Customer Churn Analysis – Telecom Industry

This project focuses on understanding and predicting customer churn in the telecom industry using machine learning techniques and data analytics. The goal is to uncover key factors influencing churn and provide actionable strategies for customer retention.

Tools & Technologies Used:

- Python (Scikit-learn, ELI5, SHAP)
- SQL for data querying and aggregation

Prepared By: Anubhuti Anurag

Project Objectives

Predict Customer Churn

Build a predictive model to accurately classify customers likely to churn.

Identify Key Churn Drivers

Analyze factors that contribute the most to churn, enabling data-driven decisions.

Customer Segmentation

Classify customers into meaningful groups based on behavior, tenure, and risk levels.

Actionable Strategies

Deliver recommendations for targeted interventions to reduce churn and boost loyalty.

Dataset Overview

7043

Customer Records

A comprehensive dataset detailing individual customer interactions and attributes.

20

Features

Including demographics, service details, and billing information for robust analysis.

The dataset includes a rich variety of data, such as:

- **Customer Demographics:** Gender, Citizen Status, Dependents, Partner
- **Service Details:** Internet service, contract type, phone service, multiple lines
- **Billing & Payments:** Monthly charges, total charges, payment methods
- **Target Variable:** 'Churn'—whether the customer has churned (Yes/No)

This mix provides both qualitative and quantitative insights necessary for analysis.

Data Cleaning & Preparation

Data Cleaning Steps:

- **Removed the** `customerID` **column:**The `customerID` was a unique identifier for each customer and held no predictive or analytical value for the churn model. Keeping it would add unnecessary noise to the dataset.

- **Handled missing values in** `TotalCharges`**:**
  - Discovered that some rows had blank entries in the `TotalCharges`field.
  - Investigated these blanks and found they mostly occurred when tenure was zero (i.e., new customers).
  - Converted these blanks into float values, typically replacing empty strings with `0.0` or assigning correct values based on the context.
- **Validated the dataset:**
  - Conducted thorough checks to ensure no missing, null, or invalid values remained in any columns.
  - Ensured all features were in the correct data types (e.g., numerical columns like MonthlyCharges and TotalCharges properly converted from string to float).

Feature Engineering & Encoding

Target Variable Encoding

Converted the target ‘Churn’ to numerical format: **1 = Yes, 0 = No** for compatibility with ML models.

Categorical Feature Encoding

Used **Label Encoding** for object-type columns such as gender, contract type, payment methods, etc.

Encoder Storage

Stored encoders in a **pickle file** to maintain consistency during model deployment and future predictions.

This step is crucial for transforming categorical variables into a machine-readable format, preparing them for robust machine learning analysis.

Exploratory Data Analysis (EDA)—Numerical Insights

Exploratory Data Analysis revealed critical numerical insights into customer behavior and churn tendencies.

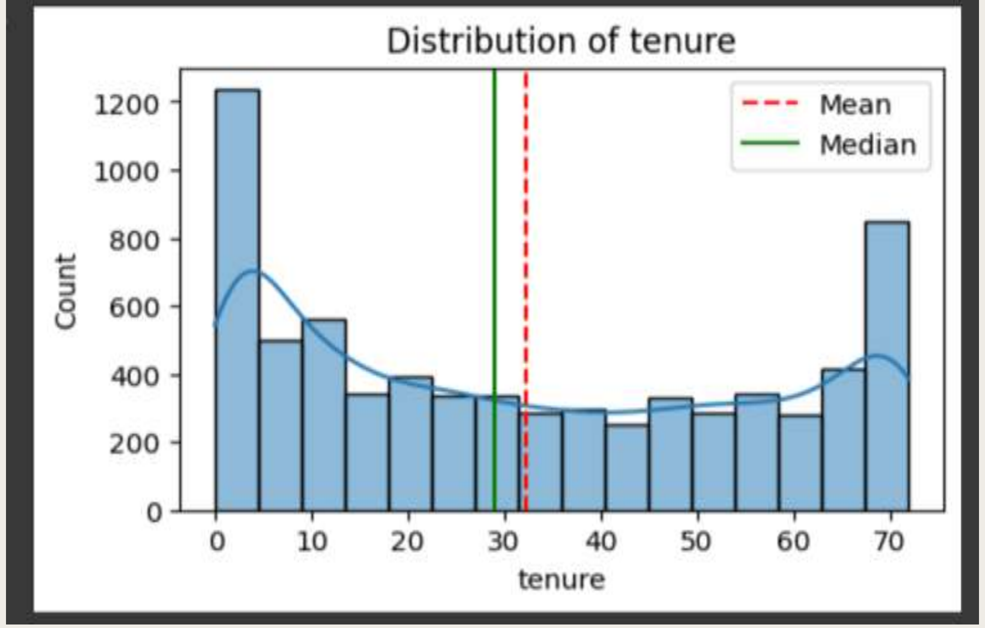
Numerical Features Explored:

