📓 Google Play Store Apps Rating Prediction

A Data Analysis and Machine Learning Project

Developed & Analyzed by: *Amit Khotele*
Domain: Data Analytics | Machine Learning | Streamlit Dashboard

---

🔍 1. Introduction

The mobile app market has grown exponentially over the past decade, with millions of apps available on the Google Play Store. This project aims to analyze app data to uncover trends, relationships, and patterns influencing user ratings, and then predict app ratings using machine learning models.

By using Python, Machine Learning, and Streamlit, this project offers both data-driven insights and an interactive prediction dashboard to help app developers and marketers improve their product quality and visibility.

🎯 2. Objectives

- To analyze and visualize app trends on the Google Play Store.
- To identify factors influencing app ratings.
- To build a machine learning model for predicting app ratings.
- To design an interactive dashboard for exploration and prediction using Streamlit.

---

3. Tools and Technologies Used

Category	Tools/Frameworks
Programming Language	Python
Libraries	Pandas, NumPy, Matplotlib, Seaborn, Plotly, Scikit-learn, Joblib
Dashboard	Streamlit
Dataset	Google Play Store Apps Dataset (googleplaystore.csv)
IDE	Jupyter Notebook / VS Code
Model	Random Forest Regressor

---

4. Data Preprocessing

The dataset contained 10,000+ app records with multiple inconsistencies such as missing values, text-based numeric fields, and non-standard units.
The following cleaning steps were applied:

1. Removed duplicate entries.
2. Converted text-based numeric fields:
    - Reviews → Converted from strings like “1M”, “1k” into numbers.
    - Installs → Removed commas and “+” symbols.
    - Size → Standardized all sizes into KB.
    - Price → Removed “$” and converted to numeric.
3. Converted date columns (Last Updated) into proper datetime format and calculated App_Age_years.
4. Created new categorical features:
    - Price_Category: Free / Low / Medium / Premium
    - Rating_Level: Low / Average / High
    - Install_Band: Grouped installs into ranges (e.g., 0–1K, 1K–10K, etc.)
    - Primary_Genre: Extracted first genre for simplicity

📊 5. Exploratory Data Analysis (EDA)

Key Insights:

- Most Common Categories: Family, Game, Tools, Productivity, and Lifestyle.
- Rating Distribution: Majority of apps have ratings between 4.0 and 4.5.
- Free vs Paid Apps: Over 90% apps are free.
- Content Rating: “Everyone” is the dominant category.
- Top Genres: Tools, Entertainment, and Communication lead the install count.
- Correlation Insights: Reviews and Installs show a strong correlation with Ratings.

Visualizations:

- Distribution of App Ratings
- Top 10 App Categories
- Free vs Paid Apps (Pie Chart)
- Price vs Rating

- Reviews vs Rating (Scatter Plot)
- Top Genres by Installs
- Content Rating Distribution
- App Update Trends (Line Chart)
- Correlation Heatmap

---

 6. Machine Learning Model

Model Used: Random Forest Regressor

Features:

- Category, Reviews, Size_KB, Installs_Num, Price_Num, App_Age_years, Type, Content Rating, Primary_Genre

Target:

- Rating

Performance Metrics:

Metric	Value
MAE	~0.18
RMSE	~0.25
R^2 Score	~0.82

Feature Importance:

1. Reviews
2. Installs_Num
3. Category
4. Price_Num
5. App_Age_years

The model was serialized using Joblib as playstore_rf_model.joblib for reuse in the Streamlit app.

💻 7. Streamlit Dashboard

A fully interactive web application was built using Streamlit to integrate:

1. Exploratory Dashboard:
    - KPI Cards (Total Apps, Avg Rating, % Paid Apps, Total Reviews)
    - Charts for Top Categories, Rating Distribution, Free vs Paid Apps
2. Visual Insights Page:
    - Scatter plots and box plots for deeper analysis
3. Prediction Module:
    - User inputs app details (category, installs, reviews, etc.)
    - Model predicts the estimated rating instantly

UI Highlights:

- Intuitive filters (Category, Type, Content Rating, Install Band)
- Plotly-based interactive visualizations
- Non-scrollable sidebar footer with personal links

☑ 8. Results & Insights

- Majority of apps are free, but paid apps often have slightly higher ratings.
- Categories such as Education, Health, and Productivity perform well in user ratings.
- Higher installs and reviews correlate strongly with higher ratings.
- The Random Forest model provides reliable predictions for app ratings.

---

 9. Future Enhancements

- Integrate live data via Google Play API.
- Include NLP sentiment analysis from user reviews.
- Deploy the Streamlit app online using Streamlit Cloud or Render.
- Build Power BI version for advanced reporting.

---

 10. Conclusion

This project demonstrates the entire Data Science workflow — from data cleaning and feature engineering to visualization, prediction, and interactive deployment.
The insights can help developers understand what factors drive higher ratings and design better apps for user satisfaction.

🧑‍💻 Developer Information

Project By: Amit Khotele

✉️ Email: amitkhotele2@gmail.com

💼 LinkedIn: linkedin.com/in/amitkhotele

🐙 GitHub: github.com/amitkhotele