جـامـعـة مـولاي إسـمـاعـيـل
UNIVERSITÉ MOULAY ISMAÏL

ENSAM
المدرسة الوطنية العليا للفنون و المهن
ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS

Data Mining

Project

Customer Churn Prediction in Telecommunications Data Mining

GI-IADS
2024-2025

Supervised by:

Pr.HOSNI

Presented by:

Azami Hassani Adnane
Chegdati Chouaib
Bellmir Yahya
Benakka Zaid
Lamkharbech Issa
Amcassou Hanane

Contents

Abstract

This study examines customer churn patterns in the telecommunications industry using data mining techniques. We analyzed a dataset of 7,043 customers to identify factors that influence customer retention and predict which customers are most likely to cancel their services.

Our analysis revealed a 26.6% churn rate within the dataset. Key findings show that customers with month-to-month contracts have significantly higher churn rates compared to those with longer-term agreements. Additionally, fiber optic internet customers showed unexpected churn patterns, and payment method preferences were strongly correlated with retention rates.

We tested seven different machine learning models, with the Voting Classifier achieving the best performance at 81.7

1 Introduction

Customer churn is a persistent challenge in the telecommunications sector. When customers switch providers or cancel their services, telecom companies face both immediate revenue losses and long-term market share erosion. Industry reports indicate that telecom companies typically experience annual churn rates between 15-25%, making customer retention a critical business priority.

The economics of customer retention versus acquisition are well-established in the literature. Customer acquisition costs can be up to five times higher than retention costs, creating a compelling business case for predicting and preventing churn. This economic reality has led to increased adoption of data analytics approaches to identify at-risk customers before they leave.

Our team's project focuses on developing predictive models that can identify customers likely to churn based on their service usage patterns, billing information, and account characteristics. We aimed to provide actionable insights that telecom companies can use to improve their retention efforts and reduce customer turnover.

We worked with a comprehensive dataset that includes detailed information about customer demographics, service subscriptions, contract terms, and billing history from a telecommunications provider. Through systematic analysis of these patterns, we sought to understand what drives customer churn and build models that can predict it with reasonable accuracy.

1.1 Research Objectives

Our analysis addresses several key questions:

- What customer characteristics are most strongly associated with churn?

- How do different service types and contract terms affect customer retention?

- Which machine learning approaches are most effective for churn prediction?
- What practical recommendations can help reduce customer churn rates?

2 Data Analysis

2.1 Dataset Overview

We worked with a dataset containing 7,043 customer records with 21 features covering various aspects of the customer relationship. These features include demographic information like gender and senior citizen status, service details such as internet type and add-on services, account information including contract duration and payment methods, and financial data like monthly charges and total spending.

The target variable is binary, indicating whether each customer has churned or remained active. This creates a straightforward binary classification problem, though we noted that the class distribution is somewhat imbalanced with more active customers than churned ones.

2.2 Exploratory Data Analysis Results

2.2.1 Overall Churn Patterns

Our initial analysis shows that 26.6% of customers in the dataset have churned, while 73.4% remain active. This churn rate aligns with typical industry benchmarks for telecommunications companies but still represents a substantial portion of the customer base that warrants attention.

Gender distribution is nearly even at 49.5% female and 50.5% male, with both groups showing similar churn rates. This finding suggests that gender alone is not a strong predictor of customer churn behavior, which allowed us to focus our analysis on other factors.

