International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Analyzing App Rating Using Natural Language Processing and Machine Learning

Y. Kamal[1], O. Sowmya[2], B. Bharath[3], B. Ramu[4]

[1,2,3,4]students, CSE department, SVP Engineering College, Andhra Pradesh, India.
Guided by: Mr. K. T. Krishna Kumar Nainar, Associate professor & T.P.O

ABSTRACT:

Daily tens of thousands of recent apps area unit other to the Google Play Store, with Associate in Nursing ever-increasing variety of designers operating alone or in teams to create them noted, all whereas facing stiff competition from around the world. as a result of most Play Store apps area unit free, the revenue model for the way in-app purchases, adverts, and Associate in Nursing memberships contribute to an app's successis hazy. As a result, instead of the quantity of cash generated, thesuccess of an Associate in the Nursing application is usually determined by {the variety the amount the quantity} of times it's been put in and also the number of client evaluations it's receivedover its life. However, thanks to inadequate or missing votes, these rating area units ofttimes compromised. what is more, their area unit respectable variations between numerical ratings and user reviews? The goal of this analysis work is to use machine learning algorithms to predict the ratings of Google Play Store apps.

KEY ELEMENTS: RANDOM FOREST, CONVOLUTIONAL NEURAL NETWORKS, K-NEAREST.

INTRODUCTION:

Users might transfer and utilize several third-party apps from the google play store. To transfer and use these applications on their humanoid phones and tablets, a legion of folks registers personal info with Google and third-party businesses. one of the fastest-growing segments of the downloaded package application trade is mobile apps. we decide on the Google play store overall alternative markets attributable to its growing quality and its fast growth. The ratings for humanoid apps would then seem to users supported the region wherever their device is registered. Developers and users primarily confirm the influence of market interactions on future technology. However, each developer and user's square measure suffers from an absence of awareness of common app markets' inner workings and characteristics. This project aims to deliver insights to grasp client demands higher and so facilitate developers' popularization.

Feedback is provided in 2 forms, text review and numeric ratings, as illustrated in Figure one. A text review, on the one hand, might contain positive or negative comments submitted by a user on a couple of specific apps or policies. This knowledge is extremely helpful for performing arts selling analysis, managing publicity, conducting product reviews, web promoter evaluation, giving product feedback, delivering client service, and so on.

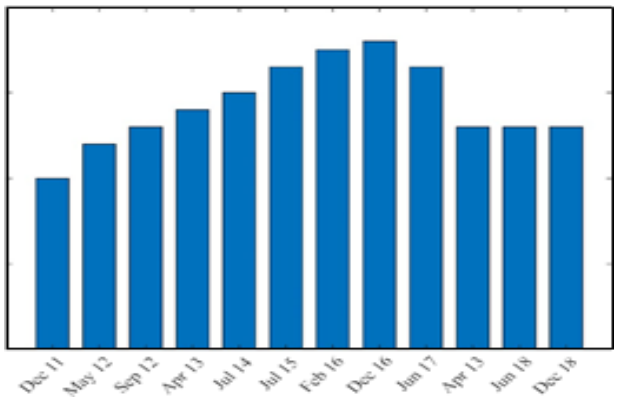
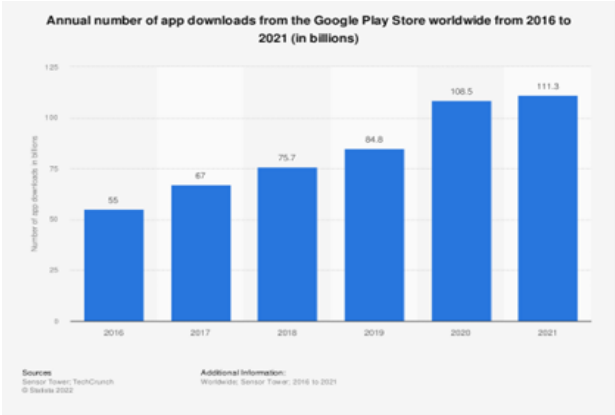


Fig1: Evolution of apps on the Google Play store [1,2]: (A) Total apps on theGoogle Play Store,



(B) Total apps downloaded from the Google Play Store

Machine learning algorithms, as well as the random forest (RF), the gradient boosting classifier (GBM), the acute gradient boosting classifier (XGB), the AdaBoost classifier (AB), and also the additional tree classifier (ET) was applied to predict numeric ratings.

