Investigating machine learning models and data scaling techniques on model performance

July 6, 2025

1 Abstract

This experiment explores the impact of data scaling techniques and machine learning models on model performance, with the goal of minimizing prediction error measured by root mean squared error (RMSE) on three different datasets. Using a *randomized blocked factorial design*, we evaluated three scaling methods (StandardScaler, MinMaxScaler, and RobustScaler) and three machine learning models (Linear Regressor, Decision Tree Regressor, and K-NN Regression) across three datasets: Audit Data, Bike Sharing Data, and California Housing Data. Our findings indicate that certain combinations of scaling techniques and models outperform others, providing valuable insights into optimal preprocessing and modeling strategies.

2 Introduction

Machine learning algorithms are crucial tools for predictiction tasks, but their effectiveness depends significantly on preprocessing choices and the characteristics of the chosen models. Among preprocessing techniques, data scaling plays a vital role in ensuring that models interpret feature magnitudes appropriately, particularly for models sensitive to feature scales or prone to optimization challenges. Likewise, machine learning models vary in their response to preprocessing due to differences in their underlying mechanics.

This project investigates how the interplay between data scaling methods and machine learning models affects predictive performance, as measured by RMSE. To achieve this, we employ a randomized blocked factorial design, systematically examining three scaling methods and three machine learning models across three datasets with distinct predictive tasks. By leveraging 5-fold cross-validation and statistical techniques such as ANOVA, we aim to determine which combinations of levels of these factors minimize prediction error and whether the observed differences are statistically significant.

The rest of the report is organized as follows: Section 3 describes the data collection process and experimental design, including details on treatment factors, blocking factors, and response variables. Section 4 presents the results, highlighting key findings and statistical analysis. Section 5 discusses the implications of the results, and provides a conclusion with a summary of findings along with recommendations for future research.

3 Details of The Experimental Design

3.1 Experimental Factors and Levels

The experiment will consist of 2 treatment factors and 1 blocking factor. The dataset will be randomly split into subsets for the treatments to be allocated to. The factors and their corresponding levels are summarized in *Table 1* below.

Factor	**Description**	**Levels**
Scaler	Preprocessing technique to scale features.	StandardScaler, MinMaxScaler, RobustScaler
Model	Type of predictive model used	Linear Regression, Decision Tree Regression, K-NN Regression
Dataset (Blocking Factor)	Dataset used for training and testing, included to reduce variability caused by dataset-specific characteristics.	California Housing Dataset, London Bike Sharing Dataset, Audit Risk Data

Table 1: Factors and their descriptions with levels.

StandardScaler (standardizes features by removing the mean and scaling to unit variance), MinMaxScaler (scales each feature to a specified range, typically [0, 1]) and RobustScaler (scales features using statistics that are robust to outliers.) were chosen to represent common preprocessing techniques with varying sensitivity to outliers and scaling ranges. The variety ensures a thorough evaluation of how scaling affects model performance.

Linear Regression, Decision Tree Regression, and KNN Regression were selected to represent diverse modeling approaches, from simple linear models to more complex and flexible methods.

Using datasets as blocking factors ensures that the variability due to dataset-specific characteristics does not confound the effects of the scaling and modeling treatments. We try to maintain the effects of other explanatory variables (that are not treatment factors) constant. The links to the datasets we used are provided in the appendix section.

3.2 Response Variable

The response variable for this experiment is the Root Mean Squared Error (RMSE), which measures the average magnitude of prediction error. RMSE is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of data points.

RMSE is reported in the units specific to the dataset's target variable. So for each dataset the target variable is given as follows:
Audit Data: Audit risk score, **Bike Sharing Data:** Count of bike rentals and **California Housing Data:** Median house house prices (USD).

3.3 Data Collection & Experimental Plan

This experiment is a 3 × 3 × 3 full factorial experiment. It is a complete randomized block design (RBD) so the treatments are complete within every block. The datasets serve as blocking factors to control variability, while the treatments consist of all combinations of data scaling methods and machine learning models.

The steps for data collection in this experiment are outlined below. The corresponding Python script can be found in the attached documents.

1. **Data Splitting:** Each dataset is split into 80% training and 20% testing sets.
2. **Scaling Application:** Each scaling method is fit to the training data and applied to both training and testing data to avoid information leakage.
3. **Cross-Validation:** For each dataset, 5-fold cross-validation is used to evaluate the performance of each combination of scaling method and model.
4. **Randomization:** The order of applying scaling-method/model combinations to the data subsets is randomized within each block. So for each Dataset, a random Scaler and a random Model is selected.

Figure 1 shows the randomized order of runs for the treatments. Each combination of scaling method and model is applied within each dataset block.

This complete design ensures each treatment combination is evaluated for each dataset block, resulting in a total of 3 × 3 × 3 = 27 runs. By employing this rigorous design, the experiment aims to isolate the effects of scaling methods and machine learning models on predictive performance and determine the most effective combinations.

