BMW Used Car Pricing Model

August 8, 2021

Contents

1.) Motivation

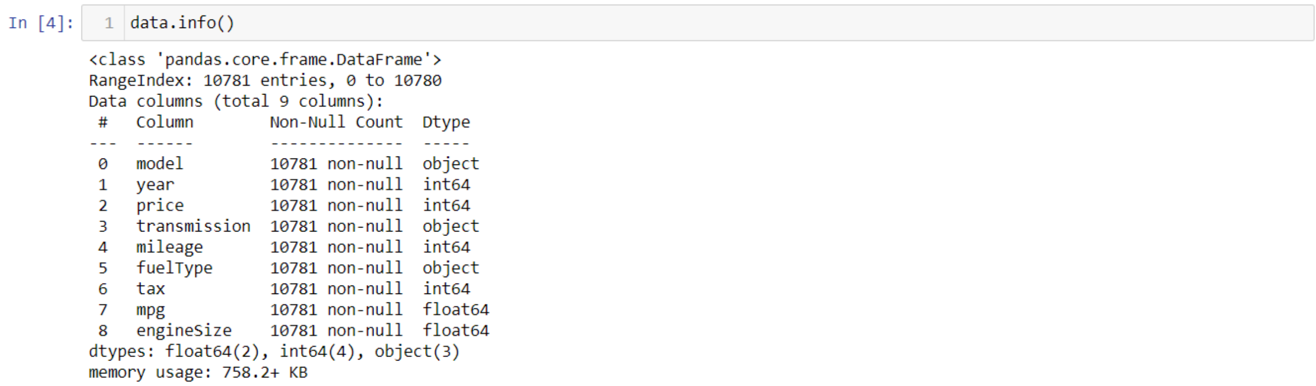
After I graduated from college, my parents decided to purchase car for me. The only criteria I had were that the car had to be a blue sedan. I test drove several models, but when the salesman let me try out a used 2007 BMW 3 Series, I knew it was the car for me. Unfortunately, it was a lemon. It had over 100,000 miles on it, the check engine light never turned off after I left the lot, and even though it put up a good fight for almost 4 years, it had almost no resale value at the end of its life.

In hindsight, I would have made a different decision had I known how to value the price of a used BMW. From personal experience, BMW's price differently than the typical car. Used BMW cars are powerful and fun to drive compared to a new Honda or Toyota, but daily maintenance, higher fuel prices because of needing a higher octane, and depreciation really eat away at the value of the car over time. Prospective owners need to take into account these hidden charges when figuring out how much to pay for a used BMW.

I am going to build the model I wish I had when buying my used BMW. This model is going to serve as a reference to prospective buyers in order for them to make a better car purchasing decision.

2.) The Dataset

The dataset has been curated by DataCamp as one of their available datasets to use for their certification process. It consists of 10,781 observations and 9 features, which are listed below. Note there are no missing values.



```
In [4]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10781 entries, 0 to 10780
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   model         10781 non-null  object 
 1   year          10781 non-null  int64  
 2   price         10781 non-null  int64  
 3   transmission  10781 non-null  object 
 4   mileage       10781 non-null  int64  
 5   fuelType      10781 non-null  object 
 6   tax           10781 non-null  int64  
 7   mpg           10781 non-null  float64
 8   engineSize    10781 non-null  float64
dtypes: float64(2), int64(4), object(3)
memory usage: 758.2+ KB
```

3.) Analysis Plan

The end user wants to know what factors impact the price of a used BMW. This is a supervised learning problem predicting a continuous feature. I will perform the following steps:

Identify a suitable performance metric to give the end user

Perform exploratory data analysis connecting the feature variables to the target and find insights on which variables influence the target.

Select suitable algorithms based on exploratory data analysis.