- The classifier accuracy was analyzed from the subsequent 2 perspectives: (a) exploitation of the matter options derived from the reviews alone, and (b) exploitation of each of the matter options and also the emoticons offered within the reviews.
- For validation, noted apps happiness to every class were accustomed compare the numeric ratings expected by ensemble learning with the user's actual ratings.

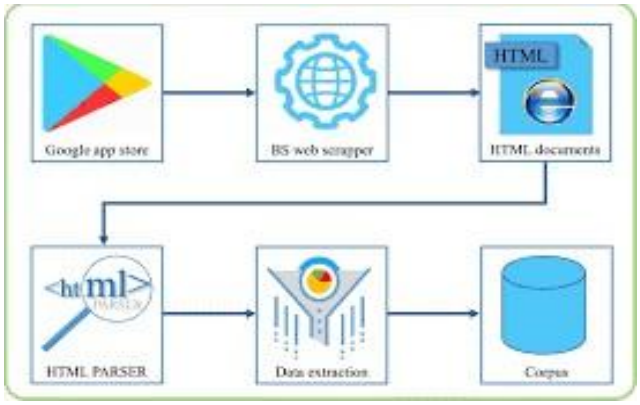


Fig 2: Flow diagram of app accessibility

METHODOLOGY:

Pandas and the Scikit Learn package were used to clean the data.

1. We encountered Null values in columns that were laterdiscarded.
2. Variables with an incorrect datatype and Inconsistentformatting was corrected.
   - 2.1 Size column has sizes in KB as well as MB.
   - 2.2 Review is a numeric field that is loaded as a string
   - 2.3 Installs field is currently stored as astring and has values of 1,000,000+.
   - 2.4 Price field is a string and has asymbol.
3. Rows with incorrect values are identified and discarded.
4. Outliers exist in columns such as installs, reviews, andpricing, which are removed using IQR.
5. Dropped unnecessary columns like “App and”, “Last Updated”, “Current Version”, and “Android -Version”.

EXISTING SYSTEMS:

App Ratings are expected to support the options provided for the app. Experiments were performed on the BlackBerry World and Samsung golem stores to gather the raw options provided for the apps, together with their value, the rank of downloads, ratings, and matter descriptions. The options were then encoded into a numerical vector to be employed in case-based reasoning and to predict the app rating. In distinction to the above-cited studies, different authors. Review-based prediction systems permit this unstructured info to bemechanically remodeled into structured knowledge reflective belief. This structured knowledge is often used as a life of users' sentiments regarding specific applications, products, services, and brands. they will thus give vital info for product- service vices refinement. this sort of sentiment analysis was conducted in the following studies.

Various sentiment analysis ways are performed to summarizethe ensembles of comments and reviews. These ways use mathematical and applied math ways (especially involving mathematician distributions) to beat the issues encountered in sentiment analysis. though these authors projected a model, it was not enforced. A recent study investigated the appliance of a machine-learning rule to a dataset covering, for instance, the appclass, the numbers of reviews and downloads, the size, type, Associate in Nursing golem version of an app, and therefore the content rating, to predict a Google app ranking. call trees, regression, supplying regression, support vector machine, NB classifiers, k-means bunch, k-nearest neighbors, and artificial neural networks were studied for that purpose.

