Predicting App Popularity/Downloads

Vaibhav Nagar
Avi Walyat
Nishant Kushwaha
Haoxiang Yi
Sian Cheng
Cheng-Ya Liou

Problem overview:

Businesses are actively seeking effective strategies and ideal parameters to amplify app downloads and user acquisition. We've delved into an extensive Google Play Store dataset featuring 13 columns and 10,841 rows to address this challenge. Through rigorous analysis, we strive to uncover unique patterns and critical variables. These insights will guide strategic decisions for increasing app downloads and shape our multi-class classification approach. This predictive model targets installations on the Google Play Store, providing a finer understanding of user acquisition nuances.

Importance of the problem:

In the ever-evolving app market landscape, businesses face the imperative of comprehending and surmounting specific challenges to foster growth. One pivotal aspect revolves around customer engagement, underscoring the significance of elevating an app's visibility. This strategic move holds the potential to extend outreach to a broader audience, thereby expanding the user base. Embracing a data-driven methodology bolsters this approach further. Through aggregating and meticulously analyzing app-specific data, enterprises can discern distinctive attributes and trends that strike a chord with users, effectively amplifying app popularity. Concurrently, the effectiveness of marketing strategies emerges as another linchpin for triumph. Intelligently embedding promotional ads within the most frequented apps can propel market penetration, guaranteeing an expanded scope and heightened brand recognition. Amid these immediate advantages, a more overarching theme of investor interest comes to the fore. Companies that can demonstrate a robust tally of app downloads showcase their product's inherent allure and emerge as compelling prospects for potential investors. Mastering the art of navigating these intricately interconnected challenges and seizing the attendant opportunities is indispensable for firms harboring aspirations of carving a distinctive niche within the bustling app marketplace.

Data Description:

In our dataset, we have initially sourced information across 13 variables: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version, and Android Version. However, to streamline our analysis and avoid redundancy, specific columns were deemed unnecessary and thus excluded. Specifically, we opted to omit the 'App' (app's name) and 'Genres' columns, as the 'Category' column provides similar categorical information. We also excluded the 'Last Updated' and 'Current Version' columns for our analysis.

Exploratory Data Analysis:

App Size and User Ratings: Striking a Balance for Optimal Engagement: In the realm of app performance and user engagement, the interaction between app size, installs, and ratings take center stage. Despite installs being distributed evenly across a range of app sizes, a consistent average rating of around 4.2 signals a predominantly positive user sentiment. Notably, apps falling within the 0-100M size bracket stand out with higher ratings, implying the existence of an ideal size range that fosters a harmonious equilibrium between user satisfaction and installation figures. This dynamic underscores the significance of optimizing app size to enhance user engagement and ratings, illuminating a crucial facet of the intricate app landscape.

Pricing and User Sentiment: The connection between app pricing and user ratings reveals a nuanced paradigm that sheds light on user behavior and preferences. Our analysis uncovers a trend where higher-priced apps receive lower ratings while moderately priced apps earn the most positive feedback. This insight underscores the discerning nature of users who prioritize perceived value and quality over monetary cost. This interplay between pricing and user sentiment emphasizes the importance of aligning app pricing with user expectations to maximize engagement and foster positive reviews, illustrating the intricate interplay between pricing strategies and overall app performance.

Android Versions and User Experience: A Complex Relationship: The correlation between Android versions and user ratings offers valuable insights into the evolving user experience landscape. While a modest positive correlation links higher Android versions with improved ratings, the predictive power of the Android version in determining ratings remains constrained. This finding underscores the complex web of factors influencing user satisfaction within the multifaceted app ecosystem. This interplay between Android versions and user sentiment highlights the need for app developers to consider a holistic approach to user experience optimization, encompassing not only

the Android version but also a range of other factors that collectively shape user engagement and app success.

Unveiling Key Correlations and Data Integration: A heatmap was constructed to encapsulate the correlations between various variables. Particularly regarding Installs, notable correlations were identified with "Paid," "Size," and "Review." While these insights provide valuable direction, all variables will be retained for model input, allowing a comprehensive understanding of their collective impact on overall predictions.

Modeling:

The "Installs" target variable was categorized into six distinct groups using a strategic approach. During the modeling phase, the dataset was divided into an 80-20 split for training and testing. This exploration involved a range of models, including both linear (Logistic Regression) and non-linear (Random Forest, Boosting) approaches.
The accuracy of Logistic Regression at 40% highlighted notable misclassifications across categories, raising concerns about its practical utility. Among the non-linear models, Boosting exhibited a substantial accuracy increase to 77%, but this came at the cost of overfitting.
The most effective model emerged as Random Forest with a 78% accuracy rate. This choice significantly improved upon the baseline accuracy of 21%, showcasing an impressive increase of nearly 300%. This underscores the model's capability in providing enhanced predictive performance across the given dataset.

Solutions and Learnings:

Insights:

- **Variable Importance:** In analyzing the dataset, three key variables have emerged as highly influential predictors for predicting app installs. These variables are Reviews, Size, and Ratings. This signifies their significant impact on the success of app installations. These variables contribute the most to the model's predictive power and indicate that they are important factors that potential users consider when deciding to install an app.

