IST 687 Applied Data Science

Google Play Store Apps Academic Project

(Group Project)

Course Learning Objectives:

* Essential concepts and characteristics of data
* Scripting/code development for data management using R and R-Studio
* Principles and practices in data screening, cleaning, and linking
* Communication of results to decision makers

Course skill sets:

* Identify a problem and the data needed for addressing the problem
* Perform basic computational scripting using R and other optional tools
* Transform data through processing, linking, aggregation, summarization, and searching
* Organize and manage data at various stages of a project lifecycle
* Determine appropriate techniques for analyzing data

The goal for this project is to implement the skills learnt from the course and extracts valuable insight from a data set through statistics analysis and visualization, and Machine Learning algorithm by using R studio.

In this project, I have understood how to manipulate data in R and clean the data into a desired format to perform statistics analysis and visualization. Further, with the introduction of Machine Learning mentioned in the course, I have a fundamental knowledge of how to implement a Machine Learning algorithm to build a model and predict the results.

The data set is Google Play Store Apps dataset from Kaggle (<https://www.kaggle.com/lava18/google-play-store-apps>). We have done many data cleansing steps on the dataset. We removed the missing values; converted Size field into the same unit and categorized into different levels; removed duplicates; parsed Genre field into two fields; converted a field from character type to numeric type and categorized it. We produced graphs to answer the business questions. And further, we implement random forest algorithm and support vector machine algorithm to build a model to predict if a published app will be popular or not (installation more than 100,000 times).