FAR EASTERN UNIVERSITY

Predicting Customer Churn

with Regression-Based and Tree-Based Methods

Jordan Dacanay

Zion John Yousef Ramilo

Institute of Arts and Sciences – Department of Mathematics

DSC1107 – Data Mining and Wrangling

Professor Frederick Gella

May 20, 2025

Introduction

This report presents an exploratory data analysis (EDA) of a customer churn dataset, specifically a subset containing 80% of the full data, referred to as churn_80. The dataset contains information on customer behavior, demographics, and usage metrics from a telecommunications company.

The objective of this analysis is to gain insights into the underlying patterns and trends that may influence customer churn—defined as whether a customer has discontinued service. By focusing on numerical variables, the report aims to explore how different factors such as call duration, charges, and total minutes are distributed across the customer base. Hence, this project aims to develop a machine learning model that accurately predicts whether a customer will churn based on usage metrics, account features, and customer service interaction data. The study also evaluates different classification models and assesses their performance on both training and test datasets.

Methodology

Data Description

This study utilized a dataset from a telecommunications company called Orange Telecom. The dataset contains cleaned customer activity data (features) along with a churn label indicating whether a customer canceled their subscription. Two datasets are provided: churn-80 and churn-20. The variables are as listed below in the table:

Variable Name	Type	Variable Name	Type
State	Categorical	Total eve calls	Numerical
Account length	Numerical	Total eve charge	Numerical

Area code	Numerical	Total night minutes	Numerical
International plan	Categorical	Total night calls	Numerical
Voice mail plan	Categorical	Total night charge	Numerical
Number vmail messages	Numerical	Total intl minutes	Numerical
Total day minutes	Numerical	Total intl calls	Numerical
Total day calls	Numerical	Total intl charge	Numerical
Total day charge	Numerical	Customer service calls	Numerical
Total eve minutes	Numerical	Churn	Categorical

Table 1. Data Variables and Type

These datasets come from the same batch but are split in an 80/20 ratio. The larger set, churn-80, is used for training and cross-validation, while the smaller set, churn-20, is reserved for final testing and evaluating model performance.

Exploratory Data Analysis

To further examine and understand the variables, we will visualize their distributions using univariate and bivariate analysis.

