Analyzing Google App Store

Author – Rishav Sinha

Objective:

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

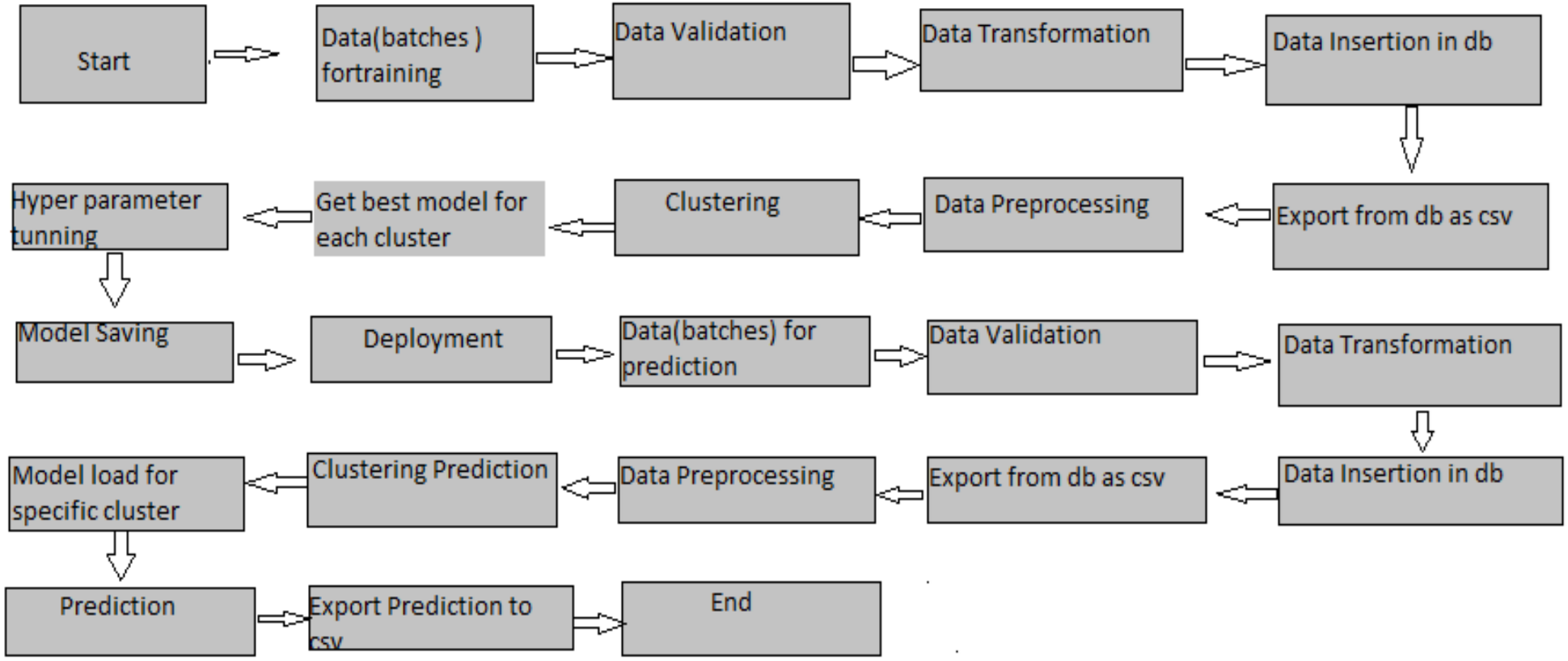
Benefits:

- Most famous app in the category
- Average app size
- Relation between category and reviews
- Installs in every category.
- Content rating 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 (Googleappstore, Googleappstore user review)
- Column names(App, Category, Review, Size, Install, Type, etc.)
- Length of time stamp(6 digits)
- Column data type(Object, Float64, int64)

Architecture



Data Validation and Data Transformation :

- Name Validation - Validation of files name as per the DSA.
- Number of Columns – Validation of number of columns present in the files.
- Name of Columns - The name of the columns is validated and should be the same as given in the schema file.
- Data type of columns - The data type of columns is given in the schema file
- Null values in columns - If any of the columns in a file have all the values as NULL or missing it is filled or cleaned by python codes.

Data Insertion in Database:

- Table creation :- Table name “t motorpv fraud” is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the “Good Data Folder” are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table.

Model Training:

- Data Export from Database:

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

- Data Preprocessing
  - Performing EDA to get insight of data like identifying distribution , outliers ,trend

    among data etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Perform Standard Scalar to scale down the values.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 5th for better Understanding

Q 4) 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.

Q 5) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q 6) 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