Applications of Data Science

Project 1

Done by:

101068721

Index

List of Figures

1. Introduction

Used cars are highly sought after. In the UK the used car market grew by 5% in 2023 alone (Lewis-Stempel, 2024). The use of machine learning models will aid in efficiency and speed. There are several supervised machine learning algorithms which can be used to predict the prices of used cars. The aim of this report is to evaluate four ML algorithms used to predict the car prices.

1.1 Machine Learning Algorithms

Four machine learning algorithms are compared subsequently in this report using a preprocessed dataset. The four machine learning algorithms used are:

i. Linear Regression
ii. KNeighbor Regression
Iii. Random forest Regression
Iv. XGBoost Regression

1.1.1. Linear Regression:

The first model used to analyse the dataset is Linear regression. Linear regression models assume a linear relationship between predictor variables and target variables (Swaminathan, 2018). It was chosen as the first model to understand the underlying relationship between the variables. Since it is a very simple model, it is easy to make inferences based on the data (Black, 2023).

1.1.2. KNN Regression:

After working on the linear regression model and plotting the residual Vs fitted graph, it is evident that the variables did not have a linear relationship, so the KNN regression algorithm was chosen. Rather than going directly to a complex model, a simple model was used so that the patterns and trends in the dataset can be studied. KNN regression is also a simple algorithm but, unlike linear regression, it does not make any assumptions about underlying patterns in data and hence it works well with nonlinear data too (Timbers, 2023).

1.1.3. Random Forest Regression:

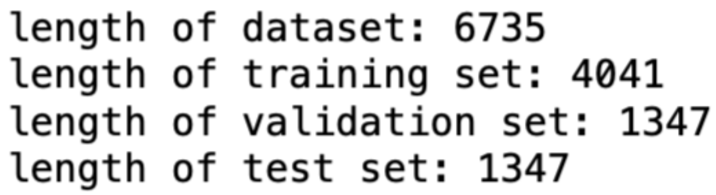
The value of mean square error and residual vs fitted graph plotted for the KNN regression model suggested the possibility of few outliers in the dataset even though the outliers were removed both manually and using local outlier factor. Random forest regression model was used to analyse the data because it is relatively not sensitive to noise and outliers (Breiman, 2001). Even though KNN regressor model worked better than linear regression on this dataset, it still was not performing well on unseen data. This also suggests the need for a more complex model. Random forest uses ensemble learning by combining predictions of different decision trees to make predictions, this makes the predictions much more accurate (Built In, 2022). Random regression was chosen next to have a complex model that understands all the underlying relationships which is also not sensitive to outliers.

1.1.4. XGBoost Regression:

XGBoost also uses ensemble learning but random forest regressor uses bagging while XGBoost uses boosting techniques. In boosting, the model corrects the errors in previous trees which results in better prediction value (Gupta, 2021). The randomisation used in XGBoost also helps in reducing chances of overfitting (Bentéjac, Csörgő and Martínez-Muñoz, 2019). This model was used to improve the r2 score and reduce MSE and MAE by boosting.

2. Results of training, fine tuning and evaluation

The dataset is divided into 80% train set and 20% test set. This train set is further divided into 80% train set and 20% validation set.



```
length of dataset: 6735
length of training set: 4041
length of validation set: 1347
length of test set: 1347
```

Figure 1: Length of dataset after preprocessing, train set, validation set and test set

Each model was trained on the train set and then evaluated using the validation set. A learning curve is made for r2 score based on train and validation set. The model was then hyperparameter tuned using grid search. Using the best parameters from the grid search another model is built and the test data is evaluated as unseen data. The validation set was used for hyperparameter tuning to avoid any possible data leaks into the test set (Sivasai Yadav Mudugandla, 2020).

Mean squared error (MSE), mean absolute error(MAE) and r^2values are calculated on the test and train data to evaluate the model. Cross validation r^2 value and train r^2 value is also checked to ensure there is no overfitting.

After the evaluation of the unseen data, residual vs fitted graph and actual vs predicted value graphs are plotted for each model using test set and predicted values.

Further in the report each model and its training, hyperparameter tuning and evaluation will be discussed in detail with graphs.

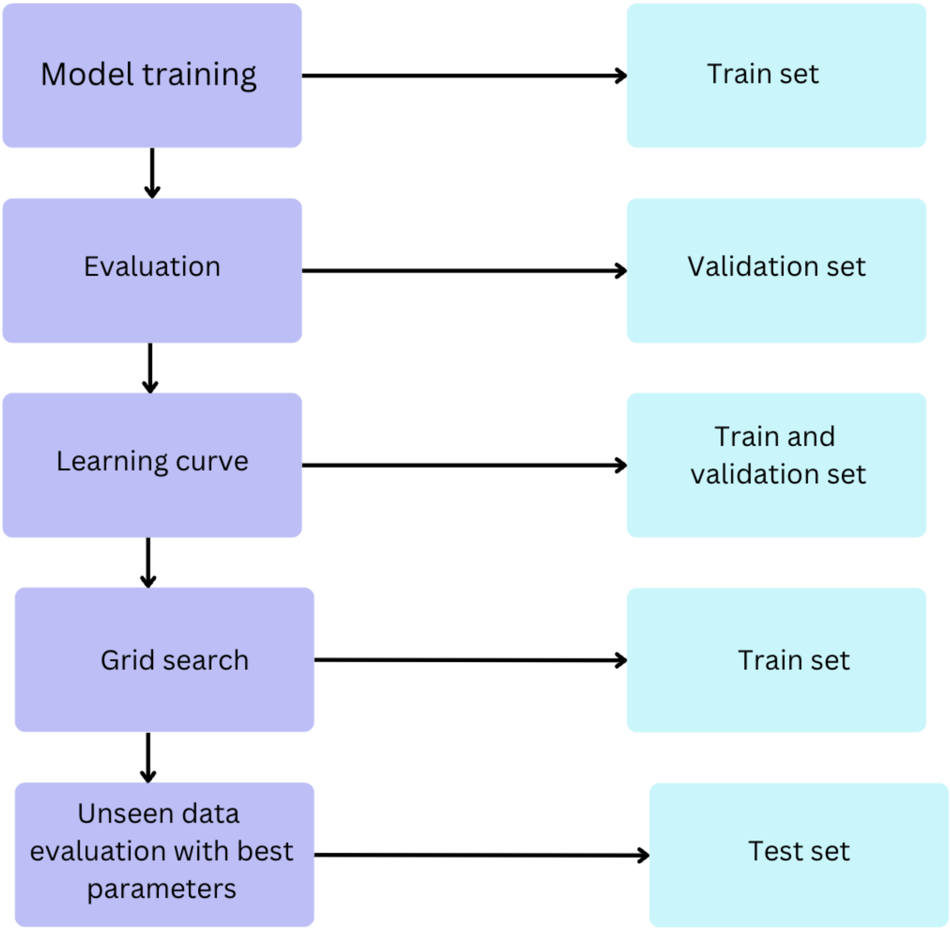


Figure 2: Training, tuning and evaluation process followed for each model

2.1 Linear Regression

The linear regression model was not performing very well with the validation set. The mean squared error and mean absolute error was quite high and the r^2 score was only 0.58. This is because linear regression models assume linear relationships in the data (Swaminathan, 2018).

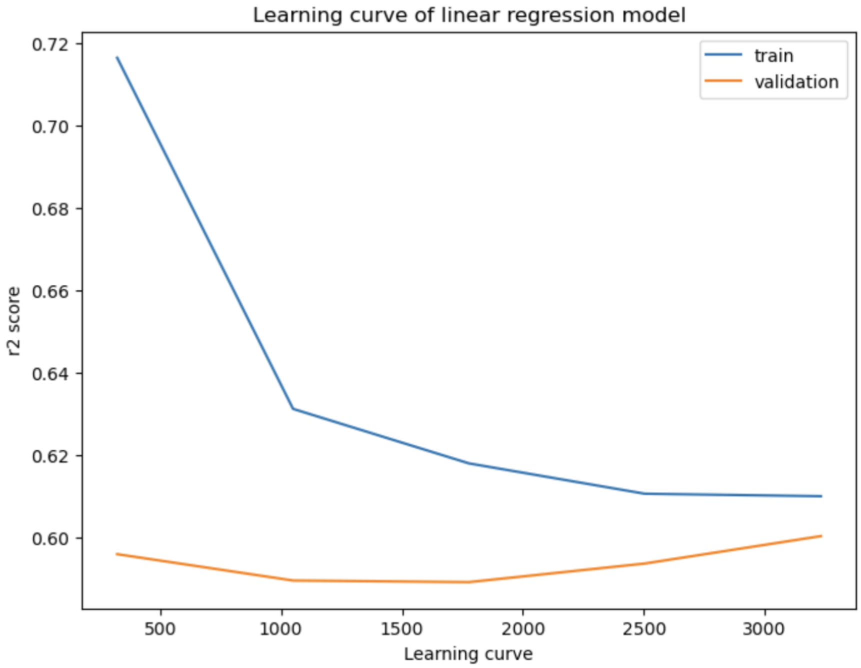


Figure 3: Learning curve of linear regression model

Checking the learning curve above, we can see that the r^2 score of the training data decreases drastically as more data is given. The variance was high in the beginning, but the variance reduced as the model was given more data. As the model is unable to learn the underlying relationships, this suggests the model is too simple and has high bias (Nischal M, 2019).

