# GOOGLE PLAYSTORE APP RATING PREDICTION

Machine Learning Project

# PROBLEM STATEMENT

📱 Google Playstore has 10K+ apps with diverse features.

⭐ Ratings are critical to app visibility & success.

💡 Can we predict ratings from metadata using ML?

# DATASET OVERVIEW

## Basic Info

- **Source:** Kaggle (Google Playstore Dataset)
- **Rows:** ~10,841 apps
- **Columns:** 13+ (App, Category, Rating, etc.)

## Key Features

- **App** – App name
- **Category** – App type (e.g., Game, Tools)
- **Reviews**, **Size**, **Installs**
- **Type** (Free/Paid), **Price**, **Rating** (Target)

## Target Variable

- **Rating** (1.0 to 5.0)
- **Goal:** Predict numeric rating using app metadata

TOOLS & TECHNOLOGIES

# TOOLS & TECHNOLOGIES

**Pandas/Numpy**

Data cleaning & manipulation

**Matplotlib / Seaborn**

Data Visualization

**Scikit-learn**

ML models, preprocessing

**XGBoost**

Gradient boosting model

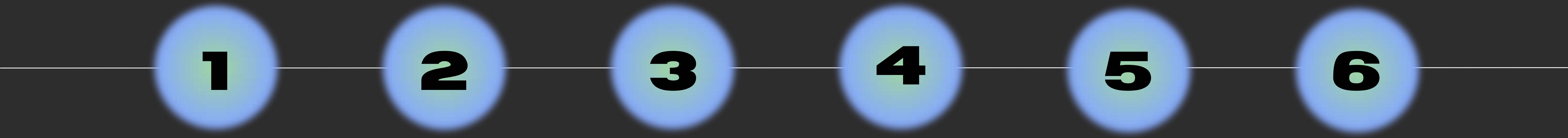
**RandomForest**

Best model for prediction

**Google Colab / VS Code**

Code environment

# DATA PREPROCESSING



1. Removed missing values
2. Cleaned numeric columns
3. Encoded categorical features using LabelEncoder
4. Log-transformed skewed features
5. Extracted time features from Last Updated
6. Handled outliers and irrelevant entries

# ML MODELS USED

ML MODELS USED

## RandomForest

Ensemble of decision trees, handles overfitting well

## Tuned RF

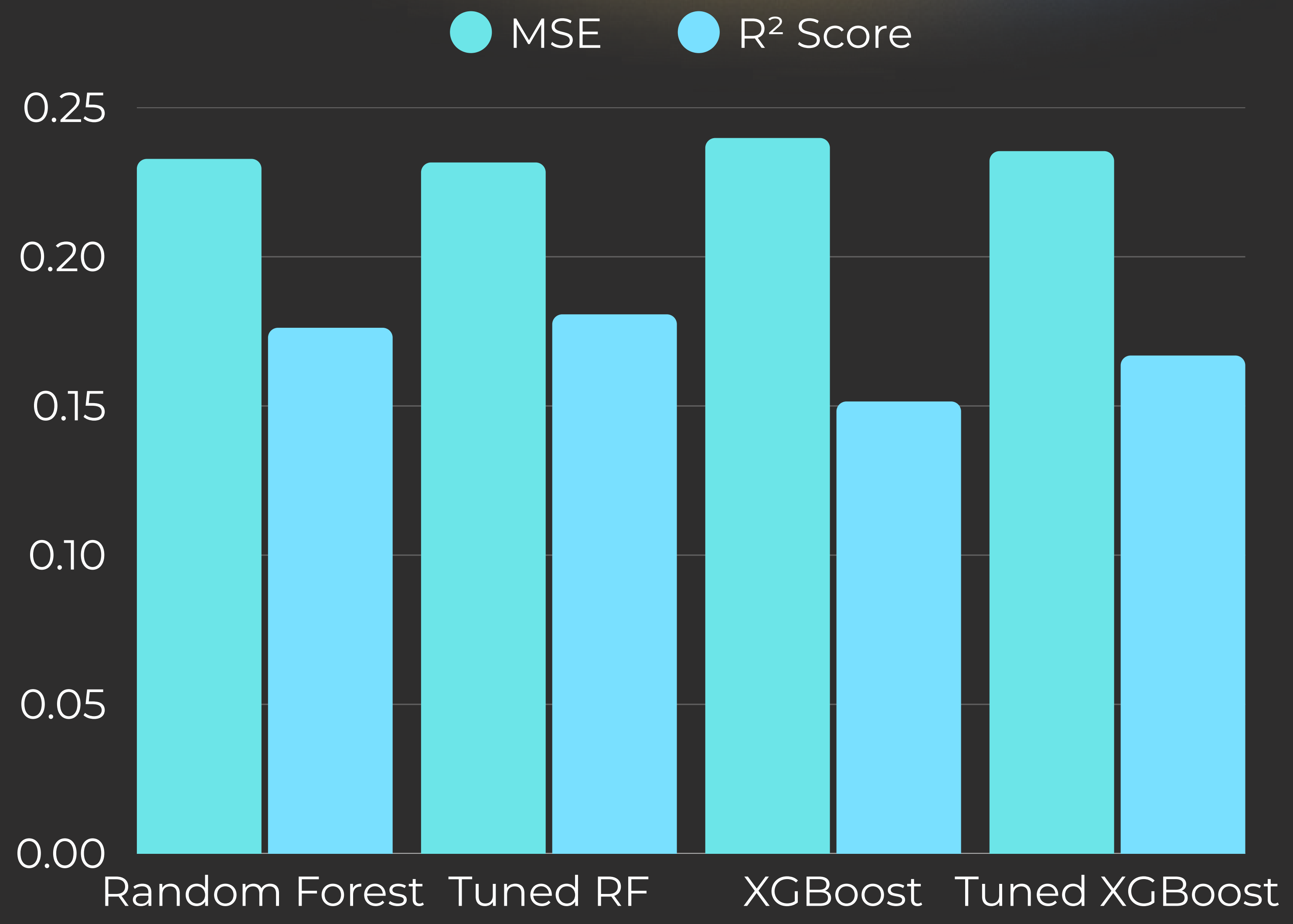
Hyperparameter tuning tool on RandomForest