Approaches are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. in keeping with our study, we decide on the supervised machine learning approach because the approach is incredibly smart atthe subsequent classes' binary and multi-class classification, regression, and aggregation.

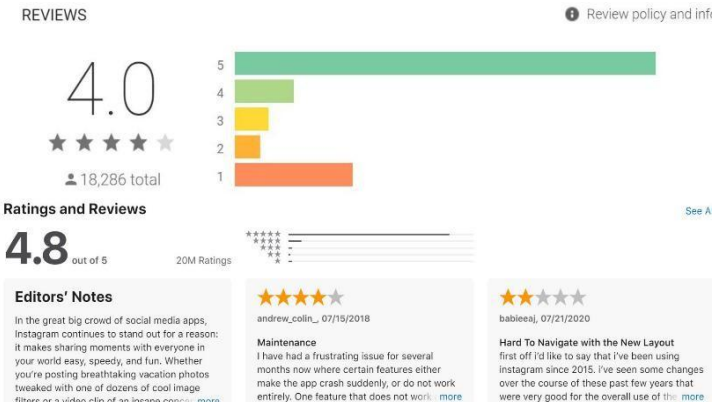


Fig3: Existing features in-app rating

RELATED WORK:

This section describes the planned approach, its modules, andtherefore the dataset employed in the experiment.

The design of the planned approach for predicting numeric ratings is printed in Figure three. It involves many subtasks, represented on an individual basis below, prompt adopting an applied mathematics analysis supported by a spin model, to extract the linguistics orientations of words. Mean-field approximations werethe accustomed reason for the approximate chance within the spinmodel. linguistics orientations area unit is then evaluated as fascinating or undesirable. A smaller range of seed words for the planned model turns out extremely correct linguistic orientations supported English lexicon. Google apps. First, text-mining techniques are unit ineffective once applied to app reviews, because it has Unicode-supported language with a restricted rangeof words. Second, those studies area unit based mostly either on rating predictions created exploitation inherent app options or on external options (e.g., price, bug report, etc.). None of these studies investigated the potential discrepancies between users' numeric ratings and reviews. To our information, this study is the initial research on such discrepancies.

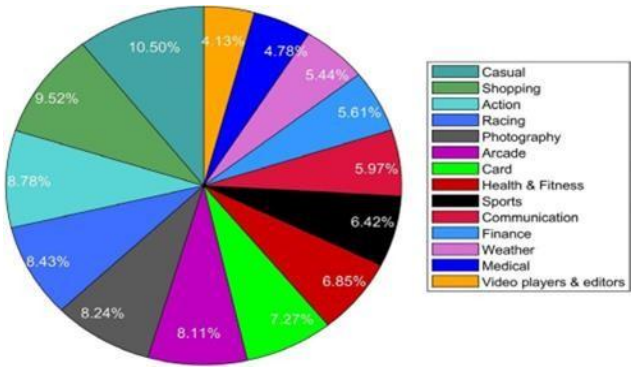


Fig3: Pie chart of user ratings

DATA SECTION:

DATA SET: The Google apps dataset was scraped from the Google Play store using the *Beautiful Soup* web scraper. The data were scraped for applications released no later than 2014 to ensure a minimum period of 5 years. The following criteria were applied:

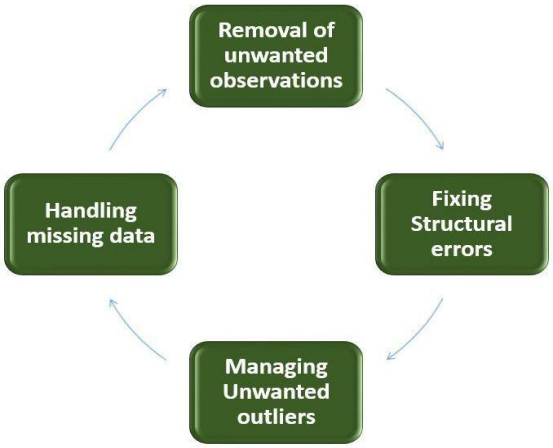
	App	Category	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up	4.1
1	Coloring book moana	ART_AND_DESIGN	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	3.9
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	4.7
3	Sketch - Draw & Paint	ART_AND_DESIGN	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	4.5
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up	4.3

