Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

DERIVING INSIGHTS INTO CUSTOMER CHURN

NAME: DHRUV VARMA

SAPID: 60009210043

Guided By

CHAPTER 1: INTRODUCTION

Customer churn refers to the phenomenon of customers discounting their relationship with the business or organization. It is a significant problem for businesses across industries as it can lead to prevent new laws and increased costs associated with acquiring new customers. Deriving insights into customer churn involves analysing customer data to identify patterns and factors that contribute to customer attrition. By understanding the reasons why customers leave business can take steps to prevent sure and improve customer retention.

CHAPTER 2: DATA DESCRIPTION

For the said problem statement, a dataset based on a Telecom service is taken into consideration. The source of the dataset is: Kaggle (https://www.kaggle.com/datasets/barun2104/telecomchurn?select=telecom_churn.csv) This dataset would be suitable for solving the problem of customer churn since the dataset contains relevant attributes, is of large sample size, based on realworld data, labelled, and of high quality. About the dataset: Each row represents a customer and each column contains attributes related to customer as described below:

Churn: If customer has cancelled the service or not

AccountWeeks: Number of weeks a customer has had an active account

ContactRenewal: If customer has recently renewed contract

DataPlan: If customer has a data plan

DataUsage: Gigabytes of monthly data usage

CustServCalls: Number of calls into customer service

DayMins: Average daytime minutes per month

DayCalls: Average number of daytime calls

MonthlyCharge: Average monthly bill

OverageFee: Largest overage fee in last 12 months

CHAPTER 3: DATA ANALYSIS

85.5% of the total customers churn while 14.5% don’t. ContractRenewal and DataPlan attributes are two important attributes in customer churn. If these two attributes are "1", the probability of customer churn is low. ContractRenewal's impact is greater than DataPlan. Increasing DataUsage decreases the probability of churn customer and decreasing other attributes decreases the probability of churn customer. I classified and averaged the Churn attribute to create the chart.

CHAPTER 4: DATA MODELLING

Standard Scaler was used to transform the values of the training and testing data. Machine Learning models like: Random Forest, Decision Tree, Gradient Boost and KNN. The prediction probabilities and accuracy scores are compared from all the Machine Learning algorithms. An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

The ROC curve for all the algorithms is plotted and Gradient Boost and Random Forest are proved to be the most efficient ones in this case.

CHAPTER 4: CONCLUSION

The telecom service must work on the following changes in order to avoid customer churning:

1. More Customer Service Calls mean the customer is more prone to churning
2. Optimize the price of talk time for customers
3. Introduce a data plan to those customers who are using data without data plans as soon as possible
4. Introduce exciting data plans for a few customers and optimizing its price can help retain customers

Gradient Boost and Random Forest are proved to be the best algorithms giving desired results with highest accuracy and a similar plot on the ROC Curve. Gradient Boost, Decision Tree, KNN are hypermeter tuned while Random Forest is not hyperparameter tuned.

Gradient Boost Algorithm is used here to decreased the Bias error.

Random Forest does not make any assumption about the data or its distribution. Hence, it generally requires minimal data transformations. Since Random Forest makes use of random subsets of features and hence it can perform quite well with a high dimensional dataset.