1. Univariate Analysis

a. Numerical Variables

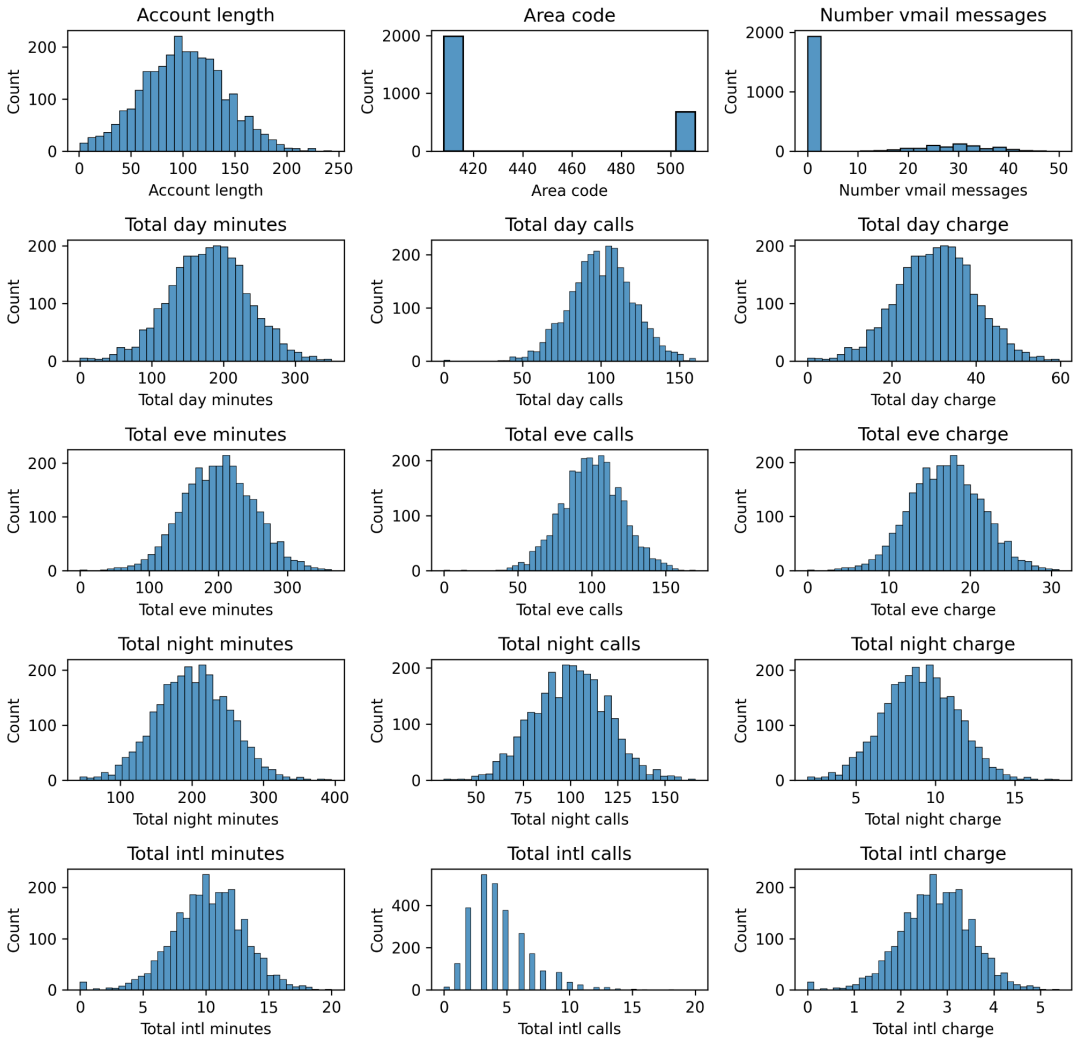

Figure 1. Univariate Distributions for Numerical Variables

As shown in the figure above, the data does not exhibit severe outliers or anomalies. Most variables demonstrate distributional properties that are favorable for modeling. Meanwhile, skewed variables like voicemail usage and international calls may benefit from transformation or stratification to enhance model performance. Overall, the distributions suggest that the data is well-behaved and suitable for statistical modeling and machine learning applications.