Fig4: Data Set of google play store

PROCESSING DATA:

Preprocessing information is split into two elements, namely: cleanup information and information reduction. information cleanup is the method of cleanup incomplete information on the attributes within the dataset to create the information additional consistent. Meanwhile, information reduction is the method of removing information on less dominant attributes so information is reduced but still manufacture correct information. within the information cleanup method, the authorclassifies and assigns the ranking information label to be high rated (> three.5) and low rated (≤ three.5), remove the k and m symbols within the size column, removes the + image within the installs column and within the information reduction method the author deletes the information that's within the attributes current version, automaton version, genre, and last updated, understand the record format and then needed to check the tuplesand attributes and same or not they are being understood and need to change to normal objects.

DATA CLEANING: In this set, if any noises or inconsistencies are there it will check and clean the data to get clear and perfect data in a perfect stream without having any missing or duplicated data.



DATA TRANSFORMATION:

Data transformation is the method of adjusting the format, structure, or values of knowledge. For knowledge analytics comes, knowledge could also be reworked at 2 stages of the info pipeline. Organizations that use on-premises knowledge warehouses usually use AN ETL most organizations use cloud- based knowledge warehouses, which may scale cipher and storage resources with latency measured in seconds or minutes.

The data should so be normalized before beginning the coaching method exploitation of the planned approach. To perceive text preprocessing, take into account the instance of a review for the mobile app. A series of preprocessing steps and their output are shown in fig4. when the completion of preprocessing, the ensemble learning classifiers are often applied to the processed dataset.

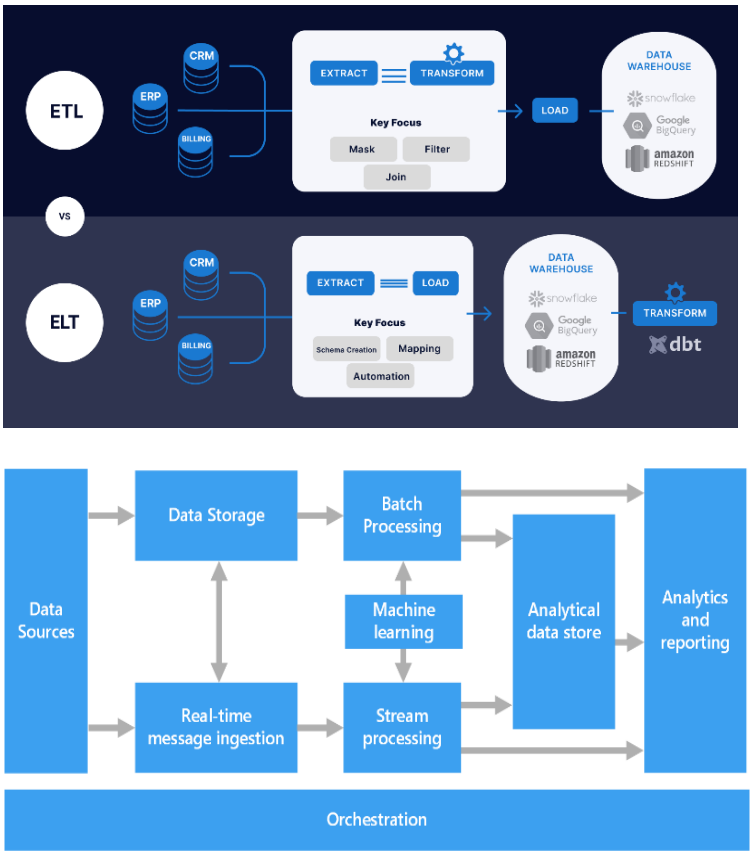


Fig5: Architecture of the proposed approach

SYSTEM REQUIREMENTS SOFTWARE REQUIREMENTS:OPERATING SYSTEM: Windows 7 CODING LANGUAGE: python

TOOL: JUPITER Notebook **HARDWARE REQUIREMENTS**HARD DISK: 256 GB

MONITOR: LED RAM: 2 GB. **ALGORITHMS**:

