Customer Churn Prediction

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Jemima I V

Meenu P V

Meenakshy S S

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Nov 2022

List of Figures

List of Abbreviations

Abbreviations	Definitions
EDA	Exploratory Data Analysis
CSV	Comma-Separated Values
PD	Pandas library in Python
NP	Pandas library in Python
XGBoost	eXtreme Gradient Boosting
SKLearn	scikit-learn

Table of Contents

Abstract

Customer churn is the number of customers who stop using a company's products or services over a given period of time. Customers in the telecom industry can choose from a variety of service providers and actively switch from one to the next. Customers are the most important resources for any companies or businesses. What if these customers leave the company due to high charges, better competitor offers, poor customer services or something unknown, which indirectly shows company's performance. Hence, Customer churn is one of the important metrics for companies to evaluate their performance. This data science project aims to predict customer churn by analyzing behavioral, demographic, and transactional data using machine learning techniques.

The project involves collecting and preprocessing data, conducting exploratory data analysis (EDA), and engineering features that capture key customer behaviors such as usage patterns, billing trends, and customer support interactions. A range of classification algorithms, including Logistic Regression, Random Forest, and XGBoost, are evaluated for their predictive performance using metrics and performing deployment using Streamlit.

The project delivers actionable insights that help telecom companies reduce churn rates, enhance customer satisfaction, and increase profitability. By integrating predictive analytics into business operations, this solution supports data-driven decision-making and sustainable growth.

.

1. Problem Definition

1.1 Overview

1.1.1 Project Objective

The goal of this project is to develop a predictive model that identifies customers likely to churn based on their demographic, usage, and account data. By accurately predicting churn, telecom companies can design targeted retention strategies to minimize customer attrition and improve their services.

1.1.2 Business Need

In the highly competitive telecom industry, customer churn poses a significant challenge. Telecom companies face the dual pressure of retaining existing customers while acquiring new ones. Customer churn not only leads to revenue loss but also increases operational costs due to the high expense of customer acquisition.

1.1.4 Timeline

The timeline given for us is one month (30/12/2024 -17/12/2024)

- (0-1 weeks): Data collection, cleaning, and exploratory data analysis (EDA).
- (2nd week): Model building, feature engineering, and model selection.
- (3rd week): Hyper parameter tuning, model validation, testing, deployment, and report generation, presentation

1.1.5 Resources

- Jupyter Notebook, VS Code
- Python
- Version control systems (Git/GitHub)

1.2 Problem Statement

This project aims to develop a predictive model that accurately identifies customers likely to churn based on their historical data, including demographic, service usage, account, and billing details. By analysing these patterns, the model will provide actionable insights that help telecom companies reduce churn rates, enhance customer retention strategies, and improve overall profitability.

The problem can be broken down into the following key objectives:

1. **Understanding Churn Drivers:** Analyse customer data to identify factors contributing to churn, such as contract type, monthly charges, and service satisfaction.
2. **Predicting Churn Risk:** Build a machine learning model to classify customers as likely to churn or not, with high accuracy and reliability.
3. **Strategic Insights:** Provide recommendations for targeted retention strategies, such as offering discounts, improving customer service, or promoting long-term contracts.

Addressing this problem will empower telecom companies to make data-driven decisions, enhance customer satisfaction, and maintain a competitive edge in the market.

.

2.Introduction

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenues to acquire new customers, upsell the existing customers, and increase the retention period of customers. However, comparing these strategies taking the value of return on investment (ROI) of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer costs much lower than acquiring a new one, in addition to being considered much easier than the upselling strategy. To apply the third strategy, companies have to decrease the potential of customer's churn,knownas"thecustomermovementfromoneprovider to another". We focused on evaluating and analyzing the performance of a set of tree-based machine learning methods and algorithms for predicting churn in telecommunications companies.

We have experimented a number of algorithms suchasDecisionTree, RandomForest,GradientBoostMachineTreeandtobuild the predictive model of customer Churn after developing our data preparation, feature engineering, and feature selection methods. Customers' churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase. Many research confirmed that machine learning technology is highly efficienttopredictthis situation.Thistechniqueisappliedthroughlearningfrom previous data.

3. Literature Survey

Churn:Achurnisdefinedascustomerattritionorlossinatelecomindustrywhencustomers terminate their contracts or their usage and switch to another service provider.There aretypeof customers that share important service features like higher bills, long distance transitions, etc. They spend more than the average rate and their expectations from the service providers is high, which make the retention of such customers more important than the rest as they are beneficial to the company. The company must allocate more resources to them to decrease the churn rate.

M.A.H. Farquad [4] proposed a hybrid approach to overcome the drawbacks of generalSVMmodelwhichgeneratesablackboxmodel(i.e.,itdoesnotrevealthe knowledge gained during training in human understandable form). The hybrid approach contains three phases: In the

