Google Play Store Apps- Data Analysis and Ratings Prediction

S Shashank[1], Brahma Naidu[2]

[1]*Student ICFAI Tech, Hyderabad*
[2]*Professor, Dept. of Computer Science, ICFAI Tech, Hyderabad, Telangana, India.*

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.*

Key Words: Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

1. INTRODUCTION

Machine learning approaches are essential for us to take care of numerous issues. In this paper, we present machine learning models and structures in detail. Machine learning has numerous applications in numerous perspectives and has incredible advancement potential.

In future, it is predictable that machine learning could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there is much information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize this points of interest to achieve more assignments.

1.1 Analysis and Prediction

In today's scenario we can see that mobile apps playing an important role in any individual's life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that in spite of the fact that it creates more than two fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android1market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding the a huge number of printed remarks.

1.2 Google Play store Dataset

The dataset consists of Google play store application and is taken from Kaggle, which is the world's largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

1.3 Data Mining

Data mining is that the process of rummaging through a knowledge set and finding correlations, anomalies and or patterns which will be of usefulness. In other words, it's having an outsized dataset filled with scattered information and trying to form sense of it by finding meaningfulness.

1.4 Python

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That's the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That's one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open source library is named panda. As we've seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

1.5 Machine Learning

Machine learning is an application of AI (AI) that gives systems the power to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the event of computer programs which will access data and use it learn for themselves

1.5.1 Supervised Learning

It is defined as a learning in which we train a machine as per our dataset or input. From that point forward, the machine is furnished with another arrangement of examples (data) so supervised learning analyses the provided data (set of preparing models) and creates a right result from given input.

1.5.2 Unsupervised Learning

In the Unsupervised learning we do not train our machine according to the present data or input. It means there is not any supervisor as a teacher in this learning. In this we allow algorithm to work on their own without any training or guidance. Here the main working of the machine is that it works on some definite patterns, similarities in the given dataset without any training or proper guidance. Therefore machine is restricted to find out the structure which is hidden in the given dataset.

1.5.3 Semi-supervised Learning

This type of learning lies between the above two learning methods.

1.6 Neural Networks

We can divide neural network into different forms such as artificial neural network, deep neural network, recurrent neural network, convolutional deep neural networks. Each form has its own importance and its own features. In neural network we have input layer, no of hidden layers, and output layer.

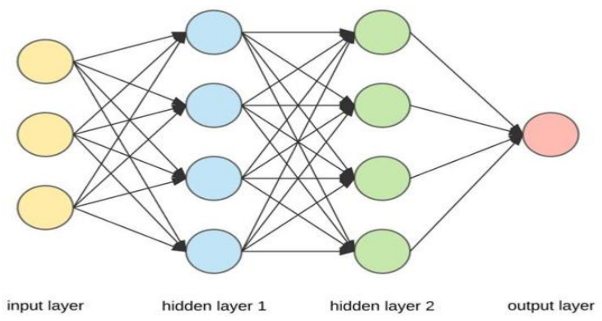


Fig -1: Neural Network

1.6.1 Deep Neural Network

A deep neural network is defined as a neural network which contains certain level of complexity, like a neural network which contains more than two layers. In the deep neural network we use some mathematical model to solve any model in a proper way using all the complexities.

A neural system, when all is said in done, is an innovation worked to reproduce the action of the human brain – explicitly, design acknowledgment and the section of contribution through different layers of deep neural associations.

In DNN, data flows forward it means from input layer to output layer without having any loopholes. At first, the DNN makes a guide of virtual neurons and allots irregular numerical qualities, or "loads", to create link between them. The loads and data sources are increased and return a yield somewhere in the range of 0 and 1. On the off chance that the system didn't precisely perceive a specific example, a calculation would alter the loads. That way the calculation can make certain parameters progressively

