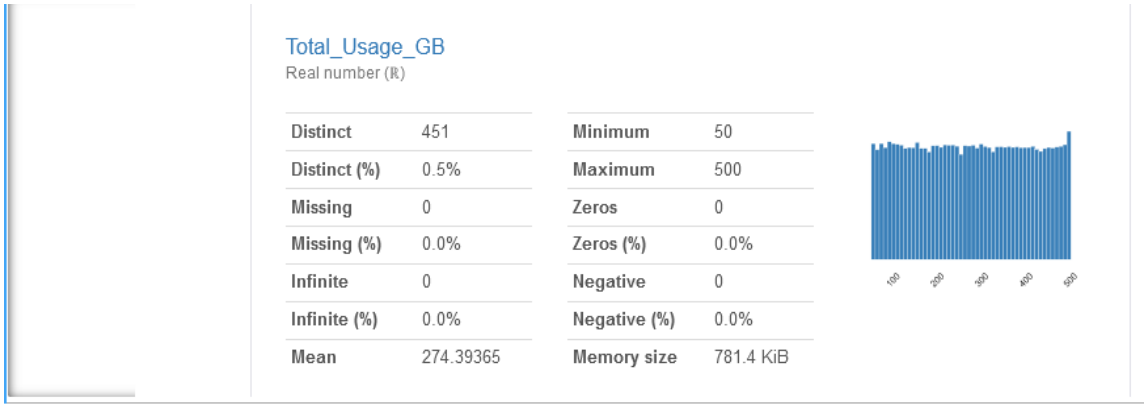


Figure 3.0: Indetail of variable Total_Usage_Bill

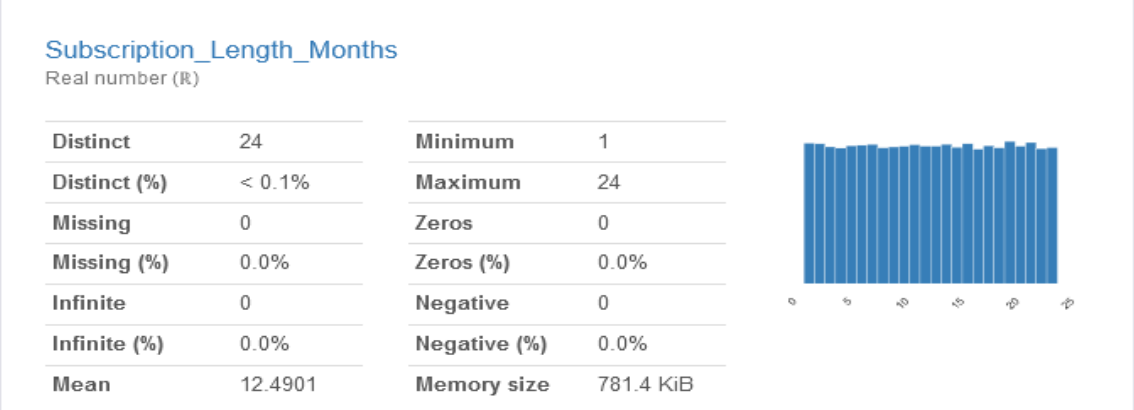


Figure 4.0: Indetail of Variable Subscription_Length_Months

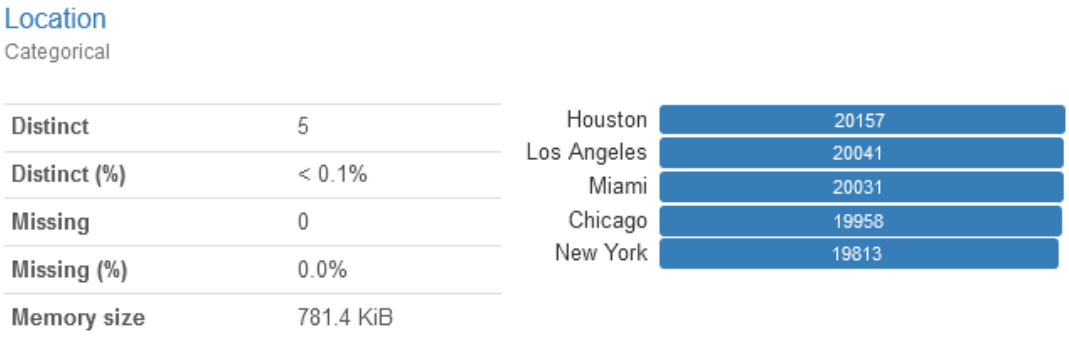


Figure 5.0: Indetail of Variable Location

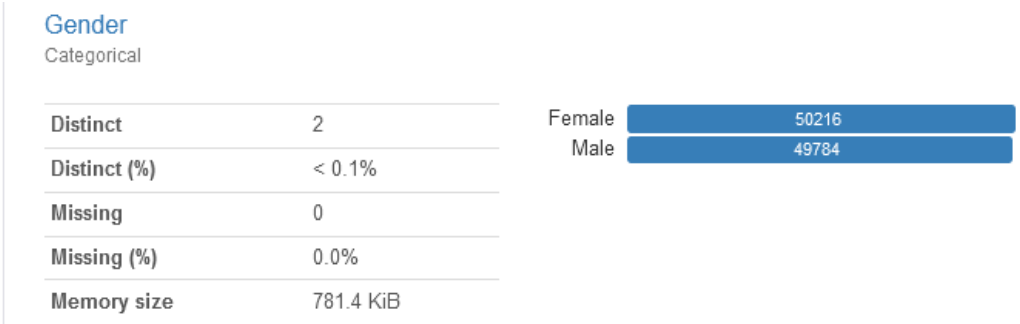


Figure 6.0: Indetail of Variable Gender

Age

Real number (ℝ)

Distinct	53	**Minimum**	18
Distinct (%)	0.1%	**Maximum**	70
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	44.02702	**Memory size**	781.4 KiB

