Jonathan Daughtry

Final Report:
Price Analysis of Used Cars

Problem Statement

The buying and selling of used vehicles is a major business, whether it is the individual selling of personal vehicles, or major businesses that are centered around the trade. The business has only grown in popularity online, with many companies providing a customer a way to buy or sell a vehicle with minimal human interaction. Obviously, the goal of both the buyer and seller is to get the best deal available. The seller wants to maximize profit, while the buyer wants to know they are not overpaying. However, every used vehicle can have a wide range of mileage, year, color, etc, so finding a comparable vehicle to measure price can be difficult.

Craigslist is a major classified advertisement website that includes the buying and selling of used vehicles. Using a dataset obtained from Kaggle that contained information on vehicles that were sold through Craigslist, my goal was to create a regression model that could be used to suggest an appropriate vehicle price for a user that wanted to sell their used vehicle. By reducing the Used Cars Dataset to ten features, I was able to use Random Forest modeling to achieve a random r squared score of 0.83 and a root mean squared error score of 5282. This process can be repeated by individuals or companies when selling their used vehicles.

Data Wrangling

The raw Used Cars dataset contained 423,856 rows, each a different vehicle, and 26 columns of features. It provided a wide variety of vehicles and their features, but it needed extensive cleaning. I started by deciding what features I wanted to keep that I would predict would be factors in determining the price. I also eliminated condition because it had over 40% null values. The determining variables that I kept were:

- Year - The year the vehicle was manufactured
- Odometer - The amount of miles the vehicle has been driven
- Manufacturer - The company that produced the vehicle
- Model - The specific type of vehicle
- Cylinders - How many cylinders the vehicle used
- Transmission - The type of transmission
- Drive - The type of drive
- Paint color - The color of the vehicle
- State - The state in which the vehicle was listed

I initially determined manufacturer, model, year, and odometer to be necessary variables to be included, so I eliminated any vehicles where there were null values for those. I eliminated prices that were below $100 and above $125,000 because those were outliers and very few, and kept cars with mileage between 1,000 and 150,000. After all of the data cleaning, I was left with 183,514 rows of individual vehicles with 10 columns of important features.

Exploratory Data Analysis of Continuous Variables

The following variables have these details:

- Price - The dependent variable, a continuous integer with a minimum of 101 and a maximum of 125,000.
- Odometer - A continuous integer with a minimum of 1,000 and a maximum of 150,000
- Year - A continuous integer with a minimum of 1919 and a maximum of 2020.
- Model - A categorical variable with 15,097 different categories.
- Manufacturer - A categorical variable with 42 different categories.
- Number of cylinders - A categorical variable that consists of integers. The 8 values are 3, 4, 5, 6, 8, 10, 12, and other.
- Transmission - A categorical variable with 4 values, automatic, manual, other, and unknown.
- Drive - A categorical variable with 4 values, front wheel drive, rear wheel drive, four wheel drive, and unknown.
- State - A categorical variable with 51 values. These were 50 states and Washington D.C.
- Color - A categorical variable with 13 values. These included custom and unknown.

I decided to eliminate model from consideration because of the number of categories, with many only having one of that category. I felt that manufacturer would be able to be a sufficient replacement with far fewer categories.

The two predictor variables that I was most interested in were the year and the mileage. I felt that these two variables would affect the price the most. I first decided to look at the correlation coefficients between those two and price and to plot them in relation to each other.

The coefficient between price and mileage was -0.51, which is considered a moderate negative relationship. This is predictable because vehicle value typically decreases as it is driven more. The line of best fit of this graph supports the negative correlation.

Figure 1: Scatterplot of Mileage of Car vs Price

The coefficient between price and year was 0.36, which is considered a weak positive relationship. This is also predictable because newer vehicles cost more, but it might not be as strong because older classic cars become more valuable. The graph supports this, with a positive slope, although not as steep. Observing the graph carefully, one can see a slight increase of prices for vehicles around 1970, which supports the theory that some classic vehicles gain value.