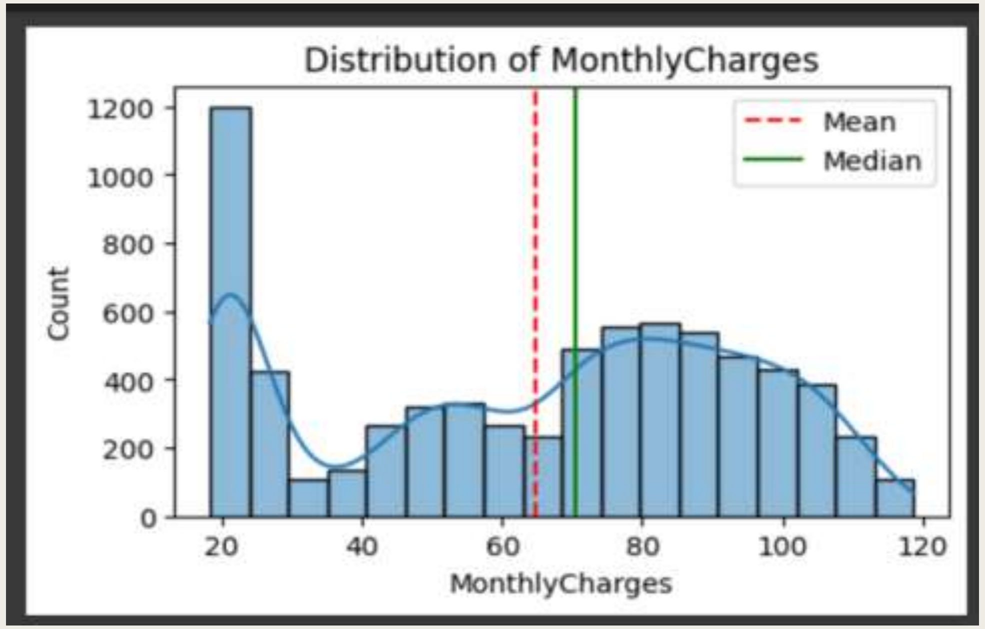
- **Tenure:** Shows customer loyalty period; lower tenure often correlates with higher churn.
- **Monthly Charges:** Higher charges sometimes lead to dissatisfaction and higher churn.
- **Total Charges:** A combination of tenure and monthly charges, reflecting overall customer value.

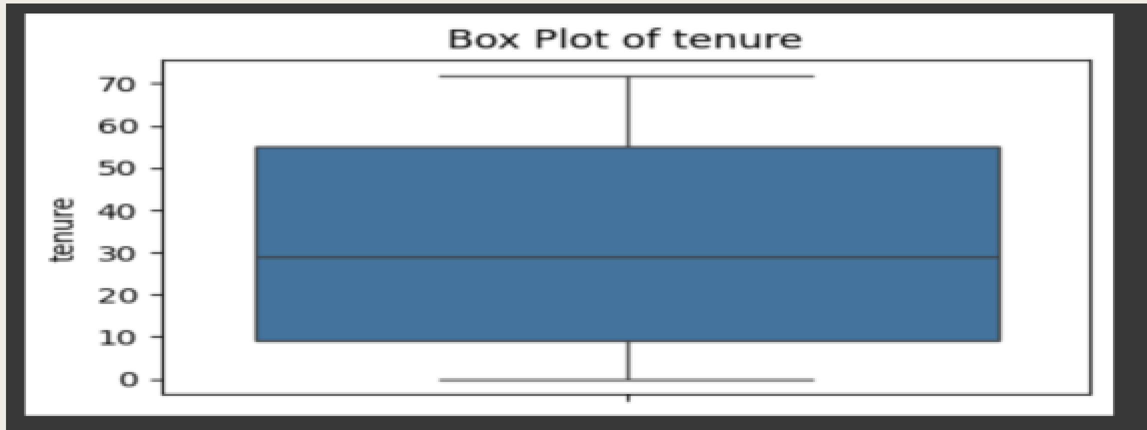
EDA Techniques Applied:

- Created **histograms** and **boxplots** to visualize distributions and detect outliers.
- Plotted a **correlation heatmap** to understand relationships between numeric variables.

