**Applying Machine Learning to the Capstone Project**

The research question at hand can be framed into a supervised classification machine learning problem. Supervised machine learning tasks, learns a function that maps an input to an output based on known and labelled data. The research questions are as follows:

1. What is the probability that a small, medium or large number of downloads would be affected by consumer ratings, number of reviews, mobile application price, mobile application size, and mobile application minimum installation operating system version.

The outcome/ dependent variable (number of installs) was broken down into an ordinal variable (AdjustedInstall2) of 1, 2 and 3. Where 1 is small number of downloads (0 - 500), 2 is medium number of downloads (1,000 - 1,000,000), and 3 is large number of downloads (5,000,000 - 1,000,000,000).

The independent variables were minimum required mobile operating system (OS) for installation (MinimumVer), average consumer ratings (Rating), mobile app size in kilobytes (SizeKB), number of consumer reviews (Reviews), and mobile app price (Price). Regression Trees was used to find the impact strength of the independent variables on the independent variable. This was because the outcome/ dependent variable (AdjustedInstall2) was an ordinal variable of 3 values. The decision tree was split on the Reviews variable as the best fit for the model (number of reviews of 43000 and above would yield app downloads of 5,000,000 - 1,000,000,000 and above, number of reviews between 14 and 43000 would yield app downloads of 1,000 - 1,000,000, and number of reviews less than 14 would yield app downloads between 0 - 500). To evaluate the success of the technique used, the following steps were taken:

1. A training and testing set were created in the ratio 70:30 respectively. The training set was used to create a regression tree model.
2. The misclassification error rates in the training and testing set were compared, and they both were 11%.
3. Furthermore, a 10 fold cross validation was used in training the data, and the trained data was used in predicting the testing set for accuracy. The predicted accuracy was 91%.
4. Training controls of number =10 and repeats = 3 was used, as well as tuning grid parameters were used when cross validating the data.
5. The model had a p-value of 2.2e-16 which is very significant since p< 0.001