**Used Vehicle Buying:**

Improving the Buyer and Seller experience through inference and prediction models

**Group Title:** Group1

**Group Members (In Alphabetical Order):**

Ralph Parlin | Patrick Prioletti | Brian Schramke | Kobi Wiseman

**Class:** IST 718 | Professor Willard Williamson

**Table of Contents**

[1. ABSTRACT 3](#_Toc57449583)

[2. DATA 4](#_Toc57449584)

[2.1 Source and Collection 4](#_Toc57449585)

[2.2 Data Cleaning and Treatment for Missing Values 4](#_Toc57449586)

[2.3 Exploratory Data Analysis 4](#_Toc57449587)

[*2.3.1 Data Transformations* 6](#_Toc57449588)

[*2.3.2 Correlation Analysis* 6](#_Toc57449589)

[*2.3.3 Principal Component Analysis* 6](#_Toc57449590)

[*2.3.4 K-Means Cluster Analysis (Of PCAs)* 6](#_Toc57449591)

[3. METHOD 7](#_Toc57449592)

[4. MODELING 7](#_Toc57449593)

[4.1 Model 1: Vehicle Feature Inference Linear Regression 8](#_Toc57449594)

[*4.1.1 Model 1 Data Transformations* 8](#_Toc57449595)

[*4.1.2 Model 1 Evaluation* 8](#_Toc57449596)

[*4.1.3 Model 1 Interpretation* 8](#_Toc57449597)

[*4.1.4 Model 1 Inference and Key Findings* 9](#_Toc57449598)

[4.2 Model 2: Price Prediction with Linear Regression 9](#_Toc57449599)

[*4.2.1 Model 2 Data Transformations* 9](#_Toc57449600)

[*4.2.2 Model 2 Evaluation* 9](#_Toc57449601)

[*4.2.3 Model 2 Interpretation* 9](#_Toc57449602)

[*4.2.4 Model 2 Inference and Key Findings* 10](#_Toc57449603)

[4.3 Model 3: Feature Importance with Random Forest 10](#_Toc57449604)

[*4.3.1 Model 3 Data Transformations* 10](#_Toc57449605)

[*4.3.2 Model 3 Evaluation* 10](#_Toc57449606)

[*4.3.3 Model 3 Interpretation* 10](#_Toc57449607)

[*4.3.4 Model 3 Key Findings* 10](#_Toc57449608)

[4.4 Model 4: Price Prediction with Random Forest 11](#_Toc57449609)

[*4.4.1 Model 4 Data Transformations* 11](#_Toc57449610)

[*4.4.2 Model 4 Evaluation* 11](#_Toc57449611)

[*4.4.3 Model 4 Interpretation* 11](#_Toc57449612)

[*4.4.4 Model 4 Key Findings* 11](#_Toc57449613)

[4.5 Model 5: Vehicle Time on Market Inference Logistic Regression 11](#_Toc57449614)

[*4.5.1 Model 5 Data Transformations* 11](#_Toc57449615)

[*4.5.2 Model 5 Evaluation* 12](#_Toc57449616)

[*4.5.3 Model 5 Interpretation* 12](#_Toc57449617)

[*4.5.4 Model 5 Inference and Key Findings* 12](#_Toc57449618)

[4.6 Model 6: Fleet Use Logistic Regression 13](#_Toc57449619)

[*4.6.1 Model 6 Data Transformations* 13](#_Toc57449620)

[*4.6.2 Model 6 Evaluation* 13](#_Toc57449621)

[*4.6.3 Model 6 Interpretation* 13](#_Toc57449622)

[*4.6.4 Model 6 Inference and Key Findings* 13](#_Toc57449623)

[4.7 Model 7: Fleet Use Random Forest 14](#_Toc57449624)

[*4.7.1 Model 7 Data Transformations* 14](#_Toc57449625)

[*4.7.2 Model 7 Evaluation* 14](#_Toc57449626)

[*4.7.3 Model 7 Interpretation* 14](#_Toc57449627)

[*4.7.4 Model 7 Inference and Key Findings* 15](#_Toc57449628)

[5. CONCLUSION 15](#_Toc57449629)

[ANNEX 1: Detailed Correlation Matrix 17](#_Toc57449630)

[ANNEX 2: Detailed PCA Plots 18](#_Toc57449631)

# 

# **1. ABSTRACT**

COVID-19 has left virtually no corner of the economy untouched, but the automotive industry has been among the hardest hit. However, according to a Wall Street Journal article from July 2020, the used vehicle market has made a tremendous rebound as economies open back up, with sales of used vehicles in the U.S. climbing back faster than that of new vehicles after dropping 38% in April[[1]](#footnote-2). Although there are divergent opinions on the causes, one thing is clear: U.S. consumers are on the hunt for used vehicles.

