**Part I: Background and Motivation**

**Problem Statement:**

Google Play, better known to many as the Google Play Store, and formally known as the Android Market is the digital distribution center developed by Google for Android users to browse and download applications for their various needs. There are multiple public datasets that provide data for the Apple Store about consumer behaviors when it comes to application downloads, however, there aren’t as many available for the Google Play Store. Many people don’t realize that the Google Play Store application data has an immense prospective for steering application-making businesses to become successful. Across the globe, Android has a substantial amount of global market share in comparison to Apple or iOS users, and within the United States of America, Apple or iOS users hold a slight majority of the market shares. According to Jack Wallen in his 2021 article, “Why is Android more popular globally, while iOS rules the US?”, the global market share for Android is 72.2% in comparison to Apple or iOS, which is a mere 26.99%, while within the United States of America, Apple or iOS has a 59.17% market share to a 40.54% market share for Android (Wallen, 2021).

The most considerable reason for Android sweeping the global market is because of the cost. People across the globe usually browse and download their applications with their cellphones, and Android cellphones are much more inexpensive than Apple or iOS. Unlike many parts of the United States of America, many countries with significantly higher population don’t have the privilege of having disposable income to spend on purchasing cell phones and downloading applications, so there is more focus on the practicality of having cheaper cell phones with free or inexpensive applications for entertainment purposes, networking, shopping, ordering a ride or a meal, or keeping up with social media. Google Play Store has done much to consider the global arena and ensure that all consumers have been catered for when it comes to their products. However, there are many other aspects of extending the popularity that Google Play Store has globally, to the United States of America and other markets where they fall short in rising above the Apple or iOS products, such as developing its brand popularity in adhering as best as possible to the what people want the most from available products and services, and sponsoring more advertising of their products, especially the use of their applications in TV shows, movies and other commercials.

As developers and managers of Google Play Store, and as the company embarks on the first quarter of 2023, we know without a doubt that the Apple Store has already begun implementing strategies to increase their growth in the United States market and maintain their strong hold over the Google Play Store. Nonetheless, Google Play Store intends to combat Apple’s continuous and rapid growth by analyzing the dataset generated from the Google Play Store so that developers can create insights to influence and propagate the Android market and become the leading application establishment that drives consumers to choose Android over Apple. To achieve such success, Google Play Store proposes the analysis of the combination of variables that both the consumers and the company can settle on improving, to obtain the maximum number of downloads within the market along with the most desired application categories. Google Play Store aims to use the information to identify the shortfalls in the United States market and capitalize on refining those avenues.

**Problem Objective:**

The objective is to discover insights to expand the United States Android Market by increasing consumers' desires to select the products and services that Google Play Store offers over their competitors. Based on the dataset, the dependent variable for the success of the Google Play Store within the Android market is the number of installs that are conducted by consumers. As a company, we must examine the factors that increase the number of installations or downloads from the applications stores and which of the application categories are more frequently installed over others to determine where the most improvements can be implemented. With such information, Google Play Store should be able to increase revenues once the needs of consumers are identified and have been met or exceeded regarding the categories consumers are gravitating towards downloading regularly and other variables associated with such desires.

Beyond predictive models, after identifying the key to increasing downloads, Google Play Store could tackle improving the brand regarding quality control of the hardware, and sway consumers to choose Android with frequent appearances of the products on their favorite television shows, movies, and other commercials.

**Part II: Methodology**

**Data Methodology:**

The developers and managers of Google Play Store scoured the internet for credible data sources to solve the business problem of how the company can maximize downloads. Kaggle, a subsidiary of Google LLC, that provides published datasets for model exploration and building was the data source that was utilized because of the credibility of its datasets. The dataset examined was Google Play Store Apps, which provided two comma-separated values file, Google Play Store and Google Play Store User Reviews. The Google Play Store file included details of the applications on Google Play. The size of the Google Play Store was 1.36 MB with 10,800 number of observations. The dataset had 13 columns containing both string and numerical variables: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating and Genres. Majority of the variables didn’t have any mismatched or missing values except for the overall user rating of the application with 1,474 mismatched elements, number of reviews for the application with 1 mismatched element, and content rating with 1 missing value.