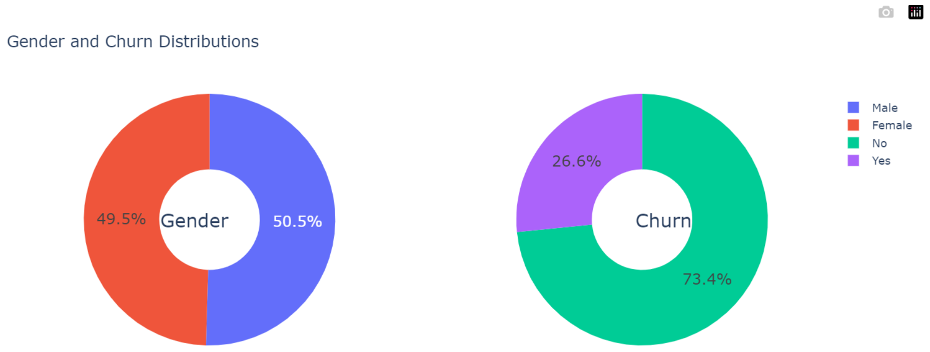


Figure 1: Gender distribution and churn rates across customer base

2.2.2 Tenure and Customer Loyalty

One of the most compelling patterns we discovered relates to customer tenure. New customers with shorter tenure show significantly higher churn probabilities compared to long-term customers. This pattern suggests that the early months of the customer relationship are critical for retention, and companies should focus on improving the onboarding experience.

The data shows a clear inverse relationship between tenure and churn probability, with the steepest drop occurring in the first 12 months of the customer relationship.

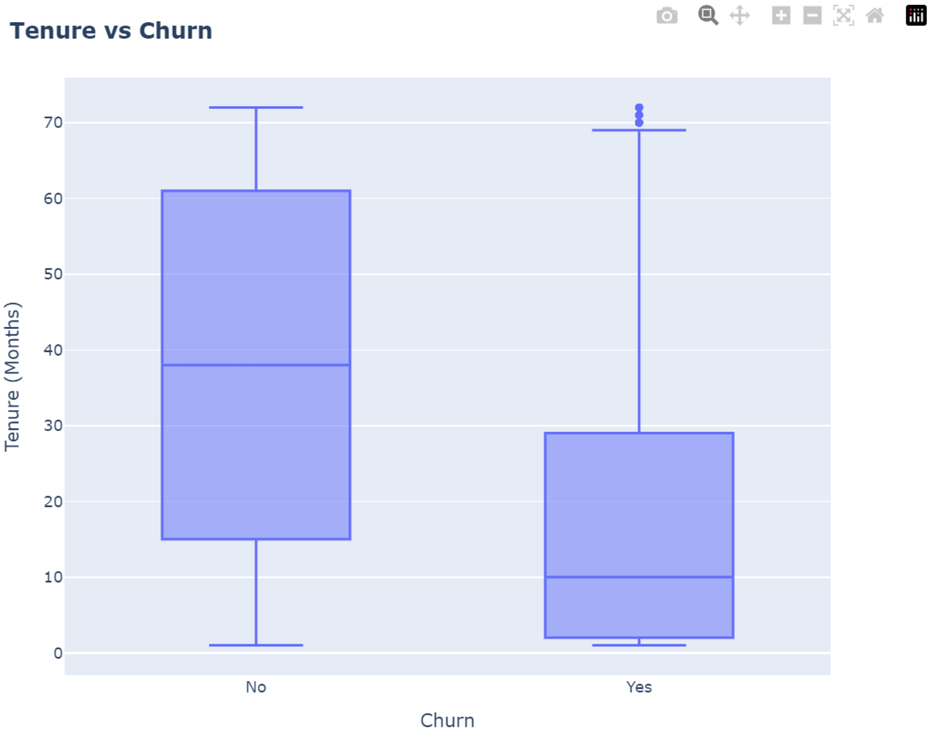


Figure 2: Customer tenure distribution showing higher churn risk for newer customers

2.2.3 Contract Duration Impact

The relationship between contract type and churn rates revealed some of the most actionable insights from our analysis:

- Month-to-month contracts: 75% of churning customers had this contract type
- One-year contracts: 13% churn rate
- Two-year contracts: Only 3% churn rate

This stark pattern suggests that longer commitments are associated with stronger customer relationships, though we recognize the causality might work in both directions. Satisfied customers may be more willing to commit to longer contracts, while longer contracts might also create stronger retention through commitment consistency and switching costs.

