Detailed Project Report on

Analysis of Google Apps Store Dataset

Prepared By:

Ishita Shetty

Project Details

Project Title	Analyzing Google Apps Store dataset in terms of App downloads and Rating
Technology	Business Intelligence
Domain	Technology
Project Difficulties level	Advanced
Programming Language Used	Python
Tools Used	Jupyter Notebook, MS-Excel, Tableau Public

Objective & Benefits

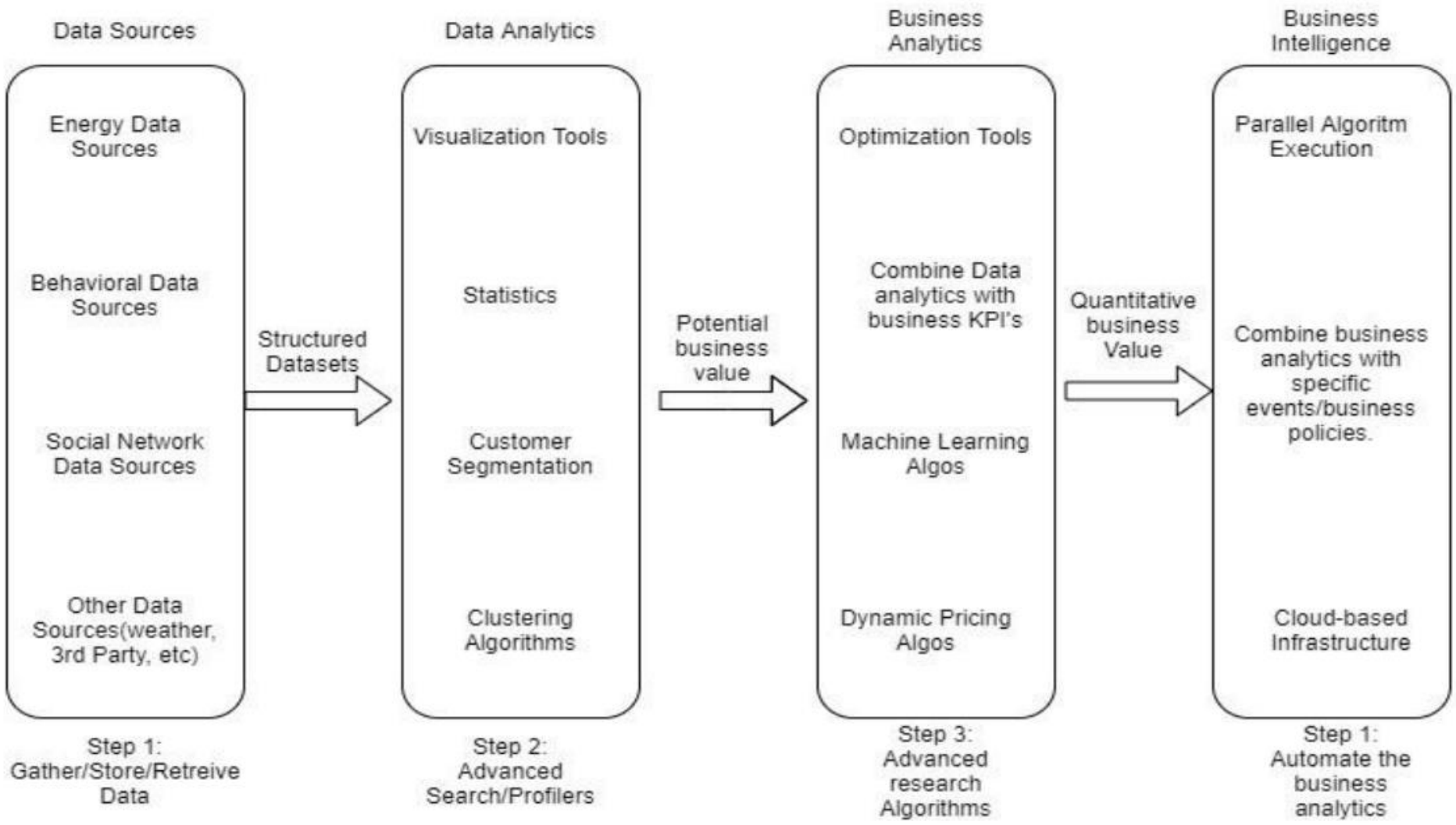
Objective:

Android is expanding as an operating system. It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. The goal is to help android developers to know what is the motivating factor for people to download an app. It will also help to find out the factors that affect someone's decision to download an app. Dive Deep in data for the factors of influences on an application, to know why and how certain applications succeed others. Also, what is required for an application to be considered as successfully topping the charts.

