**Course 3 Final Assignment**

In this Assignment, you will demonstrate your understanding of the data science methodology by applying it to a given problem. Pick one of the following topics to apply the data science methodology to:

1. Emails
2. Hospitals
3. Credit Cards

You will have to play the role of the client as well as the data scientist to come up with a problem that is more specific but related to these topics.

**Which topic did you choose to apply the data science methodology to? (2 marks)**

* *I chose to complete the data science methodology exercise on credit cards, users and payment activity.*

*Next, you will play the role of the client and the data scientist.*

**Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. (3 marks)**

**You are required to:**

1. **Describe the problem, related to the topic you selected.**
2. **Phrase the problem as a question to be answered using data.**

* *The use of credit cards makes someone capable of buying things or paying for services and fees even when the person does not have the usable assets to immediately pay. This means that card holder may hold or accrue a balance on their credit card, and the higher this balance, the higher the payment will be each month for the card holder. If this payment becomes too large, the card holder may be at risk of paying the payment partially, paying it late, or possibly not paying it at all, which creates losses to the bank as they are not recovering the money loaned, nor are they recovering the interest that is applied to the balance.*
* *Can we determine the risk factors of a credit card user of having missed or late payments based on credit usage and credit card user traits and characteristics?*

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

1. Analytic Approach

* Based on the business question, I will be using a predictive model as we are looking to leverage current data and traits to try and determine future payment activity of current cardholders. .

1. Data Requirements

* *Data for all bank users in order to separate the cardholders from the rest of the data.* 
  + *Based on this data we can pull out current and former card holders and all relevant characteristics for analysis*
  + *Data required will be:* 
    - *Cardholder status (former/current)*
    - *Payment History*
    - *Balance status (carrying balance/full payment each month)*
      * *If balance, average balance*
    - *Average spend per month*
    - *Credit Score*
    - *Age*
    - *Gender*
    - *Income (if available)*
    - *Additional Assets (checking/savings/investment)*
* *Collection method will generally be polling the internal database for users that meet all pertinent filtering criteria, no outside data* ***should*** *be required for the initial analysis*
* *We will be utilizing python and/or r for the quantitative analysis*
* *Jupyter notebooks will house the final product once deliverable.*

1. Data Collection

* *Core collection process happens, wrangle data from internal system (SQL etc), identify where we may have gaps (most all of this data is required with exception of income which isn’t always required or up to date.*

1. Data Understanding and Preparation

* *Evaluate complete data collected for preparation.* 
  + *Apply descriptive statistics to get an understanding of what we’re working with*
    - *This can give immediate insight as to the data modeling if we’re armed with trends, outliers or invalid data that will obfuscate our project.*
  + *Account for likely missing income data*
  + *Match all data records to relevant characteristics (remove PII)*
  + *Scrub data set for repeat records*
    - *More than one authorized user*
    - *Unify users with more than 1 card within the bank*
    - *Probably omit multiple payments on the account per month*
      * *Or create new/additional variable based on pay frequency*
    - *Feature Engineer (create distinct anchor/kpis to guide the learning model)*
      * *Avg Balance, Previous balance, Payment Activity, Credit Score*
  + *Any additional transformation/manipulation needed to fit the data for modeling*

1. Modeling and Evaluation

* *Create a model that takes into the account the factors of the above variables so we can feed it a test data set that may be similar in nature*
  + *A good test data set could be one that was gathered and vetted for outcomes based on the acceptance of an applicant for a credit card.* 
    - *Many of the criteria that determine someone’s eligibility for credit likely factor into ones tendency to pay on time/late/not at all.*
* *Test various algorithms to see if any variables are redundant/unnecessary/hold actual impact*
* *After this is determined, we can use visualizations, and analyses to determine model validity/viability (retest)*
* *Lastly, we will use the final model after all adjustments from test data, and iterations to determine if the model should be moved to production/deployment to solve real world problems/predict future behaviors that might be able to be fed into other models (i.e. acceptance criteria for credit eligibility)*

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.