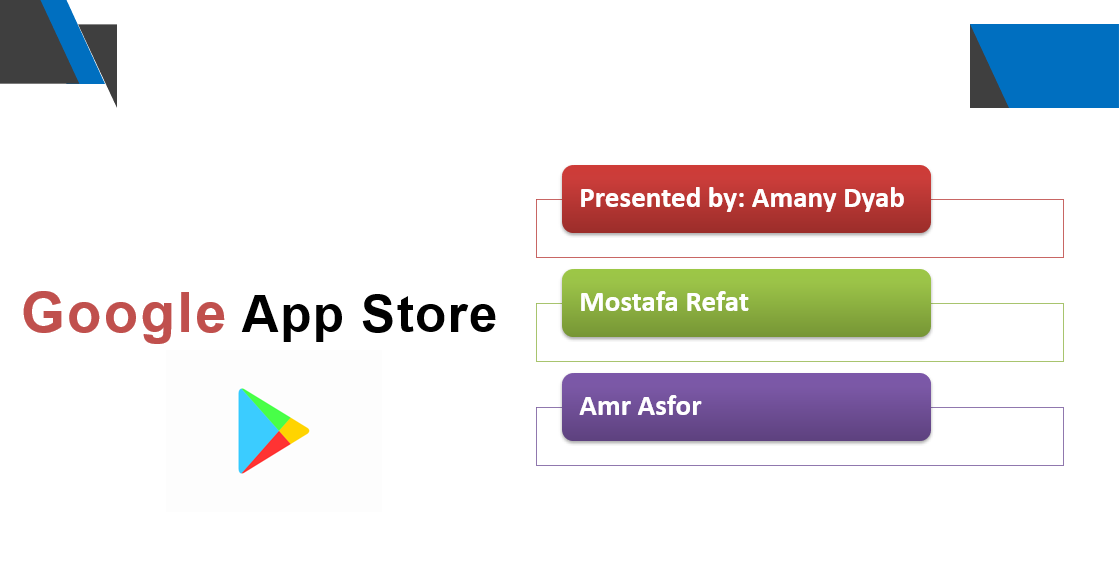
# **Google Play Store Apps- Data Analysis and Ratings Prediction**



# **Abstract**

Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps using machine learning Algorithms. I have tried to perform Data Analysis and prediction into the Google Play store application dataset that I have collected from Kaggle. Using Machine Learning Algorithms, I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application.

# **Introduction**

In this project, we investigate the different variables of Apps on Google Play Store that affect the application and the top 100 most relevant user reviews. We attempt to use our analysis to answer the following questions:

What makes an app popular?

What are some interesting trends that we can observe about user behavior & sentiment on app usage?

Can we predict app popularity if given a set of features about the app?

Mobile app market has grown ~20% in the last 4 years. Android apps market comprises 90% of its market share. Our analysis has the potential to be scaled and applied to identify and solve the following problems for the much wider mobile app industry:

Prediction of app popularity to gauge revenue generated & optimize investment strategy for app development

Identifying untapped segments in the app market

Identifying fake/junk apps that spam the play store

# **Data Cleaning**

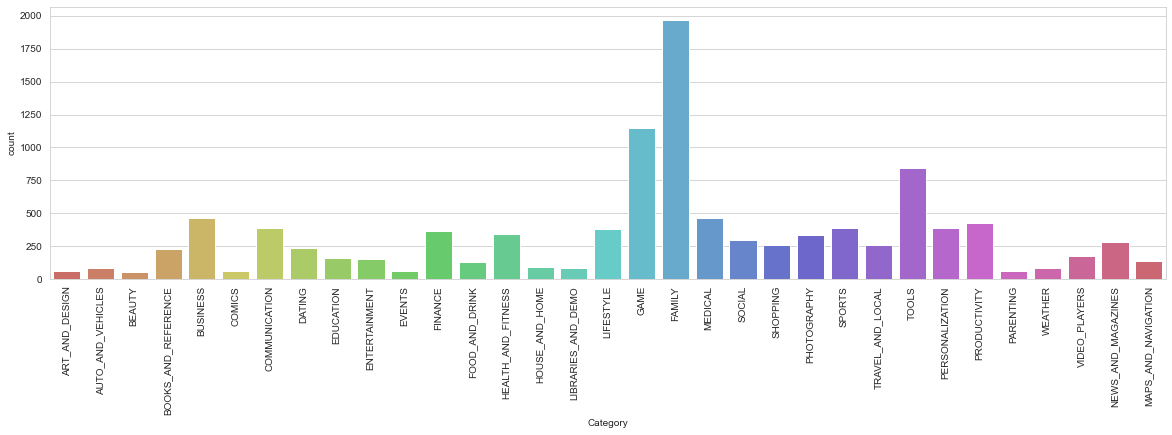
We started the analysis by creating a data dictionary to understand the structure of the dataset and what each feature represents. Post that, we handled the missing values in some of the columns by either dropping the rows or imputing them with the mode values, depending on the percentage of nulls in each feature. We transformed and some of the columns like Installs, Size and Price to numeric type for ease of analysis.

# **Exploratory Data Analysis**

After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features.

# **Android Market Breakdown**

We broke down the apps by category and found that the Family and Game categories have the highest market prevalence. The Business, Tools and Medical apps are also catching up. We also checked how each category performed in terms of number of reviews, size and installation count. We found that the gaming category has the maximum number of reviews and install count but the family apps consume more space in the play store.

