**Problem Statement**

For my first capstone project I propose to provide a solution to the WSDM KKBox’s Churn Prediction Challenge on Kaggle.com. Customer ‘churn’ refers to the situation in a subscription based business where the customer cancels or does not renew their subscription. Churn analysis is one of the more common applications of big data for business purposes whereby customer data is used to predict the probability of churn. These predictions can be used to identify customers with high churn probability and thus inform what actions the business can take to increase customer retention, which is generally more profitable than attracting new customers. Among the reasons I think this is a good choice for my first capstone project is that this is a real problem that real people care about, is relatively straightforward in approach and complexity, and uses data that is readily available and is easy to clean.

**Business Objective:**

KKBox is a music streaming app that is primarily marketed to Mandarin language smartphone users in east Asia. Specifically the goal is to predict whether a certain user will make a new service subscription transaction within 30 days after their current subscription expires. The data is provided by KKBox as csv files of columned user data as well as ‘train’ and ‘test’ files.

**Approach:**

My approach will be basically to set up a machine learning algorithm in Python to fit a prediction model to the user data and then test it. In doing this I plan on taking certain steps. First there is determining and setting up which Python libraries I will need to use and extracting the data. Then I will explore and clean the data. Next will be feature selection pertaining to what is relevant and then fitting a model. Finally, the model will be evaluated.

**Evaluation:**

The algorithm will be trained on the train dataset and tested on the test a dataset. The results will be compared to the actual values provided and the evaluation metric will be the log loss which a measure of the accuracy of the model’s predictions

**Deliverables:**

The final deliverables for my project will be the code that is well documented on github, along with a paper describing the background and findings. The paper will include three recommendations to KKbox on how to use my findings. Additionally I will provide a powerpoint presentation giving a clear overview of the project.