The App column provided the 9660 names of the applications that were found on Google Play such as “Garden Coloring Book”, “Animated Photo Editor”, “Job Korea – Career Search”, and “ROBLOX” being the most common application name. There were 33 categories belonging to the Google Play Store with ‘Family’ being 18% and ‘Game’ being 11%. Other categories included ‘Art and Design’, ‘Auto and Vehicles’, ‘Beauty’, ‘Books and Reference’, ‘Business’, ‘Comics’, ‘Communication’, ‘Dating’, ‘Education’, ‘Entertainment’, ‘Events’, ‘Finance’, ‘Food and Drink’, ‘Heath and Fitness’, ‘House and Home’, Libraries and Demo’, ‘Lifestyle’, ‘Medical’, ‘Social’, ‘Shopping’, ‘Photography’, ‘Sports’, ‘Travel and Local’, ‘Tools’, ‘Personalization’, ‘Productivity’, ‘Parenting’, ‘Weather’, ‘Video Players’, ‘News and Magazines’, and ‘Maps and Navigation’. The overall user Rating of the application contained only 9367 valid values with the ratings ranging from a minimum of 1 to a maximum of 19. The Reviews column provided the number of user reviews for the application from zero reviews to a maximum of 78.2 million reviews. The Size of the application varied with the device that was being used for the downloads, and the number of most common Installs or downloads were 1,000,000+, which accounted for 15% of the dataset, and 10,000,000+ described 12% of the dataset. The column with type provided only two categories, free or paid; 93% of the downloads were free and the other 7% were paid by customers. Zero dollars accounted for 93% of the price, equivalent to the free downloads, $0.99 accounted for 1% of the paid downloads, and 6% of the download prices varied. The content rating captured the age group the application targeted; ‘Everyone’, ‘Teen’, ‘Everyone 10+’, ‘Mature 17+’, ‘Adults only 18+’, ‘Unrated’, and nan, and the applications belong to multiple genres, for example, a musical family game could belong to Music, Game and Family genres, equivalent to the main categories.

The size of the Google Play Store User Reviews was 7.67 MB with 64,300 number of observations. The dataset had 5 columns containing both string and numerical variables: App, Translated Review, Sentiment, Sentiment Popularity, and Sentiment Subjectivity. Majority of the variables didn’t have any mismatched or missing values except for Translated Review with 5 missing values. The Google Play Store User Reviews was not used to generate any models but rather as a book of reference to solidify the interpretation of the models based on the reviews that consumers wrote in accordance with their Google Play Store application experience.

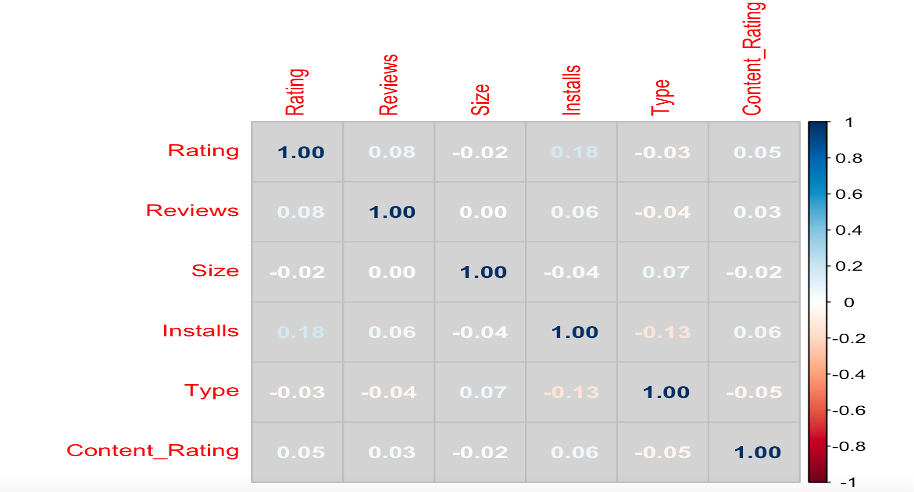
**Preprocessing and Analysis Techniques:**