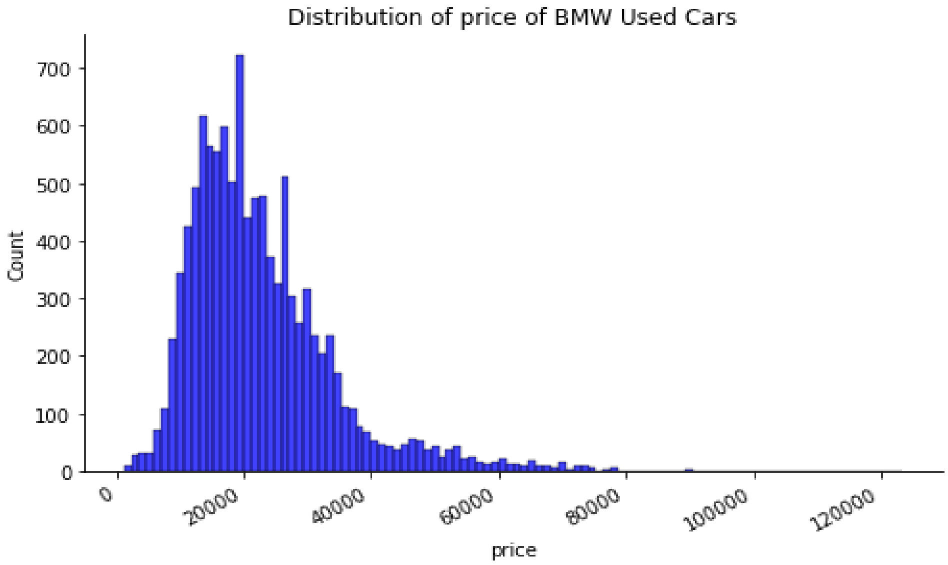
Fit, tune, and validate a model, or multiple models, to try and predict the price of a used BMW.

Discuss future improvements and suggestions for further research.

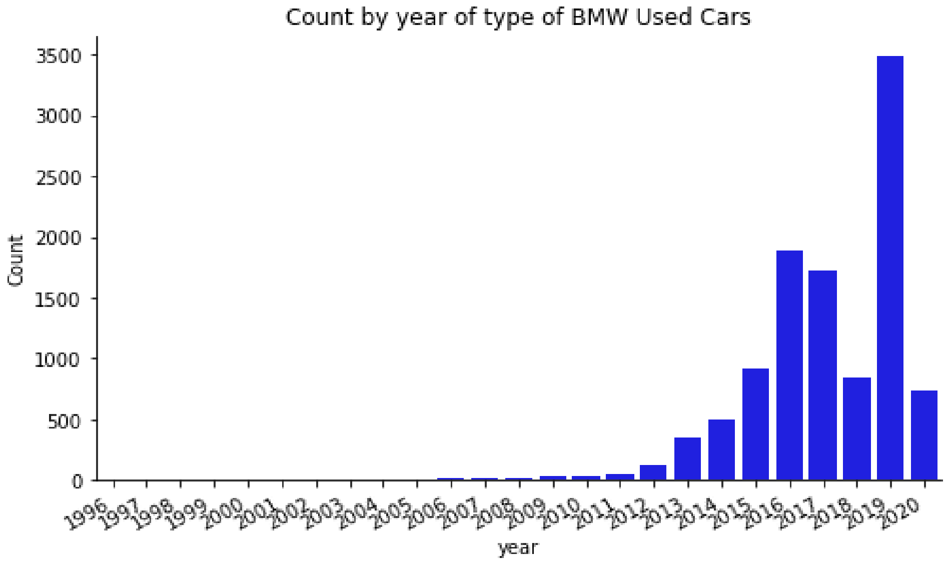
4.) Selecting Ideal Metric – RMSE

The ideal metric would give the user an interpretable estimate of how much the model differs from the actual price. In this case, the best metric to use is the Root Mean Squared Error (RMSE). This provides the squared root of the average of squared errors. Unlike Mean Absolute Error (MAE), RMSE punishes the model more for extremely errant predictions. In the case of the end user, extremely errant predictions would lead to an unusable model, particularly if the error is more than the value of the car. The squared root of the mean squared error provides for better interpretability over the standard Mean Squared Error.

