GOOGLE PLAY STORE APP RATING PREDECTION

KOLLATI RAJA KUMAR, A. DURGA DEVI
ASSISTANT PROFESSOR IN NAME DEPARTMENT OF MASTER OF COMPUTER SCIENCE, BHIMAVARAM 534202.
Email id : adurgadevi760@gmail.com
PG STUDENT, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.
Email id : rajkumarkollati1@gmail.com

ABSTRACT

Application distribution platform, for example, Google play store gets overwhelmed with a few thousands of new applications regularly with a lot progressively a 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, it is basic for a developer to know whether he is continuing the correct way. Dissimilar to making movies where the nearness of famous heroes raise the likelihood of accomplishment even before the movies are coming into the picture, it isn't the situation with creating applications. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application 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. So in this project, I have tried to perform analysis and prediction into the Google Play store application dataset that I have collected from kaggle.com. Using Big Data techniques such as Machine learning 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. And using Deep Learning I have tried to make a prediction about the user reviews that which review is positive or negative.

1. INTRODUCTION

Big Data is likewise information yet with a gigantic size. Big Data is a term used to portray a gathering of information that is big in size but then developing exponentiallywith time. In short such information is so substantial and complex that none of thecustomary information the board devices can store it or procedure it proficiently.

We can define a data is a big data with the help of these 5V‟s:

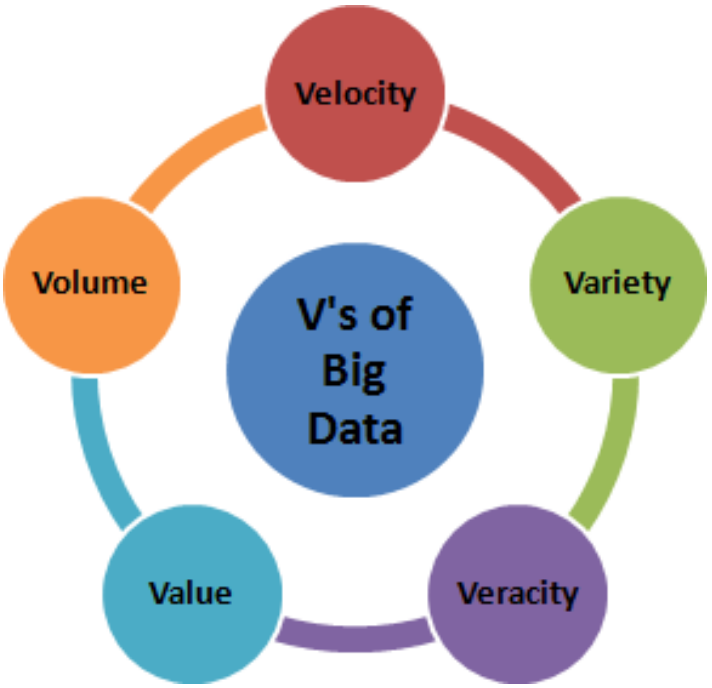


Figure 1: 5Vs of BIG DATA

- Volume: Volume means a huge amount of data that is generated in every second from any social media, cars, bank, from flights etc.
- Velocity: It means speed at which the data is generated and collected, alsoanalyzed.
- Variety: It means different types of data we are having and using it.
- Veracity: It refers to the quality and security of the data.

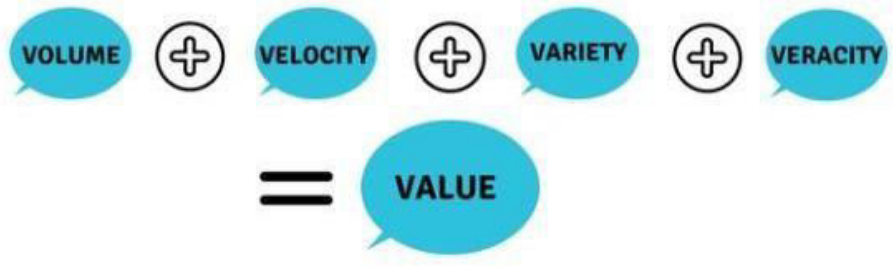


Figure 2: Value in 5Vs of BIG DATA

1.Value: In this we refer to the worth of the data that is going to be extracted.

