Machine Learning

Milestone 2 – Regression

❖ Preprocessing Techniques:

- Remove empty rows.
- Handle missing values.
  - Replace Categorical features with the MOD.
  - Replace Numerical Features with Mean value.
- Encoding Categorical Data.
  - Use label encoder to convert categorical features into numbers.
  - Use OneHotEncoder to create dummy variables for each of the categorical features.
- Apply feature selection depending on the relations between features that are shown in Correlation Heat Map.
- Split dataset into training set and testing set.
- Apply Feature Scaling to the features.

❖ Regression Techniques:

- Linear Regression.
  - Training Set:
    - MSE: 0.3468.
    - Training Time: 0.026s.
  - Testing Set:
    - MSE: 0.3610.
    - Training Time: 0.021s
- Polynomial Regression. (Degree = 4)
  - Training Set:
    - MSE: 0.48.
    - Training Time: 119.07s.
  - Testing Set:
    - MSE: 0.67.
    - Training Time: 43.8s.

❖ Used and Discard Features:

Here we let it generic as letting the user to select the suitable features according to the correlation but also, we recommend Removing the worst 5 correlated features and these features are selected or discarded based on this recommendation.

- Used Features:
  - Size bytes.
  - Rating count total.
  - Rating count version.
  - User rating version.
  - Supported devices number.
  - Languages number.
  - Vpp lic.
  - Content Rating.
  - Prime Genre.
- Discarded Features:
  - ID.
  - Track name.
  - Currency.
  - Version.
  - Price

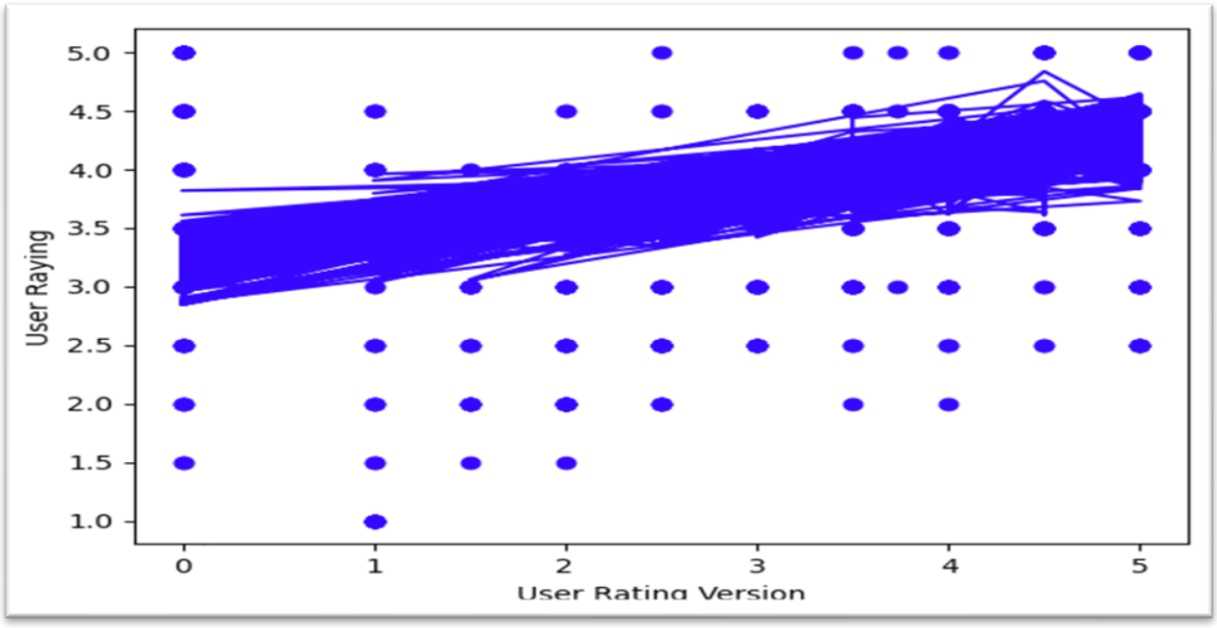
❖ Sizes:

- Training set: 66% of dataset.
- Testing set: 33% of dataset.

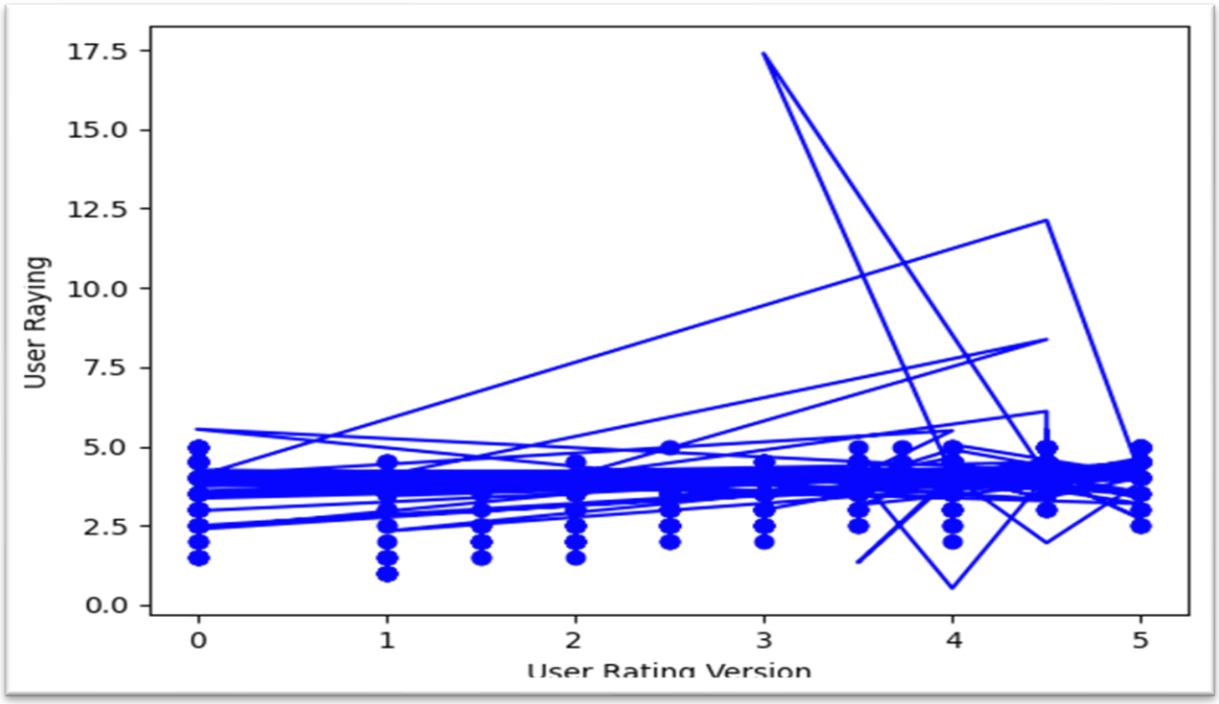
❖ Conclusion:

- Data features had weak correlations to each other, and the most affecting feature was “user_rating_ver”.

❖ Screenshots



Linear Regression Model Plotting



Polynomial Regression Model Plotting