Project Proposal

Program: Master of Data Analytics

Course: Introduction to Machine Learning

Delivered by: Dr. Dan Lizotte

Submitted by:

Dani Parayil

Garima Gambhir

Ritika Pandey

Semal Shastri

Sumedha Galgali

Western University

(2024-25)

1. Description of Applied Problem

Customer churn is a major issue for telecom companies, impacting profits and growth. This proposal aims to identify why customers are leaving a fictional telecom company and predict who is at risk. By using machine learning to analyze churn data, we will explore factors like demographics, service usage, satisfaction, and contract types. The goal is to uncover patterns that help reduce churn. For example, customers with month-to-month contracts or low satisfaction are more likely to leave, and in areas like Southern California, competitor actions may influence churn.

2. Description of Available Data

The Telco customer churn dataset covers 7,043 customers, with data organized into five tables: demographics, location, population, services, and status. Key details include customer ID, age, gender, monthly charges, contract type, churn status, satisfaction scores, and customer lifetime value (CLTV). It also includes satisfaction scores and customer lifetime value (CLTV), which help assess churn risk as well as tracks reasons for churn, providing insights into customer dissatisfaction. This dataset supports in-depth analysis of churn and customer behavior to predict who might leave.

3. Plan for Analysis and Visualization

Pre-processing:

- **Data Cleaning:** Remove missing or inconsistent data, ensuring accurate inputs.
- **Normalization:** Normalize numeric variables such as monthly charges and CLTV to standardize input for machine learning models.
- **Encoding Categorical Data:** Convert categorical variables (e.g., gender, contract type) into numerical format using techniques like one-hot encoding.

Feature Development:

- **Feature Engineering:** Create new features by combining existing ones, such as interaction terms between contract type and customer satisfaction or geographic location and churn reason.
- **Churn Risk Segmentation:** Develop a risk score based on customer demographics, satisfaction, and contract type, which will be used to group customers by churn probability.

Analysis and Machine Learning Methods:

- **Predictive Models:** Apply models such as logistic regression, random forests, and XGBoost to predict customer churn. These models will help identify the most important variables affecting churn.
- **Cross-validation:** Use k-fold cross-validation to ensure the reliability of model performance.

Visualization:

- **Churn Heatmaps:** Visualize churn risk by geographic location and contract type using heatmaps.
- **Feature Importance:** Visualize the importance of different features (e.g., satisfaction score, contract type) through bar plots and partial dependence plots.

Performance Assessment:

- **Metrics:** Evaluate model performance using accuracy, precision, recall, and F1 score. Additionally, calculate the area under the ROC curve (AUC-ROC) to assess the model's ability to differentiate between churners and non-churners.

References:

Telco customer churn (11.1.3+) (ibm.com)

Telco Customer Churn (kaggle.com)

https://scikit-learn.org/stable/modules/cross_validation.html

https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

https://builtin.com/data-science/random-forest-algorithm

https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

https://www.datacamp.com/tutorial/xgboost-in-python

https://www.nvidia.com/en-us/glossary/xgboost/

https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/