The model was hyperparameter tuned using GridSearchCV but the values did not improve for MSE, MAE or r^2. The model was tuned on fit intercept and n jobs parameters. The test data also had similar scores.

The predicted vs actual graph shown below, the closer the values are to the red line, the closer they are to the actual values. The graph shows that as the values became bigger, they moved further away from the line, this might be the reason for high MSE as MSE score is sensitive to large value errors in prediction or outliers (Jadon, Patil and Jadon, n.d.).

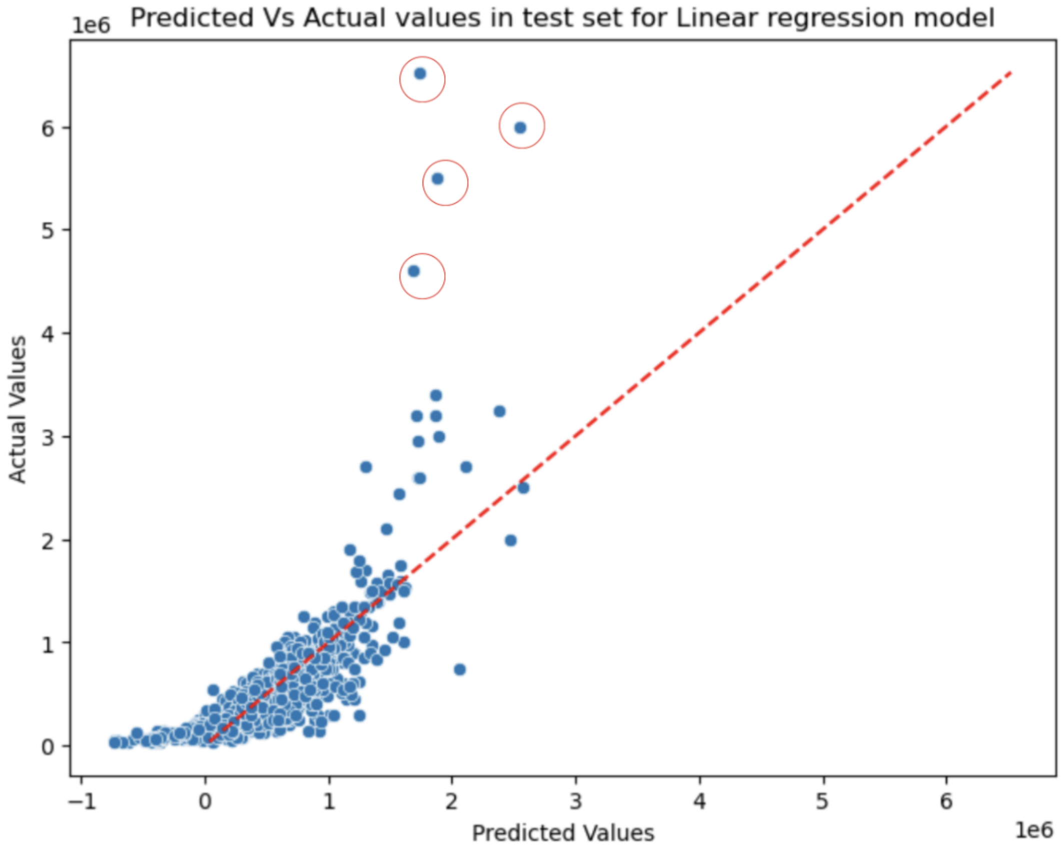


Figure 4: Actual vs Predicted graph of linear regression model (Test set)

In the residual vs fitted graph for the model, the values tend to form a “u curve” pattern, this shows that there is non-linearity in the data (Ahmed, 2020). Ideally, the values should be scattered around the y=0 line. The closer the points to the line, the closer they are to the actual value.

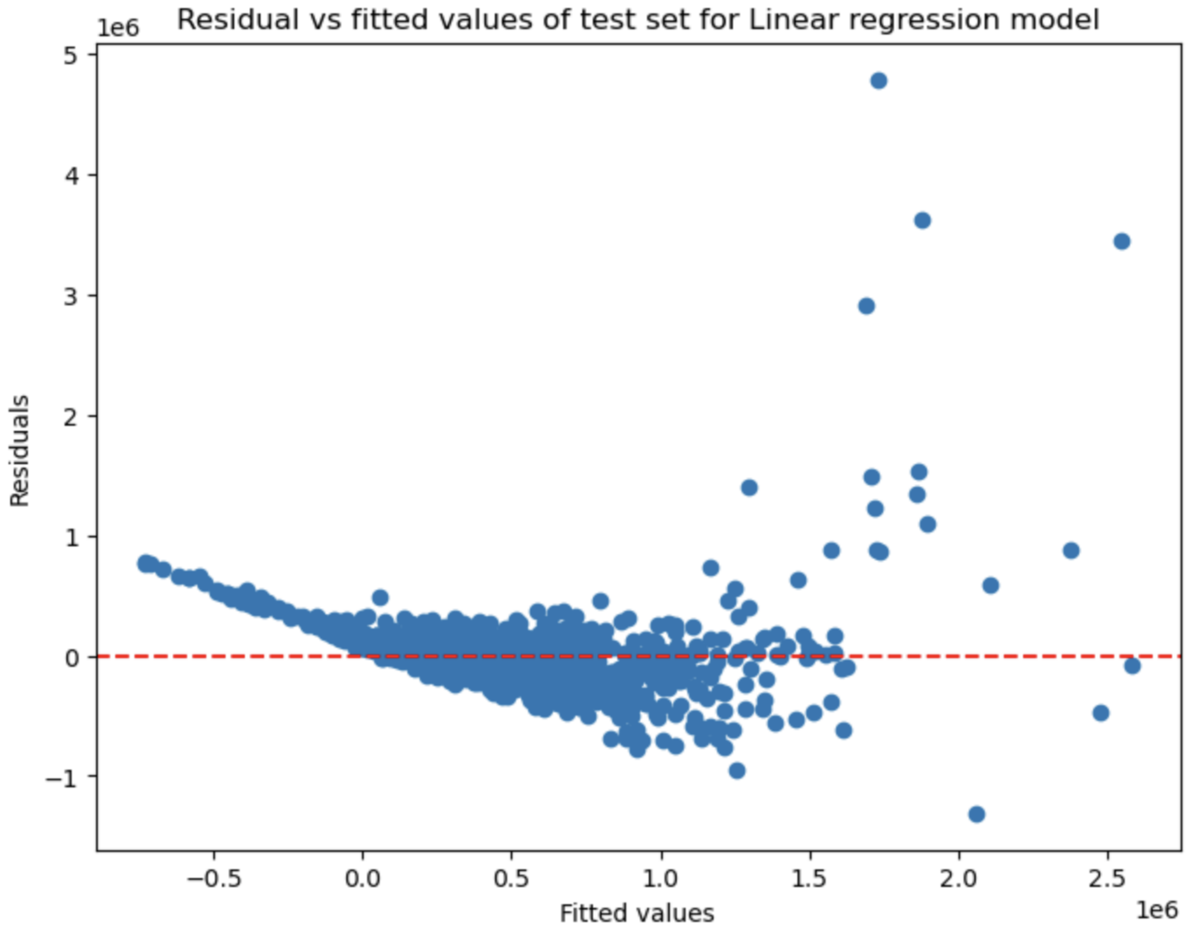


Figure 5: Residual vs fitted values for Linear regression model

2.2 KNeighbor Regression

KNN regression model performed better compared to the linear regression. The MAE and MSE score is still high but has reduced compared to linear regression model. The r^2 value has improved to 0.75.

From the learning curve provided below, we can see the train and test data seems to increase as more data is provided and this suggests that the bias is not high, and the model is learning underlying patterns. In KNN the bias increases with higher k values (T. Dorigo et al., 2022). The k value used in this model is 5. The model's r^2 scores keep increasing as more data is provided in both the train and validation set.

