Predicting Customer Behavior using Machine Learning

Final Report

Syed Abdullah Quadri – 60104641

Shaimah Mohammed - 60104699

Mohammed Hajji - 60103737

Mumin Almaghrabi – 60096897

University of Doha for Science and Technology

Machine Learning – DSAI3201 – 1

Dr. Karima Makhlouf

June 23, 2025

Table of Contents

1. Introduction

Customer churn is a major concern for companies, especially in highly competitive industries such as telecommunications. Churn occurs when a customer decides to terminate their subscription or stop using the service, directly impacting the company's revenue and growth.

The objective of this project is to build a predictive model using machine learning techniques to identify customers who are likely to churn. Early identification of these customers allows telecom companies to intervene and implement strategies to retain them.

To achieve this, we employ data-driven methods, including data preprocessing, visualization, feature engineering, and the training of multiple machine learning models. We compare their performance and interpret the results to provide actionable business insights.

2. Dataset

The dataset used in this project is a publicly available dataset representing a sample of customers from a fictional telecommunications company. It contains 7043 rows and 21 columns with details such as:

- CustomerID: Unique ID for each customer
- Gender, SeniorCitizen, Partner, Dependents: Demographic data
- Tenure, PhoneService, InternetService, Contract: Service usage information
- MonthlyCharges, TotalCharges: Billing information
- Churn: Target variable (Yes/No)

2.1 Preprocessing Steps

- Removed rows with missing or blank values (especially in TotalCharges).
- Converted categorical variables using label encoding and one-hot encoding.
- Standardized numerical features using StandardScaler for better model convergence.

2.2 Summary Statistics and Visualizations

Descriptive statistics showed that approximately 26.5% of the customers had churned. Visual analysis such as bar plots showed strong churn correlation with contract type (Month-to-Month), tenure (less than 1 year), and presence of tech support. A heatmap of correlations revealed key predictive features, guiding our feature selection process.

3. Methodology

Our methodology involves building and evaluating multiple classification models for predicting customer churn. The process is divided into several stages:

3.1 Data Preparation

The features were carefully analyzed and transformed. Categorical data such as 'InternetService' or 'Contract' were converted using OneHotEncoding to avoid ordinality assumptions. Numerical data were standardized to mean zero and unit variance using StandardScaler, which is particularly important for models sensitive to feature scales like SVM and Logistic Regression.

3.2 Model Selection

We tested the following models:

- Logistic Regression: A baseline linear classifier used for binary classification.
- Random Forest: An ensemble of decision trees trained on random subsets, reducing overfitting.
- Support Vector Machine (SVM): Effective in high-dimensional spaces, though slower to train.
- XGBoost: Gradient boosting decision trees offering high accuracy and built-in handling of missing values.

3.3 Feature Engineering & Selection

We used correlation analysis and feature importance from tree-based models to reduce the feature set and remove noise. Important features included:

- Contract type
- Monthly charges
- Tenure
- Tech support availability
- Internet service type

3.4 Hyperparameter Tuning

For each model, we performed grid search with 5-fold cross-validation to find optimal parameters:

- Logistic Regression: regularization strength
- Random Forest: number of trees, max depth
- SVM: kernel type, C value
- XGBoost: learning rate, max depth, number of estimators

4. Model Evaluation & Comparison

We applied stratified 5-fold cross-validation to ensure balanced class distribution. The following metrics were used to compare models:

- Accuracy: Measures overall correct predictions.
- Precision: Correct positive predictions over total positive predictions.
- Recall: Correct positive predictions over actual positives.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC: Captures trade-off between true positive and false positive rates.

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8	0.73	0.66	0.69	0.83
Random Forest	0.84	0.78	0.75	0.76	0.88
SVM	0.82	0.74	0.72	0.73	0.85
XGBoost	0.85	0.8	0.82	0.81	0.9

Visuals such as confusion matrices and ROC curves helped confirm XGBoost's strong recall performance.

5. Results & Discussion

After evaluating all models, XGBoost was chosen as the best-performing model for predicting customer churn. It achieved the highest accuracy, recall, and ROC-AUC scores.

5.1 Best Model Interpretation

XGBoost's ability to capture non-linear relationships and automatically handle feature interactions gave it an edge. It was especially good at identifying churners, which is crucial since recall is a priority in churn prediction.

- Important features ranked by XGBoost:
- Contract type (month-to-month contracts had the highest churn rate)
- Tenure (shorter tenure indicated higher churn likelihood)
- Internet service type (fiber optic users had higher churn)
- Tech support (absence of tech support increased churn)
- Monthly charges (higher bills correlated with higher churn)

5.2 Strengths and Limitations

Strengths of XGBoost:

- Robust performance with missing data handling
- Built-in regularization to avoid overfitting
- Strong predictive power

Limitations:

- Complex tuning process compared to simpler models
- Less interpretable than logistic regression for business audiences
- Higher training time and resource usage

5.3 Insights into Business Strategy

- Customers with month-to-month contracts and no tech support are at high risk—offering bundled contracts with added support may help retention.
- Customers with high monthly charges and low tenure could be targeted with personalized offers or discounts.
- Visual dashboards and alerts could help customer success teams act on predictions in real time.

6. Conclusion

In this project, we used machine learning techniques to predict telecom customer churn. Our process included data cleaning, visualization, feature engineering, and training multiple models.

XGBoost was the top performer, achieving strong results across all evaluation metrics. It also revealed critical insights into customer behavior, helping the company make data-driven decisions.

Key Takeaways:

- Predictive models can help reduce churn by enabling timely interventions.
- Feature importance analysis is essential for business strategy.
- Ensemble models like Random Forest and XGBoost outperform simpler models in complex datasets.

Future Improvements:

- Integrate additional data such as customer support interactions or sentiment analysis.
- Deploy the model into a real-time prediction system.
- Use SHAP values for deeper interpretability.
- Explore deep learning models if more data becomes available.

This work provides a strong foundation for telecom churn prediction systems and opens the door for future enhancement in customer analytics.