A Mini Project Report on

Customer Churn Analysis Using Data Mining Techniques

Submitted

by

Raj Nanasaheb Patil (47)
Niraj Vikram More (38)
Chaitali Pradip Patil (09)
Ishwari Shyam Yadav (67)

in partial fulfilment of the requirements for the award of the degree of

Bachelor in

COMPUTER ENGINEERING

For Academic Year 2024-2025

THE MET LEAGUE OF COLLEGES
MET Bhujbal Knowledge City
AS SHARP AS YOU CAN GET

Certificate

This is to Certify that

“Raj Nanasaheb Patil”

(Roll No: 47)

has completed the necessary Mini Project work and prepared the report on

“Customer Churn Analysis Using Data Mining Techniques”

in satisfactory manner as a fulfillment of the requirement of the award of degree of Bachelor of Computer Engineering in the Academic year

2024-2025

DEPARTMENT OF COMPUTER ENGINEERING
MET’s Institute of Engineering Bhujbal Knowledge City
Adgaon, Nashik – 422003.

Project Guide

Prof.V.P.Wani

Acknowledgement

I take this opportunity to thank all those who have contributed in successful completion of this Project work. I would like to express my sincere thanks to my ProjectGuided by **Prof. V .P. Wani**, who has encouraged me to work on this project and guided me whenever required.

We also would like to express our gratitude to our **H.O.D. Dr. P.M.Yawalkar** for giving us opportunities to undertake this project work at Met Institute of engineering, Nashik. We are extremely grateful to our **Principal Dr. V. P. Wani** for his constant inspiration and keen interest to make the project and presentation absolutely flawless.

At the last but not the least we would like to thank our Teaching staff member, Workshop staff member, Friends and family member for their timely co-operation andhelp.

By

Raj Nanasaheb Patil
Niraj Vikram More
Chaitali Pradip Patil
Ishwari Shyam Yadav

Course Objective:

1. To introduce the concepts and components of Business Intelligence (BI)
2. To evaluate the technologies that make up BI (data warehousing, OLAP)
3. To identify the technological architecture of BI systems·
4. To explain different data preprocessing techniques
5. To identify machine learning model as per business need
6. To understand the BI applications in marketing, logistics, finance and telecommunication sector

Course Outcome:

On completion of this course, the students will be able to
CO1: Differentiate the concepts of Decision Support System & Business Intelligence CO2:Use Data Warehouse & Business Architecture to design a BI system.
CO3:Build graphical reports
CO4:Apply different data preprocessing techniques on dataset
CO5:mplement machine learning algorithms as per business needs
CO6:Identify role of BI in marketing, logistics, and finance and telecommunication sector

Abstract

In today's competitive business environment, customer retention is a critical challenge, especially in the telecom industry where customers frequently switch between service providers. The cost of acquiring a new customer is significantly higher than retaining an existing one. Therefore, predicting customer churn—the process by which a customer discontinues their subscription or service is a key area of focus for business intelligence and data-driven decision-making. This mini project aims to analyze and predict customer churn using data mining techniques. We utilize the publicly available **Telco Customer Churn dataset** from Kaggle, which contains information about customer demographics, service usage, account details, and contract information. The primary goal is to build a predictive model that classifies whether a customer is likely to churn or not.

The project begins with a detailed problem definition and data exploration phase. Basic operations like shape analysis, missing value treatment, and summary statistics help understand the structure and quality of the data. Data preprocessing steps such as encoding categorical variables, handling null values, and feature scaling are performed to prepare the dataset for machine learning. A **Random Forest Classifier**, known for its high accuracy and ability to handle both categorical and numerical data, is used to train the model. The dataset is split into training and testing subsets to evaluate model performance. The model's effectiveness is assessed using metrics such as **accuracy**, **precision**, **recall**, and **F1-score**. A confusion matrix is also generated to visualize the performance.The final model reveals that certain features like contract type, tenure, and monthly charges are strong predictors of customer churn. A feature importance plot is generated to highlight these key attributes. These insights can be used by the telecom company to create targeted retention strategies.

In conclusion, the project demonstrates the power of data mining in solving real-world business problems. By predicting customer churn, companies can make informed decisions to improve customer satisfaction and reduce losses, ultimately driving business growth.

