

● CIS 5450: Big Data Analytics

Analysing Google Playstore Data

Ankita Diwan, Anant Aggarwal,
Pradipta Kinasih



Empowering
App-based Businesses
with data-driven insights

Why this dataset?

Objective: Analyze factors influencing app popularity. We aim to identify the features that correlate with higher ratings, providing insights that could be valuable to developers and marketers in optimizing their app strategies.

The project also focuses on predicting app success using key attributes such as rating counts, installs, monetization strategies etc, offering a comprehensive approach to understanding and enhancing app performance.



Why is it valuable?



For product managers

prioritize efforts on features and strategies to enhance user engagement and visibility



For developers

understand features impacting app success to guide design and content improvements



For marketers

optimize promotional strategies and target the right audience

● Data Used

Google Play Store Apps Data

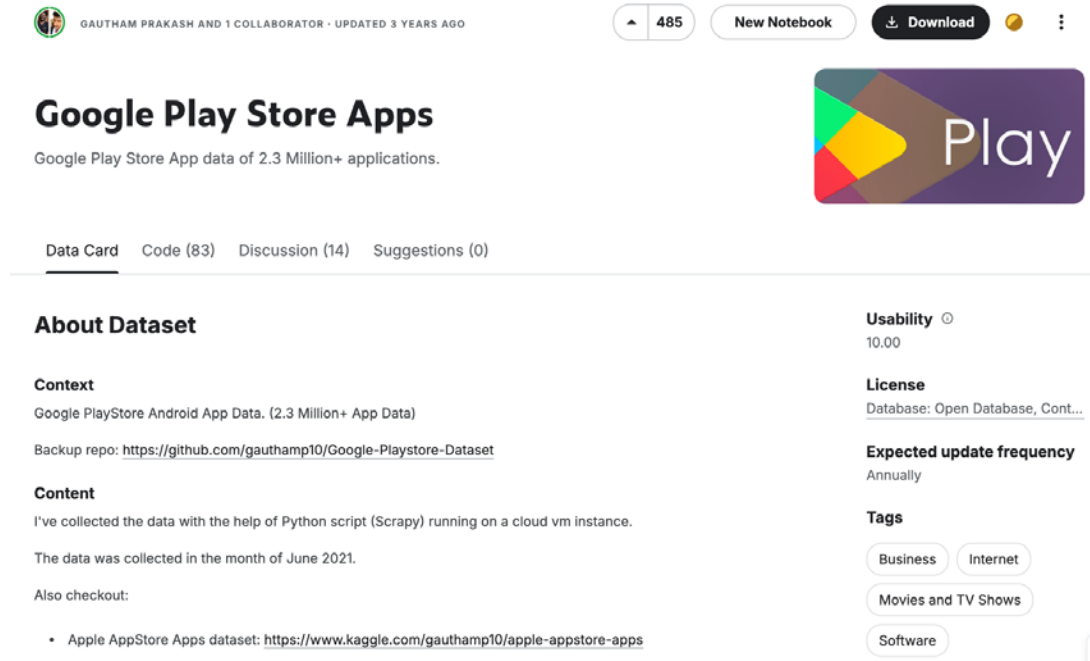
We used the **Google Play Store Apps** dataset from Kaggle, which is under the Open Database License.

The dataset includes

- 2 million+ rows
- 24 features

Like the app developer data, app rating, pricing, and the number of installs across various app categories.

[Data source link](#)



The screenshot shows the Kaggle dataset page for "Google Play Store Apps". At the top, it says "GAUTHAM PRAKASH AND 1 COLLABORATOR · UPDATED 3 YEARS AGO" and "485" views. There are buttons for "New Notebook", "Download", and a menu icon. The dataset title "Google Play Store Apps" is prominently displayed, followed by the description "Google Play Store App data of 2.3 Million+ applications." and the Google Play logo. Below the title, there are tabs for "Data Card" (selected), "Code (83)", "Discussion (14)", and "Suggestions (0)". The "About Dataset" section includes a "Context" paragraph about the data source and a "Content" paragraph about the collection method. On the right, there are sections for "Usability" (10.00), "License" (Open Database, Cont...), "Expected update frequency" (Annually), and "Tags" (Business, Internet, Movies and TV Shows, Software).

GAUTHAM PRAKASH AND 1 COLLABORATOR · UPDATED 3 YEARS AGO

485

New Notebook

Download

Google Play Store Apps

Google Play Store App data of 2.3 Million+ applications.

