GAME APPLICATION SUCCESS PREDICTION

MILESTONE 1 REPORT

Team ID: CS_43

TEAM MEMBERS DATA

NAME	ID
عبدالرحمن سيد جابر أحمد	20201701089
رقيه محمد ابراهيم مصطفي عبده	20201701253
نورهان ايمن محمد عبدالرحمن	20201700939
حنين ابراهيم امام عكاشه	20201700230
مريم احمد اسماعيل محمود	20201700800
هبه طارق كمال عبدالمطلب	20201700959

PREPROCESSING TECHNIQUES

- READING DATA

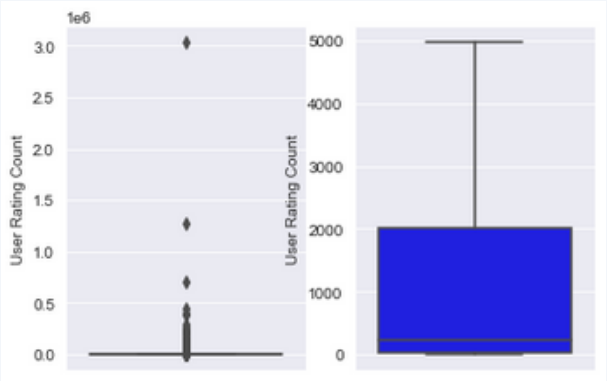
  First we read data and :
  - Drop all null rows.
  - Drop all duplicate rows.
  - Get input variables [**"URL"**, **"ID"**, **"Name"**, **"Subtitle"**, **"Icon URL"**, **"User Rating Count"**, **"Price"**, **"In-app Purchases"**, **"Description"**, **"Developer"**, **"Age Rating"**, **"Languages"**, **"Size"**, **"Primary Genre"**, **"Genres"**, **"Original Release Date"**, **"Current Version Release Date"**]
  - Drop columns from input variables:
    - [**"ID"**] because it contains unique values
  - Get output variable [**"Average User Rating"**]

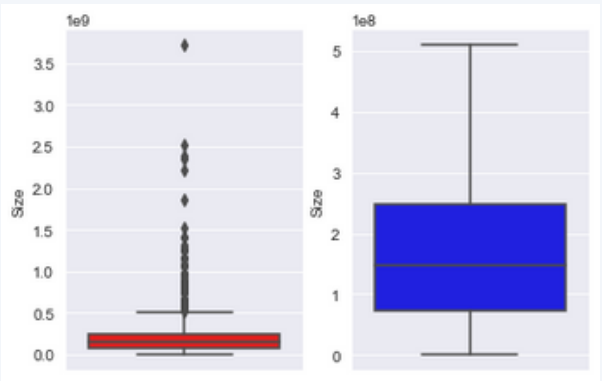
- TRAIN AND TEST SPLIT
  - split our data in 20% in testing and 80% in training
  - shuffle data
  - make random state = 10

- PREPROCESSING IN TRAIN DATA
  - ***Columns Analysis***
    - **In "Age Rating" :**
      - Remove "+" in values.
      - Convert values from **string** to **integer**.
    - **In "Languages" , "Genres" and "Primary Genre"** :
      - Get all **unique values** in each column :
        - **Languages** column has **122 unique Languages**.
        - **Genres** column has **40 unique Genres**.
        - **Primary Genre** column has **11 unique Primary Genre**.
    - **In "Original Release Date" and "Current Version Release Date"** :
      - Convert values from **string** to **date format**.
      - Convert **date format** to **integer**.

PREPROCESSING TECHNIQUES

- ***Columns Nulls***
  - Dealing with columns that contains **null values more than 50%**
    - **In "Subtitle"** :
      - **Drop column** because it contains null values more than 50%.
    - **In "In-app Purchases"** :
      - **Replace** all null values with "0" .
        - Assuming that any cell with null value does **not has any purchases**.
      - **Convert** datatype to **string** to split values in cell and replace them with it's **mean value**.
      - **Convert** datatype to **float** to get final result.
    - **In "Price"** :
      - **Replace** all null values with the most frequently value "0" .
    - **In "Languages"** :
      - **Replace** all null values with the most frequently value "EN" .
- ***Encoding***
  - **In "Price"** :
    - If value **greater than 0** then it will be **1**.
    - If value **less than or equal 0** it will be **no change**.
  - **In "Age Rating"** :
    - Map values by replacing value with its corresponding **integer value** :
      - Replace **4** by **1**.
      - Replace **9** by **2**.
      - Replace **12** by **3**.
      - Replace **17** by **4**.
- ***Outliers Detection & Removal***
  - **In "User Rating Count" and "Size"** :
    - Apply **IQR** to detect and handling outliers.
    - If value **greater than upper bound** it will be **equal upper bound**.
    - If value **less than lower bound** it will be **equal lower bound**.
    - If value **less than upper bound** and **greater than lower bound** it will be **no change**.