These insights guide feature importance and model building, providing a solid foundation for predicting churn.







Exploratory Data Analysis (EDA) – Categorical Insights

Contract Type

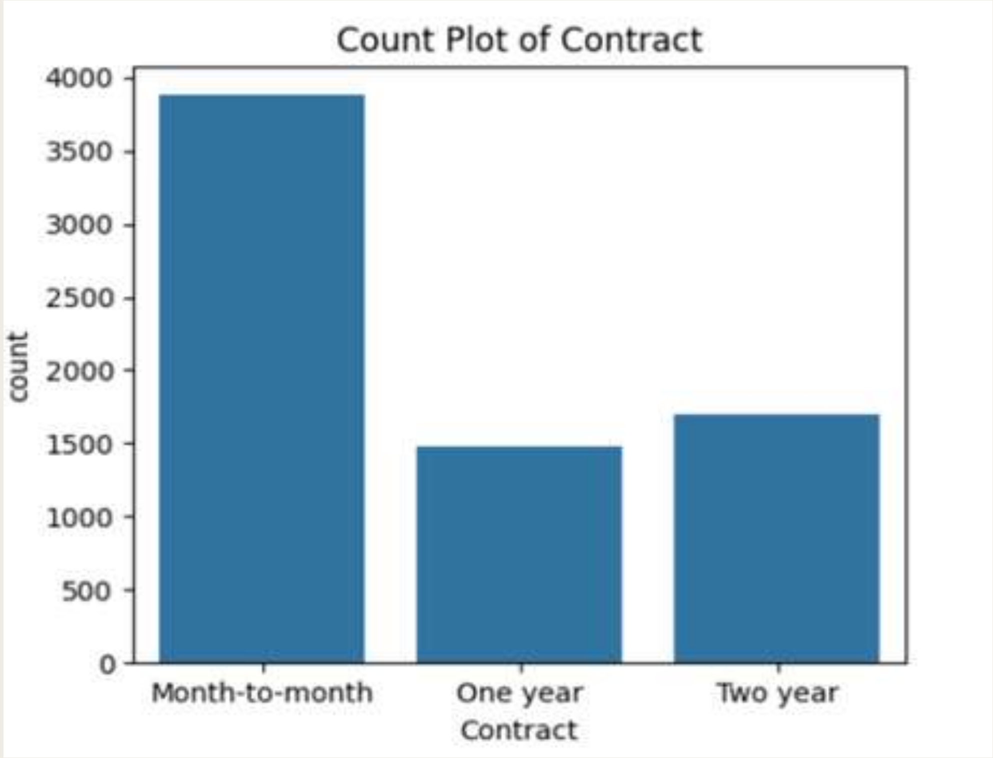
Month-to-month customers churn more frequently than those with annual or two-year contracts, indicating the importance of long-term commitments.

