Customer Churn Prediction for a Telecom Provider

1. Introduction

In subscription-based businesses such as telecommunications, customer churn—the event of a customer terminating their service is a critical driver of revenue loss. Acquiring new customers is typically more expensive than retaining existing ones, so being able to identify customers at high risk of churn enables the business to target retention campaigns more efficiently. The objective of this project is to build and evaluate predictive models that estimate the probability that a telecom customer will churn in the near future.

Using the Telco Customer Churn dataset, we develop several classification models and compare their performance using standard evaluation metrics. Beyond predictive accuracy, we also analyze which customer characteristics are most strongly associated with churn and translate these insights into actionable business recommendations.

2. Data Description and Preprocessing

The dataset contains records for 7,043 customers with 21 variables describing demographics, account information, services used, and billing details, along with a binary churn indicator. Examples of variables include gender, whether the customer is a senior citizen, the type of contract (month-to-month, one year, or two year), tenure in months, monthly charges, and total charges to date. The target variable Churn takes the values 'Yes' or 'No'.

Initial inspection showed that most variables are stored as categorical (object) fields, while tenure and charges are numeric. The TotalCharges variable was read as text because of embedded spaces for customers with very short tenure. We converted TotalCharges to numeric, treated spaces as missing values, and dropped the 11 rows with missing TotalCharges. This left 7,032 customers and 21 columns for analysis.

Figure 1 shows the overall churn distribution, confirming that the dataset is moderately imbalanced.

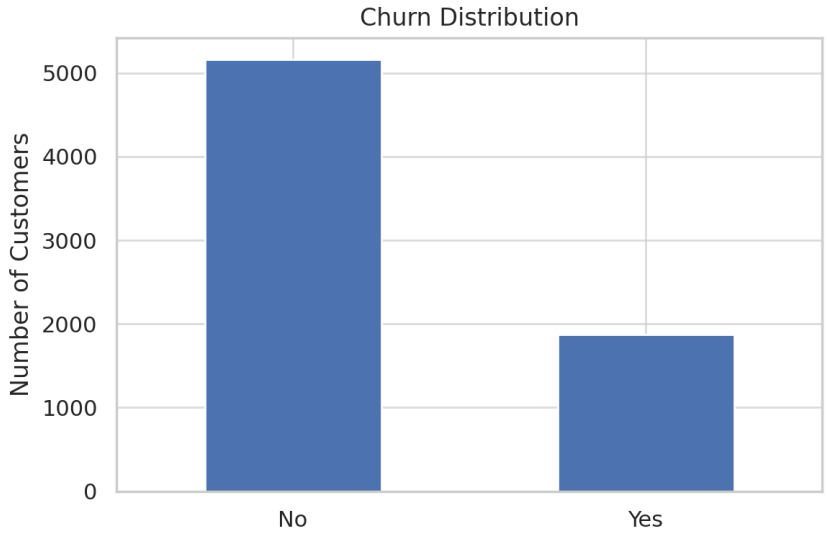


Figure 1: Overall churn distribution (count of churned vs. non-churned customers).

Roughly 26% of customers have churned, while about 74% have stayed with the company. The class imbalance is not extreme but is substantial enough that evaluation measures beyond simple accuracy, such as precision, recall, F1-score, and ROC-AUC, are necessary to properly judge model performance.

To make churn easier to model, we created a numeric indicator ChurnFlag equal to 1 for churned customers and 0 otherwise. We also constructed a tenure_group variable that bins tenure into five groups: 0–12, 13–24, 25–48, 49–60, and 60+ months. This derived variable helps summarize how churn behavior differs across stages of the customer life cycle.

3. Exploratory Data Analysis

Exploratory data analysis (EDA) was performed to understand patterns in churn as a function of contract type, tenure, and pricing. These relationships help motivate the modeling approach and provide intuition for the feature importance results later in the report.

Contract type is one of the clearest drivers of churn. Figure 2 shows churn counts by contract category.