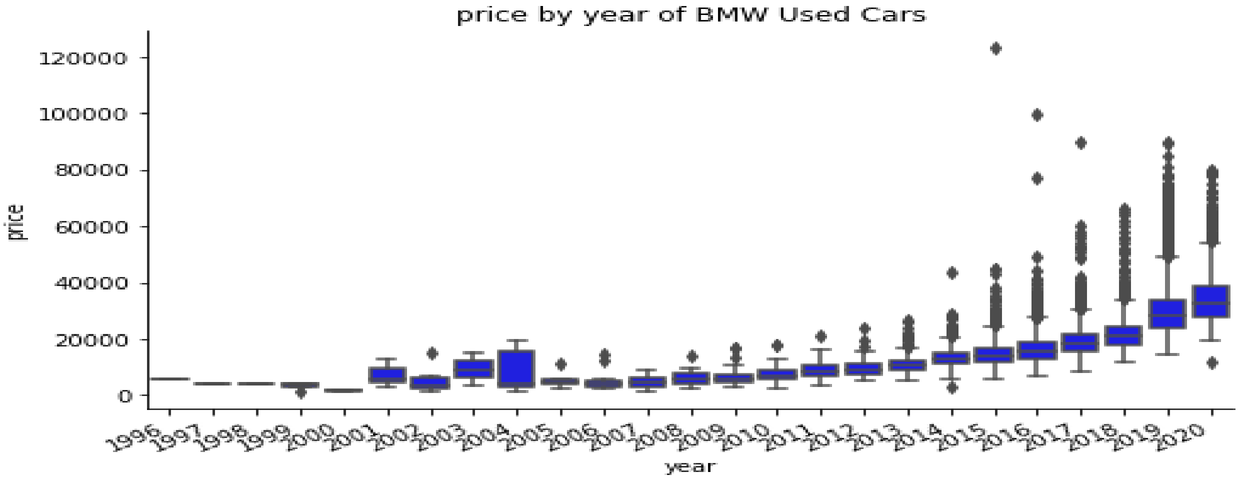
5.) Exploratory Data Analysis



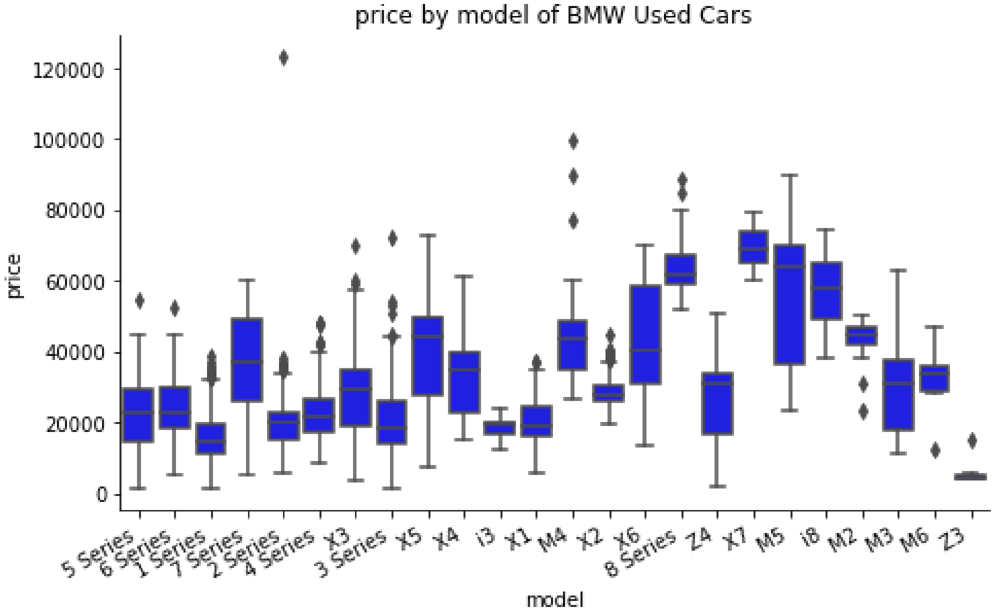
I will start by analyzing the distribution of the target variable, prices. The distribution appears to be skewed to the right by some cars with some higher prices, which is to be expected for a luxury car brand. To counteract this, I will transform the price to use the logarithm of the distribution for a more normal distribution. It is also interesting that there are BMW's that were sold for almost nothing, likely indicating a salvage title or other maintenance issues that the end user had.