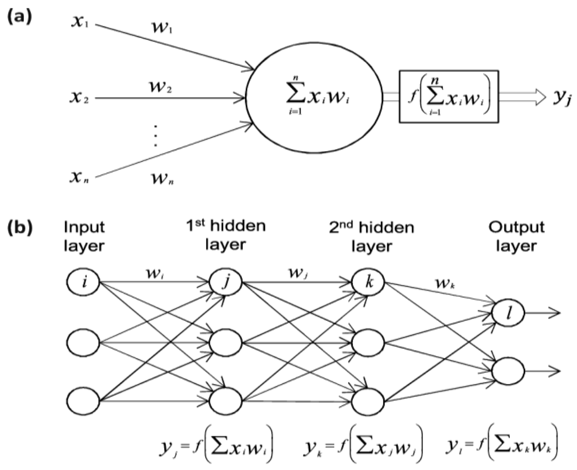


Fig -2: Neural Network Diagram

2. LITERATURE SURVEY

There has been a constant growth in the public and private information stored within the internet. This includes textual data expressing people's opinions on review sites, forums, blogs, and other social media platforms. Review-based prediction systems allow this unstructured information to be automatically transformed into structured data reflecting public opinion. These structured data can be used subsequently as a measure of users' sentiments about specific applications, products, services, and brands. They can hence provide important information for product and services refinement. This kind of sentiment analysis was conducted in the following studies.

- Kumari and other researchers [8,9,10,5] used the Naïve Bayes (NB) classifier to classify opinions as positive, negative, or neutral.

- Wang and others [11] argued that a rating is not entirely determined by a review content. For example, a user may well intend to give a positive review by employing positive words, and yet issue a comparatively lower rating.

- Dave and others [12] proposed a method for extracting the polarity in user reviews of products, expressed as poor, mixed, or good. The classifier used was Naïve Bayes (NB).

- According to Pang et al [13], although machine learning approaches perform far better for traditional topic-based categorization, they're less successful for sentiment analysis.

- Information-extraction technologies have also been explored to identify and organize opinions contained in text. For example, some authors [14] proposed a scheme for annotating a low-level representation of opinions within a text. Additionally, they described an opinion-oriented "scenario template" that summarizes the opinions expressed in a document. This approach is helpful for tasks that involve posing question from multiple perspectives.

- Other authors [15] suggested adopting a statistical analysis based on a spin model, to extract the semantic orientations of words. Mean field approximations were used to compute the approximate probability in the spin model. Semantic orientations are then evaluated as desirable or undesirable. A smaller number of seed words for the proposed model produce highly accurate semantic orientations based on the English lexicon.

- Various sentiment analysis methods have been performed to summarize the ensembles of comments and reviews [16]. These methods use mathematical and statistical methods (especially involving Gaussian distributions) to overcome the problems encountered in sentiment analysis. Although these authors proposed a model, it was not implemented.

- A recent study [17] investigated the application of a machine learning algorithm to a dataset covering, for example, the app category, the numbers of reviews and downloads, the size, type, and Android version of an app, and the content rating, to predict a Google app ranking. Decision trees, linear regression, logistic regression, support-vector machine, NB classifiers, k-means clustering, k-nearest neighbors, and artificial neural networks were studied for that purpose.

- App ratings have been predicted based on the features provided for app [18,19]. Experiments were performed on the BlackBerry World and Samsung Android stores to collect the raw features provided for the apps, including their price, rank of downloads, ratings, and textual descriptions. The features were then encoded into a numerical vector to be used in case-based reasoning and to predict the app rating.

- In contrast to the above-cited studies, other authors [20] investigated the nature of sentiments expressed in Google app reviews. Their study measured opinions and sentiments represented in user reviews through a variety 4 | UMER et al. of emojis expressing, for example, negativity, positivity, anger, or excitement. It evaluated whether those sentiments are informative for the purpose of app development and refinement.

However, the above studies are unsatisfactory in various respects and are unsuitable for predicting numeric ratings of Google apps. First, text-mining techniques are ineffective when applied to app reviews, as it has Unicode supported language with a limited number of words. Second, those studies are based either on rating predictions made using inherent app features or on external features (eg, price, bug report, etc.). None of those studies investigated the possible discrepancies between users' numeric ratings and reviews. To our knowledge, this study is the first to investigate such discrepancies and to base numeric-rating predictions for Google apps.