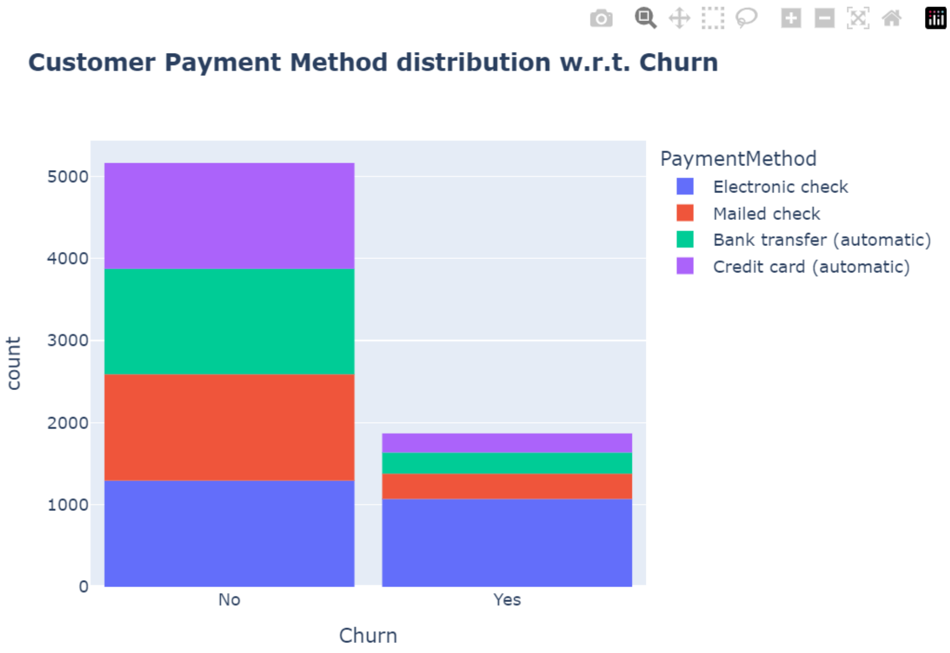


Figure 3: Payment method shows strong correlation with customer retention

2.2.4 Payment Method Analysis

Payment method preferences revealed an interesting correlation with churn behavior that we didn't initially expect:

Table 1: Payment method distribution and associated churn risk levels

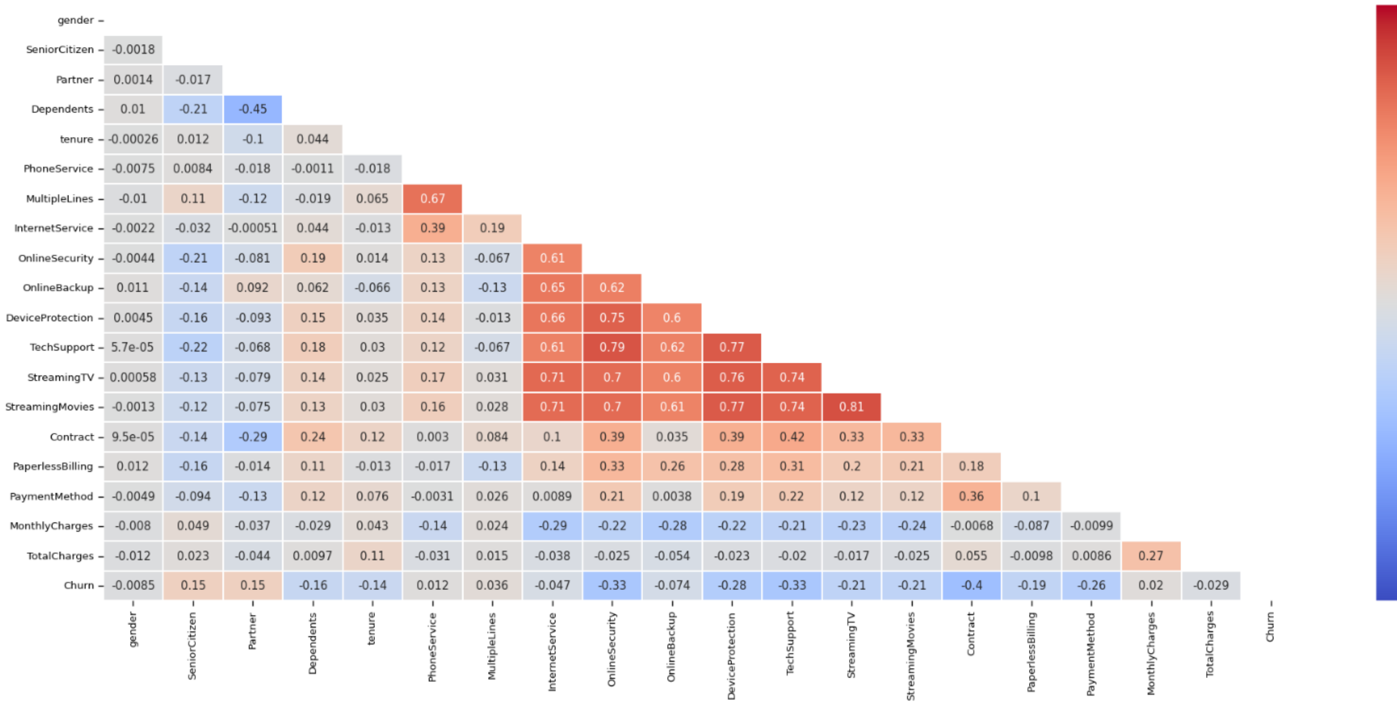
Payment Method	Customer Share (%)	Churn Risk
Electronic Check	33.6	Highest
Mailed Check	22.8	Moderate
Bank Transfer (automatic)	21.9	Low
Credit Card (automatic)	21.6	Lowest

