CHAPTER 2

TASK PERFORMED

Introduction

The telecommunications industry across the world is becoming one of the major sectors and consequently the technical growth and the ever-developing operator number increased the level of competition. Telecom firms are making an effort to subsist in this rivalry market and some measures have been formulated to bring in huge amount of revenues. To enhance the retention time of customers it is important for the companies to lessen the possibility of churn of customer, referred to as "the movement of customer from one service provider to another service provider (Ascarza, et al 2016). The churn of customers' is considered a major issue in service fields with increased cutthroat services (Ahmad, et al 2019). Many studies (Umayaparvathi and Iyakutti 2016; Jutla and Sivakumar 2005) emphasized that machine learning applications are increasingly effective to predict this situation. The underlying principle of customer churn prediction in terms of telecom industry is to calculate subscribers approximately who literally feel like to leave from a company they used so far and suggest solutions to prevent considerable churns. Recently, making an estimation of churners before they quit has become necessary in the environment of stiff competition amongst companies. The major role that the telecom industry plays made it all the more significant to build prediction mechanisms alongside the lines of churn prediction. Few studies exhibit the significance of the user retains in this industry. One of the studies exhibits that 1% intensification in the customer retains movement might essentially give rise to the 5% increase in the entire shares of the companies (Kisioglu and Topcu 2010). Huang, et al (2012) predicted that in terms of telecommunication sector, the monthly ratio of customer churn is 2.2% (Yildiz and Albayrak 2018) and the yearly rate of customer churn was 27%. Customers' retain in a telecom industry has already become a nightmare as a result of increased competitive services. Brandusoiu and Toderean (2016) proposed an advanced data mining method in order to find customer churn using algorithms of machine learning namely NN (Neural Networks) and SVM (Support Vector Machine). The findings 2 emphasized that machine learning algorithm performs better in predicting customers' churn. He et al. (2009) used a machine learning model to resolve the issue of churn among customers in big telecommunication firms. Huang et al. (2015) investigated the issue of churn among customers' in the platform of big-data, which intended to emphasize that big data significantly improve the progression of estimating the churn based on the variety, velocity and volume of the records. Ahmad, et al (2019) specified that the social network

analysis application attributes improve the outcomes of estimating the churn in telecommunication sector.

Problem statement :

The retention and acquisition of users are the major concerns in telecom industry. The fast growth of marketplace in every business is giving rise to increased subscriber base. Accordingly, companies have recognized the significance of retaining the customers who is on hand. It has become necessary for service-providers to reduce the churn rate of customers since the inattention might negatively influence profitability of the company. Churn prediction contributes to identify those users who are likely to switch a company over another. Telecom is enduring the problem of ever-increasing churn rate. Accordingly, the current study employs machine learning algorithm on big-data platform. Machine learning algorithm techniques facilitate these telecom firms to be protected with efficient approaches for lessening the rate of churn. Silent churn is one type which is considered complicated to predict since there might have such kind of users who might probably churns in the near future. It must be the aim of the decision-maker and advertisers to lessen the churn ratio since it is a recognized fact that comparatively existing customers are the most beneficial resources for companies than acquiring new one.

Objective:

The primary and secondary objectives of the study are as follows:

2.2.1 Primary objectives

i. To explore the customer churn prediction in telecom using machine learning in big data platform

2.2.2 Secondary objectives

i. To investigate the impact of customer churn in telecom industry as a whole
ii. To discuss the significance of customer churn models in telecom industry
iii. To compare the algorithms that are effective in reducing churn rate in telecom companies.

Limitations :

i. The current study is limited to telecom industry only

ii. The study does not use other techniques rather than machine learning technique

CHAPTER 3

LITERATURE SURVEY

3.1 Introduction:

A large number of researches in the subject of churn prediction are being investigated employing various statistical and machine learning algorithms since a decade. This chapter deals with the recent and most important publications on churn prediction in telecom industry in the recent period.

3.2 Impact of customer churn in telecom industry

Tanneedi, (2016) pointed out that customer churn has become a dreadful problem for the telecom industry since customers never have a second thought to leave if they don't exactly get what they are expecting. There is no benchmark model that deals with the churning issues of telecom companies precisely. The study emphasized that Big Data analytics with machine learning are considered effective as means for identifying churn. The current study makes an effort to predict customer churn in telecom employing Big Data analytics. Statistical analyses and machine learning application such as Decision trees (DT) have been used for three different datasets. From the analytics of DT, decision trees with accuracy rate of 52%, 70% and 95% have been obtained for three different data sources correspondingly. The findings pointed out that the more the quality and volume increases, the lesser the annoyance and possibility of churn can be expected in telecom industry.

Huang, (2015) exhibited in terms of telecom industry churn prediction can easily be done with big data and with 3V's such as volume, variety along with velocity. Findings emphasize that the performance of prediction has been enhanced considerably by employing a big amount of training data, a huge number of features from both operations and business support systems, as well as an increased velocity of processing new data. The study has deployed this prediction technique of churn in one of the largest mobile network operators in China. From a large number of active customers, this technique could impart prepaid customers set who are about to churn, holding 0.96 accuracy rate 8 for the top 50000 estimated churners in the list. The operations of automated matching retention with the focused essential churners considerably increase their rates of recharge, bringing about a big value

3.3 Importance of churn prediction model in telecom industry

Amin, et al (2016) make an attempt to develop the model of churn prediction in the telecom sector. The study presents a technique of rule-based decision-making, on the basis of RST (rough set theory), to obtain significant rules of decision linked with non churn and customer churn. The proposed technique efficiently executes categorization of churn from non-churn users, together with prediction of those users who will churn or might likely to churn in the near future. Experiential findings exhibit that rough set theory based on Genetic Algorithm (GA) is the most effective method for obtaining inherent knowledge in decision-based rules form from the publicly accessible, benchmark telecom information. Besides, comparative results exhibit that proposed technique provides a worldwide best solution for churn prediction in the telecom industry, when benchmarked against some high-tech techniques. In the end, the study exhibits that how attribute-level analysis could contribute to develop an effective policy of customer retention that can form an essential part of strategic process of decision-making in the telecom industry.

3.4 Comparison of various algorithms in customer churn

Yabas and Chankya (2013) specified that customer churn has become a matter of great concern of customer care management for the majority of the mobile service providers as a result of its associated costs. The current study describes our work on customer churn investigation and assessment for such services. The study has employed data mining algorithms to precisely and effectively predict subscribers who will change and ultimately switch to another service provider or competitor for the same or related service. The study makes an effort to identify alternative techniques which could correspond or enhance the recorded high scores with more effective and as well practical usage of resources. The paper also focuses on collection of meta-classifiers that have been investigated separately and chosen in line with their performances.

CHAPTER-4

PROPOSED SYSTEM

In a business scenario predicting customer churn is where a firm is attempting to retain customer which is much probable to leave the services. For reducing the rate of churn this study classifies which customers are much going to churn probably and which will not churn probably. Since obtaining new customers is challenging it is essential to retain present customers. Churn can be decreased by examining the essential customers past history systematically. Huge amount of data is managed about the customers and on carrying out appropriate examination on the same it is feasible to find probable customers that might churn. The data that is feasible can be examined in varied ways and thereby offers different ways for operators to imagine the churning of customers and avoid the same.

