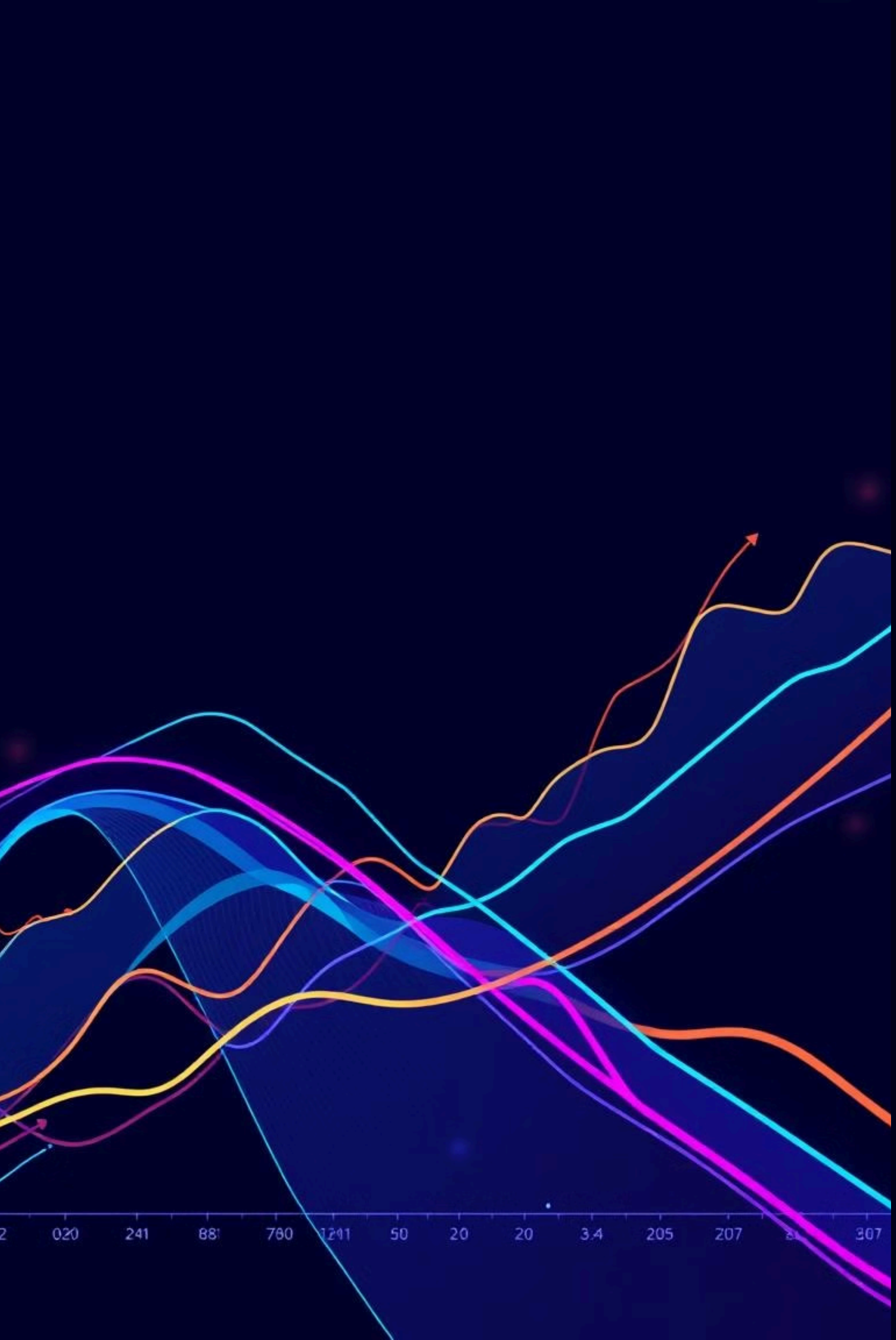

Customer Churn Prediction

Using Random Forest Classifier

A Comprehensive Analysis

K by Kaushik Kumar

Introduction

Customer churn occurs when customers stop doing business with a company. It is a critical metric for businesses, especially in industries like telecom, banking, and subscription services, as retaining customers is often more cost-effective than acquiring new ones.

This project focuses on leveraging machine learning techniques to predict customer churn, enabling businesses to understand behavioral patterns that lead to customer churn and proactively reduce churn rates by identifying customers at risk and intervening effectively.

Objective

Predict customer churn using machine learning.

Key Technology

Random Forest Classifier.

Data Preprocessing

Data preprocessing ensures that the data is clean, consistent, and ready for machine learning algorithms.

Column Removal

Dropped irrelevant columns, such as customerID.

Handling Missing Values

Imputed missing values using the mean for numerical columns.

Encoding Categorical Variables

Converted categorical features into numerical representations using one-hot encoding.

Feature Scaling

Used StandardScaler to standardize feature values.

Model Building

The Random Forest Classifier was chosen due to its ability to handle large datasets with higher dimensionality, reduce the risk of overfitting through bagging (bootstrap aggregation), and provide feature importance, making it easier to interpret the model.

1

Data Splitting

Divided the dataset into training (80%) and testing (20%) subsets.

2

Hyperparameters

Default settings for the Random Forest Classifier (e.g., number of estimators = 100).

3

Training

Trained the classifier on the scaled training data.

4

Prediction

Tested the trained model on the test dataset.

Results

The model achieved an accuracy of 80%, indicating reliable performance in predicting customer churn.

959

True Negatives

Correctly identified non-churn cases.

168

True Positives

Correctly identified churn cases.

77

False Positives

Incorrectly predicted churn.

205

False Negatives

Missed churn cases.

Conclusion & Future Work

The Random Forest Classifier demonstrated strong performance, particularly in identifying non-churn customers. The model's accuracy of 80% reflects its effectiveness in handling complex datasets. However, performance for churn prediction (Class 1) can be improved, as recall was lower for this class.

1

Model Optimization

Perform hyperparameter tuning.

2

Exploration of Other Models

Try advanced algorithms like Gradient Boosting, XGBoost, or Neural Networks.

3

Feature Engineering

Include additional features like customer interaction history, behavioral patterns, or external factors.

4

Class Imbalance Handling

Use techniques like SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset.

5

Deployment

Implement the model into a live system for real-time customer churn prediction.