Customer Churn Analysis Report

Project Overview

The primary objective of this project is to analyse customer churn patterns in a telecommunications company. Using exploratory data analysis (EDA), statistical tests, clustering, and predictive modelling, this project identifies factors influencing customer churn and provides actionable insights to improve retention strategies.

Data Overview

The dataset contains information about customer demographics, services subscribed, account information, and churn status. After preprocessing, key features include:

- **Categorical Variables:** Contract, PaymentMethod, InternetService, etc.
- **Numerical Variables:** Tenure, MonthlyCharges, TotalCharges.
- **Target Variable:** Churn (Yes/No).

Data Preprocessing

1. Converted the **TotalCharges** column to numeric format and replaced invalid entries with 0.
2. Filled missing values and encoded categorical variables using Label Encoding.
3. Removed unnecessary columns such as **gender, customerID,** and redundant service columns to create a cleaned dataset.
4. Split data into training and testing sets with an 80/20 ratio, followed by scaling numeric features using **StandardScaler.**

Key Insights from EDA

Distribution of Churn:

- **Count of Churn:**
  - Customers who churned: ~26%.
  - Customers who stayed: ~74%.
- **Churn by Gender:**
  - No significant difference in churn rates between genders.
- **Churn by Senior Citizens:**
  - Higher churn rate among senior citizens compared to non-senior citizens.

Heatmap:

Description: A heatmap showing the correlation matrix for numerical and encoded categorical variables.

Insights:

- Strong negative correlation between tenure and churn, indicating that longer-tenure customers are less likely to churn.
- Moderate positive correlation between monthly charges and churn, suggesting that customers with higher monthly bills are more likely to churn.
- Weak or negligible correlation between other variables such as SeniorCitizen and churn, showing less influence on churn behavior.
- High correlation between TotalCharges and MonthlyCharges due to the cumulative nature of total charges.

Service-Related Insights:

- Customers using month-to-month contracts have the highest churn rate.
- Customers paying via electronic checks are more likely to churn.
- Internet-related services such as online security and tech support influence churn.

Tenure and Monthly Charges:

- Long-term customers (high tenure) have lower churn rates.
- Customers with higher monthly charges are more prone to churn.

Statistical Analysis

Chi-Square Test for Categorical Variables -

- Contract: p-value = 5.86e-258.
- PaymentMethod: p-value = 3.68e-140.
- SeniorCitizen: p-value = 1.0.

ANOVA for Numerical Variables -

- Tenure: p-value = 7.99e-205.
- MonthlyCharges: p-value = 2.71e-60.
- TotalCharges: p-value = 2.13e-63.

Interpretation:

- Contract and PaymentMethod have a significant relationship with churn.
- Tenure, MonthlyCharges, and TotalCharges are also highly correlated with churn.

Predictive Modelling

A Random Forest Classifier was implemented to predict customer churn.

Model Performance:

- **Accuracy:**

  79%. The model correctly predicts customer churn with an overall accuracy of 79%.

Classification Report:

Precision: Indicates the proportion of positive identifications that are actually correct.

- For "No" churn: 82% of predicted "No" cases were correct.
- For "Yes" churn: 64% of predicted "Yes" cases were correct.

Recall: Reflects the ability of the model to identify all true positives.

- For "No" churn: The model correctly identified 90% of actual "No" cases.
- For "Yes" churn: The model identified 47% of actual "Yes" cases.

F1-Score: Balances precision and recall, showing how well the model performs.

- For "No" churn: F1-score is 0.86, indicating high reliability.
- For "Yes" churn: F1-score is 0.54, showing room for improvement in predicting churn.

Support: Indicates the number of actual occurrences of each class.

- "No": 1036 cases.
- "Yes": 373 cases.

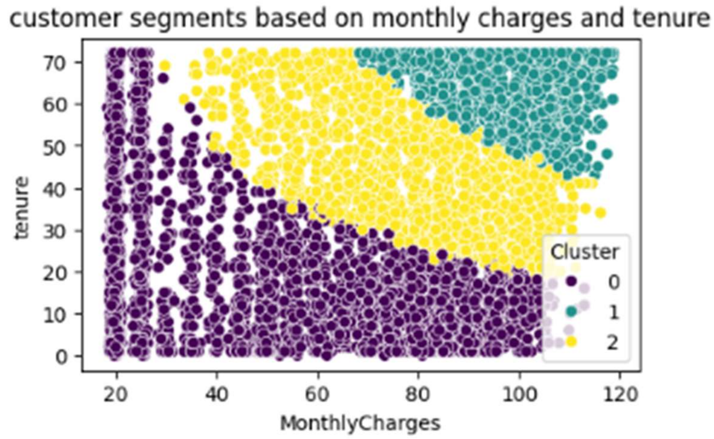
Confusion Matrix:

- Rows represent the actual classes ("No" or "Yes"), and columns represent the predicted classes.
- **True Negatives (936):** Correctly predicted as "No".
- **False Positives (100):** Predicted as "Yes" but actually "No".
- **False Negatives (199):** Predicted as "No" but actually "Yes".
- **True Positives (174):** Correctly predicted as "Yes".

