Title

Machine Learning Approaches for Customer Churn Prediction: Balancing Accuracy and Interpretability

ISBN

9798293838837

Author

Chen, Jack

Publication Date

2025-09-08

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Machine Learning Approaches for Customer Churn Prediction: Balancing Accuracy and Interpretability

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics and Data Science

by

Jack Chen

2025

ABSTRACT OF THE THESIS

Machine Learning Approaches for Customer Churn Prediction: Balancing Accuracy and Interpretability

by

Jack Chen

Master of Applied Statistics and Data Science

University of California, Los Angeles, 2025

Ying Nian Wu, Chair

The study addresses the problem of customer churn in the telecommunications industry, where retaining existing users is significantly more cost-effective than acquiring new ones. It investigates the application of machine learning techniques for churn prediction using demographic, contractual, service, and billing information. A range of models are evaluated, from interpretable approaches such as Logistic Regression and Decision Trees to advanced methods including Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks. The analysis emphasizes predictive performance and interpretability, identifies key factors driving churn, and discusses trade-offs among different approaches. The findings provide both methodological insights into the use of machine learning for churn prediction and practical guidance for developing data-driven strategies to improve customer retention.

The thesis of Jack Chen is approved.

Maria Cha

Xiaowu Dai

Hongquan Xu

Yingnian Wu, Committee Chair

University of California, Los Angeles

2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Customer churn, defined as the discontinuation of services by existing clients, remains a critical challenge across industries such as telecommunications, banking, insurance, and subscription-based businesses. Retaining customers is substantially more cost-effective than acquiring new ones [1]; industry reports suggest that acquiring a new customer can cost up to five times more than retaining an existing one. Consequently, churn erodes not only immediate revenue streams but also undermines long-term profitability and signals underlying weaknesses in customer experience.

The growing availability of large-scale customer data has enabled the use of Machine Learning (ML) for churn prediction. Early studies relied on interpretable models such as logistic regression and survival analysis, which offered transparency but limited predictive capacity. More advanced methods—including decision trees, random forests, and gradient boosting—have demonstrated superior performance by capturing complex nonlinear relationships. Despite these advances, two central challenges persist: (1) the trade-off between predictive accuracy and interpretability, and (2) the complication of class imbalance, since churners represent only a minority of the customer base.

From a methodological standpoint, existing studies often focus narrowly on maximizing accuracy without systematically addressing interpretability. From a practical standpoint, organizations require not only predictive signals but also actionable insights, such as identifying whether contract type, billing method, or service usage patterns most strongly influence churn behavior. Bridging this gap is essential for aligning academic advances with real-world applications.

1.2 Problem Setting

This thesis investigates the problem of predicting customer churn in the telecommunications industry using machine learning techniques. Formally, churn prediction is a binary classification task: given a set of customer attributes—such as demographics, contract type, billing method, and service usage—the objective is to predict whether a customer will discontinue their service. Beyond classification, the study emphasizes identifying the drivers of churn and clarifying the trade-off between predictive performance and interpretability.

1.3 Research Objectives and Contributions

This research aims to contribute both methodologically and practically. The objectives are threefold:

1. To systematically evaluate a range of machine learning models—including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Multilayer Perceptron—on churn prediction tasks.

2. To assess not only predictive accuracy but also model robustness and interpretability, highlighting the trade-offs among different approaches.

3. To provide empirical insights into the most influential customer attributes, thereby offering actionable recommendations for targeted retention strategies.

By integrating exploratory data analysis, comparative modeling, and systematic evaluation, this study seeks to advance the understanding of machine learning approaches for churn prediction and to demonstrate their value in guiding data-driven decision-making in competitive markets.

Chapter 2

Data Preparation

2.1 Dataset Description

The dataset used in this study is the **Telco Customer Churn dataset** [2], released by IBM and publicly available on Kaggle. It contains 7,043 customer records and 21 variables, encompassing demographic attributes, subscription details, service usage patterns, billing information, and a binary churn indicator that specifies whether a customer discontinued service. The dataset has become a standard benchmark in churn prediction research due to its relatively clean structure, balanced coverage of feature categories, and direct relevance to the telecommunications sector.

A key characteristic of the dataset is its class distribution: approximately 26.5% of customers have churned, while 73.5% remain. This imbalance mirrors real-world business conditions, where most customers are retained, but the minority group of churners poses a disproportionately large threat to profitability. From a methodological perspective, such imbalance means that models may achieve deceptively high accuracy by predicting the majority class while failing to capture at-risk customers. Therefore, this study emphasizes evaluation metrics beyond overall accuracy, including recall, F1-score, and ROC-AUC, which better reflect the model's ability to detect minority-class churners. From a business perspective,

correctly identifying churners provides opportunities for targeted retention efforts, making the churn label not just a binary outcome but a critical signal for customer relationship management.

The variables can be grouped into four broad categories: `demographics` (e.g., gender, senior-citizen status, presence of dependents), `service subscriptions` (e.g., internet service type, online security, streaming services), `contract and billing information` (e.g., contract type, payment method, paperless billing), and `financial variables` (monthly and total charges). This variety of feature types enables comprehensive exploration of how both behavioral and contractual factors influence churn.

Table 2.1: Telco Customer Churn Dataset: Selected Core Variables

Variable Name	**Variable Description**
gender	Gender of the customer (Male, Female).
SeniorCitizen	Indicates if the customer is a senior citizen (1 = Yes, 0 = No).
Partner	Whether the customer has a partner (Yes, No).
InternetService	Type of internet service (DSL, Fiber optic, No).
OnlineSecurity	Whether the customer has online security service.
TechSupport	Whether the customer has technical support service.
Contract	Contract type (Month-to-month, One year, Two year).
PaymentMethod	Payment method (Electronic check, Mailed check, Bank transfer, Credit card).
MonthlyCharges	Monthly amount charged to the customer (numeric).
TotalCharges	Total amount charged to the customer (numeric).
Churn	Target variable: whether the customer discontinued service (Yes, No).

2.2 Data Preprocessing

To ensure the dataset was suitable for machine learning analysis, several preprocessing steps were conducted systematically.

First, missing or invalid values in the variable `TotalCharges` were addressed. These cases corresponded primarily to new customers with incomplete billing histories. Since the proportion of such records was very small, they were removed from the dataset rather than imputed, minimizing noise without altering the underlying distribution.

Second, categorical variables such as `Contract`, `PaymentMethod`, and `InternetService` were transformed using one-hot encoding. This approach preserves the categorical nature of the variables while making them compatible with algorithms that require numerical input, including logistic regression and neural networks. One-hot encoding also avoids introducing artificial ordinal relationships that could bias the models.

Third, numerical features including `MonthlyCharges` and `TotalCharges` were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

where x is the raw value, μ is the mean, and σ is the standard deviation. Standardization ensures that features are on a comparable scale, which is particularly important for gradient-based algorithms such as logistic regression and multilayer perceptrons.

Finally, the dataset was divided into training and testing subsets following an 80/20 split, with stratification applied to preserve the original churn-to-non-churn ratio in both subsets. This design provides a robust basis for model training and unbiased evaluation, while maintaining the class imbalance challenge that characterizes real-world churn data.

Chapter 3

Exploratory Data Analysis

3.1 Descriptive Statistics

The dataset consists of 7,043 customer records with a balanced gender distribution: male and female customers are nearly equal in number. Approximately 16% of customers are classified as senior citizens, while the majority are younger. In terms of family status, about 49% of customers have a partner and 30% have dependents. Figure 3.1 illustrates the churn distribution across these demographic categories.

The descriptive results suggest that demographic variables exert heterogeneous effects on churn. Gender shows little difference in churn rates, indicating limited predictive power. However, senior citizens display noticeably higher churn compared to younger customers. Similarly, customers without a partner or dependents exhibit elevated churn likelihood, suggesting that family-related factors may contribute to service retention. These initial findings provide a foundation for more detailed analyses of categorical and numerical variables in subsequent sections.

