**Customer Churn Prediction**

**-Tanzeel Shaikh**

1. **Introduction:**

Customer churn, also known as customer attrition, is a critical metric for businesses to monitor. It refers to the percentage of customers who stop using a company's products or services during a specific period. Predicting churn allows businesses to take proactive measures to retain customers, improve customer satisfaction, and increase revenue.

This project focuses on building a machine learning model to predict customer churn using the Telco Customer Churn dataset. The dataset contains information about telecom customers, including their demographics, services subscribed, and whether they churned or not. Two models were developed: Logistic Regression and Random Forest, and their performance was evaluated using accuracy, precision, and recall metrics.

1. **Dataset Overview:**

The dataset used in this project is the Telco-Customer-Churn file, which contains:

Rows: 7,043 (customers)

Columns: 21 (features)

Target Variable: Churn (binary: Yes/No)

A screenshot of a computer

Description automatically generated

**Technologies used**:

Python, scikit-learn (for machine learning algorithms), Pandas (data manipulation), joblib.

**Key Features:**

* **Demographics**: gender, SeniorCitizen, Partner, Dependents.
* **Services**: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV,  StreamingMovies.
* **Account Information**: Contract, PaperlessBilling, PaymentMethod,  MonthlyCharges, TotalCharges, Tenure.

1. **Data Preprocessing**

**3.1 Data Cleaning:**

* The Churn column was converted to binary values:

1 for "Yes" (churned)

0 for "No" (not churned).

* The TotalCharges column contained empty strings, which were replaced with NaN and then filled with the mean value of the column.
* The customerID column was dropped as it is not relevant for modeling**.**

**3.2 Feature Engineering**

* Categorical columns were converted into numerical format using one-hot encoding.
* These columns were encoded:

gender, Partner, Dependents, PhoneService, MultipleLines,

InternetService, OnlineSecurity, OnlineBackup, DeviceProtection,

TechSupport, StreamingTV, StreamingMovies, Contract,

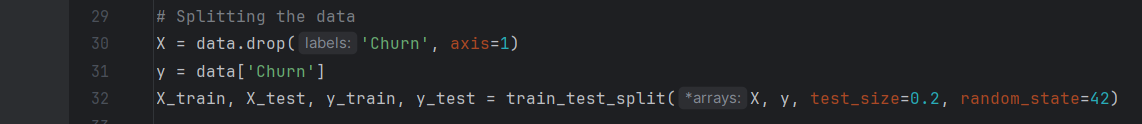
A computer screen shot of text

Description automatically generated PaperlessBilling, and PaymentMethod.

1. **Model Development:**

**4.1 Data Splitting**

* The dataset was split into training and testing sets using an 80:20 ratio:
  + Training Data: 80% of the dataset
  + Testing Data: 20% of the dataset



**4.2 Logistic Regression**

* A Logistic Regression model was trained on the dataset.
* The model's performance was evaluated using the following metrics:
  + Accuracy: 0.82
  + Precision: 0.69
  + Recall: 0.56

A screenshot of a computer program

Description automatically generated**A screen shot of a computer program

Description automatically generated**

**4.3 Random Forest with Feature Scaling**

* The features were scaled using StandardScaler to normalize the data.
* A Random Forest Classifier was trained on the scaled data.
* The model's performance was evaluated using the following metrics:
  + Accuracy: 0.79
  + Precision: 0.65
  + Recall: 0.46

A screen shot of a computer

Description automatically generatedA computer screen shot of a program code

Description automatically generated

1. **Model Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| **Logistic Regression** | **0.82** | **0.69** | **0.56** |
| **Random Forest (Scaled)** | **0.79** | **0.65** | **0.46** |

**5.2 Insights**

* The Logistic Regression model achieved slightly higher accuracy and precision compared to the Random Forest model.
* Both models had similar recall scores, indicating their ability to correctly identify churned customers.
* The Random Forest model was saved for future use using the joblib library.

1. **Conclusion**

This project successfully developed and evaluated two machine learning models to predict customer churn. The **Logistic Regression** model performed slightly better in terms of accuracy and precision, while the **Random Forest** model was saved for deployment. The results demonstrate the potential of machine learning in identifying at-risk customers and enabling proactive retention strategies.

**8. Future Work**

* **Hyperparameter Tuning**: Optimize model parameters to improve performance.
* **Feature Importance Analysis**: Identify the most significant features contributing to churn.
* **Deployment**: Integrate the model into a real-time system for churn prediction.
* **Additional Data**: Incorporate more features or external data sources to enhance model accuracy.

**9. References**

* Dataset Source: [Telco-customer-churn.csv](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)