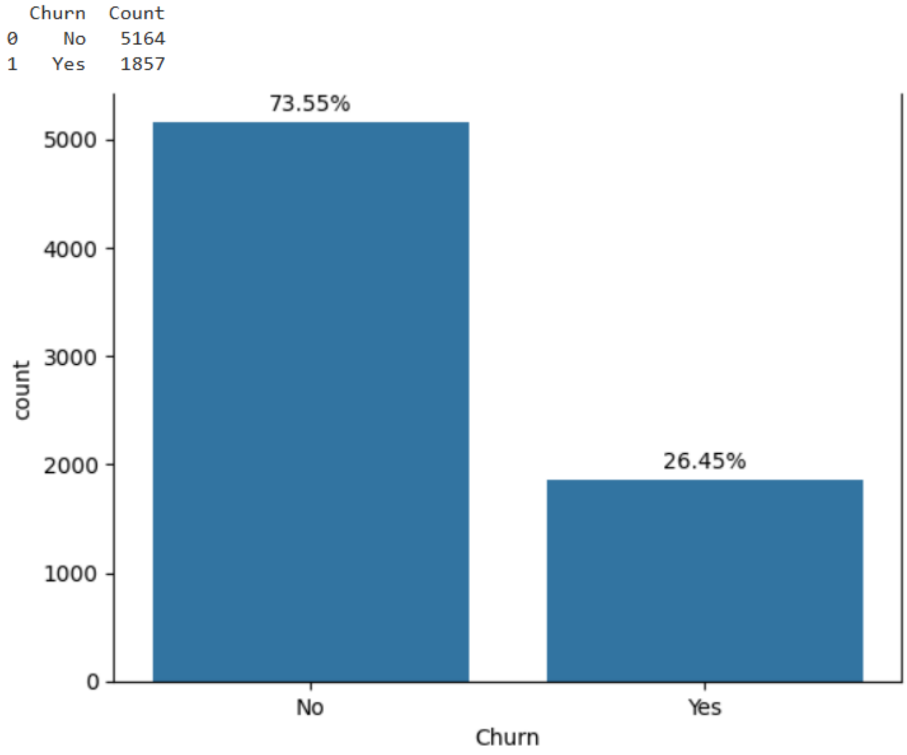
first phase, SVM-RFE (SVM-recursive feature elimination) is employed to reduce thefeature set. Inthesecond phase, dataset with reduced features is then used to obtain SVM model and support vectors are extracted. In the final phase, rules are then generated using Naive Bayes Tree (NBTree which is combination of Decision tree with naive Bayesian Classifier). The dataset used here is bank credit card customer dataset (Business Intelligence Cup 2004) which is highly unbalanced with 93.24% loyal and

6.76% churnedcustomers. The experimental showedthatthemodel does not scalable to large datsets.datsets.

Ning Lu [7] proposed the use of boosting algorithms to enhance a customer churn prediction model in which customers are separated into two clusters based on theweightassignedbytheboostingalgorithm. Asaresult,ahighriskycustomerclusterhasbeenfound.Logistic regressionisusedasabasislearner,andachurn predictionmodel is built oneachcluster, respectively. The experimental results showed that boosting algorithm provides a good separation of churn data when compared with a single logistic regression model.

4.Exploratory Data Analysis

- **Distribution of target feature**

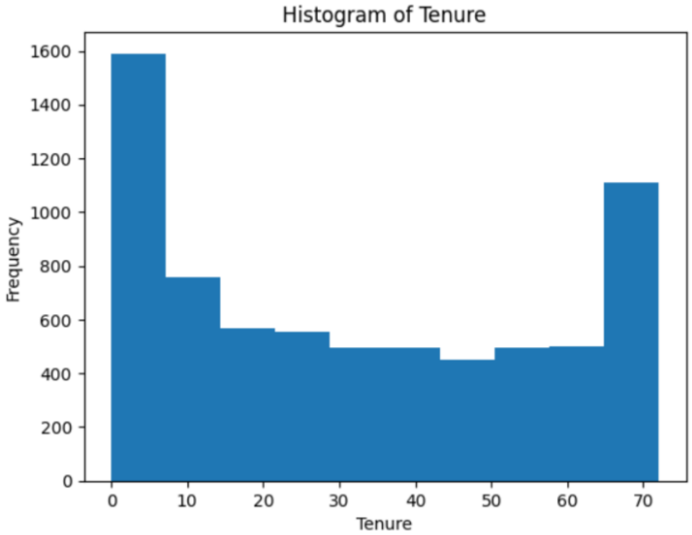


```
  Churn  Count
0    No   5164
1   Yes   1857
```

- This indicates a class imbalance, where most customers do not churn. Inpredictive modeling, addressing such an imbalance may require techniques like oversampling, under sampling.

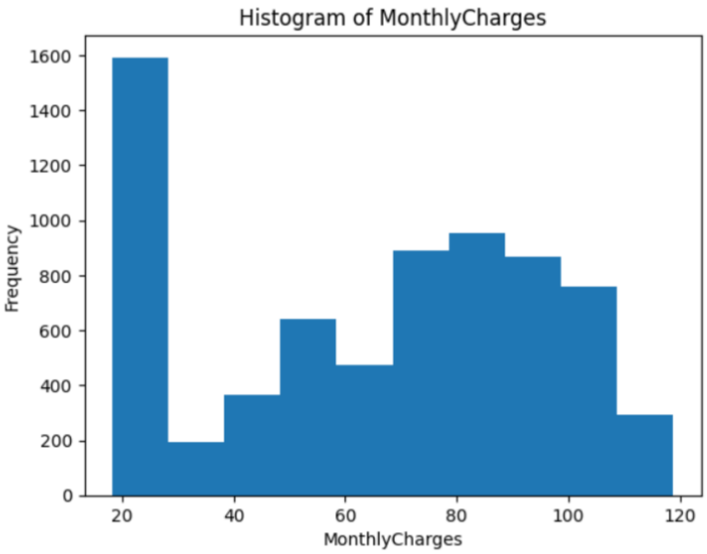
- **Univariate analysis of Numerical columns(Tenure,monthly charges,Total charges)**