b. Categorical Variables

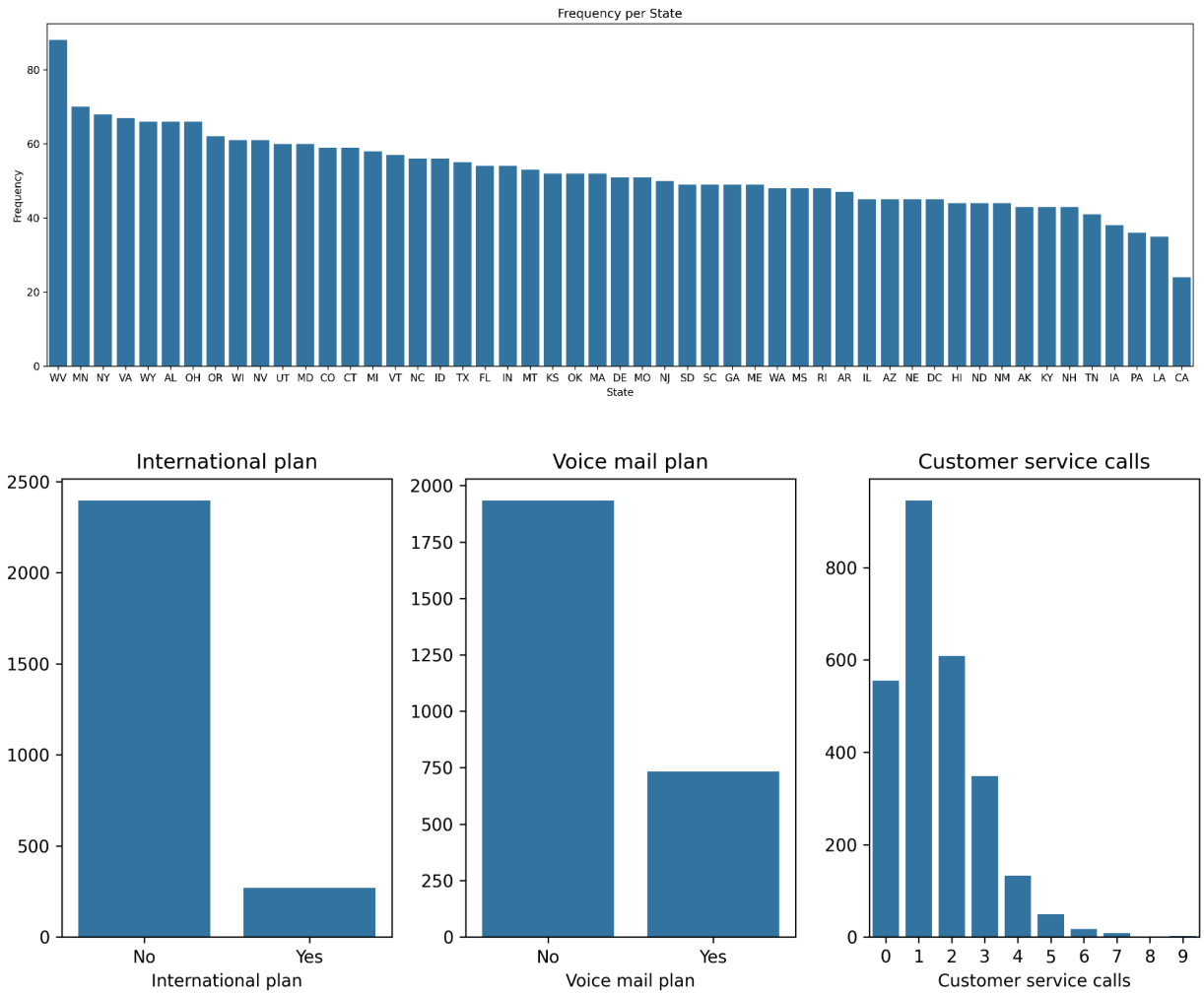

Figure 2. Univariate Distributions for Categorical Variables

The dataset exhibits an uneven distribution of customers across U.S. states, with West Virginia, Minnesota, and New York having the highest representation. This geographic imbalance may introduce regional bias in model training. Among categorical features: (1) Most customers do not subscribe to international or voicemail plans, suggesting these services are niche, and (2) Customer service calls are right-skewed, with the majority of customers making fewer than three calls. A small segment of users with high call volumes may indicate service dissatisfaction and could be predictive of churn.