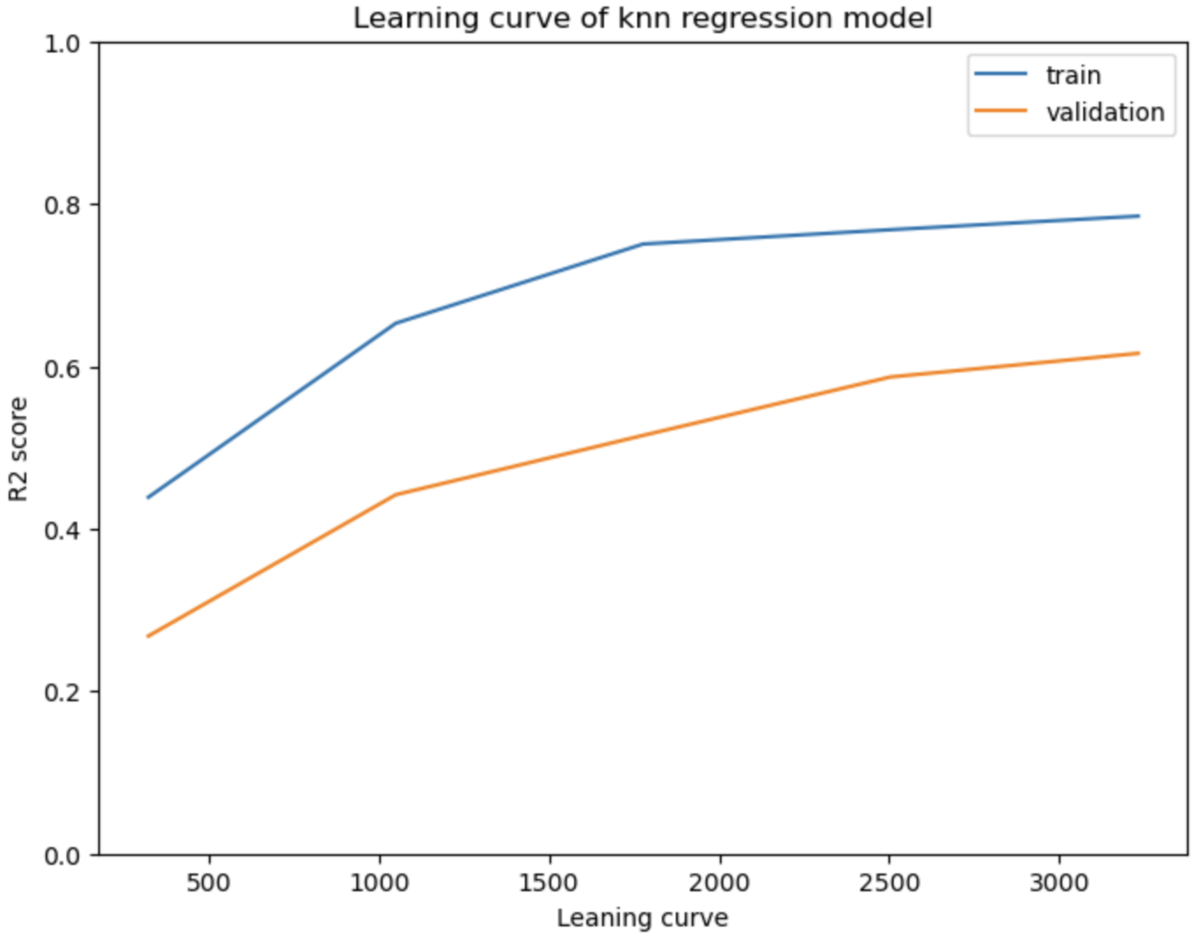


Figure 6: Learning curve of knn regression model on test data

The parameter used for hyperparameter tuning this model is the n_neighbours value from a range of 0 to 100. The best parameter suggested by grid search was 3. The r^2 validation set improved to 0.76 from 0.75. The model also performed satisfactorily when used on unseen data (test set). r^2 value in the test set was 0.79 and the cross validation r^2 score was 0.65.

The actual vs predicted graph for knn shows that most of the values have moved closer to the line, indicating that the difference between the actual values and predicted values have decreased from the previous model. The graph shows a few outliers with higher selling price which seems to be the reason for the high MSE. Both high variance and high bias would result in higher MSE (Maksym Zavershynskyi, 2017). Variance of the outliers marked in the graph will have a huge impact on the MSE value. The MSE of the test set was around 4.9e10 even though r^2 is 0.79.

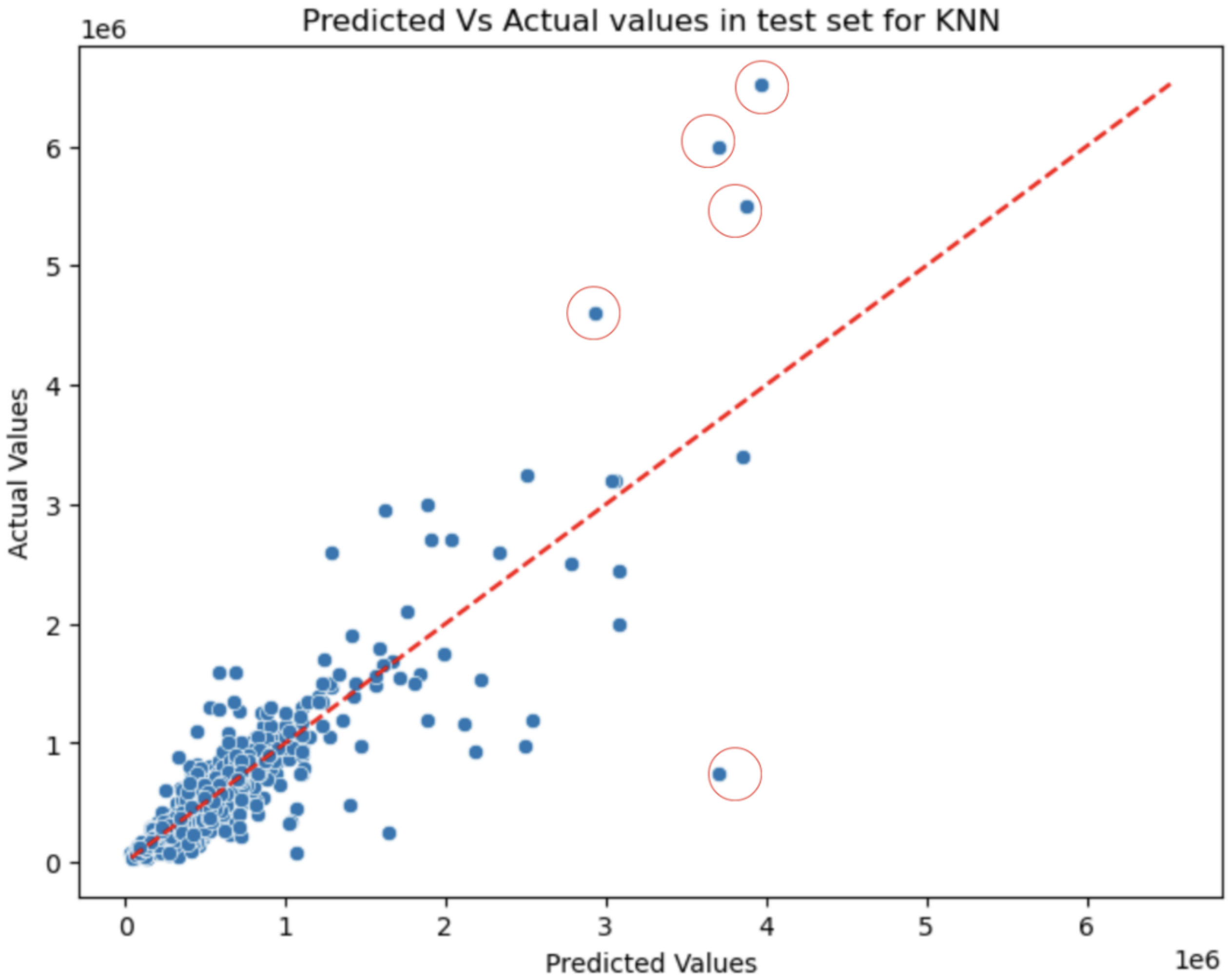


Figure 7: Predicted Vs actual values for KNN regression (test set)

Unlike the linear regression model, the KNN regression model does not seem to form a "u curve" pattern but the points are scattered around the y=0 line in the residual vs fitted graph. This shows that the model is fitting well to the dataset (KIm, 2023).

There seems to be some values which are further away from the y=0 graph. This indicates the difference of predicted value has a big difference from the actual value. This can be seen from the below graph (Frost, 2017). The variance among the values does not seem to be uniform, which suggests the data shows heteroscedasticity. Some of the data points seem to be outliers, this can be checked by using a complex model which can intricate the underlying relationship better.

