Project 3: Google Store App Rating Prediction

Details of Google Play store applications are given to predict the Rating of an Application. Given App details include Category, Size, Installs, Reviews, Type, Price, Content Rating etc.

Analysis of data is given below as summary.

Summary

- Total **10841** records are there in dataset with 13 columns (12 independent+1 dependent).
- App: Name of an application, so mostly do not have any value for the prediction of rating.
- Category and Genres are related. We can consider Genres as sub-category and might ignore for our model building.
- Similarly we can think of Android version and App version do not influence the rating of the app. As users give rating to an app based on the no. of downloads, reviews, price, type etc.
- Also last updated date does not matter for the rating of an app.
- Rating is the target column.
- Valuable columns can be
    - Category (as we can ignore sub-category, i.e. Genres)
    - Reviews
    - Size
    - Installs
    - Type
    - Price
    - Content rating
- One data record had been left shifted, which was showing wrong values for respective columns. That has been corrected by assigning values from previous column in record. Also Category for the record has been taken from the similar app category.
- **483** Duplicate records has been removed. So remaining **10358** records available.
- Data cleaning and type conversion is done for numerical data columns, such as Size, Installs, Reviews, Price, etc.
- Target column (**Rating**) had more number of Missing data (1465), which is filled with **Low** category during converting to categorical column.
- Most of the apps are in High Rating group. (High is >=3.5)
- **FAMILY** Category has the highest number of Apps.
- Missing values of **Size** column are filled with **average** app size.
- Two types of Apps are available: Free and Paid. Out of all Apps, **~7.386%** are **Paid** apps.
- The most expensive app is **"I'm Rich - Trump Edition"** and its price is **$400**.

Answers:

- ✓ **483** duplicate entries were there in dataset. Duplicate entries have been removed.
- ✓ **Category** column had a wrong category **1.9**, which occurred for a record due to shifting of data. This has been corrected by assigning category from similar apps. Also appropriate data has been assigned to respective columns.
- ✓ "**FAMILY**" category has highest number of apps.
- ✓ **Rating** column has more number of High rating apps compared to Low Rating. 79% High rating apps, 7% low rating apps and 14% missing values.. Rating column has been converted to two categories **High** and **Low**, where High >=3.5 and Low < 3.5. So Low Rating has increased to 21%.
- ✓ **Review** column has been converted to numerical.

Kishore Chandra Sahoo
kishore.cs@samsung.com

- ✓ Top 5 apps based on Reviews and their Genre are:

Top 5 Apps	
App Name	**Genre**
Facebook	Social
WhatsApp Messenger	Communication
Instagram	Social
Messenger – Text and Video Chat for Free	Communication
Clash of Clans	Strategy

- ✓ **Size** column values had **M** and **k** units, which are replaced by appropriate values as actual units. Missing values have been filled with average size of Apps. Column datatype converted to numerical.
- ✓ **Installs** column had **+** and **comma** sign, which have been removed and datatype changed to numerical.
- ✓ Approximately **7.4%** apps are **Paid** Apps.
- ✓ **$** sign has been removed from **Price** column and datatype changed to numerical.
- ✓ Most expensive app is **I'm Rich - Trump Edition** and its price is **$400**.
- ✓ Some columns have been removed such as App, Content Rating, Genre, Last updated, Current Ver, and Android Ver.
- ✓ Categorical columns such as Type, Category and Rating_Cat have been encoded with one-hot encoding.

- ✓ Data have been split at 70:30 ratio for train and test set.
- ✓ Multiple classification models are trained on the train_set and tested on the test_set for model accuracy and other classification reports. Also confusion matrix has shown in terms of heatmap.
- ✓ Various models and performance are shown below.

Decision Tree (Non-regularized)

- ✓ Full grown tree gives training accuracy of **~1**, but testing accuracy of **83%**. (i.e. Overfit model)
- ✓ It shows important features are **Reviews, Size, Installs, Price and few categories**.
- ✓ F1 score is **89%** and **58%** for High and Low respectively.

Decision Tree (Regularized) (DTR)

- ✓ Regularized DT gives model accuracy of **88%**, better than non-regularized decision tree.

- ✓ It shows important features are **Reviews, Installs, Size, few categories and Price**.
- ✓ F1 score is **93%** and **65%** for High and Low respectively.

Random Forest (RF)

- ✓ Random Forest gives model accuracy of **87%**, with Training accuracy of **98%**.

- ✓ It shows important features are **Reviews, Size, Installs, Price and few categories**.
- ✓ F1 score is **92%** and **64%** for High and Low respectively.

Kishore Chandra Sahoo
kishore.cs@samsung.com

Gradient Boost (GB)

- ✓ Gradient Boost gives model accuracy of **89%**, with Training accuracy of 88%.
- ✓ It shows important features are **Reviews, Size, Installs, Price and few categories**.
- ✓ F1 score is **93%** and **66%** for High and Low respectively.

Stacking

- ✓ Stacking of all 3 models (DTR, RF and GB models) gives model accuracy of **89%**.
- ✓ F1 score is **93%** and **67%** for High and Low respectively.

	precision	recall	f1-score	support
0	0.90	0.97	0.93	2491
1	0.81	0.57	0.67	617
accuracy			0.89	3108
macro avg	0.86	0.77	0.80	3108
weighted avg	0.88	0.89	0.88	3108

Classification Report for Stacking Model

Gradient Boost and **Stacking** gave better performance compared to other models.

Important features to make a highly rated mobile app are:

- Reviews
- Size
- Installs
- Price
- Also few categories such as Family, Health and fitness, Tools, Dating, Photography, etc.

Kishore Chandra Sahoo
kishore.cs@samsung.com