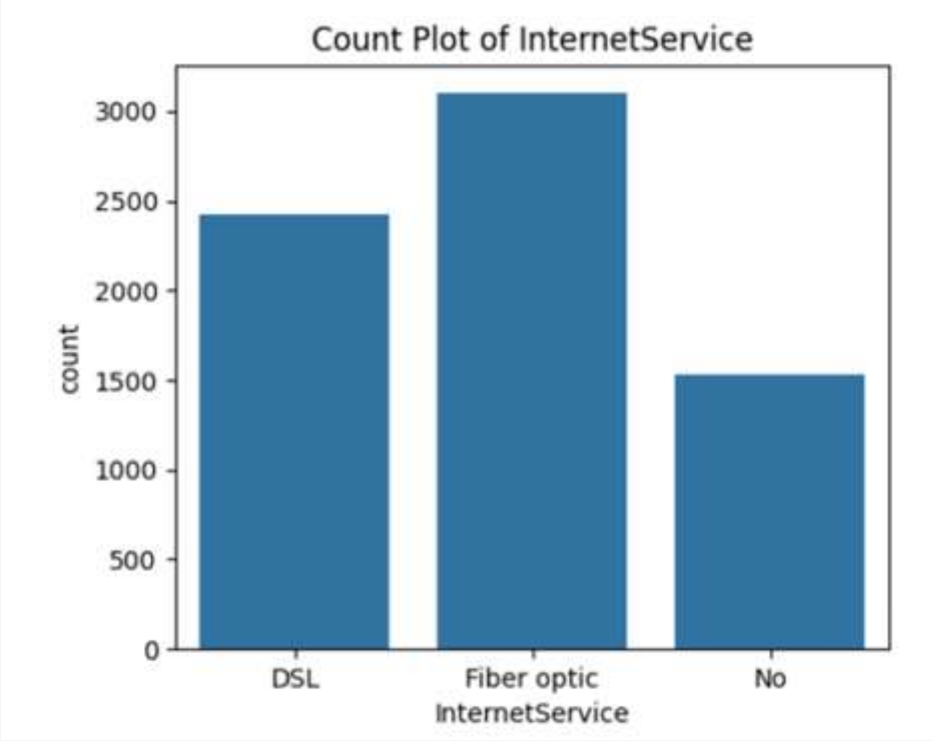
Internet Service

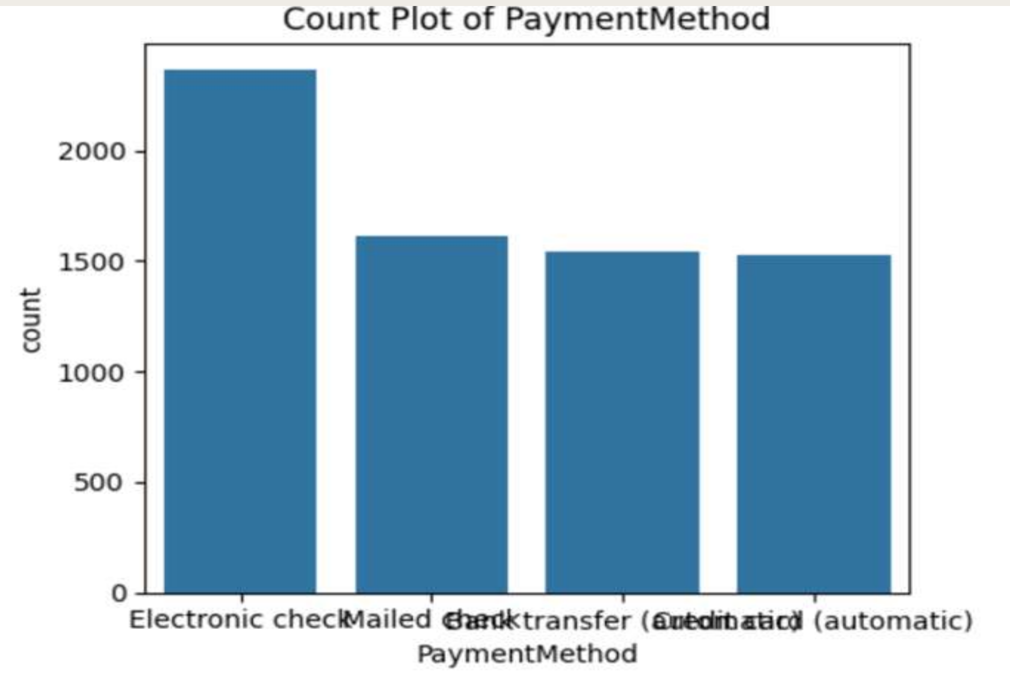
Customers with fiber-optic services showed higher churn rates, potentially due to higher charges or inconsistent service quality.