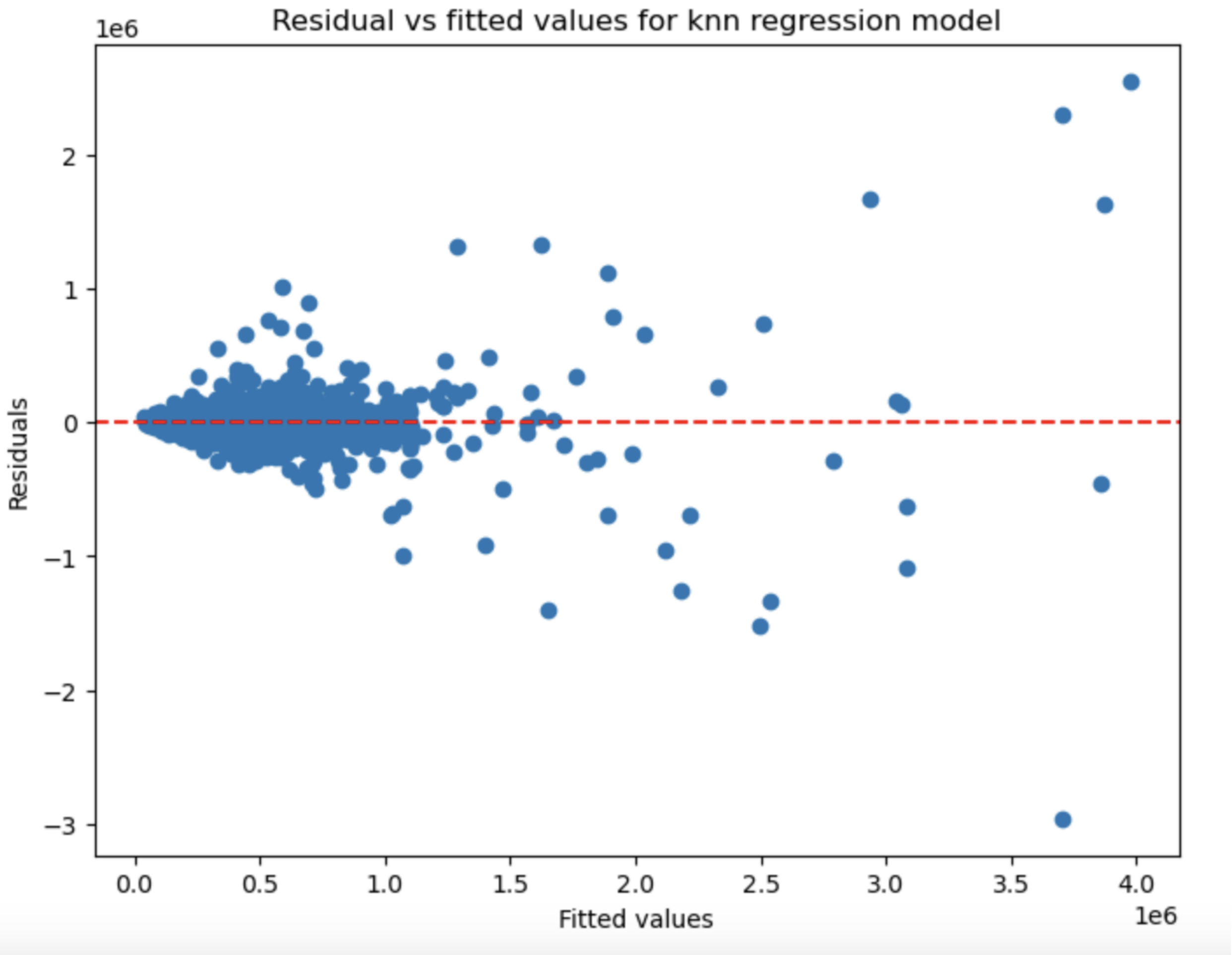


Figure 8: Residual vs fitted values for knn regression model(test set)

2.3 Random Forest Regression

Random forest regressor model has performed well in the validation set even before hyperparameter tuning. The r^2 value on the validation set was 0.86. Even though the MSE improved significantly, the value is on the higher end.

The learning curve shows that the model performs extremely well on the training set. r^2 value is close to 1 for the train set suggests that the model might be overfitting (Frost, 2017) in the beginning but as more data was given the variance reduced between train and validation data. This also shows that with more data, the overfitting issues can be removed for this model. There was good amount of increase in r^2 score until it reached 1000 data points in the validation set, then the rate at which the model performance improved reduced. The growth seems to be reaching a plateau. It can be inferred that even if more data is given the improvement of r2 score will be minimal.

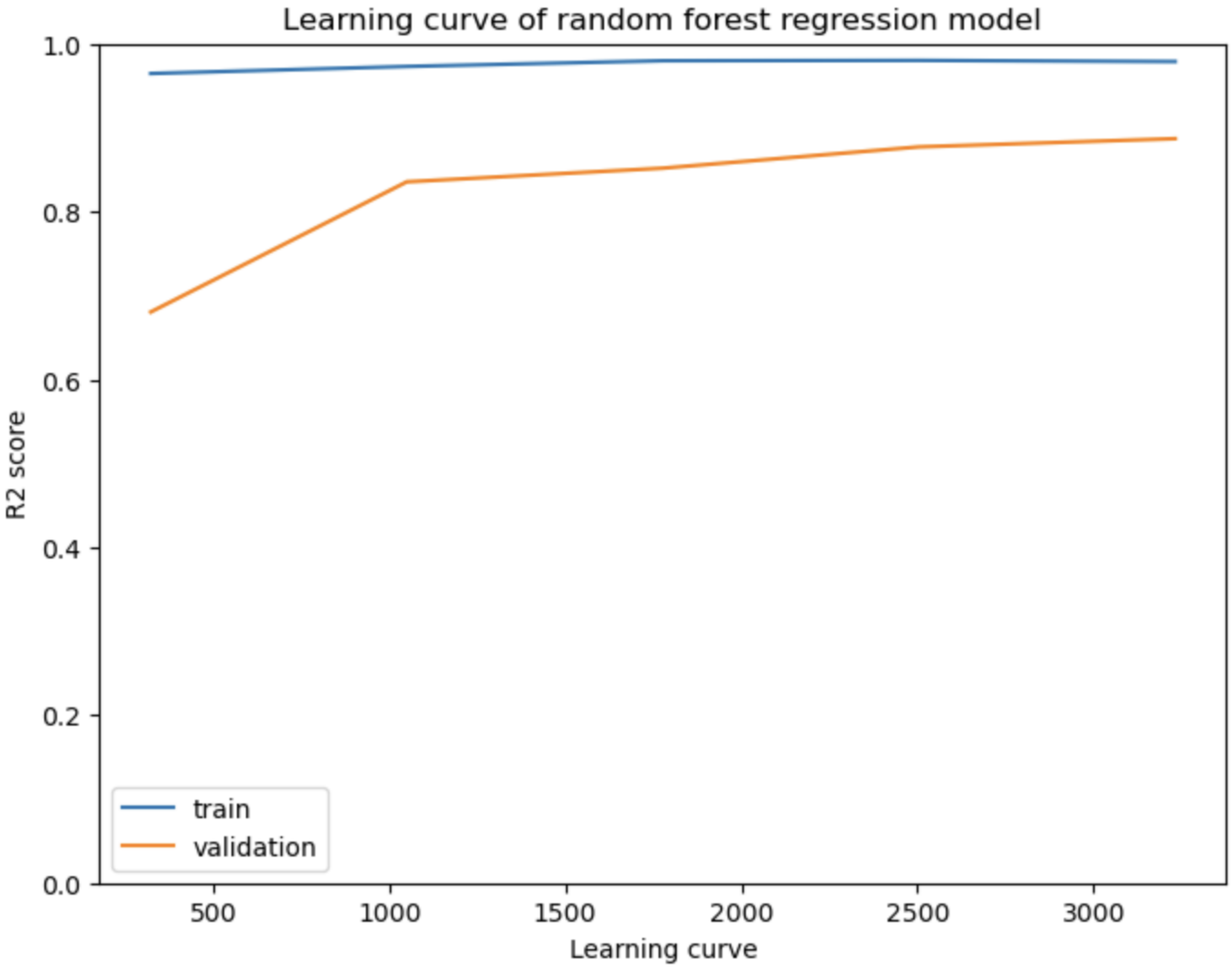


Figure 9: Learning curve of random forest regression model (test set)

The model was hyperparameter tuned on n_estimators, max_depth, min_samples_split, min_samples_split, min_samples_leaf and bootstrap. The r^2 value on the validation test increased from 0.86 to 0.88 after the grid search.

The model also performed well on the test set with 0.89 r^2 score. The validation score on the train set was also 0.88. The MSE score remained elevated with a value of approximately 2.2e10.

The below graph shows the actual vs predicted values for random forest. It shows that apart from a few outliers, the points are close to the red line, indicating that the predicted values are very close to the actual values. The distance from the center line has significantly reduced compared to the KNN regression model.