4 Analysis & Results

We fit our model as such:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \alpha_k + E_{ijk} \text{ with } i, j, k \in \{1, 2, 3\}$$

such that

$$Y_{ijk} \text{ is mean RMSE at level i,j,k}$$

$$\mu \text{ is the overall effect common to all treatments}$$

$$\tau_i \text{ is the effect of Model at level i}$$

$$\beta_j \text{ is the effect of Dataset at level j}$$

$$\alpha_k \text{ is the effect of Scaler at level k}$$

$$E_{ijk} \text{ are assumed to be independent } N(0, \sigma^2) \text{ random variables}$$

After data collection we started our analysis by first plotting the observed level means response for each factor. *Figure 2* shows the mean RMSE against all factors in one plot, one factor at a time. We observe that Dataset level means has the most variation, followed by Models and then Scaler.

4.1 Main Effects: ANOVA

To investigate the impact of the factors Dataset, Model, and Scaler on the response variable RMSE, we conducted an analysis of variance using ANOVA. We began by examining the main effects of each factor. The results of the ANOVA on the main effects are summarized in *Figure 3.*
Our hypothesis is that

$$H_0 = \tau_i = 0, i \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \tau_i \neq 0, i \in \{1,2,3\}$$

$$H_0 = \beta_j = 0, j \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \beta_j \neq 0, j \in \{1,2,3\}$$

$$H_0 = \alpha_k = 0, k \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \alpha_k \neq 0, k \in \{1,2,3\}$$

The analysis revealed that there is enough evidence to show that both factors Dataset and Model have a statistically significant effect on mean RMSE, while the Scaler factor did not show a significant influence on the response variable. Based on this, we can say that Dataset and Model are the primary factors driving the variations in mean RMSE.

The p-values for these factors were both below the 0.05, hence at a 5% significance level, we have enough evidence to reject the null hypothesis for both Models and Dataset. We do not have enough evidence to reject the null hypothesis for the factor Scaler. A deeper investigation into the effects of Dataset and Model will help identify which specific datasets and models contribute most to variations in RMSE.

4.2 Interactions Effects: ANOVA

Let us now look at the two factor interactions effects of these factors on the response. We expanded the analysis to examine potential interactions between the factors. This interaction model aimed to test whether the effect of one factor and mean RMSE depends on the levels of other factors. The ANOVA results for the interaction effects are shown in *Figure 4.*
Our hypothesis is now that

$$H_0 = \tau_i\beta_j = 0, i,j \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \tau_i\beta_j \neq 0, i,j \in \{1,2,3\}$$

$$H_0 = \tau_i\alpha_k = 0, i,k \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \tau_i\alpha_k \neq 0, i,j \in \{1,2,3\}$$

$$H_0 = \beta_j\alpha_k = 0, j,k \in \{1,2,3\} \text{ and } H_A = \text{ alteast one } \beta_j\alpha_k \neq 0, i,j \in \{1,2,3\}$$

We first observed from *Figure 3 & 4* that the parameter estimates for main effects are unchanged in both models (one with interaction terms and one without). This implies that the individual effects of the main factors (Dataset, Model, Scaler) on the mean RMSE are the same regardless of whether you include interaction terms between the factors or not in your model.

The results indicated that the Dataset:Model interaction effect was significant at a 5% significance level, suggesting that the effect of Dataset on RMSE differs depending on the model used. Hence, we have enough evidence to reject the null hypothesis that $\tau_i\beta_j = 0$. There is not enough evidence to reject the null hypothesis for the other interaction effects.

We could also consider 3 factor interaction effects however *Figure 5.* shows that is not possible. With an unreplicated experiment we cannot estimate effects for the highest-order interaction effects. Such effects would have 8 DF which would leave no DF for residuals as shown in the table. This indicates that there are not enough observations to provide precise estimates of all interaction terms. In this case, replication can increase residual degrees of freedom, improving the model's ability to detect significant effects and interactions more reliably. However, in our case, we will stick to 2 factor interactions only.

4.3 Investigating assumptions for the data and Box Cox Analysis

To further evaluate the fitted model and assess potential issues with the assumptions of the analysis, we examined diagnostic plots for the interaction model. In *Figure 6*, we present a residuals vs. fitted plot, which allows us to evaluate the assumptions of homoscedasticity (constant variance) and linearity. We observe a pattern within the plotted points which indicates that the homoscedasticity and linearity assumptions are likely to be not satisfied.

In *Figure 7.*, we display a Q-Q plot to assess the normality of the residuals. If the points significantly deviate from the straight line, it would suggest that the residuals are not normally distributed. We notice substantial deviations in the tails of the plot, which suggests that the residuals may not follow a normal distribution.

To assess whether a transformation could improve the model fit, we applied a Box-Cox transformation to the model. A Box-Cox transformation can potentially improve the validity of several statistical assumptions of a linear model. The log-likelihood plot produced by this analysis helps us determine the optimal value of λ for the transformation. *Figure 8.* shows the Box-Cox plot plotted for the model. We observe that $\lambda = 1$ (no transformation) does not fall within the 95 % confidence interval, suggesting that there is a need for transformation. After evaluating the plot, we concluded that λ around 0.125 approximately maximizes the log-likelihood.