3. EXPLORATORY DATA ANALYSIS

3.1 Free vs Paid

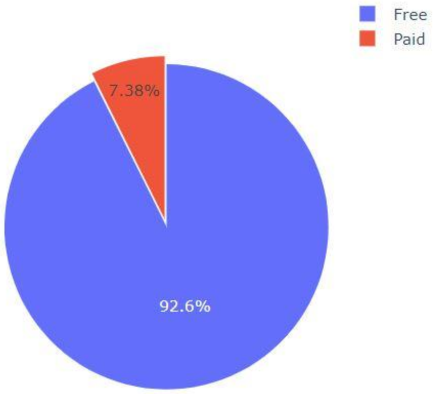


Fig -3: Free vs Paid

Here we can see that 92.6% apps are free and 7.38% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

3.2 Updated Apps

In the below plot, we plotted the apps updated or added over the years comparing Free vs. Paid, by observing this plot we can conclude that before 2011 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, By comparing the apps updated or added in the year 2011 and 2018 free apps are increases from 80% to 96% and paid apps are goes from 20% to 4%. **So we can conclude that most of the people are after free apps**

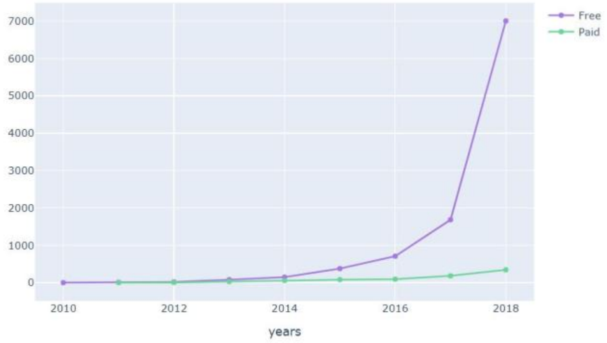


Fig -4: Updated Apps

3.3 Updated Free Apps

In this data almost 50% apps are added or updated on the month of July, 25% of apps are updated or added on the month of August and rest of 25% remaining months.

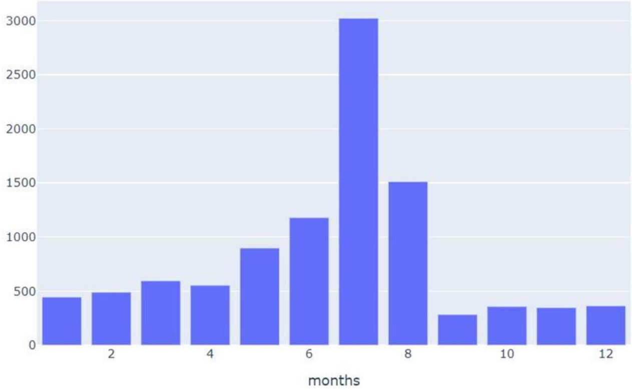


Fig -5: Updated Free Apps

3.4 Updated Paid Apps

Same as free apps most of the paid apps too updates in the month of July.

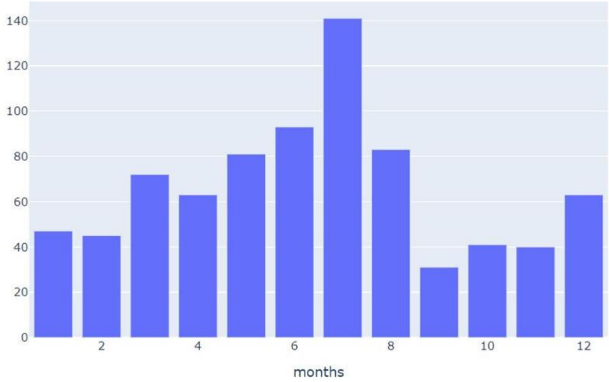


Fig -6: Updated Paid Apps

3.5 Ratings

Most Number of ratings which got on Google Play Store is given for free apps.