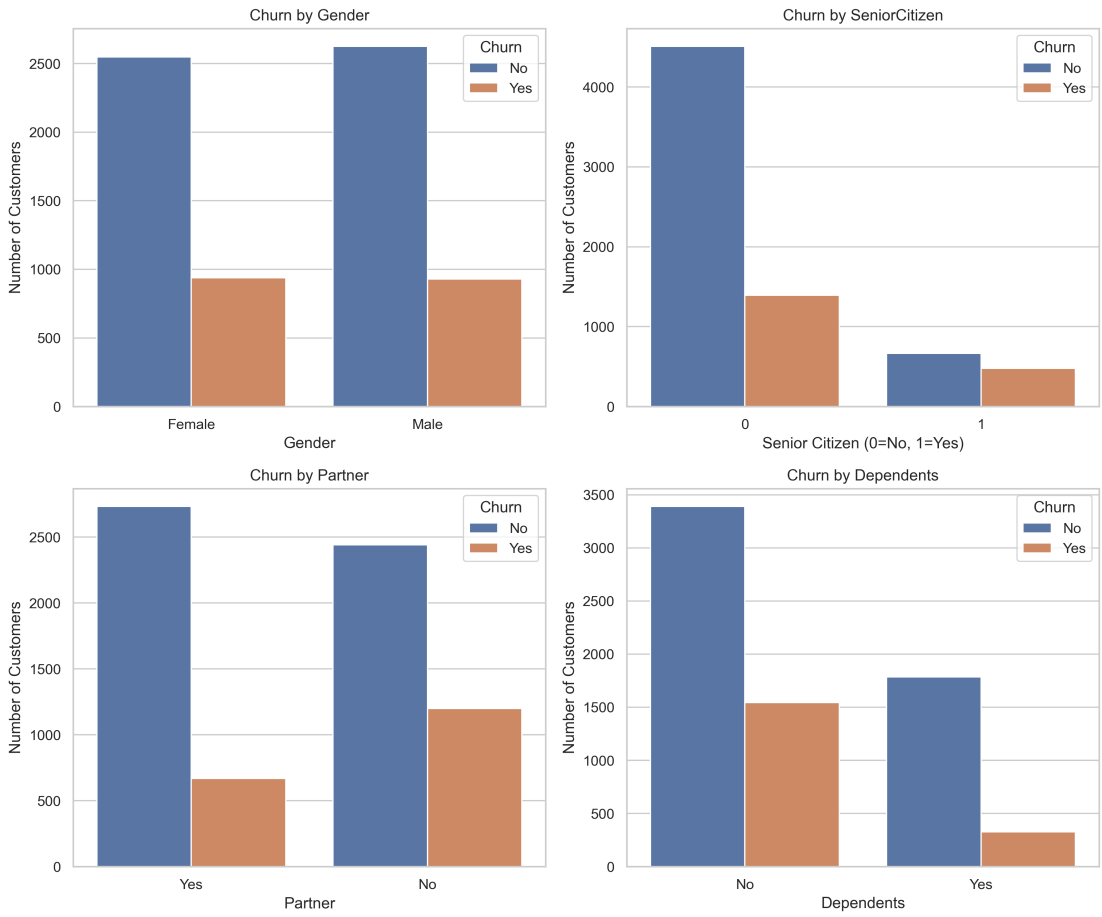

Figure 3.1: Customer distribution across demographic groups (gender, senior citizen status, partner, and dependents) and their relationship with churn.

3.2 Categorical Variables Analysis

Categorical variables such as contract type, internet service, online security, and technical support exhibit distinct relationships with customer churn. As shown in Figure 3.2, customers on month-to-month contracts are substantially more likely to churn compared to those with one-year or two-year contracts. This finding reflects the stabilizing effect of long-term agreements in reducing customer attrition.

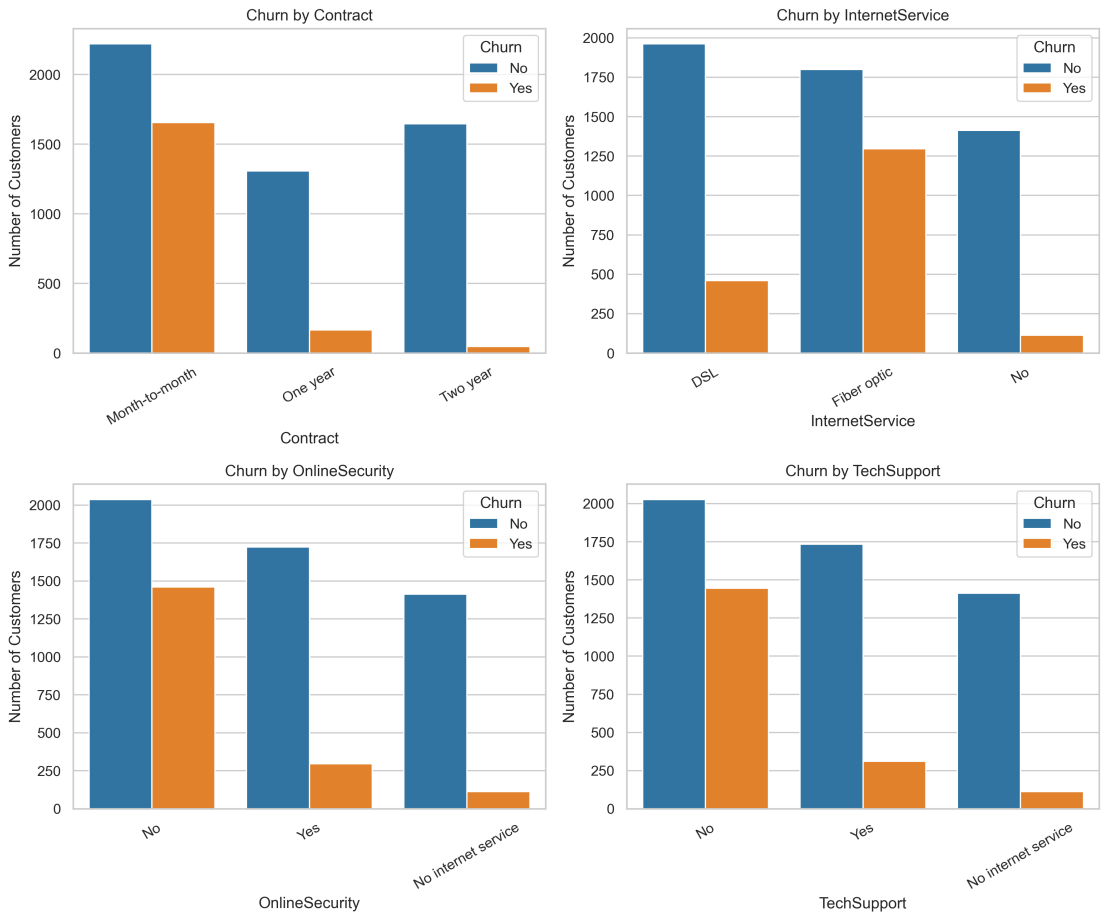

Figure 3.2: Churn distribution across categorical variables: Contract, InternetService, OnlineSecurity, and TechSupport.

3.3 Numerical Variables Analysis

Numerical variables such as `MonthlyCharges` and `TotalCharges` provide important insights into churn behavior. As shown in Figures 3.3 and 3.4, churned customers are disproportionately concentrated in higher monthly charge ranges, suggesting that price sensitivity contributes to service discontinuation. By contrast, customers with lower monthly charges are less likely to churn, reflecting the stabilizing effect of more affordable plans.

