CS 519 Applied Machine Learning I

HW5: Regression Methods Comparison

Submitted by: Indronil Bhattacharjee

1 Dataset:

California housing dataset, which is loaded using fetch_california_housing from sklearn.datasets. All the columns and all the instances in this dataset have been used for this analysis.

2 Initial Performance comparison among different linear regression methods:

Comparison among different metrics like MSE (Mean-squared error), R^2 score and time required for different linear regression methods like Linear Regression, RANSAC, Ridge, Lasso and ElasticNet Regression in the following:

(a) Mean squared error (MSE):

The mean squared error (MSE) measures the average squared difference between the predicted values and the actual values. In this context, the linear regression model and the ridge regression model achieved similar MSE values of approximately 0.556, indicating that they have comparable performance in terms of minimizing prediction errors. The RANSAC regression model, however, yielded a higher MSE of approximately 0.971, suggesting that it may not be as effective in capturing the underlying patterns in the data. Similarly, the Lasso and ElasticNet regression models produced MSE values of approximately 0.938 and 0.765, respectively, indicating moderate performance but slightly worse than linear and ridge regression.

Values for MSE values for different regression models:

```
Mean Squared Errors
--------------------------
Linear Regression MSE: 0.555891598695244
RANSAC Regression MSE: 0.9713565431329947
Ridge Regression MSE: 0.5558034669932209
Lasso Regression MSE: 0.9380337514945427
ElasticNet Regression MSE: 0.7645556403971131
```

Plots for MSE comparison for different regression models:

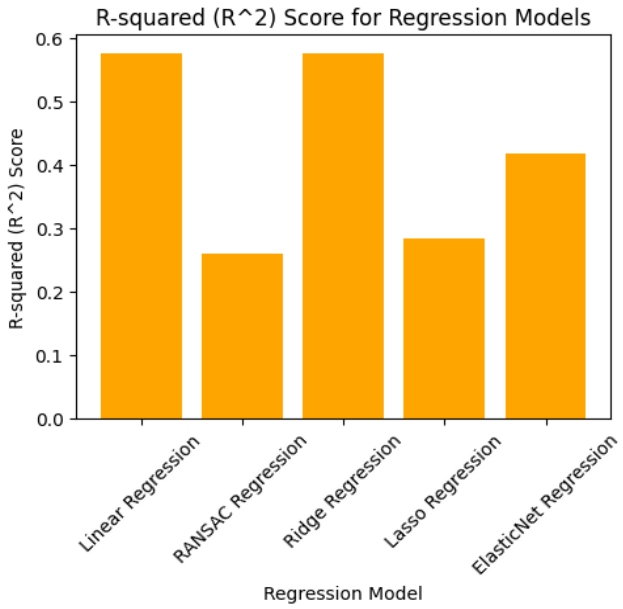
(b) R-squared (R^2) score:

The R-squared (R^2) score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 score indicates a better fit of the regression model to the data. In this analysis, both linear regression and ridge regression achieved R^2 scores around 0.576, indicating that they explain approximately 57.6% of the variance in the target variable. The RANSAC regression model had a considerably lower R^2 score of approximately 0.259, suggesting that it explains less variance in the data compared to the other models. The Lasso and ElasticNet regression models had R^2 scores of approximately 0.284 and 0.417, respectively, indicating moderate performance.

Values for R^2 score for different regression models:

```
R-squared Score
--------------------------
Linear Regression R2 Score: 0.5757877060324511
RANSAC Regression R2 Score: 0.25873787553184413
Ridge Regression R2 Score: 0.5758549611440127
Lasso Regression R2 Score: 0.2841671821008396
ElasticNet Regression R2 Score: 0.41655189098028245
```

Plots for R^2 score comparison for different regression models:



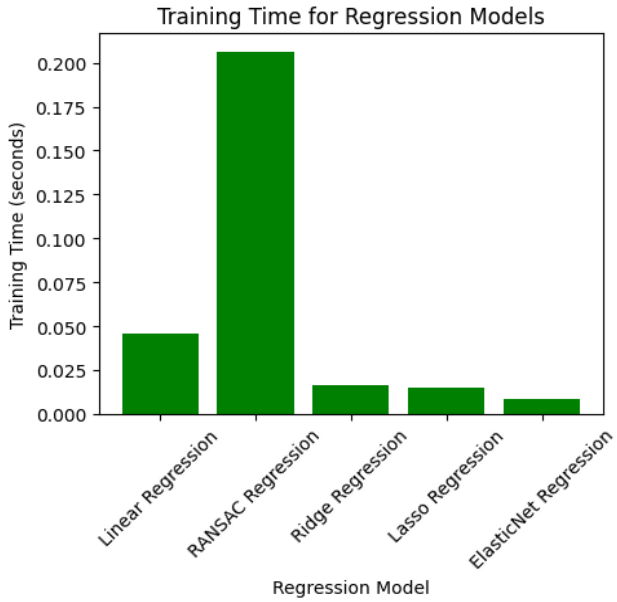
(b) Training time:

Training time represents the time taken to train each regression model on the dataset. In this analysis, the linear regression model had the highest training time of approximately 0.046 seconds, followed by the RANSAC regression model with a training time of approximately 0.206 seconds. On the other hand, the ridge regression, Lasso regression, and ElasticNet regression models exhibited much lower training times, ranging from approximately 0.008 to 0.016 seconds. These results suggest that the regularization techniques used in ridge regression, Lasso regression, and ElasticNet regression help in reducing overfitting and improving computational efficiency compared to standard linear regression and RANSAC regression.

Values for training time for different regression models:

```
Training Time
--------------------------
Linear Regression Time: 0.045546531677246094
RANSAC Regression Time: 0.20639872550964355
Ridge Regression Time: 0.016172170639038086
Lasso Regression Time: 0.01494741439819336
ElasticNet Regression Time: 0.008494853973388672
```

Plots for training time comparison for different regression models:



3 Performance of non-linear Kernel Ridge Regression with RBF Kernel:

For non-linear regression model, Kernel Ridge Regression with RBF kernel is used. Different metrics like MSE (Mean-squared error), R^2 score and time required for this model is in the following:

Values for training time for different regression models:

```
MSE of Kernel Ridge Regression with RBF Kernel: 3.5817028817952665
R2 score of Kernel Ridge Regression with RBF Kernel: -1.7332710178797495
Training Time  of Kernel Ridge Regression with RBF Kernel: 200.70035576820374
```

- For the Kernel Ridge Regression with RBF Kernel, we observe a significantly higher mean squared error (MSE) of 3.582 compared to the linear regression models. This suggests that the model's predictions deviate considerably from the actual target values, indicating poor performance in capturing the underlying relationships in the data.
- Similarly, the R-squared (R^2) score is notably negative, indicating that the model performs worse than a naive mean prediction. A negative R^2 score suggests that the model fails to explain any variance in the target variable and performs even worse than a horizontal line representing the mean of the target values.
- Furthermore, the training time for the non-linear Kernel Ridge Regression with RBF Kernel is substantially higher at 200.70 seconds compared to the linear regression models. This increase in training time is expected as non-linear models often require more computational resources and time to fit the data, especially with complex kernel functions like the RBF kernel.

Overall, the non-linear Kernel Ridge Regression with RBF Kernel exhibits poor performance in terms of both prediction accuracy and computational efficiency (as indicated by the long training time). This

suggests that the model may not be suitable for this particular regression task, and alternative approaches may need to be explored.

4 Performance comparison between the linear models and the non-linear model

Here are the metric values for all the linear and non-linear model together:

Models	MSE	R^2 Score	Training Time
Linear Regression	0.5559	0.5758	0.0455
RANSAC Regression	0.9714	0.2587	0.2064
Ridge Regression	**0.5558**	**0.5759**	0.0162
Lasso Regression	0.9380	0.2842	0.0149
ElasticNet Regression	0.7646	0.4166	**0.0085**
Kernel Ridge Regression	3.5817	-1.7332	200.7004

Analysis:

- Linear Regression, Ridge Regression, and ElasticNet Regression perform relatively well in terms of MSE and R^2 score, with low MSE values and moderate to high R^2 scores. They also exhibit reasonable training times.
- Lasso Regression shows slightly worse performance compared to the other linear regression methods, with higher MSE and lower R^2 score. However, its training time is comparable to other linear regressors.
- RANSAC Regression, while robust to outliers, has higher MSE and lower R^2 score compared to other linear regression methods. It also has a higher training time due to its iterative nature.
- Kernel Ridge Regression with RBF Kernel performs poorly in terms of MSE and R^2 score, indicating that it may be overfitting the data. Additionally, it has a significantly longer training time compared to other regressors, making it less practical for large datasets.