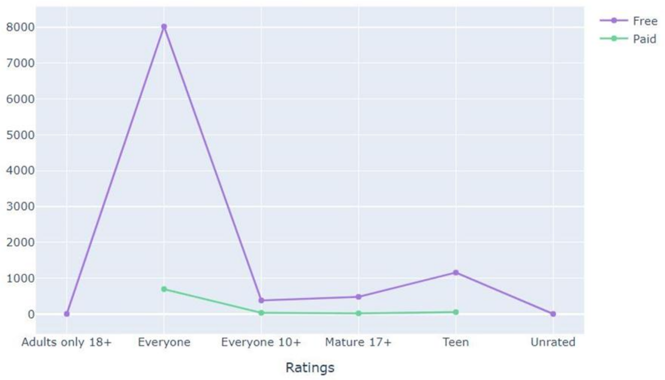


Fig -7: Ratings

3.6 Free App content Rating

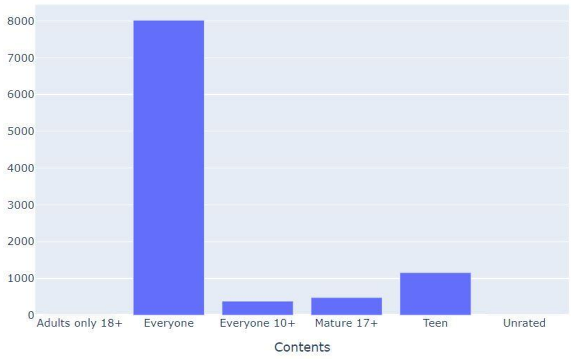


Fig -8: Free App Rating

3.7 Paid App Content Rating

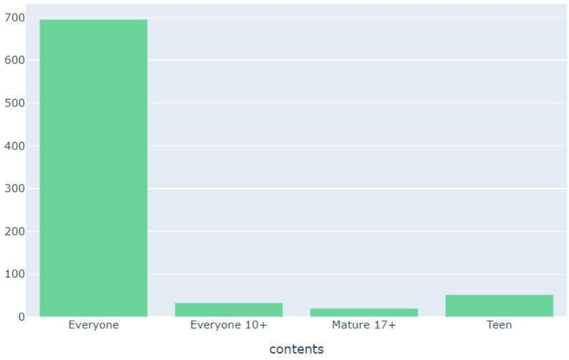


Fig -9: Paid App Ratings

3.8 Ratings of Free vs Paid Apps

Free apps are the most rated apps on the Google Play Store compared to Paid Apps

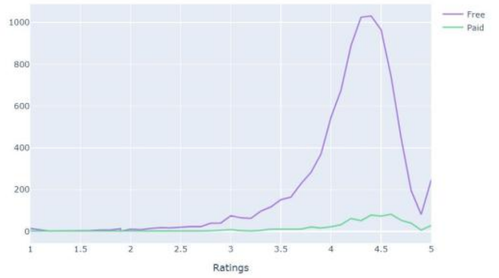


Fig -10: Free vs Paid App Ratings

3.9 Free Apps Ratings

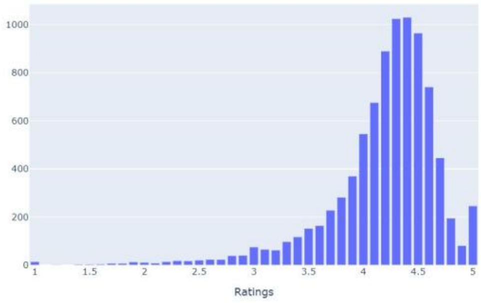


Fig -11: Free App Ratings

3.10 Paid App Ratings

Most of the paid apps on the app store are rated 4.2 to 4.8

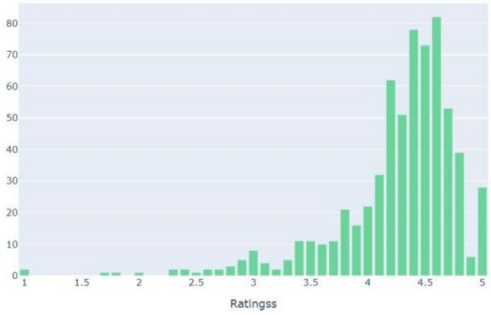


Fig -12: Paid App Ratings

3.11 Category of Apps

From the below chart we can find that most of the apps which are on Google Play Store belong to Family, Gamming and Tools.

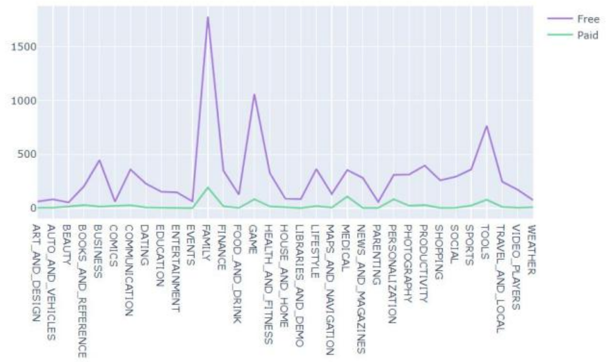


Fig -13: Paid App Ratings

3.12 Android Version

Most of the apps in Google Play store are of Android version 4.1 and up.

