Google Applications Report

Wazid Ansari

B. TECH, Delhi Technological University

ABSTRACT

The internet is a true gold mine of data. E-commerce and review are brimming with lot of untapped data with a prominent to convert into meaningful insights that can help with robust decision making. To draw meaningful insights from a data we need to perform data analysis. Data analytics is expected to radically change the way we live and do the business in the future.

Application distribution platforms like Google Play Store are routinely flooded with thousands of new apps, requiring a large number of app developers to work independently or in teams to make them successful. Because of the fierce competition that exists throughout the world, developers must ensure that they are moving in the proper path. So, in this project our goal is to draw some insights from the dataset and to predict the apps ratings. Prediction of Apps ratings will be done by using some machine learning algorithms.

INTRODUCTION

In 2020, there were 143 billion installations on Google Play, according to the worldwide mobile app industry. Whenever it comes to mobile expenditures, Google Play has more installations than the App Store.

This year, the combined sales of both stores are expected to hit 155 billion dollars. Global installations are expected to hit 230 billion by 2025, with a compound yearly growth of 10%.

From 2019 - 2026, the global mobile software market is expected to increase at a CAGR of 18.4 percent.

There are different kinds of software applications made to work on various cellphones, tablets, and computer tablets. Mobile applications often provide customers with services that are just like those available on PCs. The main purpose of the mobile application is to link customers to internet providers by enabling them to access the internet on devices.

The use of variable gadgets and the development of the client base for the e-commerce business are two main reasons driving the industry's growth. Furthermore, the market's growth is fueled by continued growth in software applications, increased attention on apps expressly utilized for health and wellness, and increased downloads and in-application buys for gaming apps.

You can easily use applications without having to give the login details every time you download an app thanks to Google Play Services. It handles and tracks Google Pay transactions to verify you get the right version of any thing you've purchased (books, movies, etc.).

Machine learning systems process large datasets quickly and provide useful insights into information, allowing for excellent healthcare services. Because the company was slow to adopt this technology, it is now swiftly catching up and providing efficient preventative and preventive healthcare services.

Healthcare organizations are increasingly leveraging computing capacity to analyze large datasets and find designs that provide significant experience from current client data to make more precise decisions and provide good quality outcomes.

Big data and deep learning algorithms swiftly process massive datasets and offer useful information for the production of high-quality pharmaceuticals.

Although the healthcare industry's use of DL and data science is still low, it is fast rising to give viable medical solutions.

Medical organizations' data has grown significantly, necessitating the computing capabilities to analyze large datasets and discover patterns from current patient data to make accurate medical advancements. Machine learning algorithms are statistical model mapping approaches that are used to discover or understand underlying patterns in data.

LITERATURE REVIEW

According to Rimisha Maredia research paper “Analysis of Google Play Store Data set and predict the popularity of an app on Google Play Store” (June 2020) Machine learning algorithms like k-nearest neighbours, Gaussian Naïve Bayes model, Decision tree model, Logistic regression. In this study Decision tree model is best and can be used by future developers and the Google Play Store team to assess the Google Play Store market and determine which app categories should be created in order to maintain the Google Play Store popular in the future.

Whereas, according to Brahma Naidu and S Shashank research paper “Google Play Store Apps- Data Analysis and Ratings Prediction” (December 2020), the exploratory data analysis of their study and this study were same. But there were differences in the best fit model for apps ratings prediction. The k-nearest neighbour model was best for their study with Accuracy score of 92%. Also, slight changes in accuracy scores were seen in their models and in ours. The SVM model accuracy score for our study is 78% whereas their accuracy score for the same model is 76%.

PROBLEM STATEMENT

One of the fastest-growing categories of the downloaded software application market is mobile applications. Apps for mobile phones are all over the place. In today's world, we can observe how mobile apps are becoming increasingly vital in people's lives. It has been observed that the growth of the mobile application market has a significant impact on advanced innovation. For this study we chose Google Play Store over all other markets because of its growing popularity and its rapid expansion. The fact that around 81 percent of the applications are free is one of the key reasons for their success. In April 2013, the market had grown to over 845900 Apps and 226,500 distinct merchants. Because most Play Store apps are free, the revenue model for how in-app purchases, advertisements, and

memberships contribute to an app's success is obscure and unavailable. As a result of this burgeoning market, more than 500 million people around the world have downloaded over 40 billion apps. The impact of market interactions on future technology is mostly determined by developers and users.

However, both developers and users are affected by a lack of awareness of the inner workings and dynamics of popular app markets. And with so much competition coming from all over the world, it's crucial for a designer to know that he/ she is on the correct track. So, we will try to identify which apps' categories are preferred by people so that organizations who want to expand their business into new horizon with an app can look at this study and can make decisions accordingly. Also, in this study, we'll look at the Google Play Store's dynamics and how we may use variables from the data set to make predictions.

