Automobile Selector: Analysis on automobile attributes to demonstrate relationships between features to help comparison between vehicle models

Priyanka Annapureddy*, Rani Sebastian**, Hsiaoan Wang***

*priyanka.annapureddy@marquette.edu, **rani.sebastian@marquette.edu,
***Hsiaoan.wang@marquette.edu

Abstract

Most of the first-time vehicle buyers struggle to find a best vehicle model they desire as the technical specifications behind automobiles can be quite overwhelming. Fuel consumption is one of the most common and critical factors that people tend to focus on while narrowing down their vehicle choices. We are presenting here a study of the several attributes of a vehicle and their relationship with fuel consumption. Exploratory data analysis captures and clarifies these relationships undeniably. The machine learning methods we used here are Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression and Ensemble by Average. We found that Ensemble by Average offers us the best predictions with an accuracy of 85.6% and that regularization tends to improve the performance.

1 Introduction

Shopping for an automobile has proven to be a challenging activity for many due to the sheer volume of options we have now. Comparing fuel consumption (mpg, or miles per gallon) is the most common tactic that buyers resort to for determining which vehicle to settle for. However, we want to highlight the fact that the fuel consumption depends on other attributes of the vehicle too, and that understanding the relationships between the attributes will help one understand their vehicle better. The mpg of the automobile need not be the sole attribute a layman should consider when making such a long-term investment. Therefore, we will be assisting the buyers in better understanding their automobile choices by helping them consider other attributes of the vehicle and by clarifying the multiple relationships between these variables. We will use the vehicle model information (model year & origin) to perform a comprehensive analysis on the relationship between the fuel consumption of the vehicle and other attributes such as horse power, acceleration speed, the number of cylinders, weight and displacement of the vehicle to help the end user identify the dream model they should invest in based on their needs.

2 Related Works

In his paper ‘K-means clustering’, Prof A, John tries to identify natural clusters in the mpg dataset and states that “*there are too many variables that separately affect the fuel economy of a vehicle*” [1] This is in sync with our goal to prove that there is a significant enough relationship between mpg and other attributes, and that based on one’s needs (single attributes or a combination) we can narrow down the models that offers better mpg. The dataset was also used in another paper ‘Combining Instance-based and model-based learning’ published by J. R Quinlan. He explored combining the two types of learning methods to see if we can get better results than by using one method alone. He goes on to show that a combination of instance and model-based learning does in-fact offer better results. [2]

A separate paper ‘Application of Neuro-Fuzzy method for prediction of vehicle fuel consumption’ written by Ramadoni Syahputra offers an informative read with a different methodology. The paper states “*predicting motor vehicle fuel consumption has become a strategic issue, because it is not only related to the issue of availability of fuel but also the problem of the environmental impact caused.*” [3] The lessons to take away remain the same, that mpg has strong ties with the other attributes.

3 Dataset

The dataset that we will be performing the analysis is the Auto-MPG dataset [4] taken from the StatLib library maintained by Carnegie Mellon University. The dataset itself is now publicly available and can be found in the UCI machine learning repository. It contains a total of 398 instances of various vehicles from various manufacturers. Each data instance consists of nine different attributes, namely –*MPG* (miles per gallon or the fuel consumption), *cylinder* count, *displacement, horsepower, weight, acceleration, model year, origin* and *car name*. The attributes are either continuous, multi-valued discrete or are strings. The dataset also contains same model vehicles but manufactured in a different year. The data is between the years 1970 through 1982. Although it has plentiful information regarding car manufacturer and attributes, the dataset comes with some missing values which may require some cleaning and replacement work.

4 Methods

4.1 Exploratory Data Analysis

We began with exploratory data analysis, allowing us to get familiar with the data and find out which variables have a strong correlation with mpg that will be good predictors for our model.

Investigating on *mpg*, we found that mpg has clear linear relationships with *horsepower, weight, displacement* (inversely proportional) and *acceleration* (directly proportional). That is, as horsepower or weight or displacement increase, the mpg will reduce but acceleration increases with mpg. These linear relationships are clearly visualized in the plot shown in Figure 1.1.