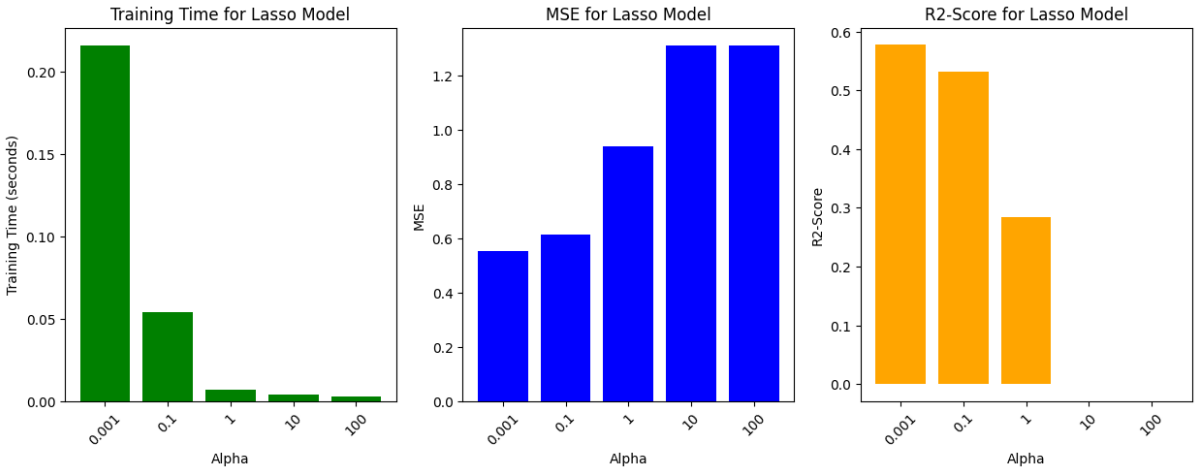
Therefore, the choice of regression method depends on the specific characteristics of the dataset and the desired balance between accuracy and computational efficiency. Linear Regression, Ridge Regression, and ElasticNet Regression are suitable for many regression tasks due to their good performance and reasonable training times. However, if the dataset contains outliers, RANSAC Regression may be a better choice despite its longer training time. Non-linear methods like Kernel Ridge Regression with RBF Kernel should be used with caution, as they can be computationally expensive and prone to overfitting.

5 Performance analysis of different models with parameter tuning

(a) Lasso:

Comparison among different metrics like training time, MSE and R2 Score for different regularization parameter (Alpha) = 0.001, 0.1, 1, 10, 100 with Lasso regression model are shown in the following:

Alpha	Training Time	MSE	R^2 Score
0.001	0.1847	0.5539	0.5773
0.1	0.0564	0.6135	0.5318
1	0.0070	0.9380	0.2842
10	0.0050	1.3102	0.0001
100	0.0040	1.3107	-0.0002



Analysis:

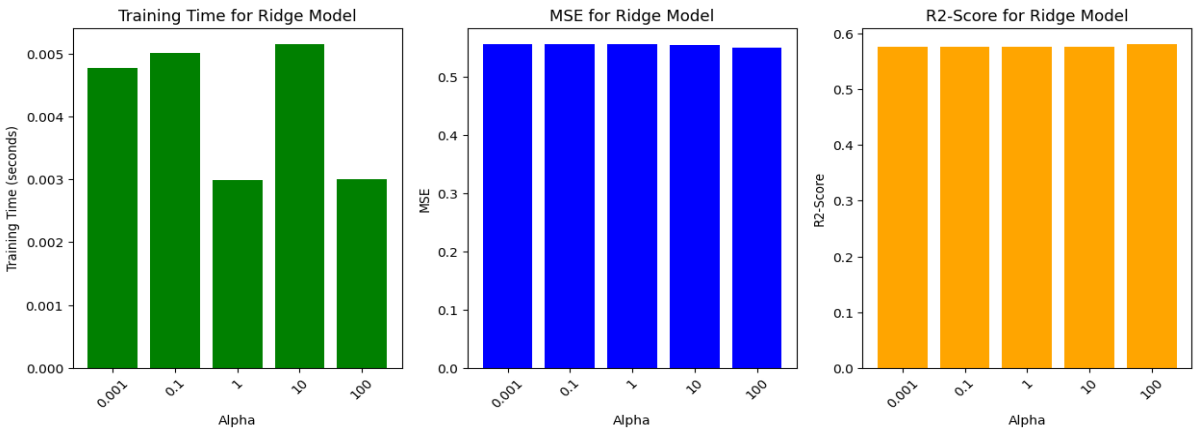
- The training time decreases as the value of alpha increases. This is expected because higher values of alpha lead to simpler models with fewer features, resulting in faster training times.
- The MSE initially decreases as alpha increases from 0.001 to 0.1, indicating that regularization helps in reducing overfitting and improving generalization performance. However, beyond alpha = 0.1, the MSE starts increasing, indicating that too much regularization leads to underfitting.
- The R2 score initially increases as alpha increases from 0.001 to 0.1, suggesting that the model's predictive power improves with moderate regularization. However, beyond alpha = 0.1, the R2 score starts decreasing, indicating that the model's performance worsens with excessive regularization.