OBJECTIVES

The objectives of our study is-

- Identify which category apps are preferred by people based on Ratings, Pricing and Reviews
- Predict future Apps ratings by creating some Machine Learning models
- Identify which model is best to predict the apps ratings

RESEARCH METHODOLOGY

Steps that will be used in our study are

1. Data Collection
2. Data Cleaning
3. Data Manipulation
4. Data Visualization
5. Data Modelling

1) DATA COLLECTION

For this study raw dataset is collected from Kaggle website. This data set comprises 13 different features that can be used to forecast whether an app will be successful or not based on its individual characteristics. This information was scrapped from the Google Play Store.

The Google Play Store is thought to be the most popular app store. The Google Play Store is flooded with tens of thousands of new applications every day, with an ever-increasing number of designers working alone or in groups to make them successful, facing great competition from all over the world. Because most Play Store apps are free, the revenue model for how in-app purchases, advertisements, and memberships contribute to an app's success is obscure and unavailable.

DATA DESCRIPTION

Attribute	Description

App	Application
Category	Category of the Application
Rating	Overall user rating of the app
Reviews	Number of user reviews for the app
Size	Size of the App
Installs	Number of user downloads/Installs
Type	Paid or Free
Price	Price of the app
Content Rating	Age group of user - Children/Adult
Genres	App's Genre

2) DATA CLEANING

Data cleaning is the process of identifying the parts of data that are wrong, incomplete, inaccurate, irrelevant, or missing, and then changing, replacing, or deleting them as needed. Data cleaning is regarded as a fundamental component of data science.

3) DATA MANIPULATION

The process of changing data to make it more organised and simpler to read is known as data manipulation. For corporate operations and optimization, data manipulation is critical. You must be able to deal with data in the way you need it to correctly use it and translate it into usable insights, such as analysing financial data, consumer behaviour, and performing trend analysis

4) DATA VISUALIZATION

The process of turning data into a chart, graph, or other visual format that aids analysis and understanding is known as data visualisation.

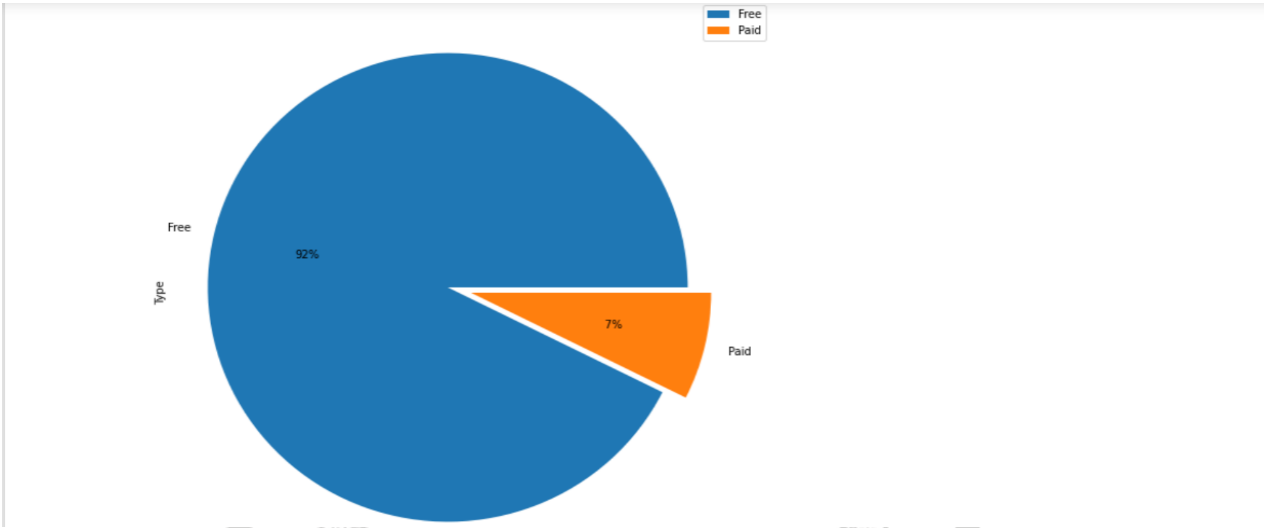
5) DATA MODELLING

The practise of examining data objects and their relationships with other things is known as data modelling. It's utilised to investigate the data requirements for various business activities. The data models are constructed to store the data in a database.

DATA ANALYSIS

The Data Analysis was performed on Python Jupyter Notebook and the libraries used are – NumPy, Pandas, Matplotlib, Scikit-learn and Seaborn.