Contents

Introduction

In the highly competitive telecommunications industry, customer satisfaction and retention play a vital role in maintaining long-term profitability. One of the major challenges faced by telecom companies is customer churn, which refers to the loss of clients or subscribers who stop using the services of a company. High churn rates not only result in revenue loss but also increase the cost of acquiring new customers. As such, understanding the reasons behind churn and being able to predict customer behavior has become a priority for telecom service providers. With the rapid growth of data availability and business intelligence (BI) tools, companies can now make data-driven decisions to enhance customer experience and improve retention strategies. Data mining techniques are particularly useful in analyzing historical customer data and identifying patterns that lead to churn. By implementing predictive models, companies can proactively address customer dissatisfaction and offer personalized solutions to retain valuable customers.

This mini project focuses on developing a Customer Churn Prediction Model using the Telco Customer Churn dataset. The dataset contains various features such as customer demographics, account information, and service usage patterns. The primary objective is to build a classification model using machine learning algorithms that can accurately predict whether a customer is likely to churn or not.The process includes data collection, preprocessing, exploratory data analysis, model training, evaluation, and visualization of results. A Random Forest Classifier is used for prediction due to its effectiveness in handling complex datasets. The findings from the model provide meaningful insights into the key factors influencing churn, which can help telecom companies make better business decisions.

Through this project, we aim to demonstrate how business intelligence and machine learning can be applied to solve practical problems and support strategic decision-making in real-world scenarios

Algorithm Used: Random Forest Classifier:

To predict churn, we use the **Random Forest Classifier**, a supervised machine learning algorithm widely recognized for its accuracy, efficiency, and interpretability. It works by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.

- **Why Random Forest?**

The Random Forest algorithm was selected for this project due to its robustness, flexibility, and high accuracy when dealing with classification problems like churn prediction. Customer churn datasets typically contain a mix of numerical and categorical variables—for example, features like tenure, MonthlyCharges, and TotalCharges are numerical, while Contract, PaymentMethod, and InternetService are categorical. Random Forest is well-suited for handling such diverse data types without requiring heavy preprocessing or transformation. Unlike other algorithms such as Logistic Regression or Support Vector Machines (SVM), which require that all input data be numeric and scaled, Random Forest can directly interpret categorical values (after simple encoding) and is not sensitive to the scale of features.Another major advantage of using Random Forest is its ability to reduce overfitting, a common issue in machine learning models, especially with decision trees. By constructing multiple trees on random subsets of the dataset and aggregating their predictions, Random Forest offers a more generalized and stable performance. This makes it particularly reliable when dealing with real-world business problems where model consistency and accuracy are crucial.Additionally, Random Forest provides insights into the relative importance of each feature in making predictions. This is especially valuable in a Business Intelligence context because it allows decision-makers to understand which factors most significantly influence customer churn. For instance, it may reveal that contract type, monthly charges, and customer tenure are key indicators of whether a customer will leave. Such insights can directly inform business strategies—for example, offering long-term contracts or discounts to high-risk customers to improve retention rates.Finally,

Random Forest is highly scalable and can handle large volumes of data efficiently, which is ideal for enterprise-scale datasets commonly encountered in telecom industries. It also supports parallel computation, which reduces training time. Due to these strengths—accuracy, interpretability, resistance to overfitting, and practical utility—the Random Forest algorithm stands out as a powerful and appropriate choice for predictive churn modeling in this BI mini project.

- Handles both **numerical and categorical** features effectively.
- Works well even with **missing or unbalanced data**.
- Reduces overfitting by averaging multiple decision trees.
- Provides **feature importance** rankings, helping in business decision-making.

Steps in Algorithm Application:

1. Data Preprocessing

The dataset is first cleaned and prepared. Missing values in columns like Total Charges are handled by converting them to numeric and filling them with the median. Categorical variables such as Contract, Internet Service, etc., are encoded using Label Encoder so the model can process them. Although Random Forest doesn't require feature scaling, we apply StandardScaler to maintain consistency and improve comparability with other models.

2. Splitting the Data

The pre-processed data is split into training and testing sets in an 80:20 ratio. The training set is used to build the model, while the test set is used to evaluate its performance on unseen data.

3. Model Training

We use a Random Forest Classifier, which creates multiple decision trees and combines their results to reduce overfitting and improve accuracy. The model learns patterns from the training data to predict whether a customer is likely to churn.

4. Prediction and Evaluation

The trained model predicts churn on the test set. We evaluate its performance using accuracy, confusion matrix, precision, recall, and F1-score—these metrics help us understand how well the model distinguishes between churned and non-churned customers.

