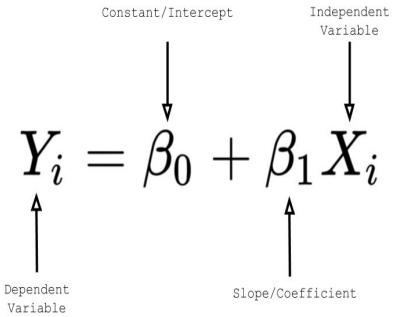
Big Data ML Capstone Project

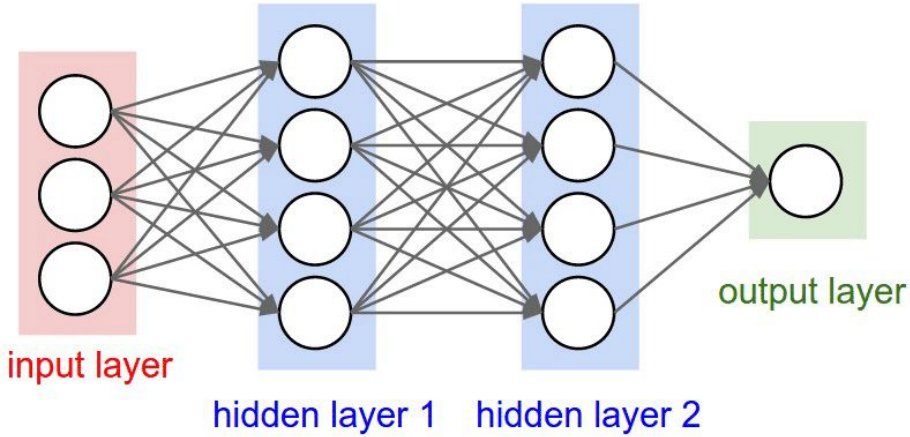
By: Shiven, John, and Zichen

Model Construction

Project - Accurate model for how the price of automobiles are connected to the horsepower and other attributes of automobile.

We used a linear regression model and a deep neural network model for loss and val loss.





Model Roadmap

Data Processing

We started by processing data into multiple columns, focusing on turning string attributes into numbers or dropping them from the dataset, then we normalized the data and started testing 4 different models

Model

We used the basic linear regression model as our baseline, and after finishing the model, we started using multiple variables, and switched to deep neural network as our final model and it improved our overall accuracy by 1,350 dollars

Evaluation

After testing all of the models 3 times, we summed their MSE averages and compared them between each other. In the end of the best model based on MSE was our Deep Neural Network with multiple variables

Test

The final, best performance that we were able to get from the Deep Neural Network with multiple variables was an MSE of 2,098 dollars, a decrease of 1,350 dollars from our Linear Regression model with one variable

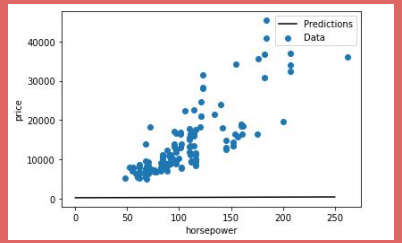
Model Settings

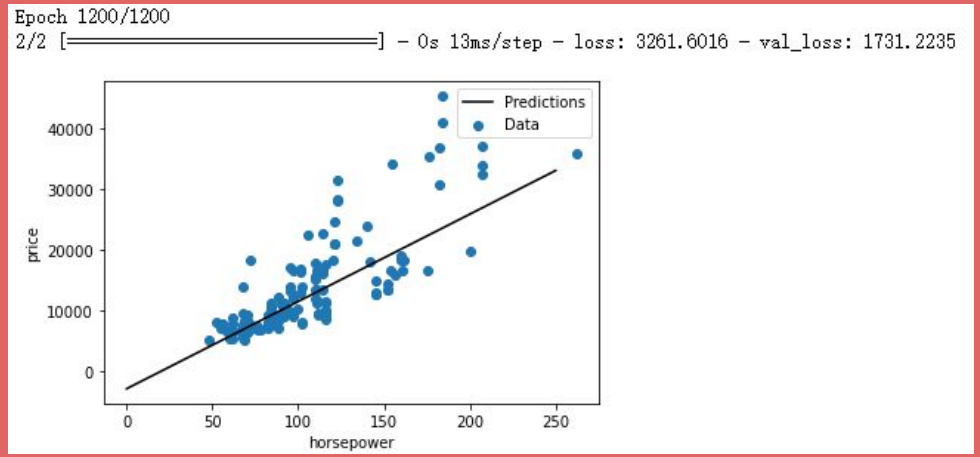
- We start by using horsepower as our independent variable, and price as our dependant variable

- Regression problems are based on the quantitative relationship between X and Y, in this case horsepower and price. The closer the model can predict Y based on X, the better it is.

- We set the learning rate to 15 because we were dealing with high prices, and used over 1000 epochs to get a result of around 3,200 MSE.