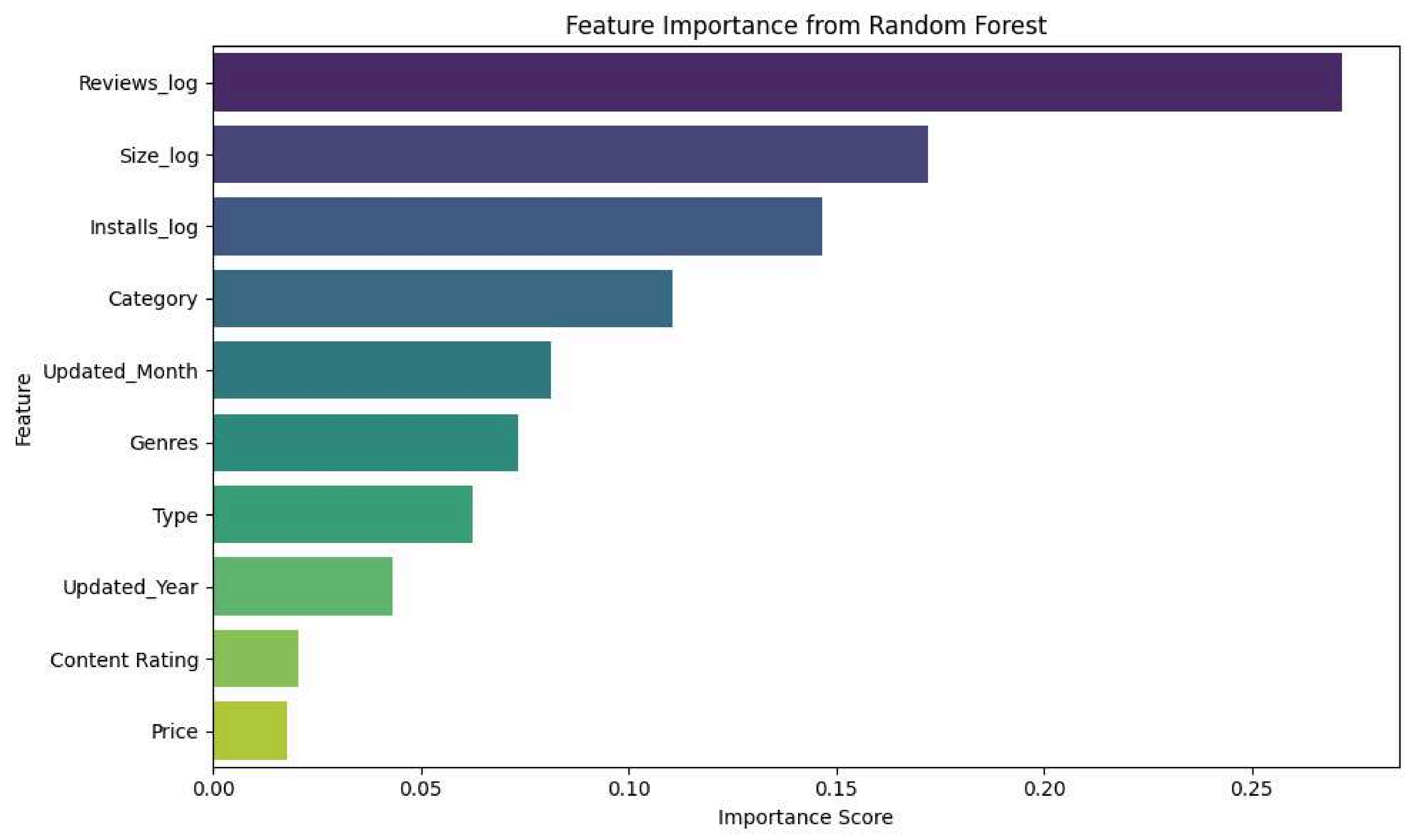
## XGBoost

Boosted tree-based model, very accurate & fast

# MODEL EVALUATION

Tuned Random Forest achieved the best overall performance, with the lowest MSE and highest R² score, making it the final model selected.