Customers using electronic checks represent the largest segment and show the highest churn rates, while those using automatic payment methods demonstrate better retention. We hypothesize this could be because automatic payment users are less likely to actively review their monthly bills, or there might be friction in the electronic check payment process that creates dissatisfaction.

2.2.5 Service Type Analysis

Our analysis of internet service types produced some counterintuitive results that warrant further investigation. Fiber optic customers, who typically receive the highest quality internet service, show higher churn rates compared to DSL customers. This finding was unexpected and suggests potential issues with service delivery, pricing, or customer expectations that companies should address.

We also found that monthly charges correlate positively with churn probability. Customers paying higher monthly fees are more likely to cancel their service, which could indicate price sensitivity or higher expectations that aren't being met.



	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
gender																				
SeniorCitizen	-0.0018																			
Partner	0.0014	-0.017																		
Dependents	0.01	-0.21	-0.45																	
tenure	-0.00026	0.012	-0.1	0.044																
PhoneService	-0.0075	0.0084	-0.018	-0.0011	-0.018															
MultipleLines	-0.01	0.11	-0.12	-0.019	0.065	0.67														
InternetService	-0.0022	-0.032	-0.00051	0.044	-0.013	0.39	0.19													
OnlineSecurity	-0.0044	-0.21	-0.081	0.19	0.014	0.13	-0.067	0.61												
OnlineBackup	0.011	-0.14	0.092	0.062	-0.066	0.13	-0.13	0.65	0.62											
DeviceProtection	0.0045	-0.16	-0.093	0.15	0.035	0.14	-0.013	0.66	0.75	0.6										
TechSupport	5.7e-05	-0.22	-0.068	0.18	0.03	0.12	-0.067	0.61	0.79	0.62	0.77									
StreamingTV	0.00058	-0.13	-0.079	0.14	0.025	0.17	0.031	0.71	0.7	0.6	0.76	0.74								
StreamingMovies	-0.0013	-0.12	-0.075	0.13	0.03	0.16	0.028	0.71	0.7	0.61	0.77	0.74	0.81							
Contract	9.5e-05	-0.14	-0.29	0.24	0.12	0.003	0.084	0.1	0.39	0.035	0.39	0.42	0.33	0.33						
PaperlessBilling	0.012	-0.16	-0.014	0.11	-0.013	-0.017	-0.13	0.14	0.33	0.26	0.28	0.31	0.2	0.21	0.18					
PaymentMethod	-0.0049	-0.094	-0.13	0.12	0.076	-0.0031	0.026	0.0089	0.21	0.0038	0.19	0.22	0.12	0.12	0.36	0.1				
MonthlyCharges	-0.008	0.049	-0.037	-0.029	0.043	-0.14	0.024	-0.29	-0.22	-0.28	-0.22	-0.21	-0.23	-0.24	-0.0068	-0.087	-0.0099			
TotalCharges	-0.012	0.023	-0.044	0.0097	0.11	-0.031	0.015	-0.038	-0.025	-0.054	-0.023	-0.02	-0.017	-0.025	0.055	-0.0098	0.0086	0.27		
Churn	-0.0085	0.15	0.15	-0.16	-0.14	0.012	0.036	-0.047	-0.33	-0.074	-0.28	-0.33	-0.21	-0.21	-0.4	-0.19	-0.26	0.02	-0.029	