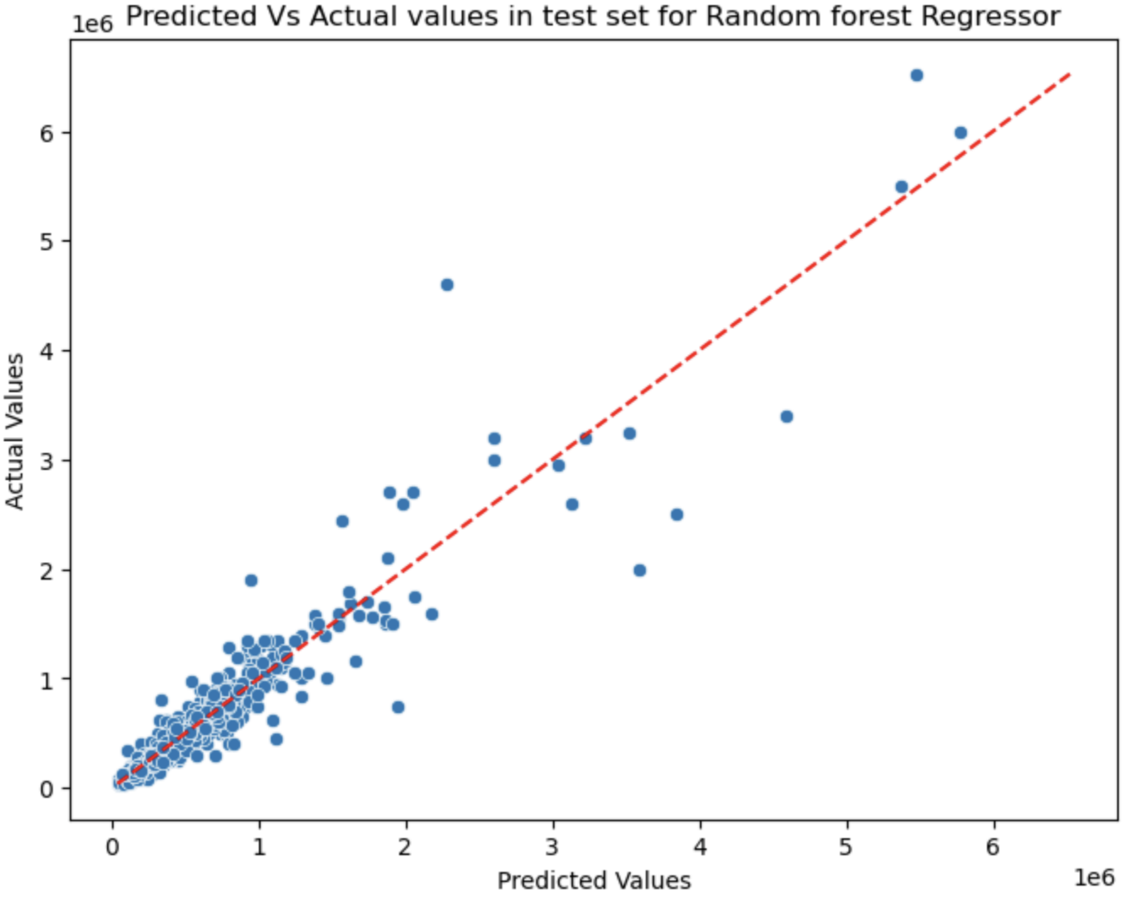


Figure 10: Predicted vs actual values for random forest regressor (test set)

The residual vs fitted graph of the random forest regressor provided below shows that the data is scattered around the y=0 line. We can see from the graph below that the variance of the residuals is not constant. There are some points on the graph which are quite far away from the y=0 line, these points can be considered as outliers. This graph also shows heteroscedasticity like the KNeighbour regression residual vs fitted graph.

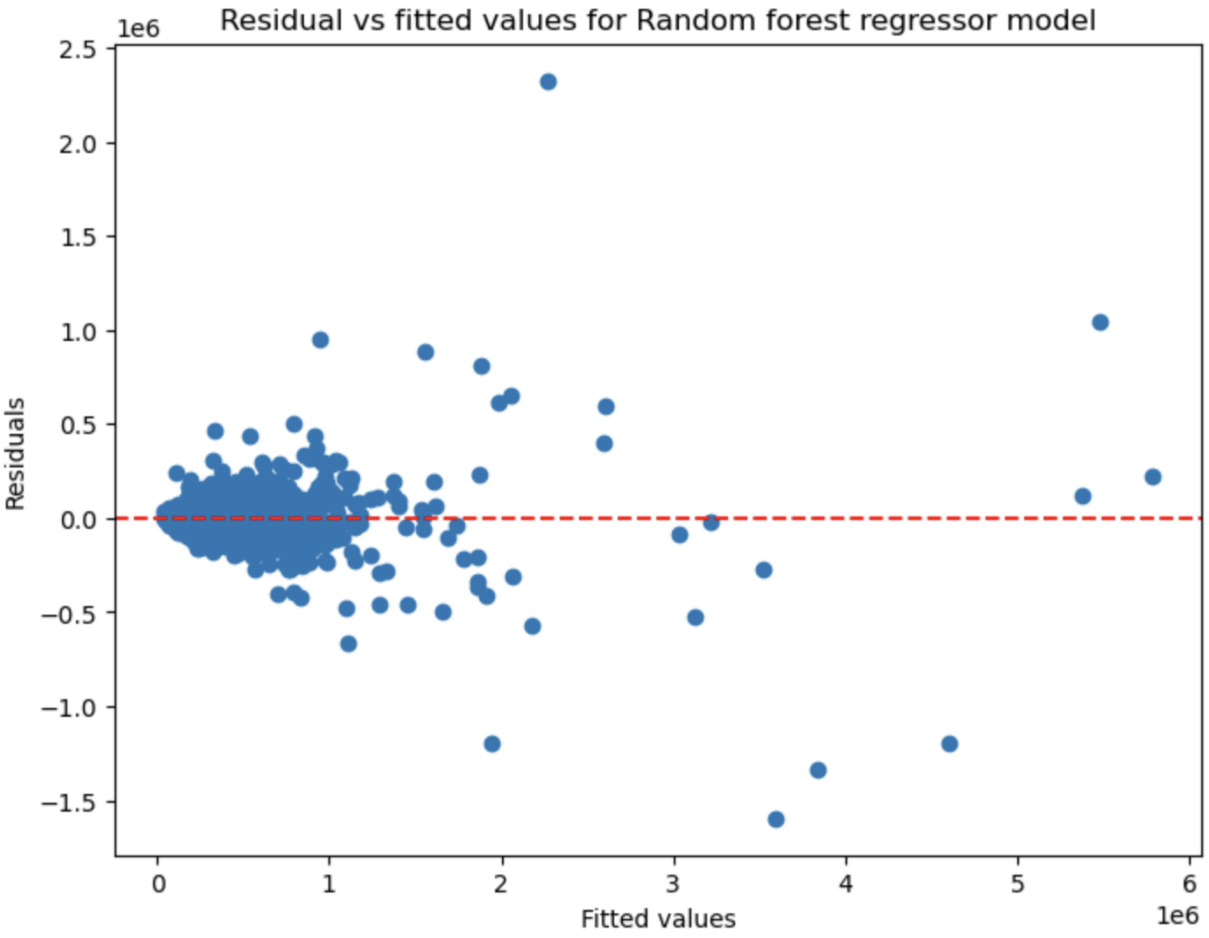


Figure 11: Residuals vs fitted graph for random forest regression (test set)

2.4 XGBoost Regression

While training the data Dmatrix was used to wrap around the data. Dmatrix optimizes memory usage and speed (Goyal, 2020). The r^2 score on the validation score was 0.88.

The learning curve shows that the model performed well in the train set. The model was overfitting when there was less amount of data samples. The model’s performance got stagnant at around 2500 data samples, this suggests that even if we provide more data to this model, the performance improvement rate will be very slow.

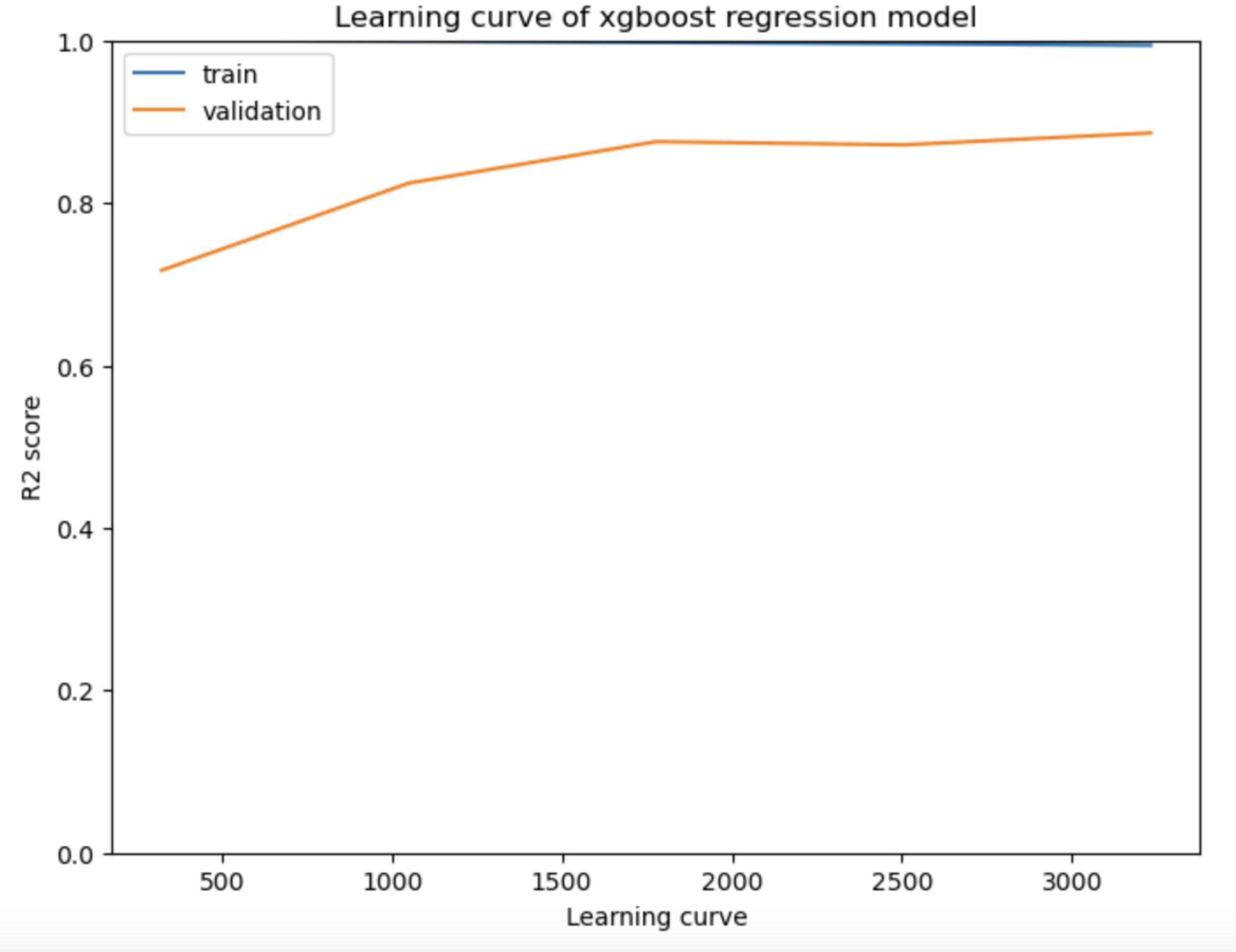


Figure 12: Learning curve of XGBoost regression model (test set)