Buying used vehicles can be challenging. Buyers are unaware of what sellers will take as a reasonable offer for their vehicle. Sellers, to include dealers, are often unaware of what price they call sell their vehicle for. This tension on price is muddied further by factors such as vehicle features, geographic region, time of year the sale is offered and alike. Although in recent years several applications such as CARFAX, Cars.com and alike have assisted, room for improvement remains with respect to informing buyers and sellers on the used vehicle market. We believe our timely analysis has helped.

Our own intuition combined with what we were seeing in the data helped us arrive at the following problem statement: *“Current processes for determining the price of a used vehicle creates pain points for both U.S. vehicle buyers and sellers as they struggle to determine a vehicle’s fair market value.  However, the valuation process can be improved through effective modeling.”*  To address the problem statement, we developed several research questions that allowed us to group analysis efforts along three primary characteristics of buying and selling vehicles: the price, how long a vehicle has sat on the market, and how the vehicle was used during its lifetime.

After rigorous exploratory data analysis, each modeler conducted individual modeling excursions to help answer our research questions. Our goal was to predict price, whether a vehicle would remain on the market more than 60 days, and predict vehicle history with respect to fleet use. With respect to inference, our goal was to determine which vehicle attributes matter most with respect to price, which vehicle attributes are most important in ensuring used vehicles don’t sit on the market, and which vehicle attributes are most important with respect to a vehicle’s fleet class.

To answer our research questions and meet our prediction and inference goals, models were created using a collection of machine learners including Regression, Logistic Regression and Random Forest. Modeling select features using linear regression proved a means of inference with respect to determining which features are most important when predicting price, while using PCA in linear regression proved more effective at predicting price. Training random forest regressors on two sets of variables, one organic and the other being principal components, we are able to extract inferential and meaningful findings relevant to vehicle price from a well performing model as well as some insight into variance-model interaction and behavior. Modeling select features using logistic regression proved a means of inference with respect to determining which features are most important when predicting whether a vehicle would be listed for more than 60 days on the market or not. Utilizing logistic regression and random forest classifiers with select features provided adequate models in predicting vehicle history with respect to fleet use and enabled inference on the most important features in predicating the fleet class.

Our analysis found that with respect to predict price, a vehicle’s horsepower, milage, and engine displacement are the leading vehicle attributes that predict price. Additionally, we found that predicting price is best achieved through our Random Forest model which rendered our lowest MSE with features that explain over 90% of the variation in price. With respect to predicting if a vehicle will remain listed on the market for more than 60 days, having a franchise dealer involved, whether the vehicle is listed as a cab or not, the vehicle's type of body, and price are the leading attributes that predict if a vehicle will remain listed that amount of time on the market. With respect to predicting if a vehicle was ever part of a commercial fleet, the vehicle’s model year, the vehicle’s mileage, and the vehicle’s owner count are the leading attributes that predict if it was part of a fleet.

# **2. DATA**

## 2.1 Source and Collection

This project utilized a used vehicle data set containing 3 million observations with 66 features and was sourced from Kaggle (<https://www.kaggle.com/ananaymital/us-used-cars-dataset>). The dataset was randomly sampled down to 600,000 observations using the “shuf” Linux command line utility to achieve a manageable file size.