When we tested transformations of the response variable with $\lambda = 0.5$, 0.25 and 0.125 we observe that the value that normalized the data properly was $\lambda = 0.25$ making it the most suitable transformation. So we transformed the RMSE variable by applying a power transformation to the data. Specifically, we raised the RMSE values to the power of 0.25.

Our model is now

$$Y_{ijk}^{\lambda} = \mu + \tau_i + \beta_j + \alpha_k + E_{ijk} \text{ with } i, j, k \in \{1, 2, 3\}$$

We plotted another log-likelihood plot in *Figure 9*, this time using our transformed RMSE values. This plot show that $\lambda = 1$ is narrowly excluded from the confidence interval, which shows that our transformation $\lambda = 0.25$ was a convenient choice.

We plot a residuals vs fitted plot and QQ-plot again so that we can verify whether the transformation we applied to the response variable was effective in meeting the assumptions of ANOVA. *Figures 10 & 11* show no pattern within the residuals and points lying on a 45degrees line respectively, indicating that the normality assumption is now satisfied. This suggests that the transformed model is a better fit for the data.

We now re-run an ANOVA analysis with our transformed RMSE and display the results in *Figure 12.* We see that we are still constistent with the results above. Dataset, Model and Dataset:Model interaction effect are still significant at a 5% significance level.

4.4 Further exploring Interaction Effects

We further explore the significant interaction effects between the Dataset and Model factors. We first make interaction plots to visualize how the relationship between these two factors influences the transformed RMSE values. These plots help identify whether the effect of one factor on the response variable depends on the level of the other factor.

We first look at the effect of the levels of factor Model when changing levels Dataset in *Figure 13.* We see that changing from Audit to California Housing increases the RMSE for all levels of Model and similarly changing from California Housing to London Bike Sharing decreases the RMSE for all levels of Model. It is hence fair to say that different levels of Model do not necessarily affect the effect of Dataset on RMSE.

Then, we look at the effect of the levels of factor Dataset when changing levels Model in *Figure 14.* We see that changing from Decision Tree to KNN makes no effect to RMSE when Dataset is Audit, but we notice a slight decrease when Dataset is London Bike Sharing and an even bigger decrease when

Dataset is California Housing. Similarly, when we switch from KNN to Linear Regression, we see that there is no effect to RMSE when Dataset is Audit and London Bike Sharing, but there is a small increase to RMSE when Dataset is California Housing.

Hence, we can conclude that Dataset levels influence RMSE when changing Model's levels to a certain extent.

4.5 Contrasts

To further investigate the effects of the Model factor on the transformed RMSE values, contrasts were set up to compare specific levels of the Model factor.
As a reminder, the Model factor contains three levels: Decision Tree, KNN, and Linear Regression. We hence definie 2 orthogonal contrasts:

1. DTvLR: Comparing Decision Tree (-1) versus Linear Regression (+1), with KNN (0).
2. DTLRvKNN: Comparing Decision Tree and Linear Regression (+1 each) versus KNN (-2).

This is better illustrated using a table.

	Decision Tree	Linear Regression	KNN
DTvLR	-1	1	0
DTLRvKNN	1	1	-2

We run another ANOVA analysis and get the results shown in *Figure 15.* The results confirm that all levels of Model are significant. We then look at the estimates of $\hat{k_1}\&\hat{k_2}$ which represent DTvLR and DTLRvKNN respectively. We retrieve the estimates and standard error and calculate the confidence interval as shown below.

$$\hat{k_1} \pm t_{8,0.975} se(\hat{k_1}) = -0.126 \pm 2.31 \text{ x } 0.00734 = [-0.143, -0.109]$$

$$\hat{k_2} \pm t_{8,0.975} se(\hat{k_2}) = 0.361 \pm 2.31 \text{ x } 0.0127 = [0.332, 0.390]$$

We are hence 95% confident that that mean RMSE was 0.109 to 0.143 less for Decision Tree than Linear Regression and that mean RMSE was 0.332 to 0.39 more for Decision Tree and Linear Regression than for KNN.

5 Conclusion

This experiment shows us our model needed a transformation to fit normal assumptions. Once that was done, ANOVA analysis revealed that out of the 3 factors we have, Model and Dataset are the statistically significant ones, which means they contribute to the RMSE. We have further shown that in terms of 2 factor interaction effects, Model:Dataset was the only significant one.

We further investigated the effects of changing the levels of one factor on another factor, on the response RMSE, namely Model and Dataset. We saw that Dataset levels influence RMSE when changing Model's levels to a certain extent but Model levels do not influence the effect of Dataset on RMSE.

Finally, we compared different levels of Model on RMSE by setting up contrasts and saw that mean RMSE for Decision Tree was less than Linear Regression and mean RMSE was more for Decision Tree and Linear Regression than it was for KNN. Since our goal is to minimize RMSE, it is fair to say that choosing the method Decision Tree is the best way to achieve that and that Scaling Method does not influence the RMSE.