After the model tuning, the r2 score on the validation set was 0.89, before the tuning it was 0.86. The r^2 score increased slightly after the model tuning. The parameters used were objective, eta, max_depth, min_child_weight, learning_rate, gamma. The model built on these best parameters was evaluated on the test set (unseen data) and r^2 score was 0.93. The MSE score decreased to approximately 1.5e10. This model provides the highest r^2 score and least MSE and MAE scores in unseen test set compared to all the other four models.

Compared to all other models evaluated before, this model has the least number of points which have deviated from the red line. This shows that most of the predicted values are close to the actual values.

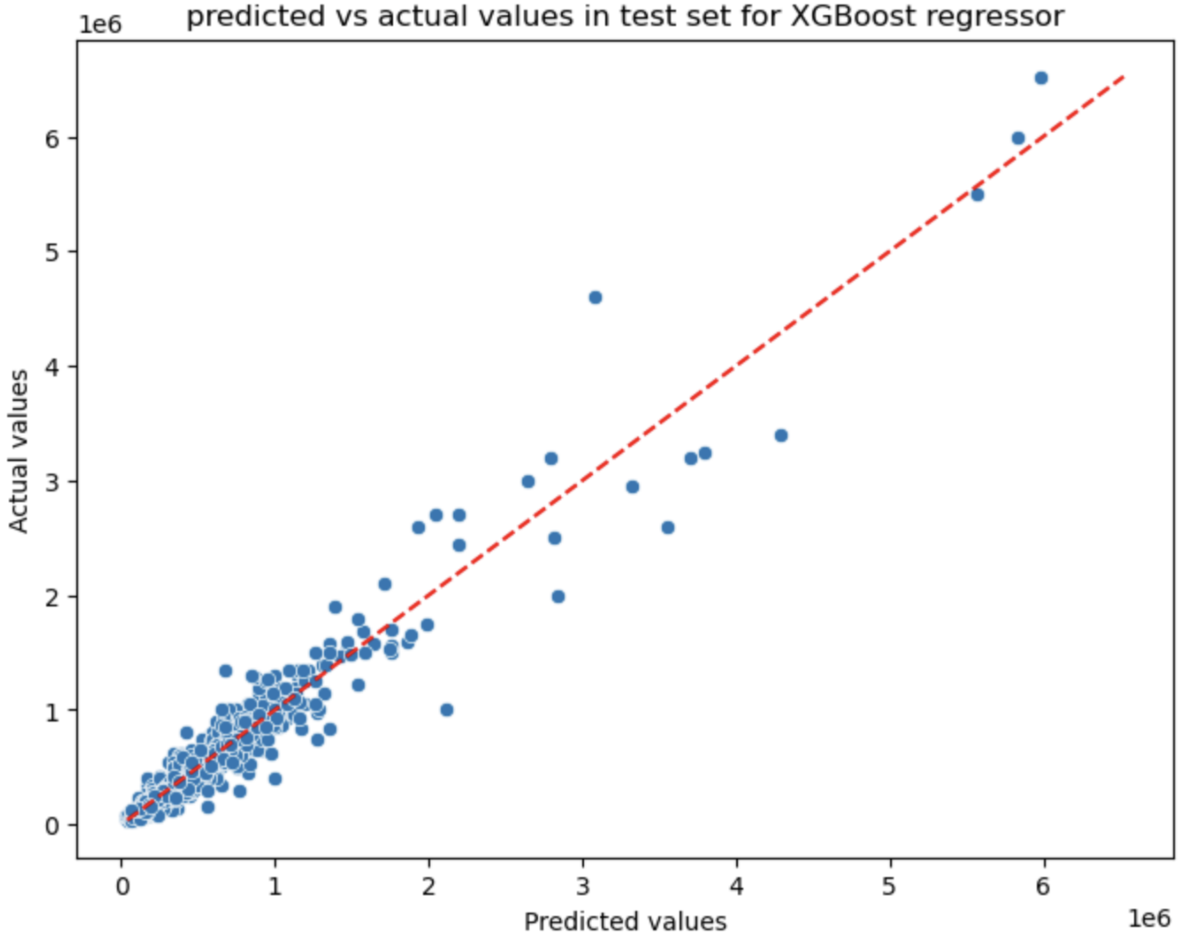
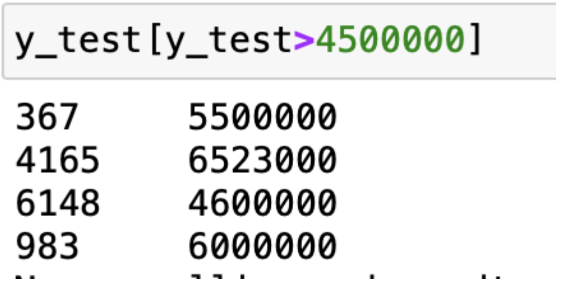


Figure 13: The predicted values vs actual values graph for XGBoost regression (test set)

2.5 Testing impact of large error on MSE

We can see that there is a value between 4000000 and 5000000 which is the point which is furthest away from the red line. This shows this value has deviated from the actual value the most compared to other points. When we check the values in y_test which are greater than 4500000, we can see the row with the value which is most deviated from the actual value in the test set is at index 6148.



```
y_test[y_test>4500000]

367     5500000
4165    6523000
6148    4600000
983     6000000
```

Figure 14: All the values in y_test which is greater than 4500000

The predicted price for selling_price of 4600000 is 3076923. To understand the impact of MSE, the predicted value of this specific row was changed to the true value to see the MSE score. The MSE score was reduced from approximately 1.5e10 to approximately 1.3e10. This shows that even if there is one large error, it impacts MSE a lot as it is sensitive to large errors (Padhma M, 2021).

3. Comparing the models

3.1 r^2 value

Out of all the four models, XGBoost has the highest r^2 score and the least MAE and MSE score. The below graph and table show all the r2 values on all all the four models on both test and train data.