Prior to any analysis, the Google Play Store dataset was cleaned using excel; replacing certain values of zeroes with N/As, so that the values within the dataset would be considered rather than be eliminated or viewed as null values, and duplicates were dropped. Python was the other external cleaning source to convert columns to strings and numerical values, while removing other special characters, so that the dataset being implemented in the models would not have any unnecessary format errors. The data cleaning in Python was further used to eliminate some extreme values from the data being scraped incorrectly, such as the 19 data points for ratings which normally on a zero-to-5-point scale, and the 34th category, not mentioned previously, of 1.9.

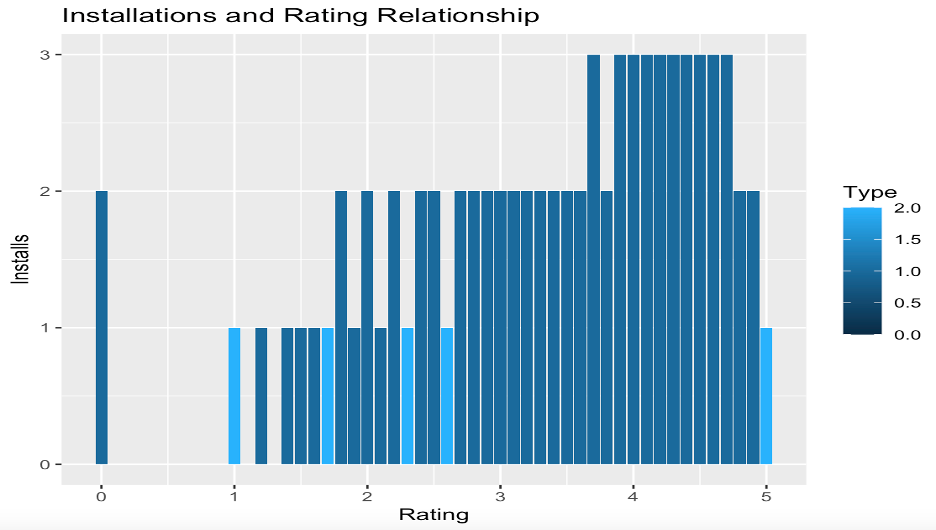
‘Content Rating’, which captured the targeted age group for certain applications, was converted from string to numerical values. ‘Teen’, ‘Everyone’ and ‘Everyone 10+’, was given the numerical value of 1, mature audience including 17+ was 2, and adults only at 18+ was 3. The genres were reduced to a single word from the original name, and ‘Type’ was assigned the value of 1 and 2 for free and paid. The ‘Last Updated’ column was converted to days and the ‘Installs’ was normalized ranking the first 33% as low success with the value 1, 33% to 66% as average success with the value 2, and the amount greater than 66% as the highest success with the value 3. The installs normalized was necessary to restrict the number of classes for smoother use in the machine learning setting.

After the dataset was cleaned, the number of observations were 10,357 with 13 variables. There were several models that were used to explore the dataset to answer the question of maximizing the number of downloads for Google Play Store. The developers and managers of Google Play Store conducted a simple and multiple linear regression on the dataset but was successful because there were too many classes and resulted in error. Simple and multiple linear regression could not have worked because the dependent variable was not a simple numerical value but had categorical classes that signified the number of successes within the downloaded ranges. With the linear regressions, the developers and managers attempted to conduct backward elimination to narrow down the subset of explanatory variables, and again was deemed unsuccessful. Random forest without boosting was another machine learning technique and ensemble learning method for classification that was unsatisfactory for what the company was trying to achieve to surpass Apple Store. Although random forest is said to outperform decision trees, the result rendered a 63% accuracy and didn’t give that much explanation as it is a “black box” and uses majority voting as the final class. Neural network took several hours of computing time and generated an even worse percentage of accuracy of 53%.