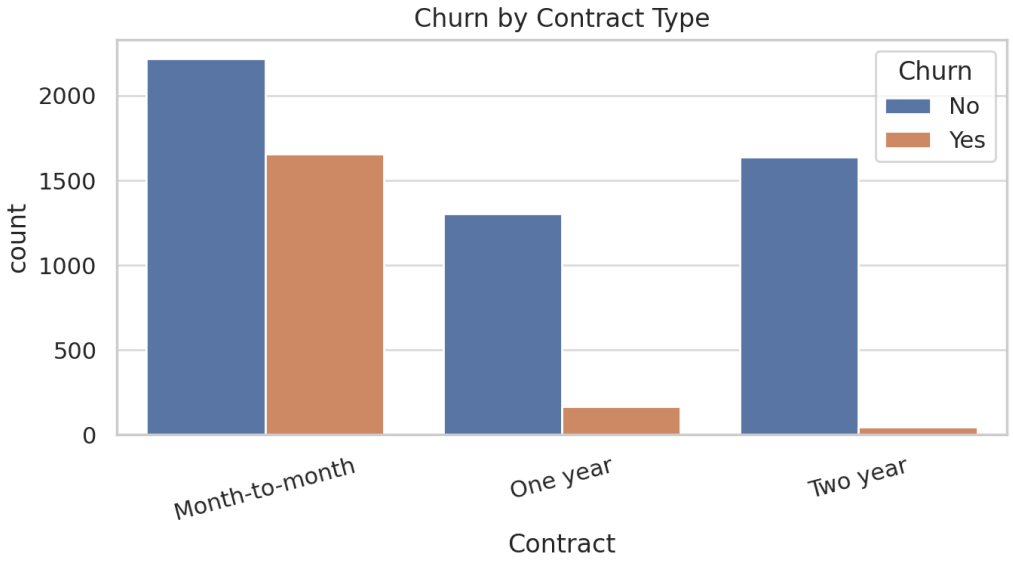


Figure 2: Churn by contract type.

Customers on month-to-month contracts exhibit substantially higher churn than those on one-year or two-year contracts. For one-year contracts, churn volume is much lower, and for two-year contracts churn is rare. This pattern is consistent with business intuition: long-term contracts lock customers in and typically come with discounts, while month-to-month customers can leave at any time with minimal switching cost.

Tenure shows a similar story from a time-based perspective. Figure 3 presents the distribution of customer tenure, separated by churn status.

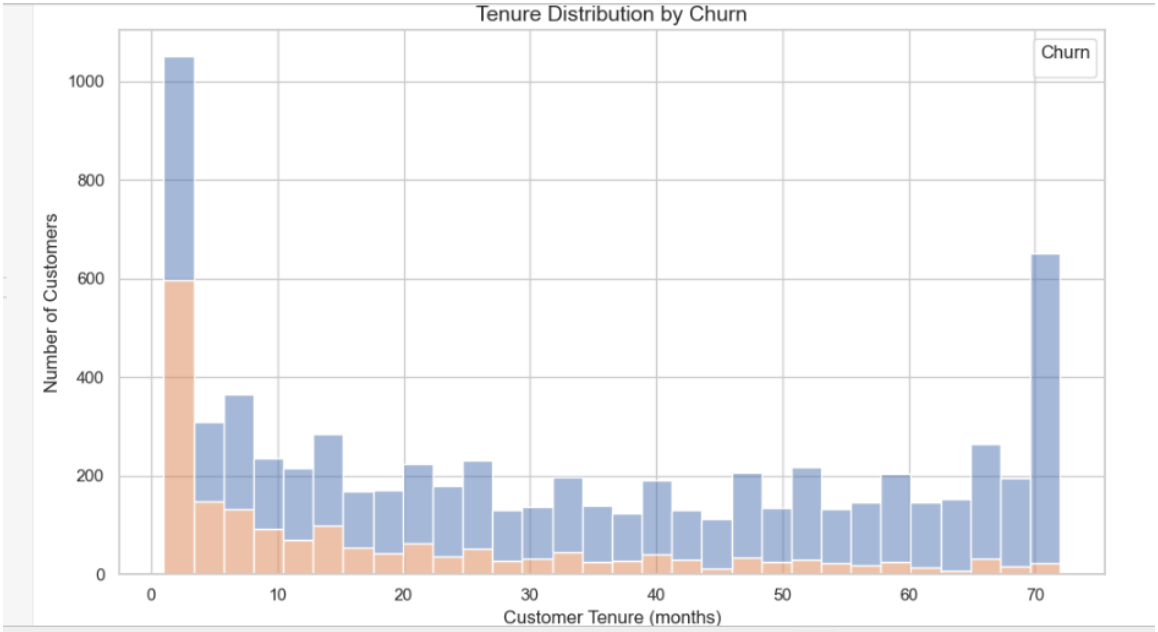


Figure 3: Tenure distribution by churn status.