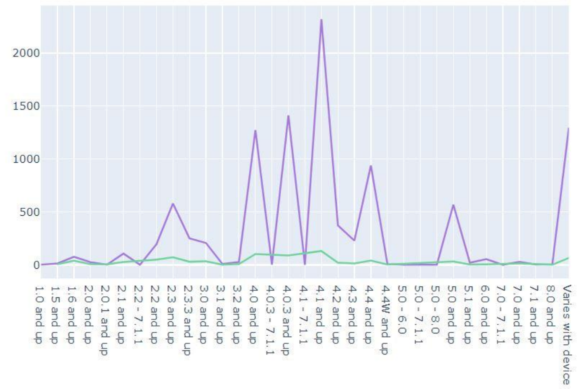


Fig -14: Paid App Ratings

3.13 Number of installations

From the below plot highest installs of the apps are crossing a Million and then 10 Million, very less apps are crossing the 500M and dream install **1B** . Some apps like Instagram, YouTube, Facebook, WhatsApp, etc. are crossing the dream install 1B.

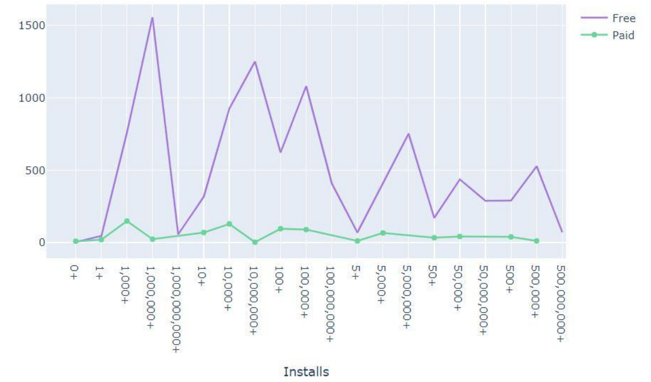


Fig -15: Paid App Ratings

3.14 Content Rating

The apps which are available for everyone are having the ratings 4 and above out of 5.

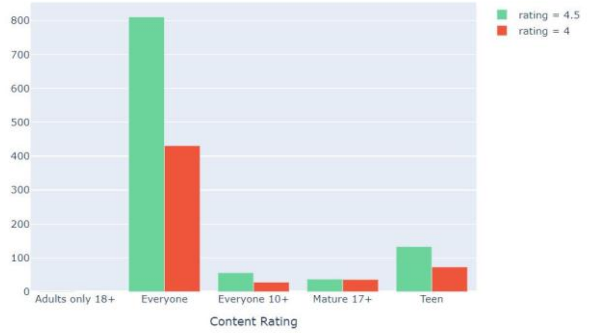


Fig -16: Paid App Ratings

3.15 Ratings over the Android Version

The Android version 4.1 and above have the ratings 4 and above.

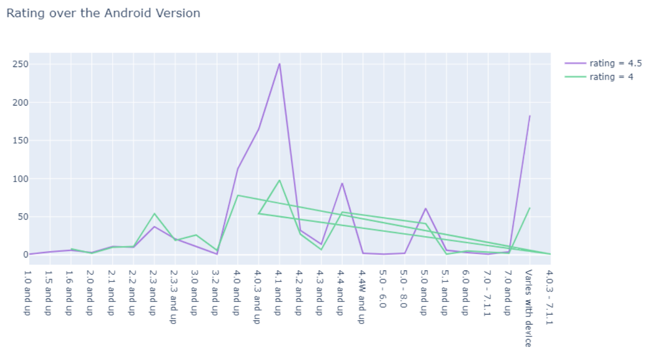


Fig -17: Paid App Ratings

3.16 Category wise Rating

The Family, Game and Tools Category has got the highest ratings i.e. 4 and above.