i. Distribution of tenure



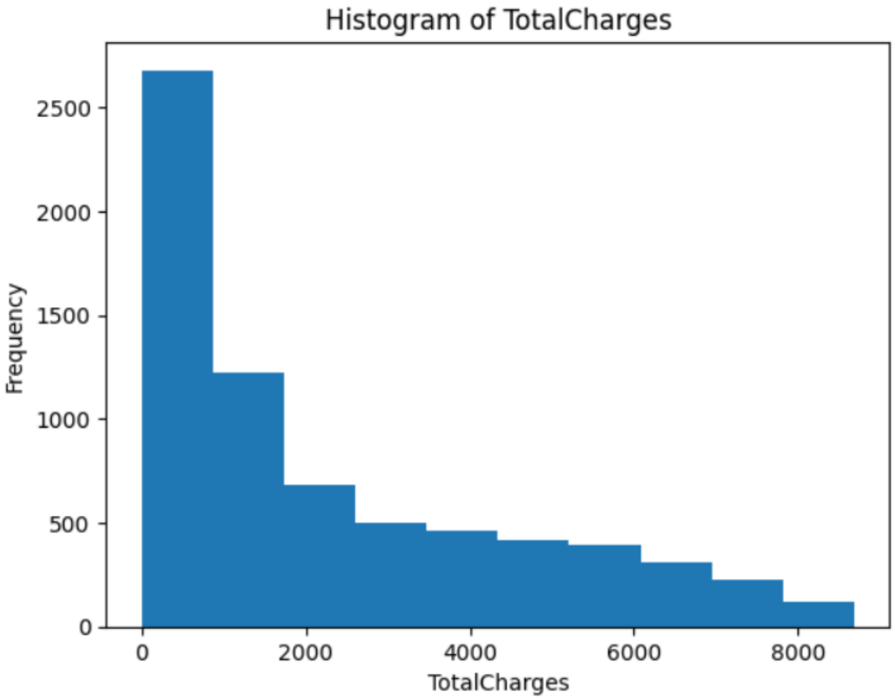
- Customers with very low tenure might represent new users who are likely still exploring the service.
- Customers with the highest tenure (long-term users) could represent a loyal base.
- The histogram suggests opportunities to investigate why customers with short tenure might churn and what keeps long-term customers engaged

ii. Distribution of monthly charges



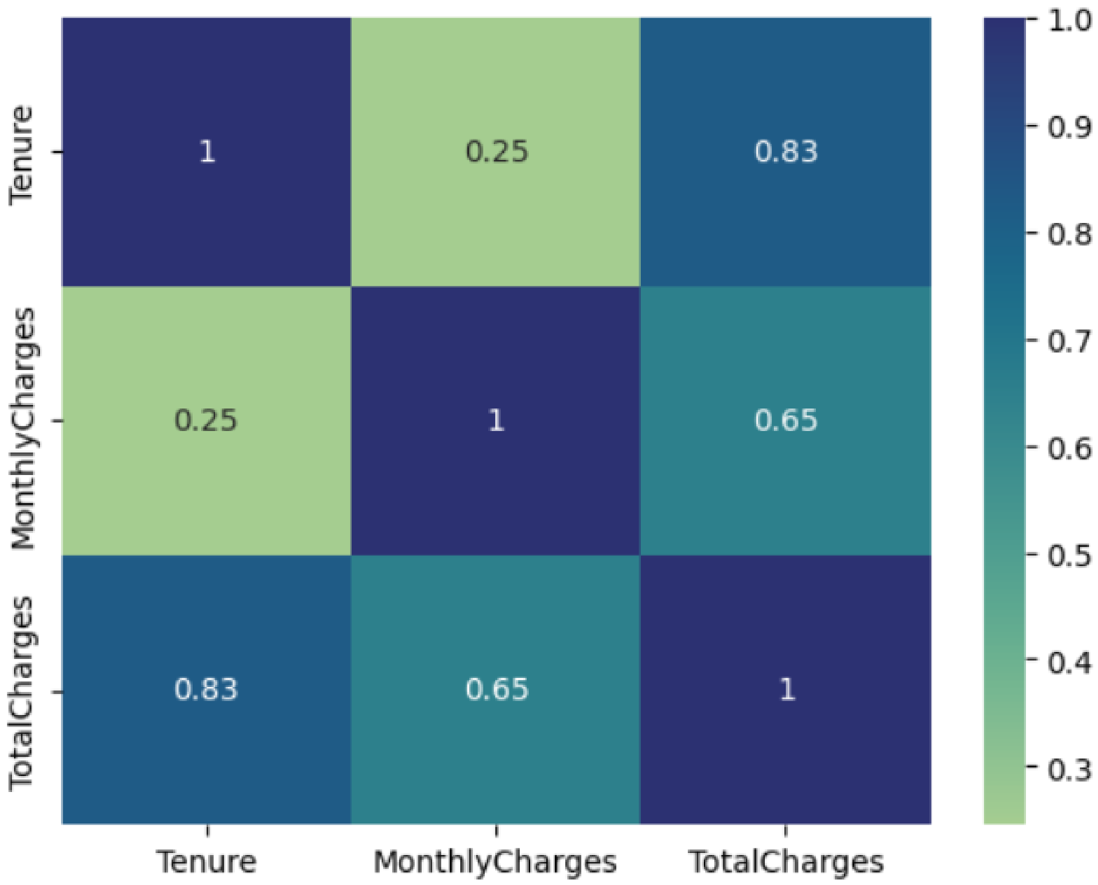
- A **normal distribution** suggests most customers have similar monthly charges.
- A **skewed distribution** (left or right) indicates customer clusters in either low or high monthly charge ranges.

iii. Distribution of Totalcharges



- A positive skew suggests most customers have lower charges, with fewer customers at the higher end.
- Low level charges customers are likely new or on basic plans.
- High level charges customers are on premium plans.

▪ Correlation between Tenure,Monthlycharges and Totalcharges

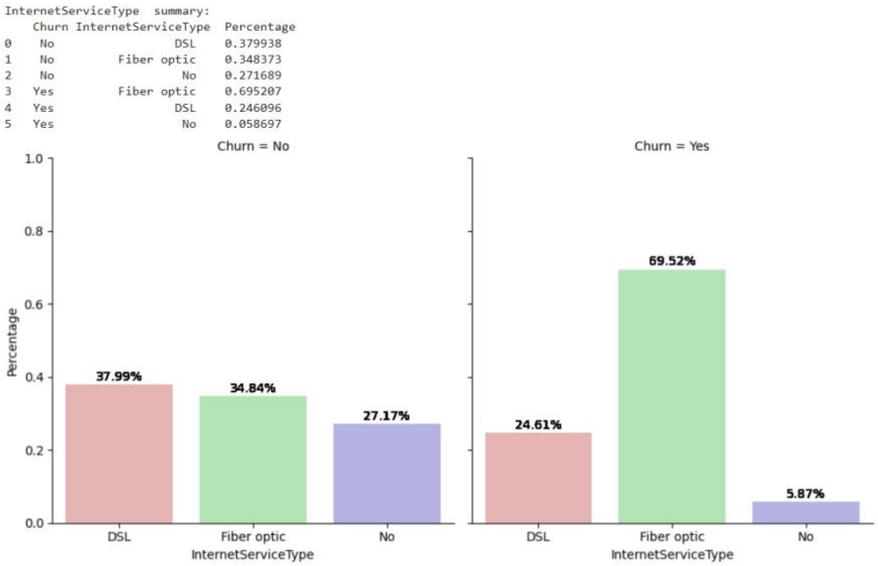


- Strong correlation between Tenure and Total Charges might indicate that retaining customers leads to higher revenue over time.