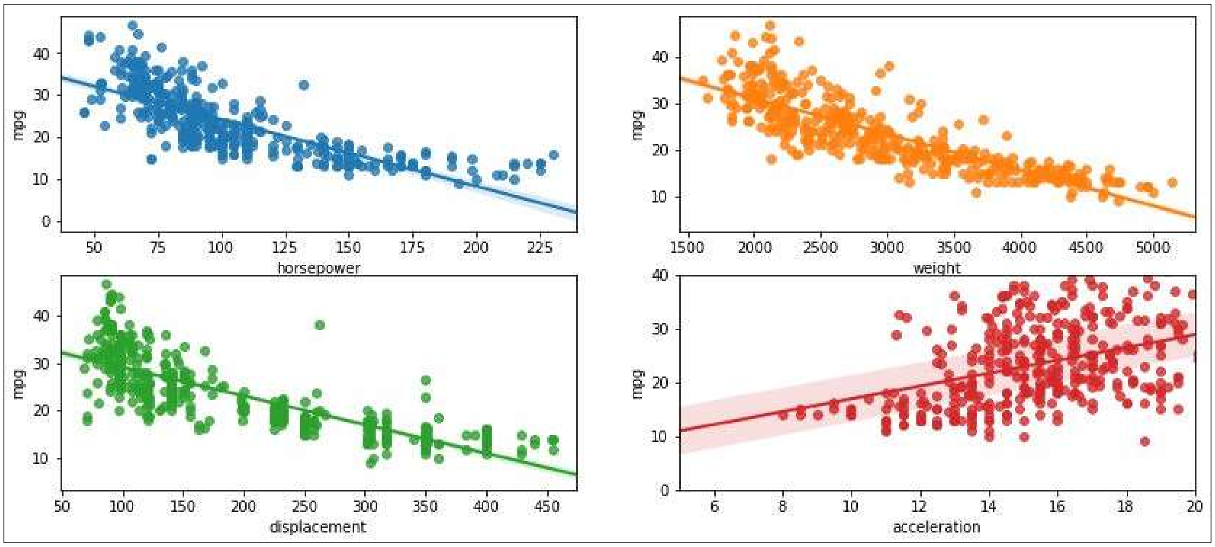


Figure 1.1 Comparison of mpg against various attributes.

The input variables themselves have strong linear relationships between themselves. Figure 1.2 highlights this –as horsepower increases, the weight tends to increase, making the vehicle heavier. This also proves that horsepower is linearly related to weight, so we used this information to impute the missing values of *horsepower* using *weight*. From this figure we can also infer that displacement and cylinder count also increases when horsepower/weight is increased.

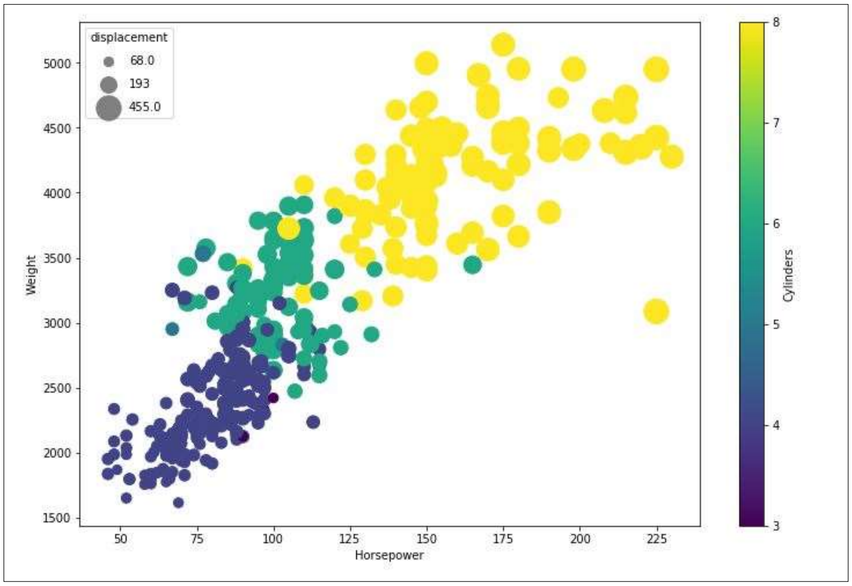


Figure 1.2 The relationships between cylinders, horsepower, weight and displacement

The number of *cylinders* have a clear impact on mpg too. Our data shows us that increasing the number of cylinders may give the vehicle more power, but it does not result in a better mpg value. The ideal cylinder count for an optimal mpg value is found to be 4 as shown in Figure 1.3.

Origin is another interesting field shown to have some correlation with mpg. Origin is 1 for USA, 2 for Europe and 3 for Japan – we find that US cars are the worst among the three and Japanese automobiles have the best mpg to offer among the three. (Figure 1.3)