Feature Importance:

- Key predictors include:
- **Tenure:** Longer-tenure customers are less likely to churn.
- **Monthly Charges:** Higher charges correlate with increased churn likelihood.
- **Contract Type:** Month-to-month contracts show the highest churn rates.
- **Payment Method:** Electronic check users are more prone to churn.

Clustering and Segmentation:



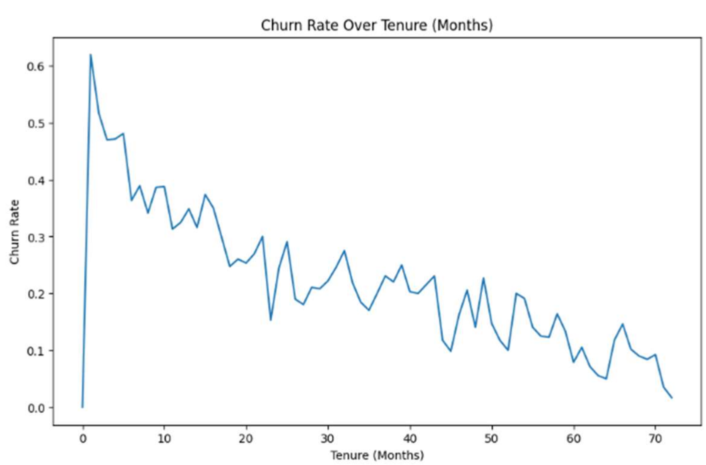
Description: This scatter plot shows the segmentation of customers into three clusters based ontenure, MonthlyCharges, and TotalCharge using K-Means clustering . Each cluster represents a distinct group of customers with similar characteristics, aiding in targeted strategies.

- **Cluster 0:** High-tenure, low-monthly charges customers.
- **Cluster 1:** Low-tenure, high-monthly charges customers (high-risk).
- **Cluster 2:** Mid-range customers.

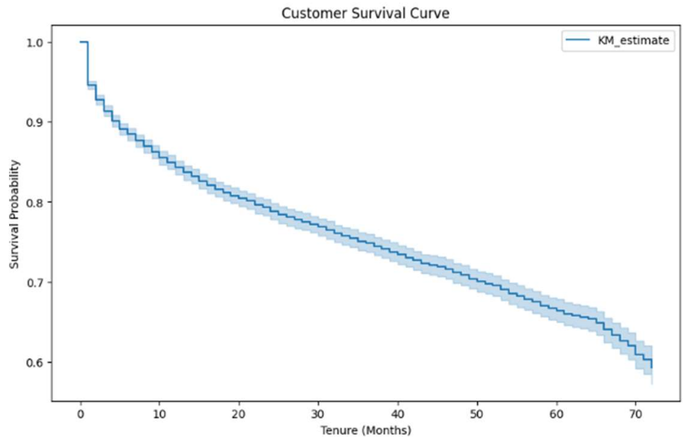
Insights: Cluster 1 represents high-risk customers with low tenure and high monthly charges.

Monthly Churn Rate -

- **Description**: A line plot depicting the average churn rate across different tenure periods.
- **Insights**: Churn rate is highest during the early months of customer tenure, emphasizing the need for retention efforts during this period.



Survival Analysis



Description: A Kaplan-Meier survival curve illustrating the probability of customers staying with the company over time.

Insights: Survival probability drops sharply in the first 10 months, stabilizing afterward.

Retention Strategies

Churn rate is highest during the early months of customer tenure, emphasizing the need for retention efforts during this period. Some Strategies should focus on retaining customers within this critical period

Based on the analysis:

1. **Target High-Risk Clusters:**

   month-to-month contracts and high monthly charges.
2. **Improve Service Bundles:**

   Encourage the adoption of online security and tech support services.
3. **Loyalty Programs:**

   Offer incentives to long-term customers or those at risk of churning.
4. **Payment Options:**

   Promote auto-pay options to reduce churn linked to electronic checks.

Conclusion

This analysis demonstrates the power of data-driven insights in identifying customer churn patterns. By implementing the suggested retention strategies, businesses can reduce churn and improve customer satisfaction. This project is a showcase of my expertise in data preprocessing, exploratory data analysis, statistical testing, machine learning, and survival analysis.