Customer Churn Prediction Project Report

1. Introduction

Customer churn, the loss of customers, is a critical concern for businesses as acquiring new customers is often more expensive than retaining existing ones. This project aims to develop a supervised machine learning model to predict customer churn based on provided telecommunications customer data. The objective is to identify customers likely to churn, enabling the company to proactively implement retention strategies.

2. Data Description and EDA Findings

The dataset contains information about telecommunications customers, including their usage patterns, account details, and whether they churned. Initial inspection revealed 2666 entries and 20 columns with no missing values. The columns include numerical, categorical (object), and boolean data types.

Exploratory Data Analysis (EDA) highlighted the following key findings:

- **Target Variable Distribution**: The dataset is imbalanced, with significantly more non-churned customers (2278) than churned customers (388). This imbalance needs to be considered during model evaluation.

- **Numerical Feature Distributions**: Most numerical features show typical distributions. Features like 'Number vmail messages' and 'Customer service calls' have a notable number of zero values.

- **Categorical Feature Relationships**: Customers with an international plan ('International plan\_Yes') have a considerably higher churn rate compared to those without an international plan. Customers with a voice mail plan ('Voice mail plan\_Yes') have a lower churn rate compared to those without a voice mail plan.

- **Numerical Feature Relationships**: Features like 'Total day charge', 'Total day minutes', and 'Customer service calls' show a positive correlation with churn. Higher values in these features are associated with increased churn probability.

- **Correlation Matrix**: Strong positive correlations exist between call minutes and their corresponding charges (e.g., 'Total day minutes' and 'Total day charge'), which is expected.

1. Data Cleaning, Preprocessing, and Feature Engineering

The following steps were taken to prepare the data for modeling:

- **Categorical Encoding**: Categorical features ('State', 'International plan', 'Voice mail plan') were one-hot encoded using pandas `get\_dummies` to convert them into a numerical format. The `drop\_first=True` argument was used to avoid multicollinearity.

- **Data Type Conversion**: The boolean 'Churn' column and the resulting boolean columns from one-hot encoding were converted to integer type (0s and 1s).

- **Feature Engineering**: New features were created to potentially improve model performance:

    -   '**Total charge**': Sum of charges across all time periods (day, eve, night, intl).

    -   '**Total minutes**': Sum of minutes across all time periods.

    -   '**Average charge per minute**': Calculated as Total charge / Total minutes, handling potential division by zero.

    -   '**Intl plan and high intl calls**': A binary feature indicating if a customer has an international plan and makes a high number of international calls ( > 3).

    -   '**VM plan and high VM messages**': A binary feature indicating if a customer has a voice mail plan and a high number of voice mail messages (> 10).

4. Model Selection Process

Several supervised machine learning models were explored for this binary classification task:

-   Logistic Regression (as a baseline)

-   Decision Tree

-   Random Forest

-   Gradient Boosting

The models were trained on the preprocessed and feature-engineered data. Their performance was evaluated on a held-out test set (20% of the data) using the following metrics:

-   **Accuracy**: Overall correctness of predictions.

-   **Precision**: The proportion of positive predictions that were actually positive.

-   **Recall**: The proportion of actual positive cases that were correctly identified.

-   **F1-score**: The harmonic mean of precision and recall, useful for imbalanced datasets.

-   **AUC (Area Under the ROC Curve)**: Measures the model's ability to distinguish between the positive and negative classes.

Initial evaluation showed that the Gradient Boosting model achieved the best F1-score among the untuned models.

5. Best Model Performance Analysis

The Gradient Boosting model was selected for hyperparameter tuning to optimize its performance. GridSearchCV with 5-fold cross-validation was used to find the best combination of hyperparameters, optimizing for the F1-score.

The best hyperparameters found were: {best\_params}

The tuned Gradient Boosting model's performance on the test set is as follows:

-   **Accuracy**: {accuracy:.4f}

-   **Precision**: {precision:.4f}

-   **Recall**: {recall:.4f}

-   **F1-score**: {f1:.4f}

-   **AUC**: {auc:.4f}

The tuned Gradient Boosting model significantly outperforms the baseline Logistic Regression model and the untuned tree-based models, achieving high accuracy and an excellent F1-score, precision, recall, and AUC.

6. Feature Importance Interpretation

Interpreting the feature importances from the tuned Gradient Boosting model provides insights into the key drivers of customer churn:

{feature\_importances\_string}

The most important features are:

- **Total charge / Total day charge / Total day minutes**: These features related to overall and daily usage and cost are strong predictors of churn. Higher usage and cost are associated with increased churn risk.