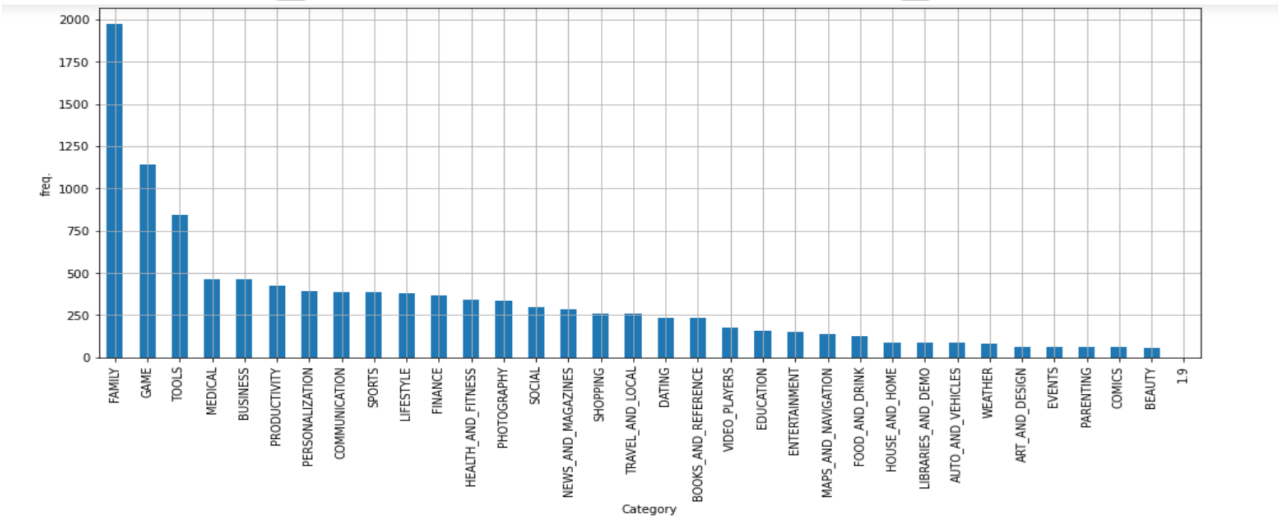
I. Free vs Paid Apps



We can observe that 92% of apps on Google Play Store are free, while 7% are purchased, implying that the majority of apps on Google Play Store are free.

II. Value_Counts()

value_counts() function returns object containing counts of unique values.



From above graph we can say Family App Category has highest frequency which means this category is preferred most by people.

Data Cleaning

First to check how many columns have missing values or non-null values I applied **info()** function of python. The info function prints the concise summary of data frame and prints the non-null values.

```
In [12]: google_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10841 non-null  object
 1   Category        10841 non-null  object
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object
 4   Size            10841 non-null  object
 5   Installs        10841 non-null  object
 6   Type            10840 non-null  object
 7   Price           10841 non-null  object
 8   Content Rating  10840 non-null  object
 9   Genres          10841 non-null  object
 10  Last Updated    10841 non-null  object
 11  Current Ver     10833 non-null  object
 12  Android Ver     10838 non-null  object
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

The output showed that there were three columns that contains non-null values, these columns were Ratings, Current Version and Android Version.

To check the check the count of the missing values or non-null values I applied **isnull().sum()** function of python.

```
In [14]: google_data.isnull().sum()

Out[14]: App                0
         Category           0
         Rating          1474
         Reviews            0
         Size               0
         Installs           0
         Type               1
         Price              0
         Content Rating     1
         Genres             0
         Last Updated       0
         Current Ver        8
         Android Ver        3
         dtype: int64
```

Data Manipulation

In this step the missing values were filled with mean, median and mode. For categorical data Mode value was used and for numerical data Median value was used. To do so one of the Python's library Panda is used. Panda provides **fillna()** function for returning values with a specific value.

```
In [26]: google_data['Type'].fillna(str(google_data['Type'].mode().values[0]),inplace=True)
         google_data['Current Ver'].fillna(str(google_data['Current Ver'].mode().values[0]),inplace=True)
         google_data['Android Ver'].fillna(str(google_data['Android Ver'].mode().values[0]),inplace=True)

In [27]: google_data.isnull().sum()

Out[27]: App             0
         Category        0
         Rating          0
         Reviews         0
         Size            0
         Installs        0
         Type            0
         Price           0
         Content Rating  0
         Genres          0
         Last Updated    0
         Current Ver     0
         Android Ver     0
         dtype: int64
```