5. Feature Importance Analysis

Random Forest provides feature importance scores that show which features most influence the prediction. This helps identify key factors like Contract, tenure, and Monthly Charges that drive churn, offering valuable businessinsights.

Objective

The main goal is to:

- Understand the reasons behind customer churn
- Use predictive analytics to classify customers as likely to churn or not
- Provide actionable insights to help the company reduce churn and retain more customers

Requirement Analysis:

Hardware Requirement:

I. Operating System – Windows 11

II. Ram - 8GB

III. Hard Disk – 1TB

IV. Processor – Intel core i3

Software Requirement:

V. Google Colab

Algorithm: Customer Churn Prediction using Random Forest

Input:
Telco Customer Churn dataset (CSV format)
Output:
Predicted class labels (Churn / Not Churn)
Evaluation metrics (Accuracy, Precision, Recall, F1-score)
Feature importance scores

Step 1: Import Required Libraries
Import Python libraries such as pandas, numpy, matplotlib, seaborn, sklearn, etc., needed for data analysis, visualization, and machine learning.

Step 2: Load the Dataset
Read the dataset into a DataFrame using pandas.read_csv().

Step 3: Data Preprocessing
Convert the TotalCharges column to numeric type and handle missing values using median imputation.
Drop columns that are not useful for prediction like customerID.
Encode all categorical variables using LabelEncoder.
Split the dataset into input features (X) and target (y = Churn).
Normalize the features using StandardScaler for consistent scale.

Step 4: Split the Dataset
Split the preprocessed data into training (80%) and testing (20%) sets using train_test_split().

Step 5: Train the Random Forest Classifier
Create a RandomForestClassifier model.
Fit the model using the training dataset: model.fit(X_train, y_train).

Step 6: Make Predictions
Use the trained model to predict churn on the test set: y_pred = model.predict(X_test).

Step 7: Evaluate the Model
Calculate evaluation metrics such as:
Accuracy using accuracy_score()
Precision, Recall, F1-score using classification_report()
Confusion Matrix using confusion_matrix()

Step 8: Analyze Feature Importance
Retrieve and visualize the most important features contributing to churn using model.feature_importances_.

Step 9: Visualize Results
Plot the confusion matrix and feature importance using seaborn and matplotlib.

System overview

Random Forest Working Process:

Think of it as an ensemble-based algorithm that builds multiple decision trees and combines their outputs to improve accuracy and stability. The process starts by taking the dataset and creating several random subsets of the data—this is called bootstrapping. For each subset, a decision tree is trained independently using a random selection of features at each split. This randomness ensures that the trees are diverse and not all biased in the same way.

The individual decision trees work like experts, each giving their own prediction (churn or not churn). When a prediction is needed, the Random Forest collects the outputs of all the trees and uses a majority voting mechanism to decide the final result. If most trees predict “Yes” for churn, the forest also outputs “Yes,” and vice versa. This model is both powerful and robust because it reduces overfitting (common in single decision trees) and adapts well to noisy or incomplete data. The final model is accurate, interpretable through feature importance, and reliable for use in real-world customer churn scenarios.

CODE:

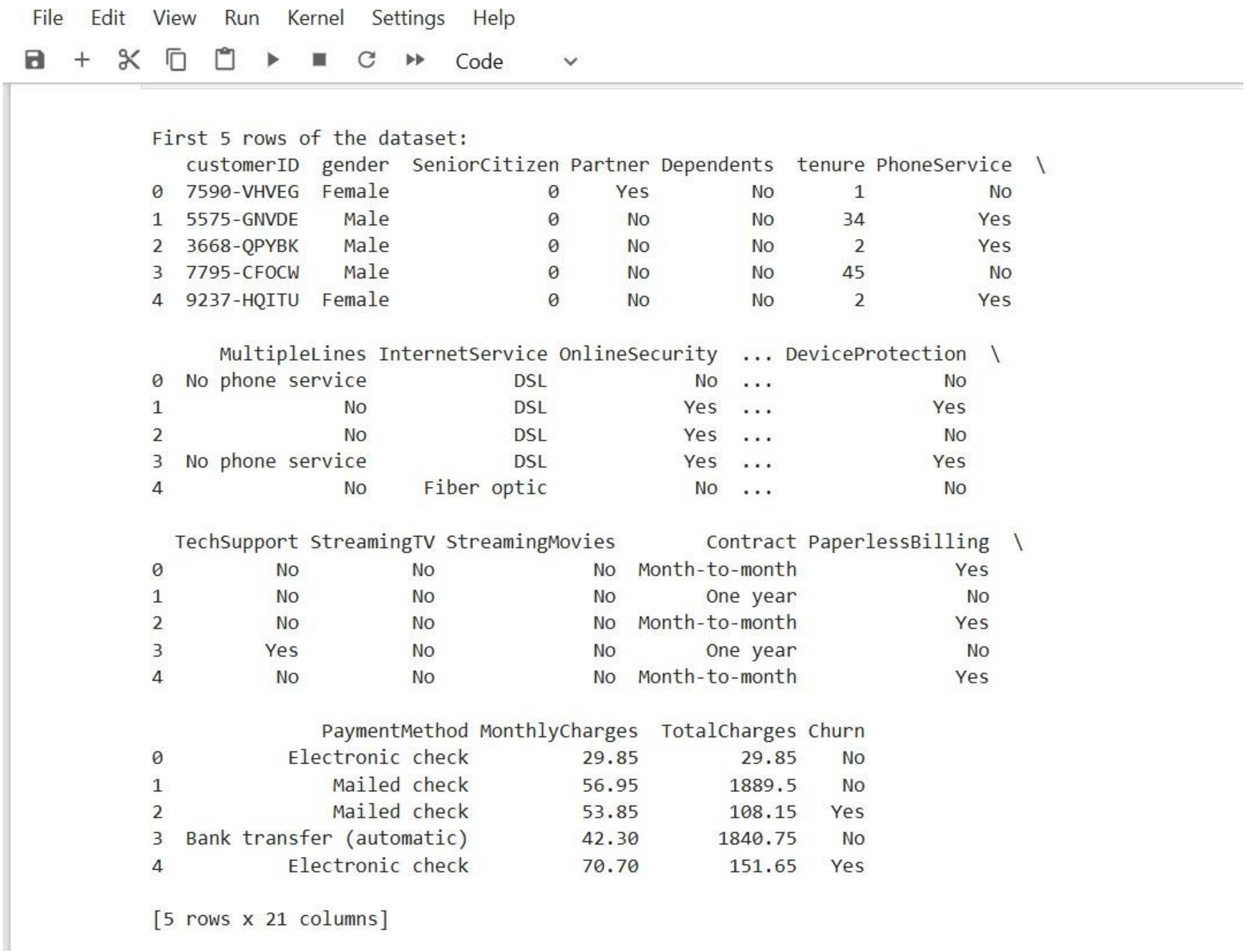
CUSTOMER CHURN ANALYSIS

```
# Step 1: Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Step 2: Load Dataset
df = pd.read_csv("Telco-Customer-Churn.csv")  # Make sure this file is in the same folder

# Basic Operations
print("\nFirst 5 rows of the dataset:")
print(df.head())

print("\nDataset shape:", df.shape)

print("\nColumn names:")
print(df.columns)

print("\nData types:")
print(df.dtypes)

print("\nChecking for missing values:")
print(df.isnull().sum())

print("\nSummary statistics:")
print(df.describe())

# Step 3: Data Preprocessing
# Remove customerID
df.drop('customerID', axis=1, inplace=True)

# Convert TotalCharges to numeric (some values might be blank)
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
df['TotalCharges'].fillna(df['TotalCharges'].median(), inplace=True)

# Encode categorical variables
le = LabelEncoder()
for col in df.columns:
    if df[col].dtype == 'object':
        df[col] = le.fit_transform(df[col])
```

```
# Step 4: Train-Test Split
X = df.drop('Churn', axis=1)
y = df['Churn']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 5: Feature Scaling
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Step 6: Model Training
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Step 7: Model Evaluation
y_pred = model.predict(X_test)
print("\nAccuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))

# Step 8: Feature Importance
importances = model.feature_importances_
feature_names = X.columns
feature_importance_df = pd.DataFrame({'Feature': feature_names, 'Importance': importances})
feature_importance_df.sort_values(by='Importance', ascending=False, inplace=True)

plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature', data=feature_importance_df)
plt.title('Feature Importance in Customer Churn Prediction')
plt.tight_layout()
plt.show()
```