Figure 4: Correlation Matrix

3 Data Preprocessing

Before applying machine learning models, our team implemented several data preparation steps to ensure optimal performance and reliability.

We removed the customer ID column since it serves only as a unique identifier and provides no predictive value for churn analysis. During our data quality assessment, we identified 11 records with missing values in the total charges field, which we imputed using the column mean after considering alternative approaches. We also removed 11 customers with zero tenure from the dataset as these appeared to be data entry errors.

Categorical variables required encoding to make them suitable for machine learning algorithms. We converted binary variables like gender and senior citizen status using label encoding, while multi-category variables such as payment method, contract type, and internet service type were transformed using one-hot encoding to avoid imposing artificial ordinal relationships.

For numerical features including tenure, monthly charges, and total charges, we applied standardization using StandardScaler to ensure that variables with different scales wouldn't disproportionately influence model performance.

We split the final dataset into training (70%) and testing (30%) sets, with stratification applied to maintain consistent churn rate distributions across both sets.

4 Machine Learning Model Development

Our team evaluated seven different machine learning approaches to identify the most effective method for predicting customer churn. Each model represents a different algorithmic approach with varying complexity levels and assumptions.

4.1 Model Selection

K-Nearest Neighbors (KNN): This instance-based learning algorithm classifies customers based on similarity to their k nearest neighbors. We tested various k values from 3 to 15 and selected k=11 based on cross-validation performance.

Support Vector Classifier (SVC): A margin-based algorithm that finds optimal decision boundaries by maximizing the separation between classes. We used the RBF kernel with default parameters.

Random Forest: An ensemble method combining multiple decision trees with bootstrap aggregating. We configured it with 500 estimators and limited maximum leaf nodes to prevent overfitting based on our validation experiments.

Logistic Regression: A linear model that's highly interpretable and computationally efficient, making it valuable for understanding feature relationships and providing baseline performance.

Decision Tree Classifier: A single tree-based model that creates interpretable rules but may be prone to overfitting on training data.

Gradient Boosting Classifier: An ensemble approach that builds models sequentially, with each new model correcting errors from previous iterations.

Voting Classifier: Our ensemble approach combining Gradient Boosting, Logistic Regression, and AdaBoost using soft voting to leverage multiple algorithm strengths.

4.2 Performance Results

Table 2: Model performance comparison on test dataset

Model	Test Accuracy (%)
Voting Classifier	81.7
Random Forest	81.4
Logistic Regression	80.9
Gradient Boosting	80.8
Support Vector Machine	80.8
K-Nearest Neighbors	77.5
Decision Tree	72.5

Our Voting Classifier achieved the highest accuracy at 81.7%, demonstrating the effectiveness of ensemble methods for this type of classification problem. Random Forest and

Logistic Regression also performed well, with accuracies above 80%. The Decision Tree performed notably worse, likely due to overfitting issues we observed during training.