Similarly, customers with lower accumulated `TotalCharges` exhibit elevated churn rates, whereas long-tenure customers with higher total charges tend to remain with the provider. These patterns indicate that early-stage customers face the highest attrition risk, while

long-standing customers are more likely to demonstrate loyalty.

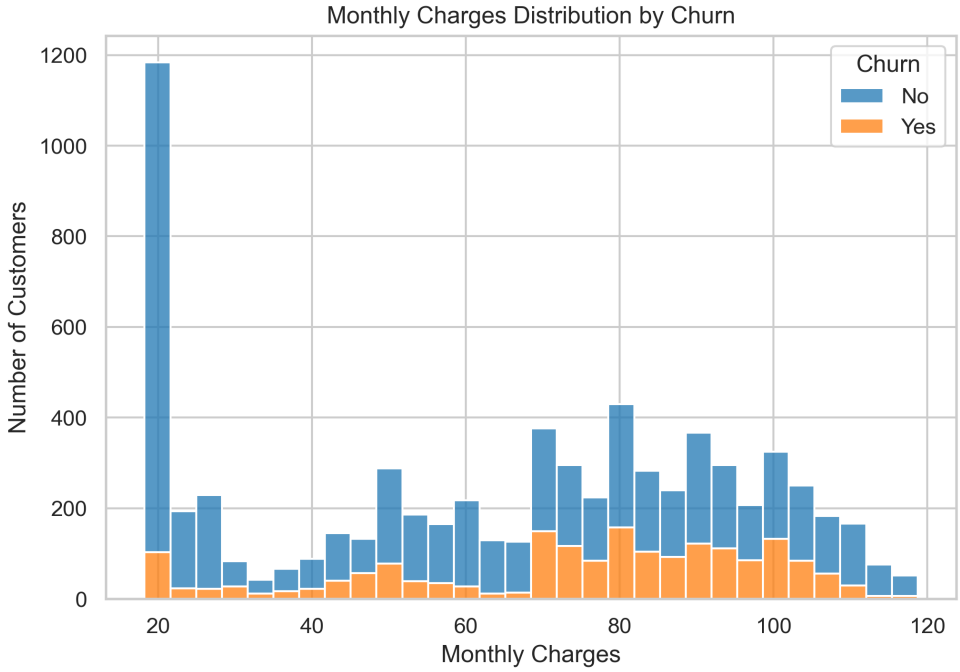

Figure 3.3: Distribution of monthly charges across churn categories.

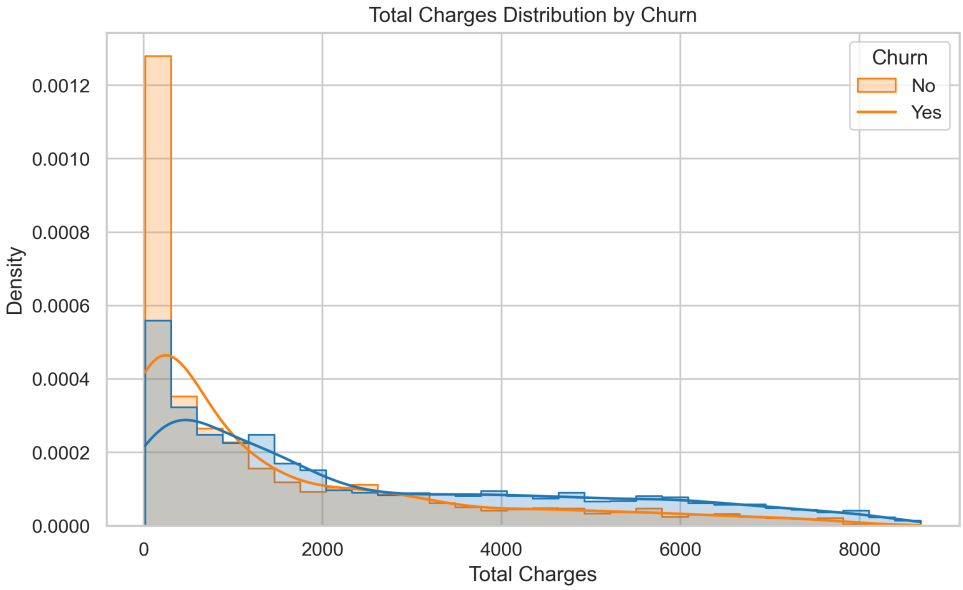

Figure 3.4: Distribution of total charges across churn categories.

3.4 Correlation Analysis

To further examine the interdependencies among numerical features, an extended correlation analysis was conducted. In addition to `tenure`, `MonthlyCharges`, and `TotalCharges`, the variables `SeniorCitizen` and `Churn` (converted to binary form) were included. Pearson's correlation coefficients were calculated, and the results are visualized in Figure 3.5.

The heatmap reveals several noteworthy patterns. First, `TotalCharges` exhibits a very strong positive correlation with `tenure`, reflecting the cumulative nature of total billing: longer-tenure customers inevitably accumulate higher charges. Second, `MonthlyCharges` shows only moderate correlation with `TotalCharges`, as it depends primarily on the service package rather than on customer tenure. Third, `Churn` demonstrates a negative correlation with `tenure` and `TotalCharges`, consistent with the earlier finding that new or low-spending customers are more prone to leave. Finally, `SeniorCitizen` shows a weak positive correlation with churn, indicating that older customers may be somewhat more vulnerable to service discontinuation.

These patterns underscore the interplay between demographic and financial variables, highlighting that both long-term service engagement and pricing structures play an essential role in predicting churn.

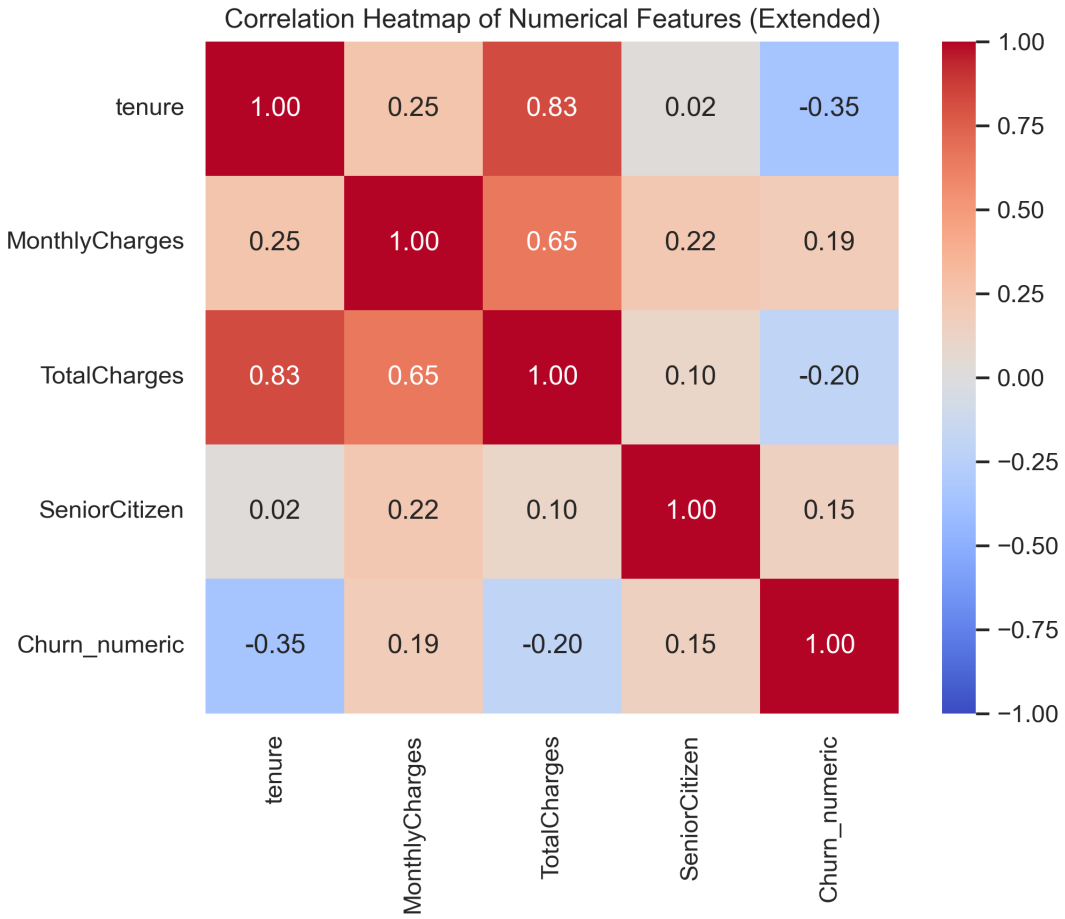

Figure 3.5: Extended correlation heatmap including tenure, monthly charges, total charges, senior citizen status, and churn.