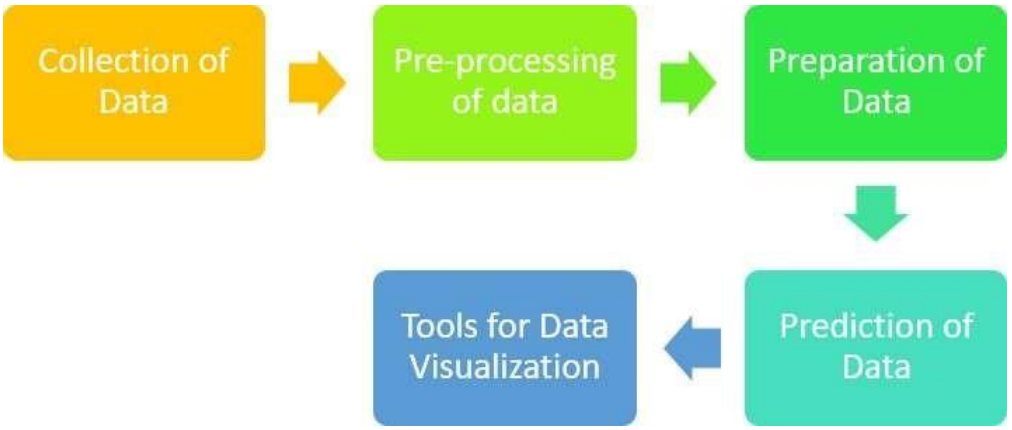


Figure 4.1: Steps used for Proposed System.

From the above figure 1 the steps used for the proposed system are collection of data, Pre-Processing of data, Preparation of data, Prediction of data and Tools for data visualization. The steps are explained below briefly:

- **Collection of Data**

  The data that is feasible for analysis in telecommunication dataset has been used and the prediction has been carried out for the same.

- **Pre-processing of data**

  The pre-processing of data involves 3 steps namely data cleaning, feature selection and data transformation. Each step is explained below:

Data transformation comprises of two explanatory variables which can be transformed from binomial form into binary form to be much applicable for the chosen models.

The data cleaning step involves missing data imputation or handling. Some of the chosen algorithms cannot manage missing data that is why missing value can be transformed by median, mean or zero. However, the replacement of missing data by computed value statistically is a better choice. The used set of data involves missing values in certain numerical variables and two categorical variables.

Before training of model, feature selection is one of the most essential factors that can influence the model's performance.

- **Preparation of data**

The main purpose of preparation of data is to improve the quality of data and enhance the performance of data analysis. The preparation of data requires to be undertaken in a much iterative way until a conclusive result is met. The processes of preparation of data involves numerical variables discretization, missing values imputation, selection of feature of most informative variables, transformation from one discrete value set to another and derivation of new variables. The process of imputation includes changing the missing values with whole data based on an estimate from finished values. Making new variables from the information is based on transformation and discretization. Two new variables were formed to estimate the voice and transformation in usage of data. Before the data can be examined the data must be cleaned and keep it prepared so that the desired outputs can be derived from it. Data must be clean so that the errors and redundancy can be eliminated because having such information will lead to improper outcomes as well. In this study a churn examination has been used on telecommunication data here the agenda is to know the feasible consumers that might churn from service provider. The end outcome provides the churn probability for each consumer. To perform the churn examination the logistic regression is used. Logistic regression is a statistical approach where the output variable is categorical rather than continuous. Logistic regression restricts the prediction to be one and zero interval

- **Prediction of data**

  The organization is concerned in the final product and it is very essential to indicate their outcome in a graphical representation such a way that is understandable and the output helps organization makes the required predictions which in turns brings profit. There are several components that helps accomplish the same.

- **Tools for Data Visualization**

  The best way to acquire the message across is to use the tools of visualization by indicating data visually it is feasible to uncover the essential patterns and the patterns that would be ignored if the statistics alone is considered. In this study Power BI is a component that is used to perform data visualization. Power BI is a business analytics component which is offered by Microsoft using which reports can be made. In this study the data is cleaned already and the output is populated in a file named prediction which will be helpful to visually display how the data seems and the effect of it.