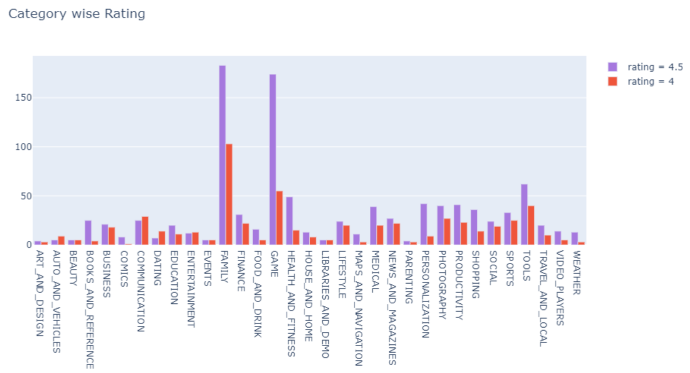


Fig -18: Paid App Ratings

3.17 Ratings over Installations

The apps which has got the 1M and 10M installations has got the ratings 4 and above.

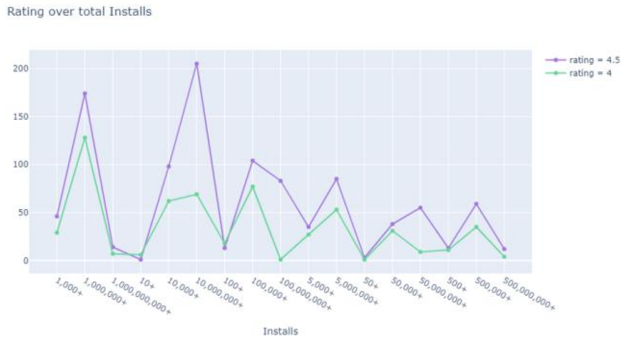


Fig -19: Paid App Ratings

4. DATA PREPROCESSING

The dataset collected from the Google Play store is semi structured or unstructured and contains significant superfluous data (defined as not contributing significantly to the prediction process). Since large datasets require longer training times, and because "stop words" reduce the prediction accuracy, text preprocessing is therefore required to overcome this limitation. Preprocessing involves various tasks including stemming, lowercase conversion, punctuation, and excluding terms.

4.1 DESCRIPTION OF DATA

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 app - Children/Adult
Genres	App's Genre

Table -1: Attributes

4.2 SOFTWARE

We Use Anaconda Navigator to launch Jupyter Notebook. Then we choose Python 3 for coding.

4.3 DATA PREPARATION

4.3.1 PRE PROCESSING

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

4.3.2 FEATURE SELECTION

Training a supervised machine learning algorithm requires textual documents to be represented in vectorial form. For this purpose, textual data must be converted into numbers without losing information.

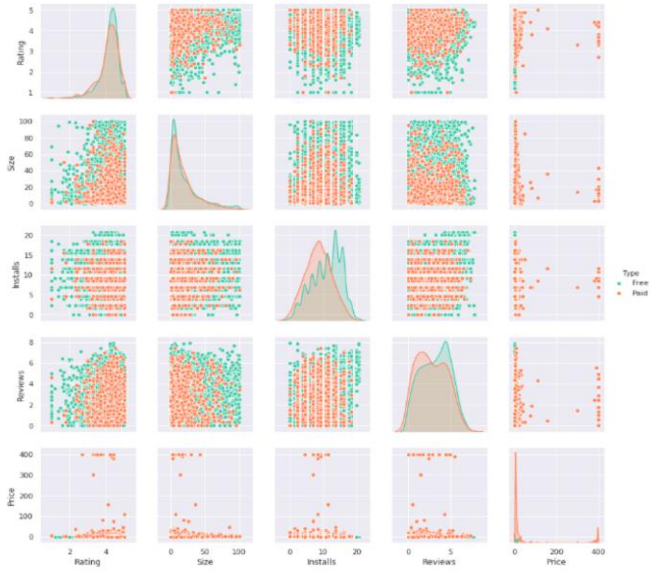
4.3.3 NORMALIZATION

The terms normalization and standardization are sometimes used interchangeably, but they typically ask various things. Normalization usually means to scale a variable to possess a worth between 0 and 1, while standardization transforms data to possess a mean of zero and a typical deviation of 1. Normalization makes training less sensitive to the size of features, so we will better solve for coefficients. Standardizing tends to form the training process well behaved because the numerical condition of the optimization problems is improved. We didn't normalize or standardize this dataset, it had been not necessary.