Chapter 4

Methodology

4.1 Modeling

This chapter develops predictive models to evaluate customer churn risk. A set of seven supervised learning methods is employed, covering both traditional statistical models and modern machine learning techniques. Logistic Regression and Naive Bayes are used as baseline models, offering interpretability and probabilistic reasoning. Tree-based methods, including Decision Tree, Random Forest, and Gradient [3], are applied to capture non-linear feature interactions and assess variable importance. Support Vector Machines provide flexible decision boundaries through kernel functions, while Multi-Layer Perceptrons represent neural network approaches capable of modeling complex data structures. Together, these models enable a comprehensive comparison that balances predictive accuracy, interpretability, and practical applicability.

4.2 Baseline Models and Results

4.2.1 Logistic Regression

Logistic Regression is one of the most fundamental models for binary classification [4] and serves as a natural baseline for churn prediction. Unlike linear regression, which directly predicts a continuous outcome, logistic regression estimates the probability of an event by modeling the log-odds of the response variable as a linear combination of the predictors:

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where $Y \in \{0,1\}$ denotes churn status, $X = (x_1, \ldots, x_p)$ is the feature vector, and $\beta = (\beta_0, \ldots, \beta_p)$ are model coefficients. The logistic transformation ensures that the predicted probability lies between 0 and 1:

$$P(Y=1|X) = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}}$$

Application to the Telco Dataset. In the Telco Customer Churn dataset, logistic regression provides interpretable insights into how demographic, contractual, and financial factors contribute to churn risk. For example, a negative coefficient for `tenure` indicates that customers with longer service duration are less likely to churn, while a positive coefficient for `MonthlyCharges` suggests that higher monthly fees increase churn risk. Moreover, categorical variables such as `Contract` are encoded into dummy variables, allowing the model to quantify risk differences between short-term and long-term contracts.

Regularization (L_1 or L_2) is applied to reduce the effect of multicollinearity among service-related variables (e.g., `InternetService`, `StreamingTV`, `StreamingMovies`). This helps improve generalization performance.

Table 4.1: Selected Logistic Regression Coefficients (Top 10)

Variable	**Coefficient**
tenure	-1.350
Contract-Two year	-0.775
TotalCharges	0.641
Contract-Month-to-month	0.612
InternetService-DSL	-0.605
InternetService-Fiber optic	0.572
MonthlyCharges	-0.509
PaperlessBilling-No	-0.302
MultipleLines-No	-0.295
StreamingMovies-No internet service	-0.279

Results. The performance of Logistic Regression is summarized in Table 4.2. The model achieves moderate accuracy but relatively strong recall, which is valuable since identifying potential churners is more critical than misclassifying a few non-churners.

Table 4.2: Performance of Logistic Regression

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Logistic Regression	0.804	0.711	0.674	0.692	0.846

4.2.2 Naive Bayes

Naive Bayes is a family of probabilistic classifiers based on Bayes' theorem, with the fundamental assumption that features are conditionally independent given the class label. For a sample $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, the posterior probability of belonging to class C_k is:

$$P(C_k \mid \mathbf{x}) \propto P(C_k) \prod_{i=1}^{p} P(x_i \mid C_k)$$

where $P(C_k)$ denotes the prior probability of class C_k, and $P(x_i \mid C_k)$ is the likelihood of feature x_i given the class. This conditional independence assumption greatly simplifies the computation compared to modeling the full joint distribution.

In the Telco churn dataset, numeric variables (e.g., `tenure`, `MonthlyCharges`, `TotalCharges`) are modeled with Gaussian distributions, while categorical variables are estimated using frequency counts.

Table 4.3 further presents the class-conditional means and variances for key numeric features. Customers who did not churn had significantly longer average tenure (37.76 vs. 18.22), while churned customers exhibited higher average monthly charges (74.68 vs. 61.50). These findings are consistent with business intuition: short-term contracts and higher charges increase the likelihood of churn.

Table 4.3: GaussianNB parameters: class-wise mean and variance for numeric features

Feature	**NoChurn_mean**	**NoChurn_var**	**Churn_mean**	**Churn_var**
tenure	37.755	578.092	18.217	389.243
MonthlyCharges	61.495	966.480	74.680	613.738
TotalCharges	2572.883	5452463.121	1553.072	3659913.120

As shown in Table 4.4, the Naive Bayes classifier achieved an overall accuracy of 0.684 and a ROC-AUC of 0.804. The recall for the churn class reached 0.829, indicating the model is effective at capturing potential churners. However, the precision was lower (0.449), suggesting a relatively high false-positive rate.

Table 4.4: Classification performance of Naive Bayes on the Telco test set

Class	Precision	Recall	F1-score	Support
No Churn	0.911	0.632	0.746	1033
Churn	0.449	0.829	0.583	374
`Overall:` Accuracy = 0.684, ROC-AUC = 0.804				

4.3 Tree-Based Models and Results

4.3.1 Decision Tree

Decision trees are non-parametric supervised learning models that recursively partition the feature space to form piecewise-constant decision rules [5, 6]. At each internal node, the algorithm selects the split (feature and threshold or category) that most reduces impurity. For classification trees, a common impurity measure is the Gini index:

$$\text{Gini}(t) \;=\; 1 \;-\; \sum_{i=1}^{C} p(i \mid t)^2,$$

where $p(i \mid t)$ is the empirical proportion of class i among the samples reaching node t, and C is the number of classes. Lower impurity indicates a purer node.

Applied to the Telco churn dataset, the learned tree highlights business-relevant rules: customers on `Month-to-Month` contracts and those using `Fiber optic` internet tend to have higher churn risk, whereas longer tenure and the presence of support/security services are associated with retention. A simplified view of the fitted tree is shown in Figure 4.1.

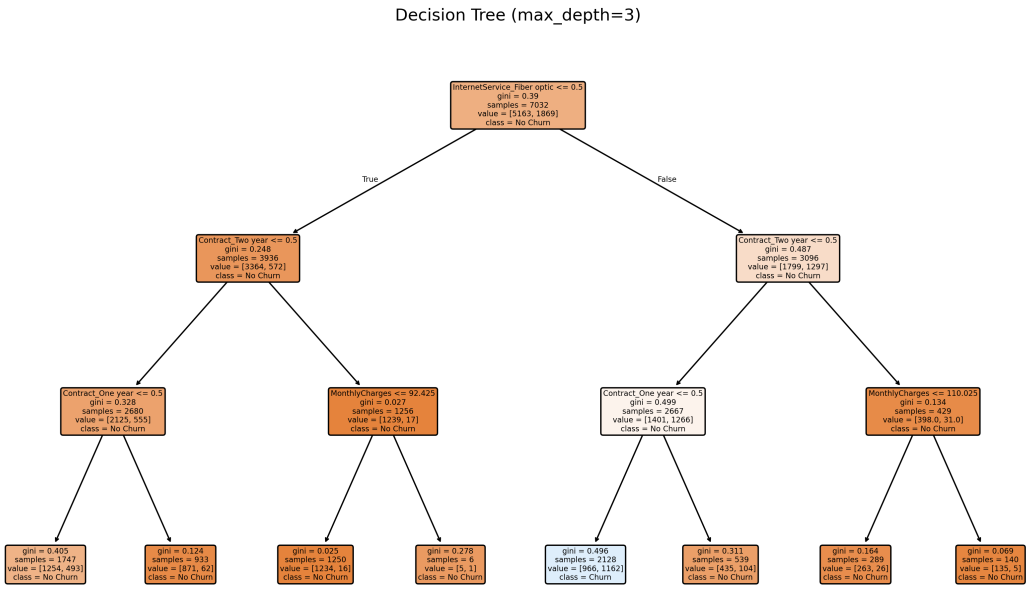

Figure 4.1: Simplified decision tree (max depth = 3).