CHAPTER 5

IMPLEMENTATION

1.DATA COLLECTION

```
# IMPORTING NECESSARY PACKAGES
import numpy as np  # array/matrix operation
import pandas as pd  # for handling dataframes
import matplotlib.pyplot as plt  # visualizations
import seaborn as sns  # advanced visualis
import streamlit as st
st.set_option('deprecation.showPyplotGlobalUse', False)
data = pd.read_csv(r"C:\Users\Prerana Biradar\Downloads\Telecom Churn (1).csv")
print('[info] data loaded successfully')
# 2. DATA PRE-PROCESSING
# 2a.BASIC EDA
print('checking for NaN values', data.columns[data.isna().any()])
data = data.dropna()  # removing rows that containg
print(data.shape)
# 2b.ADVANCE EDA
import pandas as pd
State_unique = data['State'].unique().tolist()
State_index = np.arange(1, len(State_unique) + 1, 1)
State_to_num = {i: j for i, j in zip(State_unique, State_index)}
data['State'] = data['State'].map(State_to_num)data.head(3)
data['International plan'] = data['International plan'].map({'No': 1, 'Yes': 2})
data['Voice mail plan'] = data['Voice mail plan'].map({'Yes': 1, 'No': 2})
churn_unique = data['Churn'].unique().tolist()
churn_index = np.arange(1, len(churn_unique) + 1, 1)
churn_to_num = {i: j for i, j in zip(churn_unique, churn_index)}
data['Churn'] = data['Churn'].map(churn_to_num)
data = data.dropna()  # removing rows that containg
print(data.shape)

import streamlit as st
x = data.iloc[:, :-1].values  # choosing input col
```

```
y = data.iloc[:, -1].values  # choosing output col
print('[info]data segragated to x and y successfuly')
# splitting
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2))
import streamlit as st  # to build gui webpage
from streamlit_option_menu import option_menu
st.title('INTEGRATION GUI with ML MODELS')
State = st.number_input('ENTER NUMBER OF STATE', step=1)
State = st.number_input('ENTER NUMBER OF STATE', step=1)
Account_length = st.number_input('ENTER ACCOUNT LENGTH', step=1)
Area_code = st.number_input('ENTER THE AREA CODE', step=1)
International_plan = st.number_input('ENTER INTERNATIONAL PLAN', step=1)
Voice_mail_plan = st.number_input('ENTER VOICE MAIL PLAN', step=1)
Number_vmail_message = st.number_input('ENTER NUMBER VMAIL MESSAGE', step=1)
Total_day_minutes = st.number_input('ENTER TOTAL DAY MINUTES', step=1)
Total_day_calls = st.number_input('ENTER TOTAL DAY CALLS', step=1)
Total_day_charge = st.number_input('ENTER TOTAL DAY CHARGE', step=1)
Total_eve_minutes = st.number_input('ENTER TOTAL EVE MINUATES', step=1)
Total_eve_calls = st.number_input('ENTER TOTAL EVE CALLS', step=1)
Total_eve_charge = st.number_input('ENTER TOTAL EVE CHARGE', step=1)
Total_night_minute = st.number_input('ENTER TOTAL NIGHT MINUTE', step=1)
Total_night_calls = st.number_input('ENTER TOTAL NIGHT CALLS', step=1)
Total_night_charge = st.number_input('ENTER TOTAL NIGHT CHARGE', step=1)
Total_intl_minutes = st.number_input('ENTER TOTAL INTL MINUTES', step=1)
Total_intl_charge = st.number_input('ENTER TOTAL INTL CHARGE', step=1)
Customer_service_calls = st.number_input('ENTER CUSTOMER SERVICE CALLS', step=1)
new_user_input = [ [State, Account_length, Area_code, International_plan, Voice_mail_plan,
Number_vmail_message, Total_day_minutes,Total_day_calls, Total_day_charge,
Total_eve_minutes, Total_eve_calls, Total_eve_charge, Total_night_minute,Total_night_calls,
Total_night_charge, Total_intl_minutes, Total_intl_calls, Total_intl_charge,
 Customer_service_calls]]
with st.sidebar:  # ading option menu to sidebar
 model_selection = option_menu('SELECT A MODEL', options=['LOGISTIC REGRESSION',
'KNN', 'SVM', 'XGBOOST','GRADIENT BOOST CLASSIFIERS','MODEL PARAMETERS'])
```

```
submit_button = st.button("SUBMIT")
if submit_button:
    from sklearn.linear_model import LogisticRegression
   logistic_regression_model = LogisticRegression(max_iter=1500)  # inintialization of model
    logistic_regression_model.fit(x_train, y_train)  # training log regression on preprocessed data
   st.subheader('MODEL DIAGNOSIS')
    logistic_model_output = logistic_regression_model.predict(new_user_input)
    if logistic_model_output[0] == 2:
        st.info('customer might churn')
    if logistic_model_output[0] == 1:
        st.success('customer will not churn')
# 4a.MODEL EVALUTION : LOGISTIC REGRESSION
logistic_regression_model_predicted = logistic_regression_model.predict(x_test)
# defining actual values
logistic_regression_model_actual = y_test
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics import confusion_matrix, classification_report
st.subheader('MODEL PARAMETERS')
logistic_accuracy = accuracy_score(logistic_regression_model_actual,
logistic_regression_model_predicted))
logistic_precision = precision_score(logistic_regression_model_actual,
logistic_regression_model_predicted)
st.error(f'precision of logistic regression model is :{logistic_precision}')
logistic_recall = recall_score(logistic_regression_model_actual,
logistic_regression_model_predicted)
st.success(f'recall of logistic regression model is :{logistic_recall}')
logistic_f1 = f1_score(logistic_regression_model_actual, logistic_regression_model_predicted)
st.warning(f'f1 score of logistic regression model is :{logistic_f1}')
st.subheader('CLASSIFICATION REPORT')
logistic_regression_classification_report = classification_report(logistic_regression_model_actual,
logistic_regression_model_predicted,
output_dict=True)
st.dataframe(logistic_regression_classification_report, width=1000)if model_selection == 'KNN':
from sklearn.neighbors import KNeighborsClassifier  # importing the algorithm
knn_model = KNeighborsClassifier()  # inintialization of model
    st.subheader('MODEL DIAGNOSIS')
```

```
    knn_model_output = knn_model.predict(new_user_input)
    if knn_model_output[0] == 2:
        st.info('customer might churn')
    if knn_model_output[0] == 1:
        st.success('customer will not churn')
 # 4b.MODEL EVALUTION : knn REGRESSION
    knn_model_predicted = knn_model.predict(x_test)
    knn_model_actual = y_test
     from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
    from sklearn.metrics import confusion_matrix, classification_report
    st.subheader('MODEL PARAMETERS')
    knn_accuracy = accuracy_score(knn_model_actual, knn_model_predicted)
    st.info(f'accuracy of logistic regression model is :{knn_accuracy}')
    knn_precision = precision_score(knn_model_actual, knn_model_predicted)
    st.error(f'precision of logistic regression model is :{knn_precision}')
    knn_recall = recall_score(knn_model_actual, knn_model_predicted)
    st.success(f'recall of logistic regression model is :{knn_recall}')
     knn_f1 = f1_score(knn_model_actual, knn_model_predicted)
    st.warning(f'f1 score of logistic regression model is :{knn_f1}')
    st.subheader('CLASSIFICATION REPORT')
   knn_classification_report = classification_report(knn_model_actual, knn_model_predicted,
output_dict=True)
    st.dataframe(knn_classification_report, width=1000)
  st.subheader("CONFUSION MATRIX")
    import seaborn as sns
    knn_confusion_matrix = confusion_matrix(knn_model_actual, knn_model_predicted)
    sns.heatmap(knn_confusion_matrix, color='b')
    st.pyplot()
if model_selection == 'SVM':
   from sklearn.svm import SVC  # importing the algorithm
   svm_model = SVC()  # inintialization of model
    svm_model.fit(x_train, y_train)  # training log regression on preprocessed data
   st.subheader('MODEL DIAGNOSIS')
    svm_model_output = svm_model.predict(new_user_input)
```

```
    if svm_model_output[0] == 2:
        st.info('customer might churn')
    if svm_model_output[0] == 1:
        st.success('customer will not churn')
    svm_model_actual = y_test
       from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
    from sklearn.metrics import confusion_matrix, classification_report
  st.subheader('MODEL PARAMETERS')
    svm_accuracy = accuracy_score(svm_model_actual, svm_model_predicted)
    st.info(f'accuracy of svm model is :{svm_accuracy}')
    st.subheader('CLASSIFICATION REPORT')
    svm_classification_report = classification_report(svm_model_actual, svm_model_predicted,
output_dict=True)
    st.dataframe(svm_classification_report, width=1000)
     st.subheader("CONFUSION MATRIX")
    import seaborn as sns
     svm_confusion_matrix = confusion_matrix(svm_model_actual, svm_model_predicted)
    sns.heatmap(svm_confusion_matrix, color='b')
    st.pyplot()
   if model_selection=='GRADIENT BOOST CLASSIFIERS':
    from sklearn.model_selection import train_test_split
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3)
    from sklearn.ensemble import GradientBoostingClassifier
    from sklearn.metrics import r2_score
    grad_boost = GradientBoostingClassifier()
    grad_boost.fit(x_train, y_train)
       y_pred = grad_boost.predict(x_test)
      print(y_pred)
     st.info(f'RMSE : {np.sqrt(((y_test - y_pred) ** 2).sum() / len(y_test))}')
     st.error(f'Variance score:%2f' % r2_score(y_test, y_pred))
     if model_selection == 'XGBOOST':
    data['Churn'] = data['Churn'].map({True: 1, False:
      from sklearn.model_selection import train_test_split
     x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3)
```

```
   from xgboost import XGBClassifier
 xgb = XGBClassifier()
xgb.fit(x_train, y_train)
y_pred = xgb.predict(x_test)
y_actual = y_test
from sklearn.metrics import accuracy_score
 accuracy_score(y_actual, y_pred)
if model_selection == 'MODEL PARAMETERS':
st.subheader('ACCURACY')
# accuracy
labels = ['LOGISTIC REGRESSION', "KNN", "SVM",]
accuracy = [0.85, 0.85, 0.84]
import seaborn as sns
            sns.barplot(x=labels, y=accuracy)
               st.pyplot(st.subheader('PRECISION')
              # precison
                  labels = ['LOGISTIC REGRESSION', "KNN", "SVM"]
      precision = [0.86, 0.87, 0.9]
      import seaborn as sns
      sns.barplot(x=labels, y=precision)
      st.pyplot()
    st.subheader('RECALL SCORE')
      # accuracy
      labels = ['LOGISTIC REGRESSION', "KNN", "SVM"]
      recall = [0.97, 0.96, 0.9]
      import seaborn as sns
      sns.barplot(x=labels, y=recall)
      st.pyplot()
    st.subheader('F1 SCORE')
      # accuracy
      labels = ['LOGISTIC REGRESSION', "KNN", "SVM"]
```

CHAPTER 5

TASK PERFORMED

5.1 About dataset

This study uses Kaggle website for dataset in predicting and analyzing churn. Kaggle is a site and community for hosting ML competitions. Rivalry ML can be a best way to practice and develop their skills as well as explain their abilities. Kaggle permits users to publish and find sets of data and describe models in a web-based data science surroundings, perform with other scientists of data and ML engineers and enter competition to resolve the barriers of data science. The pre-processing steps used for dataset are: 1) first the spaces are replaced with values of null in the column of total charges; 2) the values of null are reduced from the column of total charges which comprises 15 percent missing data; 3) then the data is converted to the type of float; 4) after than no internet service is replaced to no for the following columns: Device Protection, Streaming TV, Online Security, Tech Support, Streaming Movies and Online Backup; 5) the values for Senio Citizen is replaced with 0 as No and 1 as Yes; 6) Then the categorical column is made into Tenure; 7) After than the churn and non-churn customers are separated; 8) Finally the numerical and categorical columns are separated.