2. Bivariate Analysis

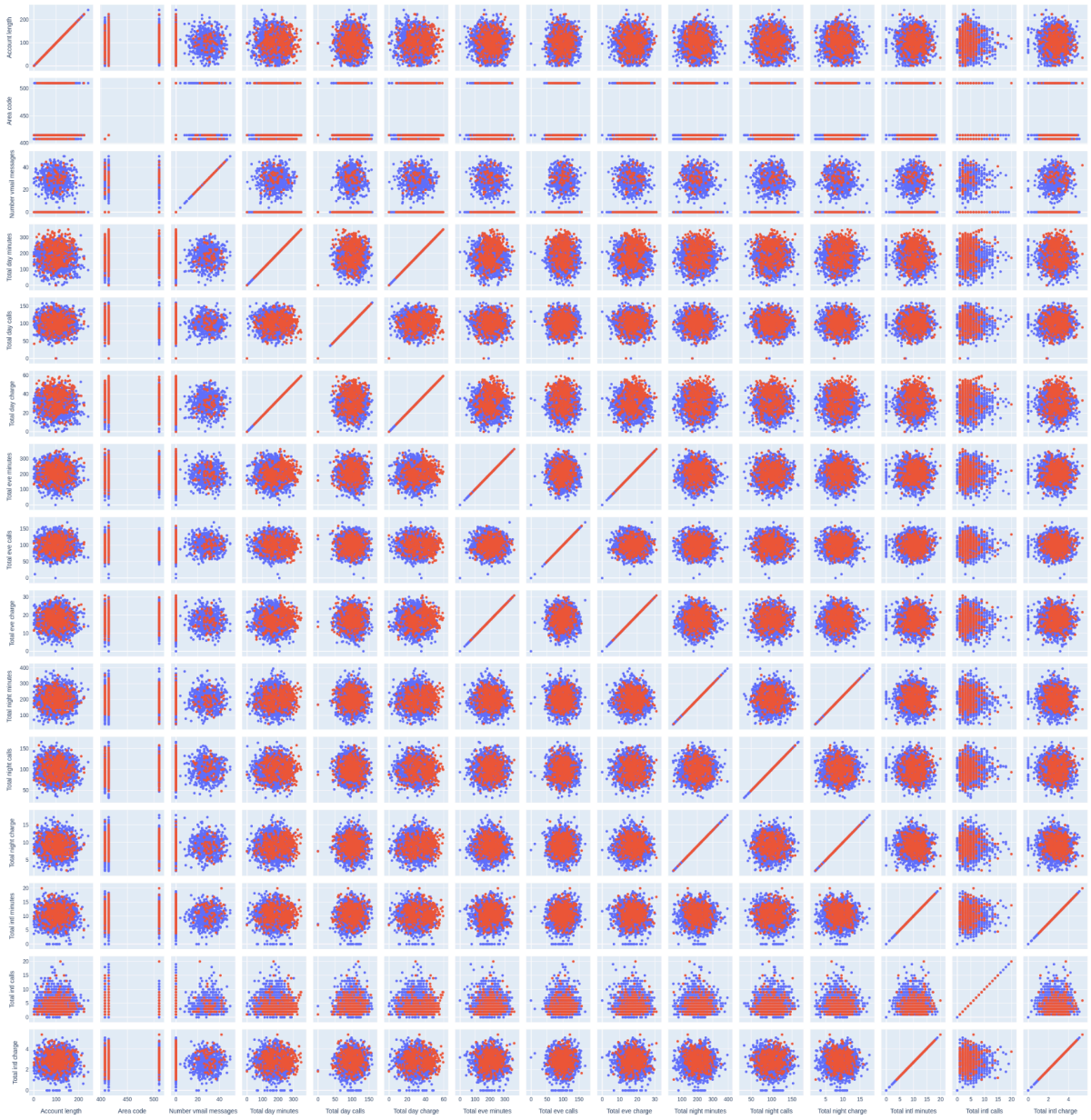

Figure 3. Scatter Plot of the Bivariate Distributions

The scatter matrix confirms the presence of highly correlated redundant variables, limited linear separability between churn classes, and the necessity for nonlinear classification models and feature engineering to improve churn prediction. This visualization serves as a foundational reference for feature selection and highlights the limitations of relying on bivariate relationships alone for classification.

Data Preparation

All data preprocessing, modeling, and evaluation procedures were conducted using the programming language Python, on which some libraries were utilized. In this study, the dataset underwent a structured series of preprocessing steps designed to transform raw inputs into a clean, model-ready format. These steps addressed the diverse data types, handled categorical encodings, ensured numerical consistency, and prepared the data for effective model learning.

1. **Data Type Adjustment**

   Several columns were explicitly cast to their appropriate types:

   - Area code and Customer service calls were converted to categorical types (object) to reflect their nominal nature.
   - The Churn column was treated as a binary label with values True or False.

2. **Feature Grouping**

   To streamline preprocessing, columns were grouped by their data characteristics:

   - **Binary features**: Columns such as International plan and Voice mail plan contain Yes/No responses. These were relabeled to numeric values using a custom FunctionTransformer, where 'Yes' became 1 and 'No' became 0.
   - **Categorical features**: Nominal variables such as State, Area code, and Customer service calls were encoded using one-hot encoding to create dummy variables. This avoided imposing ordinal relationships on non-numeric attributes.
   - **Numerical features**: All continuous variables, such as Total day minutes, Total eve calls, and Total intl charge, were passed through initially without transformation.

3. **ColumnTransformer Pipeline**

   A ColumnTransformer was implemented to apply the appropriate transformation to each group of features in parallel:

   - Binary columns passed through a relabeling pipeline.
   - Categorical columns were expanded into multiple columns using one-hot encoding.
   - Numerical columns were passed as-is in the first phase.

   After transformation, the processed data was compiled into a new data frame with updated feature names, especially for encoded columns.

4. **Target Variable Encoding**

   The Churn column, originally Boolean (True or False), was also relabeled to 1 (churned) and 0 (not churned). This allowed for compatibility with classification models that require numeric targets.

5. **Feature Scaling**

   In a secondary preprocessing stage applied during model training:

   - Numerical features were transformed using the Yeo-Johnson method to handle skewness and improve model performance.
   - Standardization was applied to ensure that features had zero mean and unit variance, which is essential for models sensitive to feature scaling, such as logistic regression and ridge regression.