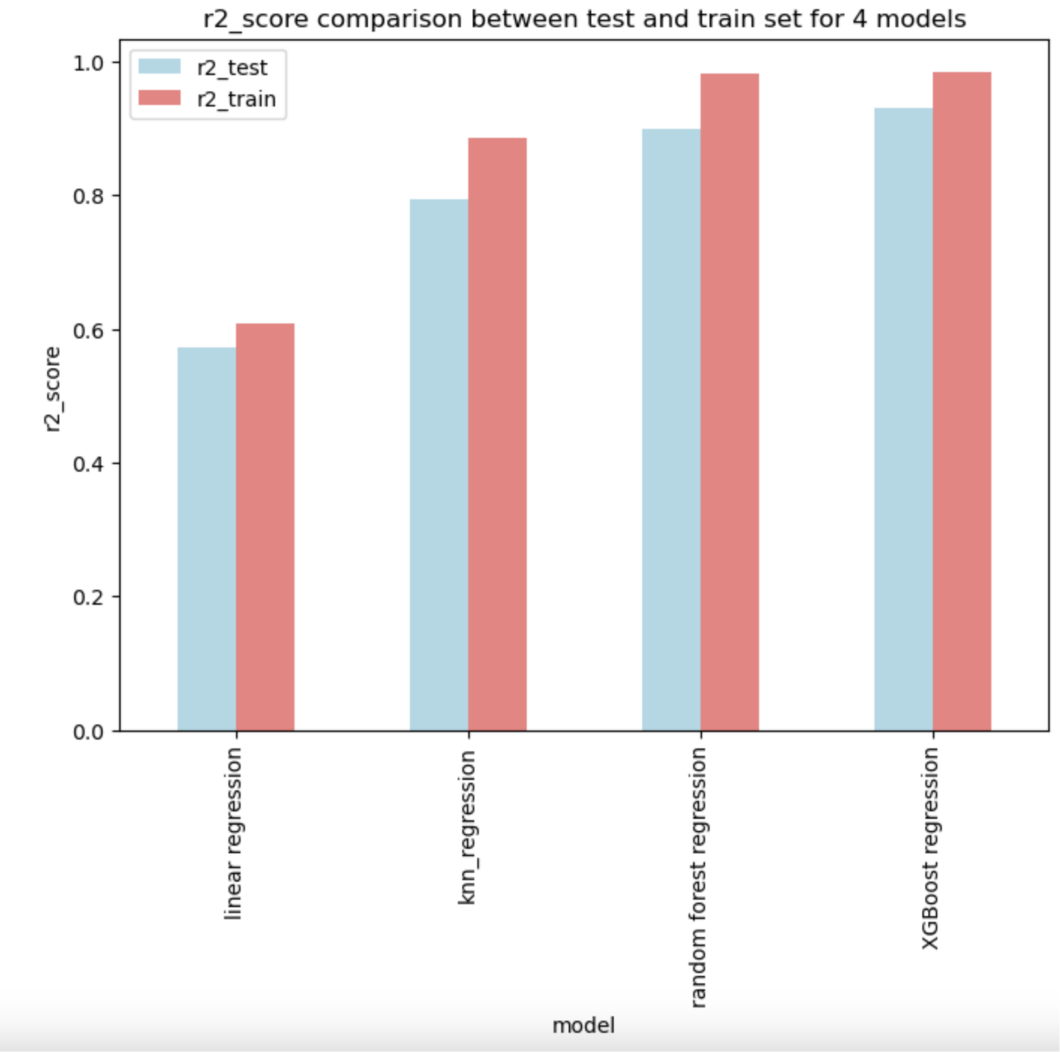
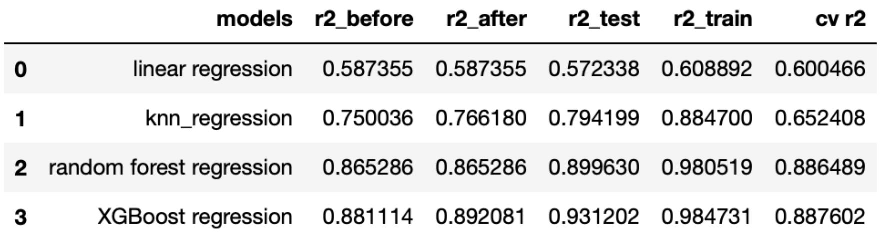


Figure 15: r2 score comparison between test and train set for all 4 models



	models	r2_before	r2_after	r2_test	r2_train	cv r2
0	linear regression	0.587355	0.587355	0.572338	0.608892	0.600466
1	knn_regression	0.750036	0.766180	0.794199	0.884700	0.652408
2	random forest regression	0.865286	0.865286	0.899630	0.980519	0.886489
3	XGBoost regression	0.881114	0.892081	0.931202	0.984731	0.887602

Figure 16: r2 score table

3.2 Mean Squared error

Even when the r2 value was high, the MSE score was relatively high too. MSE score has reduced drastically from linear regression to the XGBoost regression. The MSE score of XGBoost on unseen data is 1.5e10 and the RMSE score is 124646. The smallest value in the selling price is 29999, the maximum value is 7.2e6 and the mean of all selling prices is 5.2e5, the RMSE score is 1.2e5 is smaller compared to the mean. RMSE score is only around 23% of the mean, which shows the performance of the model is not bad.

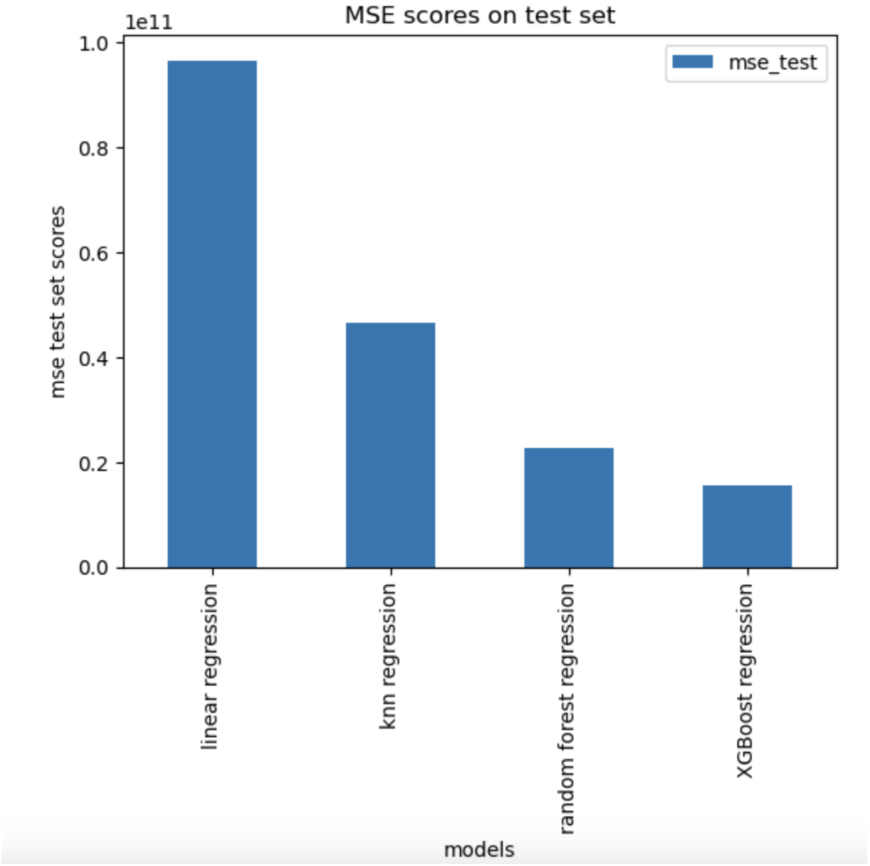


Figure 17: MSE score for all four models

3.3 Mean Absolute Error

MAE score of all models on test data plotted below shows that the XGBoost algorithm has the least MAE of all the models.