This research has been used for accuracy are Machine Learning,

Natural Language Processing, and Data Analytics:

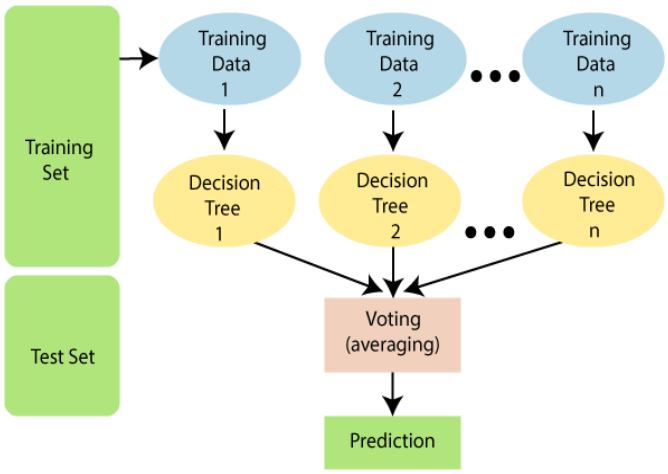
- RANDOM FOREST
- POLYNOMIAL REGRESSION
- DECISION TREE REGRESSION

RANDOM FOREST:

Random Forest could be a common machine learning rule that belongs to the supervised learning technique. It will be used for each Classification and Regression issue in a milliliter. it's supported the construct of ensemble learning, which could be a

method of mixing multiple classifiers to unravel a fancy drawback and improve the performance of the model Dependent variable: the variable, that's to be understood or to be forecasted that variable is thought because of the variable. Since the random forest combines multiple trees to predict the class of the dataset, some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- Independent variable: this factor that influences the target variable or the dependent variable and provides the information that belonging the relationship with thedependent variable or the target variable.



POLYNOMIAL REGRESSION:

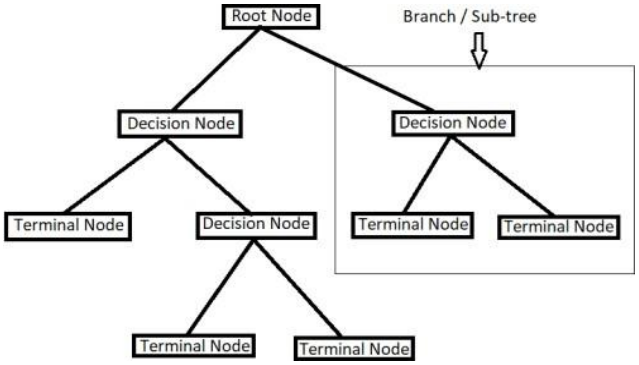
In polynomial regression, the link between the experimental variable x and therefore the variable quantity y is delineated as associate ordinal degree polynomial in x. Polynomial regression, abbreviated E(y|x), describes the fitting of a nonlinear relationship between the worth of x and therefore the conditional mean of y. it always corresponded to the least-squares methodology. consistent with the Gauss Andre Markoff Theorem, the smallest amount sq. approach minimizes the variance of the coefficients. this is often asort of regression toward the mean during which the dependent and freelance variables have a curving relationship and therefore the polynomial equation is fitted to the data; we'll check that in additional detail later within the article. Machine learning is additionally cited as a set of Multiple regressions toward themean. as a result, we have converted the Multiple regression toward the mean equation into a Polynomial regression of y on x as well as a lot of polynomial regressions.