- **Customer service calls**: A high number of calls to customer service indicates potential issues or dissatisfaction, making this a significant churn indicator.

- **International plan\_Yes**: Customers with an international plan are more likely to churn, suggesting potential issues or costs related to international services.

- **Total intl calls / Total intl charge**: Related to international usage, these also contribute to churn prediction.

- **Intl plan and high intl calls**: This engineered feature highlights that having an international plan and high international call volume is a particularly strong indicator of churn.

- **Number vmail messages / VM plan and high VM messages**: While 'Voice mail plan\_Yes' initially showed a lower churn rate, the number of voice mail messages and the combined 'VM plan and high VM messages' feature show some importance, possibly indicating specific usage patterns within the voice mail plan.

Less important features include state-specific dummy variables and some account-related features.

1. Insights gained through Analysis

- The dataset contains 2666 entries and 20 columns, including numerical, categorical, and boolean data types, with no missing values.

- The target variable 'Churn' is imbalanced, with 2278 non-churned customers and 388 churned customers.

- Customers with an international plan have a significantly higher churn rate, while those with a voice mail plan have a lower churn rate.

- Higher 'Total day charge', 'Total day minutes', and 'Customer service calls' are associated with increased churn probability.

- Strong correlations exist between call minutes and their corresponding charges (e.g., 'Total day minutes' and 'Total day charge').

- Feature engineering created new features like 'Total charge', 'Total minutes', 'Average charge per minute', 'Intl plan and high intl calls', and 'VM plan and high VM messages'.

- The data was successfully split into training (2132 samples) and testing (534 samples) sets with an 80/20 ratio.

- The baseline Logistic Regression model achieved an F1-score of 0.2931 and AUC of 0.7756 on the test set.

- Initial evaluation of Decision Tree, Random Forest, and Gradient Boosting models showed Gradient Boosting performing best with an F1-score of 0.9041.

- Hyperparameter tuning of the Gradient Boosting model using GridSearchCV improved its performance, resulting in an F1-score of 0.9103, Accuracy of 0.9757, Precision of 1.0000, Recall of 0.8354, and AUC of 0.9921 on the test set.

- Feature importance analysis of the tuned Gradient Boosting model identified 'Total charge', 'Customer service calls', and 'International plan\_Yes' as the most important features in predicting churn.

8. Conclusions and Recommendations

The tuned Gradient Boosting model is highly effective in predicting customer churn with strong performance metrics. The feature importance analysis reveals that customer usage patterns (particularly day and international usage), customer service interactions, and having an international plan are the most significant factors influencing churn.

Key Findings (Why Customers Churn):

- **High Usage/Cost**: Customers who use their phones a lot, especially during the day, and have high bills are more likely to churn.

- **Customer Service Issues**: Customers who call customer service frequently are at a higher risk of leaving, suggesting they might be having problems.

- **International Plan**: Customers with an international plan are more likely to churn.

Based on these findings, the following recommendations can be made:

- **Proactive Outreach**: Identify customers with high daily usage, high customer service call volume, or those with international plans and high international call activity. Reach out to these customers with targeted offers or support to address potential issues before they churn.

- **Analyze Customer Service Interactions**: Investigate the reasons behind high customer service call volume to identify systemic issues or common pain points that lead to dissatisfaction and churn.

- **Review International Plan Offerings**: Examine the international plan's pricing, features, and customer support to understand why it's a significant churn driver and explore ways to improve customer satisfaction in this segment.

- **Monitor Usage Trends**: Continuously monitor customer usage patterns, especially spikes in daily usage or changes in international call behavior, as early warning signs of potential churn.

- **Address Class Imbalance**: While the chosen metrics (F1-score, AUC) are more robust to class imbalance, further techniques like oversampling the minority class or using different cost-sensitive learning algorithms could be explored to potentially improve the model's ability to identify churners.

- **Collect More Data**: Gather more detailed information on customer interactions, reasons for customer service calls, and feedback to gain deeper insights into churn drivers.

**In Simple Terms**: Our best prediction model is very good at spotting customers who might leave. The main reasons customers churn seem to be related to high usage and costs, having problems that cause them to call customer service, and issues with the international plan. By focusing on these areas, we can work on keeping our customers happy.

This project provides a solid foundation for a churn prediction system. Implementing the recommendations based on the model's insights can help the telecommunications company reduce churn and improve customer retention.