# PROJECT REPORT

# Customer Churn Prediction Using Random Forest: A Data-Driven Retention Strategy


## Abstract

Customer churn is a major concern for subscription-based businesses, as it directly impacts revenue and customer lifetime value. The objective of this project is to build a machine learning model that can predict whether a customer is likely to churn based on historical customer data. The project focuses on applying the Random Forest algorithm to capture complex patterns in customer behaviour.

The workflow includes understanding the dataset, performing exploratory data analysis, preprocessing and encoding features, building a baseline Random Forest model, improving performance through hyperparameter tuning, and finally optimizing the classification threshold to align with business requirements. The outcome of the project is a churn prediction model that prioritizes identifying high-risk customers and provides insights that can be used for customer retention strategies.


**Key Takeaways**

• End-to-end implementation of a real-world ML problem
• Business-focused evaluation rather than accuracy-only metrics
• Practical model optimization using tuning and threshold adjustment

## Problem Statement

Customer churn refers to customers discontinuing their service with a company. In industries such as telecommunications, customers have multiple alternatives, and churn can occur due to pricing, service quality, contract flexibility, or lack of customer support. Most companies identify churn only after it has already happened, making it difficult to take corrective actions.

The problem addressed in this project is to build a predictive system that can identify customers who are likely to churn in advance, using historical customer data. Such a system enables organizations to proactively engage with customers and reduce churn-related losses.

**Insights**

• Churn is predictable using historical behaviour
• Early identification enables proactive retention
• Prediction is more valuable than post-churn analysis

## Objectives

The main objectives of this project are:

• To understand customer behaviour using data analysis
• To identify important factors contributing to churn
• To build a Random Forest classification model
• To improve model performance using hyperparameter tuning
• To optimize the prediction threshold based on business needs
• To interpret model results for actionable insights

## Dataset Description

The dataset used in this project is the Telco Customer Churn dataset. Each row in the dataset represents a single customer, and each column describes customer demographics, service subscriptions, billing details, or payment methods. The dataset contains both categorical and numerical features, making it suitable for tree-based models.

The target variable is “Churn”, which indicates whether a customer has left the service. The dataset reflects real-world telecom customer behaviour and includes a moderate class imbalance, which influences model evaluation and metric selection.

### Target Variable

Churn is a binary variable defined as follows:

• Churn = 1 indicates the customer has churned
• Churn = 0 indicates the customer has not churned

This makes the task a binary classification problem where misclassifying churned customers has higher business cost than misclassifying non-churned customers.

| Column Name | Data Type | Description |
|---|---|---|
| `customerID` | Categorical | Unique identifier assigned to each customer |
| `gender` | Categorical | Gender of the customer (Male / Female) |
| `SeniorCitizen` | Binary (0/1) | Indicates whether the customer is a senior citizen |
| `Partner` | Categorical | Whether the customer has a partner (Yes / No) |
| `Dependents` | Categorical | Whether the customer has dependents (Yes / No) |
| `tenure` | Numerical | Number of months the customer has stayed with the company |
| `PhoneService` | Categorical | Indicates whether the customer has phone service |
| `MultipleLines` | Categorical | Indicates whether the customer has multiple phone lines |
| `InternetService` | Categorical | Type of internet service (DSL / Fiber optic / None) |
| `OnlineSecurity` | Categorical | Whether the customer has online security service |
| `OnlineBackup` | Categorical | Whether the customer has online backup service |
| `DeviceProtection` | Categorical | Whether the customer has device protection plan |

| | | |
|---|---|---|
| `TechSupport` | Categorical | Whether the customer has technical support service |
| `StreamingTV` | Categorical | Indicates whether the customer uses streaming TV |
| `StreamingMovies` | Categorical | Indicates whether the customer uses streaming movies |
| `Contract` | Categorical | Contract type (Month-to-month / One year / Two year) |
| `PaperlessBilling` | Categorical | Whether the customer uses paperless billing |
| `PaymentMethod` | Categorical | Payment method used by the customer |
| `MonthlyCharges` | Numerical | Monthly amount charged to the customer |
| `TotalCharges` | Numerical | Total amount charged to the customer over the tenure |
| `Churn` | Categorical (Target) | Indicates whether the customer churned (Yes / No) |

**Libraries and Tools Used**

The following Python libraries were used in the project:

• Pandas for data manipulation and analysis
• NumPy for numerical computations
• Matplotlib for data visualization
• Scikit-learn for preprocessing, modelling, and evaluation

## Exploratory Data Analysis

Exploratory Data Analysis was performed to understand data distribution, detect anomalies, and identify potential churn drivers. The analysis showed that churned customers form a smaller portion of the dataset, confirming class imbalance. Tenure analysis revealed that customers with lower tenure are more likely to churn, indicating weak early customer engagement.

Contract analysis showed that customers on month-to-month contracts have significantly higher churn compared to customers with longer contract durations. Monthly charges analysis indicated that higher billing amounts are associated with increased churn probability.

Visual representations such as EDA plots, confusion matrices, and feature importance graphs were included to improve interpretability and support data-driven insights.

**Summary**

- Churn rate is imbalanced
- Customers with short tenure churn more
- High monthly charges increase churn risk
- Month-to-month contracts show highest churn
- Long-term contracts improve retention

## Data Preprocessing and Feature Engineering

Data preprocessing involved removing non-predictive columns such as customer ID. Missing values were identified and handled to ensure data quality. The target variable was encoded using Label Encoding. Categorical features were converted into numerical form using one-hot encoding.