Polynomials	Form	Degree	Examples
Linear Polynomial	$p(x): ax+b, a \neq 0$	Polynomial with Degree 1	$x + 8$
Quadratic Polynomial	$p(x): ax^2+b+c, a \neq 0$	Polynomial with Degree 2	$3x^2-4x+7$
Cubic Polynomial	$p(x): ax^3+bx^2+cx, a \neq 0$	Polynomial with Degree 3	$2x^3+3x^2+4x+6$

DECISION TREE REGRESSION:

A decision tree may be a flowchart-like structure within which every internal node represents a check on a feature (e.g., whether or not a coin flip comes up with heads or tails), every leaf node represents a category } label (decision taken when computing all options) and branches represent conjunctions of features that resultin those class labels. The methods from the foundation to the leaf represent classification rules. The below diagram illustrates the fundamental flaw of the choice tree for higher cognitive processeswith labels (Rain (Yes), No Rain (No)).

- Decision tree is one of the predictive modeling approaches used in statistics, data mining, and Machine learning.



Conclusion:

Feeding the info to the model, a big quantity of pre-processing isneeded. The quantitative score of apps within the Google Play Store might also be skew and exaggerated as a result of higher user ratings that might attract additional new users. However, because of inadequate or missing votes, this rating area is a notary and oftentimes biased. This analysis aims to use machine learning techniques to forecast the ratings of Google Play Store apps.

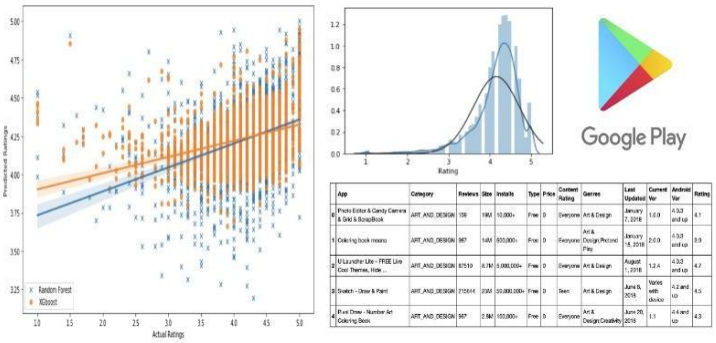
Google Play stores apps supported the user reviews for those apps. many ensemble classifiers were investigated to judgetheir performance on the reviews scraped from the Google Appstore. TF/IDF options with unigrams, bigrams, and trigrams were utilized with the chosen classifiers. the bottom truth was calculated employing a technique supported by text Blob analysis, that identifies the reviews showing a discrepancy withthe user-awarded rating. later, it had been accustomed to judging the performance of the chosen classifiers. Text Blob analysis showed that twenty-four. 72% of the user-defined app rating square measure biased. Results demonstrate that tree- based textile ensemble classifiers perform far better than boosting-based classifiers on account of their support for nonlinearity, collinearity, and tolerance to information noise. The analysis conjointly reveals that the user views square measure inconsistent with user numeric ratings which numeric rating square measure on the top of user reviews would possibly recommend. Future work includes the implementation of the deep learning technique to predict numeric ratings. The authors declare no conflict of interest. The funders had no role within the style of the study; within the assortment, analyses, orinterpretation of data; within the writing of the manuscript; or in the call to publish the results. This model gives the result mostly depending on the world healthorganization basis of data. This model gives the highest accuracy and surely be used for corrective measures.

Experiments demonstrate that the prediction accuracy is often improved. For this purpose, every application class was trained one by one exploitation RF before creating predictions showing the result obtained by coaching every class of Google apps exploitation RF. The exactitude, recall, and F-score values were obtained by averaging the ratings from every category. Averages were calculated exploitation the Scikit learn analysis metrics library. The results demonstrate that coaching every class one by one to form a prediction yields higher accuracy than once exploiting all classes combined.