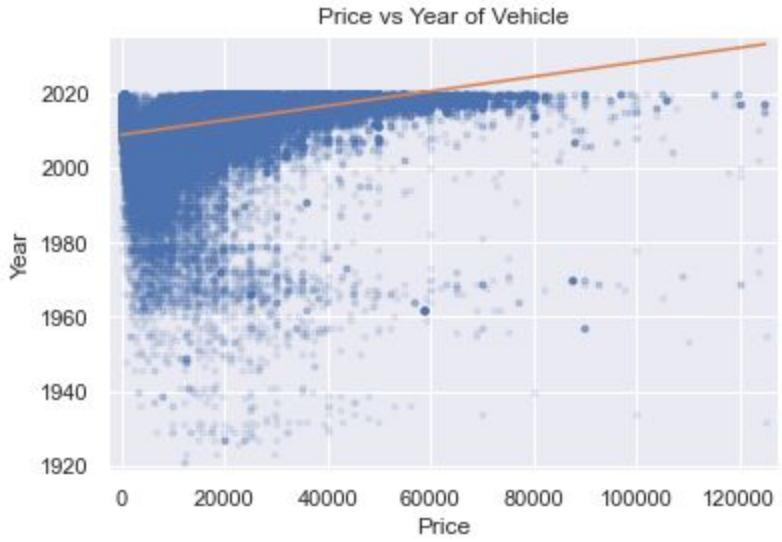

Figure 2: Scatterplot of Year Produced vs Price

Exploratory Data Analysis of Categorical Variables

The rest of the potential determining variables were all categorical, and I decided to determine if they were useful for determining price by hypothesis testing. My null hypothesis was that each one of the categorical variables would not affect price. For each determining variable, I split them by category and took their prices. I selected some of the most represented categories, then checked the p-value when their prices were compared to each other. If the p-value was below 0.05, then my null hypothesis would not be supported, which means I would use those determining variables in my modeling. The results are listed below.

Determining Variable	Categories Compared	P-value
Number of Cylinders	4 Cylinders vs 6 Cylinders	0.0
	4 Cylinders vs 8 Cylinders	0.0
	6 Cylinders vs 8 Cylinders	0.0
Transmission	Automatic vs Manual	4.86×10^{-82}
Type of Drive	Front Wheel vs Rear Wheel	0.0
	Four Wheel vs Front Wheel	0.0
	Four Wheel vs Rear Wheel	0.0
State	Texas vs California	7.04×10^{-5}
	Texas vs Florida	2.24×10^{-19}
	Ohio vs California	3.47×10^{-154}
	Ohio vs. Florida	1.72×10^{-79}
Manufacturer	Ford vs Toyota	7.65×10^{-92}
	Ford vs Nissan	0.0
	Chevrolet vs Nissan	0.0
	Chevrolet vs Toyota	4.93×10^{-12}
Color	White vs Black	1.09×10^{-26}
	White vs Silver	0.0
	Black vs Silver	1.47×10^{-233}

Table 1: P-values of Categorical Variables

All of the p-values were were well below 0.05, so all of these variables were going to be used when modeling.

Modeling

To prepare for modeling, I needed to split my data and then make dummy variables for the categorical variables. I decided to split the data into 70% for training and 30% for testing. This resulted in 128,459 vehicles that were being used for training and 55,055 for testing. Dummy variables were created for the categorical variables, and this was done separately for the training and testing sets to prevent data leakage, leaving 118 columns for each.

I decided to use three ensemble models: Random Forest Regressor, AdaBoost Regressor, and Gradient Boosting Regressor. For each of the models, I also used GridSearch cross validation on three of the hyperparameters available. For Random Forest, the hyperparameters I used were a max depth of 60, 9 max features, and 200 estimators. For Adaboost, the hyperparameters I used were a learning rate of 0.05, an exponential loss function, and 50 estimators. For the final model, Gradient Boosting, I used a learning rate of 0.25, a max depth of 10, and 200 estimators. I used four measurements to determine the effectiveness of each model: time taken to predict, the square root of the mean squared error, the mean absolute error, and the adjusted r squared. I also determined the feature importance in each model.

Random Forest Regression Model

For the Random Forest, the effectiveness measurements are given in Table 2.