Data Card Code (83) Discussion (14) Suggestions (0)

About Dataset

Context

Google PlayStore Android App Data. (2.3 Million+ App Data)

Backup repo: <https://github.com/gauthamp10/Google-Playstore-Dataset>

Content

I've collected the data with the help of Python script (Scrapy) running on a cloud vm instance.

The data was collected in the month of June 2021.

Also checkout:

- Apple AppStore Apps dataset: <https://www.kaggle.com/gauthamp10/apple-appstore-apps>

Usability 10.00

License

Database: Open Database, Cont...

Expected update frequency

Annually

Tags

Business Internet

Movies and TV Shows

Software

What process did we follow?

Cleaning the data

- Removing columns with 50% null values
- Filling missing values with median and mode
- Drop unimportant columns
- Capping Price and Rating Count to the 99th Percentile
- Log Transforming Rating Count and Size
- Currency Standardization
- Consolidating Rare Categories in Content Rating

EDA + Hypothesis Tests

- Ratings, Price, Installs
- Outlier Detection and Log Transforming of Skewed Data
- Heatmap Correlations
- Frequency distribution of categorical features
- Simplify Categories
- 11 hypothesis tests using various techniques

Modelling

- Modeling 1: Linear Regression, Random Forest, XGBoost
- Model Comparison - 1
- Feature Correlation - 1
- Update Training and Test Data
- Modeling 2: Linear Regression, Random Forest, XGBoost
- Fine-tuning XGBoost
- Model Comparison - 2
- Feature Correlation - 2

What we learnt about PlayStore Apps

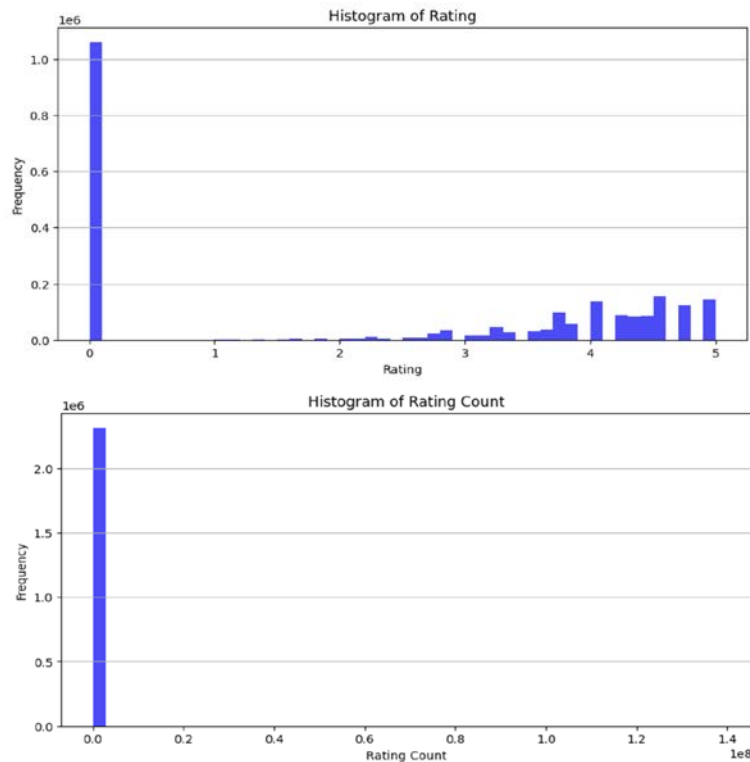
1

INSIGHT | rating and rating count

Significant disparity in user engagement and visibility across apps

A large number of apps have a rating of zero. This could indicate missing or unrated apps. Ratings above 4 are more common, which is expected for apps that have good quality.

The distribution of **rating count** is heavily skewed towards zero. Most apps have very few ratings, and a small subset of apps have an extremely high number of ratings.