The accuracy of RF and GBM reflects the discrepanciesbetween the user-specified numeric rating and reviews. Thenumeric rating is approximately 20% higher than the outcome of ensemble classifiers.

FUTURE SCOPE:

In this study, the improvement is to explore the exploitation of associate degree updated information for predictions. For future work and any improvement may be done by this prediction methodology to urge the very best correct results and may be done on varied knowledge sets like IOS apps and additional economical google play stores within the immense data set assortment given bygoogle.



	App	Category	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Rating
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up	4.1
1	Coloring book moana	ART_AND_DESIGN	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up	3.9
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up	4.7
3	Sketch - Draw & Paint	ART_AND_DESIGN	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up	4.5
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up	4.3

LEARNING FOR PROJECT:

The following aspects area unit learned whereas collaboratingon the project:

- Pandas area unit essential for knowledge validation, cleaning, selection
- Seaborn, Matplotlib is employed for visual image.
- They recognized outliers and knowledge distribution in every various victimization numerous plots like box, bar, and accumulative distribution plots for every assortment of attributes.
- Statistics area unit important in deciding the accuracy in knowledge, like mean, median, and variance, that aims to attenuate error rates.
- Victimization Inter-Quartile-Range and a log transformation to get rid of non-linearity, outliers area unit treated, and lopsidedness is reduced.
- Different algorithms were wont to take a look at prediction models. Comparing the outcomes of varied R2-scores and error rates aids in the identification of a superior model.
- I'm plotting the results on the graph victimization seaborn, visualizing the output to know the best-fitting line higher.
- Massive amounts of knowledge area unit needed to feed the rule want to observe and visualize patterns.

REFERENCES:

[1] Tressa, Eleonora MC, et al. "Mobile Applications in Otolaryngology: A Systematic Review of the Literature, Apple App Store and the Google Play Store." Annals of Otology, Rhinology & Laryngology 130.1 (2021): 78-91.

[2] Hassan, Safwat, et al. "Studying the dialogue between users and developers of free apps in the google play store." Empirical Software Engineering 23.3 (2018): 1275- 1312.

[3] Venkatakrishnan, Swathi, Abhishek Kaushik, and Jitendra Kumar Verma. "Sentiment analysis on google play store data using deep learning." Applications of Machine Learning. Springer, 2020. 15-30.

(2020).

[4] Karim, Abdul, et al. "Classification of Google Play Store Application Reviews Using Machine Learning."

[5] Reddy, Palagati Bhanu Prakash, and Ramesh Nallabolu. "Machine learning-based descriptive statistical analysis on Google play store mobile applications." 2020 Second International Conference on Inventive Research in Computing Applications (CIRCA). IEEE, 2020.

[6] Karim, Abdul, et al. "Methodology for Analyzing the Traditional Algorithms Performance of User Reviews Using Machine Learning Techniques." Algorithms 13.8 (2020): 202.

[7] Islam, Mir Riyanul. "Numeric rating of Apps on Google Play Store by sentiment analysis on user reviews." 2014 International Conference on Electrical Engineering and Information & Communication Technology. IEEE, 2014.

[8] Ali, Mohamed, Mona Erfani Joorabchi, and Ali Mesbah. "Same app, different app stores: A comparative study." 2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILE Soft). IEEE, 2017.

[9] Sadiq, Saima, et al. "Discrepancy detectionbetween actual user reviews and numeric ratings of Google App store using deep learning." Expert Systems with Applications 181 (2021): 115111.

[10] Mahmood, Ahsan. "Identifying the influence of various factors of apps on google play apps ratings." Journal of Data, Information and Management 2.1 (2020): 15-23. [11] Umer, Muhammad, et al. "Predicting numeric ratings for Google apps using text features and ensemble learning." ETRI Journal 43.1 (2021): 95-108.