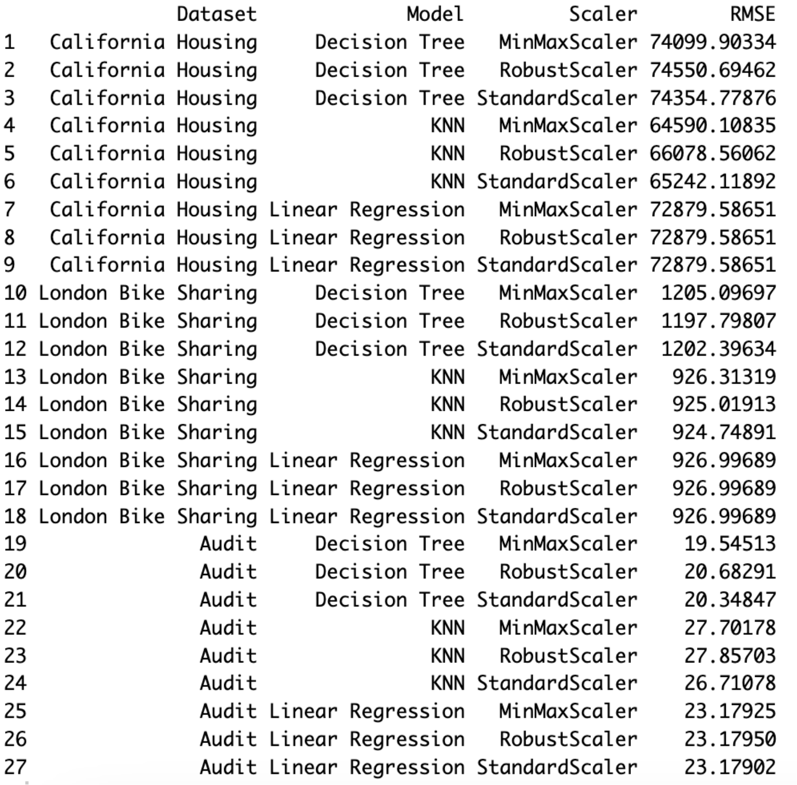
Something I might do differently would be to include more datasets with more variety and do a factorial experiment to reduce the number of runs.

6 Appendix

Here are the datasets we used for data collection, in this experiment. The script to generate our data has been submitted as well:

1. Audit Data : https://archive.ics.uci.edu/dataset/475/audit+data
2. Bike Sharing Dataset : https://www.kaggle.com/code/tomvoss/london-bike-sharing-prediction
3. California Housing Dataset : https://www.kaggle.com/datasets/harrywang/housing

Data Appendix



```
              Dataset             Model         Scaler        RMSE
1   California Housing     Decision Tree   MinMaxScaler 74099.90334
2   California Housing     Decision Tree   RobustScaler 74550.69462
3   California Housing     Decision Tree StandardScaler 74354.77876
4   California Housing               KNN   MinMaxScaler 64590.10835
5   California Housing               KNN   RobustScaler 66078.56062
6   California Housing               KNN StandardScaler 65242.11892
7   California Housing Linear Regression   MinMaxScaler 72879.58651
8   California Housing Linear Regression   RobustScaler 72879.58651
9   California Housing Linear Regression StandardScaler 72879.58651
10 London Bike Sharing     Decision Tree   MinMaxScaler  1205.09697
11 London Bike Sharing     Decision Tree   RobustScaler  1197.79807
12 London Bike Sharing     Decision Tree StandardScaler  1202.39634
13 London Bike Sharing               KNN   MinMaxScaler   926.31319
14 London Bike Sharing               KNN   RobustScaler   925.01913
15 London Bike Sharing               KNN StandardScaler   924.74891
16 London Bike Sharing Linear Regression   MinMaxScaler   926.99689
17 London Bike Sharing Linear Regression   RobustScaler   926.99689
18 London Bike Sharing Linear Regression StandardScaler   926.99689
19               Audit     Decision Tree   MinMaxScaler    19.54513
20               Audit     Decision Tree   RobustScaler    20.68291
21               Audit     Decision Tree StandardScaler    20.34847
22               Audit               KNN   MinMaxScaler    27.70178
23               Audit               KNN   RobustScaler    27.85703
24               Audit               KNN StandardScaler    26.71078
25               Audit Linear Regression   MinMaxScaler    23.17925
26               Audit Linear Regression   RobustScaler    23.17950
27               Audit Linear Regression StandardScaler    23.17902
```

Figure 1: Figure showing the combination of treatment factors for each blocking factor for our experimental runs (randomized).

