**Overview**

For this project, we were asked to develop research methods that can determine how many users on the HighNote music streaming website are most likely to switch from using the free service to the $3/month premium subscription. With this analysis, we created predictive models that can help the company generate more effective strategies for capturing premium subscriptions from their free users. First, we received a dataset from our client with data on each users’ past, current, and future features and engagement. Using the data, we were able to construct two sets of predictive models.

**Methodology**

Our primary predictive models focus on the prior features of users that make them more likely to become adopters of a premium subscription. Our secondary set of predictive models uses previous and current features to predict users’ future engagement on the website. In order to predict future engagement, we combined future engagement-related variables into a single feature named “Delta2 Engagement” and based our predictions on it. Before creating the models, we discovered that some of the data from the website is incomplete. Therefore, we modified the dataset by setting missing values to zero, allowing us to make predictions that would otherwise run into errors. For our KNN models, instead of choosing specific features or using all of them, we randomly selected five features each time and ran the model a few times.

**Analysis**

In our first set of predictive models, we found that the most accurate model was our logistic regression model with a score of about 93.2% accuracy and low variance of less than a 1% difference between the testing and training data. Therefore, using logistic regression can predict how many users will switch to premium the most accurately. By this model, we predicted that 12 users will switch to premium and 24,964 users will remain free. We only have 1,828 samples wrongly predicted.

For our second set of predictive models, we found that all models were improper predicting engagement levels with accuracy scores of around 10-30% and high variance consistently. This means that the past and current features are not good predictions for the future engagement, but we can get a general idea about how closely the features are related to engagement.

**Recommendation**

Therefore, we recommend the following: (1) using logistic regression to better predict how many users will switch to premium subscription focusing on website’s features in the past. (2) To better predict future engagement, we need more features and a more balanced dataset; The future is uncertain but we should keep our current website services for our users

**Limitations & Additional Analysis**

Lastly, we had some limitations in this project. Not only were there nulls in our dataset that led to the need to drop particular columns or replace values with zeros, but we were also limited by the size of our dataset. In the future, we would recommend that the client collects more features about their users such as number of logins or hours online per day, and datasets with more balance. These features could help them to identify how more users that are likely to become premium adopters.