Capstone Project Proposal

1. What is the problem you want to solve?

   The problem that I want to solve is to predict whether a customer churns after their subscription ends. For this dataset, each subscription lasts 30 days, and the customer can auto renew their subscription, or must make a new purchase of the service to extend the subscription. Whether a customer churns or not is therefore specified by whether the customer makes a new purchase of the service within 30 days after their membership is set to expire.

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

   The client that provided the data is KKBox. KKBox is a streaming music service in Asia. The paid subscriptions subsidize other services that they offer, so KKBox must be able to predict the churn of their paid users so that they can continue offering the other services. If KKBox can learn why users leave, they can use these insights to try and keep the customers subscribing.

3. What data are you going to use for this? How will you acquire this data?

   The data that I will use for this problem is from the Kaggle challenge WSDM - KKBox's Churn Prediction Challenge, available at https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data. The data consist of four tables in csv format:

   Train.csv: user id and whether that user has churned

   Transactions.csv: user id and info on transactions that the user has placed

   User_logs.csv: user id and info on how many songs have been played

   Members.csv: user id and demographic information for each user

4. In brief, outline your approach to solving this problem (knowing that this might change later).

   When I perform my analysis, the tables will be joined on the user id field (some users are not in the members table).

   I will engineer new features to count the number of songs played and the number of unique songs played during each subscription period. Additionally, I will engineer new features for how long a user has been a subscriber, whether the user changed the subscription plan (true/false), and average length of new subscriptions purchased (since the user can purchase subscriptions longer than 30 days).

   First, I will perform exploratory data analysis to see whether there is any correlation between the features in the dataset and the churn feature. I will be interesting if there is any correlation between demographic information and churn or the engineered features and churn. Additionally, I will observe whether there is a trend for the listening behavior (number of songs) of users vs time across the dataset. Ideally, some of the features in the

dataset or the engineered features will correlate well with churn, so that a classification model can be built to predict whether a customer will churn.

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

   My deliverables will be the code in a Jupyter notebook, and a written report that gives my findings.