Google Play Store Data Analysis and Prediction

Project By: Hassan Khan (B21F0511DS008)

Hassaan Shiraz (B21F0514DS042)

Muhammad Ali Turk (B21F0600DS026)

Introduction

This project aims to analyze and predict app ratings on the Google Play Store using machine learning techniques. The dataset was sourced from the Google Play Store, containing various attributes related to the apps.

Data Preparation:

1. Data Loading and Exploration:

- The dataset was loaded using PySpark, and initial schema inspection was performed to understand the structure and types of data.

2. Data Cleaning:

- Irrelevant columns such as Developer Email, Minimum Android, Developer Id, Developer Website, Privacy Policy, Ad Supported, In App Purchases, Editors Choice, Scraped Time, and Free were dropped.

- Data types of relevant columns were cast to appropriate formats (e.g., float for Rating, Rating Count, Minimum Installs, Maximum Installs, Price).

3. Filtering Data:

- Focused on top 8 categories: Education, Music, Business, Tools, Entertainment, Lifestyle, Food & Drink, Books & Reference.

Data Visualization

- Category Distribution: Visualized the number of apps per category using a bar plot and the percentage distribution using a pie chart.

- Content Rating Distribution: Displayed the distribution of content ratings using a count plot.

- Price Analysis: Identified the top 10 most expensive apps and their install counts using a bar plot.

Feature Engineering

1. Encoding Categorical Variables:

- `StringIndexer` and `OneHotEncoder` were used to convert categorical features (`Category` and `Content Rating`) into numerical vectors.

2. Vector Assembler:

- Combined features into a single vector column (`attributes`) for model training.

Modeling

1. Linear Regression:

- Data was split into training and testing sets.
- A Linear Regression model was trained and evaluated.
- Performance metrics:
  - RMSE: 1.452
  - MSE: 2.108
  - MAE: 0.937
  - R^2: 0.688

2. Decision Tree Regressor:

- Trained a Decision Tree model and evaluated its performance.
- Performance metrics:
  - R^2: 0.723
  - RMSE: 1.389

3. Random Forest Regressor:

- Trained a Random Forest model and evaluated its performance.
- Performance metrics:
  - R^2: 0.771
  - RMSE: 1.278

- Analyzed feature importances and the number of trees used in the model.

Results

- The Random Forest Regressor outperformed other models, achieving the highest R^2 (0.771) and the lowest RMSE (1.278), indicating better predictive accuracy.

Conclusion

- The Random Forest model is the most effective for predicting app ratings on the Google Play Store based on the given features.

- Future work could include incorporating additional features such as user reviews and app descriptions to further enhance prediction accuracy and exploring advanced models like Gradient Boosting and Neural Networks.