6. **Resampling Strategy**

   The original dataset exhibited class imbalance, with significantly more non-churned

customers (2278 observations) than churned ones (388 observations). To correct this, a two-step resampling pipeline was implemented:

- Random Under-Sampling reduced the number of majority class samples to limit bias.
- SMOTE (Synthetic Minority Over-sampling Technique) synthetically generated new minority class samples to achieve class balance.

This combined approach ensured that the training data became balanced, with both churn values having a total of 485 observations.

Model Training and Selection

Following data preprocessing and resampling, a diverse set of classification models was trained to predict customer churn. The goal was to evaluate different algorithmic approaches and identify the best model suited for accurately distinguishing churned from non-churned customers.

1. **Model Selection Rationale**

   A combination of tree-based, linear, and ensemble models was selected to cover both interpretable and high-performance algorithms:

   - **Decision Tree Classifier**

     A simple, interpretable model that captures decision rules based on feature thresholds. Useful for understanding key churn drivers.

   - **Random Forest Classifier**

     An ensemble of decision trees that reduces overfitting and improves generalization through bootstrap aggregation.

- **Gradient Boosting Classifier**

  A sequential ensemble method that builds additive models to correct previous errors. Known for strong performance in structured data tasks.

- **Logistic Regression**

  A linear model that estimates the probability of churn. Hyperparameters were tuned using GridSearchCV for optimal performance.

2. **Hyperparameter Tuning**

   For Logistic Regression, a Grid Search with 5-fold cross-validation was applied. The following parameters were explored:

   - **Penalty type**: l1, l2, elasticnet, or None
   - **Solver**: including lbfgs, liblinear, sag, saga, and newton-cholesky
   - **Regularization strength C**: values such as 0.01, 0.1, and 1
   - **Maximum iterations**: 200 to 1000

   This exhaustive search aimed to identify the best combination for minimizing error on the training set while maintaining generalization.

Model Evaluation

To assess the effectiveness of the trained models, each was evaluated using standard classification performance metrics. The evaluation process was conducted separately for both the training set and the unseen test set to ensure fairness and to examine generalization ability.

1. **Evaluation Metrics**

   The following metrics were used to comprehensively evaluate model performance:

   - **Accuracy**: Measures the proportion of correct predictions out of total predictions.

- **Precision**: Indicates the proportion of positive predictions (Churn = True) that were correct.
- **Recall (Sensitivity)**: Reflects the model's ability to correctly identify actual churned customers.
- **F1-Score**: The harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives.
- **ROC AUC (Receiver Operating Characteristic – Area Under the Curve)**: A threshold-independent metric that measures the model's ability to distinguish between the churned and non-churned classes across all classification thresholds.

2. **Evaluation on Resampled Training Set**

   Each model was trained on the balanced dataset produced through under-sampling and SMOTE. Predictions were made on this same resampled set to measure internal performance. This step helped confirm whether the models could learn the underlying patterns in the presence of balanced class distributions.

3. **Evaluation on the Original Test Set**

   The models were then evaluated on the original 20% holdout dataset, which retained its natural class imbalance. This assessed how well each model generalized to real-world distributions. A feature alignment step was performed to ensure that one-hot encoded variables were consistent between the training and test sets.

Results and Discussion

Resampled Training Set

Model Name	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest	1.000000	1.000000	1.000000	1.000000	1.000000
Gradient Boosting	0.94433	0.979955	0.907216	0.942184	0.94433
Logistic Regression	0.847423	0.854737	0.837113	0.845833	0.847423
Grid Search	0.848454	0.861111	0.830928	0.84575	0.848454

Table 2. Model Performances on Resampled Training Set

The table summarizes the performance of five classification models evaluated on the resampled training set, where class imbalance was addressed using under-sampling and SMOTE. Metrics include Accuracy, Precision, Recall, F1-Score, and ROC AUC.

1. **Decision Tree and Random Forest**

   Both models achieved perfect scores (1.000) across all evaluation metrics. While this indicates that they fit the resampled training data extremely well, such perfect performance is also a potential indicator of overfitting. These models may have memorized the training data rather than learned generalizable patterns, which can lead to a drop in performance when applied to unseen test data.