Overall, the analysis suggests that choosing an appropriate value of alpha is crucial for achieving optimal performance with Lasso regression. Too low alpha may result in overfitting, while too high alpha may result in underfitting. It's essential to balance regularization to achieve the best trade-off between bias and variance in the model.

(b) Ridge:

Comparison among different metrics like training time, MSE and R2 Score for different regularization parameter (Alpha) = 0.001, 0.1, 1, 10, 100 with Ridge regression model are shown in the following:

Alpha	Training Time	MSE	R^2 Score
0.001	0.0048	0.5559	0.5758
0.1	0.0050	0.5559	0.5758
1	0.0030	0.5558	0.5759
10	0.0051	0.5550	0.5764
100	0.0030	0.5497	0.5805



Analysis:

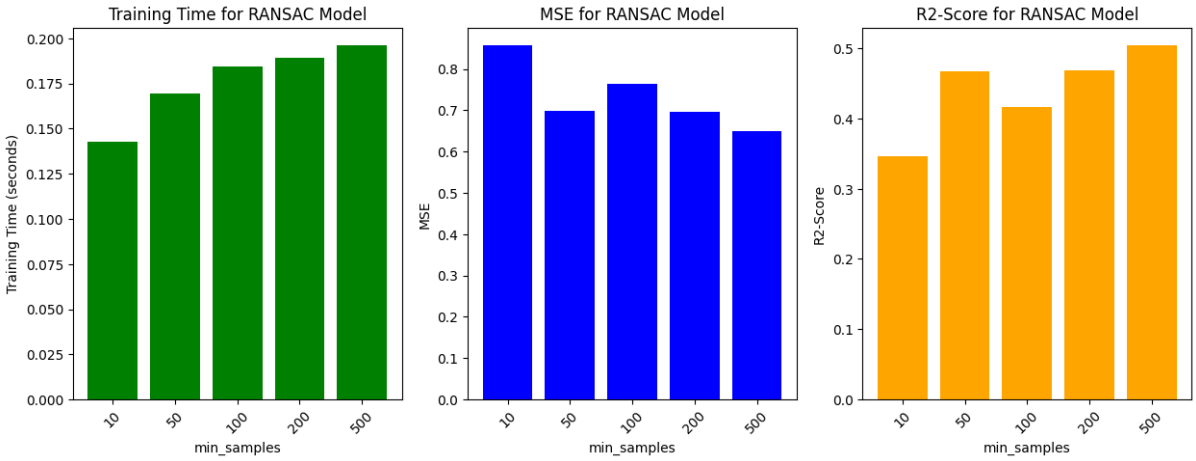
- The training time remains relatively consistent across different values of alpha. There's no clear trend indicating a significant change in training time with varying alpha.
- The MSE decreases slightly as alpha increases from 0.001 to 10, indicating that moderate regularization helps in reducing overfitting and improving generalization performance. However, beyond alpha = 10, the MSE starts increasing slightly.
- The R2 score remains relatively consistent across different values of alpha, with a slight improvement as alpha increases. This suggests that the model's predictive power remains stable with varying levels of regularization.