2

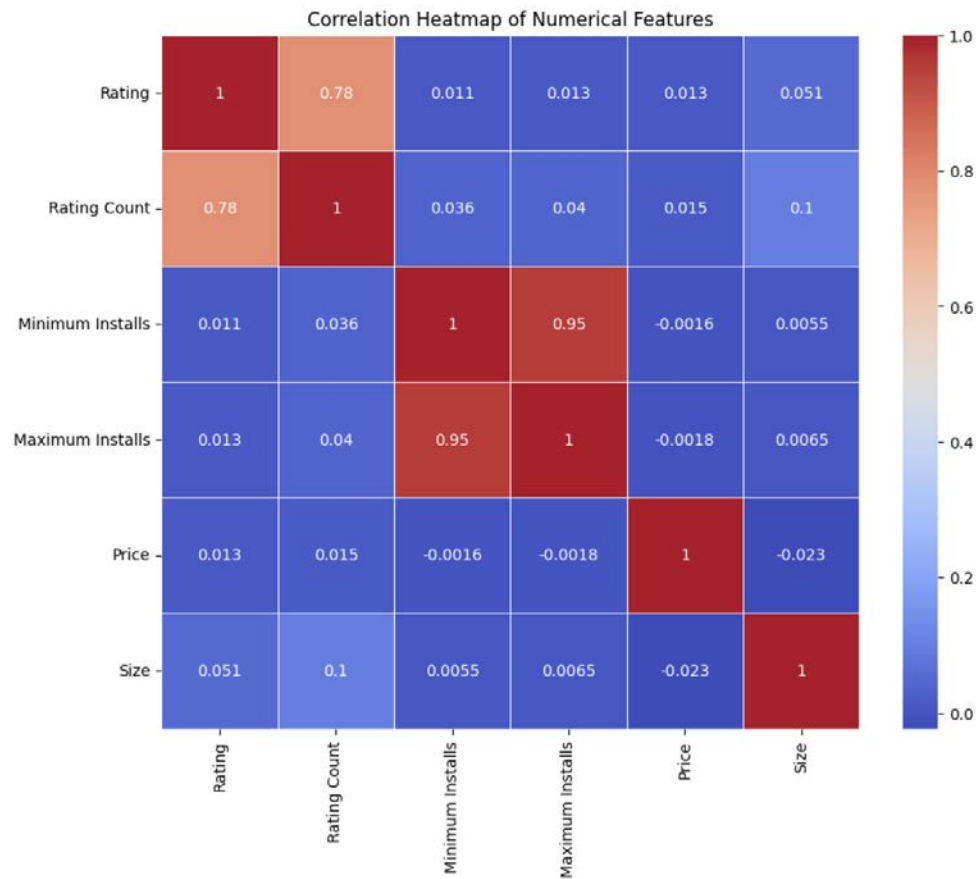
INSIGHT | correlation heatmap

Better engagement attracts more users to leave ratings

Strong positive correlation (0.78) between Rating and Rating Count i.e. more popular apps are often rated higher.

Apps with more engagement often have established credibility, leading to generally positive user feedback

High correlation (0.95) between Minimum Installs and Maximum Installs, possibly indicates that these two features are likely redundant and might not add much unique information to the model.



3

INSIGHT | rating by category

Disparities in User Engagement Across Categories

High Median Ratings:

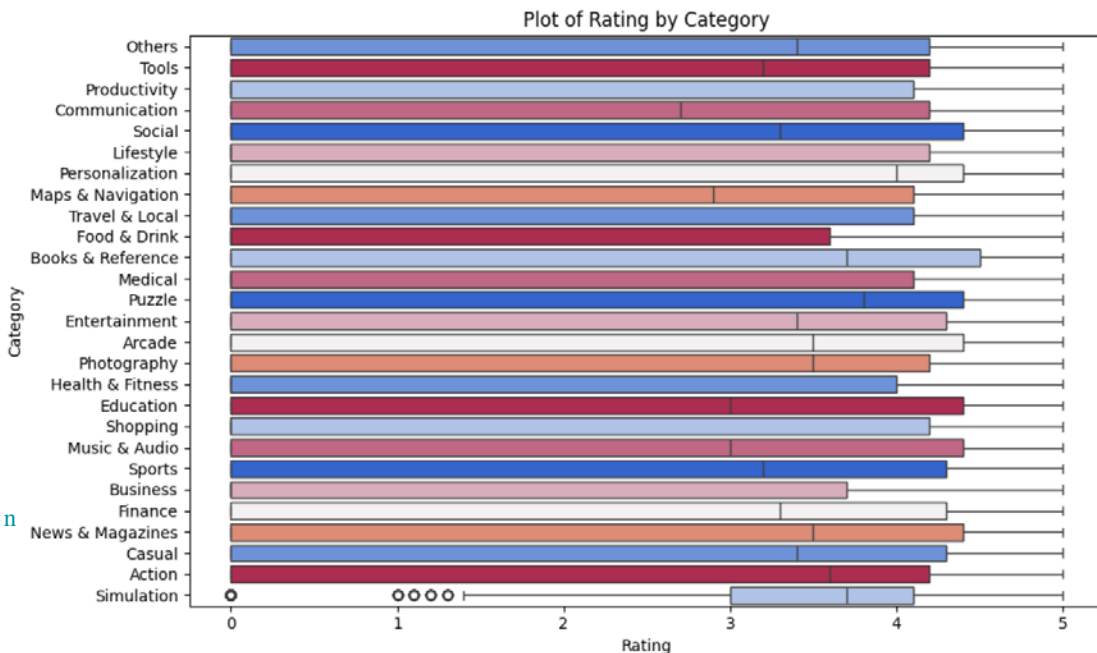
Books & Reference, Puzzle, and Personalization