```
In [31]: google_data.describe()      #Summary after Cleaning
```

Out[31]:

	Rating	Reviews	Installs	Price
count	10840.000000	1.084000e+04	1.084000e+04	10840.000000
mean	4.206476	4.441529e+05	1.546434e+07	1.027368
std	0.480342	2.927761e+06	8.502936e+07	15.949703
min	1.000000	0.000000e+00	0.000000e+00	0.000000
25%	4.100000	3.800000e+01	1.000000e+03	0.000000
50%	4.300000	2.094000e+03	1.000000e+05	0.000000
75%	4.500000	5.477550e+04	5.000000e+06	0.000000
max	5.000000	7.815831e+07	1.000000e+09	400.000000

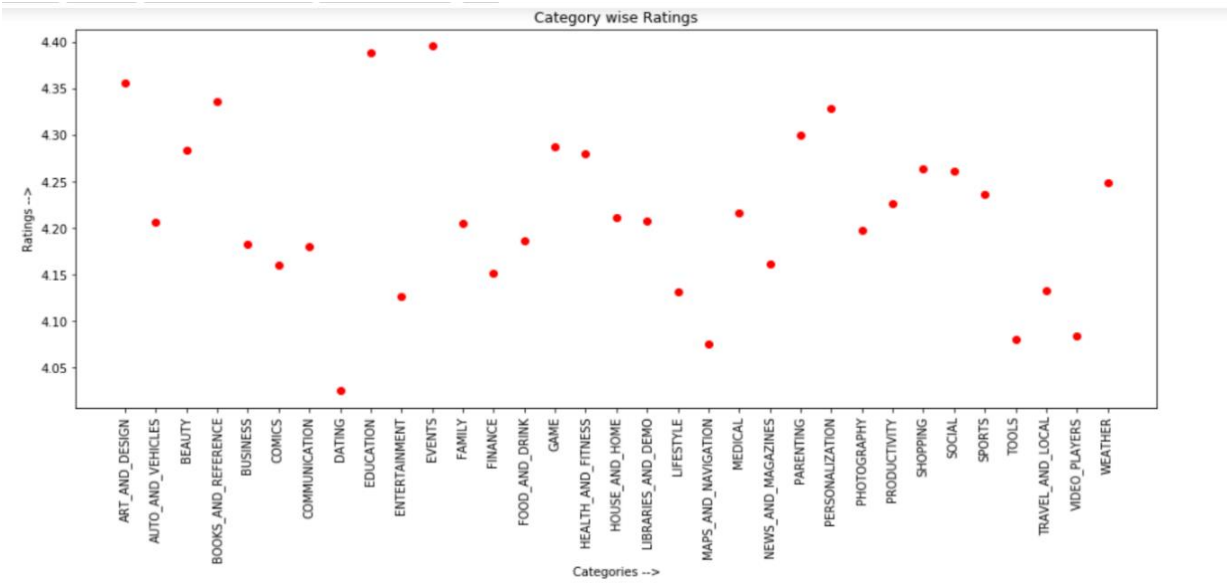
Data Visualization

To find which Category of app is preferred the most by people groupby() function is used. Parameters considered for grouping were Ratings, Price & Reviews.

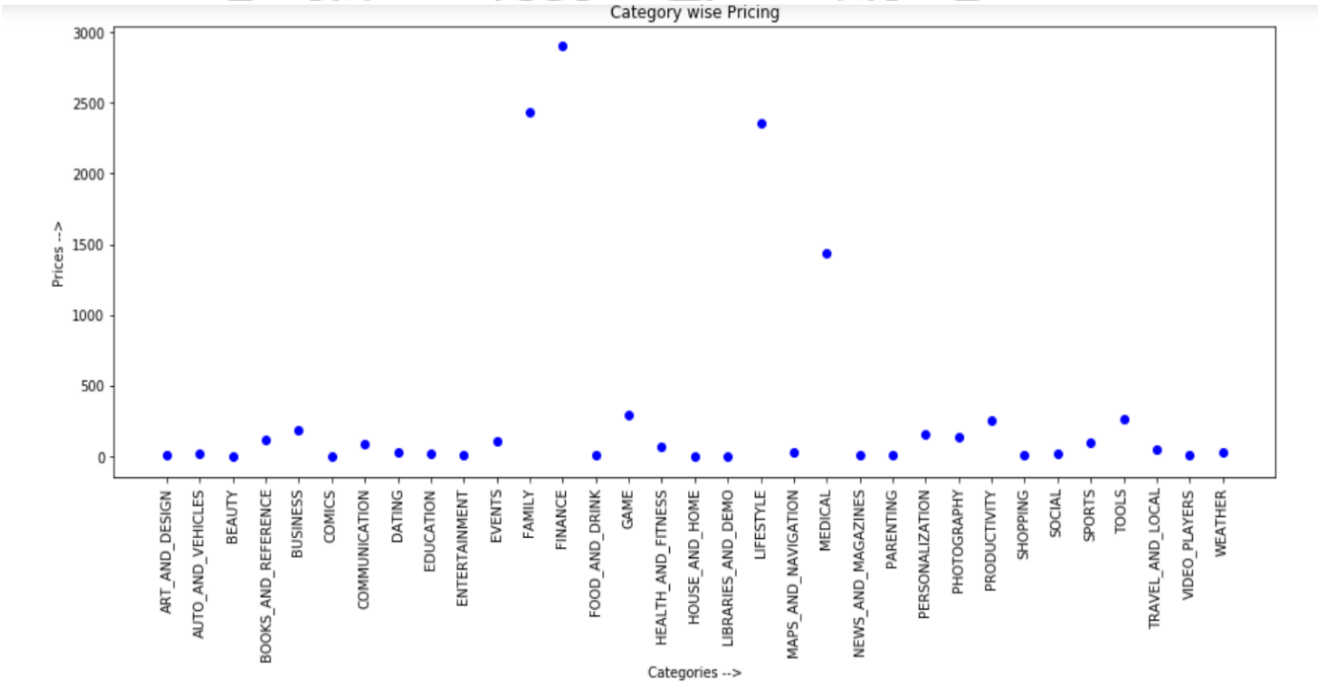
```
In [30]: #Data visualization
grp=google_data.groupby('Category')
x = grp['Rating'].agg(np.mean)
y = grp['Price'].agg(np.sum)
z = grp['Reviews'].agg(np.mean)
print(x)
print(y)
print(z)
```

