Review training and testing data sets to understand the structures of the files. Eyeballing each feature in order to look for patterns.

Data preprocessing: filling missing data in necessary columns and determine features (is\_free, genres, categories, tags, reviews etc.) which are important to modeling.

Convert genres, categories, tags to dummy columns making the inside values usable.

Ensure the alignment between columns in training and testing data (if one is missing, fill '0' for all rows) in order to prevent crash in the modeling part later.

Random Forest was chosen as the most confident prediction method after comparing with others.

Set up the final model with training data then fit the testing. Result was hence generated.

Based on what I have learnt from MSBD5001, here are major steps for this project:

Data Cleansing

Feature Engineering

Training Data Preparation

Modeling & Tuning