Time Taken in seconds (per 100 samples)	RMSE	MAE	Adj R^2
0.009	5282	2821	0.83

Table 2: Effectiveness measurements using Random Forest Regressor

I also graphed the top features by level of importance. Figure 3 shows that the most important features were mileage and price. In the figure, “Unknown” represents the number of cylinders that had unknown values, and “Unknown.1” represents the type of drive that had unknown values. “Unknown.1” could be significant because the drive category all-wheel drive was not included, so that would be included in this classification. After year and odometer, the other features that contributed the most to the random forest model were type of drive and type of cylinders. The features began to level off more after those.

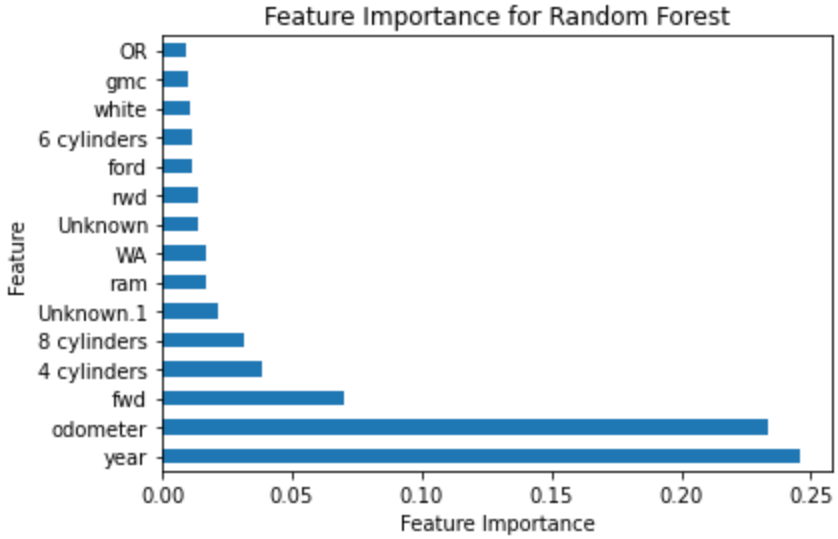

Figure 3: Feature Importance in Random Forest Regression modeling.

AdaBoost Regression Model

For the AdaBoost Regression, the effectiveness measurements are given in Table 3.

Time Taken in seconds (per 100 samples)	RMSE	MAE	Adj R^2
0.003	9372	6517	0.46

Table 3: Effectiveness measurements using AdaBoost Regressor

Figure 4 shows that, like Random Forest, the most important features were mileage and price. However, front wheel had a much greater impact on the model for Adaboost. The top 6 features were the same as Random Forest, with the only difference in order being the unknown drive and 8 cylinders. AdaBoost only emphasized 7 total features. The rest of the features were reduced to nothing.

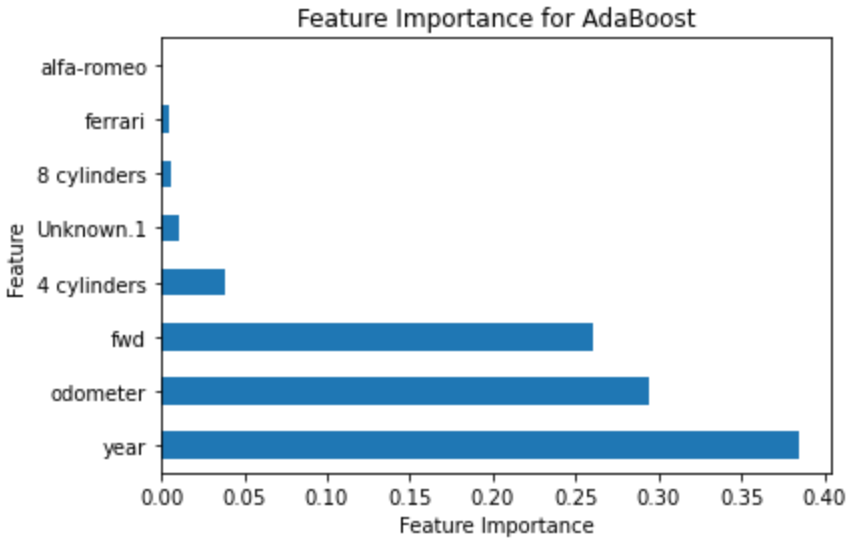

Figure 4: Feature Importance in AdaBoost Regression modeling.