Payment Method

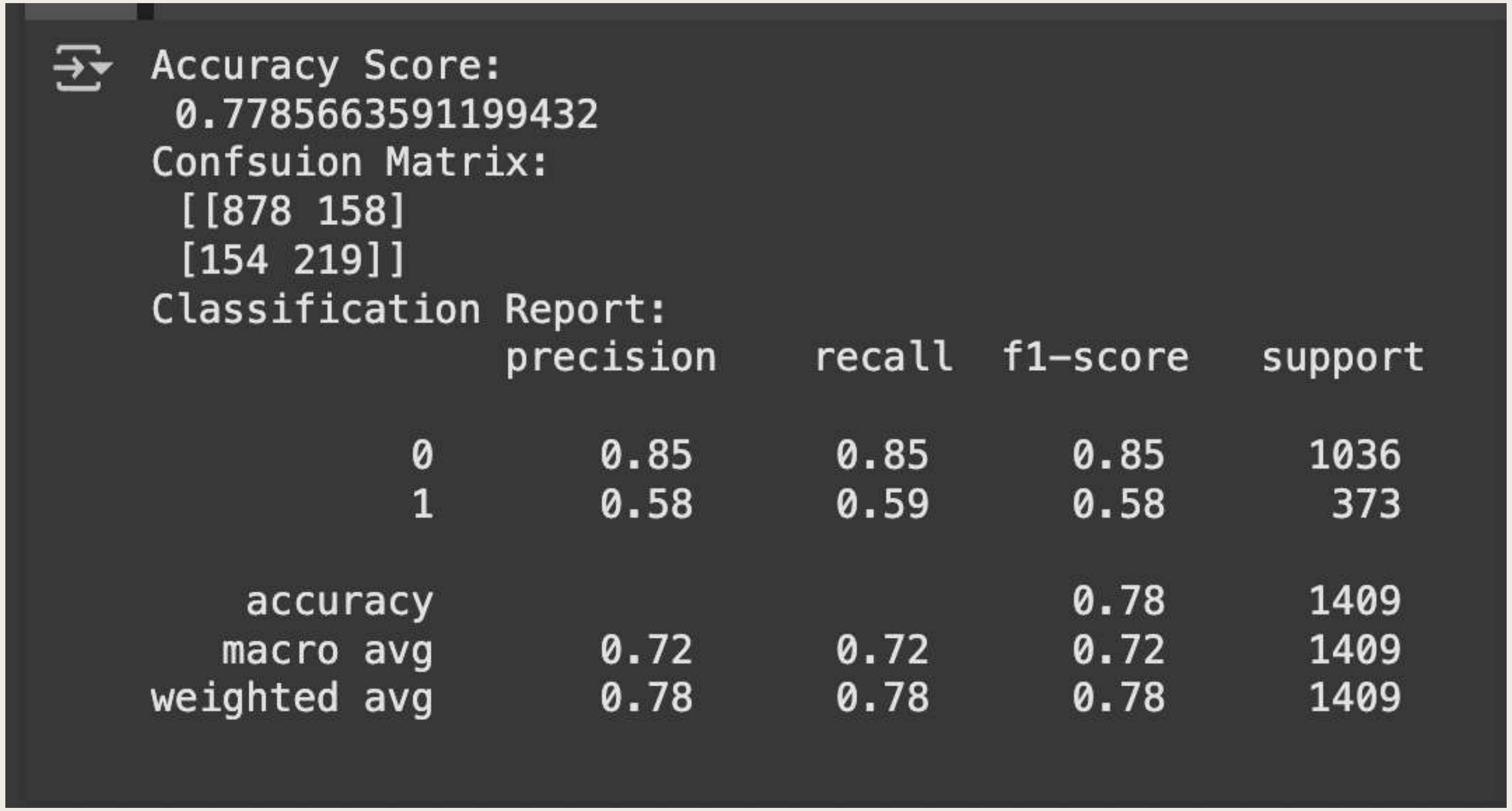
Electronic check users exhibit significantly higher churn rates compared to those using automatic payment methods like bank transfer or credit card.







Model Development & Evaluation



```
Accuracy Score:
 0.7785663591199432
Confsuion Matrix:
 [[878 158]
 [154 219]]
Classification Report:
               precision    recall  f1-score   support

           0       0.85      0.85      0.85      1036
           1       0.58      0.59      0.58       373

    accuracy                           0.78      1409
   macro avg       0.72      0.72      0.72      1409
weighted avg       0.78      0.78      0.78      1409
```