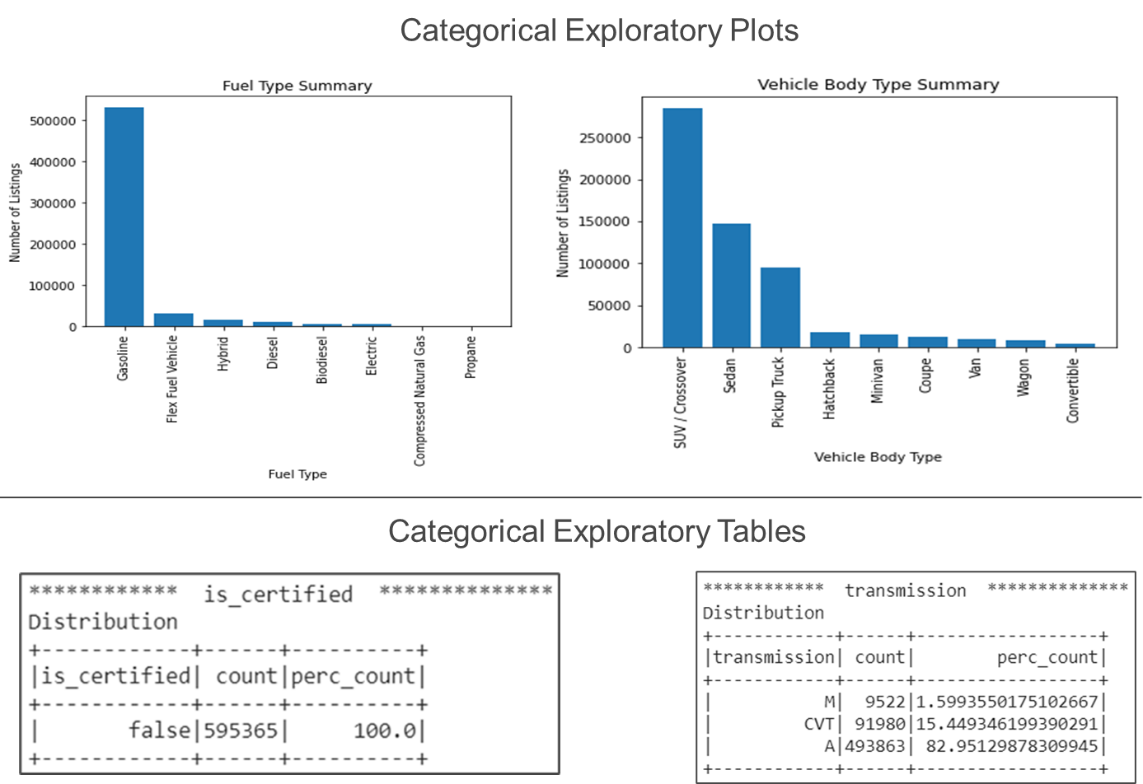
## 2.2 Data Cleaning and Treatment for Missing Values

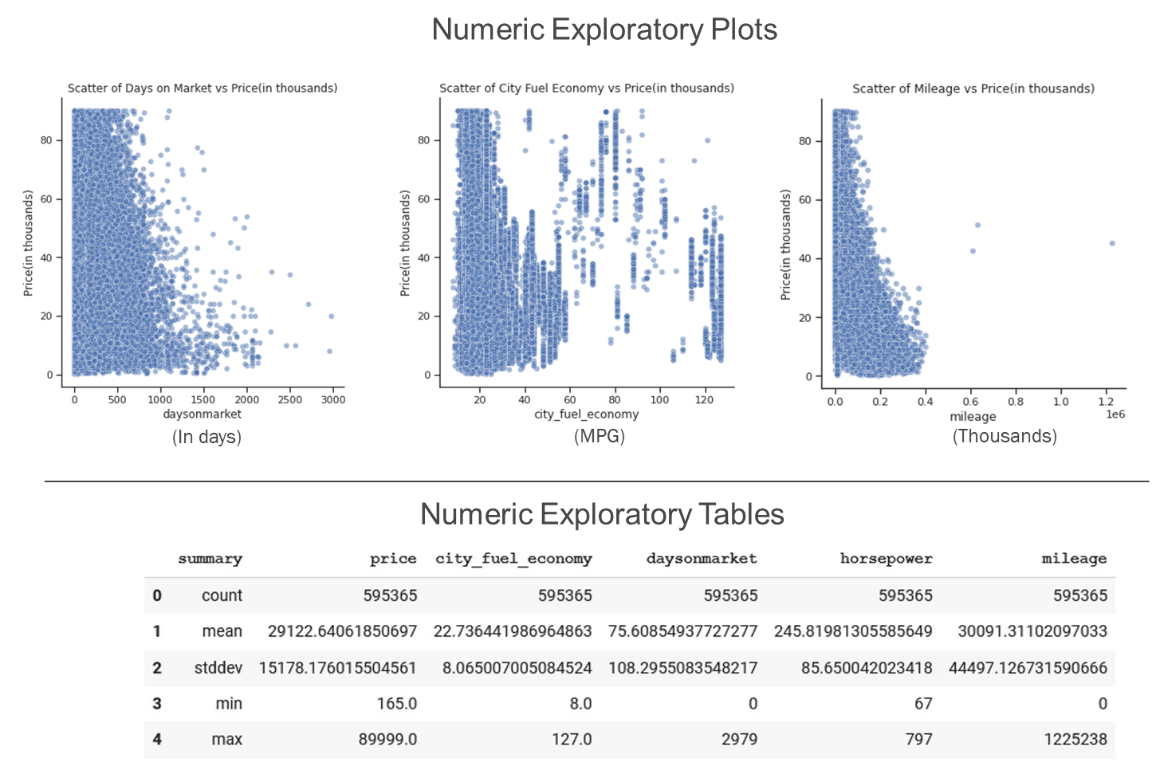
Features with zero variance or features with duplicate information were dropped. Regex expressions were used to remove unnecessary characters from many of the columns to prepare for recasting to integer values or for categorical assignment. In some instances, overlapping categorical values were consolidated into aggregated groups, for example, a “4X2” wheel-system being consolidated under the “RWD” category. NAs in categorical columns were replaced with the dominate category calculated across the data set. NAs in integer columns were replaced with the column median. NAs in float columns were replaced with column mean.