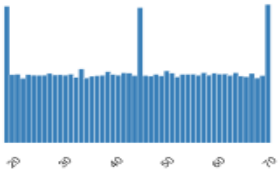


Figure 7.0: Indetail of Variable Age

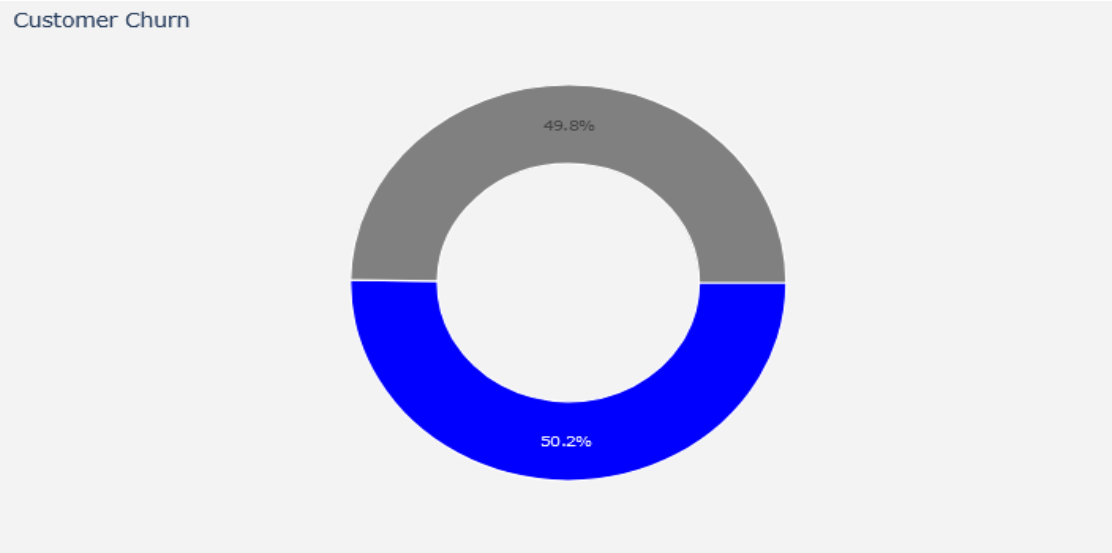


Figure 8.0: Mean of variable customer churn

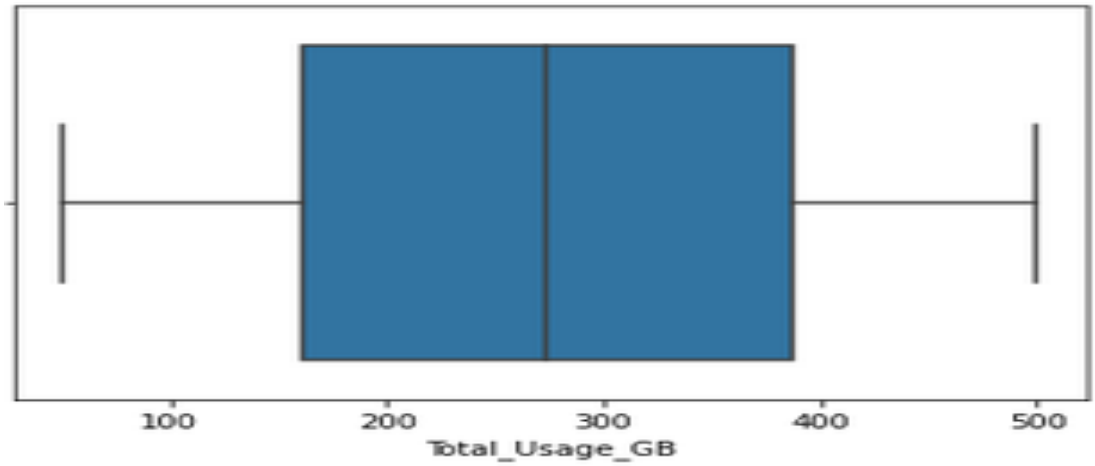


Figure 9.0: No outliers in variable Total_Usage_GB

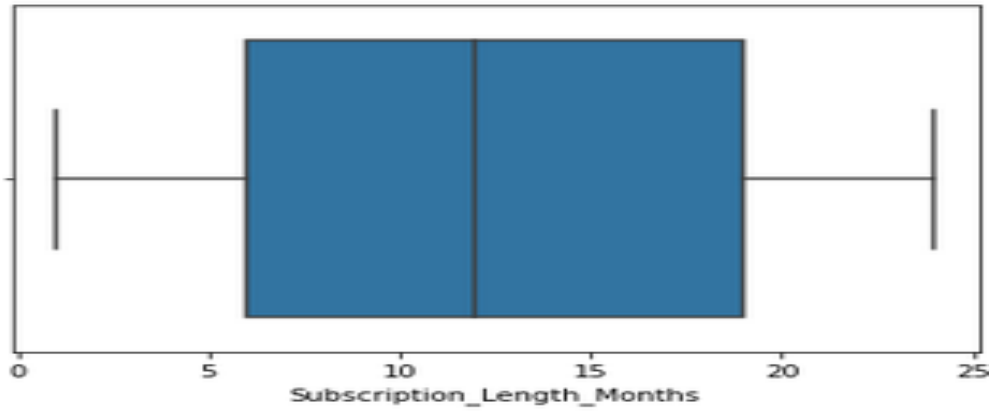