- **Price vs. Rating vs. Installs:** The investigation into the relationship between app pricing, ratings, and installs has revealed a critical

interplay. The pricing of an app has a direct impact on both the number of app installations and the ratings it receives. Higher pricing correlates with lower app installations. However, intriguingly, apps priced within the range of up to $100 demonstrate higher ratings despite their pricing. This suggests that users are willing to pay for quality apps, even at a premium price, provided they meet their expectations.

Recommendations:

- **Incentivizing Reviews:** Given the undeniable significance of user reviews in driving app success, it is recommended to implement incentives encouraging users to leave reviews for the apps. These incentives range from rewards, in-app benefits, or acknowledgments, aiming to boost user engagement and traction. By actively encouraging reviews, the apps benefit from increased visibility and credibility within the market.

- **Android Version Compatibility:** The analysis has highlighted the ongoing relevance of older Android operating system versions, which still maintain a considerable user base. Consequently, ensuring compatibility with these older Android versions is advisable when developing and updating apps. This strategy will enable the apps to tap into a broader user demographic and maximize their reach.

Conclusions:

- **More Predictors and Data Points:** To enhance the predictive model's precision and effectiveness, adding additional predictors is recommended. Factors such as the frequency of app updates, in-app purchases/subscriptions, and geographical indicators should be considered for integration. By incorporating these factors, the model will likely yield more accurate predictions and better understanding of user behavior.

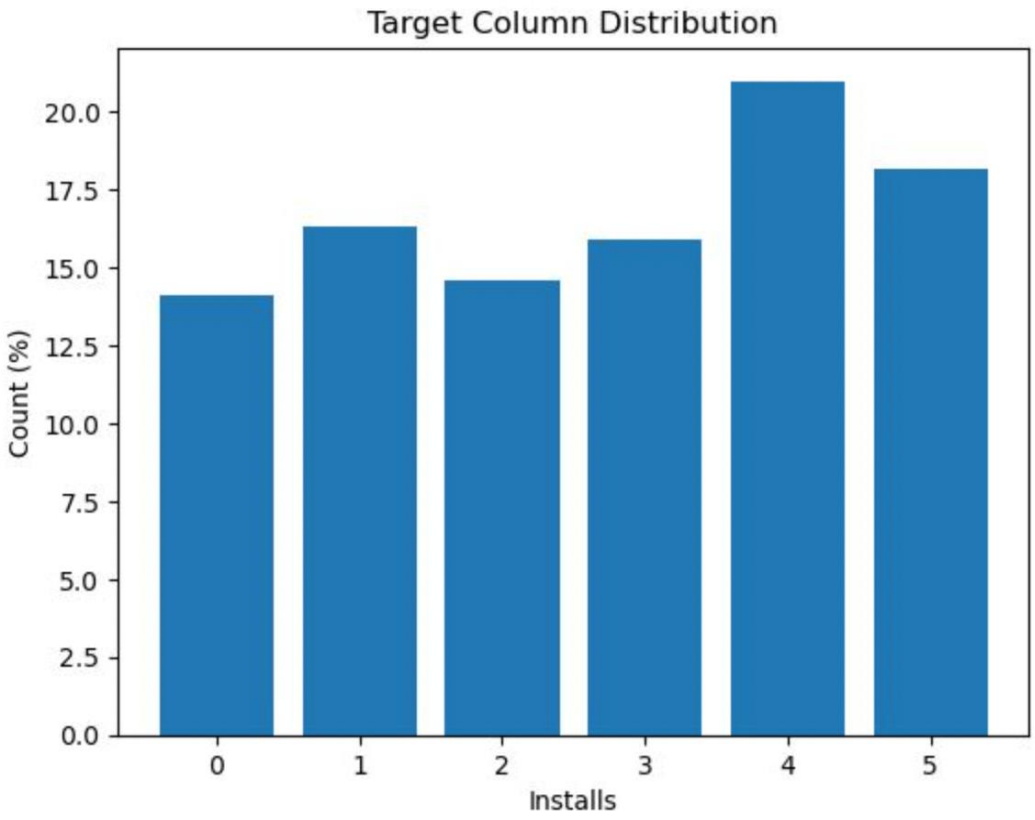
- **Conducting Sentiment Analysis:** User reviews provide a wealth of information to be harnessed for app improvement. By performing sentiment analysis on these reviews, it is possible to pinpoint specific strengths and weaknesses of the apps. This process aids in comprehending user preferences, allowing for targeted enhancements that align with user expectations. Sentiment analysis serves as a valuable tool for refining app features and functionalities.