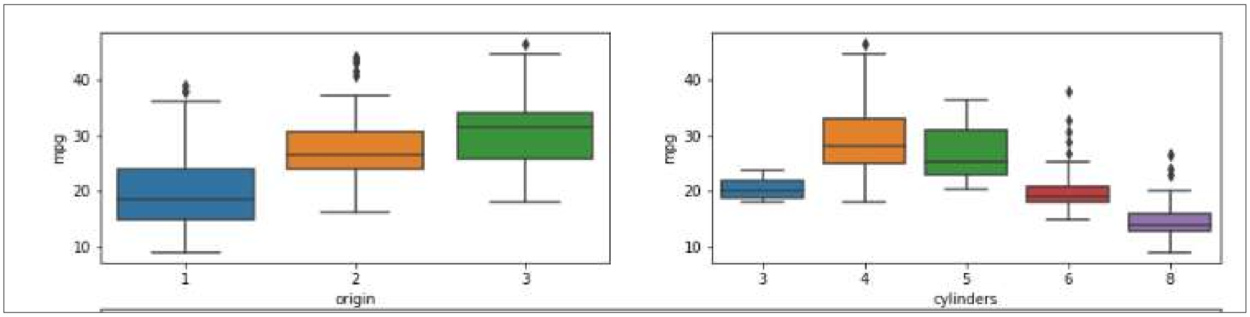


Figure 1.3 mpg VS origin, cylinders

The *model year* highlights another trend of the automobiles in general – average mpg values of the later model year (80s) vehicles are significantly higher than the mpg values of the former years (70s). These trends are shown in the Figure 1.4. It also shows the various brand names.

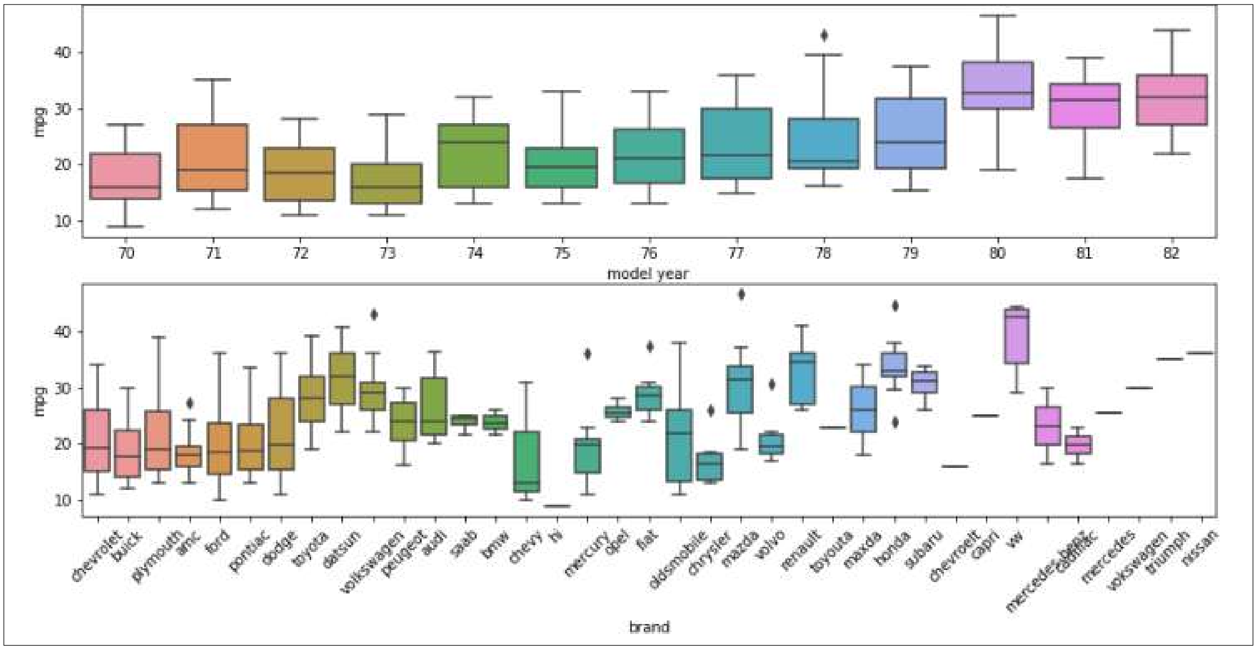


Figure 1.4 mpg VS model year and brand

The correlation between all the variables can be summed up with a heat map (Figure 1.5) and with a pair plot (Figure 1.6):