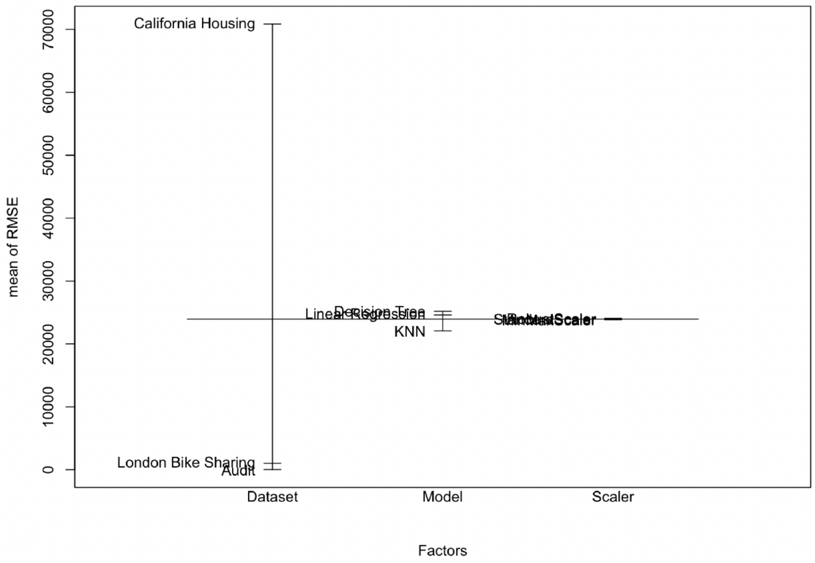
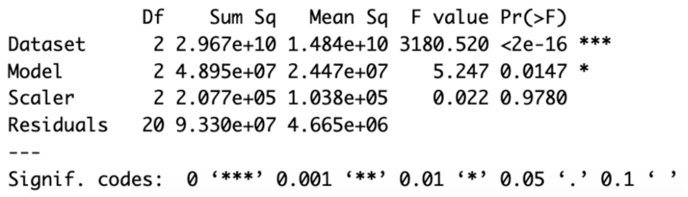
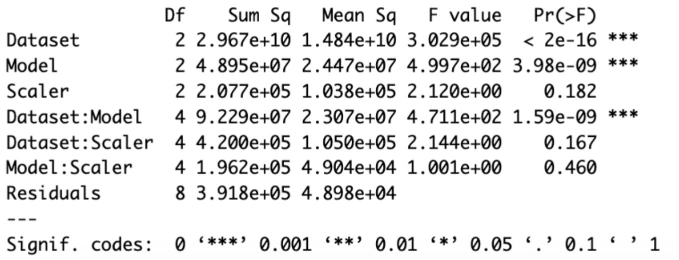


Figure 2: Mean RMSE plotted against the 3 levels of factors : Dataset, Model and Scaler



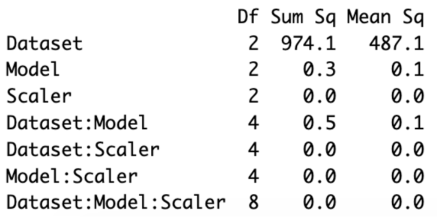
```
            Df    Sum Sq   Mean Sq  F value Pr(>F)    
Dataset      2 2.967e+10 1.484e+10 3180.520 <2e-16 ***
Model        2 4.895e+07 2.447e+07    5.247 0.0147 *  
Scaler       2 2.077e+05 1.038e+05    0.022 0.9780    
Residuals   20 9.330e+07 4.665e+06                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

Figure 3: Main effects ANOVA table.



```
               Df    Sum Sq   Mean Sq   F value   Pr(>F)    
Dataset         2 2.967e+10 1.484e+10 3.029e+05  < 2e-16 ***
Model           2 4.895e+07 2.447e+07 4.997e+02 3.98e-09 ***
Scaler          2 2.077e+05 1.038e+05 2.120e+00    0.182    
Dataset:Model   4 9.229e+07 2.307e+07 4.711e+02 1.59e-09 ***
Dataset:Scaler  4 4.200e+05 1.050e+05 2.144e+00    0.167    
Model:Scaler    4 1.962e+05 4.904e+04 1.001e+00    0.460    
Residuals       8 3.918e+05 4.898e+04                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: 2 factor interaction effects ANOVA table.



```
                     Df Sum Sq Mean Sq
Dataset               2  974.1   487.1
Model                 2    0.3     0.1
Scaler                2    0.0     0.0
Dataset:Model         4    0.5     0.1
Dataset:Scaler        4    0.0     0.0
Model:Scaler          4    0.0     0.0
Dataset:Model:Scaler  8    0.0     0.0
```