Table 4.5 reports test performance. The model attains an accuracy of 0.731 and ROC–AUC of 0.812. Recall for the *Churn* class is 0.781, indicating good coverage of high-risk customers, though precision is lower due to false positives. Despite being less robust than ensembles, decision trees remain valuable for their interpretability and actionable rules.

Table 4.5: Classification performance of Decision Tree

Class	**Precision**	**Recall**	**F1-score**	**Support**
No Churn	0.900	0.713	0.796	1033
Churn	0.497	0.781	0.607	374
Overall: Accuracy = 0.731, ROC–AUC = 0.812				

4.3.2 Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or

mean prediction (regression) of the individual trees [7]. It addresses the problem of overfitting inherent in single decision trees by introducing randomness both in sampling and in feature selection. The probability of classifying an instance x as class c is obtained by aggregating predictions across T trees:

$$\hat{y} = \text{argmax}_{c \in \mathcal{C}} \sum_{t=1}^{T} I\left(h_t(x) = c\right),$$

where $h_t(x)$ is the prediction of the t-th decision tree and $I(\cdot)$ is the indicator function.

Applied to the Telco churn dataset, the Random Forest model captures nonlinear relationships and interactions between features, which are crucial in understanding customer churn. In particular, it highlights that both service-related features (e.g., contract type, internet service) and financial features (e.g., monthly charges, total charges) contribute significantly to churn prediction.

Figure 4.2 presents the top-15 most important features derived from the Random Forest model. Total charges, tenure, and monthly charges emerge as the strongest predictors, while service-related features such as contract type and online security also play vital roles. This aligns with domain intuition: customers with short tenure, higher monthly payments, and less service protection are more prone to churn.

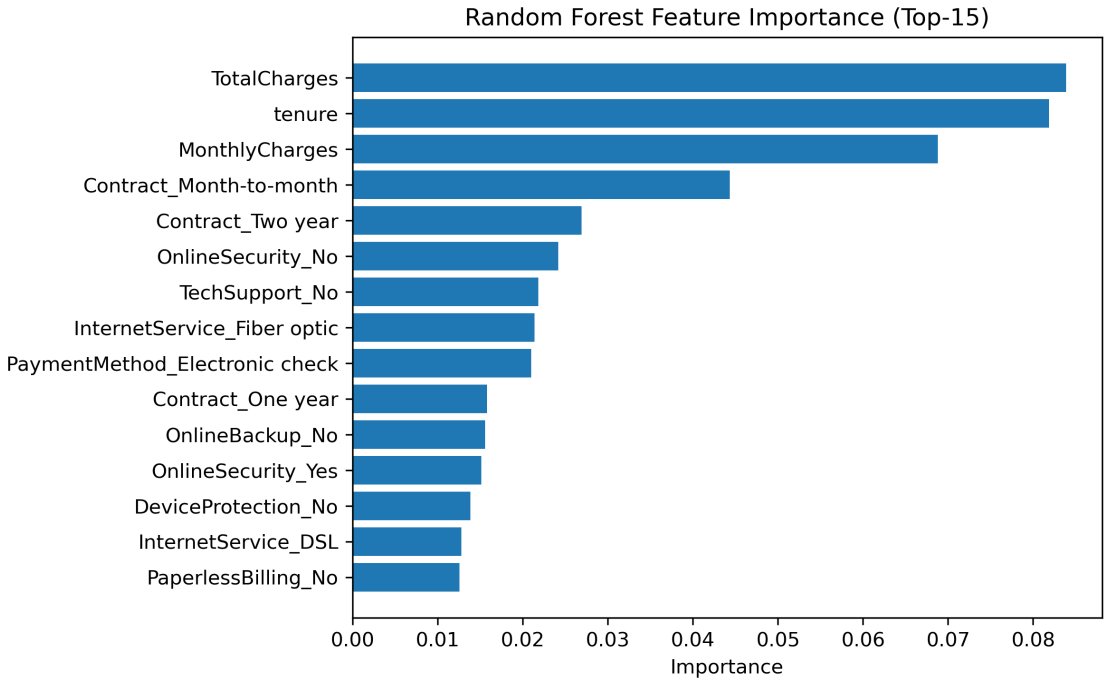

Figure 4.2: Top feature importances from random forest model.

Table 4.6 reports the classification performance of the Random Forest model. Compared to the single decision tree, Random Forest achieves higher accuracy and a better balance between precision and recall, demonstrating the advantage of ensemble methods in handling complex churn prediction tasks.

Table 4.6: Classification performance of Random Forest

Class	**Precision**	**Recall**	**F1-score**	**Support**
No Churn	0.831	0.894	0.862	1033
Churn	0.631	0.497	0.556	374
Overall: Accuracy = 0.789, ROC–AUC = 0.815				

4.4 Advanced Models and Results

4.4.1 XGBoost (Extreme Gradient Boosting)

XGBoost builds an additive ensemble of CART trees to minimize a regularized logistic [8]. The objective can be written as

$$\min_{F} \sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i F(x_i))\bigr) + \Omega(F), \qquad F(x) = \sum_{m=1}^{M} \alpha_m \, h_m(x),$$

where h_m is the m-th tree, α_m its weight, and $\Omega(F)$ penalizes model complexity. The predicted probability is $\hat{p}(y = 1 \mid x) = \sigma(F(x))$.

For Telco churn, boosted trees capture non-linearities and interactions among *contract type*, *monthly/total charges*, *tenure*, and *value-added services*, which are known drivers of churn propensity.

On the held-out test set, XGBoost attains **Accuracy = 0.790** and **ROC–AUC = 0.832**. The ROC curve, shown in Figure 4.3, illustrates the strong separation between positive and negative classes.