Low Median Ratings:

Food & Drink, Health & Fitness, and Travel & Local

Consistent Ratings:

Simulation had smaller IQR.



4

INSIGHT | rating count by category

High median and consistent distributions indicate strong user satisfaction

High Median Rating Counts:

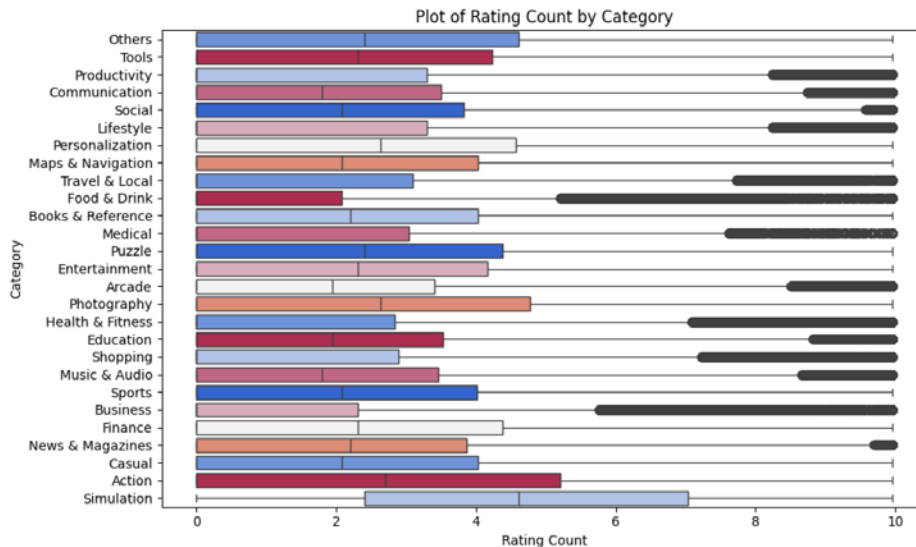
Simulation, Tools, and Puzzle

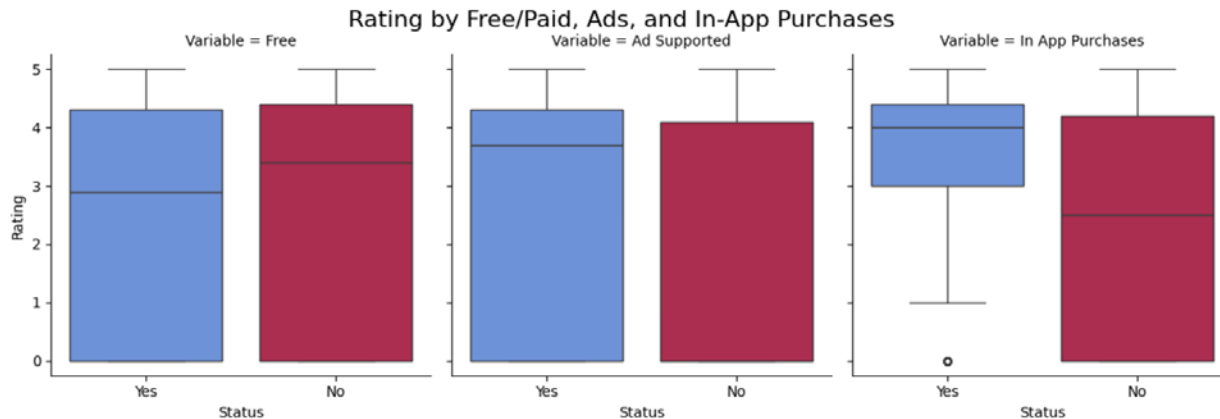
Low Median Ratings:

Food & Drink, Health & Fitness, and Travel & Local

Consistent Rating Counts:

Business, Food & Drink, and Medical have smaller IQRs,





This analysis highlights how monetization strategies such as pricing, ads, and in-app purchases impact user satisfaction, providing valuable insights for app developers and marketers.