```
Category
ART_AND_DESIGN          4.355385
AUTO_AND_VEHICLES       4.205882
BEAUTY                  4.283019
BOOKS_AND_REFERENCE     4.335498
BUSINESS                4.182391
COMICS                  4.160000
COMMUNICATION           4.180103
DATING                  4.025641
EDUCATION               4.388462
ENTERTAINMENT           4.126174
EVENTS                  4.395313
FAMILY                  4.204564
FINANCE                 4.151639
FOOD_AND_DRINK          4.185827
GAME                    4.286888
HEALTH_AND_FITNESS      4.280059
HOUSE_AND_HOME          4.211364
```

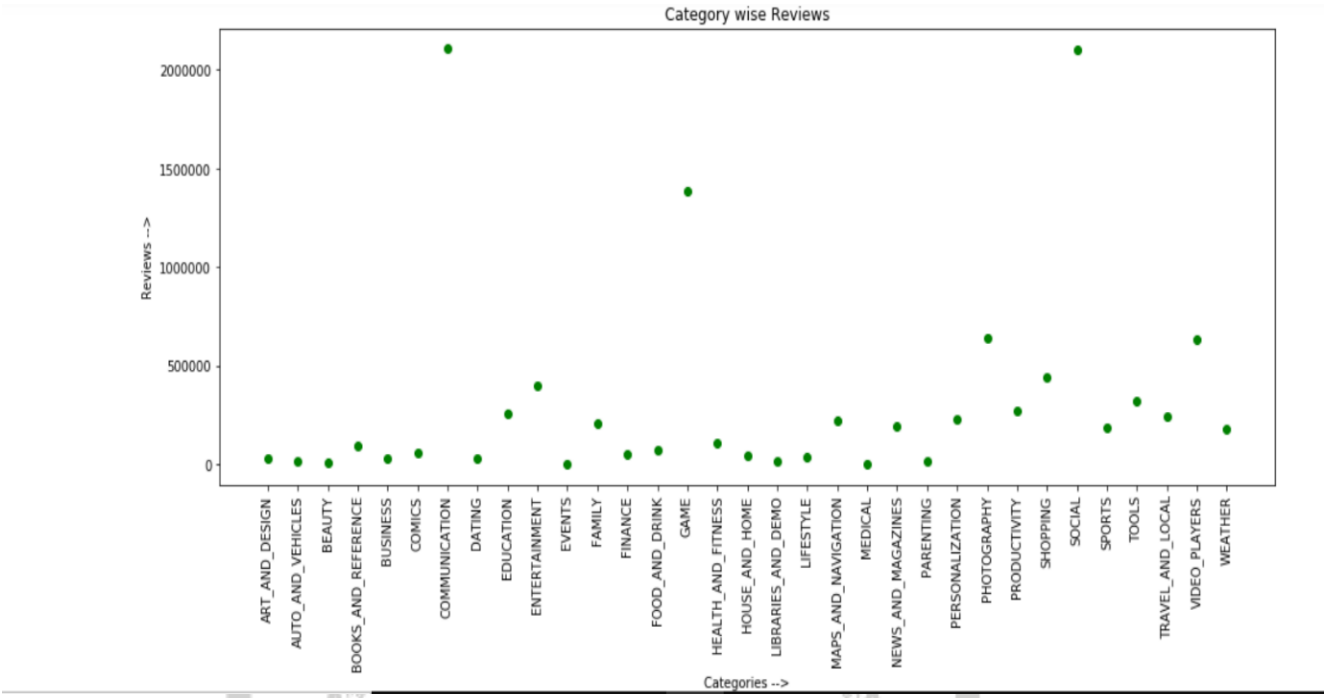
CATEGORY WISE RATINGS



CATEGORY WISE PRICING



CATEGORY WISE REVIEWS



DATA MODELLING

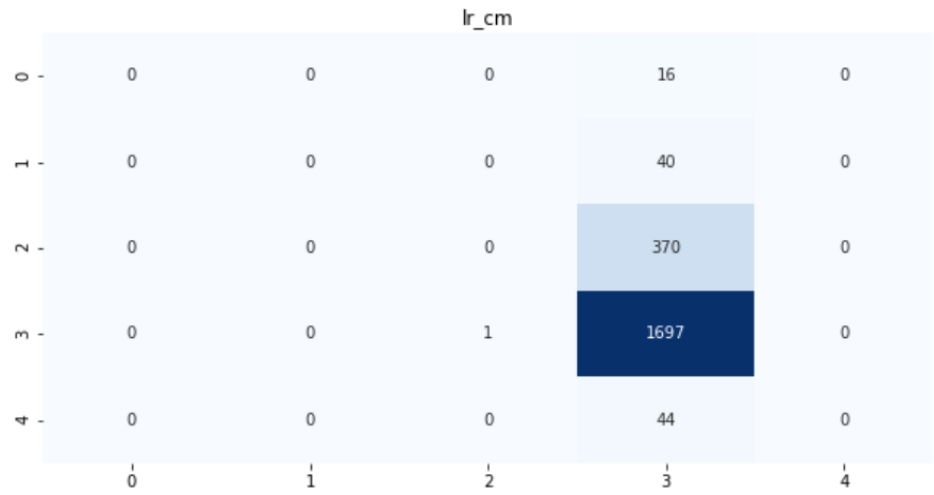
To make appropriate prediction model we first need to perform train test split function. The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results.

For our case we will split 80% of the data for training and rest 20% for testing.

After this step Machine learning models were made to predict apps ratings. To identify which model is best to predict ratings we will look at the Accuracy score of each model to compare. The model with the best accuracy score will be preferred.

LOGISTIC REGRESSION

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable.

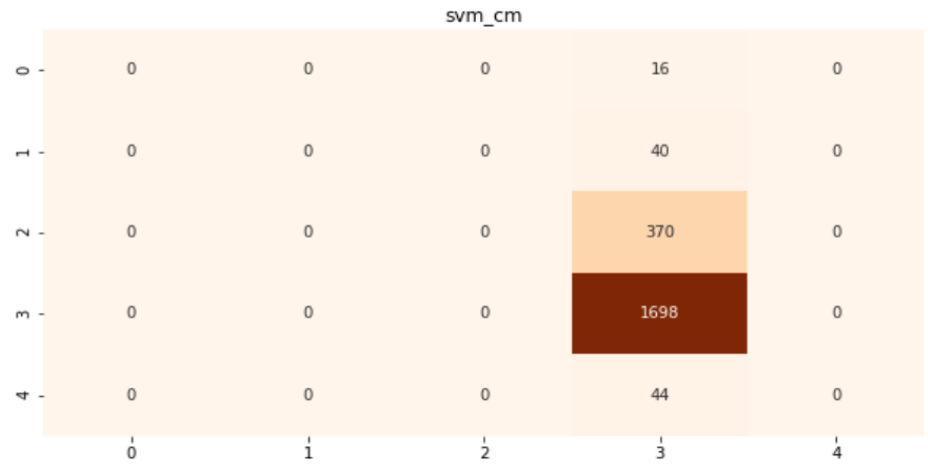