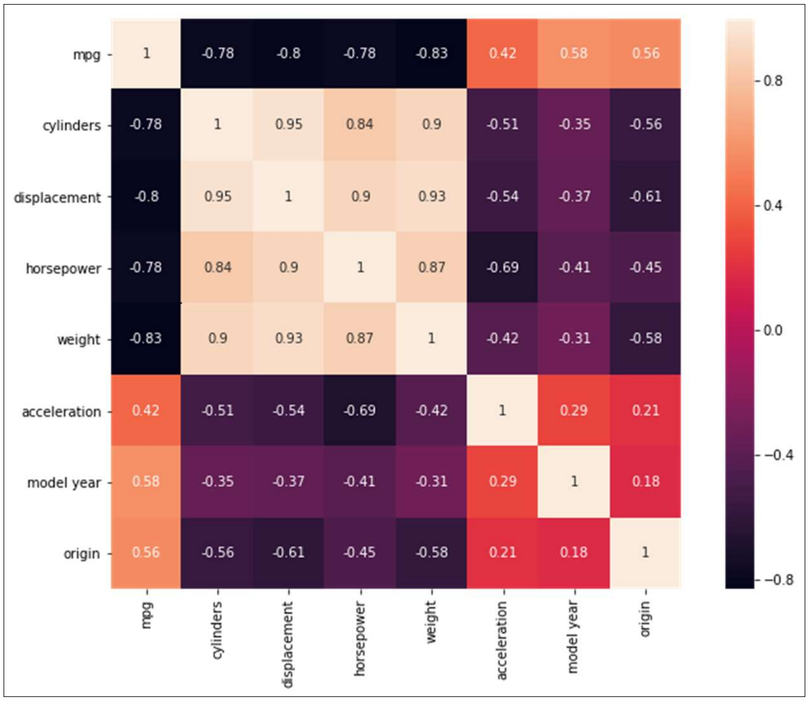


Figure 1.5 Heat Map

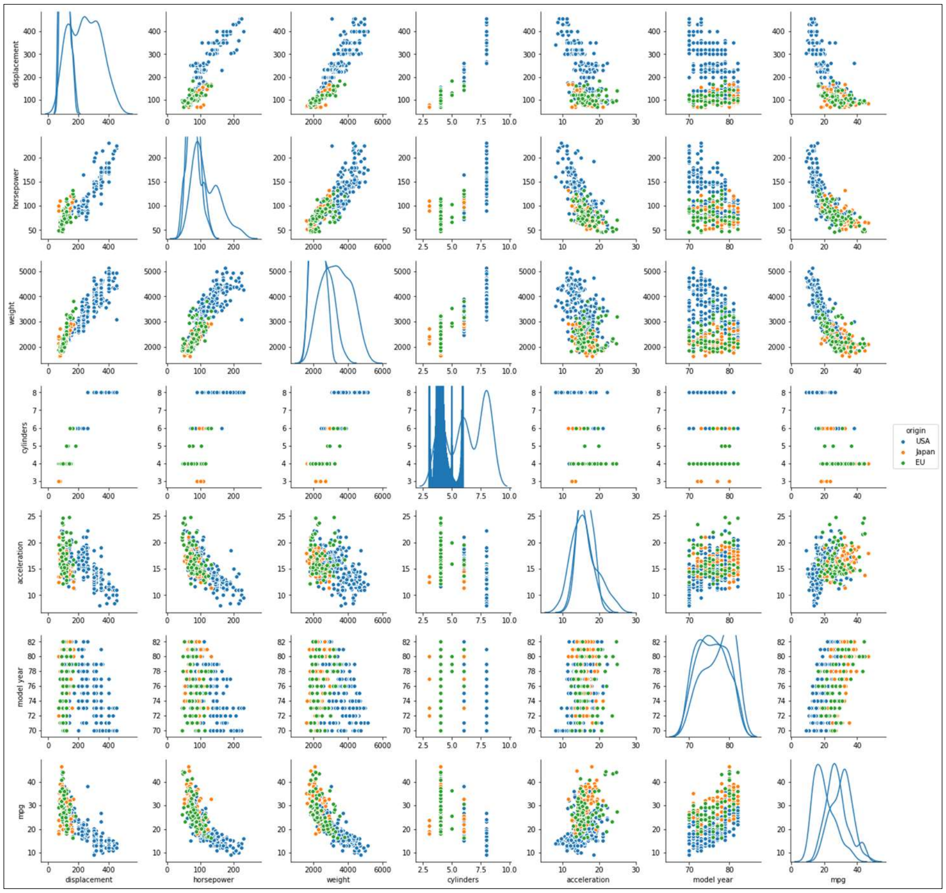
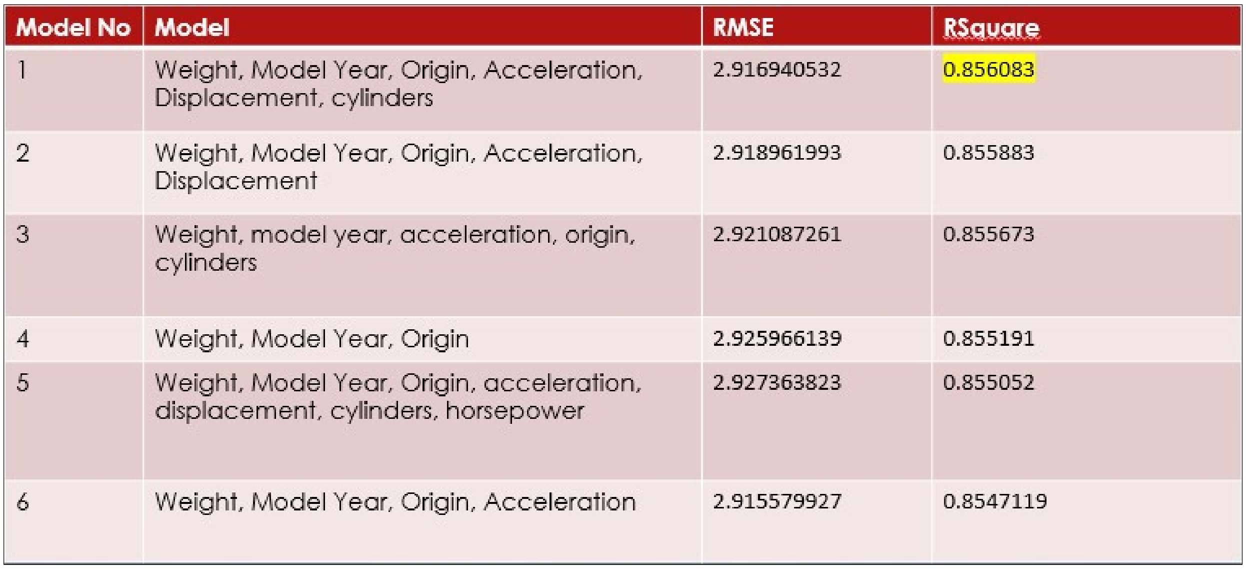


