Final Project Report
Introduction to Data Analytics

Project Title:

Predicting App Ratings in the Google Play Store: An Analytical Approach Based On Historical Data

Prepared by:

Group 8:

Ryhan
Sharmi
Shalini

AIGC-5000-IGB Fall 2023

Humber College IGS

1. Problem Statement

- **How well will an app perform in the Google Play Store based on its reviews, size, and user demographics?**

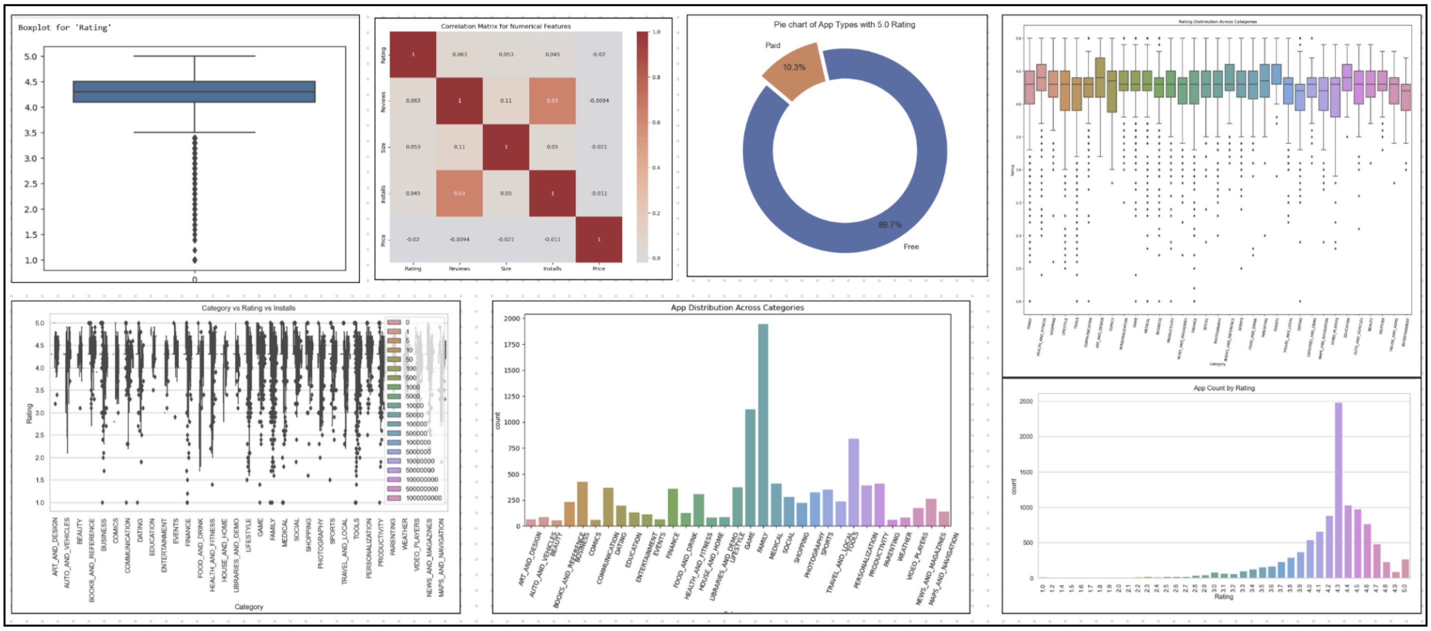
2. Dataset Description

- The data was originally scraped from the Google Play Store which includes essential details about apps. It covers the app's name, category, user ratings, number of reviews, size, download counts, whether it's free or paid, its price, target audience, genre, last update date, current version, and minimum Android version needed. This information helps analyze trends, predict app ratings, and understand what users like.
- Source: [Kaggle Dataset](https://www.kaggle.com/datasets/lava18/google-play-store-apps).
- This dataset is valuable for analyzing trends, predicting app ratings, and understanding market dynamics.

3. Dataset Analysis and Observations

Our dataset analysis from the Google Play Store indicated:

- **Data Quality**: Initial unclean data were refined.
- **Categories**: 'Games' dominate app categories, possibly affecting ratings.
- **Ratings**: Apps generally have high ratings, showing user satisfaction.
- **Reviews**: More reviews positively correlate with higher ratings.
- **App Type**: Free apps tend to get more 5-star ratings than paid ones.
- **Updates**: Frequent updates might lead to better app ratings.