The histogram indicates that churn is concentrated among customers with low tenure. Many customers churn within their first year, and churn density declines sharply for customers who have remained for several years. There is also a spike at 72 months, reflecting long-standing loyal customers who rarely churn. This suggests that early engagement and onboarding are critical: once customers stay for a long period, they tend to remain loyal.

4. Methodology

The goal is to predict the probability that a customer will churn, given their demographic and account information. The modeling pipeline includes train–test splitting, preprocessing of numeric and categorical features, model training, and performance evaluation.

We split the data into training and test sets using an 80/20 stratified split to preserve the churn proportion in both sets. Numeric features (SeniorCitizen, tenure, MonthlyCharges, and TotalCharges) were standardized using z-score scaling. Categorical features were transformed using one-hot encoding with the handle_unknown='ignore' option, which ensures robust handling of rare categories in the test set.

We trained three classification models:
1. Logistic Regression with class_weight='balanced' to compensate for class imbalance.
2. Random Forest classifier with 200 trees and class_weight='balanced'.
3. Gradient Boosting classifier with default hyperparameters.
All models were wrapped in Scikit-Learn Pipelines so that preprocessing and model estimation are performed consistently during both training and prediction.

5. Model Evaluation Results

Model performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score, and area under the ROC curve (ROC-AUC). Precision measures how often predicted churners actually churn, while recall measures how many of the true churners the model successfully identifies. F1-score balances precision and recall, and ROC-AUC summarizes the model's ability to separate churners from non-churners across all classification thresholds.

Table 1 summarizes the test-set performance of the three models.

	Model	Accuracy	Precision	Recall	F1	ROC_AUC
2	Gradient Boosting	0.796020	0.640777	0.529412	0.579795	0.838262
0	Logistic Regression	0.723525	0.487644	0.791444	0.603466	0.834402
1	Random Forest	0.785359	0.624138	0.483957	0.545181	0.817310

Table 1: Test-set performance metrics for all models.

Gradient Boosting achieves the highest overall accuracy (approximately 0.80) and the highest ROC-AUC (around 0.84), indicating strong separability between churners and non-churners. Logistic Regression performs competitively with a slightly lower accuracy but similar ROC-AUC, while Random Forest also performs well but with slightly lower recall. In many churn applications, recall on the positive class is particularly important, because missing a true churner may mean losing revenue. Logistic Regression attains the highest recall but at the cost of lower precision, whereas Gradient Boosting offers a better balance of precision and recall.

Figures 4–6 show the confusion matrices for the three models on the test set. Each matrix summarizes the counts of true negatives, false positives, false negatives, and true positives.

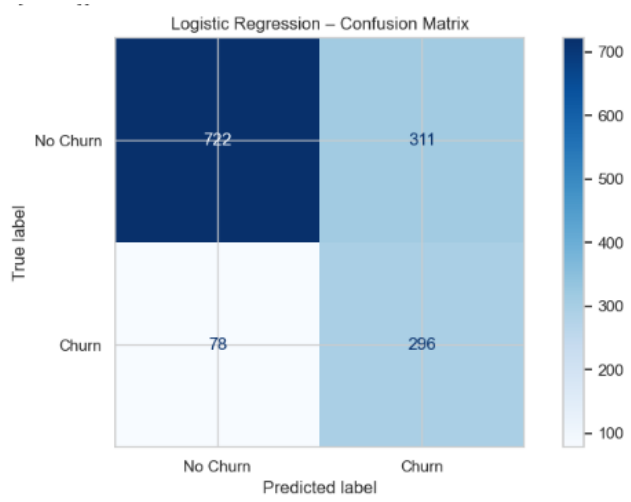


Figure 4: Confusion matrix – Logistic Regression.

