Daniel Jurado – Thinkful Final Capstone Proposal

Data Science Flex

**What is the problem you are attempting to solve?**

*Please refer to my previous two capstones as my last one will build off of both of them*

[**https://github.com/dangoML/Project-Portfolio**](https://github.com/dangoML/Project-Portfolio)

When selecting the first capstone project I wanted to find a business-related dataset that I would be able to use for all three capstones. The logic behind this was to simulate working experience as if I were working for a company where I delivered various projects. Luckily, I came across a 30GB dataset for a mobile, music streaming app called KKBOX.

For my first capstone, I took a month break to learn Apache Spark so I can effectively predict customer churn. Dealing with so much data I took this as a great opportunity to learn about distributed computing. My second capstone built off my first as I ran a customer segmentation analysis. By understanding my customers better I was able to understand what makes each customer different from one to another, and what behaviors lead to them churning.

For my last project, I want to bring this all full circle by deriving customer lifetime value and it’s effect on customer segmentation. I will be using Apache Spark in both classification and clustering capacities to verify the effect Customer Lifetime Value has on churn and our ultimate decision on which strategy we use for customer retention purposes.

**How is your solution valuable?**

As thoroughly [described in my write up,](https://medium.com/p/fd1d85464e45/edit) creating a churn prediction model is simple but choosing which model you use is completely dependent on what you are willing to spend to retain the value of those deemed high-risk. Although I was able to create several models for churn, I will need to calculate Customer Lifetime Value in order to deliver on this business goal. Understanding customer value is essential. By calculating customer lifetime value we will finally be able to understand which churn prediction model we will use for our initial customer retention strategy.

**What is your data source and how will you access it?**

The data for this exercise is [publicly available at Kaggle](https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data). It was part of a competition hosted by The 11th ACM International Conference on Web Search and Data Mining (WSDM 2018). The data contains the music listening and transaction habits of ~2,300,000 members from January 2015 to March 2017. The dataset is currently in a database I create on Google Bigquery.

**What techniques from the course do you anticipate using?**

Everything I have learned in this course will be needed for the completion of this project. Calculating CLV isn’t so straight forward and thus multiple scenarios of CLV (both conservative and aggressive) will need to be created to cover all possible assumptions. The results will then need to be taken into consideration with my first two projects to see if this information is an overall value add. As mentioned before we will be using Apache Spark to recreate our clusters and then test if there is an significant effect on our churn models.

**What do you anticipate to be the biggest challenge you’ll face?**

Learning CLV in a timely manner to deliver this project on time.