5.2 Machine Learning-Based Approaches

Below is a brief overview of popular machine learning-based techniques for anomaly detection.

1. **Logistic Regression**

   Logistic regression is the proper model of regression analysis to utilize when the dependent variable is binary. Logistic regression is a predictive examination used to describe the relation between an independent variable set and dependent binary variable. For churn of customer logistic regression has been used to estimate the probability of churn as a function of customers characters or variables set (Sahu et al, 2018). According to Hassouna et al (2016) Logistic regression is also used to find the customer churn occurrence probability. Logistic regression is based on a mathematically oriented method to examine the impact of variables on others. Prediction is made by comprising a group of equations linking values of input with the output field. The mathematical formulas for logistic regression are:

$$P(b = 1|a1, \ldots\ldots, am) = F(b) \text{ ---------- (1)}$$

$$F(b) = 1\ (1+e{-}b) \text{ ---------- (2)}$$

$$b = \beta 0 + \beta 1a1 + \beta 2a2 + \cdots + \beta mam \text{ ----------(3)}$$

Where $\beta 0$ is a constant, b is every individual e target variable, y is a binary label class one or zero, $a1.\ a2, \ldots\ldots, +am$ is the variables of predictor for every customer e from which a is to be predicted.

The datasets of customer are examined to comprise the equations of regression and an evaluation process for every customer in the set of data is then carried out. A consumer is at a risk of churn if the value of p for consumer is larger than a predefined value.

2. K-Nearest Neighbor

According to Keramatia et al (2014) K-Nearest Neighbor is one of the most useful and applicable non parametric algorithms of learning. K-Nearest Neighbor is also referred as lazy algorithm that is entire data of training is used at the phase of testing. There is no phase of training and entire points of data are used directly in the testing phase so these entire points required to be employed when it must be tested. K-Nearest Neighbor utilizes the distance between records so as to utilize it for classification. In order to estimate the distance between points K-Nearest Neighbor considers that these points are multidimensional or scalar vectors in feature space. All points of data are vectors of feature space and the label will refer their classes. The easiest case is when the class labels are binary but still it is useful on arbitrary class numbers. In K-Nearest Neighbor one parameter requires to be tuned. K is the number of neighbors/instances that are regarded for instance labeling to some class. The cross validations were carried out using different values of k. K-Nearest Neighbor does not attempt to build an internal structure and computations are not carried out until the time of classification. K-Nearest Neighbor stores only examples of the training information in feature space and the class of an example is decided based on most of the votes from its neighbors. Instance is labelled with class which is much similar among its neighbors. K-Nearest Neighbor decides neighbors based on hamming for categorical variables and distance using Manhattan, Murkowski and Euclidian measures of distance for continuous variables.

Estimated distances are employed to recognize training instances set that are nearest to the new point and allot label from these. Despite its simplicity K-nearest neighbor have been used to different kinds of application. For churn K-nearest neighbor is used to examine if a customer churns or not based on features proximity to consumers in every classes (Keramati et al, 2014).

3. Support Vector Machine-Based Anomaly Detection

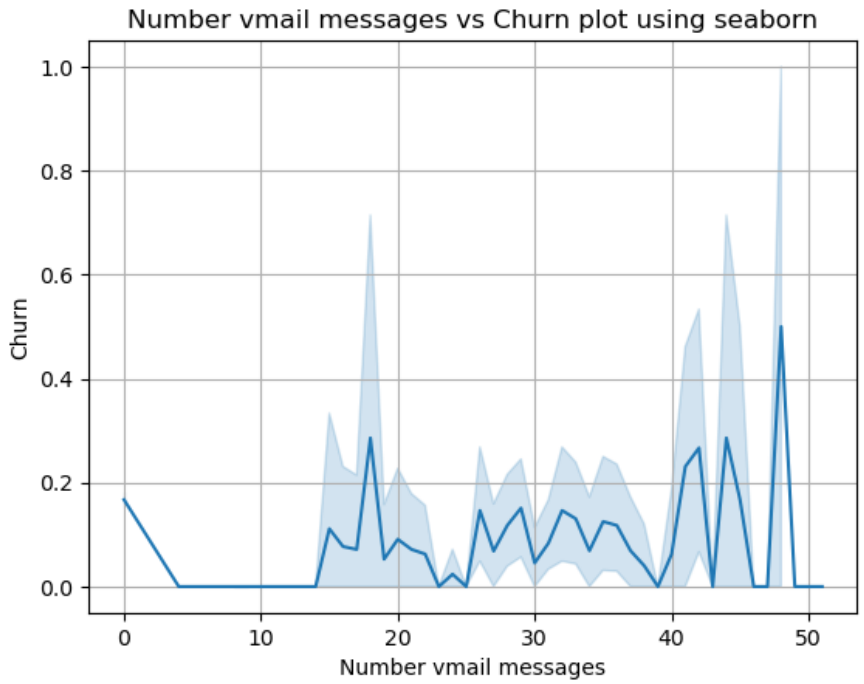
a. A support vector machine is another effective technique for detecting anomalies.
b. A SVM is typically associated with supervised learning, but there are extensions (OneClassCVM, for instance) that can be used to identify anomalies as an unsupervised problem (in which training data are not labeled).
c. The algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance, it tunes itself to identify the abnormalities that fall outside the learned region.
d. Depending on the use case, the output of an anomaly detector could be numeric scalar values for filtering on domain-specific thresholds or textual labels (such as binary/multi labels).

5.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the data before building predictive models. For the topic of predicting churn in telecom using multi-level classification, EDA involves gaining insights into the characteristics of the data related to customer behavior, usage patterns, and churn levels. By performing EDA, you'll gain a deeper understanding of your data, which will inform feature selection, model building, and interpretation of results in the subsequent stages of the churn prediction project.