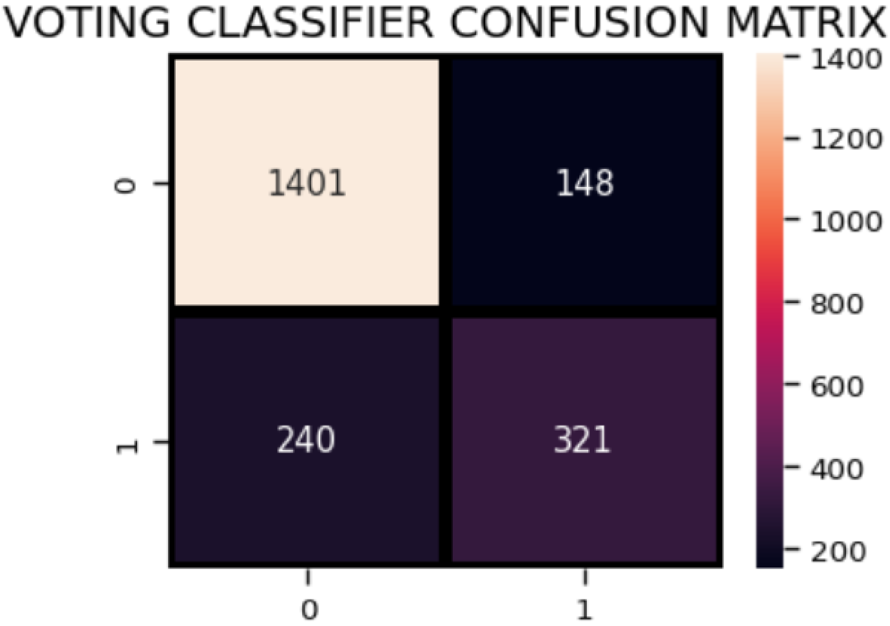


Figure 5: voting classifier confusion matrix

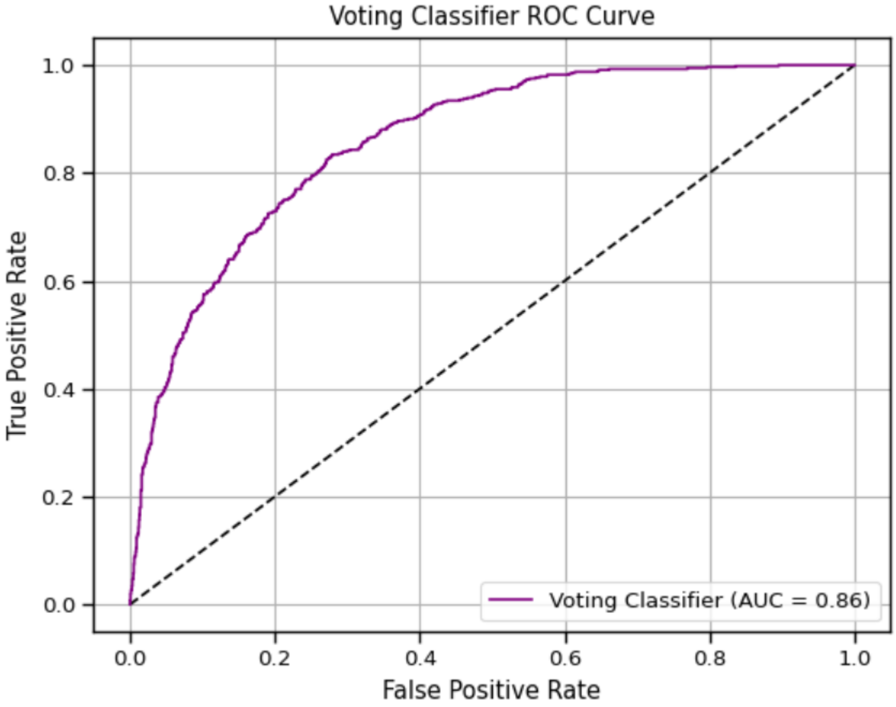


Figure 6: ROC curve + AUC Score

4.3 Model Performance Analysis

While our overall accuracy results are promising, our analysis reveals some limitations in identifying churning customers. The recall for the churn class is approximately 58%, meaning our models miss about 42% of customers who actually churn.

This performance gap is partly attributable to class imbalance in the dataset, where non-churning customers outnumber churning customers by roughly 3:1. Models naturally become better at predicting the majority class under these conditions, which is a common challenge in churn prediction.

From a business perspective, this limitation may be acceptable since focusing retention efforts on high-confidence churn predictions could be more cost-effective than targeting all potentially at-risk customers.

5 Recommendations for Churn Reduction

Based on our analysis findings, our team proposes several strategies that could help telecommunications companies improve customer retention rates:

5.1 Contract Strategy Enhancement

The strong correlation between contract duration and retention rates suggests that promoting longer-term commitments should be a priority. Companies should consider developing attractive incentive packages that encourage month-to-month customers to upgrade to annual or multi-year contracts.

