Model Exploration and Model Selection

Since this is a prediction problem where we need to predict the rating of an application based on various factors (or predictors), a regression model would be an ideal approach.

We plan to split the data in the ratio of ***80:20*** for training and testing the model respectively. Some of the regression models that we plan to use are:

1) **Linear Regression:** Our baseline model would be linear regression. LR compares input and output variables based on linear/labeled data. Since our data are mostly numeric (we can convert categories into binary using one-hot encoding), LR models fit well. We can find the association between two variables using the correlation coefficient the value of which ranges from -1 to 1.

   Some advantages of the LR model are:

   - The model is easy to use and understand, it is of low complexity
   - Training time and efficiency is great
   - If the data is truly linear, then fitting the data is perfect

   Some disadvantages of the LR model are:

   - It assumes that all predictors are linear
   - It usually tends to under-fit the data
   - If outliers exist, it's performance dips

2) **KNN Regressor:** The K-Nearest-Neighbor algorithm predicts the value based on how close it is to the nearest neighbor values. This distance can be measured using the following methods:
   a. Euclidean
   b. Manhattan
   c. Statistical

   The 'K' is the number of neighbors the training model considers. KNN generally is used for classification problems but adding a regressor to it makes it predict numerical values. The model then predicts the value based on the mean of predictor neighbor values.

   Advantages of using the KNN Regressor:

   - Training time is very less
   - It can be used for both classification and regression
   - Does not require parametric assumptions

   Disadvantages of using the KNN Regressor:

   - If the number of predictor variables is large, it tends to under-perform
   - Prediction will be skewed if range of one predictor dominates the other when using raw data. Standardization is recommended.

3) **Random Forest Regressor:** Large number of decision trees work as estimators that predict values based on rules and parameters. The outcome of all these trees is then combined and averaged to make an accurate prediction.

   Advantages of RF Regressor:

   - Handles missing data automatically
   - Scaling data is not needed, it is a rule-based approach
   - Since many estimator trees make the prediction, it increases accuracy and reduces variance

   Disadvantages of RF Regressor:

   - Training time is generally high
   - Tends to over-fit the data

4) **Support Vector Machine Regressor:** It sorts the data into two categories that it most likely belongs to. The SVM model then predicts which class the new data point for validation/testing belongs to based on predictors. The idea is to fit the error within a certain threshold – this approximates the best value without a given margin.

   Advantages of SVM Regressor:

   - Can be used for both classification and regression
   - Robust to outliers
   - Data with high dimensions can be fitted

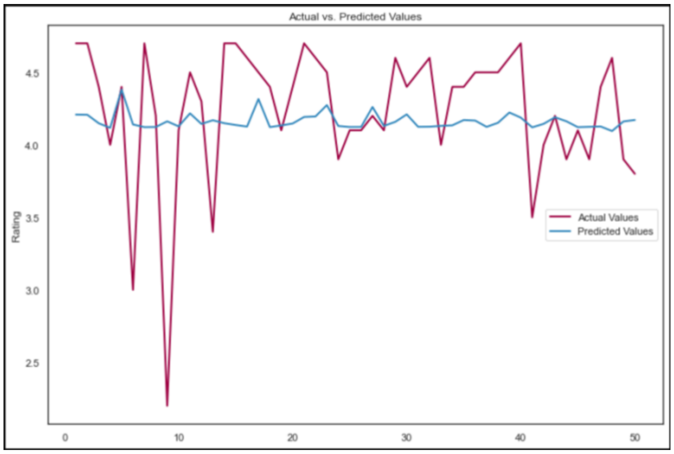
   Disadvantages of SVM Regressor:

   - Training time is high
   - Data needs to be scaled
   - Tends to over-fit the data

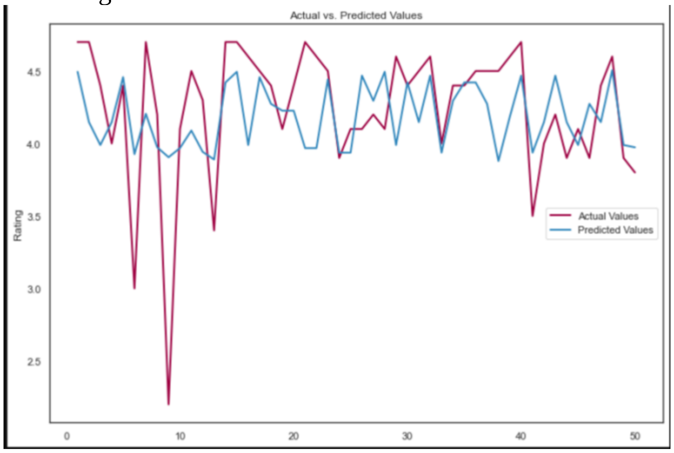
We will implement the above four models and find out which model has the best RMSE, performance, and score. We can then fine-tune the model as required by varying the parameters.

We ran the above models for our training data based on default parameters and initial observations. We will be fine-tuning the parameters and constants to ensure we get the best scoring model from the lot in the next phase. Some of the plots for actual values and predicted values are as follows for our baseline models:

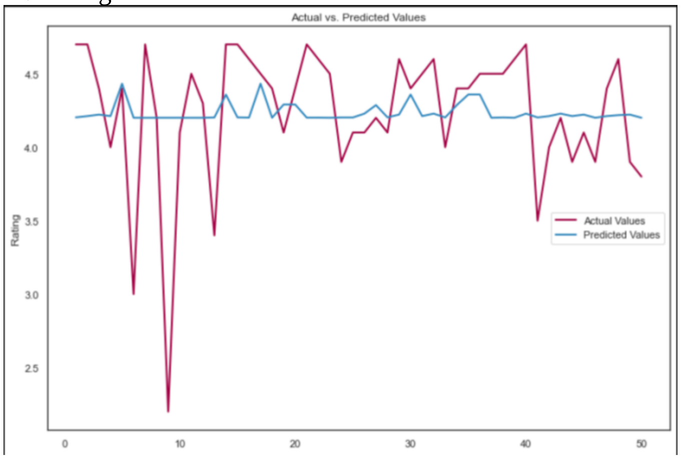
1) Linear Regression



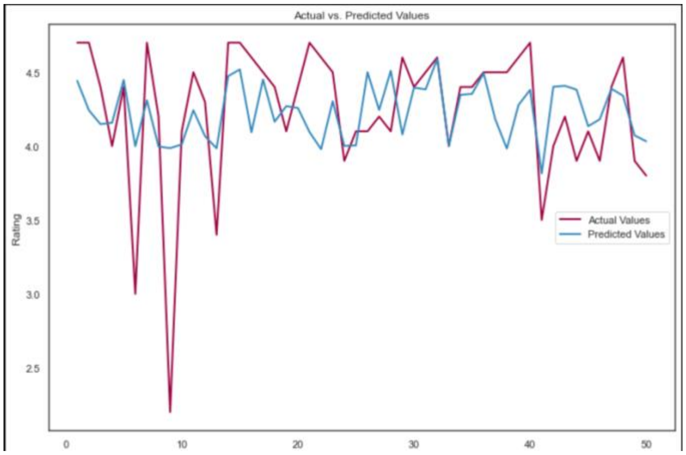
2) KNN Regressor



3) SVM Regressor



4) Random Forest Regressor


Actual vs. Predicted Values
Rating
4.5
4.0
3.5
3.0
2.5
0
10
20
30
40
50
Actual Values
Predicted Values