Google Play Store Dataset Analysis

Author : Neha Soni

Objective :

Finding key metrics and factors and show the meaningful relationships between attributes present in the dataset.

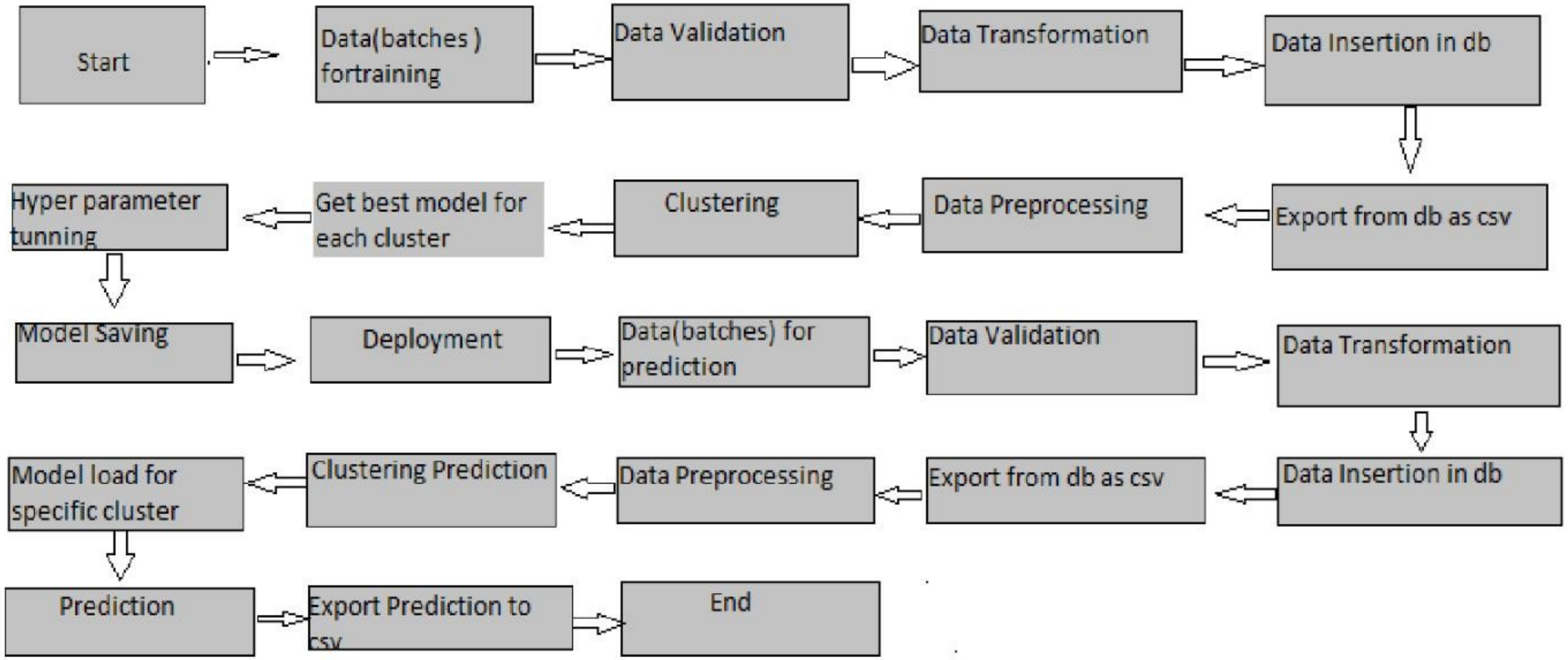
Benefits :

- Most famous app in the category
- Average app size
- Relation between category and reviews
- Install in every category
- Content writing and count
- Top genre and their number of installs
- Distribution of rating
- Ratio of paid and free apps in each category
- Sentiment review count in each category

Data Sharing Agreement :

- Sample file name (GooglePlayStore, GooglePlaystore_user_review)
- Data set has 10841 rows and 31 columns.
- Column names (app category, reviews, size , install, type, price, content writing, genres, last updated, current ver. and android ver.)
- Column datatype (Object, float 64 and int 64)

Architecture



Start

Data(batches) fortraining

Data Validation

Data Transformation

Data Insertion in db

Export from db as csv

Data Preprocessing

Clustering

Get best model for each cluster

Hyper parameter tunning

Model Saving

Deployment

Data(batches) for prediction

Data Validation

Data Transformation

Data Insertion in db

Export from db as csv

Data Preprocessing

Clustering Prediction

Model load for specific cluster

Prediction

Export Prediction to csv

End

Data validation and Data transformation :

- Name validation- Validation of files name as per DSA.
- Number of columns- Validation of number of columns present in the files.
- Name of columns- The name of the columns is validated and the should bethe same is given in the schema file.
- Data type of columns - 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 it is transformed using Python libraries such as Pandas and numpy.
- Null values in columns- if any of the file have all the values as null or missing it is a field or cleaned by python code and its libraries.

Model training:

1. Data export from database:

The accumulated data from database is exported in csv format for model training data

2. Data processing :

- Performing EDA to get inside of data like identify distribution, outliers, trend among data etc.
- Check for null values in the columns. if present impute the null values.
- Encode the categorical values with numerical values.
- Perform Standard Scaler to scale down the values.

Q1 . What is the source of data?

The data for EDA process is provided by the company.

Q2. What is the type of data?

The data was the combination of numerical and categorical values combine in csv format.

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

Refer 5th slide for better understanding.

Q4. After the file validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the achieve folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5. How logs are managed?

We are using different logs as per the steps that we follow in validation and modelling like file validation log, data insertion, model training log, prediction log etc.

Q6. What techniques you were using for data preprocessing ?

- Removal unwanted attributes
- Visualising relation of independent variables with each other and output variables
- Taking and changing distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present
- Converting categorical data into numerical values
- Scaling the data