**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
2. What is the probability that small, medium or large number of downloads would be affected by the mobile application price
3. What is the probability that small, medium or large number of downloads would be affected by mobile app categories
4. What is the probability that small, medium or large number of downloads would be affected by mobile app minimum installation 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 will be mobile app categories (Categories), 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). Ordinal logistic regression will be used to find the impact strength of the independent variables on the independent variable. This was because the outcome/ dependent variable (number of installs) is an ordinal variable of 3 values.

To evaluate the success of the technique used, the following steps will be taken:

1. A training and testing set will be created. The training set will be used to create an ordinal logistic model and prediction, while this model and prediction would be tested on the testing set.
2. For each model combination, check the AIC to establish the best model. The lower the AIC, the better the model.
3. Use significant independent variables, with estimates t-value >2 and p-value < 0.05
4. Ensure the independent variable signs make sense in real life.
5. Compare the misclassification error rates in the training and testing set, and ensure they are similar
6. The false positive (FP), false negative (FN), true positive (TP), true negative (TN) ratios should be similar in both the training and testing sets.