```
Accuracy Score :  0.7827490774907749
```

SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

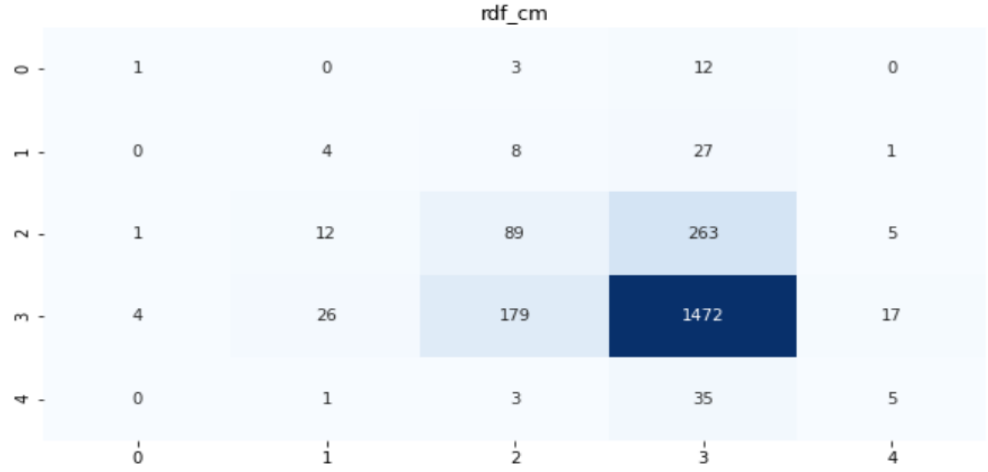
```
SVM_regressor_accuracy: 0.783210332103321
```



RANDOM FOREST

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