1 We can perform any type of Analysis using Big Data with the following way:

2 Collection of Data

3 Classification of Data

4 Identification of Pattern

5 Finally the Prediction

6 Visualization

2. LITERATURE SURVEY AND RELATED WORK

In the recent years, enormous work is carried out in the domain of Weather forecasting. Weather forecasting is one of the applications to predict state of climate in future at agiven location.

2.1"BIG DATA TECHNIQUES FOR EFFICIENT STORAGE AND PROCESSING OF WEATHER" [1]

This Research paper proposes an efficient Big Data technique for storage and processing of weather data. In general Apache Hadoop framework is most popularly use for storing and processing of enormous dataset. During this study, Apache Spark and Cassandra integration is experimented to judge the time taken to efficiently store any datasets and process it and therefore the result is evaluated with Hadoop Map Reduce

Weather datasets is collected from National Climatic Data Centre (NCDC). In weather forecasting the raw information is received through satellite delivered over to the various weather stations and this data stored in cluster. Traditional Database like SQL are not best to handle unstructured data or weather data. Input datasets contain field like location, date, temperature, humidity, pressure, rain, wind etc.

METHODOLOGY USED IN THIS RESEARCH PAPER:-

- Hadoop Map Reduce implementation:-Hadoop Framework works on parallel processing distributed system which are conventionally based on map reduce jobs. The Input data is splitted into split. These splits are passed to mapper and the resultoutput is given as input to reducer. Hence in this research paper spark used to process weather data compared to Hadoop Map Reduce.
- Spark implementation: - Due to its in-memory computation it can perform 10x better than Map Reduce. Core concept in Apache Spark is RDD which act as a table in database and can hold various type of data and store on different partition.
- NoSQL Database Cassandra:-NoSQL Database provides a provision for storage and retrieval of unstructured data unlike the traditional database which use tabular relations. NoSQL Database are being efficiently used for real time web application and high speed online transactional data.
-

2.2 "COMMERCIAL PRODUCT ANALYSIS USING HADOOP MAP REDUCE" [2]

It examines how an association can find certified open entryways in solidifying disengaged and online data to give cleverness on how consolidating separated and online data can be helpful. Associations use proposition estimations which have the above favorable circumstances. Proposition computations are best seen for their use on online business Web destinations. Here they use customer's interests as a commitment to make a record of endorsed things.

The first one is called content based sifting. Content based separating can moreover be called as intellectual sifting, which endorses things dependent on an examination between the substance of the things and a customer profile.

Additionally, the next one is community oriented sifting. It relies on not just the attributes of the things yet rather how person's for example various customers respond to comparable articles. Affiliations need to get all of the data characteristics, detached and on the web, into a lone database, which would be moreover refined by front line examination procedures, and use the solidified data for exactness concentrating on.

3. METHODOLOGIES

MODULE

There are six modules in the project are:

1. Descriptive Analysis
2. Exploratory Data Analysis
3. Data Preprocessing and Visualization
4. Feature Engineering
5. Model building and Evaluation
6. Model Deployment

1) Descriptive Analysis:

In this module, Categorical data are first described by counting the number of observations in each category, then expressing them as a percentage of the overall sample size. The dataset is first previewed followed by identifying the dimensions (rows, columns) of the dataset and the data types contained within it, such as target variables, unique elements within each attribute. A

statistical summary of the dataset is then obtained. Null and unique values should be checked and the range of the target variable should be determined.

2) Exploratory Data Analysis:

First, some basic data exploration is done and then inference is made based on the assumptions. Univariate analysis is performed initially on the target variable along with numerical predictors and categorical predictors respectively. Then, bivariate analysis is conducted on numerical variables. Multivariate analysis is also performed for one's reference, and all the plots that we obtained are analysed. This module serves the purpose of taking a closer look at the data, including any irregularities, as well as correcting any inconsistencies for the next module, Data Pre-Processing.

3) Data Preprocessing and Visualization:

Categorical features are processed through label encoding, and different types of graphs are plotted accordingly. The missing values are filled in and the unnecessary columns are removed. For the pre-processed data, we identify the top columns which have the greatest influence on the Purchase amount.

1) Feature Engineering:

In this module, the features are scaled and transformed. Then, variables and interaction variables are derived. A function is created to count features. At last, the dataset is divided into test and training set, all the unnecessary columns are dropped and files are exported as modified versions.