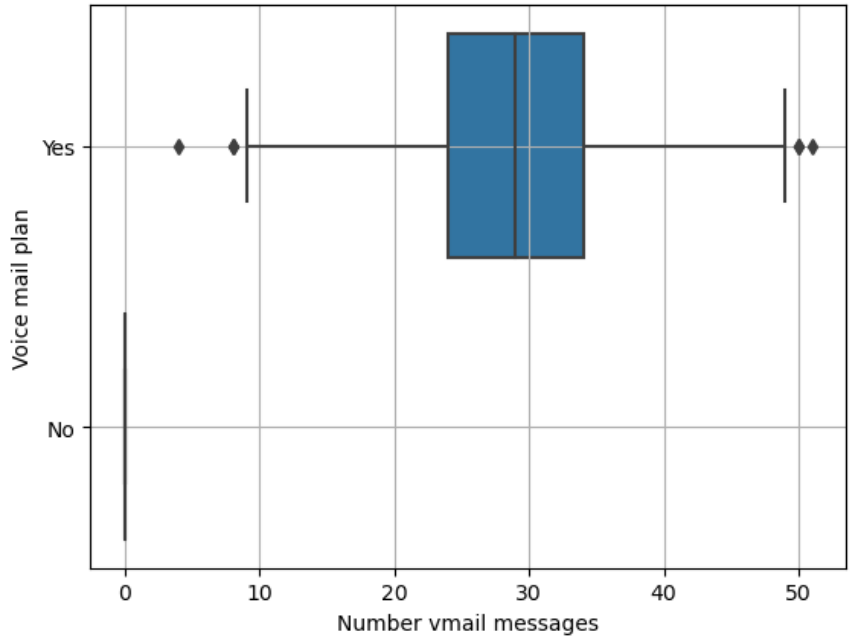
- **ADVANCE EDA**

  - **LINE PLOTS**



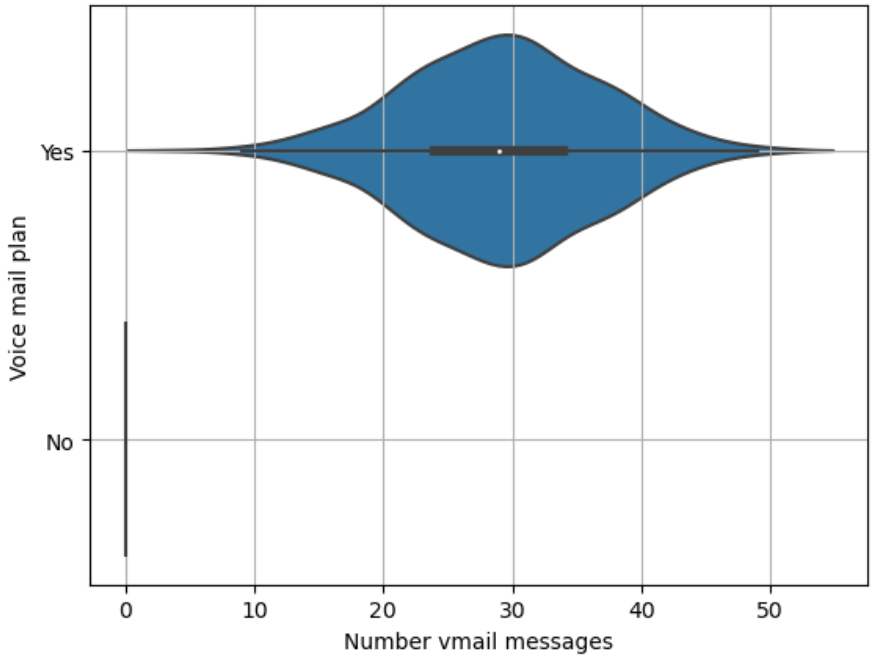
Graph (a) : Line plots

- **BOX PLOTS**



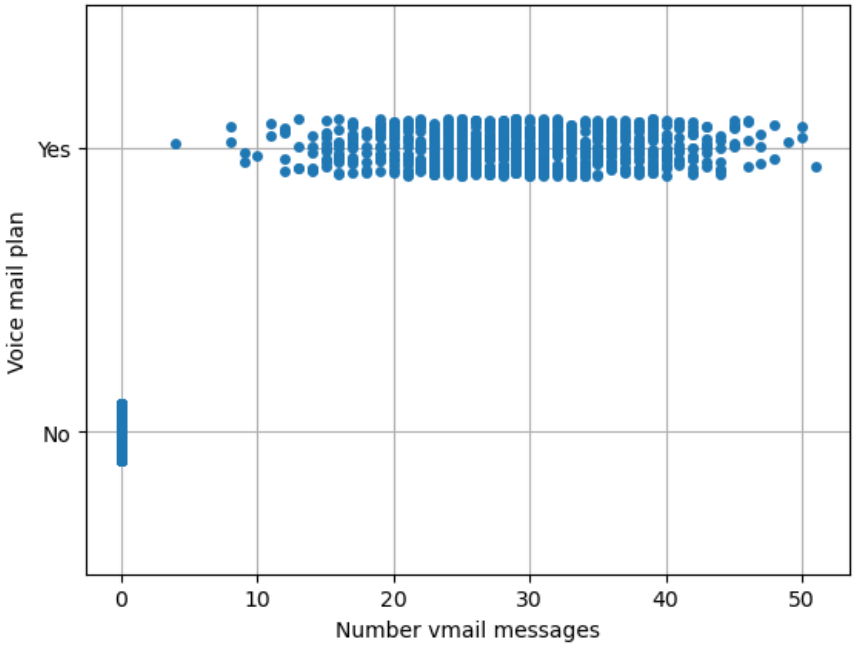
Graph (b) : Box plots

- **VIOLIN PLOTS**



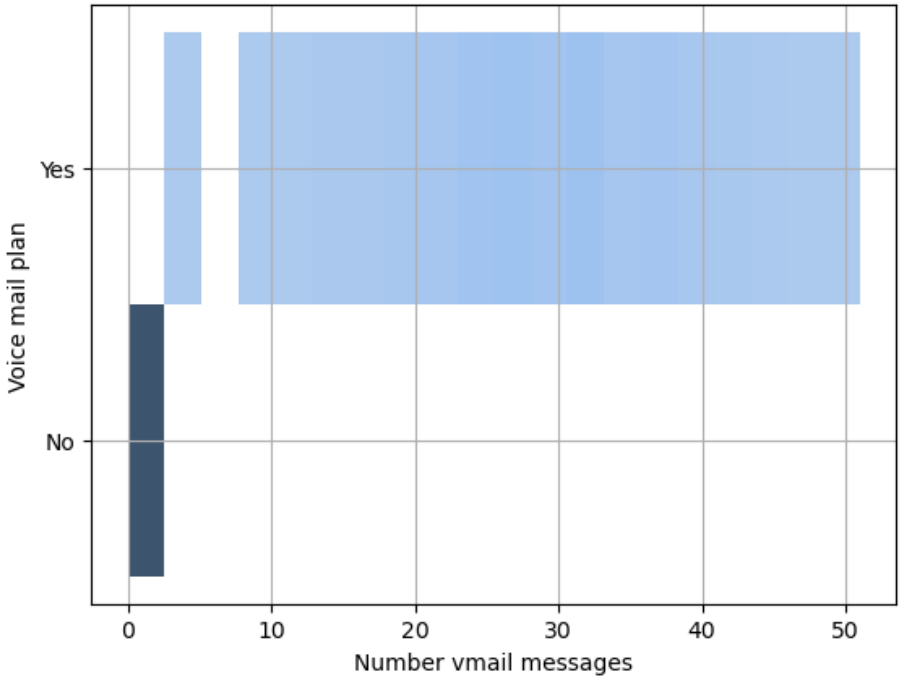
Graph (c) : Violin plots

- **STRIP PLOTS**



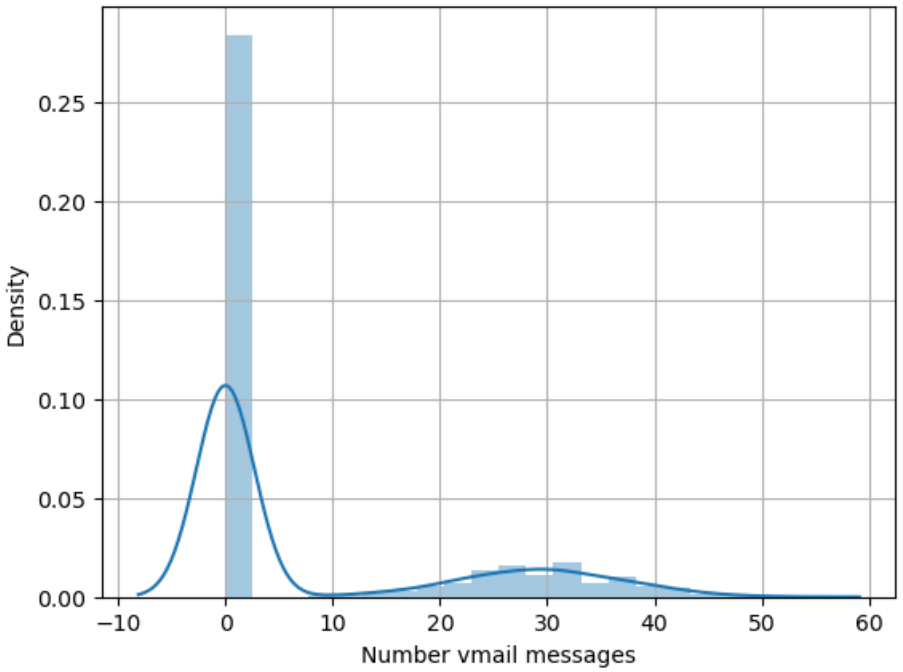
Graph (d) : Strip plots

- **HISTOGRAM PLOTS**



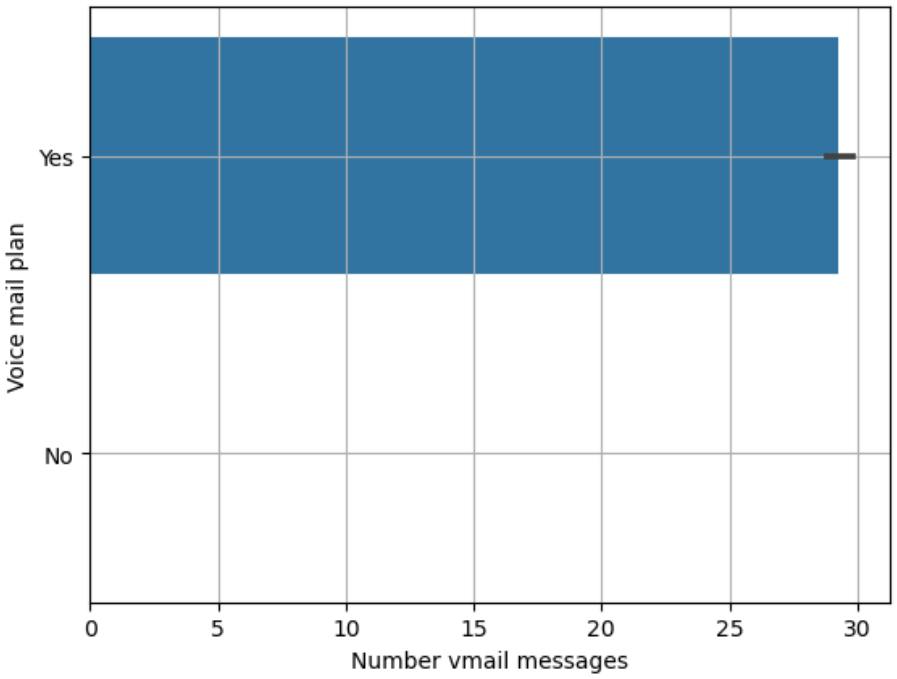
Graph (e) : Histogram plots

- **DISTRIBUTION PLOTS**



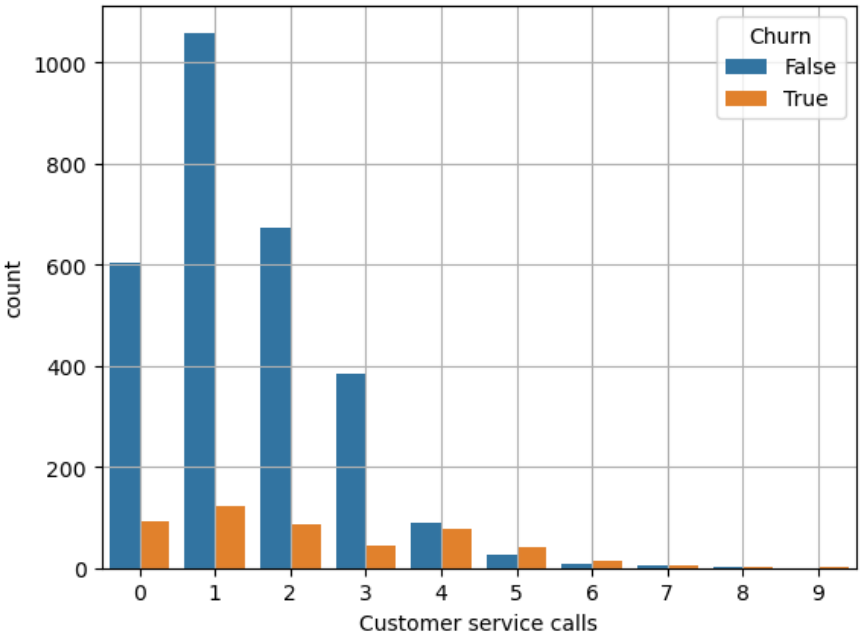
Graph (f) : Distribution plots

- **BAR PLOTS**



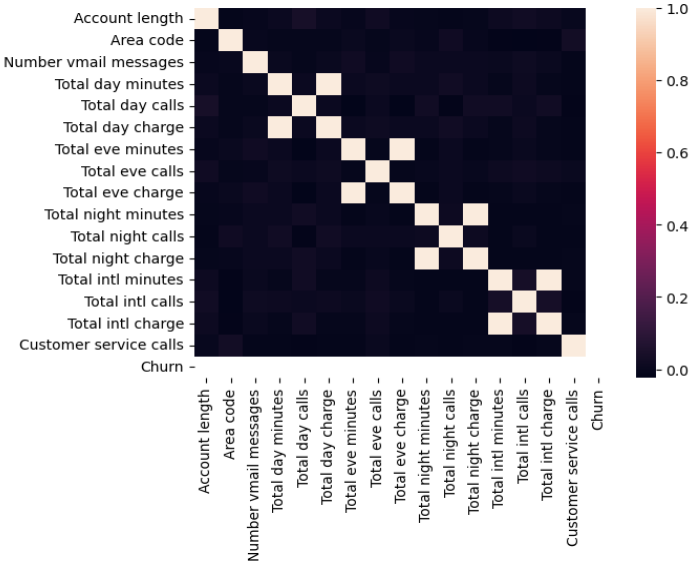
Graph (g) : Bar plots

- **COUNT PLOTS**