```
RandomForest_accuracy: 0.7246309963099631
```

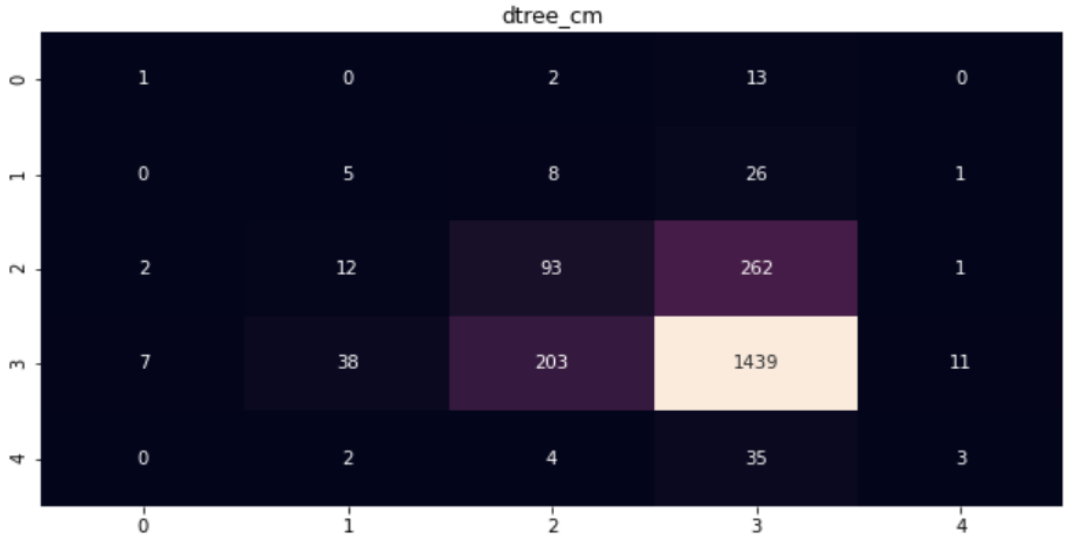


DECISION TREE CLASSIFIER

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

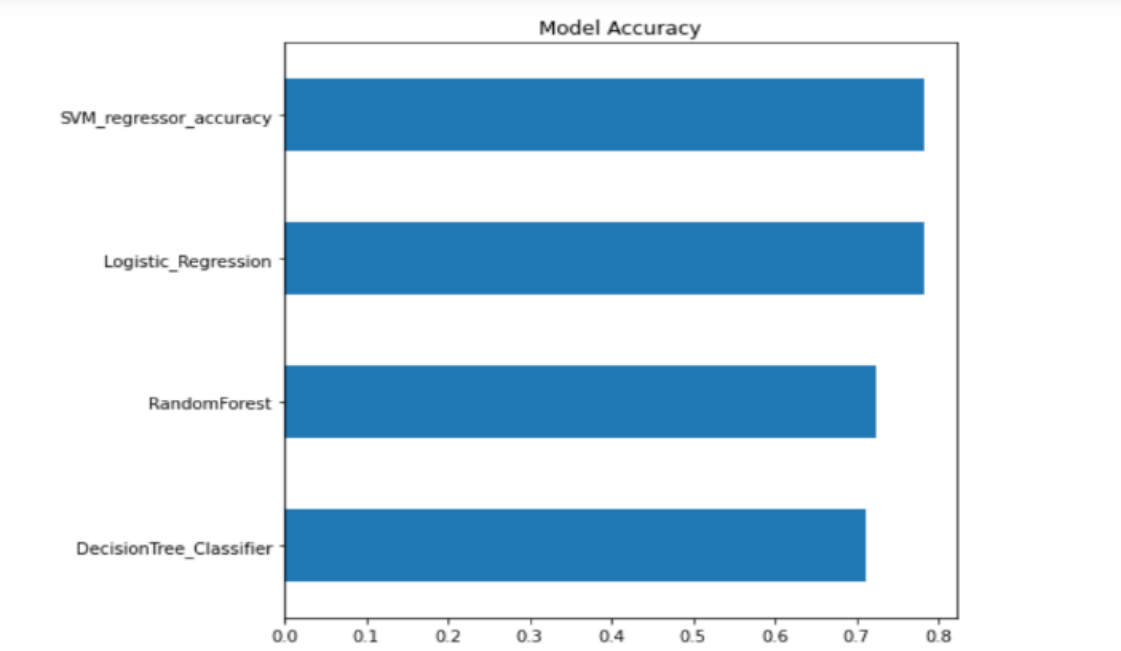
The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