## 2.3 Exploratory Data Analysis

The exploratory data phase consisted of three distinct activities. The first was to provide summary statistics for each of the features in the dataset, the second activity was to create visualizations for both numeric and categorical features, and lastly was to examine the scatterplots and correlations between each feature. Each of these actives helped us better understand the features, their relationships, and provided us some initiation as we prepared to move into the modeling phase. An example of the data exploration and summary statistics can be seen below in Figure 1 below.

Figure 1: Exploratory Analysis





The EDA phase of this this analysis was critical in scoping the project and it greatly informed our analysis. For example, the original dataset included used vehicles with prices well beyond $1 million dollars. In an effort to provide a more useful model to average consumers, we capped the values of a vehicle at $90,000. Our exploration also lead us to discover that many of the observations included in the data were for vehicles categorized as “new”. These records were also removed from the data. After dropping columns not relevant to the analysis, addressing missing values, addressing outliers, and conducting relevant feature engineering, our data set had 303,798 observations across 66 features.

### *2.3.1 Data Transformations*

In addition to numerical data being available for analysis, we elected to apply categorical-to-numerical transformations in a multitude of ways. Primarily, any data which can be represented ordinally, such as engine size, should be converted with some rules, and so we create a numerical version of the variable maintaining manual control over the meaning of the order. Other variables which did not require ordinal properties to be maintained used a simple string indexer, which did not label any observation with a particular order, simply an arbitrary number. Depending on the algorithm and engineer preference, other transformations such as scaling or re-configurations of data for interpretability purposes were performed.

### *2.3.2 Correlation Analysis*

We can see, detailed by the correlation matrix of variables after some categorical encoding transformations, there are multiple columns which belong to one class and are therefore dropped. Additionally, we get a glimpse into what variables may be used as predominant predictors of dependent variables in addition to which variables may be under-valued or usurped during an analysis depending on algorithmic tendencies. See Annex 1 for a detailed correlation matrix.

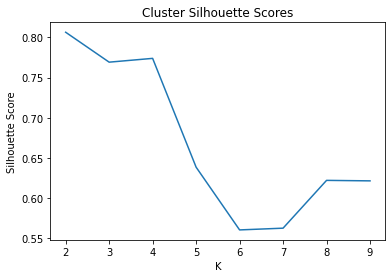
### *2.3.3 Principal Component Analysis*

Through a principal component analysis depicted in the associated plots, we can see that a relatively low number of principal components describe almost all of the variance, not including vehicle price. Additionally, we see that a large proportion of principle components account for virtually no proportion of the variance. See Annex 2 for plots.