2. **Gradient Boosting**

   This model demonstrates a strong balance between precision and recall, with high ROC AUC, indicating effective discrimination between churned and non-churned customers while maintaining robustness against overfitting.

3. **Logistic Regression**

   This model offered competitive performance while remaining highly interpretable. Its slightly lower recall (0.84) suggests a modest trade-off in sensitivity compared to

tree-based models, but its generalization potential is likely stronger due to its simplicity and regularization.

4. **Grid-Search-Tuned Logistic Model**

   The grid-tuned logistic regression model achieved metrics almost identical to the untuned version, with slight improvements in precision (0.86) and F1-score (0.85). This consistency suggests that the logistic model is relatively stable and benefits marginally from hyperparameter tuning.

Overall, while the Decision Tree and Random Forest models appear flawless on the resampled training set, their perfect scores raise concerns about overfitting. Gradient Boosting provided a more realistic and generalizable performance profile. The logistic regression models, though slightly less performant in raw metrics, present a strong balance of performance and interpretability, especially suitable for practical deployment and stakeholder communication.

The ROC AUC values confirm that all models, particularly Gradient Boosting and Logistic Regression, are capable of discriminating between churn and non-churn across thresholds, reinforcing their reliability for churn prediction tasks.

Original Test Set

Model Name	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.797601	0.398990	0.831579	0.539249	0.811768
Random Forest	0.866567	0.520000	0.821053	0.636735	0.847589
Gradient Boosting	0.878561	0.546667	0.863158	0.669388	0.872138
Logistic Regression	0.799100	0.397906	0.800000	0.531469	0.799476
Grid Search	0.806597	0.406593	0.778947	0.534296	0.795068

Table 3. Model Performances on Original Test Set

This table presents the evaluation metrics of five classification models on the original test set, which retains its real-world class imbalance. Metrics include Accuracy, Precision, Recall, F1-Score, and ROC AUC, allowing a comprehensive comparison of model generalization.

The best-performing model on the original test set is Gradient Boosting, excelling in all key metrics, particularly Recall, F1-Score, and ROC AUC. Random Forest is a strong runner-up with a good balance of accuracy and class separation. While Decision Tree shows high recall, its precision is low, suggesting overfitting. The Logistic Regression models, despite lower scores, remain useful for interpretable insights and quick deployment.