Feature scaling was not applied since Random Forest models are insensitive to feature magnitude. The dataset was split into training and testing sets using stratified sampling to preserve churn distribution.

**Insights**
• Clean data improves model reliability
• Encoding avoids ordinal bias
• Stratified split ensures fair evaluation

## Baseline Random Forest Model

A baseline Random Forest classifier was trained using default parameters. The model achieved good overall accuracy but showed relatively low recall for churned customers. This indicated that although the model performed well on non-churn customers, it failed to identify a significant number of churn cases.

This observation highlighted the limitation of accuracy as a standalone metric and justified further model optimization.

**Insights**
• Baseline establishes reference performance
• Accuracy alone is misleading
• Churn recall needs improvement

## Hyperparameter Tuning

Hyperparameter tuning was carried out using RandomizedSearchCV to explore multiple parameter combinations efficiently. The tuning process focused on improving recall for the churn class and controlling overfitting.

After tuning, the model showed a noticeable improvement in churn recall and achieved a ROC-AUC score of approximately 0.85, indicating strong separation between churned and non-churned customers.

**Insights**
• Tuning improves minority class detection

• ROC-AUC confirms model robustness
• Random Forest handles non-linear patterns well

## Threshold Optimization

By default, classification models use a probability threshold of 0.5. However, for churn prediction, missing churned customers is more costly than generating false alarms. Threshold optimization was applied by lowering the decision threshold.

This adjustment significantly improved recall for churned customers while maintaining acceptable precision. The optimized threshold better aligned the model with real-world retention objectives.

**Insights**
• Business cost drives threshold selection
• Lower threshold captures more churn
• Practical models outperform default settings

## Model Evaluation and Comparison

The tuned and threshold-optimized model was compared against the baseline model. Although overall accuracy decreased slightly, the optimized model correctly identified a significantly higher number of churned customers. The confusion matrix confirmed a reduction in false negatives, making the model more suitable for business use.

Insights
• Accuracy trade-off is acceptable
• Business value increased
• Optimized model is deployment-ready

## Feature Importance Interpretation

Feature importance analysis showed that tenure, contract type, monthly charges, total charges, and service-related features such as technical support and internet service type are the most influential predictors of churn.

Customers with short tenure, month-to-month contracts, and higher monthly charges were more likely to churn. These insights can be directly used to design targeted retention strategies.

Insights
• Model decisions are explainable
• Insights are actionable
• Supports strategic planning

## Final Insights

- Customers with shorter tenure exhibit a significantly higher likelihood of churn, highlighting the importance of strong onboarding and early engagement strategies.
- Month-to-month contract customers show the highest churn rates compared to long-term contract holders, indicating that contract duration plays a critical role in customer retention.
- Higher monthly charges are strongly associated with increased churn probability, suggesting that pricing and perceived value influence customer loyalty.
- Hyperparameter tuning and threshold optimization significantly improved the model's ability to identify churned customers, achieving high recall and a ROC-AUC of approximately 0.85.
- Feature importance analysis revealed that tenure, contract type, billing amount, and support-related services are the most influential factors driving churn.
- The optimized Random Forest model provides actionable business insights, enabling targeted retention campaigns and data-driven decision-making to reduce customer attrition.

## Conclusion

This project successfully demonstrates the application of machine learning techniques to address a real-world business problem of customer churn prediction in the telecommunications domain. By leveraging historical customer data and applying the Random Forest algorithm, the project delivers a predictive system capable of identifying customers who are at risk of discontinuing services before churn actually occurs.

The study began with a thorough exploration of the dataset to understand customer behaviour, data distribution, and churn patterns. Exploratory Data Analysis revealed that churn is not random but strongly influenced by customer tenure, contract type, billing amounts, and service-related factors. These insights guided the selection of features, evaluation metrics, and model optimization strategies, ensuring that the modelling approach remained grounded in business relevance.

A baseline Random Forest model was developed to establish initial performance benchmarks. While the baseline model achieved reasonable overall accuracy, it showed limitations in identifying churned customers, highlighting the inadequacy of accuracy as the sole evaluation metric for imbalanced classification problems. This motivated the use of hyperparameter tuning to enhance model performance, particularly with respect to churn recall.

Hyperparameter tuning using RandomizedSearchCV resulted in a significant improvement in the model's ability to detect churned customers and improved the overall discriminatory power, as reflected by a strong ROC-AUC score. To further align the model with real-world business

objectives, threshold optimization was applied. By adjusting the probability threshold, the model successfully reduced the number of missed churn cases, prioritizing customer retention over marginal accuracy gains.

Feature importance analysis added transparency to the model by identifying the key drivers of churn. Factors such as short customer tenure, month-to-month contracts, higher monthly charges, and lack of support services were found to be strong indicators of churn. These insights transform the model from a predictive tool into a decision-support system, enabling businesses to design targeted retention strategies such as improved onboarding, contract incentives, pricing optimization, and enhanced customer support.

Overall, this project highlights the importance of aligning machine learning solutions with business goals rather than optimizing purely for technical metrics. The final churn prediction system emphasizes interpretability, business impact, and practical deployment considerations. The approach and results demonstrate a strong understanding of end-to-end machine learning workflows and provide a solid foundation for further enhancements, including real-time prediction systems, advanced explainability techniques, and integration with customer relationship management platforms.