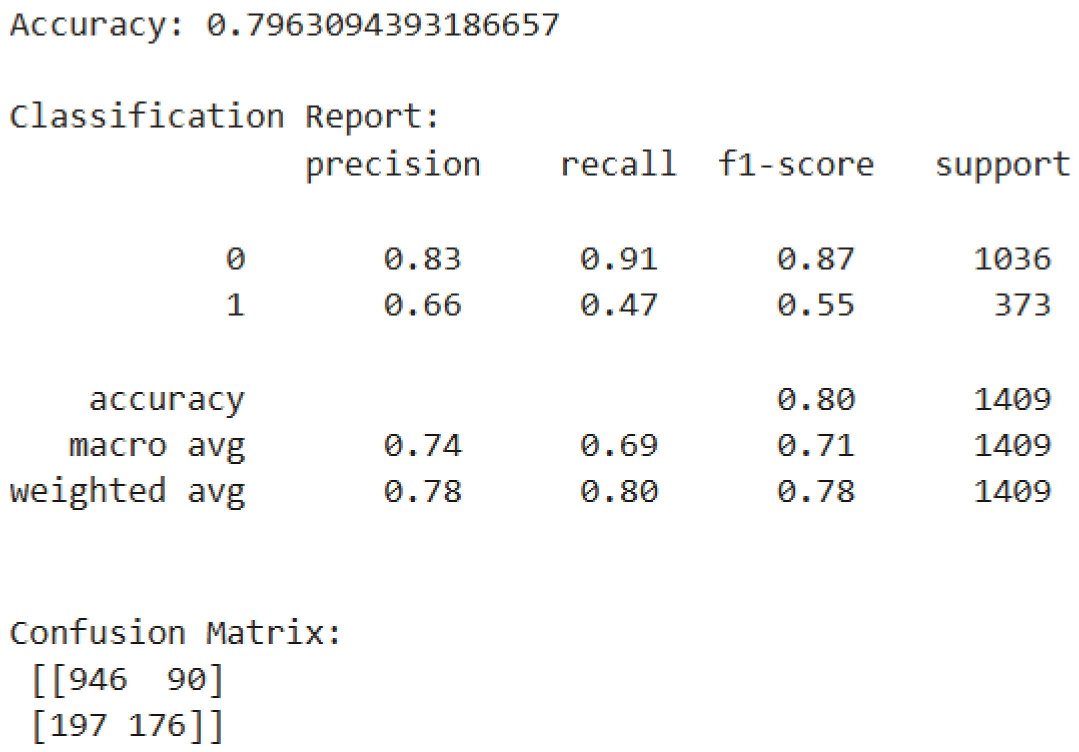
OUTPUT:

