Telco Customer Churn Prediction

Xinyuan Wang(xw648) Qiuyu Bao(qb33) Xiyi Fan(xf227)

Introduction

In modern days, Telecommunications is undoubtedly one of the essential industries in our daily life, considering that we need to rely on their service to keep up with the fast-paced information and keep in touch with our beloved ones. The business of Telecommunication is undoubtedly very competitive around the world. For a company to stay in business and survive in the heated competition, one of the crucial things is to keep the customers. Thus, an important subject that needs to be done is predicting customer behavior, analyzing all relevant customer data, and developing focused customer retention programs. With the knowledge of statistics and data science, we aim to provide some analysis focusing on explaining what types of customers are more likely to stay and leave, creating some prediction models to make customer churn predictions, and ultimately helping telecom companies to survive and thrive among all competitions.

Data

We are using the Telco Customer Churn dataset from Kaggle, which contains 7043 customers' data. This data have 21 features, which include customers who left within the last month, services that each customer has signed up for (phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies), customer account information such as how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges, and customer demographic info about customers – gender, age range, and if they have partners and dependents.This dataset surely covers enough features and information sufficient for us to investigate customer behavior on staying or churning and to build good quality predictive models,

Method

We would first use some exploratory data analysis to understand better and visualize the data set. Since we are dealing with 21 features, we could use principal component analysis to extract the main components that explain the majority of the data set. After which, we could use various approaches to build models for prediction, such as k-means clustering, decision trees, random forest, and logistic regression. We are aiming to see not only the prediction accuracy of the model but we are also looking forward to good interpretability, which can help us better understand the causal relationship between the customer's features and the outcome of churn or not.