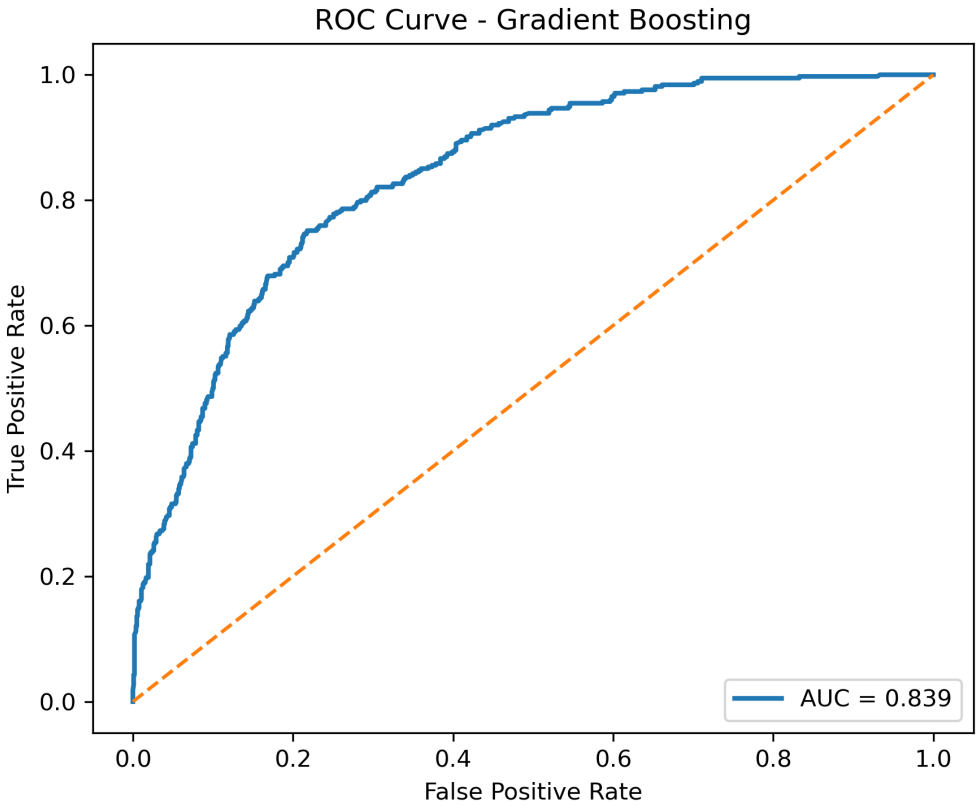

Figure 4.3: ROC curve of XGBoost.

The classification performance is summarized in Table 4.7. The model shows strong specificity (recall = 0.881 for the *No Churn* class), while the *Churn* class remains more challenging (precision = 0.620, recall = 0.537).

Table 4.7: Classification performance of XGBoost

Class / Average	**Precision**	**Recall**	**F1-score**	**Support**
No Churn (0)	0.840	0.881	0.860	1033
Churn (1)	0.620	0.537	0.576	374
Macro Avg	0.730	0.709	0.718	1407
Weighted Avg	0.782	0.790	0.785	1407
Overall Accuracy		0.790		
ROC–AUC		0.832		

4.4.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) are powerful supervised learning algorithms that aim to find the optimal hyperplane that separates different classes with the maximum margin [9]. For a binary classification problem with data $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$, the primal optimization problem of the SVM can be formulated as:

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i,$$

where w and b define the separating hyperplane, $\phi(x)$ maps the data into a higher-dimensional feature space, ξ_i are slack variables, and $C > 0$ is the regularization parameter. The use of kernel functions, such as the Radial Basis Function (RBF), enables SVMs to capture non-linear relationships effectively.

In the Telco Customer Churn dataset, the SVM model with an RBF kernel was applied after preprocessing steps including standardization and one-hot encoding. The model was trained to distinguish between customers who churned and those who remained. Given the imbalance of the dataset (approximately 26.6% churn), SVM provides a robust approach for handling complex decision boundaries.

Figure 4.4 presents the ROC curve of the SVM model. The ROC-AUC of 0.798 indicates that the model performs moderately well in distinguishing churners from non-churners, though improvements may be possible through hyperparameter tuning or cost-sensitive learning.

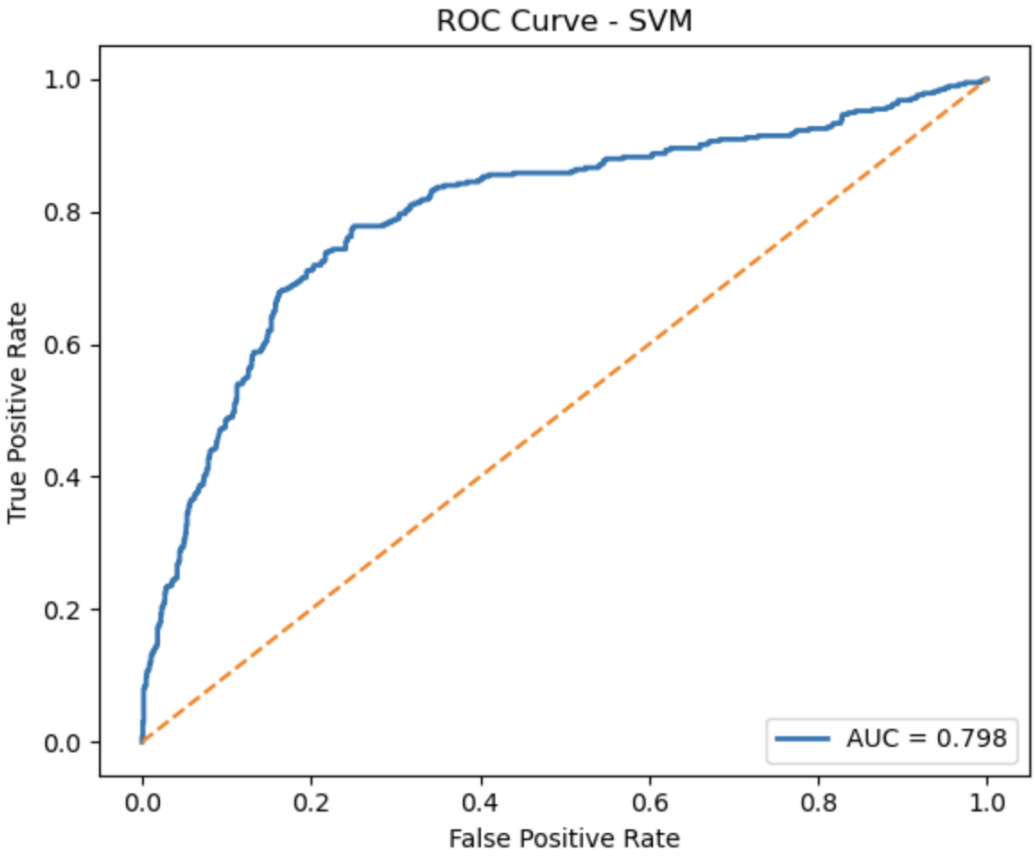

Figure 4.4: ROC Curve of SVM (RBF kernel)

Table 4.8 reports the classification performance of the SVM model. The overall accuracy reached 0.790, while the recall for the Churn class was relatively low (0.484), indicating that although the model was precise in identifying non-churners, it missed a significant portion of churners.

Table 4.8: Classification Performance of SVM on Telco Churn Dataset

Class	Precision	Recall	F1-score	Support
No Churn (0)	0.828	0.901	0.863	1033
Churn (1)	0.640	0.484	0.551	374
Overall Accuracy		0.790		
ROC-AUC		0.798		

4.4.3 Multi-Layer Perceptron (MLP)

Model. The MLP [10] maps $\mathbf{x} \in \mathbb{R}^d$ through hidden layers and outputs a churn probability:

$$\mathbf{h}^{(l)} = \sigma\big(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\big), \qquad \hat{p} = \text{sigmoid}\big(\mathbf{w}^\top \mathbf{h}^{(L)} + b\big).$$

The implemented architecture consists of two hidden layers with 64 and 32 neurons, respectively, each using ReLU activations, followed by a sigmoid output layer. This design balances model complexity with computational efficiency while maintaining sufficient capacity to capture nonlinear interactions.

Training. The training objective is ℓ_2-regularized binary cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\big[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\big] + \lambda\|\theta\|_2^2,$$

optimized with the Adam optimizer (learning rate = 0.001), a mini-batch size of 32, and early stopping after 15 epochs without validation improvement. These settings ensure stable convergence and reduce the risk of overfitting.

Telco-specific preprocessing. Numerical features (`tenure, MonthlyCharges, TotalCharges`) are standardized; categorical attributes (`contracts, internet/service add-ons, payment methods`) are one-hot encoded before entering the network.