Figure 1.6 mpg Pair plot

4.2 Linear Regression

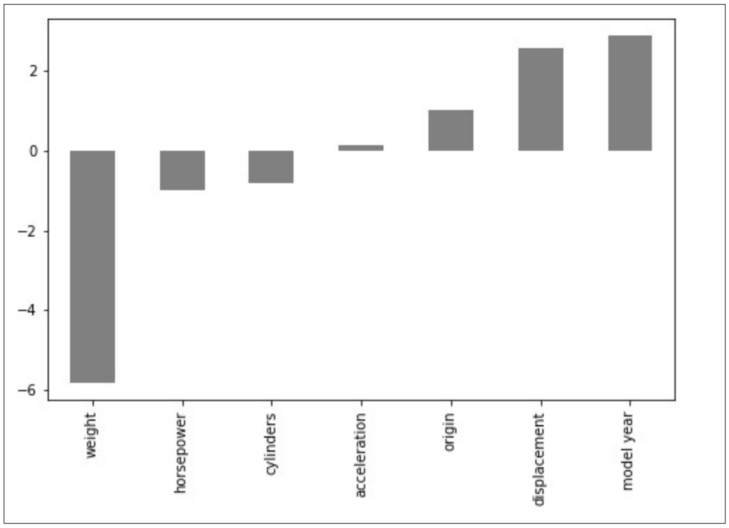
Linear Regression establishes a linear relationship between dependent variable (Y) and one or more independent variables (X1, X2, ... Xn). Y for us is mpg and X1, X2 etc. are our predictor variables. We started Linear Regression with Forward Feature Selection using RMSE and R-square as comparison metrics. We began with individual variables and selected the best from the lot (highest R-squared value). We then proceeded to combine the best variable from the first round with the other variables to produce a new model with better values for RMSE and R-Square. We had to go through seven rounds of feature selection until we were left with the final model. However, the best model we arrived at after performing linear regression was not the all-inclusive 7th model but rather the model that contained all the variables except the horse power. The top six linear regression models from forward feature selection is as shown in Table 1 (sorted based on R-squared).

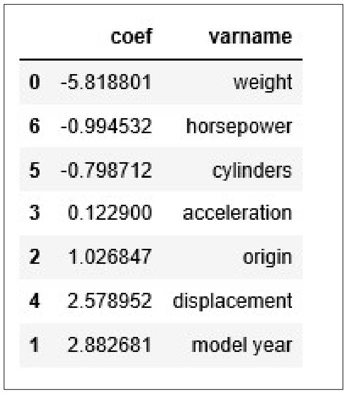


Model No	Model	RMSE	RSquare
1	Weight, Model Year, Origin, Acceleration, Displacement, cylinders	2.916940532	0.856083
2	Weight, Model Year, Origin, Acceleration, Displacement	2.918961993	0.855883
3	Weight, model year, acceleration, origin, cylinders	2.921087261	0.855673
4	Weight, Model Year, Origin	2.925966139	0.855191
5	Weight, Model Year, Origin, acceleration, displacement, cylinders, horsepower	2.927363823	0.855052
6	Weight, Model Year, Origin, Acceleration	2.915579927	0.8547119

