A COMPARATIVE STUDY OF VARIOUS MACHINE LEARNING ALGORITHMS FOR CHURN PREDICTION PROBLEM

Aatish Kayyath   Aston Glen Noronha   Rohith Sure

1. PROBLEM STATEMENT

Predict with high accuracy which customers are about to churn and compare the performance of various machine learning algorithms for the same.

2. INTRODUCTION

Losing valuable customers is never a pleasant experience for any company. It is common for companies to focus on acquiring new clients at the beginning, then grow by offering additional products or getting existing clients to use them more frequently.

If all is going well, there comes a point when the company is large enough that it must also choose a slightly more defensive strategy and focus on retaining existing customers. Despite the best user experience, there will always be a group of clients who are not satisfied and decide to leave.

As a result, the company must find a way to prevent these (voluntary) departures in the most effective way possible. This is where the churn model, among others, comes to the rescue.

“Churn is defined in business terms as ‘when a client cancels a subscription to a service they have been using.’” A common example is people cancelling Spotify/Netflix subscriptions. The purpose of Churn Prediction is to predict which of your clients will cancel a subscription, i.e., leave a business.

Obtaining this information is necessary from a company's perspective since acquiring new customers can be arduous and costly. In this way, Churn Prediction enables them to focus more on customers at a high risk of leaving.

Technically, it is a binary classifier that divides clients into two groups (classes) — those who leave and those who do not. Along with assigning them to one of the two groups, it will usually tell us how likely it is that they belong to that group.

For this project, our Team will be using the data available for Telecom Industry. Churn is one of the biggest problems in the Telecom Industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%. We will use the dataset available from IBM and Kaggle and try to mine and merge more data sources in the coming weeks.

3. DATASETS TO BE USED

a. IBM Telecom Churn Data:

https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

b. Kagle Telecom Churn Data:

https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data

4. IMPLEMENTATION PLAN

SPRINT 1: Team will read research papers and other resources and prepare a Literature Survey to understand the current solutions for the problem. The team will try to find more sources of datasets which if helpful we will mine and merge datasets.

SPRINT 2: In Sprint 2 Team will start working on the data. The team will complete the EDA within a week and try to gain insights and draw conclusions from the same. The team will complete the data cleansing, work on detecting and treating outliers, reduce dimensionality, remove correlated features, and prepare derived variables. The team will further work on extensive data pre-processing.

SPRINT 3: Once the data is prepared and ready each member of the team will start working on different ML (Machine Learning) algorithms. Each member will try to understand the algorithms thoroughly, check the performance, and accuracy and prepare a basic structure.

SPRINT 4: In the final week, each member will work on increasing the accuracy and optimizing their models by Hyperparameter Tuning. The team will compare the working and results of each algorithm and prepare a report on the best one. Finally, the Team will work on creating the final consolidated report, making visualizations and any revisions if required.

5. **TEAM RESPONSIBILITIES**

A) Data Cleansing and Pre-processing: Entire team will work together for data collection, data cleaning and data pre-processing. Reporting and Visualization will also be done together.

B) Literature Survey: Each member of the team will research a minimum of 5 papers related to the topic

C) ML Algorithm development: Each member of the team will be responsible for developing his own machine learning algorithm for the problem.