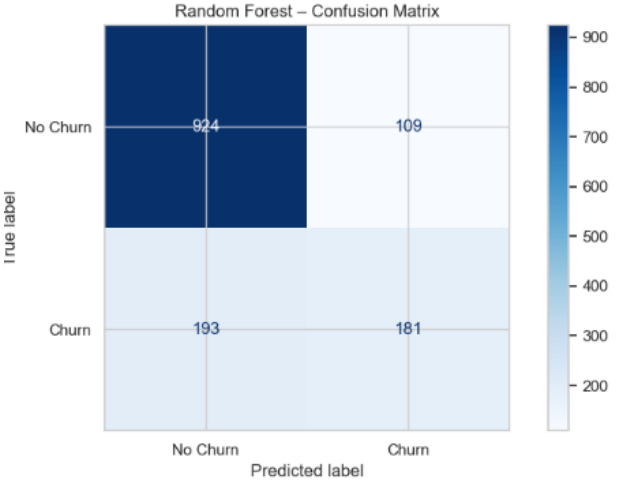


Figure 5: Confusion matrix – Random Forest.

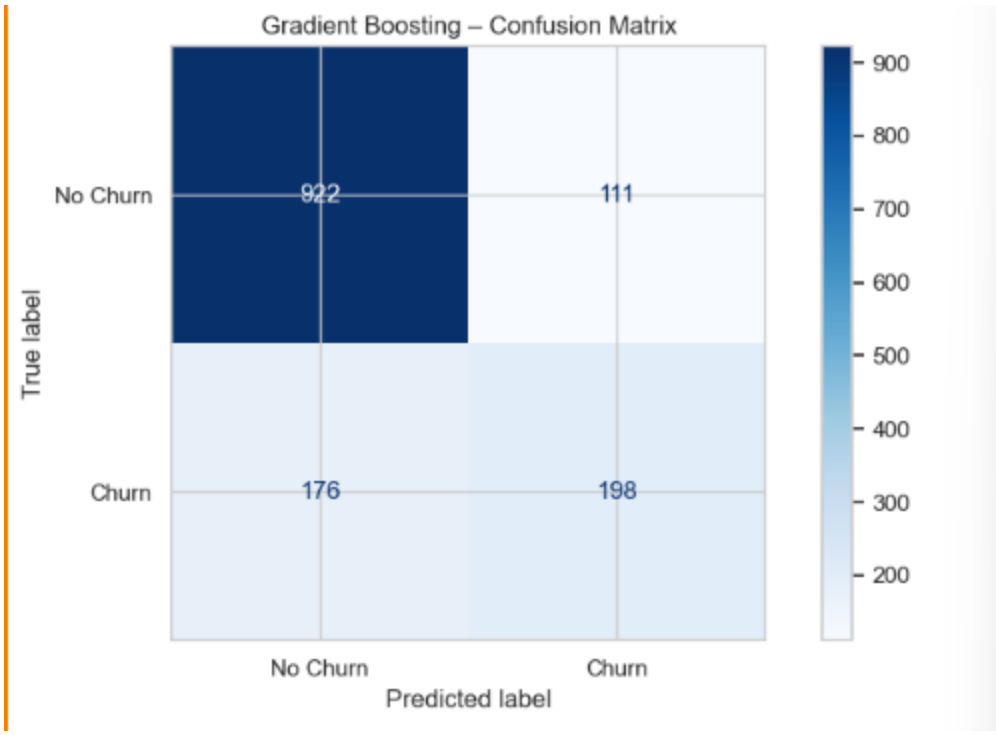
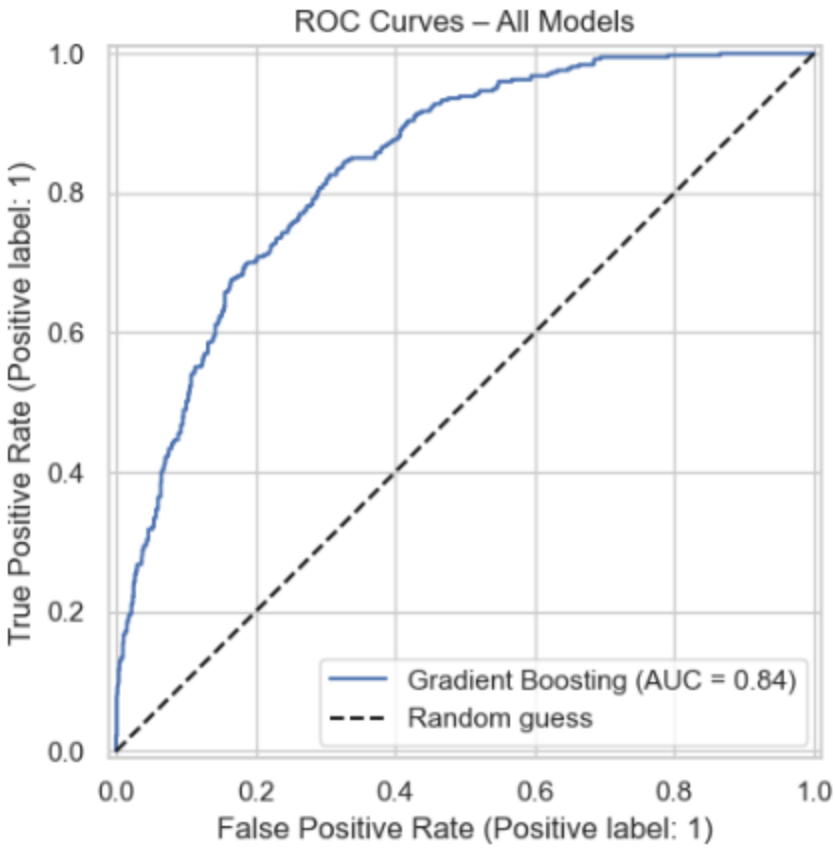
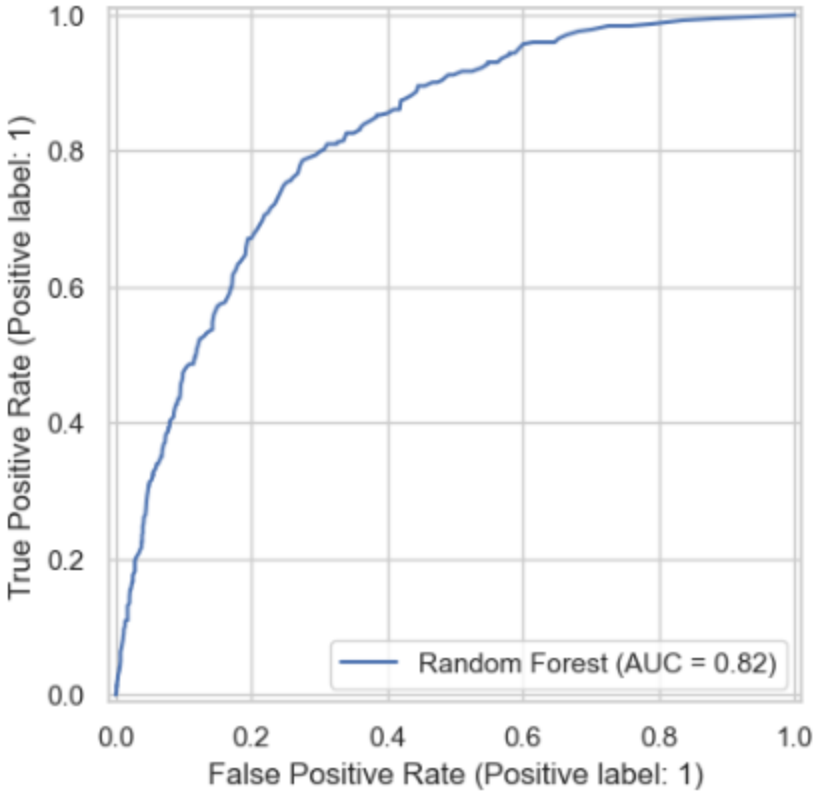


Figure 6: Confusion matrix – Gradient Boosting.

Across all models, the majority of customers are correctly classified as non-churners. Logistic Regression produces relatively more false positives but fewer false negatives, reflecting its higher recall. Gradient Boosting strikes a balance, with a good number of true positives while keeping false positives moderate. Random Forest tends to be more conservative, with fewer positive predictions and thus lower recall.

Receiver operating characteristic (ROC) curves offer another way to compare models across a range of thresholds. Figure 7 overlays the ROC curves for all three models.


