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

# 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 |

# 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).