```
DecisionTree_Classifier_accuracy: 0.7107933579335793
```



COMPARISON MODEL

We compared all the above models based on the Accuracy Score.



By looking at the above graph we can see that SVM Classifier and Logistic Regression have the highest Accuracy Score. So, we can say that both these models can be used to predict the apps ratings.

LIMITATIONS

- The dataset used for analysis is not the latest
- Another limitation can be Paid Reviews. Companies now pay people to give good reviews and to promote their apps. So, we don't know that the reviews are completely honest or not and it can affect the decision making
- Although the mobile application industry is in demand and people are now becoming tech savvy, but there are many competitors already in the market. For every category there are thousands of apps available in the market with almost same features so this will be a challenge for the company to produce something different with limited resources.

RECOMMENDATIONS

- Based on the results obtained from the analysis the Communication and Social apps have the highest reviews and Events and Education have highest ratings. So, if any organization want to expand their horizons and want to launch any new app so they can go for these categories
- Besides SVM and Logistic Regression Models other Machine Learning Algorithms can also be used to predict apps ratings

CONCLUSION

This data set includes a substantial quantity of information that can be utilised for a variety of reasons. It can be utilised to boost the worth of a business or the Google Play Store in general. We have drawn some meaningful insights from this dataset by performing Exploratory Data Analysis and by applying various Machine learning algorithms.

Through Exploratory Data Analysis we observed that people were more drawn to Free Apps. The Events and Education category of apps have highest ratings. Whereas Family and Finance apps have high prices. Communication and social apps have the highest reviews, and it is because social apps like WhatsApp, Facebook, Instagram, Snapchat etc have users in great numbers. These users must have given their positive and negative reviews on the Play store.

When it comes to predictions, we can say that we can make predictions using this dataset. We applied different machine learning algorithms to predict apps ratings and found that Support Vector Machine and Logistic Regression Algorithms were best fit for this study. The Accuracy Score of these two models were highest. The reason we took Accuracy Score is because it talks about the ratio of number of correct predictions to the total number of input samples.

So based on our problem statements and objectives we can conclude that organizations who want to expand their business horizons can go for Social, Communication and Family category of app because these apps were more preferred by people. And for predictions Support Vector Machine and Logistic Regression algorithms should be used.

References

[1] Brahma Naidu, S Shashank (Dec 2020), International Research Journal of Engineering and Technology (IRJET), [online] https://www.irjet.net/archives/V7/i12/IRJET-V7I1248.pdf [Accessed 24th December 2021]

[2] Lavanya (2019), “Google Play Store Apps” [online] https://www.kaggle.com/lava18/google-play-store-apps [Accessed 26th December 2021]

[3] Michelle Atkinson (January 2015), “An analysis of apps in Google Play store”[online] https://www.pewresearch.org/internet/2015/11/10/an-analysis-of-apps-in-the-google-play-store/ [Accessed 24th December 2021]

[4] Mihhail Matskin, M.T. Rahman, Shahab Mokarizadeh (January 2013), “Mining and Analysis of Apps in Google Play,” [online] https://www.researchgate.net/publication/290102532_Mining_and_analysis_of_apps_in_google_play [Accessed on 24th December,2021]

[5] Muhammad Umer, Imran Ashraf, Arif Mehmood, Saleem Ullah, Gyu Sang Choi (July 2020), “Predicting numeric ratings for Google apps using text features and ensemble learning”, [online] https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2019-0443 [Accessed 24th December]

[6] Rimisha Maredia (June 2020) “Analysis of Google Play Store Data set and predict the popularity of an app on Google Play Store”, [online] https://www.researchgate.net/publication/343769728_Analysis_of_Google_Play_Store_Data_set_and_predict_the_popularity_of_an_app_on_Google_Play_Store [Accessed 24th December 2021]

[7] Tutorials Point, “Data Analysis Process” [online] https://www.tutorialspoint.com/excel_data_analysis/data_analysis_process.htm [Accessed 24th December 2021]