1) **Basic Operations**



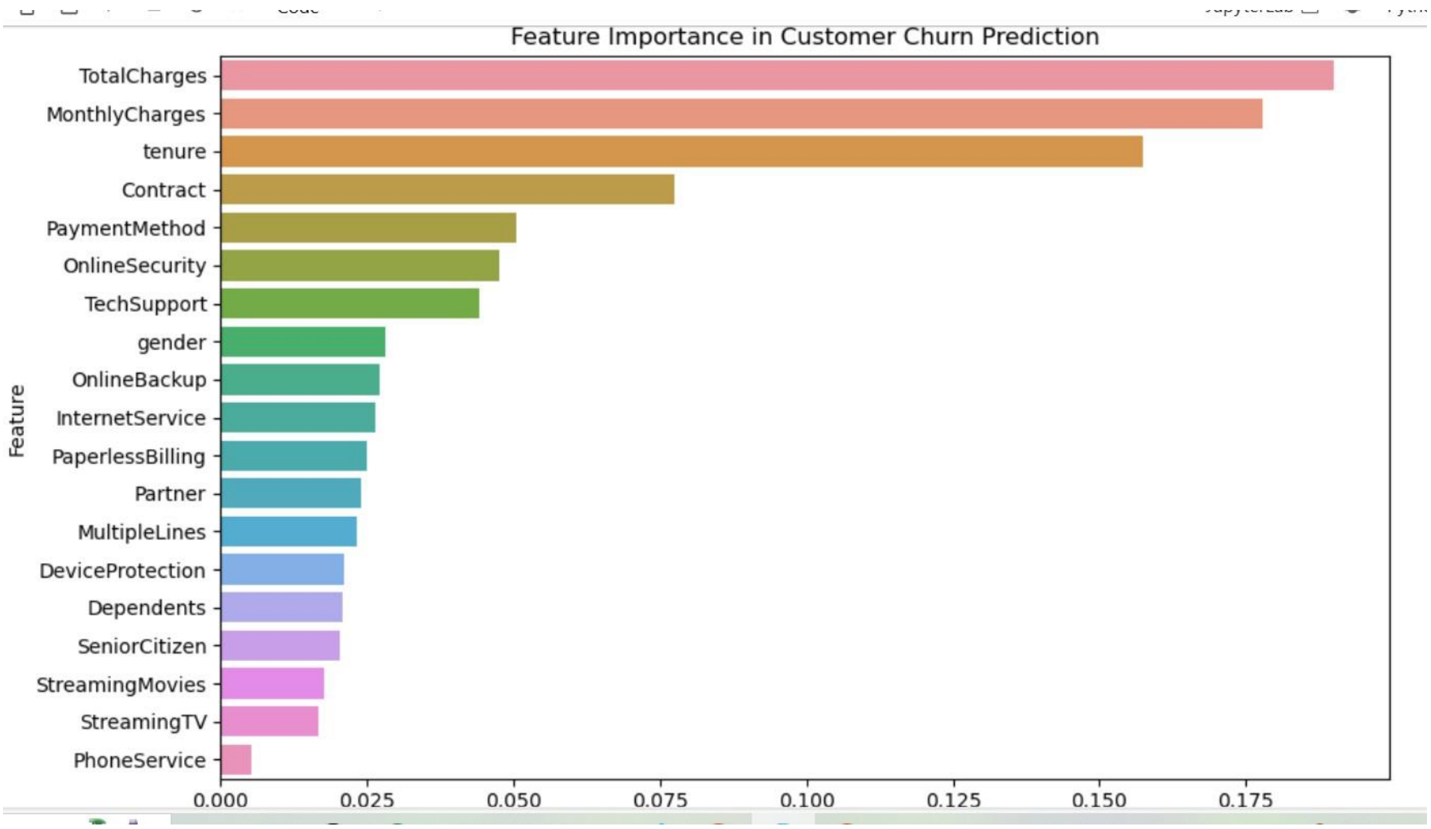
```
First 5 rows of the dataset:
   customerID  gender  SeniorCitizen Partner Dependents  tenure PhoneService  \
0  7590-VHVEG  Female              0     Yes         No       1           No   
1  5575-GNVDE    Male              0      No         No      34          Yes   
2  3668-QPYBK    Male              0      No         No       2          Yes   
3  7795-CFOCW    Male              0      No         No      45           No   
4  9237-HQITU  Female              0      No         No       2          Yes   

      MultipleLines InternetService OnlineSecurity  ... DeviceProtection  \
0  No phone service             DSL             No  ...               No   
1                No             DSL            Yes  ...              Yes   
2                No             DSL            Yes  ...               No   
3  No phone service             DSL            Yes  ...              Yes   
4                No     Fiber optic             No  ...               No   

  TechSupport StreamingTV StreamingMovies        Contract PaperlessBilling  \
0          No          No              No  Month-to-month              Yes   
1          No          No              No        One year               No   
2          No          No              No  Month-to-month              Yes   
3         Yes          No              No        One year               No   
4          No          No              No  Month-to-month              Yes   

               PaymentMethod MonthlyCharges  TotalCharges Churn  
0           Electronic check          29.85         29.85    No  
1               Mailed check          56.95        1889.5    No  
2               Mailed check          53.85        108.15   Yes  
3  Bank transfer (automatic)          42.30       1840.75    No  
4           Electronic check          70.70        151.65   Yes  

[5 rows x 21 columns]
```