Figure 5: 3 factor interaction effects ANOVA Table

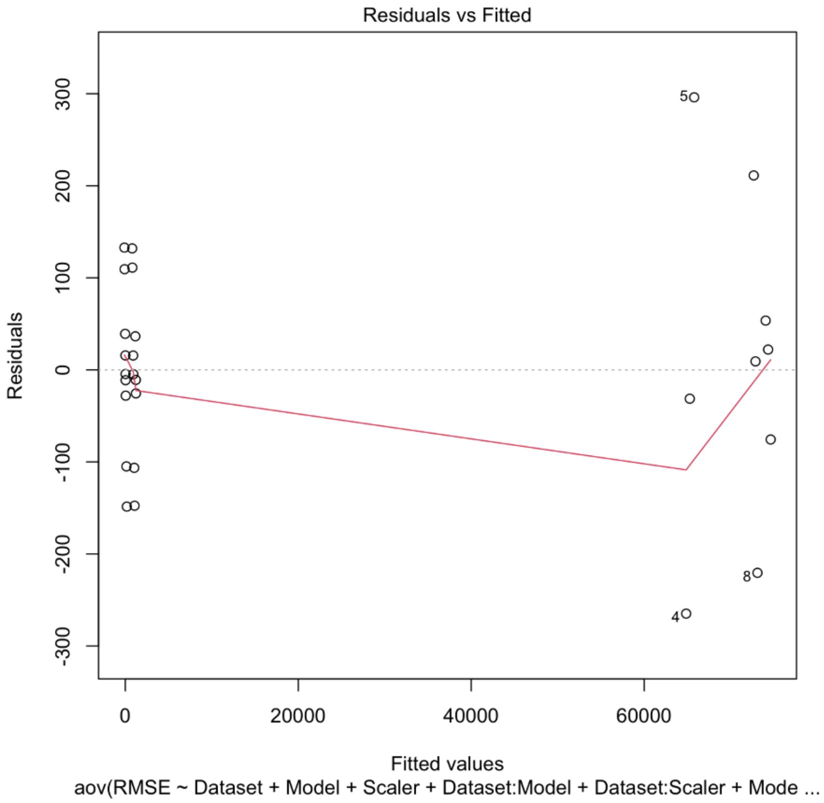


Figure 6: Residuals vs Fitted Values for the 2 factor interaction model

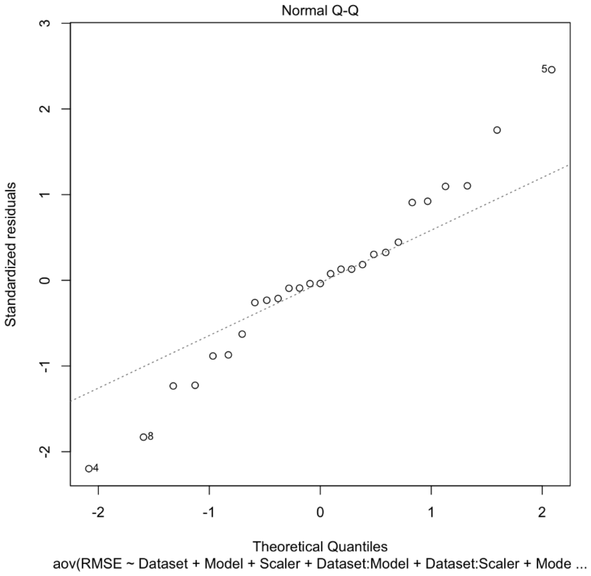


Figure 7: QQ plot for the 2 factor interaction model

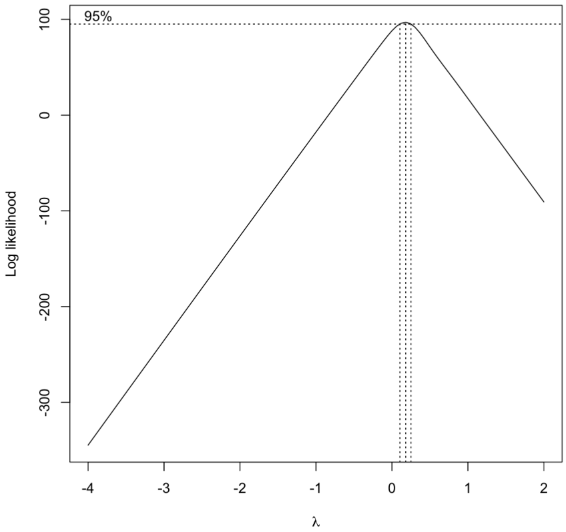


Figure 8: Log-likelihood plot produced for Box-Cox analysis

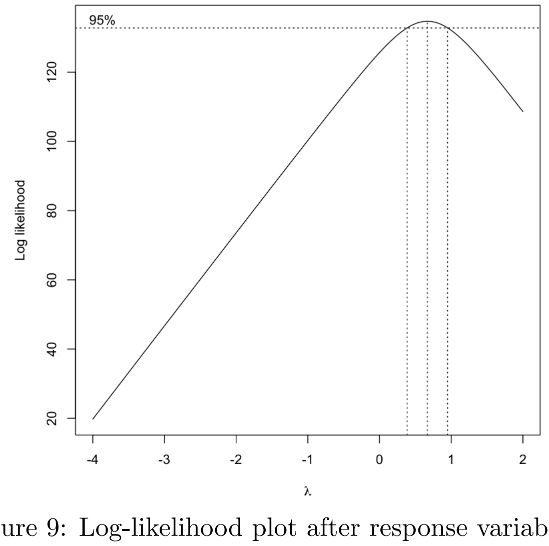


Figure 9: Log-likelihood plot after response variable is transformed with lambda = 0.25