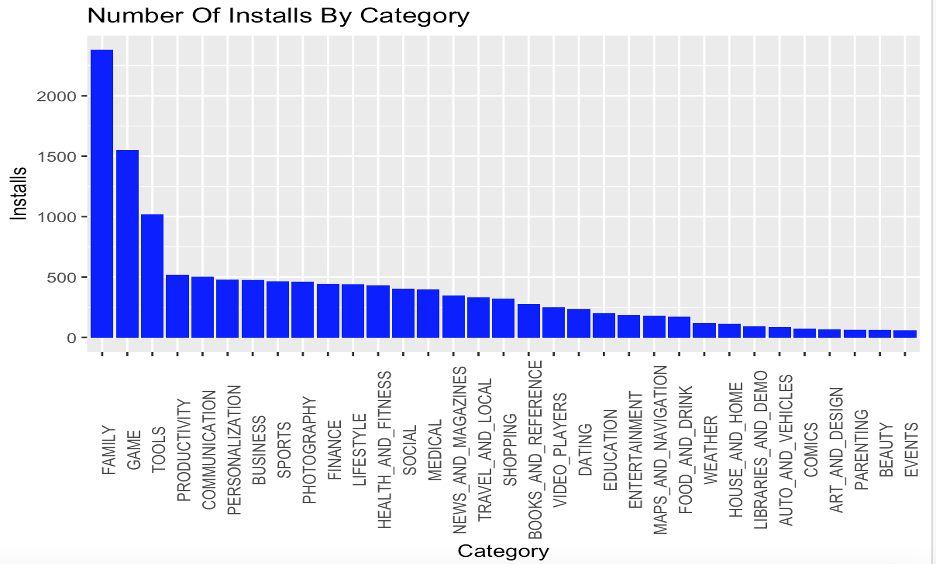
Finally, the machine learning technique that was used and resulted in the highest success rating was the classification tree. The completed classification tree with specified parameters generated a large list of classification rules but were reduced to a few variables that explained most of the accuracy within the dataset. With the success in the analysis technique, the developers, and managers of Google Play Store, further examined the correlation of the reduced variables along with a few others, the relationship between the number of installs and the type of application being downloaded by consumers according to the dataset, and the rating and installs.



The variables that have the most correlation are installs and rating at 18%.



Installations with the highest rate of success have a rating between 3.6 to 4.7.