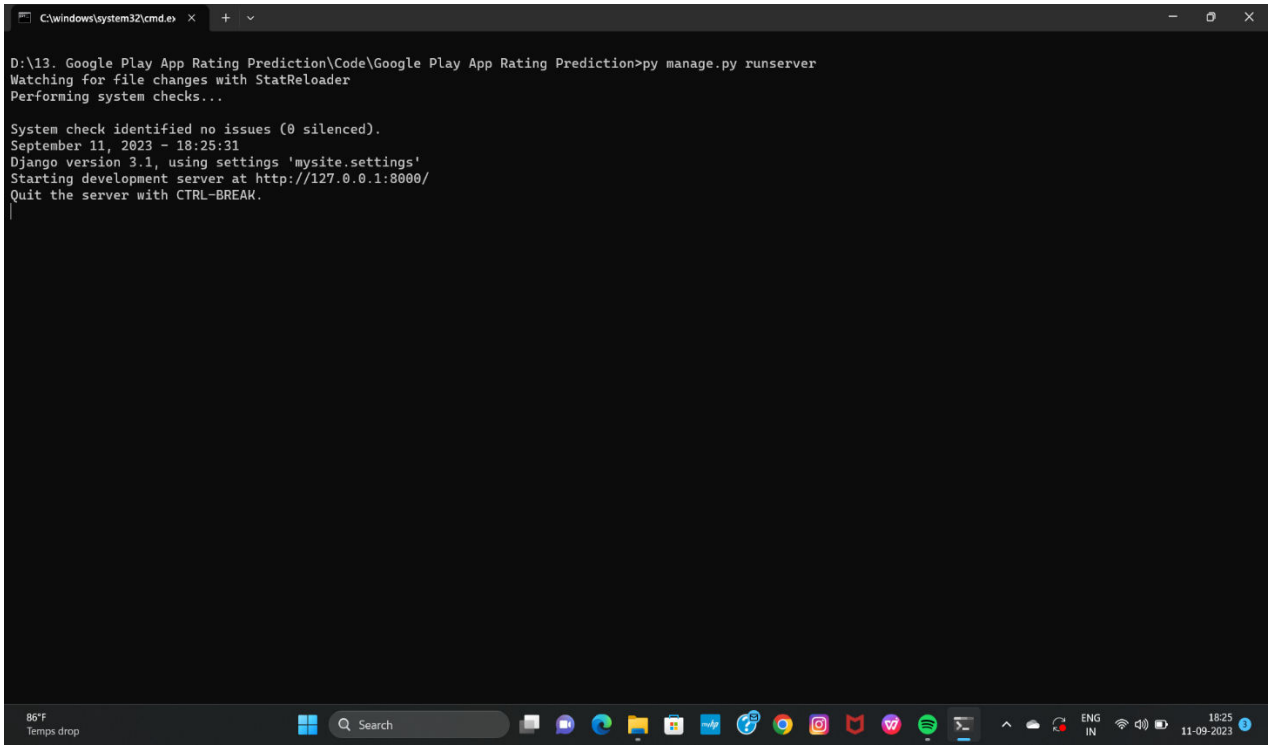
2) Model building and Evaluation:

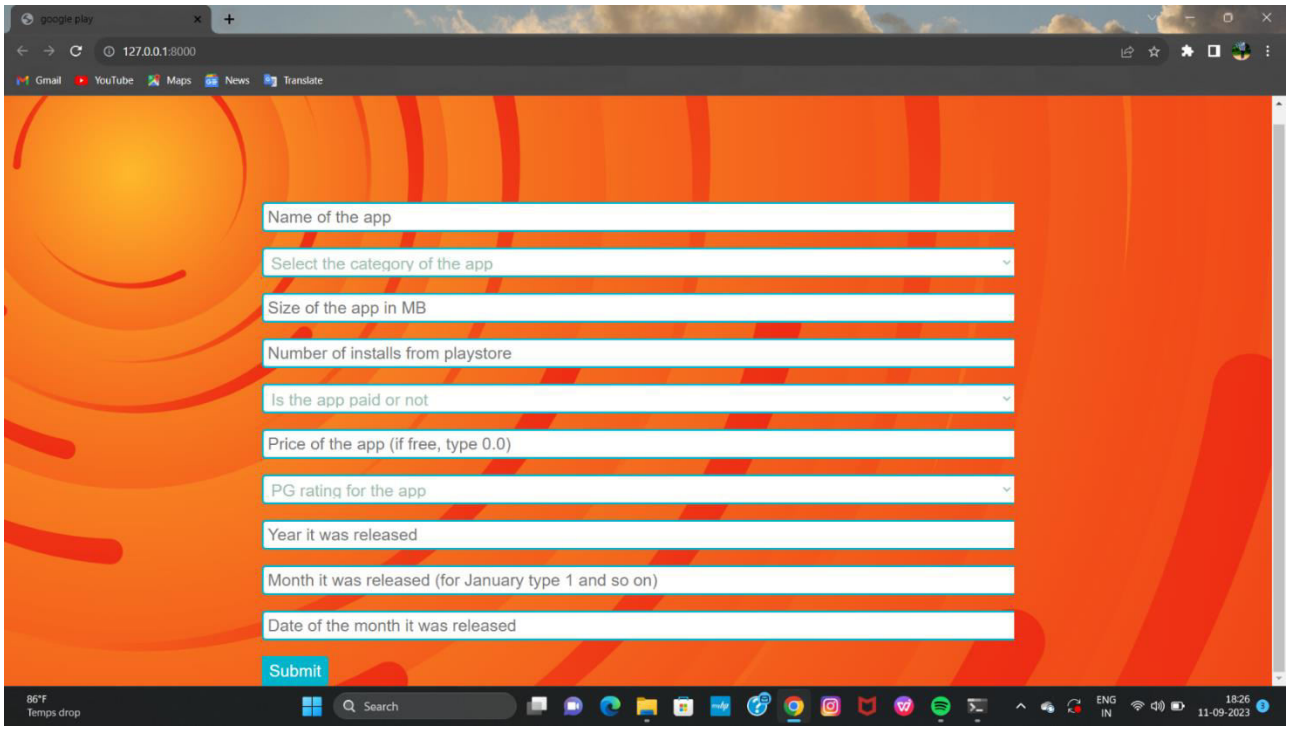
As many models are built, a generic function will be defined which takes the algorithmand data as input and makes the model, performs cross-validation and generatessubmission. The models include Ridge Regression, Decision Tree, Random Forrest, ADABoost, XG Boost.In order to evaluate the model, we will use two metrics: root mean square error (RMSE)and R squared score (r2 score). In statistics, the root of the square root of the variance of the errors is the RMSE.A model with a lower RMSE value is better. The R squared is a statistical measure of howclose the data are to the fitted regression line. Its value is between 0 and 1, and the higher the value the better fit the model is.

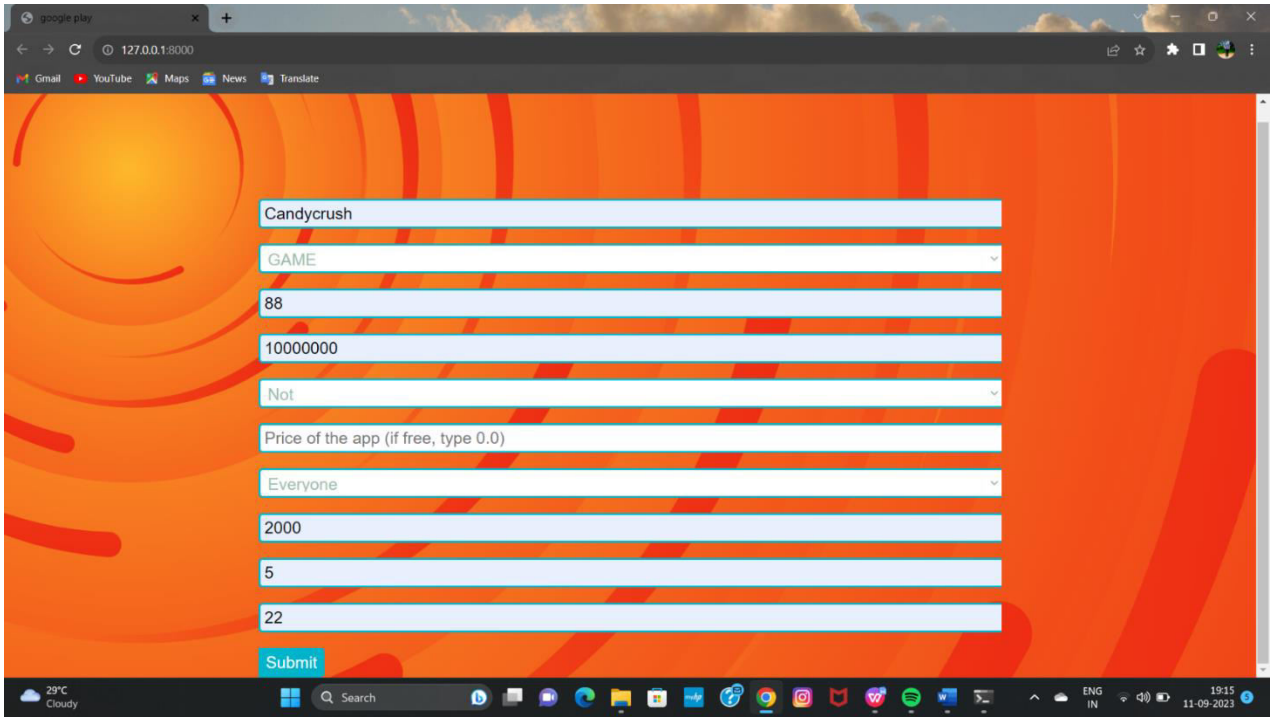
3) Model Deployment:

The best-fit model among all the 5 models is deployed to the User interface. Theuser will be able to able to enter the numerical and categorical values to predict purchaseamount. The predictor should able to validate data if it is within the given range anddisplay the purchase amount.

4. RESULTS AND DISCUSSION SCREEN SHOTS




google play
127.0.0.1:8000
Gmail
YouTube
Maps
News
Translate
Name of the app
Select the category of the app
Size of the app in MB
Number of installs from playstore
Is the app paid or not
Price of the app (if free, type 0.0)
PG rating for the app
Year it was released
Month it was released (for January type 1 and so on)
Date of the month it was released
Submit
86°F
Temps drop
Search
ENG
IN
18:26
11-09-2023



5. CONCLUSION AND FUTURE SCOPE

The Google Play Store is the biggest application advertises on this earth. It produces more than the download of the Apple App Store, yet not profits as the App Store. We scratched information from the Play Store to lead examine on it. In this project I have used Big data technique such as Machine learning to analyze the different attributes of the given dataset of Google playstore application such as top free apps, top paid apps ,most reviewed apps, apps under editor"s choice with the help of Machine learningQL and displayed the results as shown above.

Moreover in the rest of my project I have tried to predict the positive or negative user review on the basis of given dataset using deep neural network. First I have read the given dataset which is in the form of text and converted it into mathematical form then pass this dataset in the deep neural network and calculate the output compare it with actual output and then train the model and adjust the weights to minimize the error using backpropagation this process is performed several times. Finally by comparing the output after training process with the set range to find out which review is positive or negative.

Accuracy of model increases to 84.3% from 50% after training the model which is far better than the previous model.

6. REFERENCES

[1] K.Anusha and K.Usha Rani, "Big Data Techniques for efficient storage and processing of weather." International Journal for Research and Applied Science & Engineering Technology (IJRASET).

[2] Kshitij Jaju1, Vishal Nehe2,Abhishek Konduri3,‟ Commercial Product Analysis Using Hadoop MapReduce‟, International Research Journal of Engineering andTechnology (IRJET, Volume: 03 Issue: 04 | April-2016 , 2016 IJSRSET | Volume 2 |

[3] Nishant Rajput , Nikhil Ganage ,and Jeet Bhavesh Thakur,"Review Paper on Hadoop and Map Reduce", IJRET: International Journal of Research in Engineering and Technology, Volume: 06 Issue: 09 | Sep-2017

[4] Andrzej Romanowski, Michal, and Skuza, "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction", 2015 FederatediConference on, pp. 1349-1354. IEEE,i2015.

[5] K. R. Srinath,‟ Python – The Fastest Growing Programming Language‟ International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12 | Dec- 2017.

[6] Ayon Dey,‟ Machine Learning Algorithms: A Review‟, Vol. 7 (3), 2016, ISSN:0975-9646.

[7] Xuedan Du, Yinghao Cai, Shuo Wang and Lejie Zhang," An Overview of Deep Learning " International Research Journal of Engineering and Technology (IRJET),
Volume: 02 Issue: 13 | Nov-2016

[8] Kleiner Perkins Caufield and Byers",Power of Mobile Applications" International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 23 | Sep- 2016.