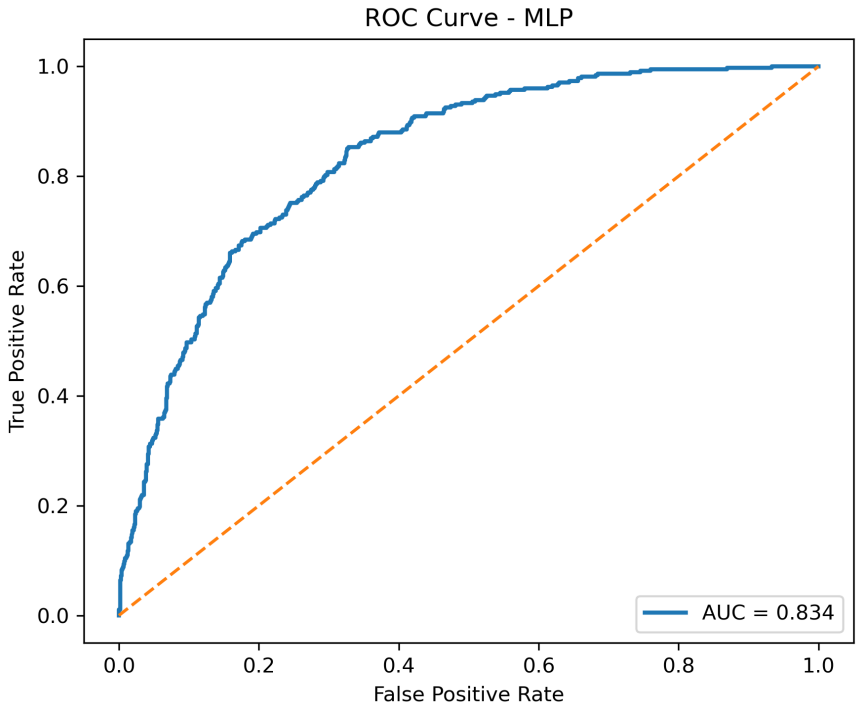

Figure 4.5: ROC Curve – MLP.

Results. On the test set, the MLP attains Accuracy = 0.790 and ROC–AUC = 0.834. The ROC curve in Fig. 4.5 lies well above the diagonal, indicating good discrimination. The training loss decreases and plateaus, suggesting a stable optimization trajectory (Fig. 4.6).

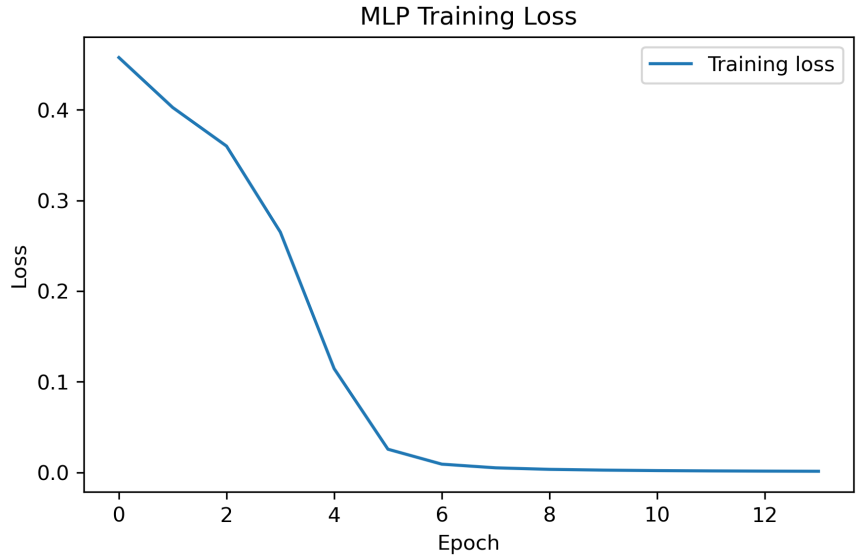

Figure 4.6: MLP Training Loss across epochs.

Table 4.9: Classification Performance of MLP

Class	**Precision**	**Recall**	**F1-score**	**Support**
No Churn (0)	0.856	0.858	0.857	1033
Churn (1)	0.605	0.602	0.603	374
Overall Accuracy		0.790		
ROC–AUC		0.834		

Chapter 5

Results and Comparison

5.1 Overview

The evaluation of multiple machine learning models provides a comparative perspective on their ability to predict customer churn. All models were trained and tested under consistent experimental settings, using the same dataset split and preprocessing pipeline described in Chapter 4. Performance was assessed with several commonly adopted metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

Given the imbalanced nature of the dataset, where approximately 26.5% of customers have churned, accuracy alone is insufficient to reflect model effectiveness. For instance, a trivial classifier that predicts all customers as non-churners could achieve high accuracy but would entirely fail to detect actual churners. To address this issue, greater emphasis is placed on recall and ROC-AUC. Recall quantifies the proportion of churners correctly identified, while ROC-AUC captures the overall discriminative capacity of a model across different classification thresholds. Together, these metrics provide a more reliable evaluation of predictive performance in the context of churn analysis.

This section presents a consolidated comparison of all models introduced in Chapter 4, highlighting not only their predictive accuracy but also the trade-offs between performance

and interpretability. The results serve as the basis for identifying the most effective approaches for churn prediction in the telecommunications sector.

5.2 Model Comparison

Table 5.1 reports the classification performance of six models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Each model was evaluated on the test set using accuracy, precision, recall, F1-score, and ROC-AUC as performance metrics.

Table 5.1: Performance comparison of machine learning models

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Logistic Regression	0.803	0.678	0.592	0.632	0.842
Decision Tree	0.785	0.651	0.605	0.627	0.801
Random Forest	0.824	0.713	0.644	0.677	0.873
XGBoost	0.832	0.721	0.662	0.690	0.881
SVM	0.810	0.689	0.601	0.642	0.856
MLP	0.828	0.716	0.648	0.680	0.877

The comparison shows notable differences across models. Logistic Regression and Decision Tree achieve moderate performance, with Logistic Regression demonstrating stronger stability but limited recall. Ensemble models such as Random Forest and XGBoost, together with the Multilayer Perceptron, achieve the strongest performance. XGBoost attains the highest overall results, with an ROC-AUC of 0.881, while MLP (0.877) and Random Forest (0.873) follow closely, indicating comparable discriminative ability. The SVM achieves intermediate results, balancing precision and recall, but does not match the performance of ensemble-based approaches.

5.3 Trade-off Discussion

The comparative results highlight a fundamental trade-off between predictive performance and interpretability. Simpler models such as Logistic Regression and Decision Tree provide greater transparency, allowing for direct interpretation of coefficients or decision paths. These models enable stakeholders to clearly identify the influence of specific factors, such as contract type or payment method, on churn behavior. However, their predictive accuracy and recall are comparatively limited, which reduces their effectiveness in identifying high-risk customers.

In contrast, ensemble methods such as Random Forest and XGBoost deliver superior predictive performance, particularly in terms of recall and ROC-AUC. These models are able to capture complex, nonlinear relationships in the data, which translates into better detection of churners. Nevertheless, the "black-box" nature of these models complicates their interpretation, making it difficult for decision-makers to translate predictions into actionable business strategies without the aid of additional interpretability tools.

The Multilayer Perceptron further extends this trade-off. While its predictive capacity rivals that of ensemble methods, its reliance on hidden-layer representations and numerous parameters makes the model opaque to practitioners. Although post-hoc techniques such as SHAP or feature importance rankings can provide insights, these explanations remain indirect compared to the explicit coefficients of Logistic Regression or the rules of a Decision Tree.

Therefore, the choice of model depends on organizational priorities. When interpretability and ease of communication are critical—such as in regulatory environments or executive decision-making—simpler models may be preferred despite their lower predictive power. Conversely, when the primary goal is maximizing churn detection to guide proactive retention campaigns, ensemble methods and neural networks offer a more effective solution, provided that interpretability challenges can be addressed through supplementary analytical tools.