APPENDIX

Data Description

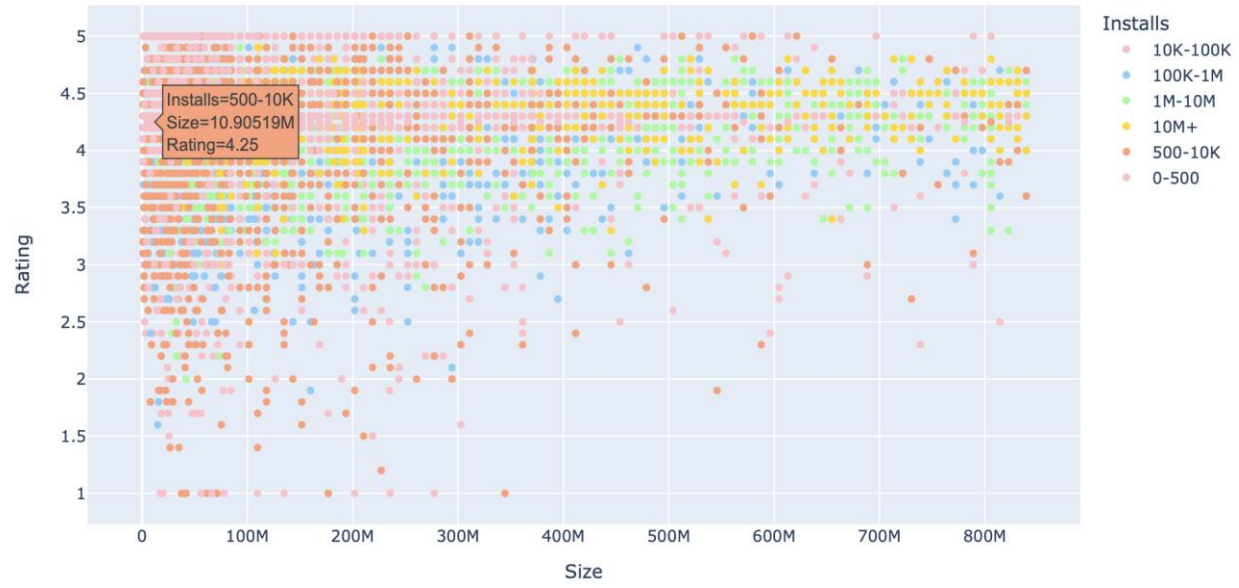
				Target Variable
CATEGORY	**RATING**	**REVIEWS**	**SIZE**	**INSTALLS**
Category the app belongs to ex: Family, Sport etc.	Overall user rating of the app ex: 4.1, 4.6 etc.	Number of user reviews for the app	Size of the app (Measured in bytes)	Number of app downloads/installs in 6 categories
PRICE	**CONTENT RATING**	**ANDROID VER**	**PAID**	**FREE**
Price of the app (Measured in dollars)	Convert into Binary: 0 = Restricted 1 = Everyone	Only keep minimum Android version, ex: 4.0 or 3.8 etc.	Dummy variable for paid app	Dummy variable for free app