5 INSIGHT

Ads or In-app purchases does not mean lower ratings

Free vs. Paid Apps:

Paid apps have slightly higher median ratings compared to free apps

Ad-Supported Apps:

The presence of ads may not strongly influence overall user satisfaction

In-App Purchases:

Apps with in-app purchases tend to have higher median ratings

Hypothesis Testing

We performed **11 hypothesis tests**, utilizing a wide variety statistical techniques to identify significant differences and associations across various app features such as pricing, size, installs, ratings, and categories.

Methods Used:

- T-Test
- ANOVA
- Permutation Test
- Bootstrapping
- Chi-Square Test

Key Questions:

- Do free and paid apps differ in size, installs, or ratings?
- Are there differences across app categories?
- Are content ratings linked to pricing models?

Hypothesis 1:

Is there a significant difference in average **ratings** between free and paid apps?

- H_0 : No difference between the average ratings of free and paid apps.
- *method*: Independent T-Test
- *result*: T-statistic: -15.8 and P-value: 0.0

Null hypothesis rejected!



Pricing impacts ratings

Hypothesis 2:

Is there a significant difference in the **average rating counts** among different app categories?

- H_0 : No difference in average Rating Counts across categories
- *method*: One-way ANOVA
- *result*: F-statistic: 3535.17, P-value: 0.0

Null hypothesis rejected!



Pricing impacts rating counts

Hypothesis 3:

Is the observed difference in mean Rating Count between categories statistically significant or by random chance?

- H_0 : Difference in mean Rating Counts across categories is due to random chance
- *method*: Permutation Test
- *result*: Observed Mean Difference: 4.00 and P-Value: 0.0
Null hypothesis rejected!

 User engagement (Rating Count) varies by category

Hypothesis 4:

Is there a significant association between apps rated for Everyone and their pricing model (Free or Paid) ?

- H_0 : Content Rating ("Everyone 10+") and Free/ Paid status are independent (no association).
- *method*: Chi-Square Test
- *result*: Chi-Square Statistic (X^2): 1071.50 and P-Value: 0.0
Null hypothesis rejected!

 Pricing models for apps vary across target content rating

Hypothesis 5:

Are there are significant differences in the number of maximum installs across app categories?

- H_0 : There is no significant difference in the number of installs across categories
- *method*: One-way ANOVA
- *result*: F-Statistic: 11.74 and P-Value: 0.0

Null hypothesis rejected!



Certain app categories attract significantly more users

Hypothesis 6:

Is there a significant difference in the average size of free and paid apps

- H_0 : There is no significant difference in the average size of free and paid apps
 - *method*: Permutation Test
 - *result*: Observed Mean Size Diff: 0.25 and P-value: 0.0
- Null hypothesis rejected!



App size is different between free and paid apps

Data Modeling

We developed predictive models to forecast app ratings, starting with **Linear Regression** as a baseline and progressing to **Random Forest** and **XGBoost**.

GOAL

To evaluate model performance using MSE and R^2 scores while identifying key features driving **app ratings**.

- **Feature Engineering:** Clean, preprocess, and encode data.
- **Modeling:** Compare Linear Regression, Random Forest, and XGBoost
- **Feature Importance:** Analyze top predictors and their correlation with ratings.
- **Insights:** Extract actionable findings for app development and marketing.

approach

Models Comparison

Model		MSE	R ² Score
Linear Regression	Baseline model to predict app ratings using selected features.	1.679	0.618
Random Forest	Non-linear ensemble model with stratified downsampling.	0.208	0.95
XGBoost	Gradient boosting model leveraging GPU for efficiency	0.194	0.96

Insights

- **XGBoost:** Best performance, capturing complex patterns effectively.
- **Random Forest:** Strong performer, slightly below XGBoost.
- **Linear Regression:** Too simplistic for the dataset's complexity.