2) Model Training



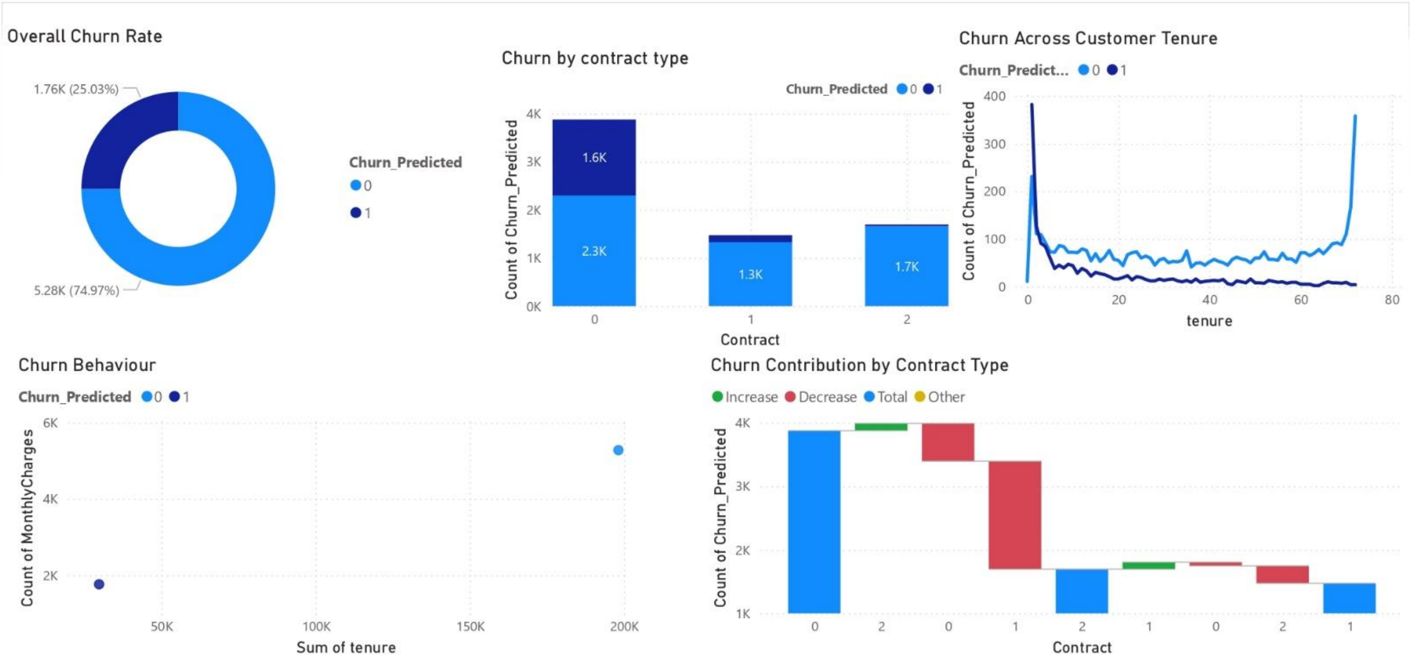
```
Accuracy: 0.7963094393186657

Classification Report:
               precision    recall  f1-score   support

           0       0.83      0.91      0.87      1036
           1       0.66      0.47      0.55       373

    accuracy                           0.80      1409
   macro avg       0.74      0.69      0.71      1409
weighted avg       0.78      0.80      0.78      1409


Confusion Matrix:
 [[946  90]
 [197 176]]
```

3) Feature Importance



4. Visualization :-

1. Pie Chart
2. Stacked Column Chart
3. Line Chart
4. Scatter Plot
5. Waterfall Plot



Conclusion

Customer churn is a critical challenge faced by telecom industries, and identifying potential churners in advance helps businesses retain their valuable customers. In this project, we implemented and analyzed a data mining approach using the Random Forest classification algorithm to predict customer churn. With a computational complexity that balances accuracy and efficiency, Random Forest handles large datasets and diverse feature types effectively.Through structured preprocessing, model training, and evaluation, the system extracts hidden patterns in customer behavior. The algorithm’s ensemble nature eliminates overfitting often found in single decision trees and provides reliable results backed by feature importance insights. The implemented model was evaluated using multiple metrics, ensuring the system is robust and applicable in real-world business scenarios.Thus, we have implemented and analyzed the performance of the Random Forest classifier in predicting customer churn and highlighted its effectiveness as a data mining approach in a Business Intelligence context.

References

- **Telco Customer Churn Dataset (Kaggle)**
  https://www.kaggle.com/datasets/blastchar/telco-customer-churn

- **Random Forest Classifier - Scikit-Learn Documentation**
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- **Understanding Classification Metrics**
  https://scikit-learn.org/stable/modules/model_evaluation.html

- **SHAP (SHapley Additive Explanations)**
  https://shap.readthedocs.io/en/latest/

- **LabelEncoder - Scikit-Learn**
  https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

- **StandardScaler - Scikit-Learn**
  https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

- **Business Intelligence Concepts**
  https://www.investopedia.com/terms/b/business-intelligence-bi.asp

- **Ensemble Learning Techniques**
  https://towardsdatascience.com/ensemble-learning-bagging-boosting-and-stacking-c9214a10a205