Final Project Proposal

11/12/2019

Team Name: Team Unknown

Team members: Entong Li, Jingyu Xie, Xindi Lu, Zijia Cao

Research Question:

For this project, we want to figure out which variables could predict the rating of Google Play Store apps.

Dataset Descrition:

The data set we use is for Google Play Store apps. There are overall 10841 observations and 13 variables in this data set and rating is the response variable.

Data cited from: https://www.kaggle.com/lava18/google-play-store-apps

Variables will be used:

1. App: Application name
2. Category: Category the app belongs to
3. (respond variable) Rating: Overall user rating of the apps
4. Reviews: Number of user reviews for the app
5. Size: Size of the app
6. Installs: Number of user downloads/installs for the app
7. Type: Paid or free
8. Price: Price of the app
9. Content Rating: Age group the app is targeted at 10.Genres: An app can belong to multiple genres

Variables may be explored if have time:

1. LastUpdated: Date when the app was last updated on Play Store
2. Current Ver: Current version of the app available on Play Store
3. Android Ver: Min required Android version

Analysis:

Step 1: Filter all NA values and values larger than 10 in the response variable “rating”.

Step 2: Clean the data and include the predict and respond variables only.

Step 3: Make a ggpairs to see the relationship between variables, exclude some predictors if they have very weak correlation with respond variable.

Step 4: Fit a weighted model and unweighted model to decide which one is better.

Step 5: Fit additive model and interactive model to decide whether we need an interaction between any two predictor variables

Step 6: Finalize our model to predict app ratings.

Step 7: Evaluation of the visualization.