- Investigating outliers in these relationships (e.g., high Total Charges but low Tenure) can uncover interesting patterns, such as premium or short-term customers.

- **Bivariate analysis of Features Vs Churn**

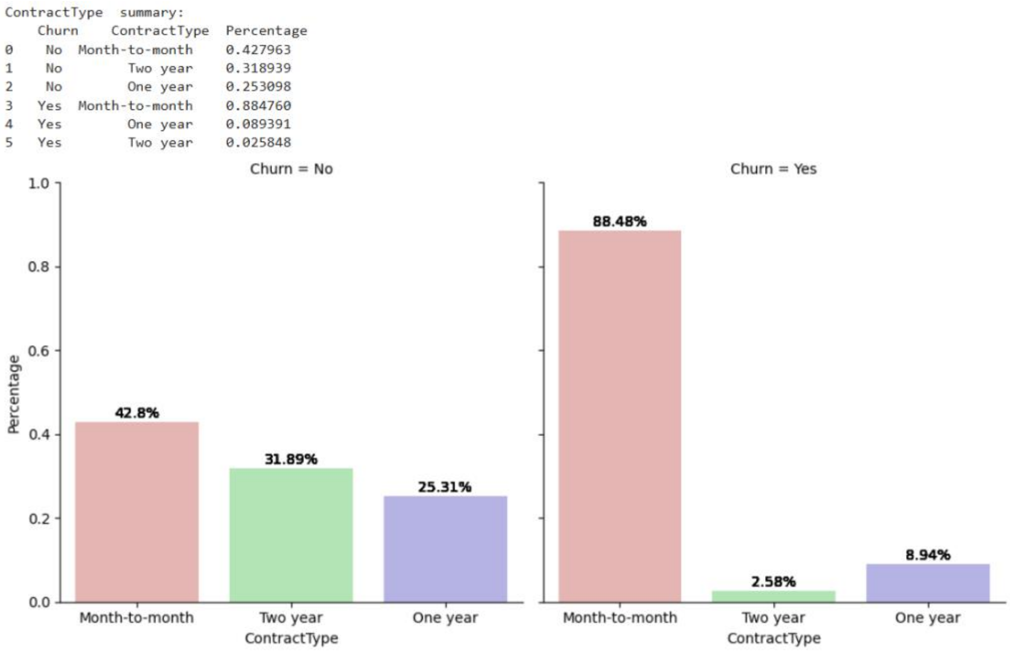
I. Distribution of Internet service type Vs churn



```
InternetServiceType  summary:
   Churn InternetServiceType  Percentage
0    No                 DSL    0.379938
1    No         Fiber optic    0.348373
2    No                  No    0.271689
3   Yes         Fiber optic    0.695207
4   Yes                 DSL    0.246096
5   Yes                  No    0.058697
```

- Churn subscriber likely to have fiber optic (70%) internet service rather than DSL service.

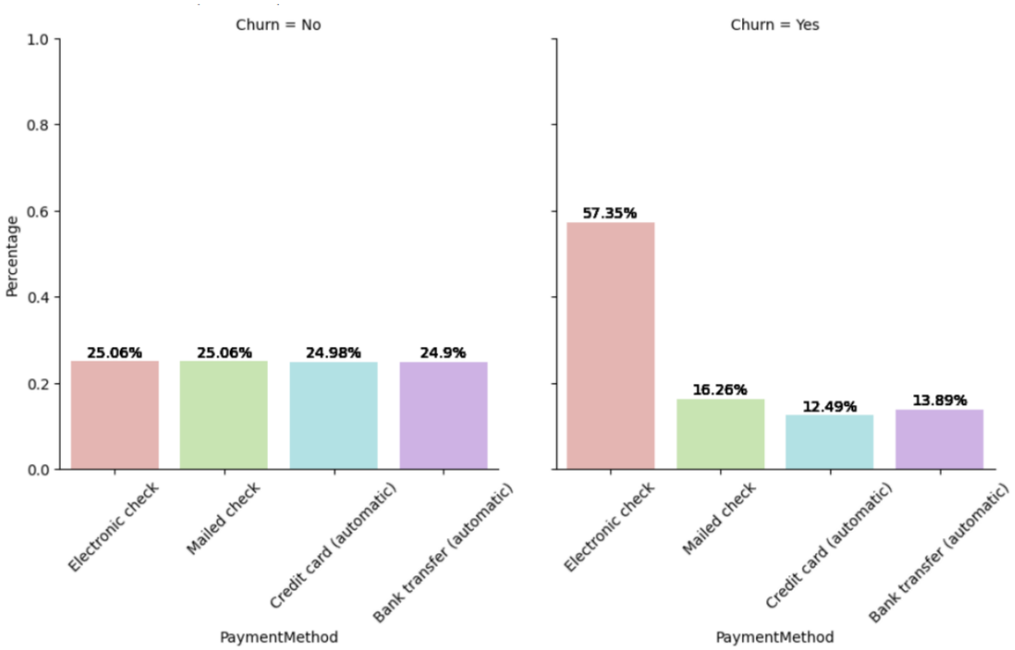
II. Distribution of Contract type Vs churn



```
ContractType  summary:
   Churn    ContractType  Percentage
0    No  Month-to-month    0.427963
1    No        Two year    0.318939
2    No        One year    0.253098
3   Yes  Month-to-month    0.884760
4   Yes        One year    0.089391
5   Yes        Two year    0.025848
```