PREPROCESSING TECHNIQUES

- ***Dealing With Categories***
  - **In "Primary Genre" , "Genres" and "Languages"** :
    - Apply **one-hot encoding** to the columns.
    - **Replace** original columns by one-hot encoded columns.
    - **Drop** any column contain zeros more than 90%.

- PREPROCESSING IN TEST DATA
  - ***Columns Analysis***
    - **In "Age Rating" :**
      - Remove "+" in values.
      - Convert values from **string** to **integer**.
      - Map values by replacing value with its corresponding integer value :
        - Replace 4 by 1.
        - Replace 9 by 2.
        - Replace 12 by 3.
        - Replace 17 by 4.
    - **In "Original Release Date" and "Current Version Release Date"** :
      - Convert values from **string** to **date format**.
      - Convert **date format** to **integer**.
  - ***Columns Nulls***
    - Dealing with columns that contains null values more than 50%
      - **In "Subtitle"** :
        - Drop column because it contains null values more than 50%.
      - **In "In-app Purchases"** :
        - **Replace** all null values with "0" .
        - Assuming that any cell with null value does **not has any purchases**.
        - **Convert** datatype to **string** to split values in cell and replace them with it's **mean value**.
        - **Convert** datatype to **float** to get final result.
      - **In "Price"** :
        - Replace all null values with the most frequently value "0" .
      - **In "Languages"** :
        - Replace all null values with the most frequently value "EN" .

PREPROCESSING TECHNIQUES

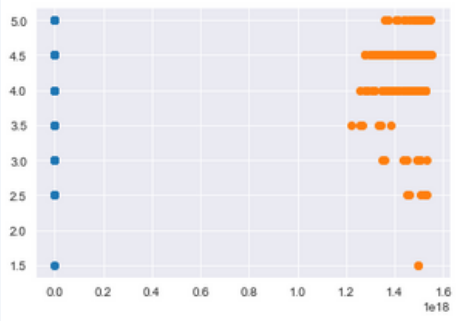
- *FEATURE TRANSFORMATION*
  - **In "URL"** :
    - Extract country name and rename column to "Country".
    - Drop "Country" column because there is **no unique values**.
  - **In "Icon URL"**
    - Extract colors and rename column to "Color".
  - **In "Name"** :
    - Tokenize name.
    - Apply part of speech tagging.
    - Filter out stop words and check if the word is a noun or verb and calculate it's frequency.
    - Get the 50 most frequent words.
    - Replace each word in the "Name" column with 1 if it matches one of the 50 most frequent words.
    - Rename column to "frequent words in Name".
  - **In "Description"** :
    - Convert text to lowercase.
    - Remove URLs, punctuations, stop words and special characters.
    - Remove frequently words that occur more than 2000 times (most frequently 14 word) in document.
    - Remove rare words in documents.
    - Extract game difficulty from "Description" column in "Game Difficulty" column.

- *CORRELATIONS*
  - Get correlation between input variables and output variable
  - Correlation value between two features:
    - "Size" and "Average User Rating" :
      - Correlation = -0.04
    - "Age Rating" and "Average User Rating" :
      - Correlation = 0.00
    - "In-app Purchases" and "Average User Rating" :
      - Correlation = -0.01
    - "Price" and "Average User Rating" :
      - Correlation = -0.08

PREPROCESSING TECHNIQUES

- FEATURE SELECTION
  - After Compute the **correlation** between **input variables** and **output variable** we get top features according to "Average User Rating" value if it greater than **0.1**
  - We get "User Rating Count" and "Original Release Date" as a **top features**

- LINEAR REGRESSION MODEL
  - ***TRAIN MODELS***
    - Accuracy Test = 0.011568518878795842
    - Mean Square Error Test = 0.29740971410321826
  - ***EVALUATE MODELS***
    - Accuracy Test = 0.014374589812016114
    - Mean Square Error Test = 0.3382383496767792
  - ***Train Models with the whole data set***
    - Accuracy Test = 0.01222150503429198
    - Mean Square Error Test = 0.3055913878508464



- Random Forest Model
  - *Train Models*
    - *Accuracy Test = 0.851311921957106*
    - *Mean Square Error Test = 0.851311921957106*

  - *Train Models with the whole data set*
    - *Accuracy Test = 0.08940913200299794*
    - *Mean Square Error Test = 0.3124886485661008*

- Lasso Regression Model
  - *Train Models*
    - *Accuracy Test = 0.011568518878795842*
    - *Mean Square Error Test = 0.29740971410321826*

  - *Train Models with the whole data set*
    - *Accuracy Test = 0.014374589812016114*
    - *Mean Square Error Test = 0.3382383496767792*

PREPROCESSING TECHNIQUES

- Ridge Regression Model
  - *Train Models*
    - *Accuracy Test = 0.0555160741402525*
    - *Mean Square Error Test = 0.2841863090463304 7*
  - *Train Models with the whole data set*
    - *Accuracy Test = -3.460474705487327*
    - *Mean Square Error Test = 1.53070688677892*