4. Proposed Analytical/Prediction Model

Our predictive model utilizes the **Random Forest algorithm**, an ensemble learning method known for its accuracy and robustness. **Why** we chose this mode is because, the model:

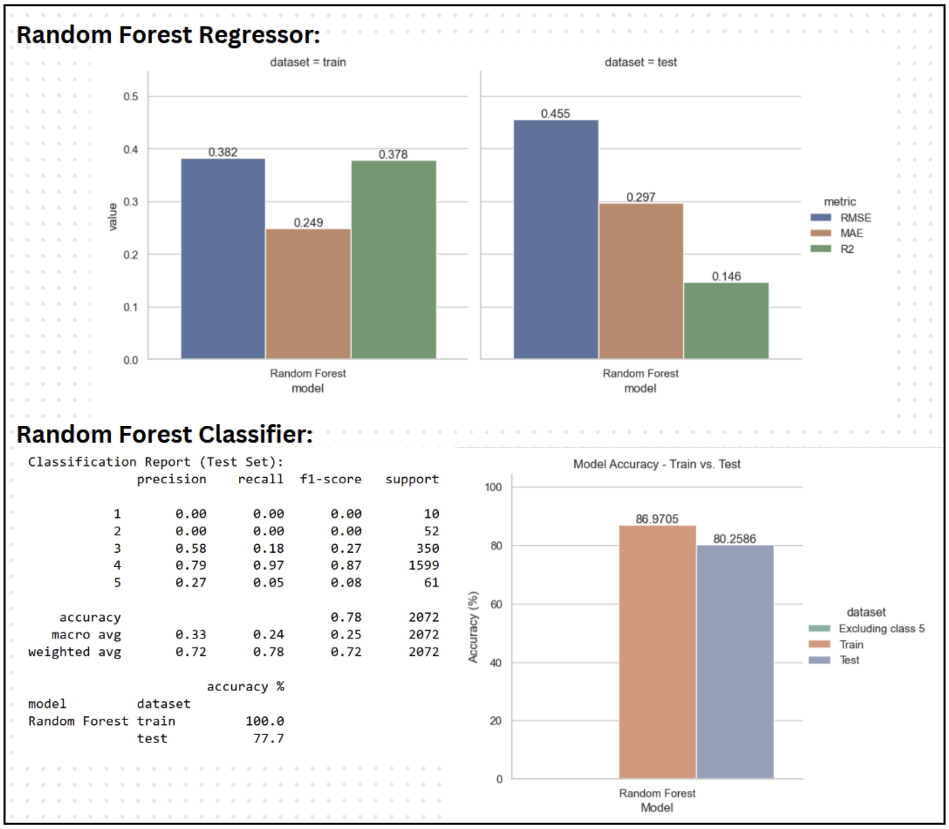
- **Combines Multiple Decision Trees:** This reduces the risk of overfitting and improves prediction accuracy.

- ➔ **Handles Various Data Types and Features:** It can efficiently process our dataset's diverse range of features.
- ➔ **Provides Insights into Feature Importance:** This is crucial for understanding which aspects most significantly impact app ratings.

5. Results and Discussions

We tried both Random Forest Regressor and Random Forest Classifier to compare which model fits the best (regression vs classification). Our conclusion was that the Random Forest Classifier performed far better achieving an **80.3% accuracy**.

- ➔ **Random Forest Regressor**:
  - ◆ RMSE of 0.382 (train) and 0.455 (test).
  - ◆ Consistent MAE and R2 values across train and test sets.
- ➔ **Random Forest Classifier**:
  - ◆ 100% accuracy on train set; overfitting indicated.
  - ◆ 77.7% accuracy on test set; lower performance on unseen data.
- ➔ **Hyperparameter Tuning**:
  - ◆ Improved test accuracy to 80.3% after excluding class '5'.
  - ◆ Optimized parameters: max_depth 20, min_samples_leaf 2, min_samples_split 10, n_estimators 300.
- ➔ **Conclusions**:
  - ◆ Reasonable prediction accuracy with room for improvement.
  - ◆ Model bias towards majority class detected.
  - ◆ Further tuning needed to enhance generalization.



Random Forest Regressor:

Random Forest Classifier:

```
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        10
           2       0.00      0.00      0.00        52
           3       0.58      0.18      0.27       350
           4       0.79      0.97      0.87      1599
           5       0.27      0.05      0.08        61

    accuracy                           0.78      2072
   macro avg       0.33      0.24      0.25      2072
weighted avg       0.72      0.78      0.72      2072

                        accuracy %
model         dataset
Random Forest train          100.0
              test            77.7
```