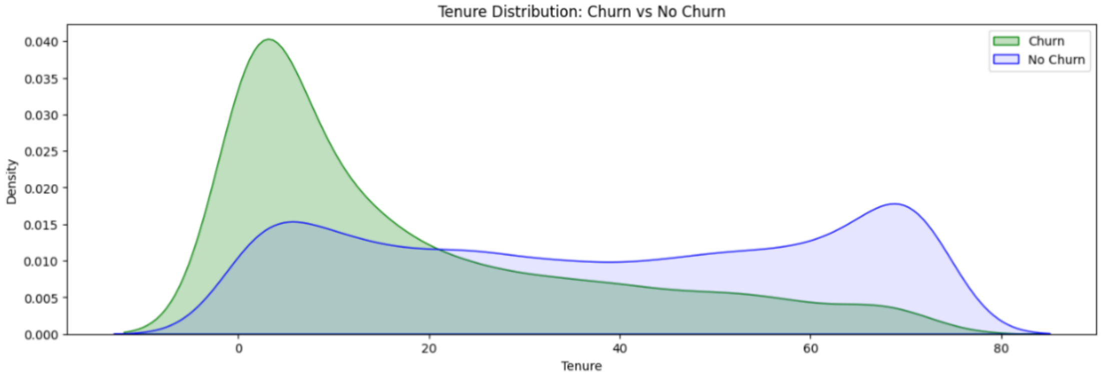
- 88% of churn subscriber has Month-to-month service, means not contracted with company

III. Distribution of Payment method Vs churn



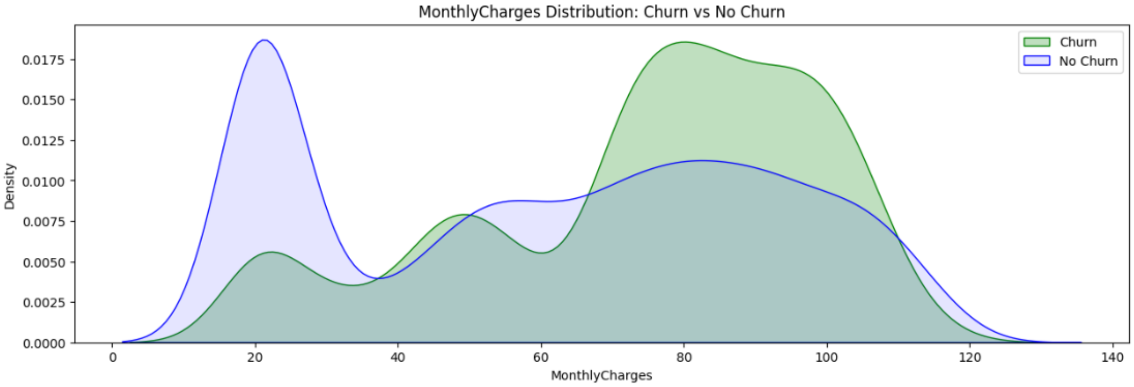
- Churn subscriber most likely to have Elctronic Check service (57%).

IV. Distribution of Tenure with respect to Churn



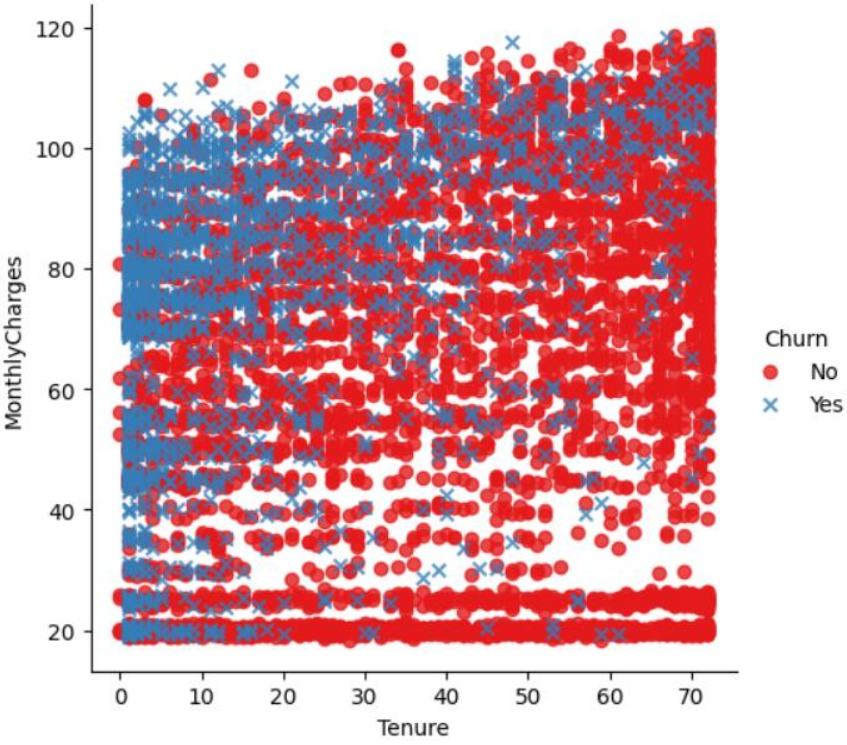
- Tenure Distribution shows that customers who has tenure around less than a year left the brand more

V. Distribution of Monthlycharges with respect to Churn



- Customers who has churned have monthly charges around 80-100 USD.

▪ Relation between Monthly Charges and Tenure



- People who tend to have short tenure and high monthly charges are more likely to churn

Customer Churn Prediction Project Code Explaination

Introduction

In this section, we will explain the steps and logic followed to predict car prices using machine learning. The process includes data importation, preprocessing, exploratory data analysis (EDA), and model building.