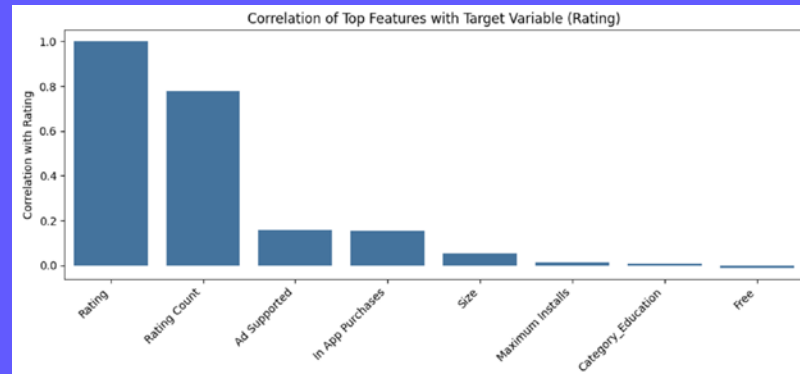
Feature Importance

Key Features (XGBoost & Random Forest):

- **Rating Count:** Strongest predictor of ratings
- **Maximum Installs & Size:** Moderate importance
- **Ad Supported & In-App Purchases:** Indicate monetization influences

Correlation

- **Rating Count (0.7796):** apps with more ratings tend to receive higher ratings due to greater user engagement.
- **Ad Supported (0.1583) and In-App Purchases (0.1536):** monetization features slightly influence ratings.
- **Size (0.0513) and Maximum Installs:** minimal impact on ratings.
- **Category_Education and Free:** negligible positive



Models Comparison

Model		MSE	R ² Score
Linear Regression	Baseline model to predict app ratings using selected features.	0.647	0.748
Random Forest	Non-linear ensemble model with stratified downsampling.	0.264	0.89
XGBoost	Gradient boosting model leveraging GPU for efficiency	0.240	0.897

Insights

- **XGBoost:** Refined features improved focus and representation, but reduction in performance suggests some predictive information may have been excluded.
- **Random Forest:** While feature selection and downsampling improved the dataset's balance and focus, some information critical to the model's accuracy may have been lost during refinement
- **Linear Regression:** Indicates that model fits the data better after the feature adjustments

Feature Importance

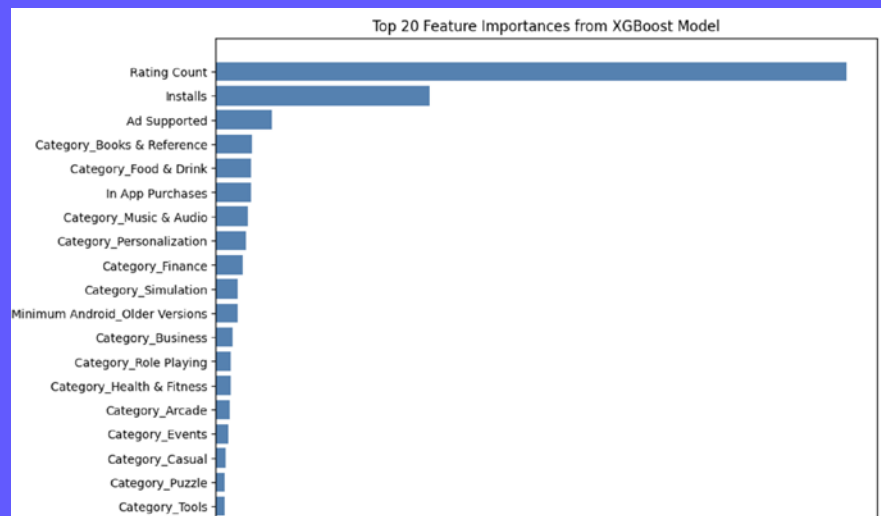
Key Features (XGBoost):

- **Rating Count:** Most significant predictor. Highlights the correlation between user engagement and high ratings.
- **Installs:** A major feature with 14% importance, emphasizing that popular apps often receive higher ratings.
- **Ad Supported:** Moderate importance (3.7%), suggesting ads have a noticeable but smaller influence.

Categories like Books & Reference, Food & Drink, and Personalization: Contribute between 2-2.4%, showing domain-specific trends in user preference.

