TeamTwo - ORIE 5741 Proposal

Shahrukh Showkath John Guo

In this project, we aim to predict customer churn for a music streaming service. The dataset we will use is from Sparkify, which is a dataset created by Udacity. We hope to be able to predict which users are at risk of leaving the service given an instance of their session data, with an extension of this being able to predict the likelihood of a user leaving in the future given a time series of their session data.

By being able to predict user churn prior to it happening, we can provide a window of opportunity to intervene and, through the use of targeted marketing and other incentives, retain the user. This is important as repeat customers are often more profitable than acquiring new customers, and since attracting new customers may be more expensive than the cost of providing incentives to retain existing users.

This dataset will allow us to create such a system as it contains a large amount of user session data, such as user location and what tier of service they are using. Our approach will be to first perform preliminary data analysis to understand the data, such as whether there is class imbalance or potential data leakage through the use of visualizations such as histograms and scatter plots, as well as summary statistics. Then data preprocessing will be applied to transform the data into a format that can be used by computer models. At this stage, a secondary analysis will be performed by using simple models to try and find further problems with the data, such as data leakage that was not apparent in the previous step before using a more expressive model. This may also help with determining which features are most important for predicting churn and if further feature engineering is necessary.