Customer Churn Prediction

Documentation

By – Sayan Kashyap

Contents

1. Problem Statement:

Customer churn is a critical issue for businesses, especially in the telecom sector, where acquiring new customers is often more expensive than retaining existing ones. The objective of this project is to predict and prevent customer churn using historical data. By implementing data-driven strategies, we aim to enhance customer retention.

2. Dataset Overview:

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

Description: This dataset contains customer data, including demographic details, subscription details, service usage, and whether the customer has churned.

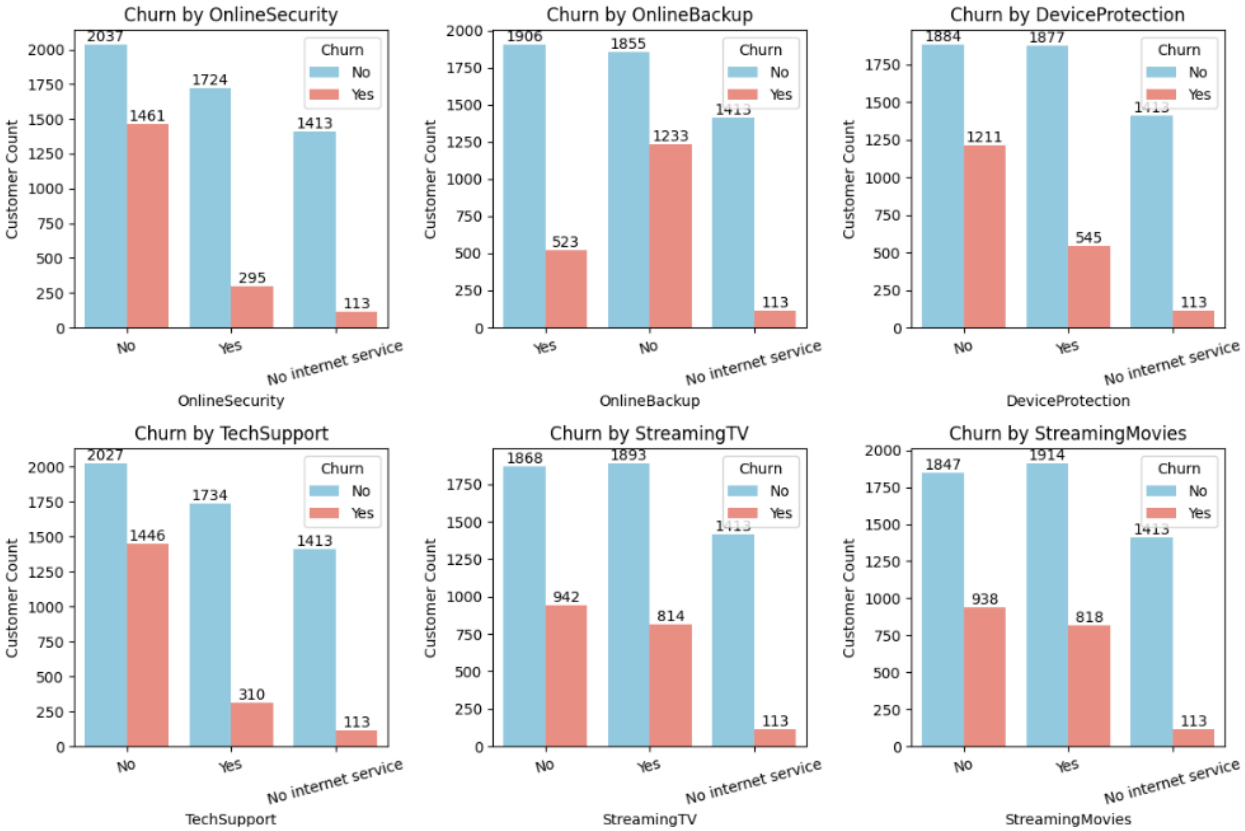
Key Attributes:

- **CustomerID**: Unique identifier for each customer
- **Gender, SeniorCitizen, Partner, Dependents**: Demographic details
- **Tenure, Contract, Payment Method**: Subscription and billing information
- **MonthlyCharges, TotalCharges**: Financial details
- **Churn**: Target variable indicating if the customer has left