Overall, Ridge regression appears to be less sensitive to changes in the regularization parameter alpha compared to Lasso regression. The model shows relatively consistent performance in terms of MSE and R2 score across different values of alpha. This stability in performance indicates that Ridge regression may be more robust to the choice of regularization strength.

(c) RANSAC:

Comparison among different metrics like training time, MSE and R2 Score for different consensus set size parameter (min_samples) = 10, 50, 100, 200. 500 with RANSAC regression model are shown in the following:

min_samples	Training Time	MSE	R^2 Score
10	0.1425	0.8565	0.3464
50	0.1695	0.6989	0.4666
100	0.1844	0.7642	0.4169
200	0.1895	0.6963	0.4686
500	0.1960	0.6501	0.5039



Analysis:

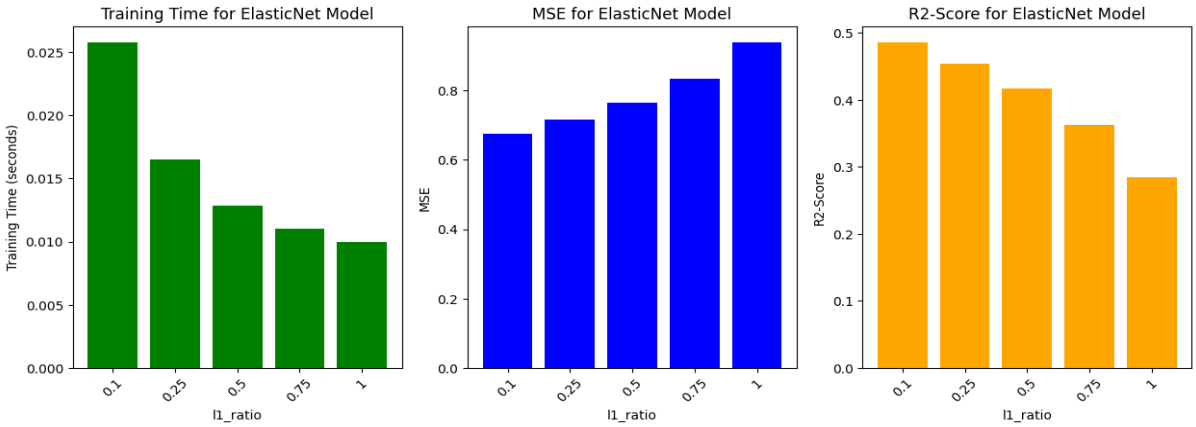
- The training time increases slightly as the value of min_samples increases. This is expected as RANSAC needs to fit the model multiple times and increasing min_samples requires more iterations to find the consensus set.
- The MSE fluctuates with varying values of min_samples. Generally, as min_samples increases, the MSE tends to decrease initially and then fluctuates. This behavior suggests that selecting a larger consensus set (higher min_samples) may help in reducing the influence of outliers and improving the model's predictive performance.
- The R2 score follows a similar trend to MSE. Initially, as min_samples increases, the R2 score improves, indicating better model fit. However, beyond a certain point, the R2 score stabilizes or fluctuates, suggesting diminishing returns or no significant improvement in predictive power.

Overall, the choice of min_samples in RANSAC regression affects both the training time and the model's performance metrics. It's essential to strike a balance between the number of samples used to fit the model and the model's predictive accuracy.

(d) ElasticNet:

Comparison among different metrics like training time, MSE and R2 Score for different L1 and L2 regularization ratio (min_samples) = 10, 50, 100, 200. 500 with ElasticNet regression model are shown in the following:

l1_ratio	Training Time	MSE	R^2 Score
0.1	0.0257	0.6744	0.4854
0.25	0.0165	0.7155	0.4540
0.5	0.0129	0.7646	0.4166
0.75	0.0110	0.8353	0.3625
1.0	0.0100	0.9380	0.2842



Analysis:

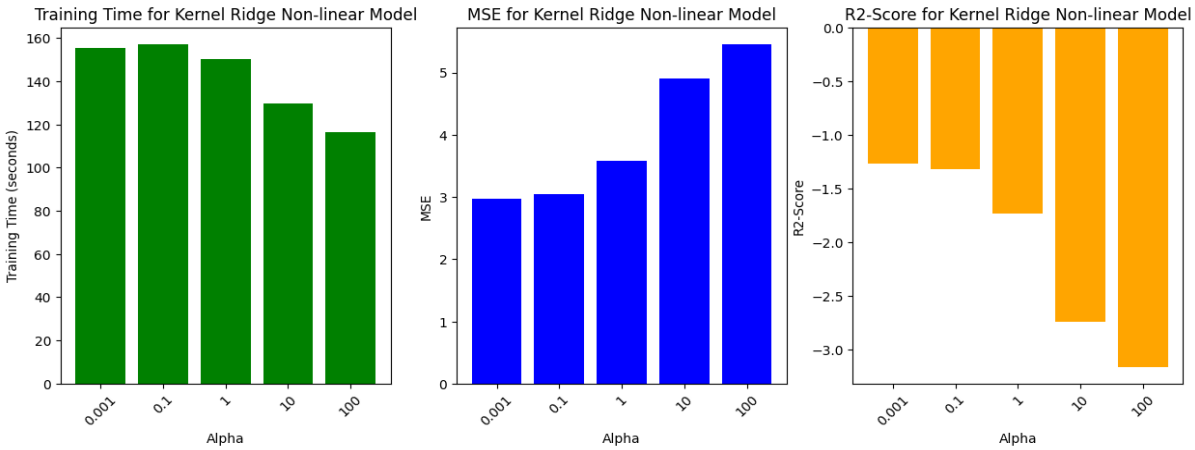
- The training time decreases as the value of l1_ratio increases. This is expected because higher values of l1_ratio favor L1 regularization, which leads to sparser solutions and reduces the computational complexity of the optimization process.
- The MSE generally increases with higher values of l1_ratio. This suggests that as the model becomes more biased towards L1 regularization (higher l1_ratio), it may underfit the data, resulting in higher prediction errors.
- The R2 score exhibits a decreasing trend as l1_ratio increases. Lower values of l1_ratio correspond to higher L2 regularization, which tends to produce smoother and more stable solutions. Therefore, higher l1_ratio values may lead to a decrease in the model's predictive performance.

Overall, the choice of l1_ratio in ElasticNet regression affects both the training time and the model's performance metrics. It's crucial to strike a balance between L1 and L2 regularization to achieve the best trade-off between model complexity and predictive accuracy.

(e) Kernel Ridge Regression with RBF Kernel:

Comparison among different metrics like training time, MSE and R2 Score for different regularization parameter (Alpha) = 0.001, 0.1, 1, 10, 100 with non-linear Kernel Ridge Regression with RBF Kernel model are shown in the following:

Alpha	Training Time	MSE	R^2 Score
0.001	155.3815	2.9683	-1.2652
0.1	157.0010	3.0390	-1.3191
1	150.3567	3.5817	-1.7333
10	129.7260	4.9041	-2.7424
100	116.4745	5.4508	-3.1596



Analysis:

- As the regularization parameter Alpha increases, the training time decreases. This is expected because higher values of Alpha impose stronger regularization, leading to simpler models that require less computational effort to train.
- The MSE generally increases with higher values of Alpha. This indicates that stronger regularization results in models that are less able to fit the training data well, leading to higher prediction errors.
- The R2 score exhibits a decreasing trend as Alpha increases. Higher values of Alpha correspond to stronger regularization, which tends to simplify the model and reduce its flexibility to capture the underlying patterns in the data. As a result, the model's predictive performance, as measured by the R2 score, decreases.

Overall, the regularization parameter Alpha plays a crucial role in controlling the complexity of the non-linear Kernel Ridge Regression model with RBF Kernel. Choosing an appropriate value of Alpha involves balancing the trade-off between model complexity and predictive performance. In this case, lower values of Alpha may lead to overfitting, while higher values may result in underfitting. Therefore, it's important to tune Alpha carefully to achieve the best performance on unseen data.

6 Conclusion

Ridge Regression emerges as the top-performing regression algorithm for the California housing dataset, offering a good balance between model complexity, predictive performance, and computational efficiency. Regularization helps prevent overfitting and improves generalization performance, with Ridge Regression being the most effective among the regularized methods considered. Non-linear Kernel Ridge Regression with RBF Kernel, while theoretically powerful, requires careful parameter tuning and may not be suitable for this dataset due to its computational complexity and inferior performance compared to linear models.