Gradient Boosting Regression Model

For the final model used, Gradient Boosting Regression, the effectiveness measurements are given in Table 4.

Time Taken in seconds (per 100 samples)	RMSE	MAE	Adj R^2
0.002	5629	3355	0.81

Table 4: Effectiveness measurements using Gradient Boosting Regressor

Figure 5 shows a similar pattern with top features as the other models, with differences in emphasis on each feature. Year was more than double any other feature, with mileage and front wheel drive following. Year, mileage, drive, and number of cylinders are the most important features before the features start leveling off.

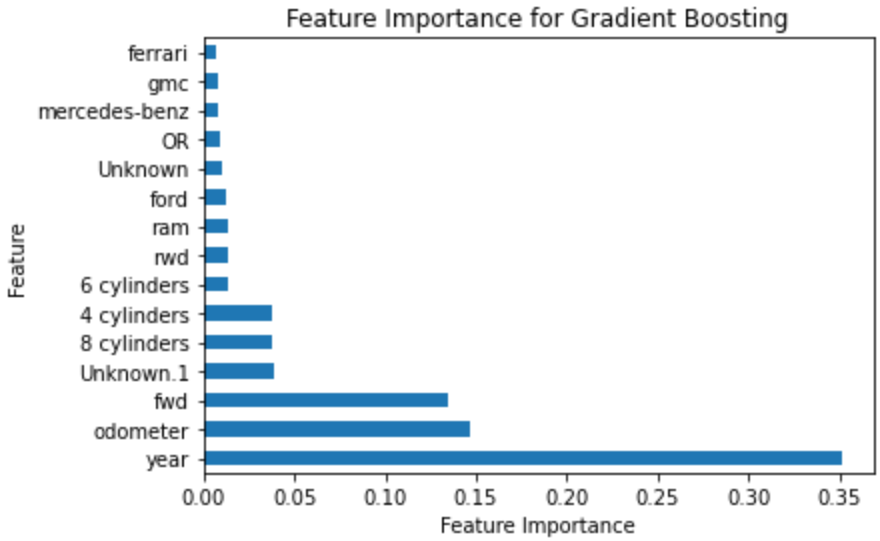

Figure 5: Feature Importance in Gradient Boosting.

Comparison of Models

I graphed the results against each other in order to get a visual comparison of how they compared.

Time

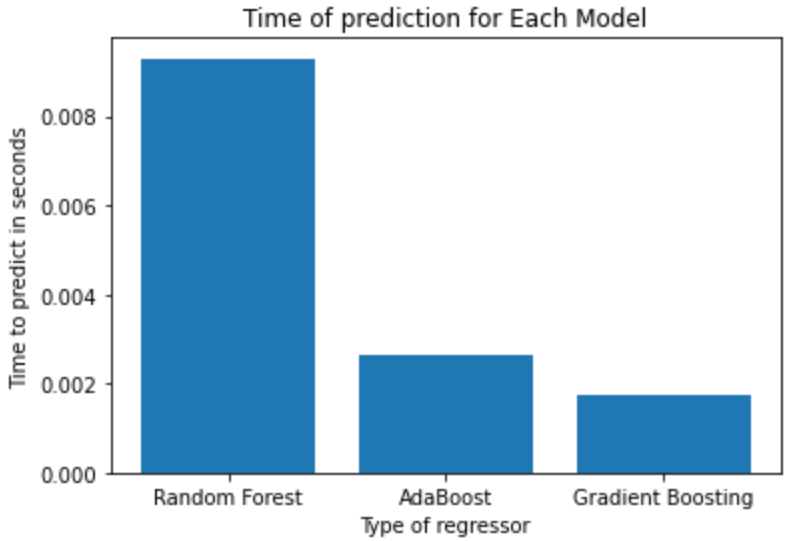

Figure 6: Time Comparison of Three Models

The most noticeable difference is between the Random Forest Model and the other two models, with Random Forest taking three times as long as the next closest model. However, this graph is a little misleading because all of the times are very low. If any of these models were being used by individuals selling a single vehicle or even a company selling a low or moderate number of vehicles, the differences in time are negligible. The number of vehicles that are being tested would have to be substantially large before the amount of time became an issue, so I considered time to be the least important in determining the best model.

