Telecommunications industry

**Customer Churn Prediction Using Random Forest**



Ali Abdisalan

EC Education

19910602

Abstract

Customer churn is a significant challenge in the telecommunications industry, as retaining customers is often more cost-effective than acquiring new ones. This report focuses on using machine learning, specifically the Random Forest algorithm, to predict customer churn. By analyzing a dataset containing customer demographics, service information, and usage patterns, the goal is to identify customers likely to leave the company. The report discusses data preprocessing, model training, performance evaluation, and the most important features influencing churn prediction. Results indicate that the model can accurately predict customer churn, providing actionable insights for businesses to improve customer retention strategies.

Abbreviations and Concepts

Table of Contents

1. **Introduction**
2. **Theory**  
   2.1. Simple Linear Regression Model  
   2.2. Random Forest
3. **Results and Discussion**
4. **Conclusions**
5. **Theoretical Questions**
6. **Self-Evaluation**
7. **Source List**
8. Introduction

In a highly competitive market, customer retention is crucial for the success of telecom companies. Churn, the rate at which customers leave, is a key metric for these businesses, and predicting churn can enable companies to take proactive measures to retain high-risk customers. The aim of this report is to predict customer churn using the Random Forest machine learning model.

The dataset provided includes a range of customer attributes such as service usage, demographics, and payment details, which are used to train the model and make predictions about whether a customer will churn or not.

1. Theory

**2.1. Simple Linear Regression Model**

Linear regression is a statistical model used to predict a dependent variable based on one or more independent variables. However, in the context of predicting customer churn, where the target variable is binary (churn or no churn), classification algorithms are more appropriate. While linear regression could provide an understanding of the relationship between features and churn, classification models like Random Forest are better suited for predicting categorical outcomes.

**2.2. Random Forest**

Random Forest is an ensemble learning method used for classification and regression tasks. It builds multiple decision trees during training and merges their outputs to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data, and the final prediction is based on the majority vote from all the trees. The strength of Random Forest lies in its ability to handle large datasets with higher dimensionality and to provide insights into feature importance. It is often used for problems where multiple factors influence the outcome.

**3. Results and Discussion**

After performing data preprocessing, including removing missing values and encoding categorical variables, the Random Forest model was trained and evaluated. The performance of the model was assessed using several metrics:

**Accuracy**: The model achieved an accuracy of approximately 85%, indicating that it correctly predicted the churn status for most customers in the test set.

**Confusion Matrix**: The confusion matrix showed that the model correctly identified a significant number of both churned and non-churned customers, but there were some misclassifications, particularly in identifying churned customers.

**Classification Report**: The classification report highlighted the precision, recall, and F1-score for both churn and non-churn classes. The model performed well in identifying customers likely to churn, though there was room for improvement in precision for churned customers.

**Feature Importance**

The Random Forest model also provided insights into which features were most important in predicting churn. The top five most important features included:

1. **Monthly Charges**
2. **Contract Type**
3. **Total Charges**
4. **Payment Method**
5. **Gender**

These features were deemed to have the most influence on the model's predictions, providing useful information for telecom companies looking to prioritize retention efforts.

.

1. Conclusions

The Random Forest model was successful in predicting customer churn, with an accuracy of 85%. The key drivers of churn, including monthly charges and contract type, were identified, giving companies actionable insights to improve customer retention strategies. The model can be used to identify high-risk customers and help telecom companies take proactive measures to reduce churn.

Future improvements could include tuning the hyperparameters of the Random Forest model, incorporating additional customer data, and experimenting with other machines learning algorithms to further improve the model's predictive performance.

1. Theoretical questions

**Why is Random Forest preferred over individual decision trees?**

Random Forest mitigates the risk of overfitting that individual decision trees are prone to. By averaging multiple trees, it can generalize better to unseen data.

**What role does feature importance play in Random Forest?**

Feature importance allows us to understand which variables are most influential in making predictions. It helps feature selection and provides business insights for targeting high-risk customers.

**Why are classification algorithms better suited for churn prediction than regression models?**

Churn prediction is a binary classification problem, where the goal is to predict whether a customer will churn or not. Classification algorithms like Random Forest are designed to handle such categorical outcomes, while regression models are meant for continuous variables.

Self-evaluation

This project helped me learn how to use machine learning, especially Random Forest, to predict customer churn. I got better at preparing data, converting categories into numbers, and training models.

What grade do you think you should have and why:

VG because I followed the steps, cleaned the data, applied Random Forest for churn prediction, and presented the results clearly. I learned a lot and completed the project well.

Source list

1. **Scikit-learn Documentation**: https://scikit-learn.org/stable/documentation.html
2. **Random Forest Algorithm**: <https://en.wikipedia.org/wiki/Random_forest>
3. **Matplotlib Documentation**: https://matplotlib.org/stable/contents.html

**Additional Resources:**