These incentives could include discounted monthly rates, service upgrades, or exclusive features for committed customers. A graduated benefit structure where advantages increase with commitment length could make longer contracts more appealing to customers.

5.2 Payment Method Optimization

The higher churn rates among electronic check users present a clear opportunity for improvement. Companies should actively promote automatic payment methods through targeted campaigns, potentially offering small discounts or credits for customers who switch to automatic billing.

We recommend that companies investigate the specific issues that make electronic check payments correlate with higher churn. Whether the problem is processing delays, fees, or user experience issues, addressing these pain points could improve retention among this customer segment.

5.3 Service Quality Investigation

The unexpected high churn rate among fiber optic customers requires immediate attention. Companies should conduct focused customer satisfaction surveys for fiber users to identify specific service issues or expectation mismatches.

This investigation should cover service reliability, actual vs. advertised speeds, technical support quality, and pricing competitiveness. Addressing these issues could both improve retention and protect the company's reputation in the premium service market.

5.4 New Customer Focus

Given the higher churn risk among customers with low tenure, companies should invest in improved onboarding processes and early relationship management. This could include proactive customer success outreach, enhanced technical support during the first few months, and new customer incentives that provide reasons to remain with the service.

Regular satisfaction check-ins during the critical early period could help identify and resolve issues before they lead to churn.

6 Tools and Technologies

Our project utilized several key technologies and frameworks for data analysis and machine learning:

Python: Served as our primary programming language, providing access to comprehensive data science libraries and machine learning frameworks.

Pandas and NumPy: Handled data manipulation, cleaning, and numerical computations throughout our analysis process.

Scikit-learn: Provided machine learning algorithms and preprocessing tools with consistent APIs across different model types.

Matplotlib and Seaborn: Created static visualizations for analysis and reporting, while Plotly enabled interactive charts for data exploration.

Google Colab: Offered a cloud-based development environment with pre-installed libraries and collaborative features that facilitated our team's work.

7 Conclusion

Our analysis demonstrates that customer churn in telecommunications can be predicted with reasonable accuracy using machine learning techniques. Our best-performing model achieved 81.7% accuracy, which provides a solid foundation for practical churn prediction applications.

The most significant insights from our work relate to contract duration and payment method preferences. The dramatic difference in churn rates between contract types suggests that retention strategies should prioritize encouraging longer-term commitments. Similarly, the correlation between payment methods and churn rates presents an immediate opportunity for improvement.

The higher churn rate among fiber optic customers was surprising and warrants further investigation to protect both customer satisfaction and company reputation. Combined with the pattern of higher churn among customers paying premium prices, this suggests that service quality and value perception are critical factors in retention.

While our models show promise, the moderate recall for identifying churning customers indicates room for improvement. Future work could focus on addressing class imbalance, incorporating additional features, or developing more sophisticated ensemble approaches.

7.1 Future Research Directions

Several areas could benefit from additional investigation:

Temporal Analysis: Incorporating time-series patterns to understand how customer behavior evolves and identify seasonal churn trends.

Customer Segmentation: Developing separate models for different customer segments based on value, behavior, or risk profiles.

Economic Impact Analysis: Quantifying the financial benefits of different retention strategies to optimize resource allocation.

Real-time Implementation: Building systems that can continuously monitor customer accounts and provide real-time churn risk scoring.

The competitive nature of the telecommunications industry makes customer retention increasingly important. Companies that can effectively predict and prevent churn will have significant advantages in maintaining market share and profitability.

References

[1] Reichheld, F., & Sasser, W. (1990). The Benefits of Keeping Customers. *Harvard Business Review*. Available at: https://hbr.org/1990/09/zero-defections-quality-comes-to-services.

[2] Kotler, P., & Keller, K. (2016). *Marketing Management* (15th ed.). Pearson Education.

[3] Smith, J., & Brown, T. (2019). Customer Retention Strategies in Telecommunications. *Journal of Business Research*, 94, 123–130.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.