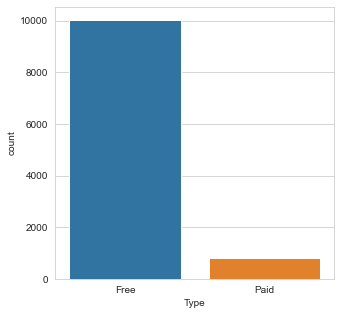


# **Sizing Strategy**

We analyzed the sizing distribution of the top-rated apps (rating greater than 4.5) and observed that most top-rated apps are optimally sized between ~2MB to ~40MB i.e., neither too light nor too bulky. We found that the Game and Family categories have the highest number of bulky apps. We also observed that despite this, these bulky apps are fairly highly rated indicating that they are bulky for a purpose. Majority of the paid apps that are highly rated have small sizes which implies that most paid apps are designed and developed to cater to specific functionalities and hence are not bulky. Users prefer to pay for apps that are light-weighted. A paid app that is bulky may not perform well in the market but it also depends on the category of the app.

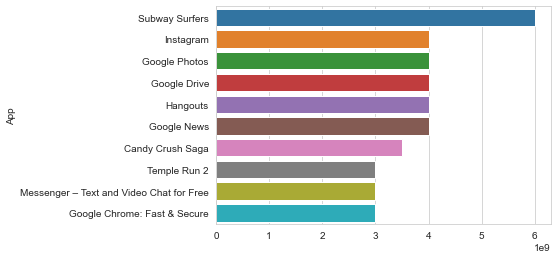
# **Pricing Strategy**

The initial intuition for pricing strategy for apps to be popular would be to make them free. That is confirmed by the data as free apps dominate the store at 92.6%, making it over 12.5 times their paid counterparts. It follows that cheaper apps are also more popular which is supported by the fact that 90% of paid apps are less than $10 and 80% of apps are less than $5. Categorically, the most expensive apps are Family and Medical apps which would suggest that people are willing to pay more for medical apps, possibly due to a belief in a reliability or highly specialized applications. When cleaning the data, one of the major outliers we found were useless apps priced over $250 which we considered ‘junk’ apps. The names of these apps were typically something to the effect of “I am rich” which would do nothing other than prove the user paid an exorbitant amount it. An interesting insight for app developers for games is that all of the games are within $20.



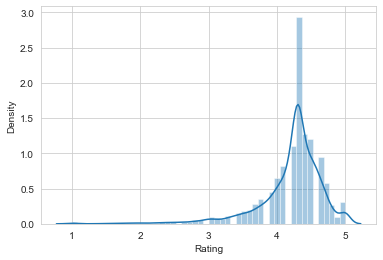
# **Install Count and Rating**

To answer the question of whether highly installed apps are rated better, we plotted the install count and rating. We saw that there was no pattern in apps that had installs less than 500 Million. However, for apps with installs greater than this threshold, the minimum rating rises to 3.7 out of 5. To confirm this, we bucketed the apps into bins based on number of installs and plotted the minimum rating for each bucket. We observed that the lowest rating for an app increases with the number of installs. We can say that popular apps are less likely to be rated badly.



# **Rating Distribution**

The rating distribution revealed that most apps perform reasonably well with an average rating of 4.17. We broke down the average rating by category to check if any category performs exceedingly good or bad. We conducted a One-way Anova Test and confirmed that the average ratings across categories is statistically different. The Health and Fitness and Books and Reference produce the best apps with 50% apps having a rating greater than 4.5. Interestingly, half of the Dating apps have a rating lower than the average.



# **Machine Learning**

We proceeded to predict the popularity of app based on its features and chose the install count as the measure of popularity. We split the number of installs into four buckets – Not so popular, Intermediate Popular, Popular and Extremely Popular, so that we have a classification problem at hand.

# **Feature Selection**

We plotted a Pearson Correlation Matrix which showed a moderate positive correlation exists between the number of reviews and the number of installs. Our initial analysis also revealed that the size of the app affects the rating. Most of the top-rated apps were sized between 2 Mb and 40 Mb. Intuitively, the number of reviews and size of the app would explain the majority of the popularity of the app. The rating distribution showed that categories had statistically different mean ratings which in turn contributes to the popularity. We decided to use the features - ‘Installs’, 'Reviews', 'Category', 'Content Rating', 'Type', 'Genres' and 'Size'. We dropped the features – Android Ver, Current Ver and Last Updated since these did not have any significant relationship with the install count.

# **Implementation**

We implemented three classification models to predict the popularity of the app, amongst which Random Forest with tuning gave us the best accuracy with 93.4 % accuracy. KNN Regression – 93.1% accuracy and Linear regression– 92.7% accuracy. We used Randomized Search CV to loop over a set of random values for the parameters of the model. We ran the grid search and chose the optimal number of tress, number of features at each split, maximum number of levels and the method of selecting samples. Using these parameters, we predicted the popularity of the app on our test set and obtained an accuracy of 93.4%.

# **Findings and Conclusion**

Developers should aim to keep the apps optimally sized (between 2 MB to 40 MB) to increase the likelihood of it being popular.

Paid apps designed to fulfil specific functions tend to be small sized and are more likely to meet user expectations.

Majority of the apps should be free. For apps with no ads, apps can be priced under $10.

Free apps tend to outperform paid apps both in volume, install counts and ratings.

There is a positive correlation between installs and ratings

The category (Game) is a potential unsaturated space for app developers. Apps in this category tend to have high installs.

Highly installed apps such as Communication, Productivity & Photography are not as highly rated indicating dissatisfaction among customers and potential untapped market segments.

For paid apps, users review harshly.

Basic sentiment analysis revealed some common issues like loading time and inconvenience with apps may contribute to a negative rating. Usability was one of the main reasons apps were rating positive.