Feature Importance

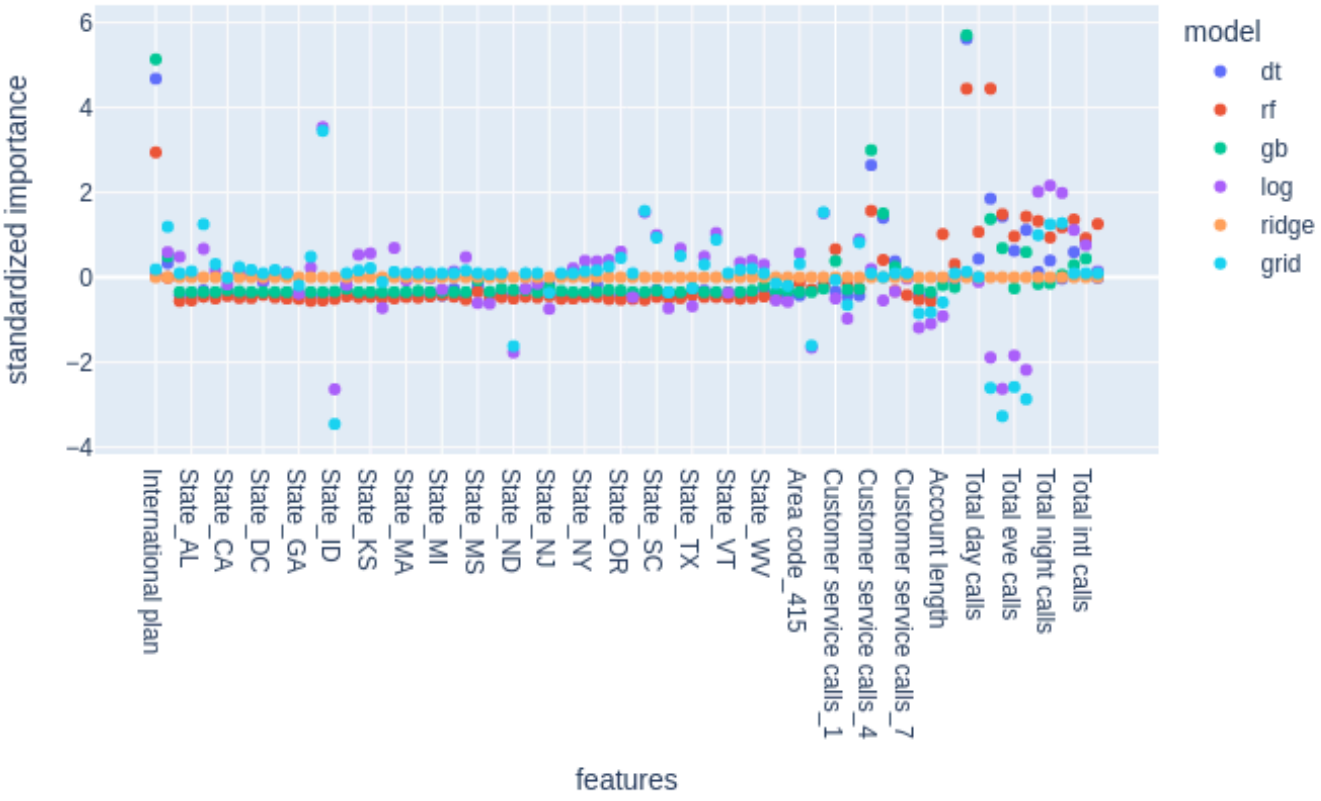

Figure 4. Feature Importance Visualization

The figure presents a comparative analysis of standardized feature importances across six models: Decision Tree (dt), Random Forest (rf), Gradient Boosting (gb), Logistic Regression (log), Ridge Regression (ridge), and Grid Search-Tuned Logistic Regression (grid). Each point represents a standardized importance score assigned to a specific feature by a given model.

The analysis reveals that customer behavior metrics, particularly service usage and number of customer service calls, are the most influential features in predicting churn. While feature importance magnitudes vary by model type, the general pattern is consistent, validating the relevance of these predictors. Linear models provide stable and interpretable weights, while ensemble models highlight sharp decision splits for select features.

This plot underscores the utility of comparing feature importance across multiple models to gain deeper insight into what drives predictive performance and how different algorithms interpret feature relevance.

This study investigated customer churn prediction using a structured machine learning pipeline, combining exploratory data analysis, systematic preprocessing, resampling for class imbalance, model training, and performance evaluation. Multiple classification algorithms were applied and assessed on both a balanced training set and a real-world test set to identify patterns indicative of customer attrition.

Conclusion

This study investigated customer churn prediction using a structured machine learning pipeline, combining exploratory data analysis, systematic preprocessing, resampling for class imbalance, model training, and performance evaluation. Multiple classification algorithms were applied and assessed on both a balanced training set and a test set to identify patterns based on customer attrition.

The exploratory analysis revealed that features like service usage—especially total minutes, call charges, and customer service interactions—played an important role in predicting

churn. The preprocessing pipeline ensured consistent transformation across categorical and numerical data types, enabling the models to learn from a clean and standardized dataset. Moreover, addressing class imbalance through a combination of Random Under-Sampling and SMOTE proved essential for improving model sensitivity, particularly for detecting churned customers.

Among the models evaluated, Gradient Boosting consistently delivered superior performance, achieving the highest recall, F1-score, and ROC AUC on the original test set. Random Forest also performed well, while Logistic Regression provided a great balance of interpretability and predictive capability, especially after hyperparameter tuning through GridSearchCV. On the other hand, models such as Decision Trees, while achieving perfect performance on the resampled training data, demonstrated signs of overfitting when evaluated on unseen data.

Finally, the feature importance analysis revealed that behavioral indicators, particularly customer service interactions and usage intensity, are the most influential drivers of churn. By evaluating multiple models and comparing their learned patterns, this project demonstrates not only the feasibility but also the practical utility of predictive modeling in customer retention strategies.

In the end, this study emphasizes how effective ensemble models are for predicting churn and highlights the importance of careful data preparation, balancing, and evaluation in creating meaningful, real-world solutions for the telecom industry.