Appendix

# Glossary

* Normalised: the data is shifted so that its mean is zero and rescaled to move its standard deviation to 1. I am excluding the zero/one categorical columns from this, as it doesn’t make sense to include them.
* Linear Regression: an algorithm which attempts to fit an equation of the form Y = wX + c to data so that its error amount to the true value is minimised as far as possible
* Random Forest Regression – a collection of decision tree regressors (hence the forest) that behaves as a single model as the average prediction is returned from the collection of decision tree results. Each individual tree is trained using “bagging and random features” [1]. A random subset of features is given to each tree.
* Hyperparameter – a parameter of the machine learning model (so named as machine learning models are used to find the parameters of a model)
* Bayesian Optimisation – a function used to run through an array of different hyperparameters for a machine learning model
* Decision Tree Regression – splits the data into distinct regions with the region’s mean being the response to a prediction [2] (paraphrased)
* Bagging – using sampling with replacement from the data to create artificial training data (bootstrap samples)

# Intermediate results

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Hyper-Parameters** | **RMSE** | **MAE** |
| Linear Regression | Default | 4810.33 | 3387.65 |
| Linear Regression | Default (columns: model, year, mileage, fuel type, MPG, engine size) | 4810.35 | 3387.66 |
| Linear Regression | Default (columns: model, year, mileage, fuel type, MPG, engine size), Normalised | 2376.54 | 1706.78 |
| Linear Regression | 'Lambda', 0.0002972, ...  'Learner', 'leastsquares', 'Regularization', 'ridge', ...  'Solver', 'bfgs'  (columns: model, year, mileage, fuel type, MPG, engine size), Normalised | 1799.95 | 1364.07 |
| Linear Regression | 'Lambda', 0.0002972, ...  'Learner', 'leastsquares', 'Regularization', 'ridge', ...  'Solver', 'bfgs'  (columns: model, year, mileage, fuel type, MPG, engine size), Normalised, with validation data | 1809.63 | 1365.87 |
| Linear Regression | 'Lambda', 0.0002972, ...  'Learner', 'leastsquares', 'Regularization', 'ridge', ...  'Solver', 'bfgs'  (columns: model, year, mileage, fuel type, MPG, engine size), Normalised, with validation data. Hybrid MPG changed from 201.8 to 50.3 | 1781.80 | 1355.62 |
| Linear Regression | 'Lambda', 0.0002972, ...  'Learner', 'leastsquares', 'Regularization', 'ridge', ...  'Solver', 'bfgs'  (columns: model, year, mileage, fuel type, MPG, engine size), Normalised | 1799.9457 | 1364.072 |
| Linear Regression | 'Lambda', 0. 000010015, ...  'Learner', 'leastsquares', 'Regularization', 'ridge', ...  'Solver', 'bfgs'  (columns: model, year, mileage, fuel type, MPG, engine size), Normalised | 1803.4871 | 1364.3257 |
| Random Forest | Default (columns: model, year, mileage, fuel type, MPG, engine size) | 1205.83 | 867.67 |
| Random Forest | Default (columns: model, year, mileage, fuel type, MPG, engine size). Hybrid MPG changed from 201.8 to 50.3 | 1264.37 | 892.84 |
| Random Forest | MinLeafSize, 31, Method, LSBoost, NumLearningCycles, 193, LearnRate 0.38119. (columns: model, year, mileage, fuel type, MPG, engine size). | 1150.40 | 835.21 |
| Random Forest | MinLeafSize, 31, Method, Bag, NumLearningCycles, 193, (columns: model, year, mileage, fuel type, MPG, engine size). | 1340.52 | 965.67 |
| Random Forest | MinLeafSize, 1, Method, Bag, NumLearningCycles, 499, (columns: model, year, mileage, fuel type, MPG, engine size). | 1113.79 | 821.45 |

# Implementation Details

For the linear regression I am excluding tax and transmission type columns and normalising the data. By normalising the data, the residuals then jumped into a proper normal distribution (before they were rather significantly more on the negative side). There are still many residuals to the left, giving the histogram a very slight right skew (in picture NormalisedLRResidualPlot.jpg in case the final plot is different).

Using the script OptimizeLinearRegression I got the general hyperparameters for the Linear Regression model. I then tried using Bayesian optimisation to find the lambda, but that figure performed worse than that found by the auto-tuning (though not by much).

# References

1. Schapire Robert (2001) *Random Forests*, *Random Forests*. Available at: <https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf> (Accessed: 1 December 2021).
2. James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert. *An Introduction to Statistical Learning with Applications in R*. 2nd ed. Vol. 1. 1 vols. Chapter 8: Springer Science+Business Media LLC, 2021. <https://ebookcentral.proquest.com/lib/city/reader.action?docID=6686746>.