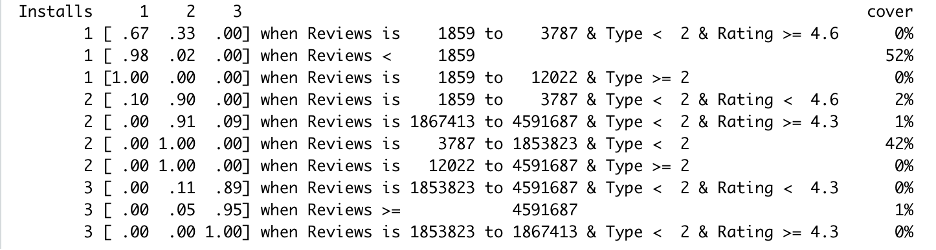


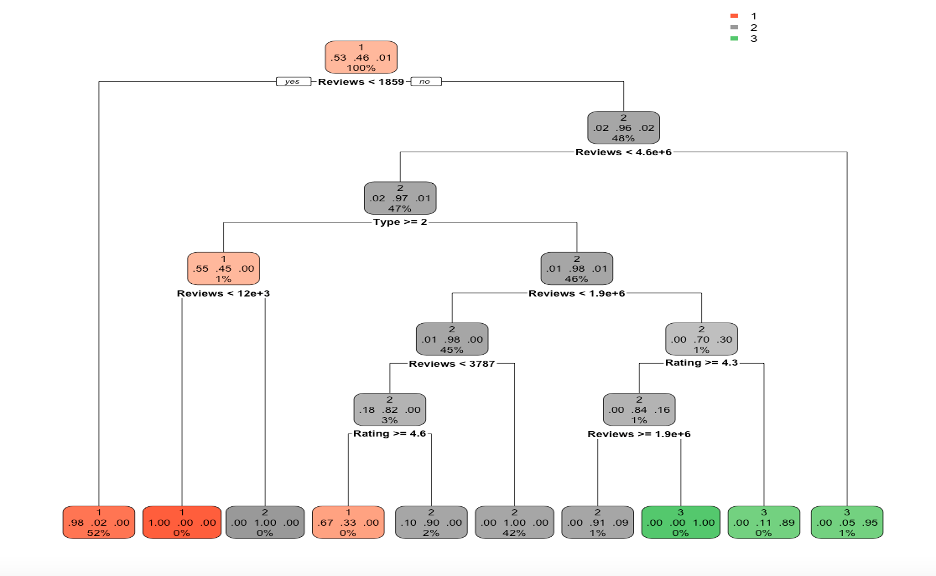
The top three most installed categories are family, game, and tools.

**Part III: Performance Results**

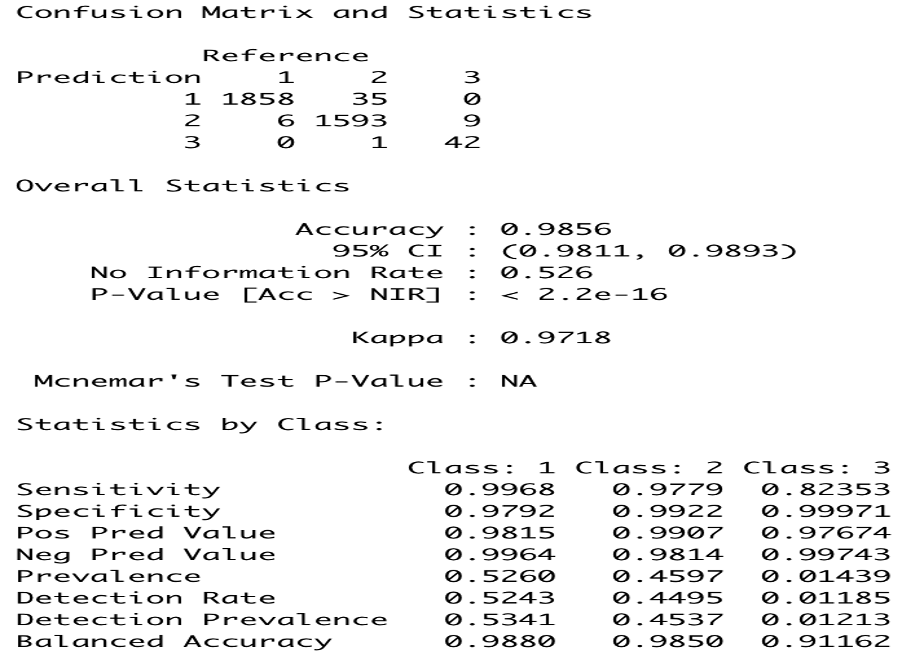
**Discussion and Interpretation of Results:**

The classification pruned tree models produced the best result for presenting actionable insights that can be drawn for developers to influence and expand the Android Market within the United States of America by predicting the number of downloads for the application for the Google Play Store. The accuracy for the unpruned classification tree was 98.34%, however we pruned the tree to avoid overfitting the data, and the pruned classification tree was 98.56%, which was among the highest from the entire analysis.