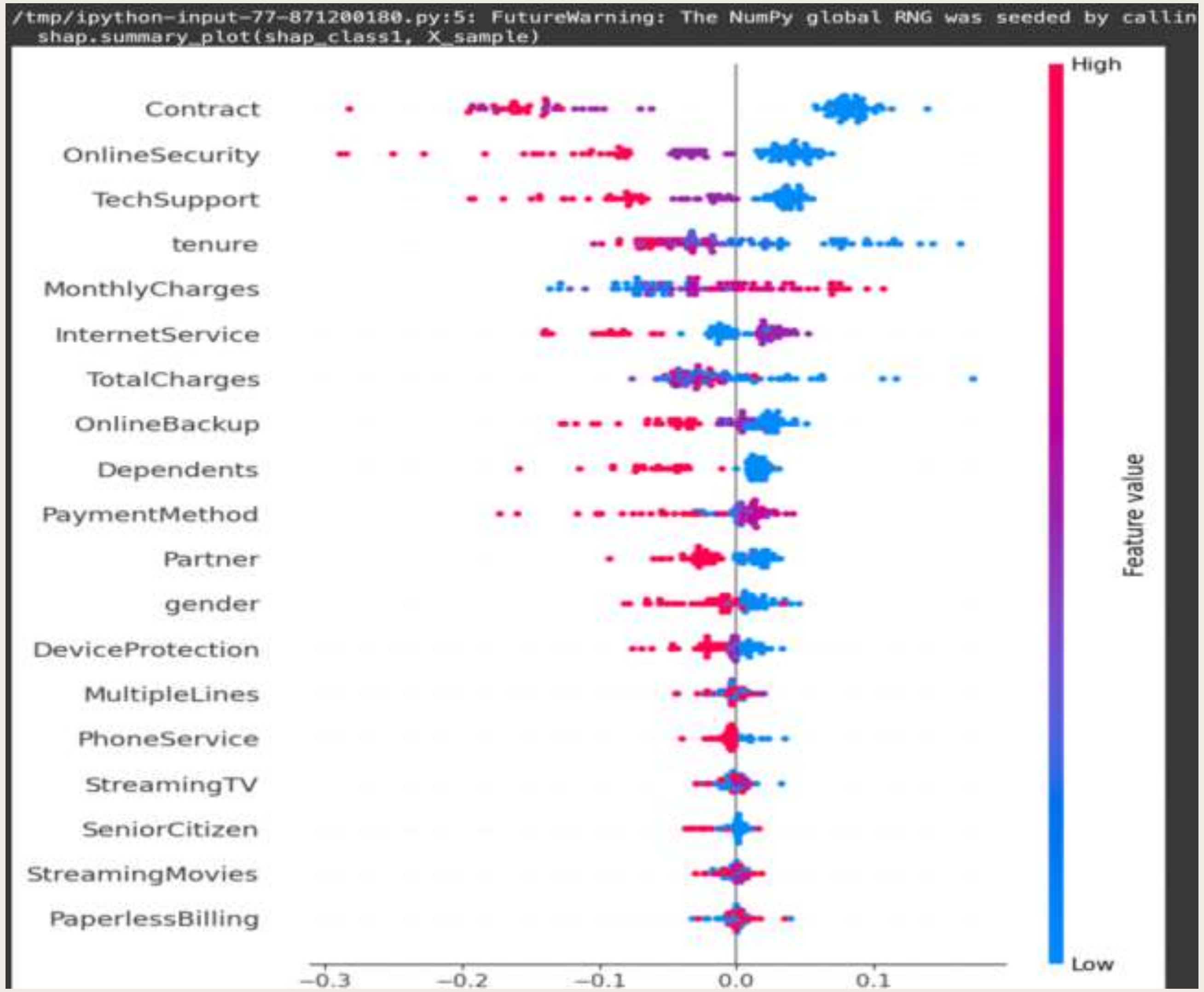
- **Handling Class Imbalance:** Churn (~26.5%) vs. non-churn (~73.5%) shows imbalance. Applied **SMOTE** (Synthetic Minority Oversampling Technique) to balance classes in the training dataset.
- **Train-Test Split:** The dataset is split into **80% training** and **20% testing**.
- **Models Evaluated:** Decision Tree, Random Forest, and XGBoost. **Random Forest** delivered the best performance with ~**84% accuracy**, proving robust and generalizing well.
- **Validation Method:** 5-fold cross-validation was used to ensure model reliability and prevent overfitting.