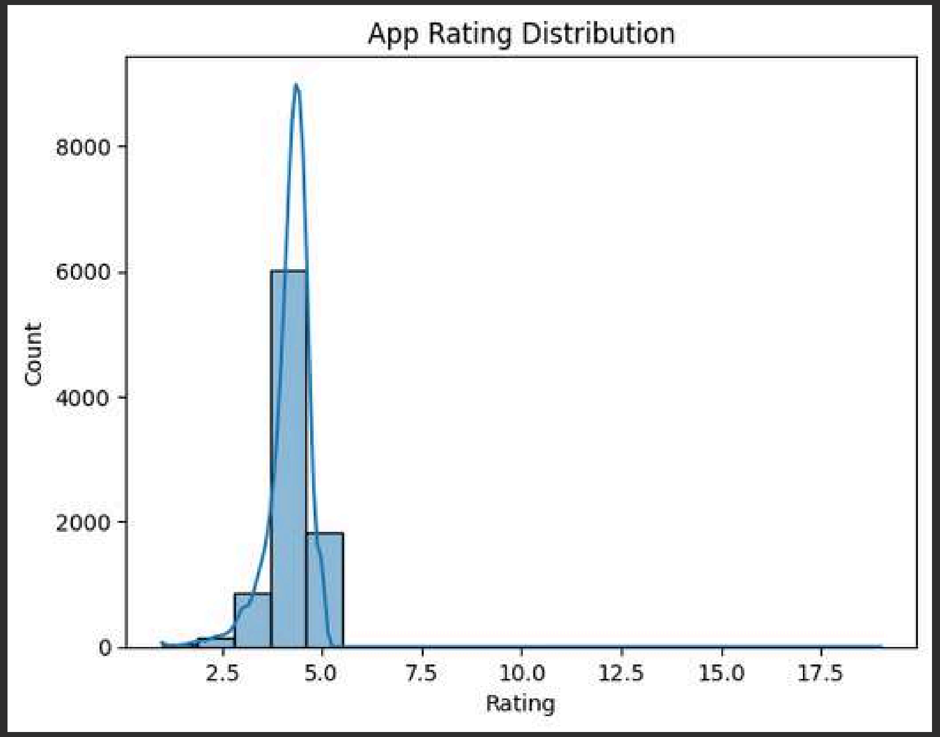


- **Reviews_log, Size_log, and Installs_log are the top 3 features**
- **Content-based and temporal features (Category, Updated_Month) also contribute**
- **Price and Content Rating had the least impact**