Root Mean Squared Error

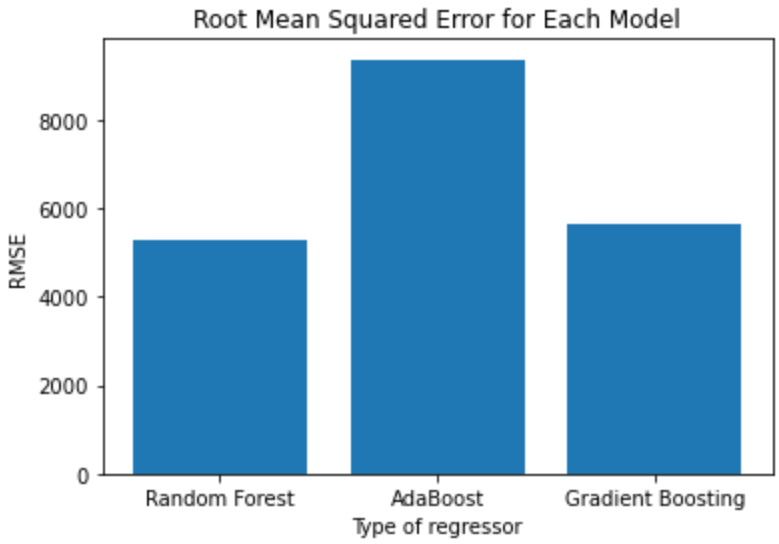

Figure 7: Root Mean Squared Error Comparison of Three Models

The root mean squared error was the first way I determined how effective the model was by measuring how its predictions differed from the observed values. Since lowest value signals a better model, according to Figure 7, Random Forest performed the best. However, Gradient Boosting was a value less than 400 more, which could be argued to not be significant. However, AdaBoost had a value of almost 4000 greater than the next highest, Gradient Boosting. I reserved judgement on whether it could be ruled out as a model I would not use until I could see how it measured up in the other measurements that were being used.

Mean Absolute Error

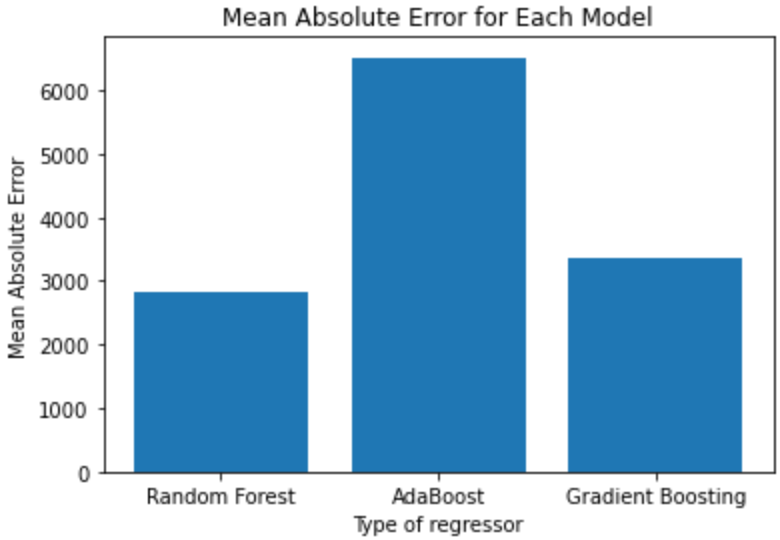

Figure 8: :Mean Absolute Error Comparison of Three Models

Mean absolute error scores similar to root mean squared error in that the lowest score shows a better model. The pattern of scores followed the same order and tendencies as root mean squared error. According to figure 8, Random Forest performed the best, followed closely by Gradient Boosting. The score difference was about 500, which again could be argued to not be significant. AdaBoost had an error of 3000 more than Gradient Boosting, and I was able to rule it out as the best model that could be used.