Figure 9.0: No outliers in variable Subscription_Length_Months

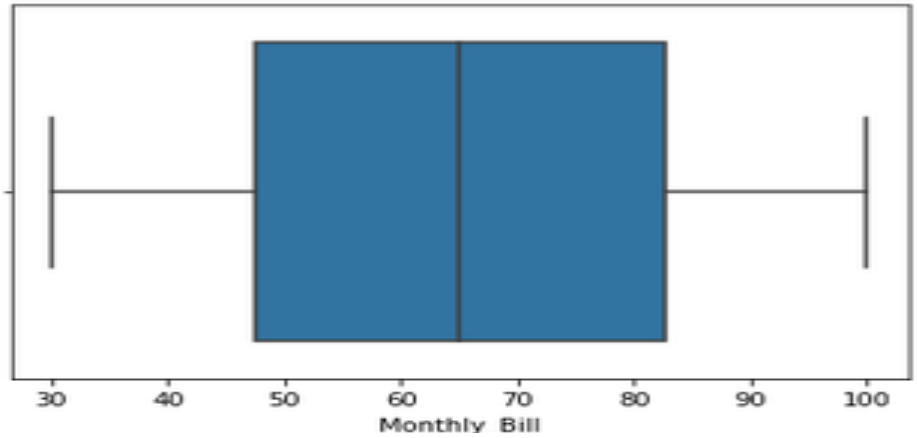


Figure 10.0: No outliers in variable Monthly_Bill

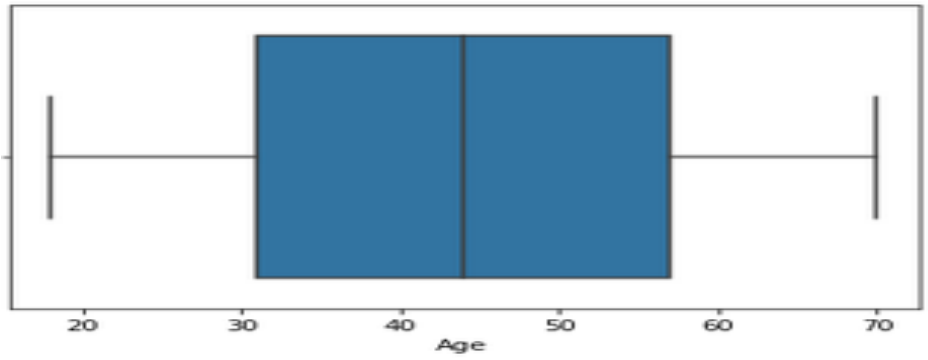


Figure 11.0: No outliers in variable Age

Links:

- GitHub: https://github.com/pushkara20/Customer_Churn_Prediction