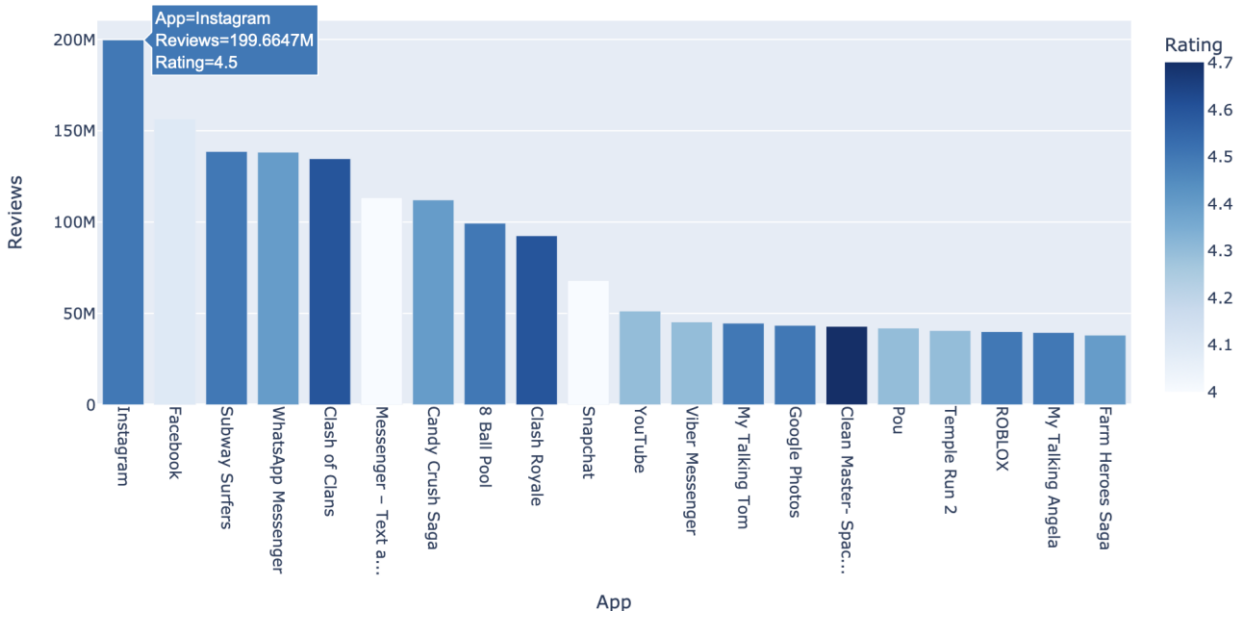
Target Variable Distribution



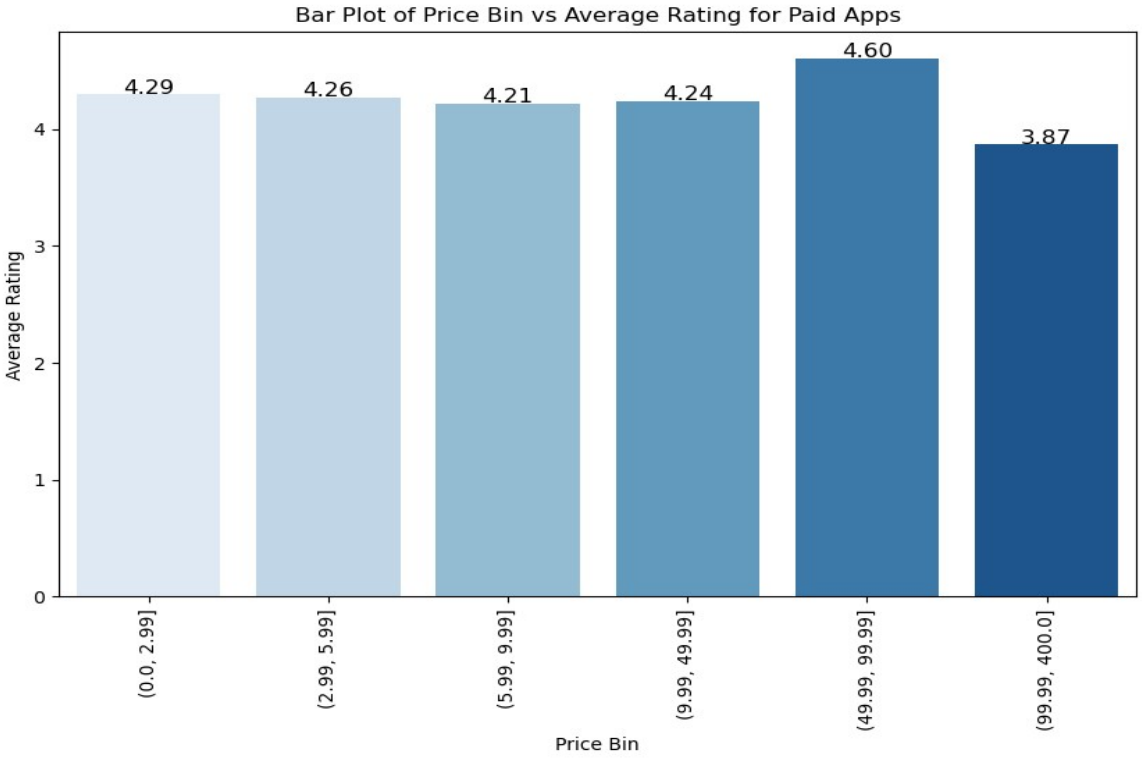
Exploratory Data Analysis

- **App Size and User Rating/Reviews**

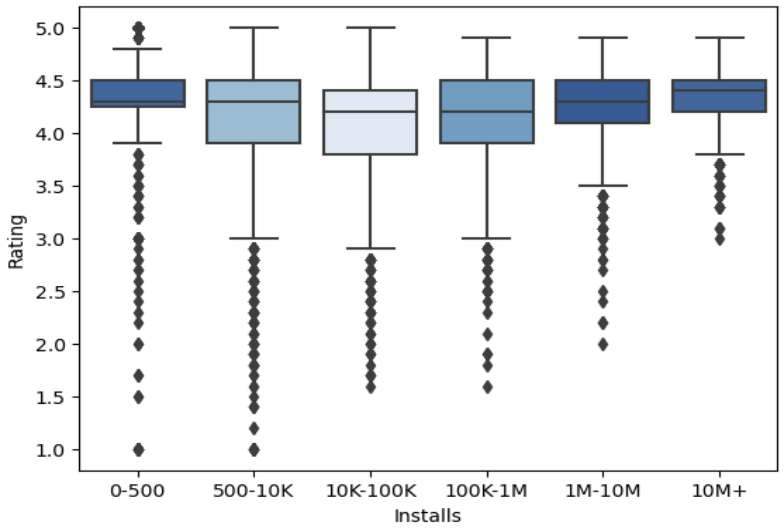