# FEATURE IMPORTANCE

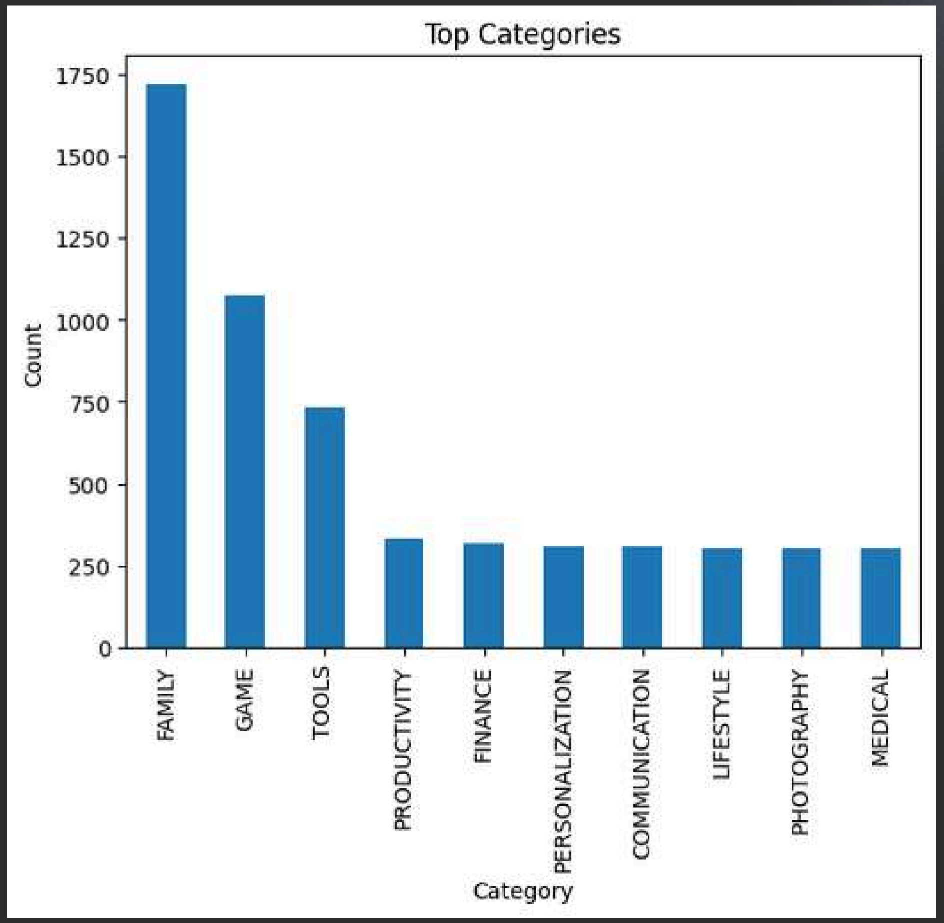
High review count and install volume strongly influence app rating predictions, indicating user engagement is a key driver.