Table 1 Top 6 models from Forward feature selection

We also plotted the coefficients to understand the weightage of each variable. Figure 2.a graphically illustrates the coefficients of each of the variables and figure 2.b highlights the numeric weights of the coefficients.





	coef	varname
0	-5.818801	weight
6	-0.994532	horsepower
5	-0.798712	cylinders
3	0.122900	acceleration
2	1.026847	origin
4	2.578952	displacement
1	2.882681	model year

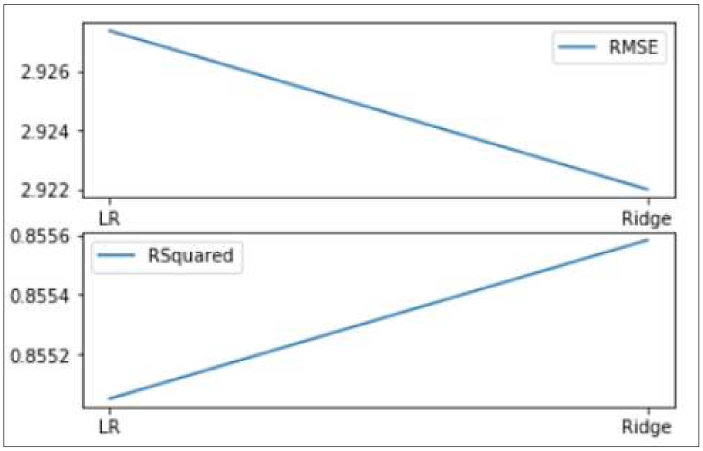
Figure 2.a Variable Coefficients

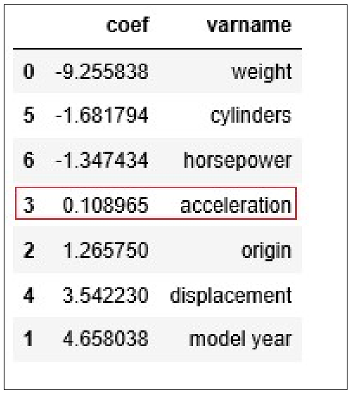
Figure 2.b Coefficient Weights

4.3 Ridge Regression

The second method we tried was Ridge Regression. Ridge Regression is a regularization technique used when the data suffers from multicollinearity (independent variables are highly correlated) and may result in overfitting. In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge reduces the standard errors. Ridge performs L2 regularization, i.e. adds a penalty equal to square of the magnitude of coefficients. [5]

After applying the Ridge regression, we got marginally better R-squared values (0.8555 compared to 0.8550 from Linear Regression model that includes all 7 variables), and this is highlighted in fig 3.a. An important feature to observe here is that ridge tends to minimize the impact of irrelevant variables, but it will not shrink them to zero. This can be seen by comparing the coefficients of acceleration we got from linear regression (Figure 3.b):





	coef	varname
0	-9.255838	weight
5	-1.681794	cylinders
6	-1.347434	horsepower
3	0.108965	acceleration
2	1.265750	origin
4	3.542230	displacement
1	4.658038	model year

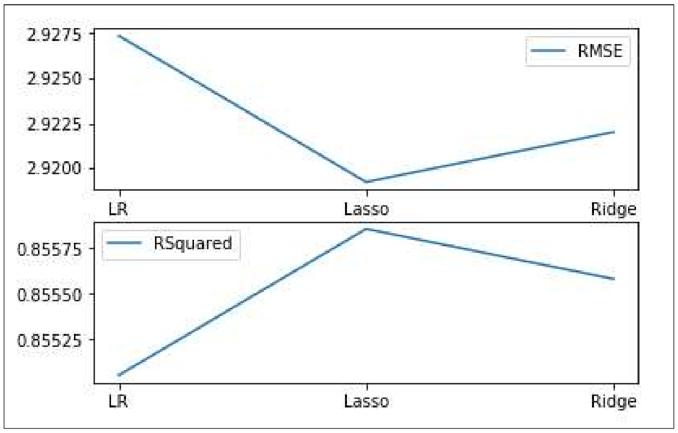
Figure 3.a LR vs Ridge

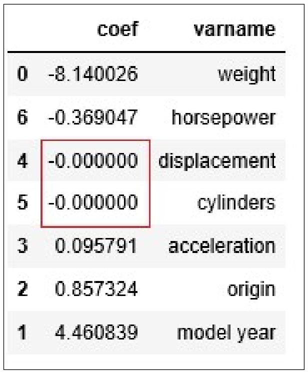
Figure 3.b Ridge Coefficients

4.4 Lasso Regression

We tried Lasso next. Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. Lasso, as opposed to Ridge, performs a L1 regularization, and adds sum of the absolute value of coefficients to introduce the bias. [5]

