**Guide for 650 Final Project**

In practice, there are many problems that are solved using data and analytic models. When assigned a data analytic project, understanding what question is being asked of the data drives the type of algorithm that is used. This guide is designed to provide DAT-650 students a roadmap through the milestones and final project. This roadmap aligns with CRISP-DM step of Data UnderstandingData PrepModelingEvaluation iterations.

When beginning a project, we do not go through 1 pass of these steps and have a model. Instead, we iterate many times through a variation of these steps for Descriptive Statistics and Predictive Modeling.

The DAT-650 project is a Pilot. In a pilot, we are being asked to perform a “feasibility study”. This means that we use limited resources to determine if the business case is “feasible”. If we find it is, then a funded project will be initiated by the company.

Considering the limited resources, this means that we just have a sample of the data and a general idea of what we want to predict. In the Use Cases for 650, they all are Classification problems as they have a binary target variable.

The pilot will use tools like Excel, Tableau to explore the data. Run statistics on the data to understand the data distributions of the different variables. What are their min/max/average/median/range. What are the values of the Percentiles for 90th/10th as well as a couple other like 75th/25th or 95th/5th – performing these statistical descriptive models for different variables in the data will tell you a lot about the data. Further, running histograms also tell us about the data and frequency of data. This is where the focus of data understanding and data prep is for the pilot. The data prep is an outcome of the data understanding.

Note that a Descriptive model is the base case of the predictive model. It is a very important part of your overall model design and analysis.

While there are many different types of predictive models, for the pilot, we want to leverage the one that gives us a good idea of predictive power but does not require the extensive design and development that a production model would. This model type would allow us to leverage the data understanding and perform data prep by selecting different variables to model and then continue to iterate through these same steps until we find a pilot model that supports the “feasibility” of the predictive model.

For the DAT-650 project, the datasets provided with the use cases all are Classification problems. We want to determine – Credit Default Y/N, Employee Attrition Y/N, Customer Churn Y/N and Sales Win/Loss.

Supervised Learning is the type of Predictive Model family of algorithms that would be used to solve Classification problems. In a Supervised Learning model, a characteristic is that we know the variable we want to predict.

Use the information below to select 1-3 algorithms to use for your pilot model. Each data set will perform differently using the models below. Additionally when performing analysis, selecting different variables that seem reasonable from a business perspective and applying different models can further refine the model outcome so you supporting data for your final recommendations.

The model algorithms for the Classification datasets are:

* + - * Decision Trees
        + Perform splits based on best variable at each decision node
        + Good
        + Easy to interpret
        + Does not require data preparation
        + Greedy-never look back
        + Unstable – small changes can produce large change in tree
      * Naïve Bayes Classifier
        + Simple version of Bayes
        + Assumes all inputs are independent of one another
        + Fast, single pass algorithm
        + Susceptible to redundant features
        + Requires categorical data
      * Support Vector Machines (SVM)
        + Uses support vectors to determine boundaries
        + Accurate but slow to train
        + Data size is not an issue
        + Creates non-linear decision boundaries

Resources to read:

Discussion of Algorithms

<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>

Descriptive Statistics

<https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9>

<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/>