Google Play Store Apps Data Analysis & Rating Prediction

1. Introduction

This project was completed as part of the Unified Mentor Data Analytics Program. The goal of this project is to perform end-to-end data analysis and predictive modeling on Google Play Store app data. The dataset provides insights into app categories, installs, ratings, user reviews, and other metadata that can be used to understand trends and predict app performance.

2. Dataset Details

Two datasets were provided: - googleplaystore.csv: Contains app metadata such as category, reviews, installs, price, content rating, etc. - googleplaystore_user_reviews.csv: Contains user reviews along with sentiment polarity and subjectivity.

3. Data Cleaning & Preprocessing

Key steps in data cleaning included: - Handling missing values without dropping data (imputation). - Removing duplicate entries. - Encoding categorical variables using Label Encoding. - Feature engineering: creating new variables such as log_reviews, log_installs, is_free, and review_rate. - Converting installs and price to numeric values for analysis.

4. Exploratory Data Analysis (EDA)

The exploratory analysis revealed: - Top app categories and their popularity. - Distribution of installs and reviews across apps. - Pricing strategies (most apps are free, but premium pricing also exists). - Distribution of content ratings across age groups.

5. Sentiment Analysis

Sentiment analysis was performed on the user reviews dataset. Sentiment polarity and subjectivity scores were derived using natural language processing techniques. This provided insights into how users perceive apps and whether reviews were positive, negative, or neutral.

6. Time Series Analysis

App update dates were analyzed to understand the impact of updates on app ratings. The analysis showed that frequent updates generally correlate with better user ratings and retention.

7. Predictive Modeling

Two approaches were attempted: - Regression: RandomForest, XGBoost, and Gradient Boosting were applied to predict exact app ratings. However, results were not strong, with R^2 scores around 0.10–0.14. - Classification: Ratings were grouped into categories (Low, Medium, High). A RandomForestClassifier achieved ~66% accuracy, with strong performance in predicting Medium ratings.

8. Challenges Faced

- Imbalanced dataset (most apps rated around 4.0–4.5). - Presence of noisy and extreme values in installs, reviews, and prices. - Regression models struggled due to low variance in target variable.

9. Conclusion & Learnings

This project provided hands-on experience in handling real-world datasets, performing sentiment analysis, feature engineering, and applying machine learning techniques. It highlighted the challenges of imbalanced data and the importance of preprocessing in model performance.

10. Future Work

Future improvements could include: - Using oversampling techniques such as SMOTE to balance classes. - Hyperparameter tuning of classification models for better accuracy. - Exploring deep learning approaches such as LSTMs for review sentiment analysis. - Building a dashboard for interactive visualization of app trends.

11. References

1. Google Play Store Dataset (Kaggle) 2. TextBlob documentation for sentiment analysis 3. Scikit-learn documentation 4. XGBoost documentation