# **COGS 108 - Final Project Proposal**

# **Names**

* Ziqian Cui
* Rui Lu
* Chenxi Du
* Qiuyi Wang

# **Group Members ID**

* A14383826
* A14684839
* A14876523
* A15525077

**Research Question**

We use supervised learning models to identify customers who are likely to stop using service in the future. Furthermore, we will analyze top factors that influence user retention.

## **Background and Prior Work**

Nowadays, customer satisfaction is one of the most important features that many companies pay the most attention to. After reviewing each survey made by the customers, those companies will analyze them and improve their user experience from interior to exterior. In our group, we decided to look into Netflix customer churn rate and see what kinds of features cause them to unsubscribe. By looking at the churn rate, we are able to provide feedback back to Netflix (if possible) to improve their user experience.

We will first divide customers into two groups, one group contains paid customers, the other group contains customers who cancel after free trial. In each group, we will be looking at what kinds of shows or movies they have been watching. In group one (paid customers), we will pay attention to how long they subscribed and how active they are. In group two (free trial customers), we will also pay attention to what shows or movies they have been watching, but with this group, we will also be looking at what part of Netflix they disenjoy that causes them to not continue with the service.

After searching online, we found two useful resources. First source is a data analysis surrounding the topic of Sparkify, a music streaming service. In this report, it gives us the insight on how to import data from Spark, how to EDA, feature engineering, modeling and evaluation with F1 score and precision. After a series of analysis, it gets the result of which feature best attracts users which will help us to figure out what to do in our project. Second source is a customer churn prediction. In this dataset, it contains a customer’s college, income, overage and so on. From a customer's personal information, we can predict that he will continue to subscribe or inscribe in the future.

References (include links):

* 1) An example of churn’s analysis from a music streaming service called Sparkify. https://medium.com/@karvsmech/understanding-customer-churn-2a07a639720f
* 2) A github example of a user's income, college might determine if a user might be churn. https://github.com/vutsalsinghal/Churn-Prediction

# 

# **Hypothesis**

# We hypothesize that there are many factors that influence user retention in American technology and media service providers. For example, customer service, the number of movies updated per month, the time that users surf on the app. We try to figure out lots of related factors and find the most important three as our research target. We predict that the higher evaluation on customer service, the more movies updated per month, and higher the user surfing time on the app will improve user retention.

# **Data**

The ideal dataset we would like to use is a dataset that contains features such as user satisfaction directly from Netflix.

1. Dataset features assumptions:

· Users\_on\_free\_trial: True/False (boolean)

· Users\_subscription\_length: list of numbers (int)

· Users\_experience: list of comments (strings)

· Users\_names: list of names (strings)

· Movies\_title: list of names (strings)

· Movies\_length: list of numbers (int/float)

· Show\_title: list of names (strings)

· Shows\_length: list of numbers (int/float)

· Show\_seasons: list of numbers (int)

2. The dataset would likely be stored in CSV for easier access and analysis by using pandas.

3. The dataset would likely have at least 5,000 observations. Since a small amount of observations would contain biases, a roughly large number of observations will provide a precise result.

4. The observations will be users who subscribe to Netflix and who have been unsubscribed to Netflix.

5. The dataset will be analyzed by different methods such as SVM, linear regression and so on.

**Ethics & Privacy**

We have permission to use our data set, because we choose datasets from public resources in Kaggle, a platform for data scientists to share data. Some data scientists in Kaggle also used these datasets to train their models. The datasets had already stripped out any private information. Regarding privacy, the datasets will be fine.

The potential biases in our datasets may include look-ahead bias. When we do churn modeling, we may conclude data that we don’t have at the time of prediction in our model. For example, we cannot include the number of purchases in December, if we predict on November 1 which customer will churn next year. We need to make target leakage in advance to help us avoid the bias and perform our model as expected.

# 

# **Team Expectations**

Rui Lu Cleaning, collecting data from datasets and predictions based on other models.

Ziqian Cui Predictions based on given dataset and decides our assumptions and hypothesis

Chenxi Du Collecting data,setting up basic mathematical models and provide reasonings

Qiuyi Wang Reporting thesis and conclusions, providing with detailed examples

Due to the fact that 2 members of our team are in different time zones right now, we majorly use remote ways to have discussion about our projects like deciding topics, correcting details and ensuring our project could be worked and data could be collected easily online. Since we have different majors, the perspectives would be pretty different or even challenging but at the same time, different majors offer us with more knowledge of different fields which helps us to ensure our projects could cover more aspects.

# **Project Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **Meeting Date** | **Meeting Time** | **Completed Before Meeting** | **Discuss at Meeting** |
| 10/9 | 5 PM | Read & Think about COGS 108 expectations; brainstorm topics/questions | Determine best form of communication; Discuss and decide on final project topic; discuss hypothesis; begin background research; Get to know our teammates |
| 10/14 | 5 PM | Do background research on topic | Discuss ideal dataset(s) and ethics; draft project proposal |
| 11/4 | 5 PM | Edit, finalize, and submit proposal; Search for datasets | Discuss Wrangling and possible analytical approaches; Assign group members to lead each specific part |
| 11/18 | 5 PM | Import & Wrangle Data (Ant Man); EDA (Hulk) | Review/Edit wrangling/EDA; Discuss Analysis Plan |
| 12/2 | 5 PM | Finalize wrangling/EDA; Begin Analysis (Iron Man; Thor) | Discuss/edit Analysis; Complete project check-in |
| 12/9 | 5 PM | Complete analysis; Draft results/conclusion/discussion (Wasp) | Discuss/edit full project |
| 12/12 | 5PM | Continue making progress on the project | Final push;Discuss as a group what else to add to compete our project. |
| 12/16 | Before 11:59 PM | NA | Turn in Final Project & Group Project Surveys |