4.4 STANDARD DEVIATION

The standard deviation is square of variance. That's one among the ways to live the info. To calculate the quality deviation is first find the mean. Then each number are going to be subtracted from mean. Then take the mean of these squared differences. Then take the squared difference then we'll be done.

Below these plots can help us learn more about our data.



This is the basic exploratory analysis to look for any evident patterns or relationships between the features.

Fig -20: Whole Data Charts

4.5 CORRELATION MATRIX

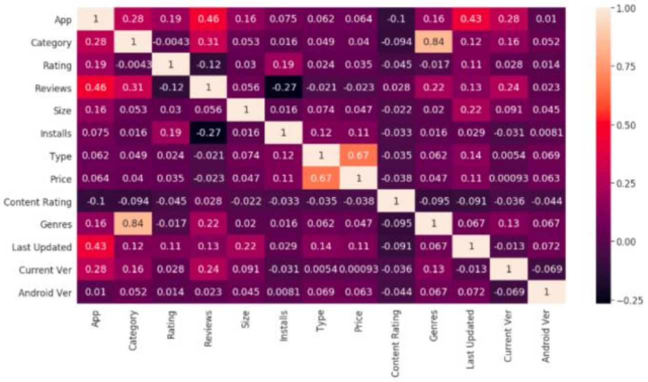


Fig -21: Correlation Matrix

5. ALGORITHMS

5.1 Random Forest

Random forest regression is applied to all the variables the results of random forest determine the importance of all the variable and their influence on the rating. The results of random forest regression are evaluated using Mean Square Error. Random forest model is the first model that is applied to the dataset and the results of Random forest classification are computed for a number of variables to find the importance of these variables.

5.2 Support Vector Regression

As Support Vector Regression (SVR) is a promising regression model for continuous variables, it is used to find the importance of all the numeric variables. In this model, only numeric values are used so the importances of these variables are computed with the rating. As SVR is usually used for continuous numeric data, this model is applied only on the numeric variables so the importance of those variables can be find out.

5.3 Linear Regression

Linear Regression model is also used to find the variable importance of different variables with ratings. Although linear regression model is a simple regression model, it sometimes produces better results than other complex models. In this model, only numeric values are used. Therefore, when this model is applied to the dataset, only the numeric variables are considered.

5.4 K Nearest Neighbors

KNN is easiest supervised machine learning algorithm. It's the foremost basic machine learning algorithm you'll find on scikit-learn. We will use KNN solve complicated problems. With the assistance of KNN we will do pattern recognition and data processing. KNN defines the similarity. From the given dataset KNN finds common groups between attributes. We split the info into training and test set. Then we will see what proportion similarity it becomes on the result.

5.5 K Means Clustering

K-Means Clustering Algorithm works by comparing or using similarity for every datum. This is often commonly used with Unsupervised Learning. The K term means the amount of groups clustered. Then we get the results of the centers of the clusters. Then we will plot this also as our original clustered data. For instance, K-Means you ought to be an expert in your data. This mean clustering numbers should be known before you run your algorithm. K-Means works when the info is in additional of a flipped cone shape.

Algorithm	Accuracy
Random Forest	73.55%
SVR	76.49%
Linear Regression	72,45%
K- Nearest Neighbor	92.22%
K-Means Clustering	69.56%

Table -2: Accuracy of Algorithms

6. CONCLUSIONS

After undergoing these algorithms and process, we concluded that our hypothesis is true. Meaning you can predict the app ratings, however significant preprocessing must be done before you start the classification and regression processes.

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market! This shows that given the

Size, Type, Price, Content Rating, and Genre of an app, we can predict about 92% accuracy if an app will have more than 100,000 installs and be a hit on the Google Play Store.

User reviews are limited to identifying polarity and subjectivity. However, the massive increase in review-based data implies a requirement to focus also on performing predictions. This process is challenging yet fruitful, as user reviews are qualitative while ratings are essentially quantitative. The numeric scoring of apps within the Google App store could also be biased and overrated because higher ratings given by users potentially attract several new users disproportionately. This study therefore investigated the utilization of ensemble classifiers to predict numeric ratings for Google Play store apps supported the user reviews for those apps. Several ensemble classifiers were investigated to guage their performance on the reviews scraped from the Google App store. Future work includes the implementation of the deep learning technique to predict numeric rating.