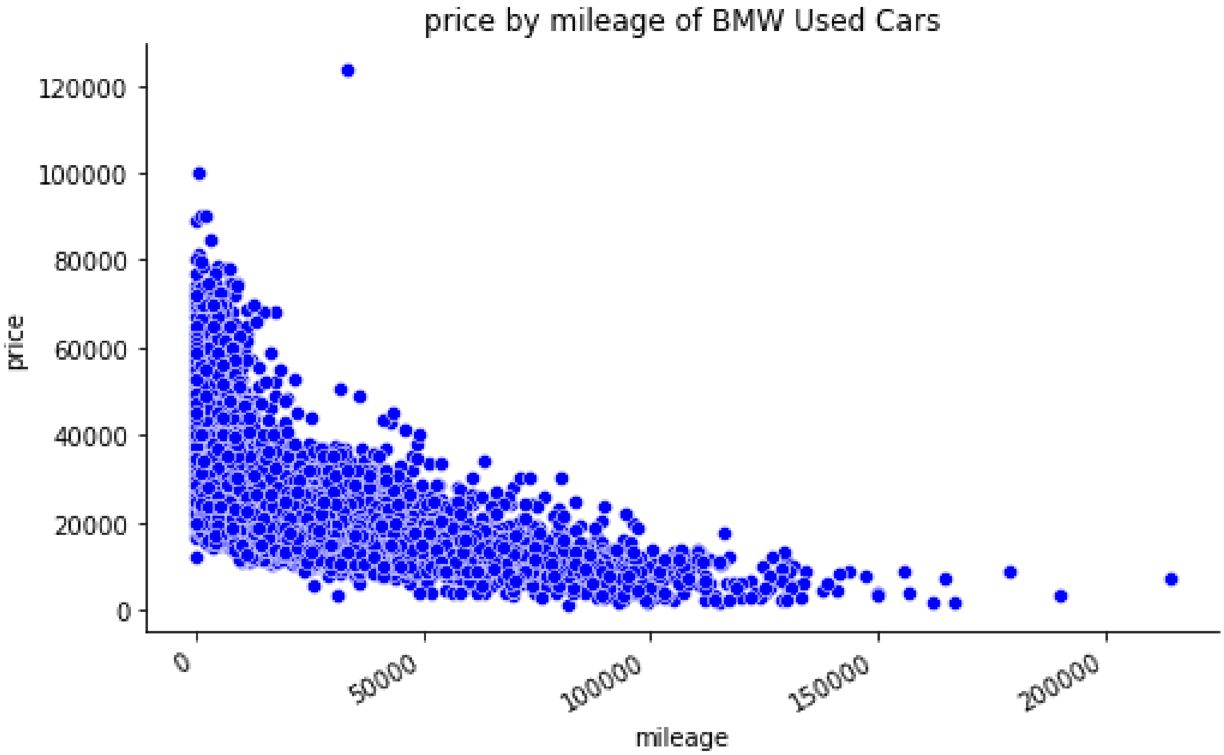
Next, I will look at the number of cars in the dataset by model year. Like most car models, there are more newer cars available than older ones. What is striking, however, is that most BMW’s do not last very long compared to a typical used car. Most cars in the dataset have been on the road fewer than 8 years assuming this dataset was collected in 2021. Cars older than that are likely either off the road, with another owner, or not in good enough condition to sell for anything. The vast majority of used cars are 2019 models, which likely comes from when leases expire or are models that never sold in the first place.