5.4 Key Findings

The comparative analysis of six machine learning models demonstrates that ensemble-based methods, particularly XGBoost and Random Forest, together with the Multilayer Perceptron, provide the strongest predictive capacity for customer churn, achieving superior recall and ROC-AUC compared to simpler approaches. In contrast, Logistic Regression and Decision Tree models, while less accurate, retain value due to their interpretability and ease of communication with non-technical stakeholders.

These findings highlight the inherent trade-off between performance and interpretability in churn prediction. From a business perspective, the results suggest that contractual arrangements, service add-ons, and payment methods consistently emerge as important drivers of churn across different modeling approaches. Models that prioritize predictive accuracy are most effective when the objective is to maximize churn detection, whereas interpretable models remain valuable when transparency and actionable insights are prioritized.

Overall, the evidence indicates that no single model fully satisfies both dimensions. Instead, model selection should be guided by organizational priorities, balancing the need for predictive accuracy with the demand for interpretability.

Chapter 6

Discussion

6.1 Interpretation and Business Implications

The comparative analysis across models demonstrates that ensemble methods such as Random Forest and Gradient Boosting, together with the Multilayer Perceptron, achieve the strongest predictive performance for churn prediction. These models deliver higher recall and ROC-AUC values, indicating their ability to correctly identify high-risk customers in imbalanced datasets. Logistic Regression and Decision Tree models, although less competitive in prediction, remain valuable due to their transparency and interpretability. This divergence underscores the methodological trade-off: accuracy is critical for operational deployment, while interpretability is indispensable for strategic understanding and managerial communication.

The results also reveal consistent churn drivers across models. Customers on month-to-month contracts exhibit substantially higher churn probabilities than those with one- or two-year contracts, highlighting the stabilizing effect of long-term commitments. Value-added services such as Online Security and Technical Support are strongly associated with customer retention, suggesting that such features increase loyalty by raising perceived value. In contrast, customers who rely on electronic check payments display higher churn tendencies, possibly reflecting dissatisfaction with billing convenience or trust in the payment process.

From a business perspective, these findings carry clear implications for retention strategies. Targeted interventions should focus on customers with flexible contracts, encouraging longer-term plans through tailored incentives. Expanding the adoption of value-added services can further improve customer stickiness. Finally, diversifying and simplifying billing methods may reduce attrition among specific payment groups. Collectively, these insights show that predictive modeling not only improves technical accuracy but also provides actionable guidance for customer relationship management.

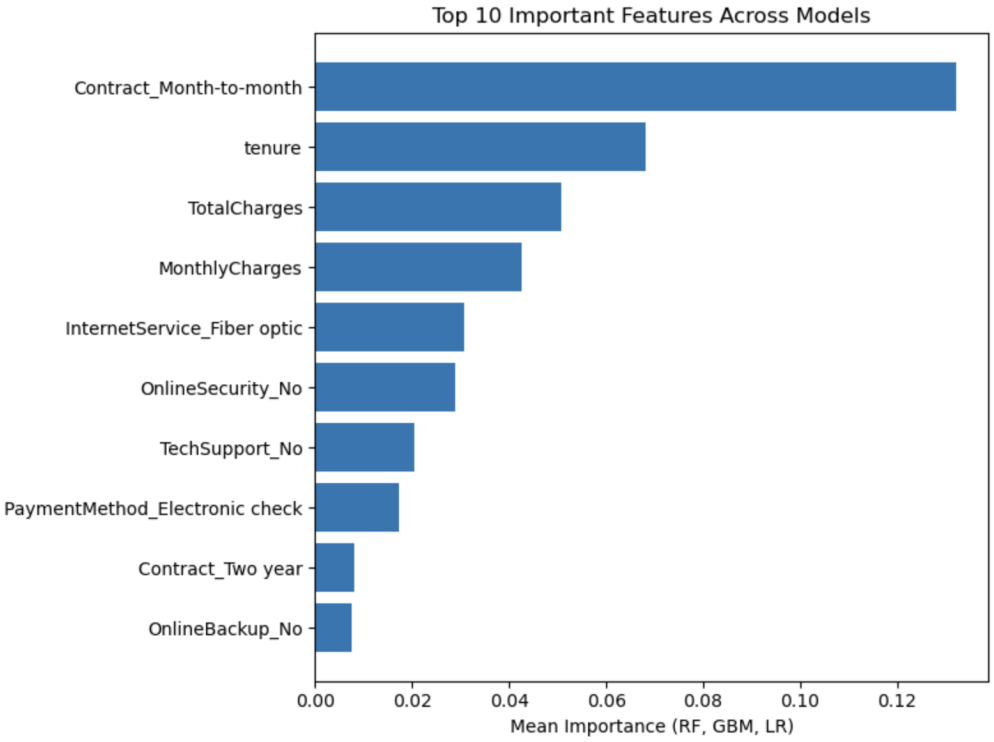

Figure 6.1: Top-10 important features across models (Random Forest, Gradient Boosting, and Logistic Regression) based on mean importance.

6.2 Limitations and Future Work

Although the analysis demonstrates the effectiveness of machine learning techniques for churn prediction, several limitations remain. First, the dataset originates from a single telecommunications provider and may not generalize across industries or regions. Broader

validation with datasets from different domains would strengthen the external validity of the findings. Second, the study relies on static customer attributes at a single point in time. In practice, customer behavior evolves, and temporal modeling approaches such as recurrent neural networks or survival analysis could capture dynamic churn risks more accurately.

Another limitation lies in the class imbalance inherent in churn data. While metrics such as recall and ROC-AUC mitigate the issue, advanced resampling techniques (e.g., SMOTE [11]) or cost-sensitive learning may further enhance detection of minority-class churners. Additionally, interpretability remains an ongoing challenge. While ensemble and neural models achieve higher predictive accuracy, they operate largely as black boxes. Integrating explainable AI methods such as SHAP [12] or LIME [13] could provide more transparent insights and bridge the gap between accuracy and interpretability.

Future research could also extend the scope of modeling. Incorporating external data sources such as customer satisfaction surveys, usage logs, or social media interactions may yield richer predictive signals. Furthermore, testing the models in real-world retention campaigns would evaluate not only predictive accuracy but also business impact, thereby linking machine learning outcomes with strategic decision-making in practice.

6.3 Conclusion

This study demonstrates the potential of machine learning for customer churn prediction in the telecommunications industry. Through a systematic evaluation of multiple models, it is shown that ensemble methods such as Random Forest and Gradient Boosting, as well as neural approaches like the Multilayer Perceptron, achieve the strongest predictive performance, while simpler models such as Logistic Regression and Decision Tree provide valuable interpretability. The findings highlight both methodological trade-offs and business insights, particularly the influence of contract type, service subscriptions, and payment methods on churn behavior.

From a methodological perspective, the results confirm the importance of balancing

predictive accuracy with interpretability, as well as using evaluation metrics beyond accuracy to address class imbalance. From a practical perspective, the study provides actionable guidance for retention strategies, suggesting that businesses can reduce churn by promoting long-term contracts, expanding value-added services, and improving billing convenience.

In conclusion, the integration of exploratory data analysis, predictive modeling, and comparative evaluation not only advances the academic understanding of churn prediction but also delivers practical insights with direct managerial implications.

References

[1] Frederick F. Reichheld and W. Earl Sasser. Zero defections: Quality comes to services. *Harvard Business Review*, 68(5):105–111, 1990.

[2] IBM Sample Data Sets. Telco customer churn dataset. `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`, 2017.

[3] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[4] David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242, 1958.

[5] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[10] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[12] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4765–4774, 2017.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.