Model Interpretation & Customer Segmentation

Model Explainability (SHAP Values)

Used **SHAP (SHapley Additive Explanations)** for a detailed feature importance analysis. The top influential features identified were:

- **Tenure:** Length of time a customer has been with the service.
- **Monthly Charges:** The amount billed to the customer each month.
- **Contract Type:** The nature of the customer's service agreement (e.g., month-to-month, annual).
- **Internet Service:** The type of internet service subscribed to by the customer.



```
/tmp/ipython-input-77-871200180.py:5: FutureWarning: The NumPy global RNG was seeded by callin
  shap.summary_plot(shap_class1, X_sample)
```

Customer Segmentation Based on Churn Risk:

At Risk

Customers with high churn probability (~40% churn rate) need proactive retention strategies.

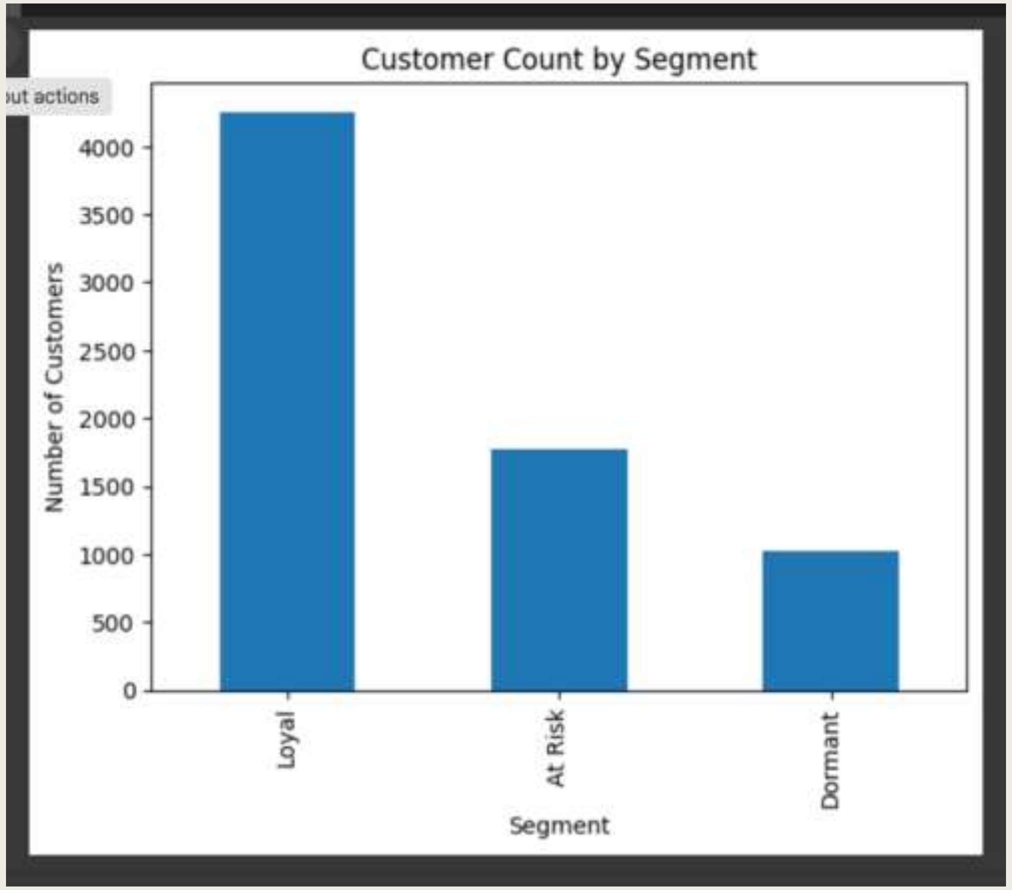
Dormant

Customers with low activity or engagement (~18% churn rate); potential for win-back campaigns.

Loyal

Customers with long tenure and stable payments (<2% churn risk). Focus on rewards and continued engagement to retain them.

This segmentation is crucial for crafting targeted marketing and retention strategies, maximizing impact and efficiency.



References

- **Dataset Source:** Telco Customer Churn Dataset – Kaggle
- **Python Libraries:** Scikit-learn, Pandas, Matplotlib, Seaborn, ELI5, SHAP
- **Research Papers & Documentation:**
  - SMOTE for handling class imbalance
  - SHAP values for model interpretation
- **SQL queries** used for customer behavior aggregation