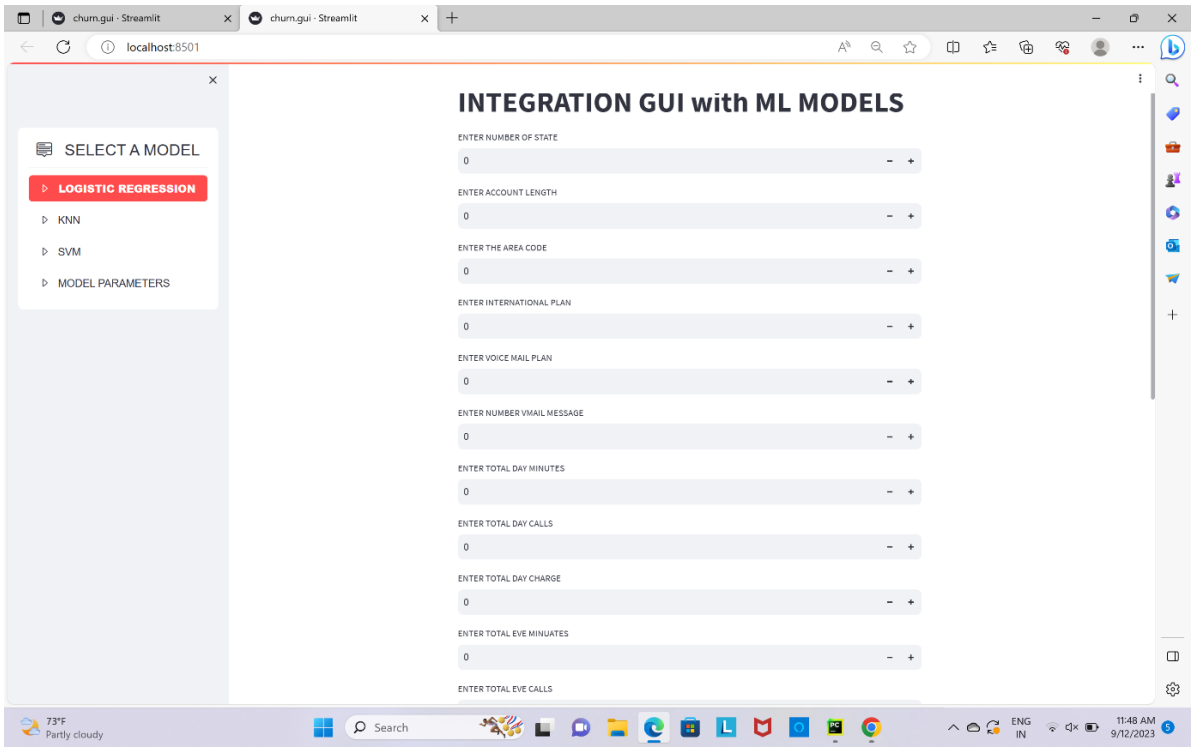
Graph (h) : Count plots

- **HEAT MAPS**



Graph (i) : Heat Map

5.4 Logistic regression



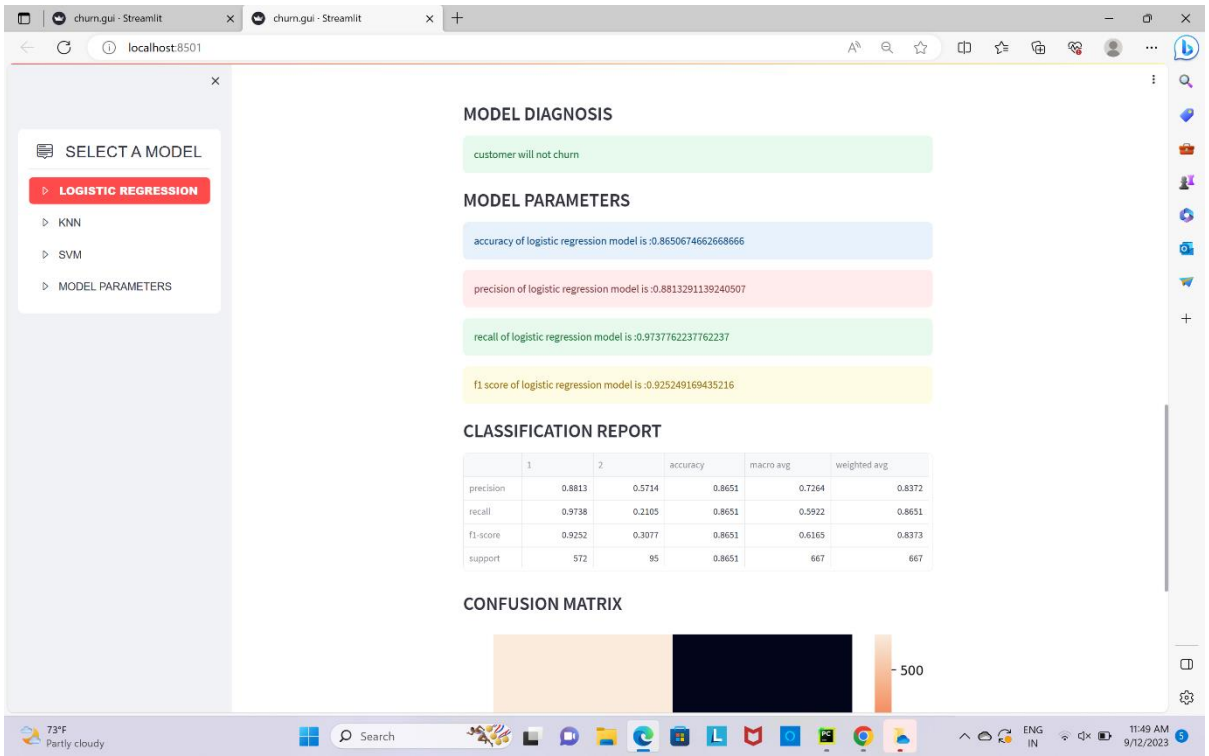


Fig : 5.4

5.5 KNN

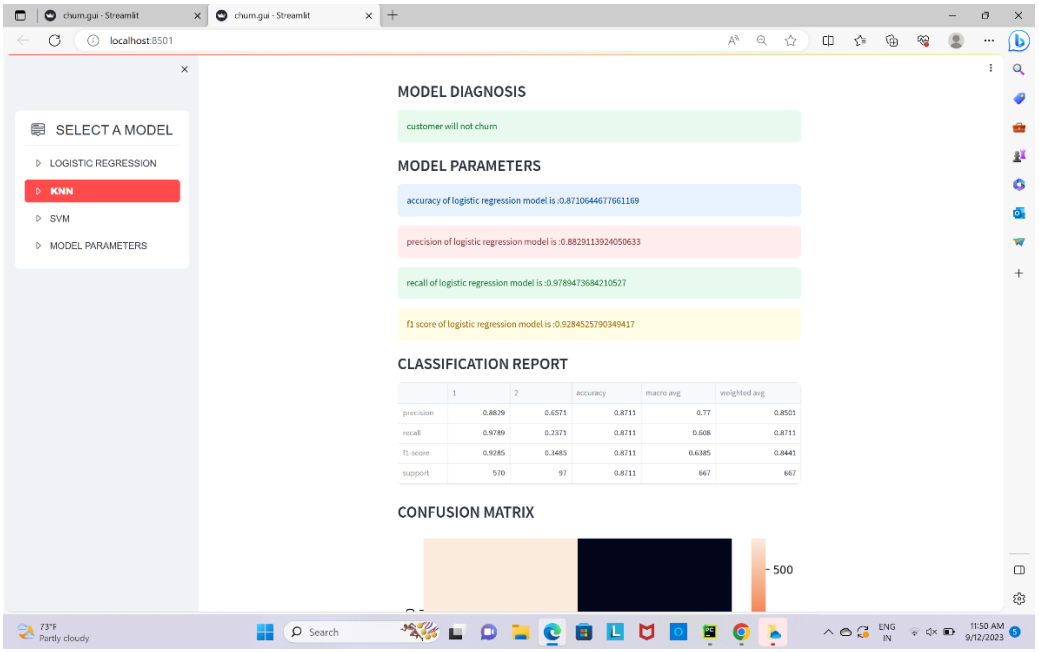


Fig : 5.5

5.6 SVM

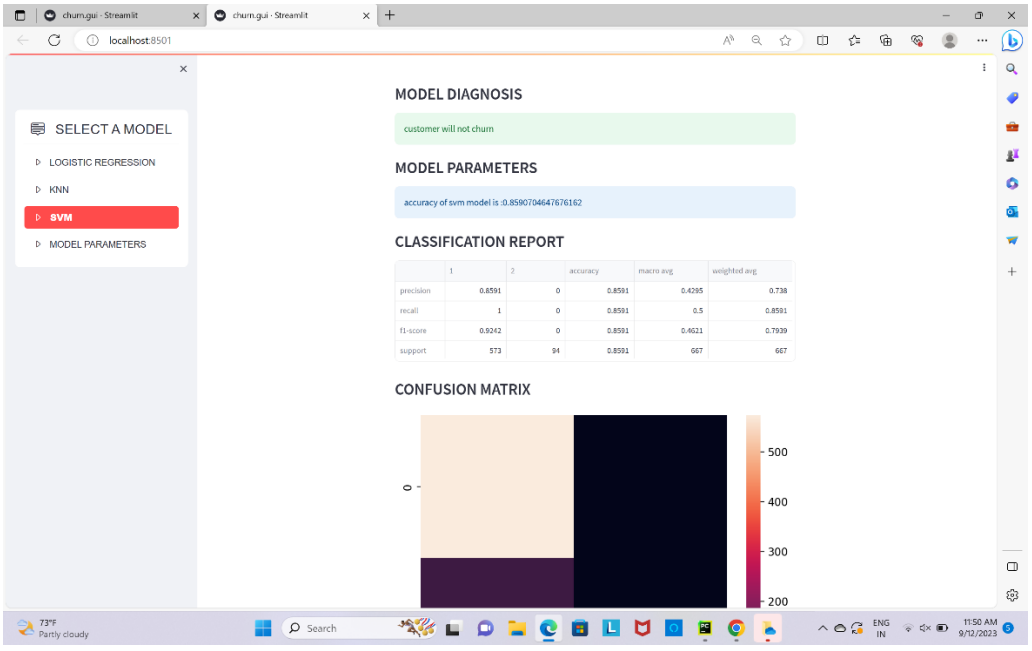


Fig : 5.6

5.7 MODEL PARAMETERS

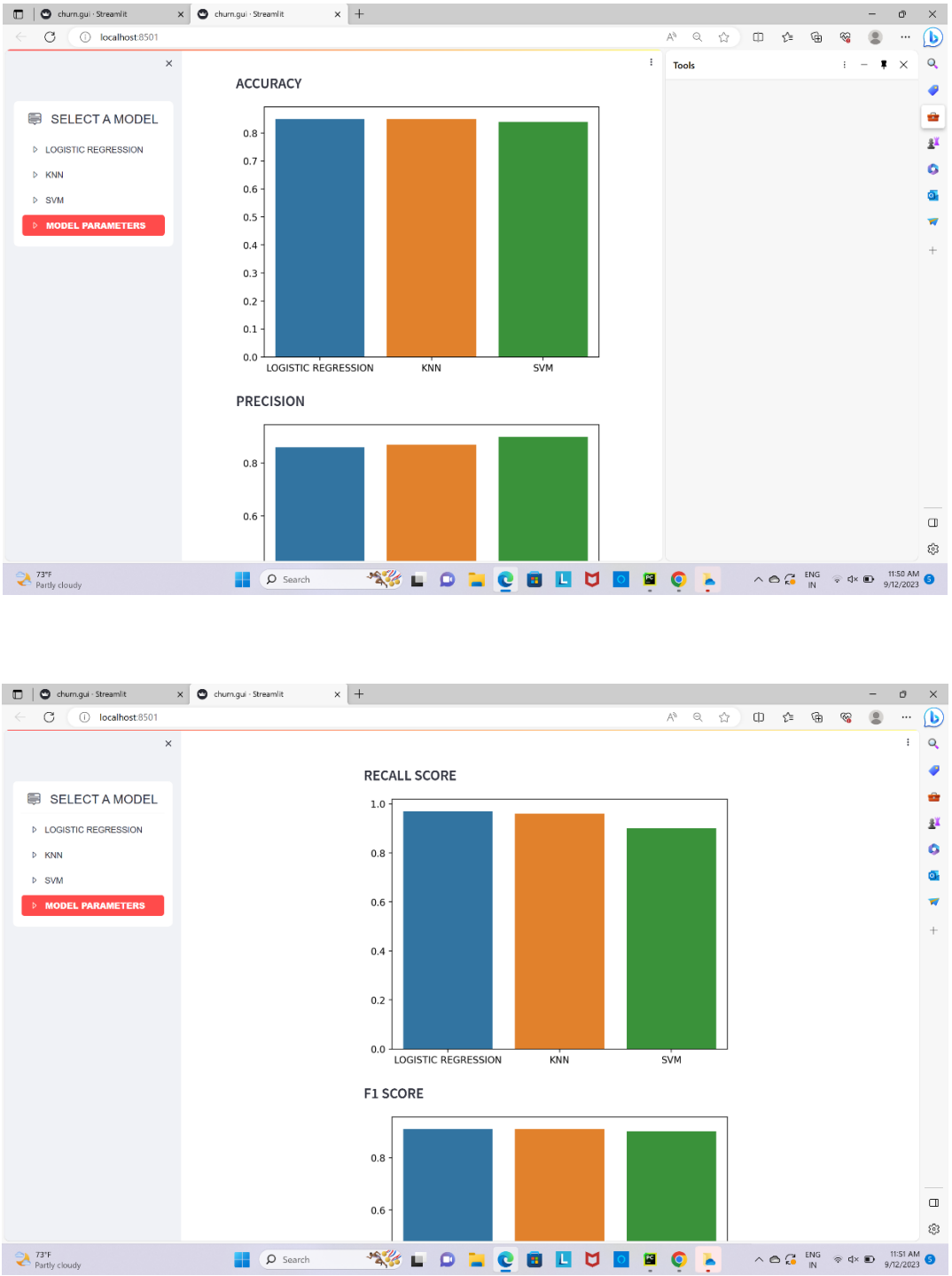


Fig 5.7

CHAPTER 6

ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- Granular insights.
- Targeted interventions.
- Improved accuracy.
- Better resource allocation.
- Enhanced customer segmentation.

DISADVANTAGES

- Complexity.
- Data imbalance.
- Integration complexity.

CHAPTER 7

CONCLUSION

In the competitive telecom sector standardization and public policies of mobile communication permits customers to switch over from one carrier to another carrier easily resulting in a competitive market. The prediction of churn or the task of recognizing customers who are probable to discontinue service use is a lucrative and essential issue of telecom sector. Customer churn is often a critical problem for the telecom sector as customers do not delay to leave if they do not predict what they are viewing for. Customers mainly need value for money, competitive cost and greater service quality. Customer churning is associated directly to satisfaction of customer. It is a known fact that the customer acquisition cost is larger than customer retention cost that makes the retention a difficult prototype of business.

REFERENCES

[1]. Tanneedi, N.N.P.P. (2016). Customer Churn Prediction Using Big Data Analytics. Master Thesis, Blekinge Institute of Technology.

[2]. Huang, B, Kechadi, M.T., Buckley, B. (2012). Customer churn prediction in telecommunications." Expert Systems with Applications, 39(1), pp. 1414-1425.

[3]. Amin, A, Anwar, S, Adnan, A, Nawaz, M, Alawfi, K, Hussain, A, Huang, K. (2016). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, volume 237, pp. 242-254.

[4]. Yabas, U, Chankya, H.C. (2013). Churn prediction in subscriber managementfor mobile and wireless communications services. IEEE Publications.