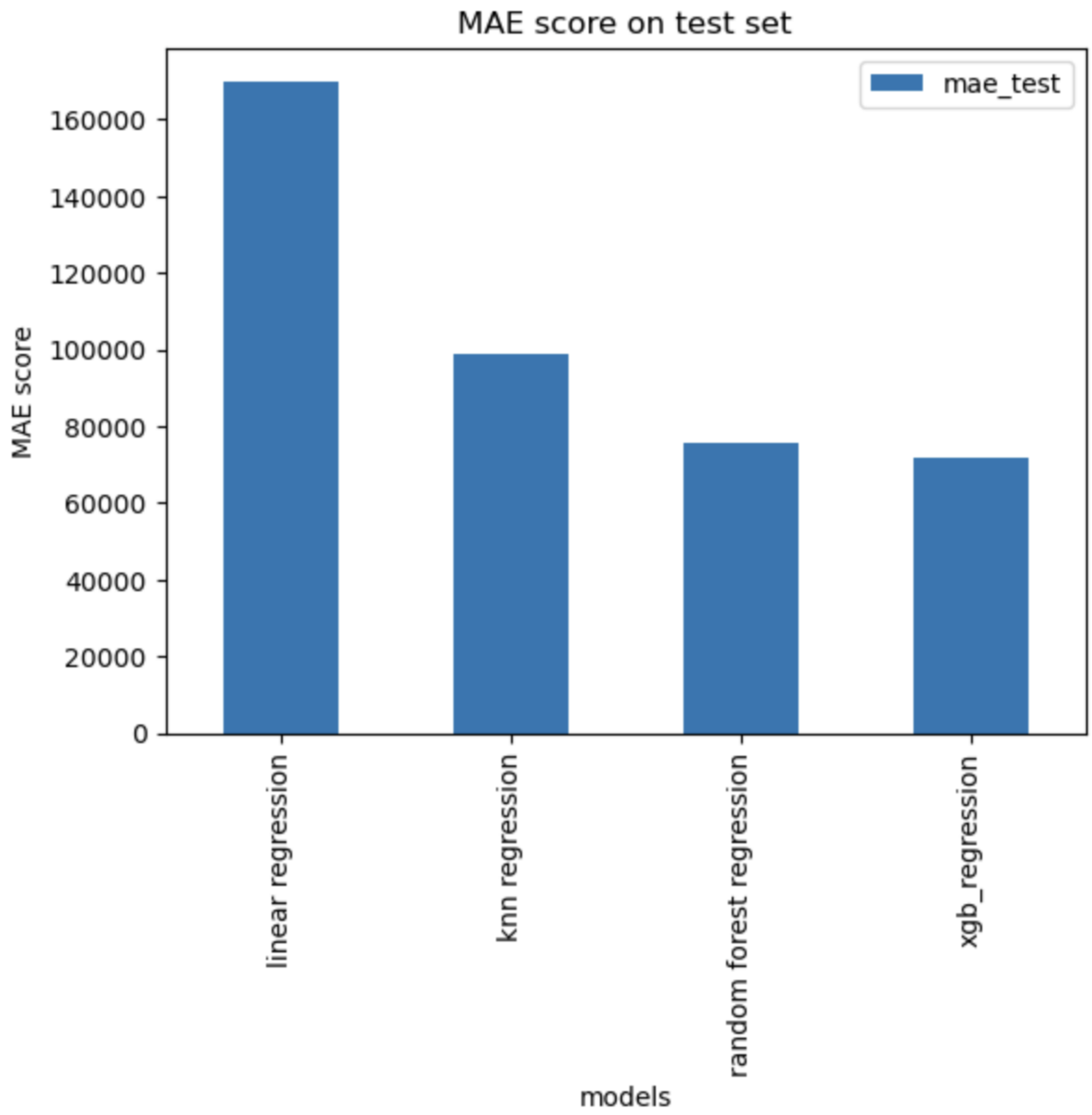
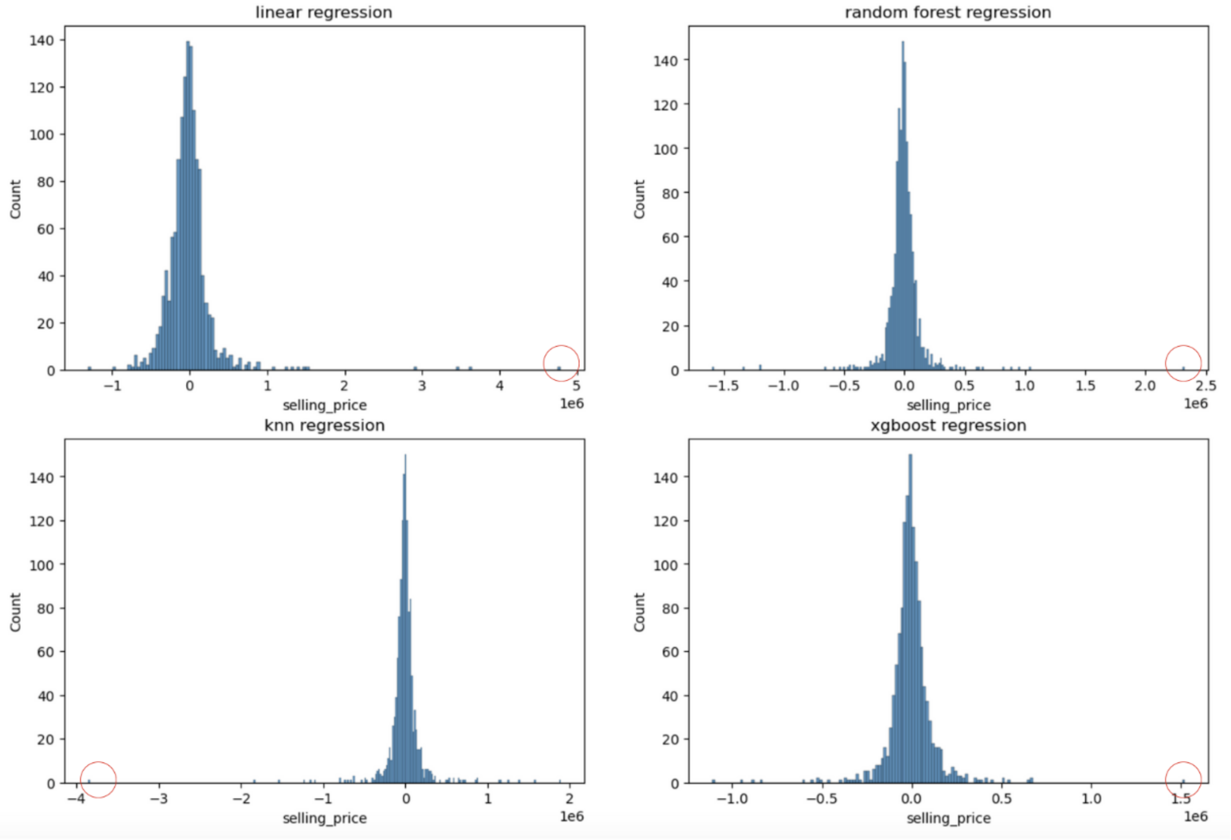


Figure 18: MAE score for all the models

3.4 Residual histograms

The residual histogram of linear regression, random forest regression and XGBoost regression have the predicted value with the biggest difference from the actual value towards the right side of 0 on x axis; it shows that these values were overpredicted or greater than the actual value. Meanwhile, the KNN regression model has the largest residue towards the left side of 0 mark on x axis, which indicates that the value was underpredicted or the predicted value was less compared to the actual value.



4. Reflection

While working on this project, I learnt a lot about different types of models which can be used for regression-based problems. Since I have never worked on regression-based graphs it was quite hard for me to interpret them in the beginning. I had to go through various resources to understand each graph which was plotted for each model. This project has helped me read model performance graphs better.

It confused me that the value of MSE was unusually high even when the r2 score was high. I got to study about the impact of large errors on MSE value. It was also harder to interpret the MSE score because I had not normalised or transformed the target variable even though the range of values were quite high. I would also like to try using log transformation for transforming the target variable in the future.

Since four models were being compared, I was confused on ways to structure all the data in a way that it is easy to read the findings of each model. I contemplated on whether all the models should be evaluated on validation set and then a common grid search needs to be used for all the models but eventually I ended up sequentially moving from one model to another after finishing evaluation on validation set, hyper tuning and evaluation on test set. This made me understand how much time goes into understanding the way information needs to be structured.

There were several roadblocks while I was trying to analyse the data using different models however this project helped me understand a lot of concepts. It also made me sections I need to work on.

5. References

Ahmed, S. (2020). *Metrics and Plots for Analyzing Linear regression models*. [online] Medium. Available at: https://medium.com/ml-course-microsoft-udacity/metrics-and-plots-for-analyzing-linear-regression-models-43b533547574 [Accessed 12 Jul. 2024].

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*. [online] Available at: https://arxiv.org/pdf/1911.01914.

Black, R. (2023). *Why Use Linear Regression? Discover the Impacts of This Analysis.* [online] Medium. Available at: https://medium.com/@reinaldoblack/why-use-linear-regression-discover-the-impacts-of-this-analysis-9fda98a4377e [Accessed 11 Jul. 2024].

Breiman, L. (2001). *Random Forests*. [online] Available at: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

Built In. (2022). *Random Forest: A Complete Guide for Machine Learning | Built In*. [online] Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 11 Jul. 2024].

Frost, J. (2017a). *Five Reasons Why Your R-squared can be Too High*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/regression/r-squared-too-high/ [Accessed 12 Jul. 2024].

Frost, J. (2017b). *Heteroscedasticity in Regression Analysis*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/regression/heteroscedasticity-regression/ [Accessed 12 Jul. 2024].

Goyal, S. (2020). *Boosting performance with XGBoost - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/boosting-performance-with-xgboost-b4a8deadede7 [Accessed 12 Jul. 2024].

Gupta, A. (2021). *XGBoost versus Random Forest - Geek Culture - Medium*. [online] Medium. Available at: https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30 [Accessed 11 Jul. 2024].

Jadon, A., Patil, A. and Jadon, S. (n.d.). *A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting*. [online] Available at: https://arxiv.org/pdf/2211.02989.

KIm, D. (2023). *How to plot Predicted vs Actual Graphs and Residual Plots*. [online] Medium. Available at: https://dooinnkim.medium.com/how-to-plot-predicted-vs-actual-graphs-and-residual-plots-dc4e5b3f304a#:~:text=A%20Predicted%20vs%20Actual%20plot,with%20a%20slope%20of%201. [Accessed 12 Jul. 2024].

Lewis-Stempel, F. (2024). *Britain's used car market is thriving as prices settle, stock improves and cheaper second-hand EVs...* [online] This is Money. Available at: https://www.thisismoney.co.uk/money/cars/article-13060403/Britains-used-car-market-thriving-prices-settle-stock-improves-cheaper-second-hand-EVs-enter-market-experts-say.html [Accessed 11 Jul. 2024].

Maksym Zavershynskyi (2017). *MSE and Bias-Variance decomposition - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/mse-and-bias-variance-decomposition-77449dd2ff55 [Accessed 12 Jul. 2024].

Nischal M (2019). *Bias and variance in linear models - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/bias-and-variance-in-linear-models-e772546e0c30 [Accessed 11 Jul. 2024].

Padhma M (2021). *A Comprehensive Introduction to Evaluating Regression Models*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#:~:text=Mean%20Squared%20Error%20(MSE)%20is,MSE%20for%20robust%20outlier%20handling. [Accessed 12 Jul. 2024].

Sivasai Yadav Mudugandla (2020). *Data Leakage in Hyperparameter Tuning | Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/data-leakage-with-hyper-parameter-tuning-c57ba2006046 [Accessed 12 Jul. 2024].

T. Dorigo, Guglielmini, S., J. Kieseler and Strong, G. (2022). *Deep Regression of Muon Energy with a K-Nearest Neighbor Algorithm*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/359080822_Deep_Regression_of_Muon_Energy_with_a_K-Nearest_Neighbor_Algorithm [Accessed 11 Jul. 2024].

Timbers, T. (2023). *Chapter 7 Regression I: K-nearest neighbors* | *Data Science*. [online] Datasciencebook.ca. Available at: https://datasciencebook.ca/regression1.html [Accessed 11 Jul. 2024].