Lasso gives us even better R-squared values even though the difference is again only marginal (R-squared 0.8558 compared to 0.8550 from Linear Regression). The comparison of Lasso with LR & Ridge is highlighted in fig 4.a. We saw Ridge only minimizes the impact of irrelevant variables, but Lasso has the drawback of shrinking them to a complete zero (which is equivalent to removing them from the model) resulting in loss of some information. The coefficient values as shown in figure 4.b highlights this:





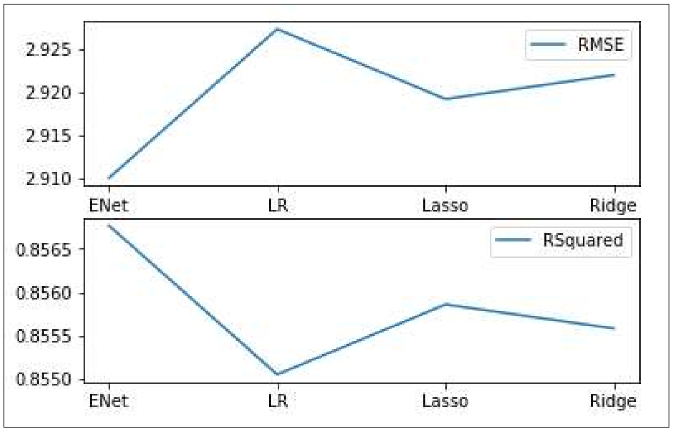
	coef	varname
0	-8.140026	weight
6	-0.369047	horsepower
4	-0.000000	displacement
5	-0.000000	cylinders
3	0.095791	acceleration
2	0.857324	origin
1	4.460839	model year

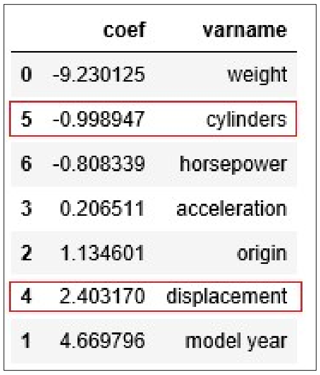
Figure 4.a LR vs Ridge vs Lasso

Figure 4.b Lasso Coefficients

4.5 ElasticNet Regression

ElasticNet Regression, or E-Net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. [6] E-Net also gives us only a marginal improvement (R-squared 0.8567 compared to 0.8550 from Linear Regression). The comparison of E-Net with Lasso, LR & Ridge is highlighted in figure 5.a. The coefficients list highlights that the variables that lasso had shrunk to zero are re-included by E-Net (figure 5.b):





	coef	varname
0	-9.230125	weight
5	-0.998947	cylinders
6	-0.808339	horsepower
3	0.206511	acceleration
2	1.134601	origin
4	2.403170	displacement
1	4.669796	model year

Figure 5.a E-Net vs LR vs Ridge vs Lasso

Figure 5.b E-Net Coefficients

4.6 Ensemble (by Average)

Ensemble uses multiple methods to obtain better predictions. Ensemble by Average is one of the many Ensemble methods. Ensemble Averaging is the process of creating multiple models and combining them to produce a desired output, as opposed to creating just one model. Frequently an ensemble of models performs better than any individual model, because the various errors of the models "average out." [7]

Ensemble by average offers us comparatively better results compared to all our previous methods (R-squared 0.8568 compared to 0.8550 from Linear Regression). The comparison of all 4 previous models and Ensemble average method is as shown in figure 6:

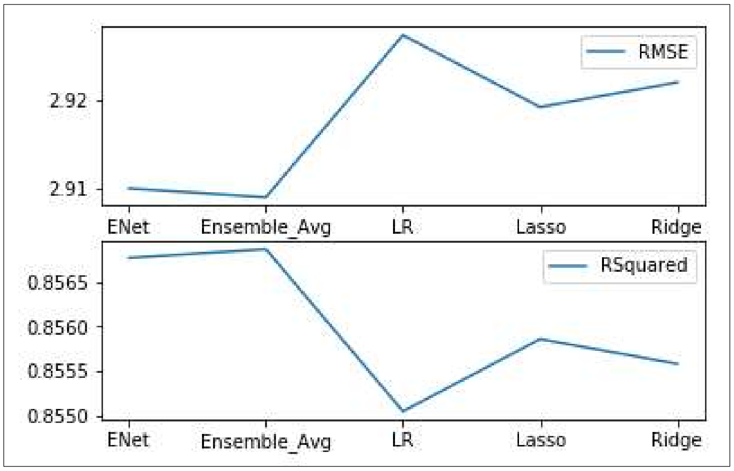


Figure 6 Ensemble Average vs E-Net vs LR vs Ridge vs Lasso

5 Results

Ensemble by Average gave us the best R-squared value of 85.6% and RMSE score of 2.90. Higher R-squared values gives the goodness of fit of the model. Fig 7 is plot of actual vs predicted mpg with actuals on x-axis and ensemble predicted values on y-axis. Graph shows that predicted values are along the regressed diagonal line between 15 to 30 and the model has under estimated the actual values between 10 to 15 and above 30.