Benefits:

- ✓ Analyse consumer trends and determine which type of apps are the most popular and profitable.
- ✓ Classify applications based on their categories.
- ✓ Present the growth of applications through years.
- ✓ Compare different categories of applications based on the Android version.
- ✓ Compare the rates in different kinds of applications.
- ✓ Assess supported Android version with numbers of reviews based on different categories.

Architecture



Data Details

Data Sharing Agreement :

- Dataset file name: googleplaystore.csv and googleplaystore_user_reviews.csv
- Length of time stamp(6 digits)
- This dataset contains 13 different features that can be used for predicting whether an app will be successful or not.
- These columns are: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Ver, Android Ver .
- Column data type: Text and Numbers

Data Validation and Data Transformation :

- Name Validation - Validation of files name as per the DSA. We have created a regex pattern for validation. After it checks for date format and time format if these requirements are satisfied, we move such files to "Good_Data_Folder" else "Bad_Data_Folder.“
- Number of Columns – Validation of number of columns present in the files, and if it doesn't match then the file is moved to "Bad_Data_Folder.“
- Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad_Data_Folder".
- Data type of columns - The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad_Data_Folder".
- Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad_Data_Folder".

Data Description:

Variable	Significance
App	Name of the App
Category	Category of the app
Rating	Over all user rating of the app out of 5 on the Play Store
Reviews	Number of user reviews for the app
Size	Size of app
Installs	Number of user downloads/installs for the app
Type	Paid or Free
Price	Cost of the App
Content Rating	Age group the app is targeted at
Genres	An app can belong to multiple genres (apart from its main category)
Last updated	Date when the app was last updated on Play Store
Current Ver	Current version of the app available on Play Store
Android Ver	Minimum required Android Version

Methodology

Our analysis approach is divided into four phases:

1. **Data Extraction and Preparation**
2. **Data Cleansing and Data Mining**
3. **Data Imputation & Manipulation**
4. **Exploratory Data Analysis, Visualization & Dashboarding**

1. Perform Extract-Transform-Load: The dataset collected from the Google Play store is semi structured or unstructured and contains significant superfluous data. It contains 13 different features that can be used for predicting whether an app will be successful or not. It has information such as app name, category, rating, and more. And the other is a list of reviews for each app with the sentiment if that particular content of the review was positive, neutral, or negative.

Data preparation is the process of filtering and transforming raw data prior to processing and analysis. It is an important step and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

2. Data Cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. In this project, we have counted the number of missing values in the Dataset, checked for Outliers in our data, removed columns that are 90% empty and so on.
Data Mining is 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.

3. Data Manipulation refers to the process of adjusting data to make it organized and easier to read. Data manipulation adjusts data by inserting, deleting and modifying data in a database such as to cleanse or map the data. In this project, we filled the null values with appropriate values using aggregate functions such as mean, median or mode to retain most of the data/information of the dataset. We then converted certain attributes into numerical values for further analysis and finally displayed a summary statistics after imputation.

4. In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.

Key Performance Indicator (KPIs)

Key indicators give us a summary of the apps in Google Play Store and its relationship with different metrics;

1. Content Ratings by Reviews
2. Top 10 Apps by Ratings
3. Top earning apps
4. Top 10 Apps by Installs
5. Highest and Lowest rated genres
6. Count of application in each category differentiated by their type
7. Percentage of Review Sentiments

Conclusion

After undergoing these analysis, we concluded that our hypothesis is true. Meaning you can predict the app ratings, however significant pre-processing must be done before you start the analysis. The Play Store apps data has enormous potential to drive app-making businesses to success. 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. From the results and process we have implemented, we can conclude that we have achieved this project's objectives which are analysing the Google Play Store apps and determining trends of the Google Play Store.

Q & A

Q1. What's the source of data?

- The Dataset was taken from iNeuron's Provided Project Description Document. The main source of both the datasets is Kaggle.

Q2. What was the type of data?

- The data was the combination of Numerical and Categorical values.

Q3. What is the complete flow you followed in this project?

- The Refer to slide 4 & 8 for better understanding

Q4. What were the libraries that you used in Python?

- I used Pandas, NumPy and Matplotlib

Q & A

Q5. What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data

Q6. What tool did you use for presenting your analysis?

- I have used Tableau for preparing a Dashboard for this project.

Thank You!