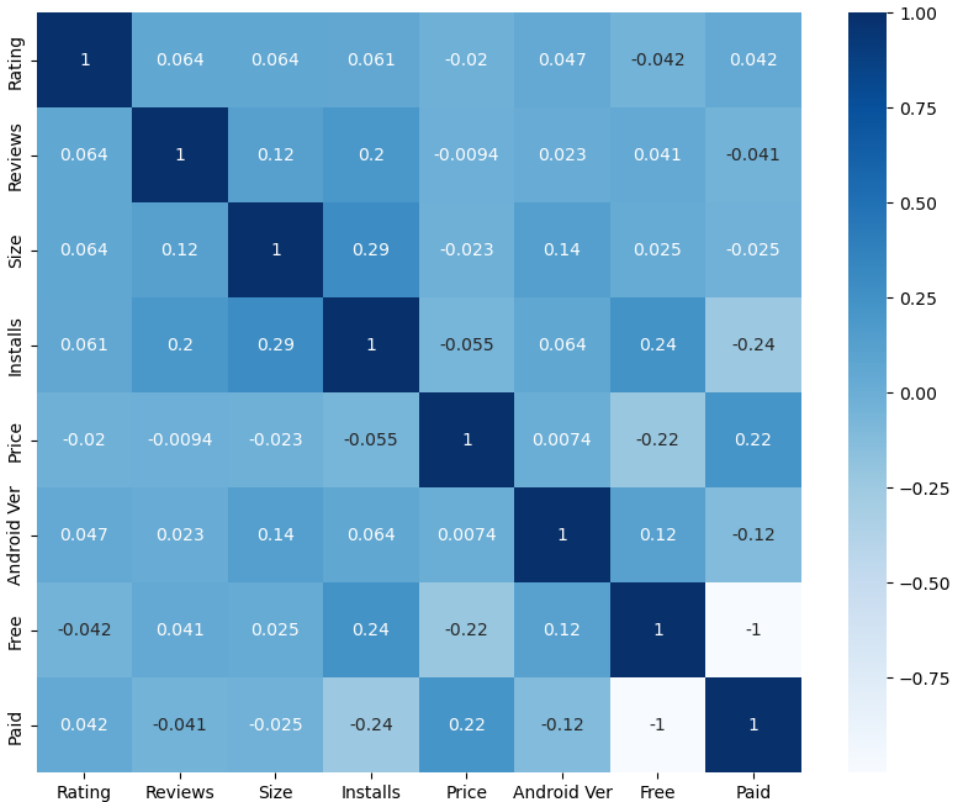
- **Pricing Bin and User Rating**



- **Android Versions and User Experience**



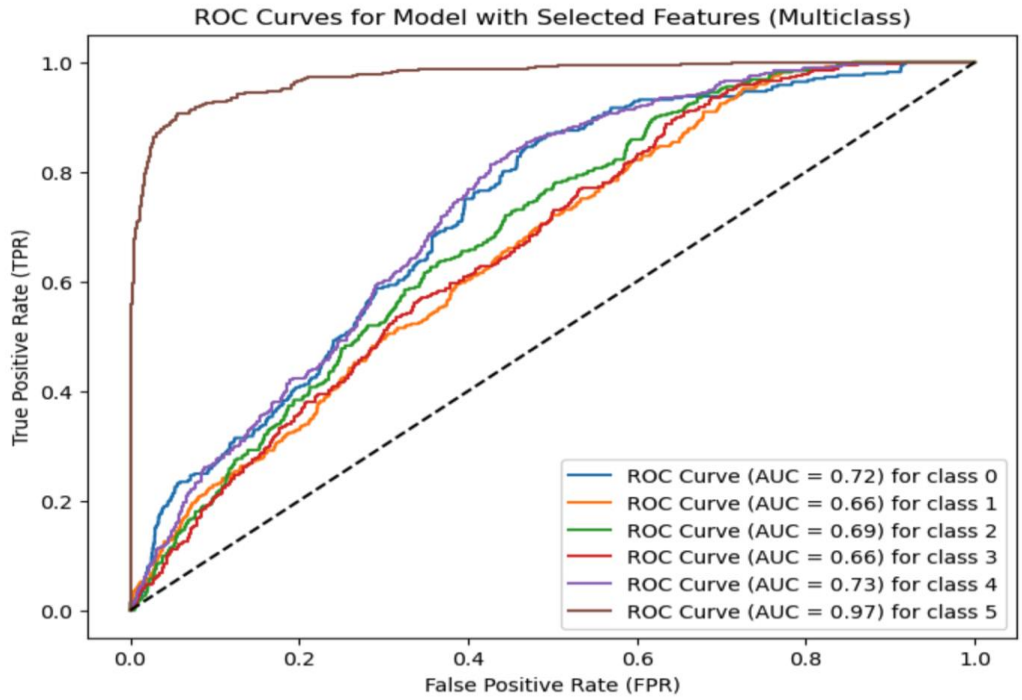
- **Unveiling Key Correlations and Data Integration**