ROC Curves – All Models
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Gradient Boosting (AUC = 0.84)
Random guess
0.0
0.2
0.4
0.6
0.8
1.0


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Random Forest (AUC = 0.82)
0.0
0.2
0.4
0.6
0.8
1.0

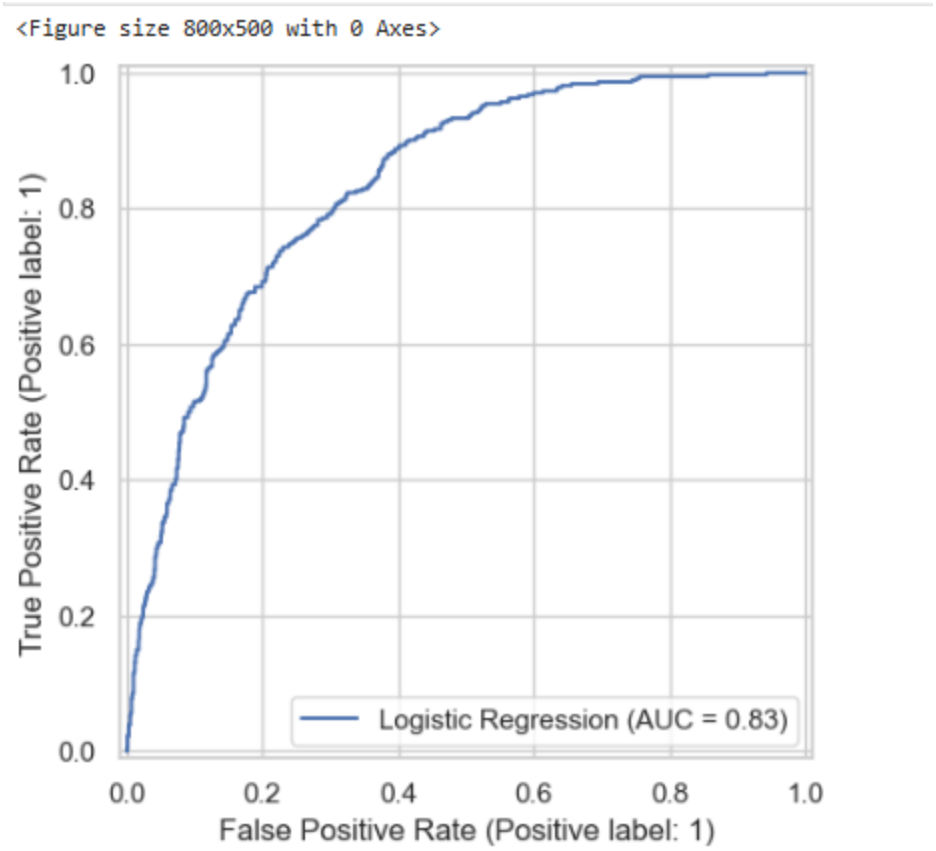


Figure 7: ROC curves for Logistic Regression, Random Forest, and Gradient Boosting.

All three models substantially outperform random guessing, as indicated by curves well above the diagonal. Gradient Boosting has the largest area under the curve, consistent with the summary metrics, followed closely by Logistic Regression. This reinforces the conclusion that Gradient Boosting is the best-performing model overall, while Logistic Regression remains a strong, more interpretable baseline.

6. Feature Importance and Drivers of Churn

To better understand which factors drive churn, we examined feature importances from the Gradient Boosting model. Tree-based models such as Gradient Boosting provide a natural measure of how much each feature contributes to reducing classification error.

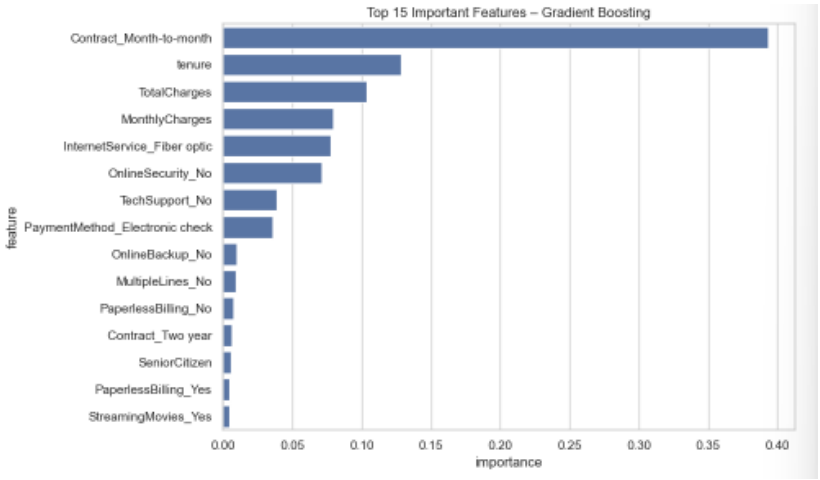


Figure 8: Top 15 most important features according to the Gradient Boosting model.