3. Exploratory Data Analysis (EDA):

EDA was performed to gain insights into the dataset and identify key trends:

- **Data Cleaning**:
    - Missing values in TotalCharges were handled appropriately.
    - Categorical data was encoded for model compatibility.
- **Data Distribution & Trends:**
    - Churn rate distribution was analyzed.
    - Relationships between churn and variables like tenure, contract type, and payment method were explored.
- **Visualizations:**
    - Histograms and boxplots were used to analyze numerical features.
    - Bar charts showed the distribution of categorical variables.
    - Correlation heatmaps identified feature relationships.



4. Feature Engineering:

To enhance model performance, feature engineering techniques were applied:

- **New Features Created:**
  - Average Monthly Spend derived from TotalCharges and Tenure
  - Tenure Category based on duration of subscription
- **Encoding:**
  - One-hot encoding for categorical variables
  - Label encoding for binary categorical features
- **Scaling:**
  - Standardization applied to numerical features

5. Machine Learning Modelling:

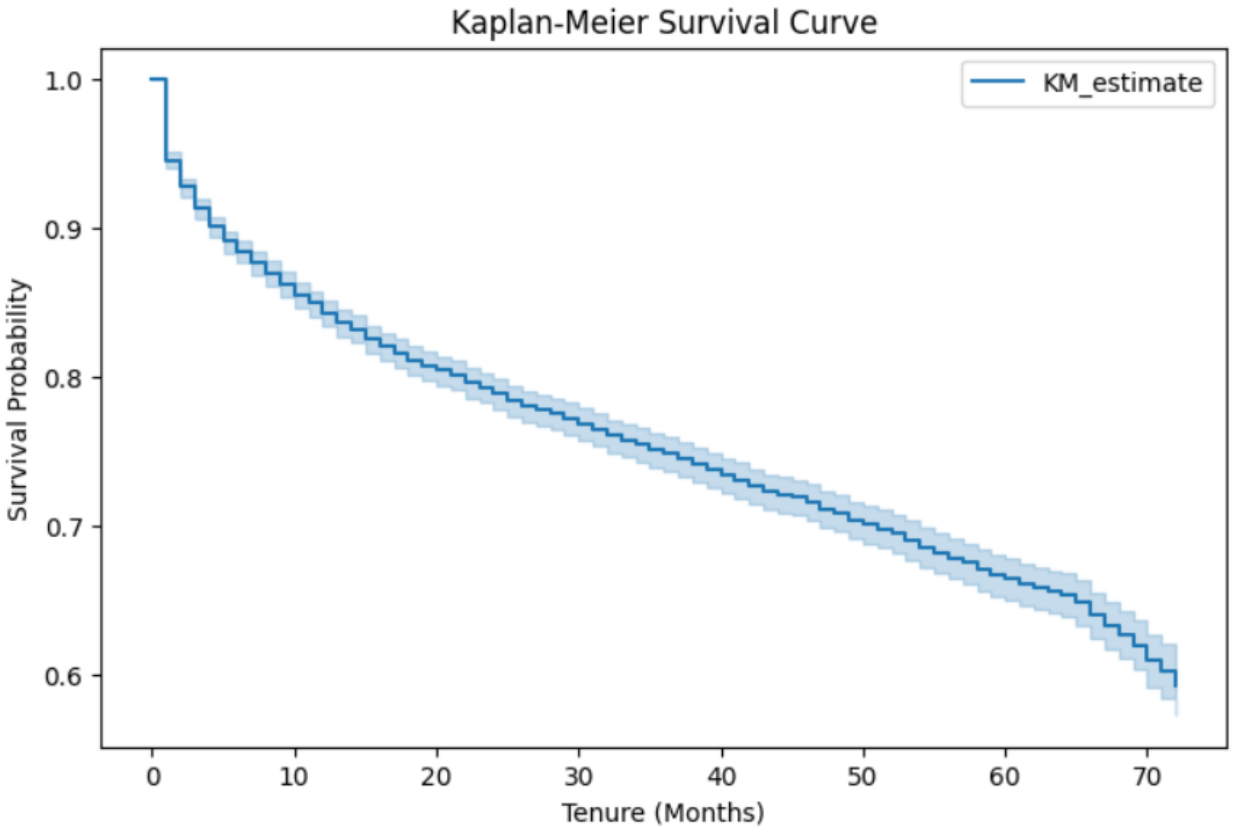
The following steps were followed for churn prediction:

- **Model Selection:**
  - Baseline models: Logistic Regression, Decision Tree
  - Advanced models: Random Forest, Gradient Boosting (XGBoost), Neural Networks
- **Model Training & Evaluation:**
  - Models were trained on the dataset with train-test split (80-20 ratio).
  - Evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC
  - Best performing model: **XGBoost with an AUC-ROC score of 0.85**
- **Hyperparameter Tuning:**
  - Grid Search and Random Search were used for optimization.
  - Best parameters identified for XGBoost resulted in improved performance.

6. Survival Analysis:

Survival analysis technique is used to estimate the time until an event (e.g., customer churn) occurs. It helps analyze retention rates over time. It was applied to understand customer retention probabilities over time:

- **Kaplan-Meier Curve:**
    - Estimated customer survival probability based on tenure.
- **Cox Proportional Hazards Model:**
    - Identified key factors influencing customer churn.
- **Key Insights:**
    - Customers with month-to-month contracts had higher churn risks.
    - Long-term contracts and automatic payments reduced churn likelihood.



7. Automated Customer Scoring Workflows:

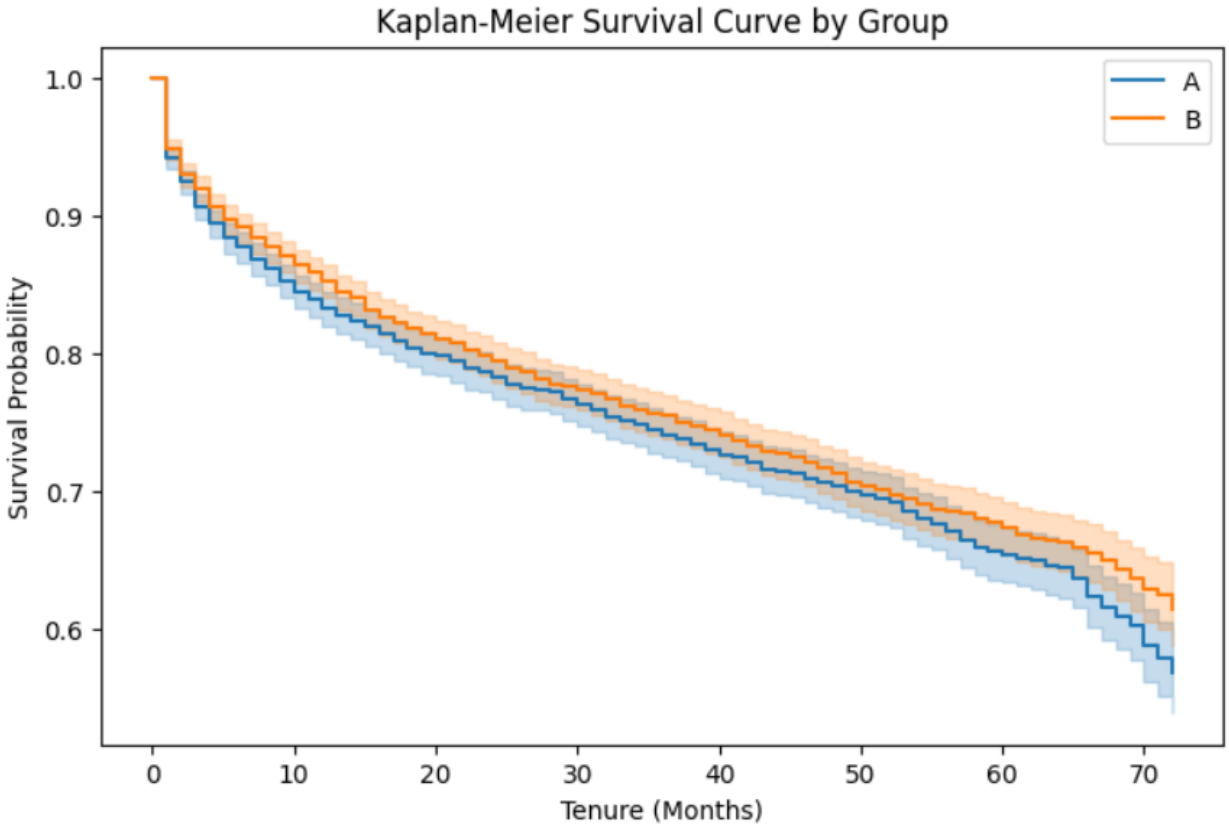
An automated workflow was created to predict customer churn risk in real-time:

- **Model Deployment:**
  - The best-trained model (Logistic Regression) was saved and deployed.
- **Prediction Function:**
  - A function was implemented to predict churn probability based on customer attributes.
- **Process:**
  - Input data is preprocessed (categorical encoding, scaling).
  - The trained model predicts churn probability.
  - High-risk customers can be flagged for retention strategies.

8. A/B Testing for Retention Strategies:

To evaluate the effectiveness of retention strategies, we conducted an A/B test:

- **Experiment Setup:**
  - Customers were randomly assigned to two groups: Control (A) and Treatment (B).
  - Group B received retention incentives (e.g., discounts, special offers).
- **Churn Rate Comparison:**
  - A chi-square test was conducted to compare churn rates between groups.
  - **Result:** The p-value determined whether the strategy significantly impacted churn.
- **Kaplan-Meier Survival Analysis:**
  - Survival curves for both groups were plotted.
  - **Key Insights:**
    - If p-value < 0.05, the retention strategy significantly reduced churn.
    - If p-value ≥ 0.05, the strategy had no significant effect.



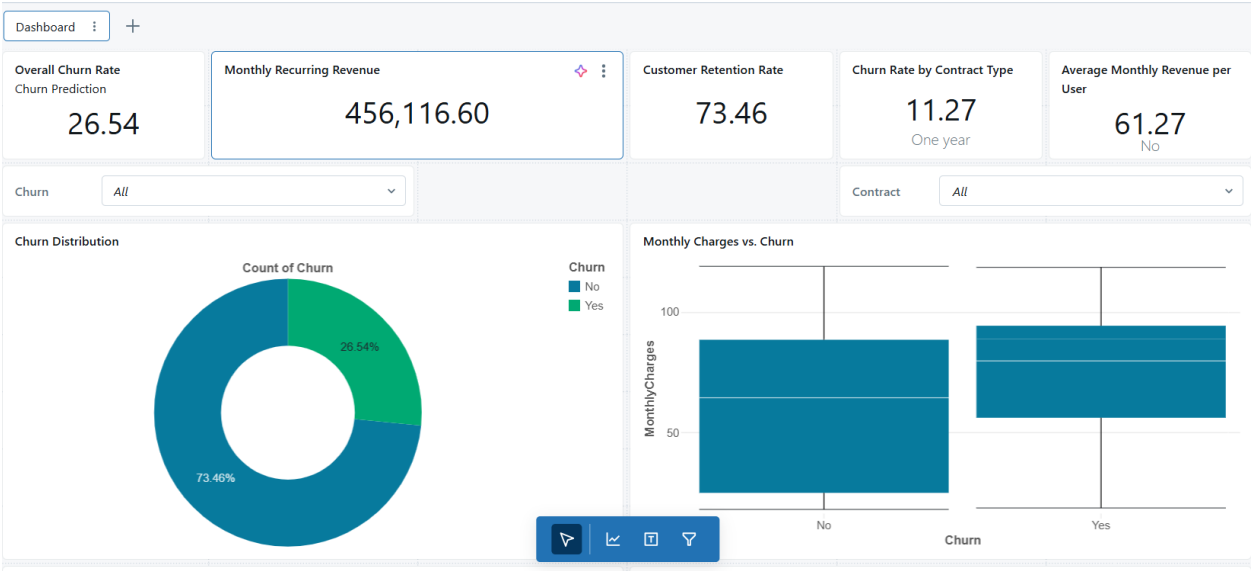
9. Model Drift Monitoring:

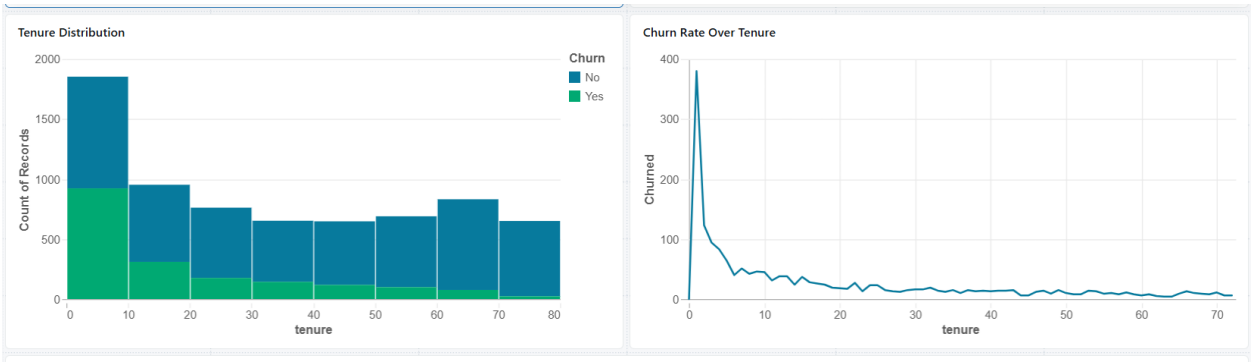
To ensure continuous model performance, we implemented a model drift monitoring system:

- **Dataset Splitting:**
    - The dataset was divided into old (past) and new (recent) data to simulate temporal changes.
- **Drift Detection Methods:**
    - **Numerical Features:** Two-sample Kolmogorov-Smirnov (KS) test
    - **Categorical Features:** Chi-Square test for statistical significance
- **MLflow Tracking:**
    - Drift metrics (p-values) were logged in MLflow for continuous monitoring.
    - The trained logistic regression model was logged into MLflow for version control.
- **Key Insights:**
    - If a p-value is low, it indicates significant drift and the need for model retraining.
    - This system ensures real-time drift detection and mitigation.

10. Dashboard:

- **Overview:** The dashboard provides insights into churn patterns, customer retention, and revenue trends.
- **Key KPIs:** Tracks Overall Churn Rate, Churn by Contract Type, Customer Retention Rate, ARPU, and Monthly Recurring Revenue.
- **Visualizations:** Includes pie charts, bar charts, line graphs, heatmaps, and pivot tables to analyze customer behavior.
- **Key Insights:**
    - Month-to-Month contracts and high monthly charges lead to higher churn.
    - Long-term contracts improve retention and reduce churn.
    - **Filters:** Allows deeper analysis by churn status and contract type.





11. Challenges & Solutions:

Challenge	Solution
Imbalanced Data	Applied oversampling and class weighting
Missing Values	Handled using median imputation and business logic
Total Charges Data Type Issue	Converted object type to numeric and handled missing values by filling NaN values with the median due to skewed distribution
Senior Citizen Column Encoding	Converted 0/1 values into Yes/No for better visualization
Model Overfitting	Regularization and feature selection were used
High Cardinality Categorical Data	Used target encoding and grouping
Model Drift	Implemented MLflow tracking and drift detection

12. Results & Insights:

- **Key Findings:**
    - Customers on month-to-month contracts were more likely to churn.
    - Higher tenure customers had a lower churn rate.
    - Electronic check payment method had a higher churn rate.
- **Best Performing Model:**
    - XGBoost with an AUC-ROC score of 0.85 was the most effective.
- **Next Steps:**
    - Implementing an automated customer scoring system.
    - Applying A/B testing for retention strategies.
    - Deploying models in production with monitoring for drift.

13. Conclusion:

This project successfully demonstrated a data-driven approach to predicting customer churn. The insights gained can help businesses design better customer retention strategies and optimize services to reduce churn rates. By implementing predictive modeling, survival analysis, and A/B testing, companies can take proactive measures to minimize churn and maximize customer lifetime value. And the dashboard further enhances these insights by providing real-time visualization and data-driven decision-making capabilities.