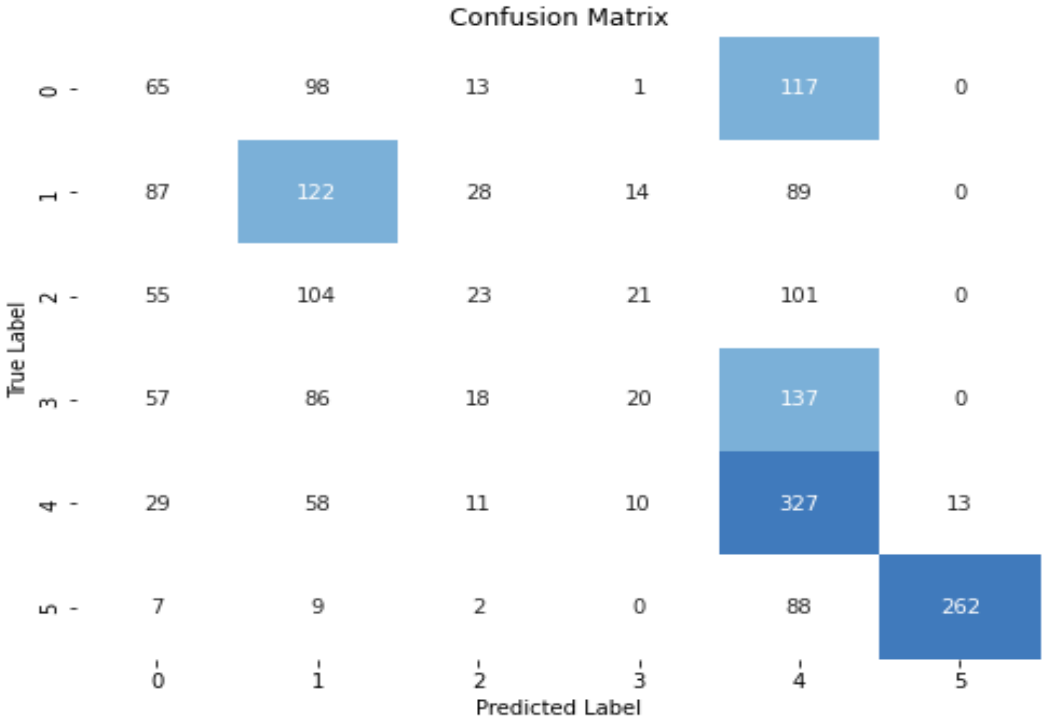


Models

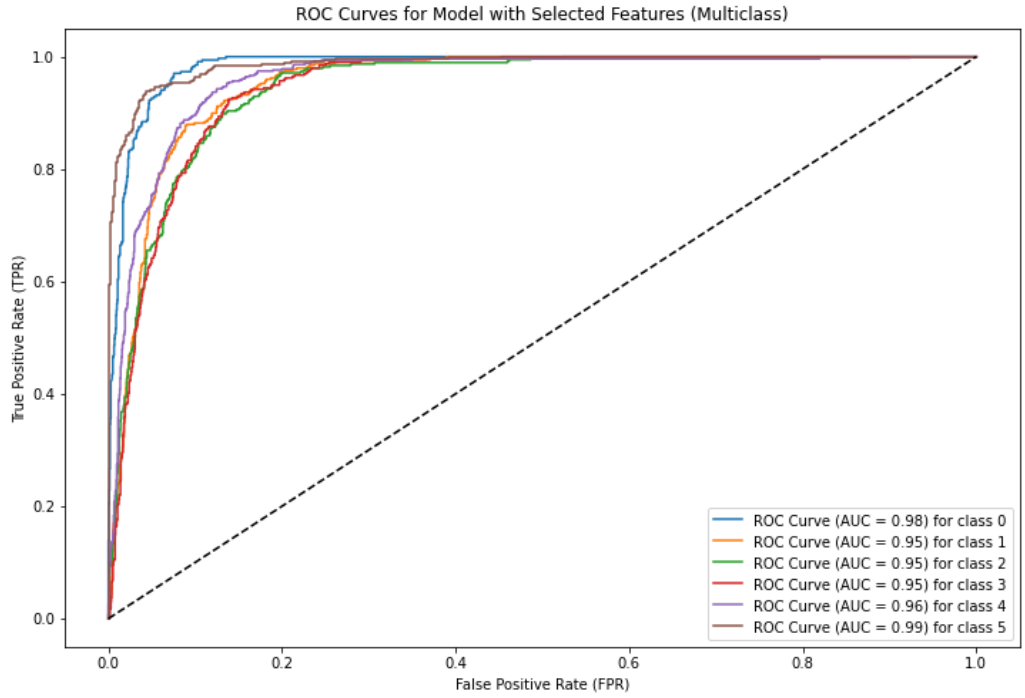
Metric	Logistic Regression	Boosting	Random Forest
Accuracy	0.4019	0.7706	0.7812
Precision	0.4097	0.7591	0.7632
Recall	0.3672	0.7565	0.7593

