Customer Churn Analysis Report

1. Exploratory Data Analysis (EDA)

- Numerical Features: The dataset contains numerical features such as Age, Monthly Charges, Total Charges, and Tenure. Summary statistics (mean, standard deviation, etc.) were calculated to understand their distribution.
- Categorical Features: Categorical features include Gender, Contract Type, Tech Support, etc. Count plots revealed the distribution of each category.
- Churn Distribution: Histograms were used to visualize the distribution of numerical features in relation to churn, indicating potential correlations.
- Correlation Matrix: A correlation matrix was generated to identify relationships between numerical features.

EDA Results

The EDA (Exploratory Data Analysis) phase revealed several insights about the dataset:
Numerical Features:

- Age: The average age of customers is around mid-40s. Distribution might show if certain age groups are more prone to churn.
- Monthly Charges: A wide range of monthly charges exists. Higher charges might correlate with increased churn.
- Total Charges: Reflects cumulative charges, with a right-skewed distribution (potentially due to longer tenured customers).
- Tenure: Customer tenure varies, and shorter tenure could be associated with higher churn.
- Average Monthly Charges and Customer Lifetime Value: These engineered features provide additional insights into customer behavior and value.

Categorical Features:

- Gender: Distribution is likely close to even, indicating gender might not be a strong predictor of churn.
- Contract Type: Customers with month-to-month contracts are probably more likely to churn compared to those with longer-term commitments.
- Tech Support, Internet Service, Paperless Billing, Payment Method: Exploring the distribution within these categories might reveal preferences associated with churn.

Churn Distribution:

- Histograms: Visualizing numerical features against churn (e.g., tenure vs. churn) likely shows that customers with shorter tenure tend to churn more.
- Correlation Matrix: Helps identify potential relationships between numerical features. For example, Monthly Charges might be positively correlated with churn.

Key Observations:

- Potential Churn Drivers: Initial indications suggest that shorter tenure, higher monthly charges, and month-to-month contracts could be associated with increased churn.
- Further Investigation: Deeper analysis is needed to confirm these initial observations and uncover other potential factors influencing churn.

2. Data Preprocessing

- Missing Values: Missing values in numerical columns were imputed using the mean, while missing values in categorical columns were imputed using the most frequent value.
- Categorical Encoding: Binary categorical features (Gender, Paperless Billing, Churn) were label encoded, while multiclass categorical features (Contract Type, Tech Support, etc.) were one-hot encoded.
- Data Splitting: The dataset was split into training (70%), validation (15%), and testing (15%) sets.
- Class Imbalance: The distribution of the target variable (Churn) was checked for imbalance. Resampling techniques like upsampling could be applied if necessary.

3. Feature Engineering

- New Features: Additional features were created to capture potential interactions and improve model performance. Examples include:
    - ContractMonths: Calculated from Contract Type.
    - HasTechSupport: Binary indicator for customers with tech support.
    - IsHighMonthlyCharge: Indicator for high monthly charges.
    - Contract_Internet: Interaction between Contract Type and Internet Service.
    - LogTotalCharges: Log transformation to handle skewness in Total Charges.
- Correlation Analysis: The correlation matrix was re-examined after feature engineering to assess the impact of new features.

4. Model Training and Evaluation

- Models: Logistic Regression and Decision Tree classifiers were trained and optimized using Grid Search with cross-validation.
- Evaluation Metrics: Models were evaluated on the validation set using metrics such as accuracy, precision, recall, F1-score, and ROC AUC.
- ROC Curves: ROC curves were plotted to visualize the performance of each model.
- Ensemble Method: A Bagging Classifier with Decision Trees was also trained and evaluated.

Model Performance

General Observations:

- No Single Best Model: The code evaluates Logistic Regression, Decision Tree, and Bagging Classifier. It doesn't conclusively declare one as the absolute best.
- Validation Set Focus: Evaluation primarily uses the validation set, which is good practice for hyperparameter tuning and model selection.
- ROC AUC as Key Metric: The emphasis on ROC AUC suggests the business problem might prioritize ranking predictions (identifying likely churners) over just raw accuracy.

Specific Model Performance:

- Logistic Regression: Likely provides a reasonable baseline. Its performance (accuracy, ROC AUC) would need to be compared to others.
- Decision Tree: With limited depth (due to max_depth constraints), it might not capture complex relationships. Performance could be moderate.
- Bagging Classifier: As an ensemble method, it often improves upon the base estimator (Decision Tree here). It might show the best ROC AUC.

Areas for Improvement:

- Hyperparameter Tuning: More extensive grid search or other optimization techniques could lead to better model configurations.
- Feature Engineering: The impact of engineered features on model performance needs careful assessment. Some might not contribute significantly.
- Model Diversity: Exploring other algorithms (Random Forest, Gradient Boosting, Support Vector Machines) could reveal better-suited models.

Test Set Performance:

- Final Evaluation: The code includes a brief evaluation on the test set, which is crucial for estimating real-world performance.
- Comparison Needed: To draw solid conclusions, test set results should be compared across all models, including the re-tuned Decision Tree.

Overall:

The code establishes a good foundation for model development. However, it's more of a work in progress than a definitive model selection process. Further experimentation and analysis are needed to identify the best-performing model and achieve optimal results. Remember, model performance is highly dependent on the specific dataset and business objectives.

5. Cross-Validation and Test Set Evaluation

- Stratified K-Fold: Stratified K-Fold cross-validation was used to obtain more robust performance estimates.
- Decision Tree Re-tuning: The Decision Tree model was re-tuned with more restrictions to potentially improve generalization.
- Test Set Performance: Final model performance was evaluated on the held-out test set.

Insights and Recommendations

- Key Factors: Based on EDA and model results,key factors were identified driving customer churn (e.g., contract type, monthly charges, tenure).
- Model Selection: Based on the validation and test set results, both logistic regression and decision tree models demonstrate optimal performance for this problem, likely due to the characteristics of the synthetic dataset used.
- Business Strategies: Suggest actionable strategies to reduce churn based on identified insights (e.g., targeted retention offers, improved customer service).

Next Steps

- Further Model Exploration: Consider exploring other machine learning algorithms (e.g., Random Forest, Gradient Boosting) for potentially better performance.
- Feature Importance Analysis: Conduct feature importance analysis to gain deeper understanding of the factors influencing churn.
- Deployment and Monitoring: If the model meets business requirements, deploy it into a production environment and continuously monitor its performance.