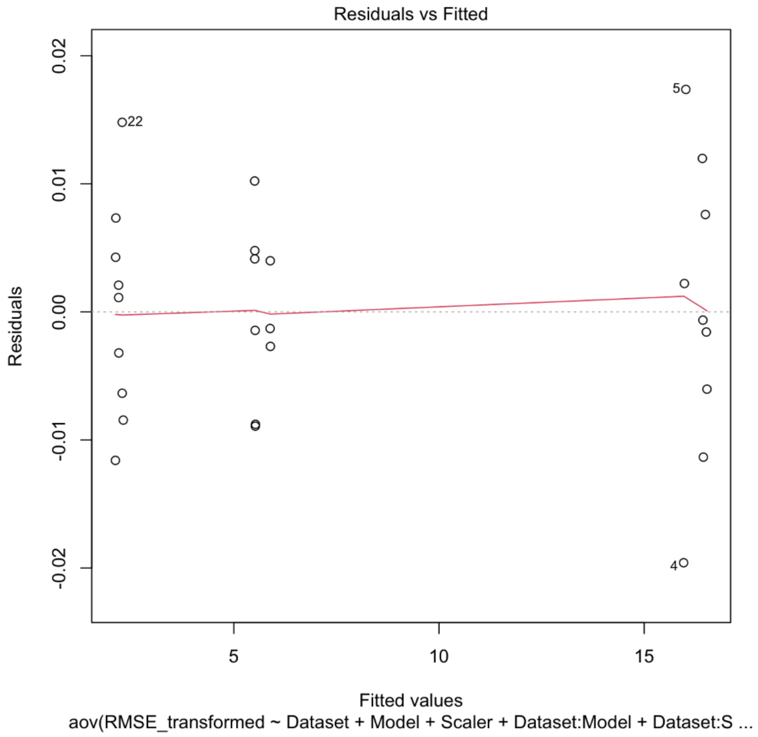


Figure 10: Residuals vs Fitted Values(Transformed response variable)

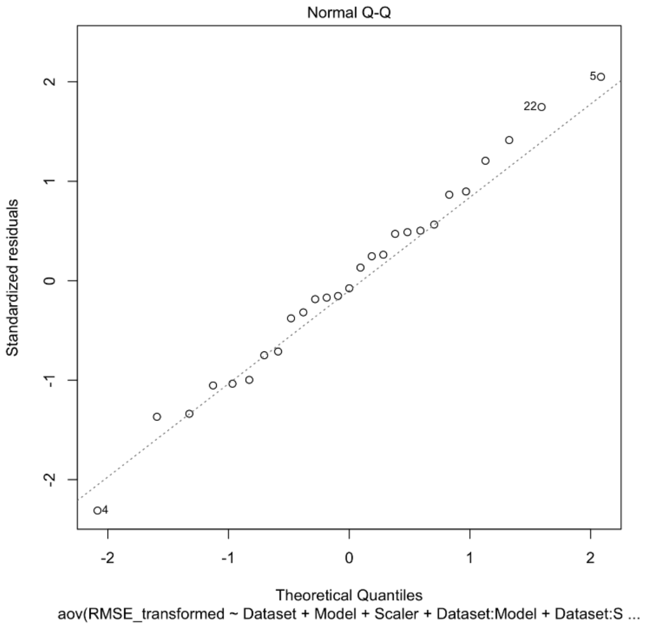
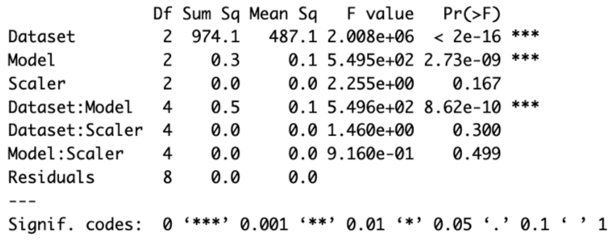