REFERENCES

[1] Statista, Number of available application in the Google Play store from December 2009 to March 2019, https://www.statista.com/ statistics/266210/number-of-available-applications-in-the-googl e-play-store/, Online: accessed 22 May 2019.

[2] Statistaa, Number of mobile app downloads worldwide in 2017, 2018 and 2020 (in billions), https://www.statista.com/statistics/ 271644/worldwide-free-and-paid-mobile-app-store-downloads/, Online: accessed 22 May 2019.

[3] J. Horrigan, Online shopping, pew internet and American life project, Washington, DC, 2018, http://www.pewinternet.org/Repor ts/2008/Online-Shopping/01-Summary-of-Findings.aspx Online: accessed 8 Aug. 2014.

[4] D. Pagano and W. Maalej, User feedback in the appstore: an empirical study, in Proc. IEEE Int. Requirements Eng. Conf. (Rio de Janeiro, Brazil), July 2013, pp. 125–134.

[5] T. Chumwatana, Using sentiment analysis technique for analyzing Thai customer satisfaction from social media, 2015.

[6] T. Thiviya et al., Mobile apps' feature extraction based on user reviews using machine learning, 2019.

[7] H. Hanyang et al., Studying the consistency of star ratings and reviews of popular free hybrid android and ios apps, Empirical Softw. Eng. 24 (2019), no. 7, 7–32.

[8] N. Kumari and S. Narayan Singh, Sentiment analysis on e-commerce application by using opinion mining, in Proc. Int. Conf.- Cloud Syst. Big Data Eng. (Noida, India), Jan. 2016, pp. 320–325.

[9] R. M. Duwairi and I. Qarqaz, Arabic sentiment analysis using supervised classification, in Proc. Int. Conf. Future Internet Things Cloud (Barcelona, Spain), Aug. 2014, pp. 579–583.

[10] H. S. Le, T. V. Le, and T. V. Pham, Aspect analysis for opinion mining of vietnamese text, in Proc. Int. Conf. Adv. Comput. Applicat. (Ho Chi Minh, Vietnam), Nov. 2015, pp. 118–123.

[11] H. Wang, L. Yue, and C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (Washington, D.C., USA), July 2010, pp. 783–792.

[12] K. Dave, S. Lawrence, and D. M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proc. Int. Conf. World Wide Web (New York, USA), 2003, pp. 519–528.

[13] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in Proc. ACL-02 Conf. Empirical Methods Natural Language Process. (Stroudsbrug, PA, USA), 2002, pp. 79–86.

[14] C. Cardie et al., Combining low-level and summary representations of opinions for multi-perspective question answering, New directions in question answering, 2003, pp. 20–27.

[15] H. Takamura, T. Inui, and M. Okumura, Extracting semantic orientations of words using spin model, in Proc. Annu. Meeting Association Comput. Linguistics (Ann Arbor, MI, USA), 2005, pp. 133–140.

[16] A. Buche, D. Chandak, and A. Zadgaonkar, Opinion mining and analysis: a survey, arXiv preprint arXiv:1307.3336, 2013.

[17] M. Suleman, A. Malik, and S. S. Hussain, Google play store app ranking prediction using machine learning algorithm, Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms, 2019.

[18] F. Sarro et al., Customer rating reactions can be predicted purely using app features, in Proc. IEEE Int. Requirements Eng. Conf. (Banaf, Canada), Aug. 2018, pp. 76–87.

[19] S. Aslam and I. Ashraf, Data mining algorithms and their applications in education data mining, Int. J. Adv. Res. Computer Sci. Manag. Studies 2 (2014), no. 7, 50–56.

[20] D. Martens and T. Johann, On the emotion of users in app reviews, in Proc. IEEE/ACM Int. Workshop Emotion Awareness Softw. Eng. (Buenos Aires, Argentina), May

2017, pp. 8–14.

[21] G. Hackeling, Mastering machine learning with scikit-learn, Packt Publishing Ltd, 2017.

[22] Scikit learn, Scikit-learn classification and regression models, http://scikitlearn.org/stable/supervised_learning.html# supervised -learning/, 10 Apr. 2019

BIOGRAPHIES

S SHASHANK
Student
ICFAI Tech Hyderabad

BRAHMA NAIDU
Professor
ICFAI Tech Hyderabad