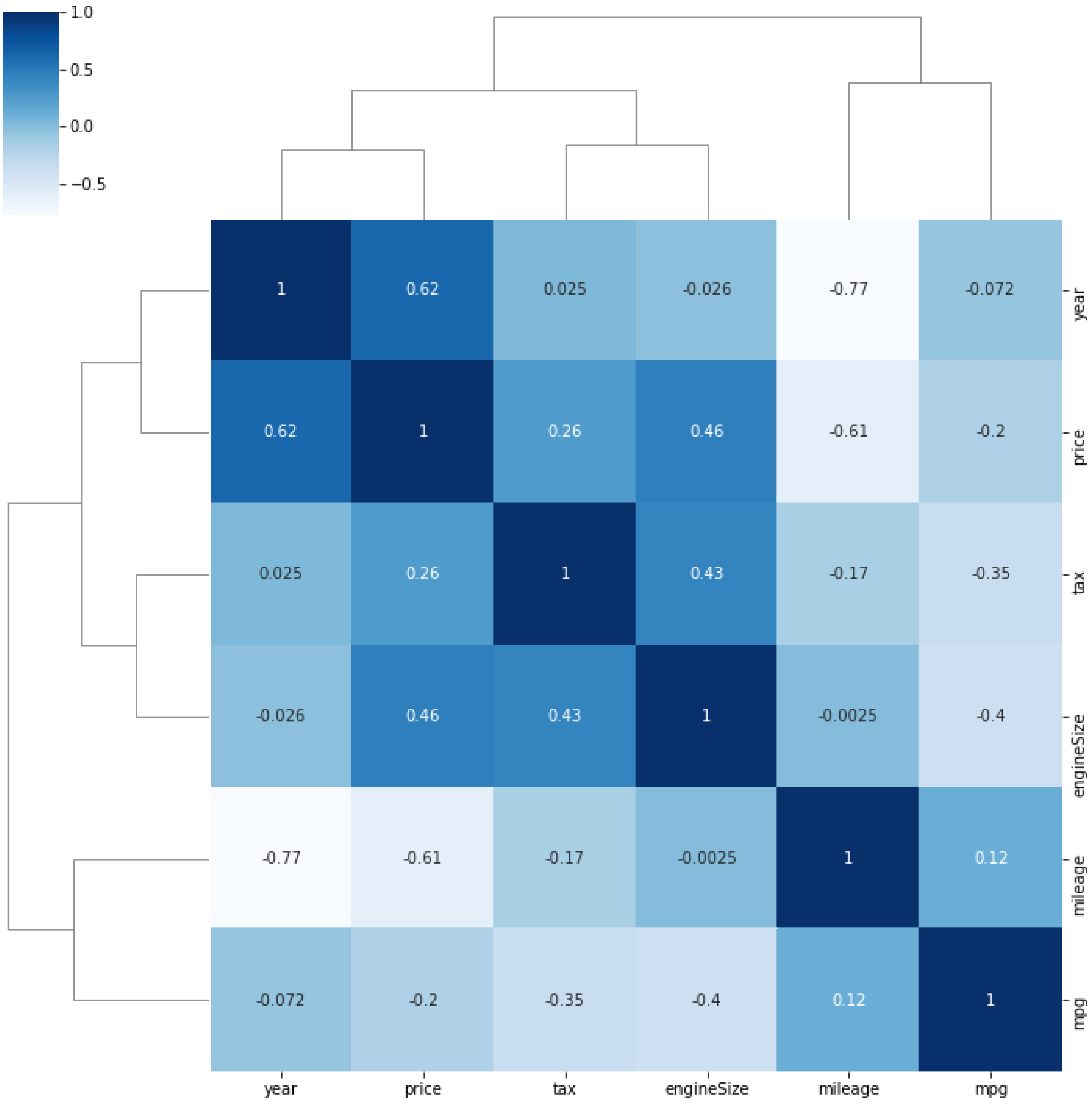
Here is a series of boxplots comparing the price by year. As the year gets closer to today, the more cars are available, thus there’s a wider range of values. There is also a consistent pattern where the newer the car, the higher the price on average. I did remove the one outlier of a car worth more than $120,000 during modeling.



This plot shows the price by model. The data shows a wide variation of price between models and within models. Some models, such as the M5 and X7, have a higher median price than the Z4, for example. With 24 different factors, model has the most cardinality of any feature in the dataset.



This last plot shows the price given the mileage, and it is useful for two purposes. One is that it shows a pattern where the more mileage on the car, the lower the price. The second is that the outlier of the car for over $120,000 really sticks out. The number of cars with zero mileage indicates that not every car in this dataset is “used”.



Above is a clustermap showing the correlation between the feature variables. It is similar to a heatmap, except that variables that are closer to each other are correlated with each other. For example, year and price are near each other because the newer the year, the higher the price.

From all of these charts and the data description, we can conclude the following:

- There are no missing values in the dataset.
- One outlier needs to be removed.
- The dataset is quite large, with 10,780 observations after removing that outlier.

- The distribution of price is skewed right and must be adjusted when modeling.
- As cars age and gain mileage, they depreciate in value.
- Car model, while important, does not determine price alone.
- There are no heavily correlated variables, suggesting no multicollinearity.

6.) Model Fitting

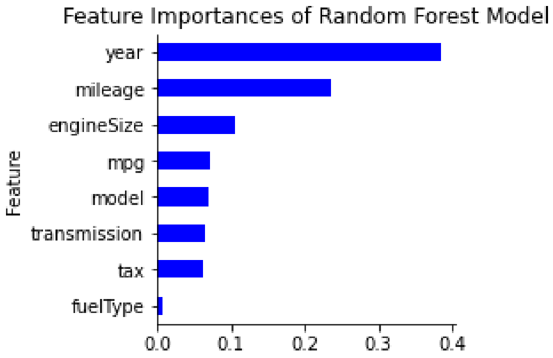
I fitted three different models for this problem: linear regression, random forest, and XGBoost. They all have their different strengths. The linear regression is the most basic model available, and this will be used as a baseline model to validate the random forest and XGBoost models. A random forest works well because the dataset contains both categorical and numerical features. It creates a series of decision trees using only a subset of the features each time to determine the most important features. The XGBoost model is similar to a random forest in that it builds a series of decision trees with the features given, but each model builds on top of each other. It comes with more hyperparameters to tune to tune at the cost of increased complexity.