Reviews, rating, and type are the variables that determine the number of downloads that consumers will gravitate towards for the Google Play Store applications. The classification rules and the decision tree for the pruned tree shown above illustrates that the if reviews are less than 1856, 52% of the class is covered and will most likely generate low success rate of downloads (98% chance that there will be a low number of downloads and 2% will be an average number of downloads). Additionally, if reviews are between 3787 to 1853823 and free to download, 42% of the class is covered and will result in an average number of downloads with 100% of that 42% being a pure leaf with an average success of downloads for applications from the Google Play Store. Examining the decision tree more closely, reviews are greater than 1856, there is an increase in the chances that there will be an average number of downloads with class 2 carrying 96% versus a 2% chance for a low success of downloads and a 2% chance for the highest success of downloads by consumers. With the reviews being greater than 1856, but less than 4591687, and there is a payment requirement, downloads will most likely plummet to a low and average success of occurring. However, if the reviews for the downloads that require payment is greater than 12022, 100% of the downloads will fall in the average range of being completed by customers. Likewise, if reviews are greater than 1856 and 4591687, the 1% that is covered by the model will fall between the ranges of an average or highest number of downloads.



As stated, the confusion matrix deemed the accuracy to be 98.56% for the pruned classification tree. The model correctly classified the 1858 installs ranking in the first 33% as low success in downloads for the applications at the Google Play Store, and incorrectly classified the 6 installs as ranking within the 33% to 66% range as average success in downloads for the applications at the Google Play Store. For average success in downloads for the applications at the Google Play Store, the model correctly predicted 1593, and incorrectly predicted 35 installs to be ranked in the first 33% as low success in downloads, and 1 install ranked in the highest success in downloads above 66%. The value of 3, which represents the highest success in downloads for the applications at the Google Play Store was correctly predicted 42 installations by the classification tree, and incorrectly predicted 9 installs to be ranked within the 33% to 66% range as average success in downloads for the applications at the Google Play Store.

**Conclusion:**

Following the completion of the analysis it has been determined by the developers and managers of Google Play Store that reviews, ratings, and type are the variables that govern the number of downloads by consumers within the market. Concentrating on what consumers are looking for regarding application downloads requires reading through reviews to see the likes and dislikes that are voiced about specific applications, checking the ratings that are given with each review, and scrutinizing the applications that customers are paying for just so that they can get the experience. Although many customers tend to download mostly free applications, there is still 7% of the population that pays for applications. The model predicted that the majority of those downloads will yield an average success, which is where Google Play Store can capitalize and exploit the market to dominate in the United States.

The members of the development and management teams at Google Play Store found the project to be intriguing and challenging. We live with the era where applications are used in all aspect of our lives, and many choose to use Android products over Apple and vice versa. The dataset took some time to clean so that we could determine the best model regarding accuracy and predictability. We were challenged on remembering how to interpret the result of the complex models and whether it was better to relinquish some accuracy for more precise predictions, however, we when it was all said and done, we were able to incorporate the best accuracy and prediction for the number of downloads for the applications at the Google Play Store developers to influence and expand the Android Market within the United States of America, and provide actionable recommendations for the company.

**Reference Page**

Gupta, Lavanya. "Google Play Store Apps." *Kaggle*, 28 Nov. 2018,

www.kaggle.com/datasets/lava18/google-play-store-a

apps/code?datasetId=49864&searchQuery=EDA&select=googleplaystore.csv. Accessed

24 Nov. 2022.

Wallen, Jack. "Why is Android More Popular Globally, While iOS Rules the US?"

*TechRepublic*, 12 May 2021, www.techrepublic.com/article/why-is-android-more-popular-globally-while-ios-rules-the-us/. Accessed 24 Nov. 2022.