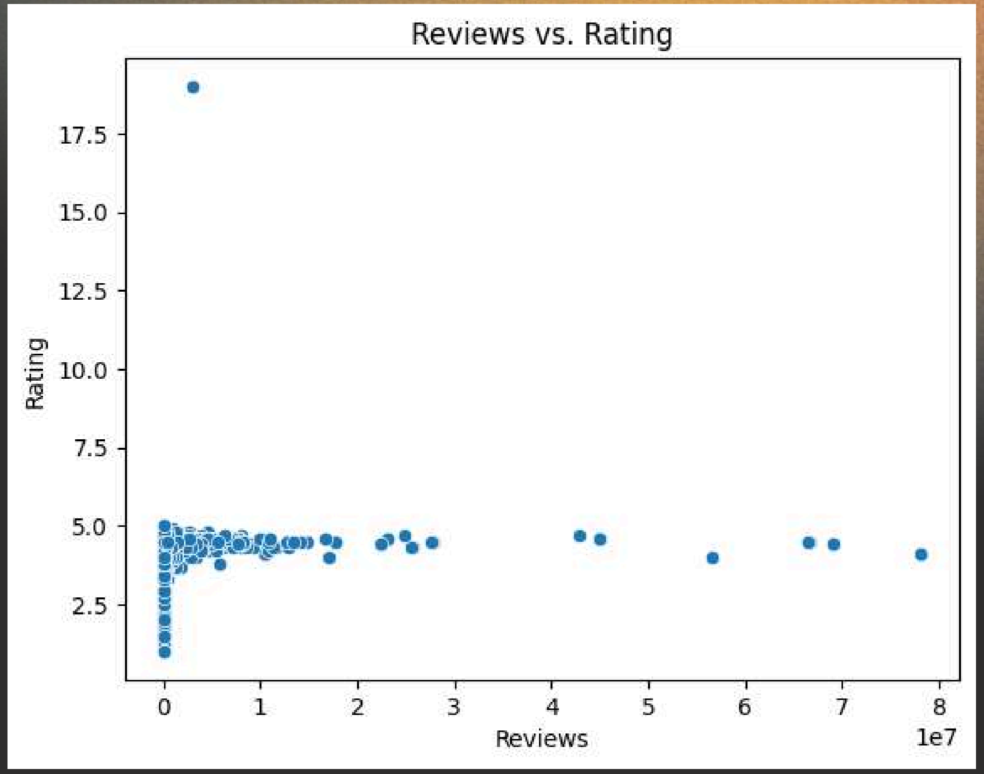
# INSIGHTS FROM EDA



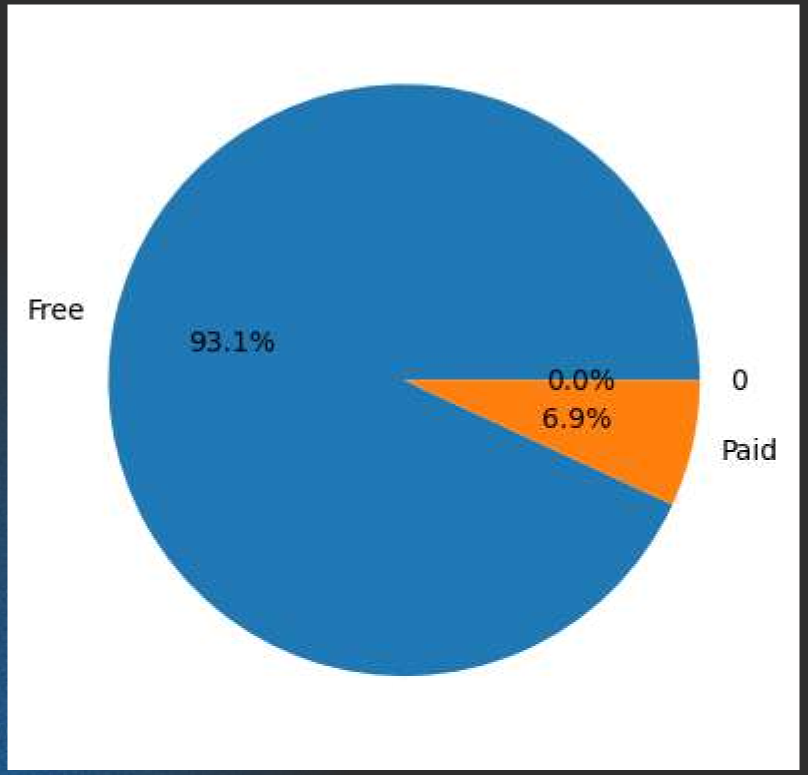
Most app ratings cluster between 4.0–4.5, indicating generally positive feedback.

The FAMILY, GAME, and TOOLS categories dominate in volume.





Apps with more reviews generally have stable ratings, though outliers exist.



A striking 93% of apps are free, showing strong user preference.

# CONCLUSION

- Built an ML pipeline to predict app ratings using Google Playstore data
- Preprocessing included handling missing values, transforming skewed data, and encoding categories
- Random Forest with GridSearchCV achieved the best performance:
  - MSE: 0.2316
  - $R^2$ Score: 0.1807
- Key features: Reviews, Size, Installs, Category

# FUTURE SCOPE

- INCORPORATE TEXTUAL DATA LIKE APP DESCRIPTIONS, USER REVIEWS (NLP)
- INCLUDE MORE METADATA SUCH AS PERMISSIONS, UPDATE FREQUENCY, OR DEVELOPER REPUTATION
- TEST OTHER MODELS LIKE NEURAL NETWORKS OR CATBOOST
- DEPLOY THE MODEL AS A REAL-TIME WEB APP (USING FLASK OR STREAMLIT)

THANKS

# THANKS

Thank you for your time!
Looking forward to your feedback.

| Name | Email Address |
|---|---|
| Srshti Jain | srshtijain17@gmail.com |