Milestone 2 Report: Advanced Data Analysis and Feature Engineering

1. Data Analysis Report

Section: Advanced Data Analysis

Data Source:
The dataset used in this analysis was imported from the processed results of Milestone 1, specifically from the file Milestone1_result_df.csv.

Objective:
To identify statistically significant relationships between customer features and churn behavior using various statistical methods.

Techniques Applied:

1. **Independent Samples t-Test**
   - Used to compare the mean values of numerical features between churned and non-churned customers.
   - The null hypothesis assumes that there is no significant difference in feature means between the two groups.
   - Features with a p-value < 0.05 were considered statistically significant and potentially influential in predicting churn.
2. **Chi-Squared Test for Independence**
   - Used to evaluate the association between categorical variables and churn.
   - A contingency table was constructed for each categorical feature against the churn flag.
   - Features with statistically significant associations ($p < 0.05$) were considered for inclusion in the model.

Summary of Insights:

- The statistical tests revealed that several numerical and categorical features had strong associations with customer churn.
- These insights were used to guide both feature selection and the creation of new derived features in the subsequent engineering phase.

2. Enhanced Visualizations

Section: Advanced Visualizations and Dashboards

Objective:

To visually explore and communicate churn-related trends, customer behaviors, and the importance of various features using advanced graphical techniques.

Techniques and Tools Used:

1. **Churn Distribution and Segmentation Analysis**
    - Bar plots and pie charts were used to show the distribution of churned vs. non-churned customers.
    - These visualizations offered a clear overview of class imbalance and highlighted churn rates across key demographic segments (e.g., gender, segment, marital status).
2. **Correlation Matrix (Heatmap)**
    - A heatmap was created to visualize the correlation between numerical features.
    - This helped identify multicollinearity and informed which features to retain or drop.
3. **Customer Segmentation Visualization**
    - Customers were segmented based on behavioral and demographic factors, then visualized using grouped bar charts and box plots.
    - These visualizations revealed patterns such as higher churn rates among customers with low interaction frequency or specific account types.
4. **Feature Importance Visualization**
    - A logistic regression model's coefficients were visualized to show the magnitude and direction of influence for each feature.
    - Positive coefficients indicate increased churn likelihood, while negative coefficients suggest customer retention indicators.
5. **Interactive Dashboards (Prototype)**
    - A dashboard prototype was created using plotly and dash to allow dynamic filtering of churned customers based on features like segment, age, and account tenure.
    - Though basic, this dashboard demonstrated the potential for business users to interactively explore churn patterns.

Output Highlights:

- Visualizations clearly illustrated key churn drivers like tenure, complaints, and account types.
- The feature importance chart validated the statistical analysis and informed feature selection.

3. Feature Engineering Summary

Section: Feature Creation and Transformation

Objective:
To enhance the predictive performance of churn models by generating new features and transforming existing ones for better representation of customer behavior and engagement.

Techniques and Steps Applied:

1. **New Feature Creation**
   - **Customer Tenure:**
     A tenure feature was derived based on account duration, indicating how long a customer has been active. This helped capture customer loyalty patterns.
   - **Interaction Frequency:**
     Features representing how frequently customers engaged with services were created (e.g., number of complaints, service usage count).
   - **Engagement Metrics:**
     Additional features such as average balance per active month and normalized income-to-loan ratios were engineered to represent financial activity levels.
2. **Handling Missing Data**
   - Missing values in categorical features were handled using placeholder encoding (e.g., "Unknown" or -1) and later treated during encoding.
   - For numerical features, either median imputation or leaving them for model handling (if tree-based models are used) was applied.
3. **Categorical Encoding**
   - **One-Hot Encoding:**
     Applied to categorical variables such as marital status, customer segment, and education level to avoid ordinal misinterpretation.
   - **Label Encoding:**
     Used for binary features (e.g., gender) where only two categories were present.

4. **Numerical Transformation**
    - **Log Transformation:**
      Applied to skewed numerical features such as Balance and Outstanding Loans to normalize their distributions.
    - **Feature Scaling:**
      StandardScaler was applied to numerical features to ensure uniform scale, especially for models sensitive to feature magnitude (e.g., Logistic Regression).
5. **Outlier Handling**
    - Outliers in features such as Income, Credit Score, and Age were capped using the 1.5×IQR rule to mitigate their influence on the model.

Summary of Feature Engineering Impact:

- Enhanced feature expressiveness led to better model precision and recall.

- Scaling and encoding contributed to model stability and accuracy.