Adjusted r squared

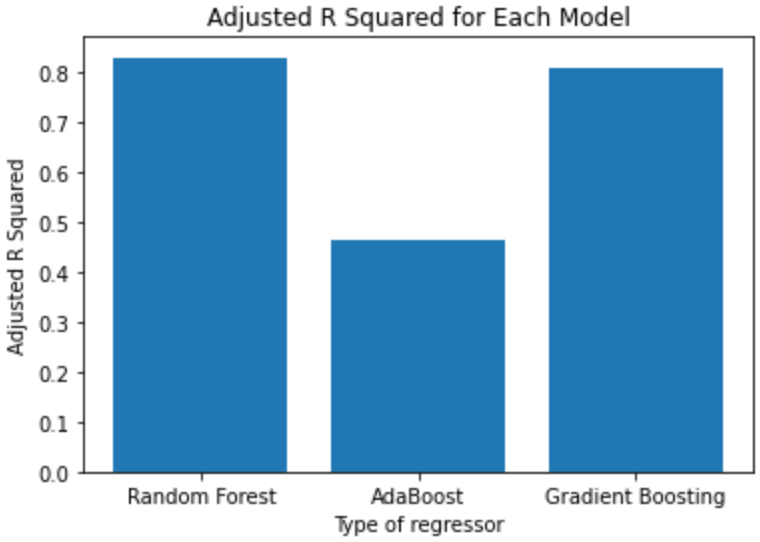

Figure 9: :Adjusted R Squared Comparison of Three Models

Unlike the previous methods used to determine effectiveness of models, models with higher values are considered better. I preferred using adjusted r squared instead of r squared because adjusted takes into account the number of features used. AdaBoost had the lowest score. Random Forest had the highest score at 0.83, but Gradient Boosting was very close with a score at 0.81. Once again, this score difference could be within the margin of error.

Takeaways

The initial takeaway is that AdaBoost consistently performed the worst of the three models that were used. Random Forest consistently performed the best, but it was followed closely in each model by Gradient Boosting. These models performed close enough that their effectiveness could be affected by further adjustments of the hyperparameters, but Random Forest was the most reliable model and would be the first model that I would suggest using.

Feature reduction could be used when modeling because the features year, mileage, number of cylinders, and type of drive were the most significant for feature importance. More specifically, 4 cylinders, 8 cylinders, front wheel drive, and unknown drive affected the models the most. However, when removing the other features completely, does have a noticeable negative effect on model performance, which was apparent in AdaBoost.

The time that each model took to predict was not significant. I determined that since buying and selling vehicles are considered significant purchases and can cost a lot of money, getting a fair price is by far more important, even if the model takes a little longer to make a prediction. Also, the time differences in the models were very small.

Future Research

The high correlation of price with mileage and year of the car was predictable. However, I should be wary of the multicollinearity of mileage and year. Naturally, as cars get older, one can predict that they will gain miles. There was a negative correlation between these in my initial statistical testing, but it was not considered significant. It might be beneficial to run models only using mileage or year, but not both, or to use models that punish multicollinearity. Also, I would consider reducing the dimensionality of the variables since four played a much more significant role in determining the models.

The models might benefit from improvement by further adjustment and expansion of their hyperparameters. For example, I adjusted by increasing the hyperparameters for Gradient Boosting in my GridSearch cross validation, and the model improved. Increasing it more would likely improve the model further. I could also add hyperparameters that I did not include initially.

I could also adjust some of the ways that I clean the dataset. In particular, I could use the model variable to make it a more useful categorical variable but classifying each model as a sedan, SUV, truck, etc. However, this would take extensive time and research to go through over 15,000 different values to classify them.

The original dataset had a variable called condition that would have potentially been useful. However, I did not use this because it had a high percentage of null values, and I did not want that to be misleading in the model. Also, that category is the most subjective of the variables, and those that are selling their car on Craigslist would be motivated to give it a better condition than it really is in order to increase value.

This dataset is continuously being updated as new vehicles are being sold. Cleaning and adding the new vehicles to the dataset periodically could include the model's performance.