Libraries Used:

The following Python libraries were utilized in this project to streamline the workflow and ensure efficient data analysis, preprocessing, and modelling:

- **Pandas**:
  Pandas is a powerful data manipulation and analysis library. It was used to load the dataset, handle missing values, explore data patterns, and perform transformations such as grouping, filtering, and cleaning.

- **Matplotlib**:
  Matplotlib is a versatile library for creating static, interactive, and animated visualizations. It was used for generating basic plots like bar charts, line graphs, and scatter plots to understand data trends and distributions.

- **Seaborn**:
  Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive statistical graphics. It was used to create more advanced visualizations like heatmaps, boxplots, and pair plots, which helped in understanding correlations and feature distributions.

- **Scikit-learn (sklearn)**:
  Scikit-learn is a comprehensive library for machine learning. It was used extensively for preprocessing and modelling:

  - **OneHot Encoding**: This technique was used to convert categorical variables into a numerical format by creating binary columns for each category, ensuring the data could be effectively processed by machine learning algorithms.

  - **Standard Scaling**: This method was applied to normalize numerical features by scaling them to have a mean of 0 and a standard deviation of 1, ensuring that features with different scales do not disproportionately influence the model.

  - **Pipeline**: A key feature of Scikit-learn used in this project to combine multiple preprocessing steps, such as encoding and scaling, into a single streamlined workflow, making the process more efficient and less error-prone.

Data Loading and Inspection

The dataset obtained from Kaggle consists of 7043 rows and 21 columns, including the target variable, 'Churn'. The following steps were performed:

1. Data Inspection and Conversion:

Inspected the dataset to identify numerical and categorical columns.
Converted object-type columns into numerical and categorical formats as needed.

2. Exploratory Data Analysis (EDA):

Performed EDA on both numerical and categorical columns to understand the data distribution and relationships.
Conducted outlier inspection for numerical columns, and no significant outliers were detected.

3. Handling Columns:

Checked and managed data types, null values, and duplicates.
Dropped the 'CustomerID' column as it was not relevant for analysis.

Handling Null Values

After converting the "TotalCharges" column to numeric, we observed the presence of some null values.

To handle these:

1. Data Visualization: We plotted the data to understand the distribution and behavior of "TotalCharges" with respect to other variables.

2. Mathematical Imputation: We filled the null values using the formula:

TotalCharges = Tenure × MonthlyCharges.

3. Validation: After imputation, we rechecked the graph, and it remained consistent with the original distribution. Thus, we proceeded with this approach confidently.

Feature Engineering and Preprocessing

To prepare the data for modeling, the following preprocessing steps were performed:

1. One-Hot Encoding and Standard Scaling:

A pipeline was used to apply One-Hot Encoding to categorical features and Standard Scaling to numerical features to ensure they are in the same scale for model training.

2. Column Transformer:

A ColumnTransformer was used to combine the transformations into a single step, applying the respective encodings and scalings to the appropriate columns efficiently.

Model Building and Evaluation

Three different models were evaluated for predicting customer churn: Logistic Regression, Decision Tree, and Random Forest. Each model was trained using the preprocessed data and evaluated using the classification report to assess their performance on various metrics such as precision, recall, F1-score, and accuracy.

Logistic Regression

Classification Report:

	Precision	**Recall**	**F1-score**	**Support**
0	0.85	0.88	0.86	1053
1	0.60	0.54	0.57	352
Accuracy			0.79	1405
Macro avg	0.72	0.71	0.72	1405
Weighted avg	0.79	0.79	0.79	1405

Decision Tree

Classification Report:

	Precision	Recall	F1-score	Support
0	0.89	0.84	0.86	1117
1	0.49	0.60	0.54	288
Accuracy			0.79	1405
Macro avg	0.69	0.72	0.70	1405
Weighted avg	0.81	0.79	0.80	1405

Random Forest

Classification Report:

	Precision	Recall	F1-score	Support
0	0.90	0.84	0.87	1134
1	0.47	0.62	0.54	271
Accuracy			0.79	1405
Macro avg	0.69	0.73	0.70	1405
Weighted avg	0.82	0.79	0.80	1405

Advanced Models

XGBoost

XGBoost outperformed all other models in terms of accuracy.

The performance of XGBoost was significantly better due to its ability to handle complex relationships in the data through boosting and regularization techniques.

Classification Report:

	Precision	**Recall**	**F1-score**	**Support**
0	0.89	0.85	0.87	1098
1	0.53	0.61	0.57	307
Accuracy			0.80	1405
Macro avg	0.71	0.73	0.72	1405
Weighted avg	0.81	0.80	0.80	1405

Hyperparameter Tuning for XGBoost

To optimize the performance of the XGBoost model, hyperparameter tuning was performed using techniques such as grid search or randomized search. The goal was to find the best set of hyperparameters to maximize the model's accuracy.

Effect on Accuracy:
After tuning the hyperparameters, the accuracy decreased by approximately 1% compared to the baseline model.

Decision:
Despite the slight drop in accuracy, we decided to proceed with the tuned XGBoost model. This decision was based on other factors such as model stability, precision, recall, and the ability to handle imbalances in the dataset.

Classification Report:

	Precision	**Recall**	**F1-score**	**Support**
0	0.89	0.84	0.86	1113
1	0.49	0.60	0.54	292
Accuracy			0.79	1405
Macro avg	0.69	0.72	0.70	1405
Weighted avg	0.81	0.79	0.80	1405

Model Serialization

Once the XGBoost model was finalized, it was saved as a serialized pickle file. This allows the model to be easily deployed and integrated into a web application for real-time customer churn prediction.

Saving the Model:

The XGBoost pipeline was serialized into a pickle file, which stores the trained model and all preprocessing steps (like encoding and scaling). This ensures that the model can be loaded and used without needing to retrain it each time.

Model Integration and Deployment with Streamlit

After saving the trained XGBoost model as a pickle file, it was integrated into a Streamlit application. The application provides an interactive interface for users to input customer data and receive churn predictions.