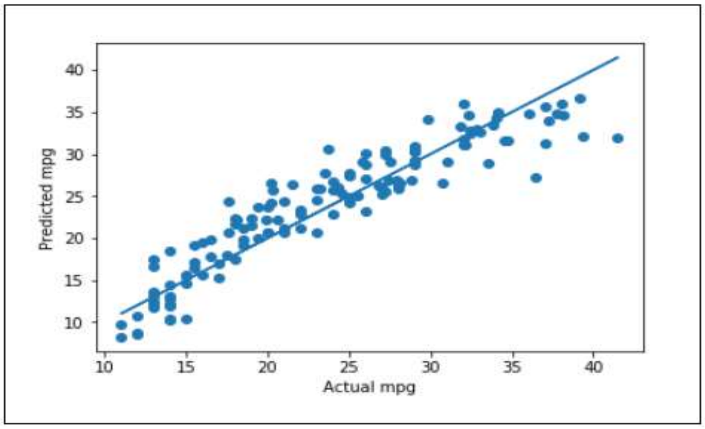


Figure 7 Predicted mpg vs Actual mpg for Ensemble by Average

6 Discussion

All implemented models performed reasonably well. After seeing initial results from exploratory data analysis that highlighted the strong linear relationship mpg has with other variables, we expected accuracy baselines ranging between 90-92%. However, multicollinearity (strong linear relationship between the individual predictors) could be the reason why we did not get an accuracy of 90% or above. Still, we can conclude that 85.6% is a fair accuracy value, if not good. Thinking about how to improve accuracies of linear models is what prompted us to explore independent topics of regularization methods using Ridge, Lasso, ElasticNet and finally Ensemble by Average.

Linear Regression (LR) offered some interesting observations for us. The best LR model (highest R-Squared) does not include all 7 predictor variables – it leaves out horsepower. The LR model that includes all 7 variables is only the fifth best. Also, displacement is inversely proportional to mpg, as proven in exploratory analysis. But LR coefficient values marks displacement as the second best positive coefficient as we can see in figures 2.a & 2.b (logically it should be a negative or a near-zero coefficient). Lasso model seems to consider this and shrinks displacement coefficient to zero (figure 4.b).

Ridge, Lasso, ElasticNet and Ensemble by Average were methodologies that were not covered in the class syllabus. However, we are glad we explored them as we now have a better understanding of data and models in general and this will aid us in our future projects immensely.

Overall, we can see that regularization improves the accuracies of the model, and therefore we can assume that if the dataset was large there is a possibility it may suffer from overfitting if no regularization is applied during training.

7 Conclusion

Certain newer datasets (which we came across later during the project) include latest vehicle models and more variables in them, like Transmission (auto or manual), Guzzler type, Air Aspiration Method, Number of Gears etc. Given more time, we would like to apply the concepts we explored in this project with these larger datasets and verify if our core observations still hold.

ACKNOWLEDGMENT

We would like to express our sincere thanks to Professor Dr. Praveen Madiraju for introducing to us the concepts of Data Mining and making it very interesting.

Appendix A: Team Member Contributions

Each team member's contribution to the project is listed below. We also actively discussed on the theory together as we worked through the models.

Priyanka: Data fetching, Exploratory Data Analysis, Comparison of Linear, Ridge & Lasso Regressions, Elastic Net Regression, Ensemble by Average, Final Report, Slides.

Rani: Mid-way report, Linear Regression, Ridge Regression, Lasso Regression, Final Report, Slides.

Justin: Mid-way report, Interpretation of data plots and attribute relationships, Final Report, Slides.

Appendix B: References

[1] K- Means Clustering Exercise – Prof Dr. John Aleshunas

[2] Combining Instance-based and model-based learning- J.R. Quinlan

Combining
Instance-Based and

[3] Application of Neuro-fuzzy method for prediction of vehicle fuel consumption - Ramadoni Syahputra

[4] https://archive.ics.uci.edu/ml/datasets/auto+mpg

[5] https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/

[6] https://en.wikipedia.org/wiki/Elastic_net_regularization

[7] https://en.wikipedia.org/wiki/Ensemble_averaging_(machine_learning)