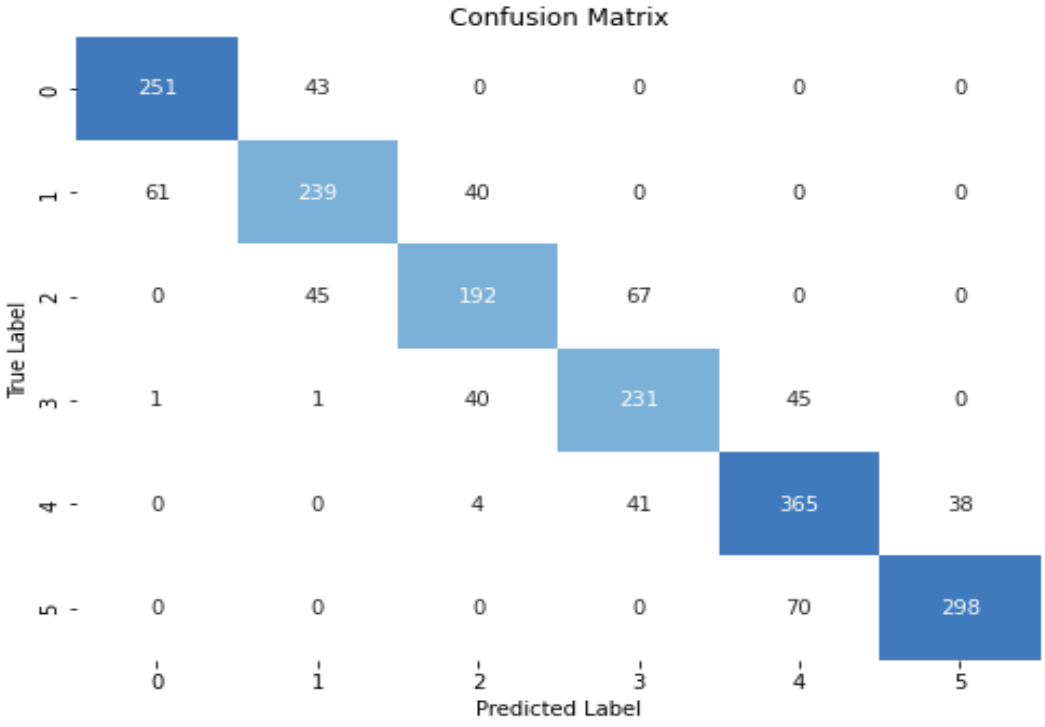
- **Logistic Regression**



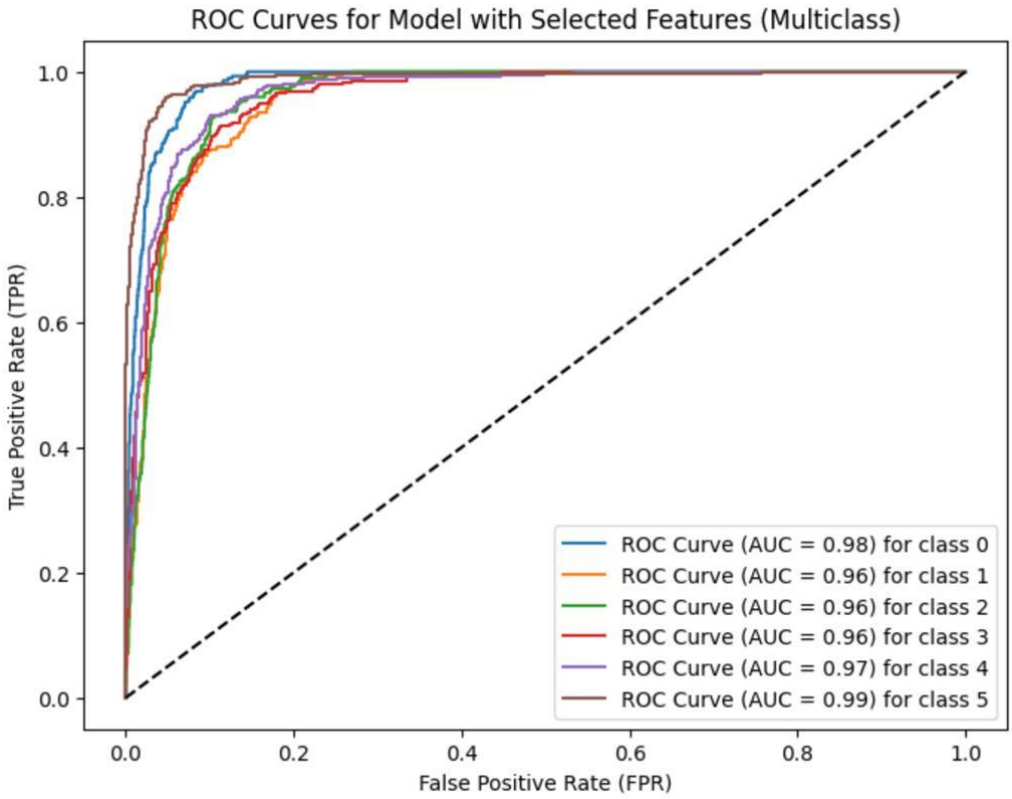


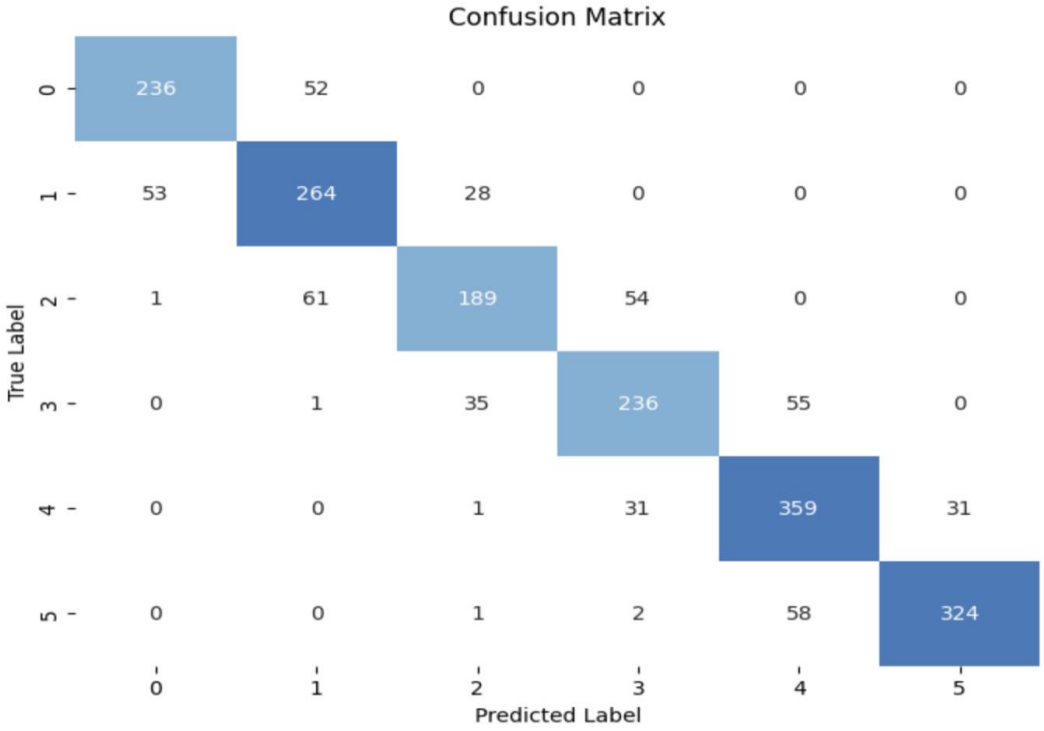
- **Boosting**

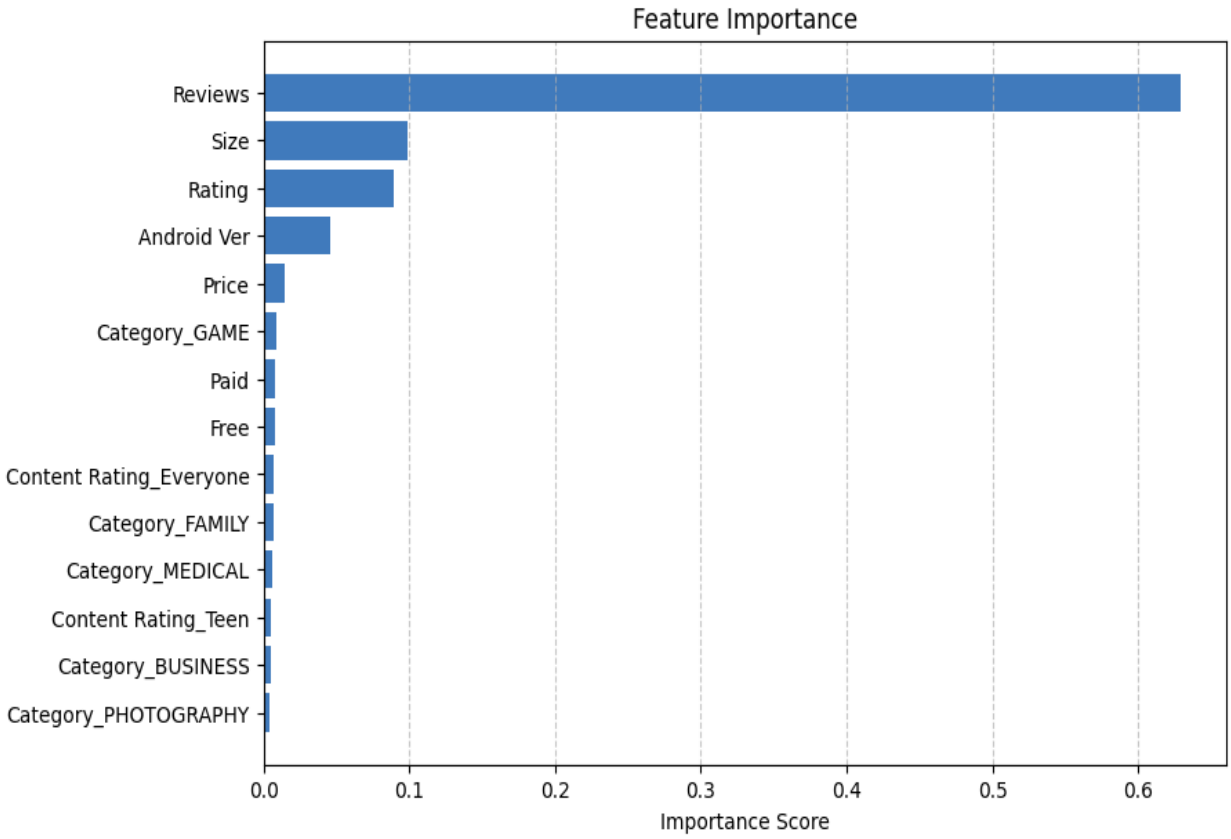

ROC Curves for Model with Selected Features (Multiclass)
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.98) for class 0
ROC Curve (AUC = 0.95) for class 1
ROC Curve (AUC = 0.95) for class 2
ROC Curve (AUC = 0.95) for class 3
ROC Curve (AUC = 0.96) for class 4
ROC Curve (AUC = 0.99) for class 5


Confusion Matrix
True Label
Predicted Label
0 251 43 0 0 0 0
1 61 239 40 0 0 0
2 0 45 192 67 0 0
3 1 1 40 231 45 0
4 0 0 4 41 365 38
5 0 0 0 0 70 298
0 1 2 3 4 5

- **Random Forest**







Sample Model Run

Rating	# Reviews	Size(MB)	Price	Android Version	Free	Paid	Predicted_Class	Actual_Class
3.8	44636	27.9	0	4	1	0	1M-10M	1M-10M
3.7	492	4.3	0	2.2	1	0	10K-100K	10K-100K
4.3	8	13.0	0	4.1	1	0	0-500	500-10K
4.3	2151039	75.0	0	4	1	0	10M+	10M+
4.3	355921	17.0	0	4	1	0	10M+	10M+
3.8	5868	20.0	0	4.1	1	0	1M-10M	100K-1M
4.2	19	5.4	0	4.1	1	0	500-10K	500-10K
4.8	216	19.3	$6.99	4	0	1	500-10K	10K-100K
4.5	787107	27.9	0	4	1	0	10M+	10M+
4.4	1690802	58.0	0	4	1	0	10M+	10M+