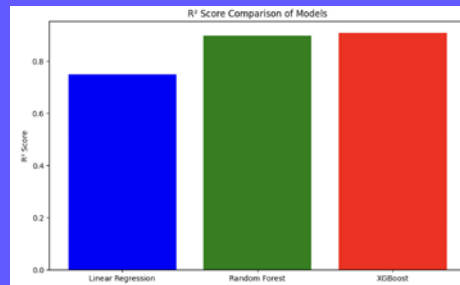
Insights

- Features like "Music & Audio" and "Finance" have minimal influence but reflect niche trends.
- High-performing features (e.g., Rating Count, Installs) dominate predictive power.
- The results guide developers on leveraging these impactful features for app improvement.



Models Comparison | after fine tuning

Model	MSE	R ² Score
Linear Regression	0.647	0.748
Random Forest	0.265	0.897
XGBoost	0.2382	0.907



To address hardware limitations in Google Colab, downsampling was applied for Random Forest and XGBoost, ensuring balanced representation across categories.

Insights

- **Linear Regression:** Captures ~75% of the variance ($R^2 = 0.7483$), performing reasonably well but limited in handling complex patterns.
- **Random Forest:** Performs better ($R^2 = 0.8971$) but faces memory constraints on the full dataset.
- **XGBoost:** Achieves the best performance ($R^2 = 0.9074$) with effective generalization and efficient GPU-based training, leveraging boosting and regularization to handle complex relationships and mitigate overfitting.

Actionable Insights

Insight 1: Engagement Drives Ratings

Features like Rating Count and Installs were identified as the most influential predictors of app ratings.



Incentivize user ratings through in-app prompts or rewards.

Optimize app listing to improve discoverability and encourage downloads.

Insight 2: Monetization and Perception

Ads or in-app purchases do not significantly harm user perceptions and can even be seen positively if implemented thoughtfully.



Design non-intrusive ad placements and ensure in-app purchases provide genuine value.

Highlighting these features in app descriptions can set the right user expectations.

● Implications and insights

Insight 3: Data -Driven Prioritization

Features such as app size, free/paid status, and category have minimal impact on ratings compared to user engagement metrics.



Prioritize fast, responsive design to enhance user satisfaction.

Include features that facilitate user interaction, such as social sharing or loyalty rewards.

Insight 4: Non-Linear Feature Effects

Simple linear assumptions about user behaviors may not capture complex interactions between features and outcomes.



Advanced analytics tools needed to explore nuanced relationships in user data.

Incorporate these findings into predictive models for targeting high-value users effectively.

Challenges we faced

Data Size and Imbalance

- **Challenge:** Over 2 million records caused memory issues, particularly for Random Forest.
- **Resolution:** Downsampling techniques to maintain class balance for categorical variables and prevent performance degradation.

Limited Computational Resources

- **Challenge:** Google Colab's memory constraints limited training on the full dataset.
- **Resolution:** Enabled GPU runtime for efficient training with XGBoost. For Random Forest, downsampling was applied due to memory limitations.

● Challenges and limitations

Feature Encoding and Non-Numeric Data

- **Challenge:** Non-numeric features like "Currency" and app categories required transformation.
- **Resolution:** One-hot encoding to convert categorical variables into numerical formats suitable for ML models.

Residual Bias in Categories

- **Challenge:** Systematic biases in residuals for categories like "Books & Reference" and "Finance."
- **Resolution:** Highlighted need for further refinement - feature interactions and advanced modeling techniques - to capture unique category characteristics.

Overfitting Concerns

- **Challenge:** Risk of overfitting with complex models like XGBoost.
- **Resolution:** Cross-validation, hyperparameter tuning, and regularization techniques to ensure generalization and prevent overfitting.

Library Compatibility Issues

- **Challenge:** Deprecated or incompatible library functions, such as `plot_partial_dependence`.
- **Resolution:** Switched to SHAP analysis for feature interpretation, enabling insights into feature contributions while maintaining compatibility.

What can we do in future?

Incorporate Additional Data Sources

- Include social media sentiment, user feedback, or app update details to improve the model's ability to predict app ratings more accurately.
- Use real-time data for predictive models to provide dynamic insights into app performance.

App Categories based Custom Insights

- Identify what drives high ratings for niche categories like "Books & Reference" or "Finance."
- Customize strategies for categories exhibiting systematic biases, improving their market performance.