Process:

1. Loading the Model: The pickle file containing the trained XGBoost model was loaded into the app_model.py script.

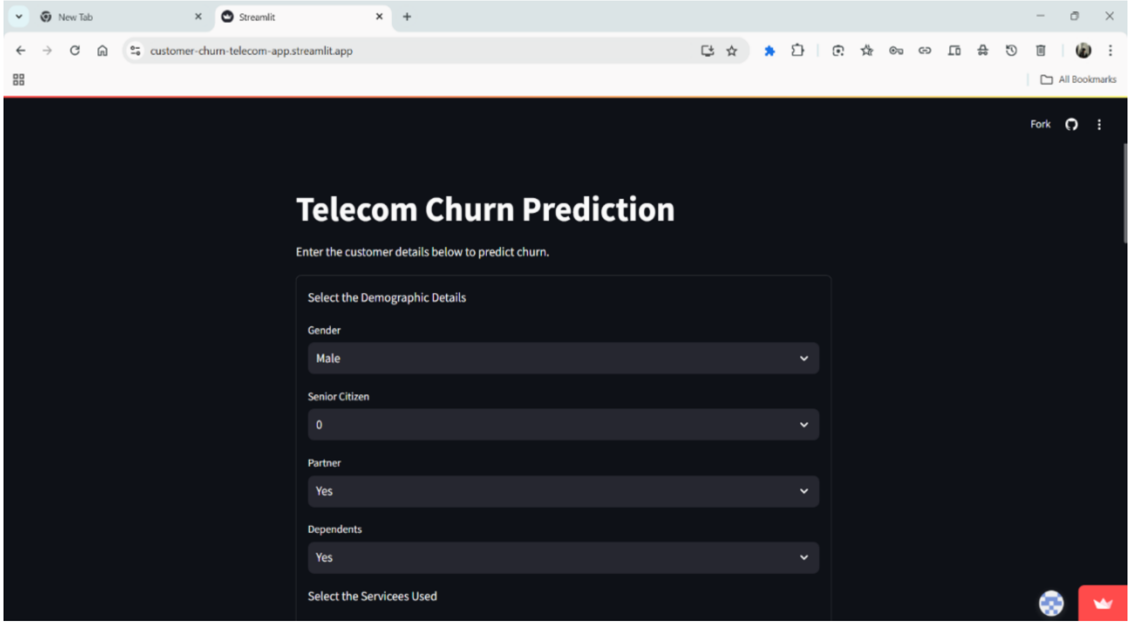
2. User Input: Customers provide their details through the Streamlit interface.

3. Prediction: Upon clicking the 'Predict' button, the app predicts whether the customer will churn and provides the probability of churn.

Streamlit Interface:

The interface is simple and user-friendly, designed to allow easy input of customer data and display the churn prediction results in real time.

Result



Telecom Churn Prediction

Enter the customer details below to predict churn.