The most influential feature by a wide margin is Contract_Month-to-month, indicating that being on a month-to-month contract is a strong risk factor for churn. Tenure, TotalCharges, and MonthlyCharges also play important roles, consistent with the EDA findings. Customers with low tenure and low cumulative charges are more likely to leave, while very long-tenured customers are much less likely to churn. Service-related features such as having fiber optic internet, lacking online security or technical support, and using electronic check as the payment method also contribute meaningfully to churn risk.

The importance of payment method suggests that customers paying by electronic check may experience more friction or dissatisfaction compared to those on automatic credit card or bank transfers. Similarly, the importance of security and support features indicates that bundling these services or emphasizing their value could help reduce churn.

7. Business Recommendations

Based on the modeling results and feature importance analysis, several actionable recommendations emerge for the telecom provider.

First, customers on month-to-month contracts should be the primary focus of retention efforts. The company should design targeted campaigns that encourage these customers to migrate to one-year or two-year contracts by offering discounts, loyalty points, or value-added bundles. Locking customers into longer-term agreements directly addresses the strongest churn driver identified in the model.

Second, early-tenure customers represent a vulnerable segment. Many customers churn within their first year, so the company should strengthen onboarding programs, provide proactive check-ins during the first few months of service, and quickly resolve any

service issues. Welcome offers, tutorials, and personalized support may help new customers see value early and reduce early churn.

Third, pricing and bill transparency are crucial. Higher monthly charges are associated with churn, especially for customers who may not fully understand their bill. The company should review its pricing structure and consider offering usage-based or tiered plans that better match customer needs. Proactively contacting customers whose bills have recently increased could prevent dissatisfaction from turning into cancellations.

Fourth, the analysis suggests that customers paying by electronic check are more likely to churn. Encouraging these customers to switch to automatic payments via credit card or bank transfer—perhaps with small incentives—could reduce churn while also improving revenue predictability.

8. Limitations and Future Work

Despite the strong performance of the Gradient Boosting model, several limitations should be acknowledged. The dataset is historical and may not fully reflect future customer behavior if market conditions, competitors, or internal policies change. The feature set is also limited: important drivers such as customer satisfaction scores, network quality, call center interactions, or promotional campaign exposure are not included.

Additionally, hyperparameter tuning for the models was limited. A systematic search over model hyperparameters, for example using grid search or randomized search with cross-validation, could yield further improvements in predictive performance. Another avenue for future work is cost-sensitive modeling: incorporating the financial value of retaining a customer versus the cost of retention offers to optimize decisions based not only on churn probability but also on expected profit.

Finally, while Gradient Boosting provides strong accuracy, it is less interpretable than simpler models such as Logistic Regression. Techniques like partial dependence plots or SHAP values could be used in future analyses to provide more granular explanations of how individual features affect churn probability.

9. Conclusion

This project developed and evaluated several predictive models for telecom customer churn using the Telco Customer Churn dataset. After appropriate preprocessing and feature engineering, Logistic Regression, Random Forest, and Gradient Boosting models were trained and compared. Gradient Boosting achieved the best overall performance with an ROC-AUC of approximately 0.84, while Logistic Regression provided a strong, interpretable benchmark with higher recall.

The analysis confirmed that month-to-month contracts, short tenure, higher charges, lack of support or security services, and electronic check payment are key risk factors for

churn. These insights translate directly into business actions, including incentivizing longer-term contracts, strengthening early-life customer engagement, reviewing pricing and billing practices, and promoting automatic payment methods. Implementing these recommendations, supported by an ongoing churn prediction system based on the Gradient Boosting model, can help the telecom provider reduce churn, stabilize revenue, and improve customer lifetime value.