Project Proposal

Customer Churn Prediction

Zhiqi Yang, 2110635
Elise Wolf, 1828204
Xi Liu, 2113820
Shiqi Zhou, 2036915
Yenan Chen, 2113612

October 2024

Submitted to
Data and Web Science Group
Prof. Dr. Sven Hertling
University of Mannheim

Contents

1 What is the problem you are solving? ii
2 What data will you use? ii
3 How will you solve the problem? ii
4 How will you measure success? (Evaluation method) iv
5 What do you expect your results to look like? iv

1 What is the problem you are solving?

The goal is to develop a machine learning model to predict whether telecom customers are likely to churn. Customer churn is a costly issue for telecom companies, as retaining customers is cheaper than acquiring new ones. Accurate churn prediction allows companies to focus retention efforts on at-risk customers, offering targeted solutions like discounts or improved services.

Key questions include:

- Which factors (e.g., contract type, tenure, monthly charges) are most related to churn?
- How effective are machine learning models in predicting churn?
- What actions can telecom companies take to reduce churn based on the predictions?

2 What data will you use?

We will use the Telco Customer Churn dataset, available on Kaggle, which contains data from a telecommunications company on customer demographics, account information, services used, and whether the customer has churned.

Data source: Telco Customer Churn Dataset (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Data Gathering: The dataset is readily available for download in CSV format from Kaggle. After downloading, we will inspect and preprocess the data for analysis. The dataset is already well-structured, but we will perform necessary preprocessing steps.

3 How will you solve the problem?

We will approach the problem by building and evaluating multiple machine learning models that predict customer churn based on the available features. Our approach will consist of the following steps:

Preprocessing Steps Required

- **Data Cleaning**: We will first check for and handle any missing or inconsistent data entries. Missing values in numerical fields will be imputed using the median, while categorical fields may have missing values filled with the mode.
- **Feature Encoding**: Since the dataset contains categorical variables (e.g., gender, contract type), we will use techniques such as one-hot encoding to transform these into numerical representations.
- **Scaling**: Numerical features such as *MonthlyCharges* and *TotalCharges* will be scaled to ensure that all features contribute equally to the model.

Algorithms and model specifics

- **Naive Bayes**: Naive Bayes can be used for customer churn prediction by treating each feature (e.g., customer behavior, demographics, usage statistics) as independent. The model calculates the likelihood of a customer churning based on these features.
- **Logistic Regression**: Logistic Regression is often used in customer churn prediction for binary classification (churn or no churn). It models the probability of a customer churning based on input features like the number of complaints, subscription length, and usage frequency.
- **K-Nearest Neighbors (KNN)**: KNN can predict churn by finding the most similar customers (neighbors) and using their churn behavior to predict whether a current customer will churn. If a majority of similar customers have churned, the model predicts that the customer is likely to churn.
- **Decision Trees**: Decision trees can be used to predict customer churn by splitting data into branches based on decision rules (e.g., if a customer's usage is below a certain threshold, predict churn). The tree is built by recursively selecting the feature that best separates churners from non-churners.
- **Random Forest**: Random Forest uses an ensemble of decision trees to improve churn prediction accuracy. Each tree is built from a random sample of data and a random subset of features. The final prediction is made by averaging the predictions of all the trees.
- **Support Vector Machines (SVM)**: SVM can be applied to customer churn prediction by finding the best hyperplane that separates churners from non-churners. SVM can also be used with kernels to capture non-linear relationships in the data.

- **Multilayer Perceptron (MLP)**: MLP is a type of ANN with multiple layers of neurons, which can be used to predict customer churn by learning from complex, high-dimensional data. The MLP architecture is well-suited for capturing non-linear patterns in the data.
- **Artificial Neural Networks (ANN)**: ANNs can model complex relationships between input features (e.g., demographics, usage data, behavior patterns) and the likelihood of churn. It can capture non-linear relationships that other models might miss.

4 How will you measure success? (Evaluation method)

We will evaluate our models using standard classification metrics:

- **Accuracy**: The proportion of correctly classified customers.
- **Precision and Recall**: Precision measures how many of the predicted churn customers actually churn, while recall measures how many of the actual churn customers were correctly identified.
- **F1-Score**: A balanced measure of precision and recall, especially useful when the dataset is imbalanced.
- **ROC-AUC Score**: The area under the ROC curve will provide insight into how well our model distinguishes between churn and non-churn customers across various threshold values.

5 What do you expect your results to look like?

- **A predictive model**: The output will include a classification model that predicts the likelihood of churn for each customer based on their features.
- **Feature Importance Analysis**: Using models like Random Forest or XGBoost, we will be able to rank the features that most influence customer churn. This will give the telecom company actionable insights into which factors they should address to reduce churn.
- **Churn Probability Dashboard**: We may create a simple interface or dashboard that visualizes the probability of churn for individual customers, along with the top contributing factors.
- **Actionable Recommendations**: Based on our findings, we will provide recommendations to the telecom company about which customer segments are most at risk of churn and what strategies they might employ to retain these customers.