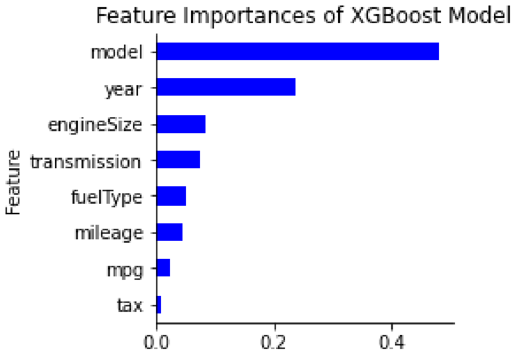
For the random forest and XGBoost models, I built a pipeline to preprocess the data, encode categorical features, scale numerical features, and fit the model using a 5-fold cross validated grid or randomized search on the training set. Hyperparameters tuned for the random forest include the max depth of the trees and the maximum number of features for each tree. Decreasing the maximum number of features reduces the possibility of overfitting. Number of estimators was not tuned because more estimators is almost always better, and I find it faster to have a consistent high number of estimators than grid search a series of smaller estimators.

Random Forest	Hyperparameters
Num_Estimators	100
Max Depth	19
Max Features	0.6
Random State	42
Root Mean Squared Error	$2,424.33

XGBoost	Hyperparameters
Num_Estimators	100
Max Depth	10
ETA (learning rate)	0.17
Colsample_bytree	0.55
Alpha	0.1
Lambda	3
Random State	42
Num Iterations	50
Root Mean Squared Error	$2,281.24

Above are tables showing the hyperparameters tuned, the final values of the fitted model, and the Root Mean Squared Error of the final model on the test set. The root mean squared errors are the average difference between the model's prediction and the actual price. For comparison's sake, the baseline linear regression model had a root mean squared error of $3,388.13. The mean squared errors of the training sets ($1,161 and $1,541) are much lower than these, indicating some overfitting.





7.) Model Evaluation

The baseline linear regression model had a root mean squared error of $3,388, while the random forest had a root mean squared error of $2,424 and the XGBoost had a root mean squared error of $2,281. This means that on the test set, the models were off on average by $3,388, $2,424, and $2,281, respectively. The differences between the baseline model and the tuned models suggest the model selection and hyperparameter tuning had an impact in building a better model.

Both the random forest and the XGBoost model have similar root mean squared errors, yet their feature importances differ. The XGBoost model, for example, places much more importance on the model of the car than the random forest does. Conversely, the random forest model places more emphasis on the mileage of the car than XGBoost does. For the end user, I would suggest looking at three categories as the most important: model, year, and mileage. The model is the car's type and theoretically, certain cars of the same type are similar. As a car gets older and users put more miles on it, the car becomes less valuable.

Model	year	transmission	mileage	fuelType	Tax	mpg	engineSize	RF_Price	XGB_Price
3 Series	2011	Automatic	103000	Petrol	100	23	3	$8,850	$9,761
3 Series	2007	Automatic	116068	Petrol	100	23	3	$5,947	$4,260

I figured it would be best to apply the model to predict the price of an actual BMW. I found a 2007 version online with 116,000 miles with very similar specifications and used the model to predict its price. I also predicted the price I should have paid for a 9-year-old BMW with the specifications I remember having. The models priced the nine-year-old BMW at $8,850 and $9,761, and the 2007 BMW at $5,957 and $4,260. It suffices to say my parents and I paid much more than that in maintenance alone. The large difference doesn’t mean the models are wrong. Models provide structure in thinking, and they would have said buyer beware in this case.

8.) Conclusions and Suggestions for Further Research

This report examines factors involved in predicting the price of a used BMW. The results of this model can be used to predict the price of used BMWs for prospective buyers. Using the available dataset, three models were fitted: a baseline linear regression, a random forest, and an XGBoost model. The best model was the XGBoost model with a root mean squared error of $2,281. This model probably overfits, but is still a solid result considering the average price of a used BMW in the dataset is just over $22,000.

Future research should focus on four efforts. The first is to stack the random forest and XGBoost models together to come up with a model that will likely perform better than either model alone. The second is to update the data to include 2021 and 2022 models. The third would be to look at predicting the longevity of a BMW considering that most of the cars in the study were fewer than 10 years old. Lastly, it would be nicer to have better clarity on where these cars were sold, as BMW is an international brand and prices would vary in different locations.

References

https://www.carfax.com/vehicle/WBAVC53567FZ78038