Select the Demographic Details

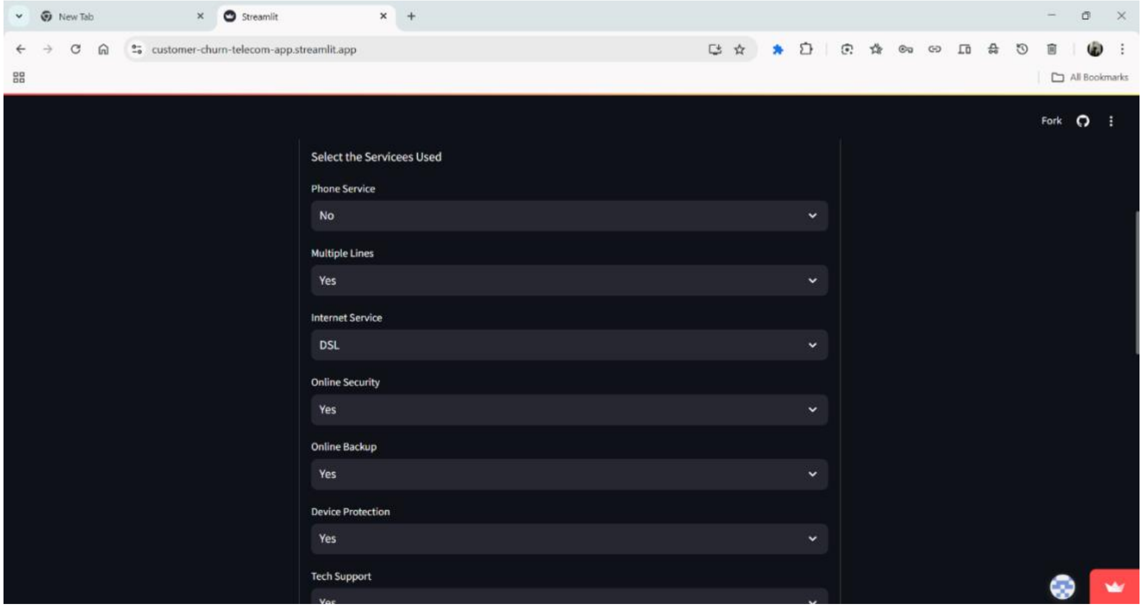
Gender: Male

Senior Citizen: 0

Partner: Yes

Dependents: Yes

Select the Servicees Used



Select the Servicees Used

Phone Service: No

Multiple Lines: Yes

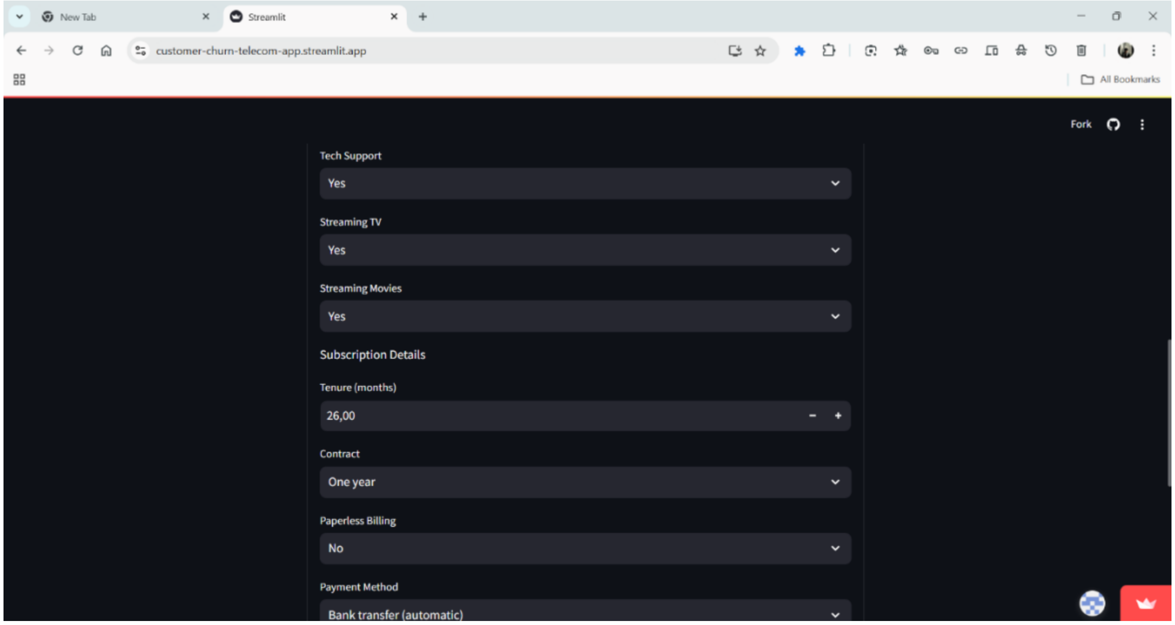
Internet Service: DSL

Online Security: Yes

Online Backup: Yes

Device Protection: Yes

Tech Support: Yes



Tech Support: Yes

Streaming TV: Yes

Streaming Movies: Yes

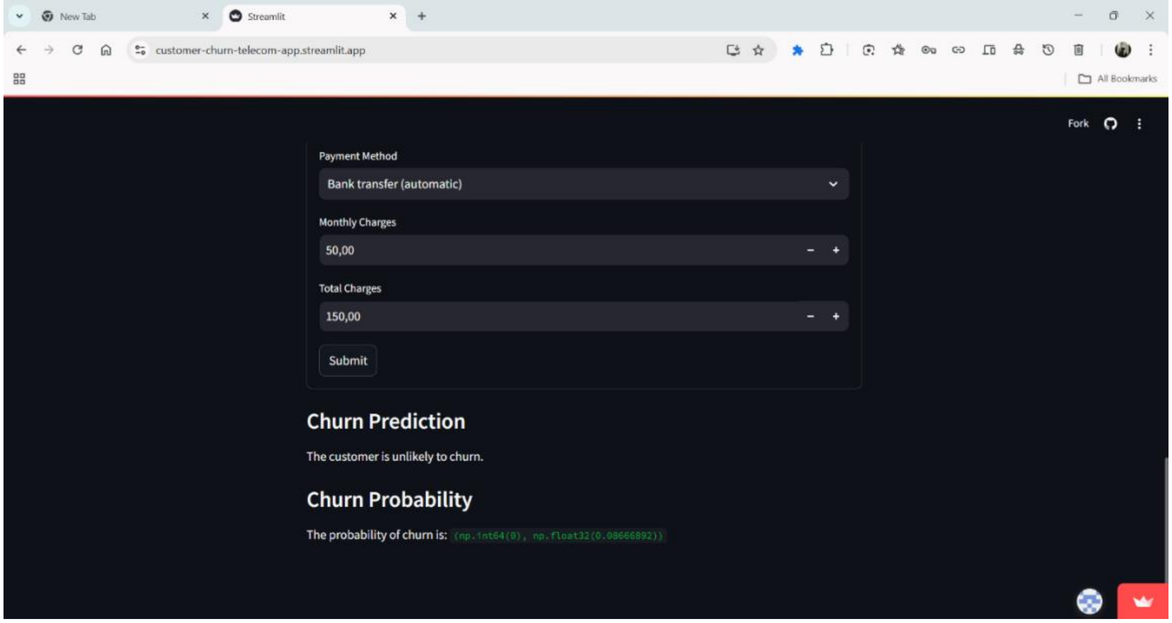
Subscription Details

Tenure (months): 26,00

Contract: One year

Paperless Billing: No

Payment Method: Bank transfer (automatic)



Payment Method: Bank transfer (automatic)

Monthly Charges: 50,00

Total Charges: 150,00

Churn Prediction

The customer is unlikely to churn.

Churn Probability

The probability of churn is: `(np.int64(0), np.float32(0.08666892))`

Conclusion

In this project, we developed a customer churn prediction model using a dataset obtained from Kaggle. The dataset was preprocessed with data cleaning, encoding, and scaling, and several machine learning models were evaluated, including Logistic Regression, Decision Tree, Random Forest, and XGBoost.

Model Performance:
The models were evaluated using accuracy and classification reports. Among all the models tested, XGBoost delivered the highest accuracy and showed the most promise for churn prediction. Although hyperparameter tuning led to a slight drop in accuracy, the overall performance of XGBoost, including its precision and recall, was deemed optimal for this task.

Cross-Validation and Stability:
Cross-validation was performed on the Random Forest model to ensure its stability and reliability. The results showed consistent performance, but XGBoost still outperformed it in the final evaluation.

Deployment:
The trained XGBoost model was serialized using pickle and deployed into a simple Streamlit web application. This web app allows users to input customer data and receive real-time predictions on whether a customer will churn, along with the associated probability.

Future Improvements:

While the current model is robust, there is still room for improvement, particularly in fine-tuning the XGBoost model and exploring additional feature engineering and model enhancements. Further research into hyperparameter optimization and additional models may help achieve even higher performance.

References

https://www.researchgate.net/publication/310757545_A_Survey_on_Customer_Churn_Prediction_using_Machine_Learning_Techniques