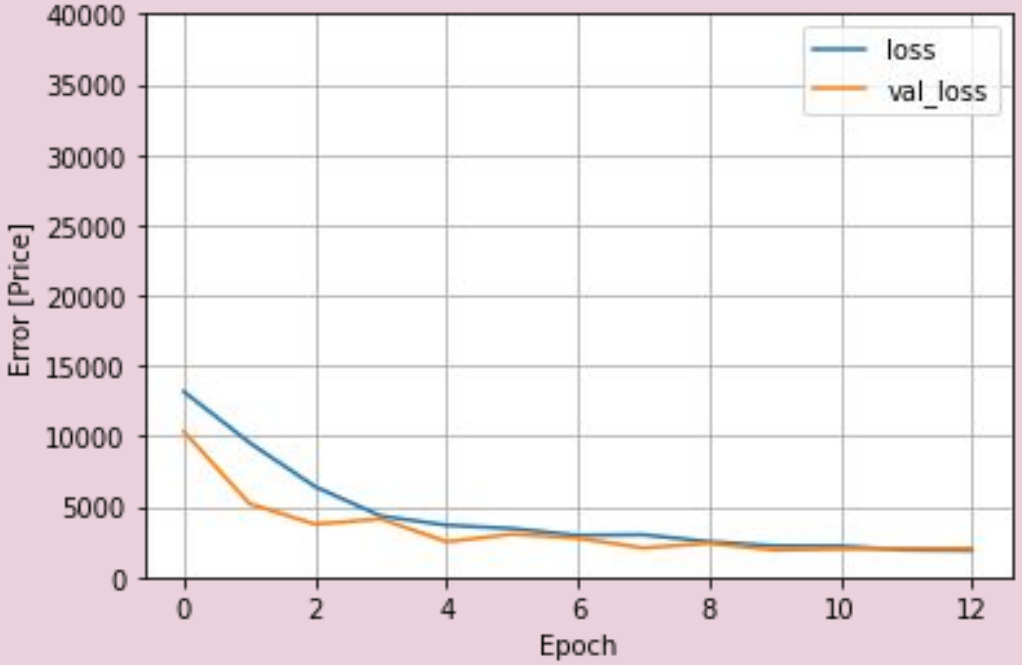
- We saw that if the epoch # × learning rate is too low, then the machine would hardly learn anything. If this value is too high then the model of overfit

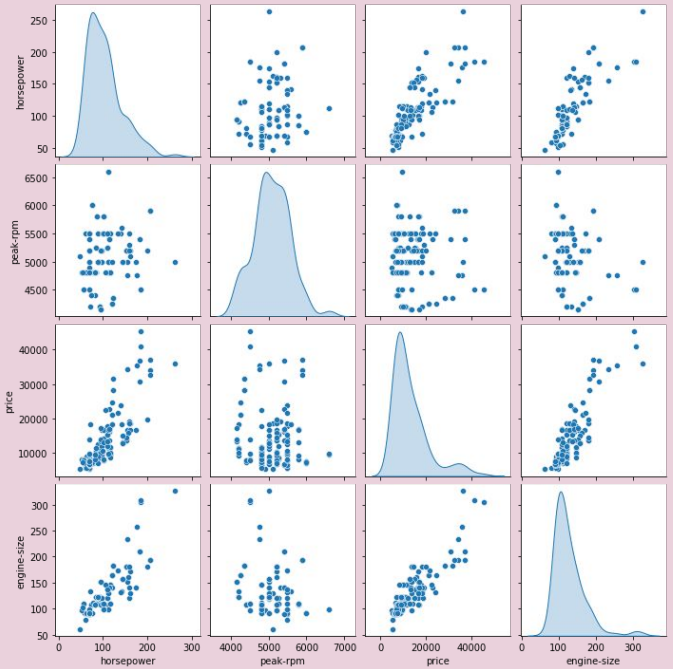




Data Visualization

	symboling	fuel-type	doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
0	3	2	2.0	5	3	1	88.6	168.8	64.1	48.8	...	130	6	3.47	2.68	9.0	111.0	5000.0	21	27	13495.0
1	3	2	2.0	5	3	1	88.6	168.8	64.1	48.8	...	130	6	3.47	2.68	9.0	111.0	5000.0	21	27	16500.0
2	1	2	2.0	4	3	1	94.5	171.2	65.5	52.4	...	152	6	2.68	3.47	9.0	154.0	5000.0	19	26	16500.0
3	2	2	4.0	3	2	1	99.8	176.6	66.2	54.3	...	109	6	3.19	3.40	10.0	102.0	5500.0	24	30	13950.0
4	2	2	4.0	3	1	1	99.4	176.6	66.4	54.3	...	136	6	3.19	3.40	8.0	115.0	5500.0	18	22	17450.0





Performance

Model	Training 1	Training 2	Training 3	Average
Linear Regression with one variable	3,262	3,257	3,258	3,259
Linear Regression with multiple variables	2,490	2,428	2,392	2,437
Deep Neural Network with one variable	3,478	3,456	3,545	3,493
Deep Neural Network with multiple variables	2,098	2,388	2,181	2,223

Conclusions

- The best model in our experiment had a means absolute error of 2,223 dollars on average when predicting the price of a car

- This is 1,036 dollars smaller than the baseline model which is linear regression model with one variable

- Based on this experiment, we have learned that as the cost of a car increases, the relationship between price and horsepower becomes smaller, and that several others attributes factor in when predicting the price of a car. We have also learned that the bigger the numbers being predicted are, the higher the learning rate and/or epoch should be for the model.