Figure 11: QQ plot (Transformed response variable)



```
               Df Sum Sq Mean Sq   F value   Pr(>F)    
Dataset         2  974.1   487.1 2.008e+06  < 2e-16 ***
Model           2    0.3     0.1 5.495e+02 2.73e-09 ***
Scaler          2    0.0     0.0 2.255e+00    0.167    
Dataset:Model   4    0.5     0.1 5.496e+02 8.62e-10 ***
Dataset:Scaler  4    0.0     0.0 1.460e+00    0.300    
Model:Scaler    4    0.0     0.0 9.160e-01    0.499    
Residuals       8    0.0     0.0                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12. Interaction effects ANOVA table. (Transformed response variable)

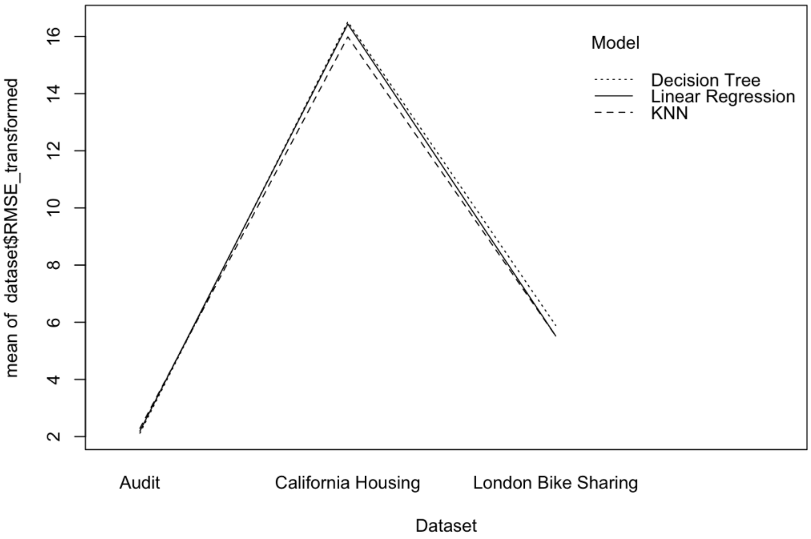


Figure 13: Interaction plot (Mean transformed RMSE vs Dataset)

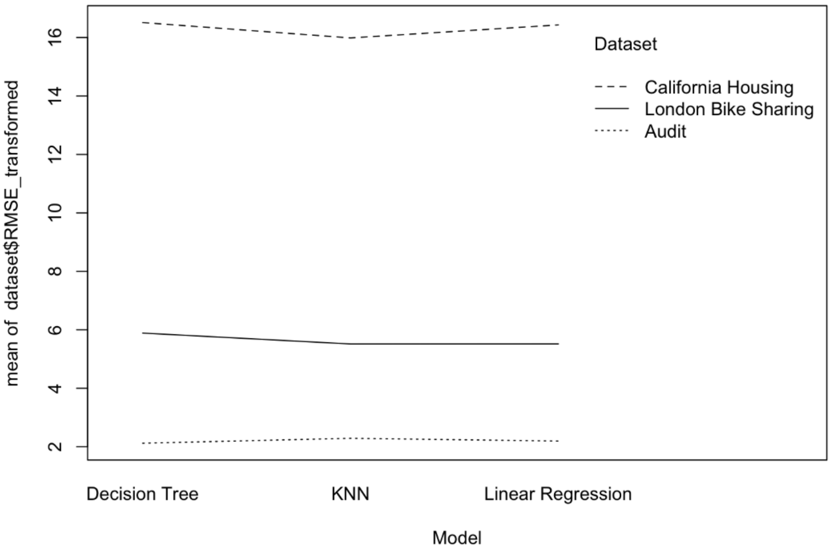
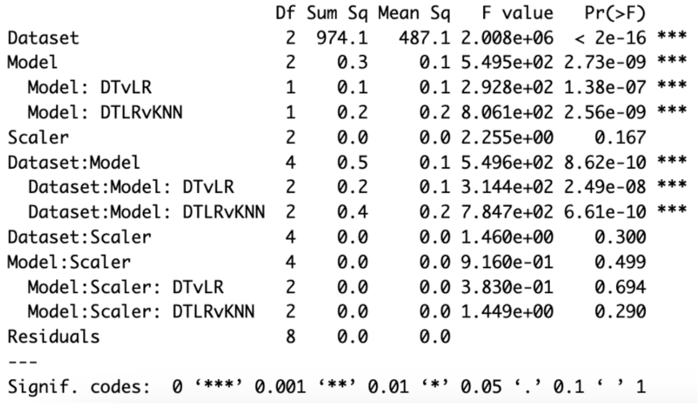


Figure 14: Interaction plot (Mean transformed RMSE vs Model)



```
                           Df Sum Sq Mean Sq   F value   Pr(>F)    
Dataset                     2  974.1   487.1 2.008e+06  < 2e-16 ***
Model                       2    0.3     0.1 5.495e+02 2.73e-09 ***
  Model: DTvLR              1    0.1     0.1 2.928e+02 1.38e-07 ***
  Model: DTLRvKNN           1    0.2     0.2 8.061e+02 2.56e-09 ***
Scaler                      2    0.0     0.0 2.255e+00    0.167    
Dataset:Model               4    0.5     0.1 5.496e+02 8.62e-10 ***
  Dataset:Model: DTvLR      2    0.2     0.1 3.144e+02 2.49e-08 ***
  Dataset:Model: DTLRvKNN  2    0.4     0.2 7.847e+02 6.61e-10 ***
Dataset:Scaler              4    0.0     0.0 1.460e+00    0.300    
Model:Scaler                4    0.0     0.0 9.160e-01    0.499    
  Model:Scaler: DTvLR       2    0.0     0.0 3.830e-01    0.694    
  Model:Scaler: DTLRvKNN    2    0.0     0.0 1.449e+00    0.290    
Residuals                   8    0.0     0.0                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 15: ANOVA with contrasts