### *2.3.4 K-Means Cluster Analysis (Of PCAs)*

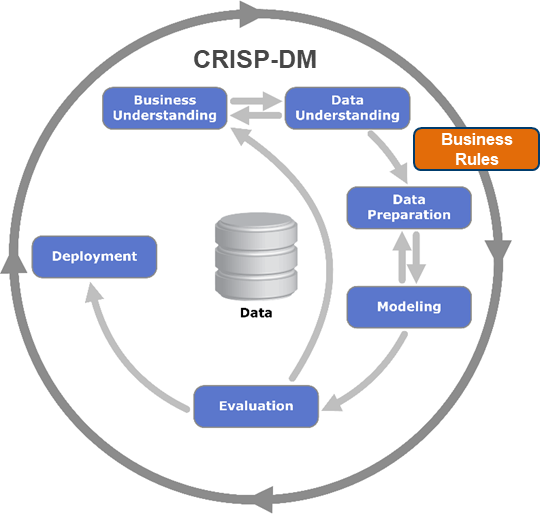
Utilizing the aforementioned 47 principal components we can see that unsupervised distance-based clustering yields the following silhouette scores. This can be observed in the plot, where we observe 2 centroids give the highest silhouette score. Both the PCA values as well as the cluster assignment may yield some understanding in later analyses, however, on their own they give insight only into how well separable the independent features are with no score to be applied through a dependent variable.

Figure 2. Cluster Silhouette Scores



# **3. METHOD**

Figure 3. Process

Our entire process followed the Cross-industry Standard Process for Data Mining (CRISP-DM) methodology. However, we augmented CRISP-DM slightly by applying a set of business rules to guide us. These business rules consisted of how we treated features, how we spoke about the features, and assigned units of measure for select features. We applied these rules to a base dataset, and from there each modeler was able to conduct the necessary excursions needed to performed modeling that helped answer their assigned research question. Once complete with individual modeling efforts, the team came back together to share findings, evaluate models and integrate findings into the report. The team collaborated through the use of Google drive, Google Colab, GitHub, and MS Teams.

# **4. MODELING**

Our own intuition combined with what was learned during EDA to arrive at the following research construct and questions:

1. Price Setting and Price Expectations:
   1. How can sellers set fair prices and how can buyers benchmark for value determinations? **[Prediction: Regression and Random Forest]**
   2. What are the most important features that help predict the price of a used vehicle? **[Inference: Regression and Random Forest]**
2. Vehicle Time on Market:
   1. What impact does a vehicle having been in an accident have on the time the vehicle spends on the market? **[Inference: Logistic Regression]**
   2. Determine how a vehicle’s price (which will be used as a predictor, unlike the linear model) on the probability for a vehicle to be listed on the market for a long time. **[Inference: Logistic Regression]**
3. Previous Vehicle Usage and History:
   1. How can we determine is a vehicle was previously part of a commercial fleet or a taxi to help buyers avoid heavy use vehicles? **[Prediction: Logistic Regression and Random Forest]**

These questions guided our process and provided the framework for which we created our division of labor. These three sections speak to the root of our problem statement and they allowed us to group analysis efforts along three primary characteristics of buying and selling vehicles: the price, how long a vehicle has sat on the market, and how the vehicle was used during its lifetime.

## 4.1 Model 1: Vehicle Feature Inference Linear Regression

Model 1 was designed to help answer research question number one concerning vehicle price setting and price expectations. Model 1 used linear regression analysis to provided inference on which sort of features best explain the variation in vehicle price.

### *4.1.1 Model 1 Data Transformations*

In addition to the features cleaned in the preprocessed data, Model 1 also used two features created to represent vehicle size categorically from the numeric dimensions of length and width. The length feature was indexed to “compact”, “midsize”, “fullsize”, and the width feature was indexed to “narrow”, “regular” and “wide”. This transformation was conducted to convert the values to categories more easily understood and to improve inference. For example, a vehicle buyer may not understand what a 110 inch wide vehicle is, but they most likely understand what a compact vehicle is. The other transformation created an age feature based on the year as a vehicle’s age is more informative and easier to interpret than a vehicle’s year. Keeping with the interpretability theme, this modeling effort only included features determined to be understood by average consumers. For example, features that included highly technical vehicle specifications, such as torque-grade, were excluded from this inference model. In total, 21 of the total available features were used. Pipelines were used to index the categorical features, assemble the feature vector, and scale the features using a standard scale transformer with default parameters.

### *4.1.2 Model 1 Evaluation*

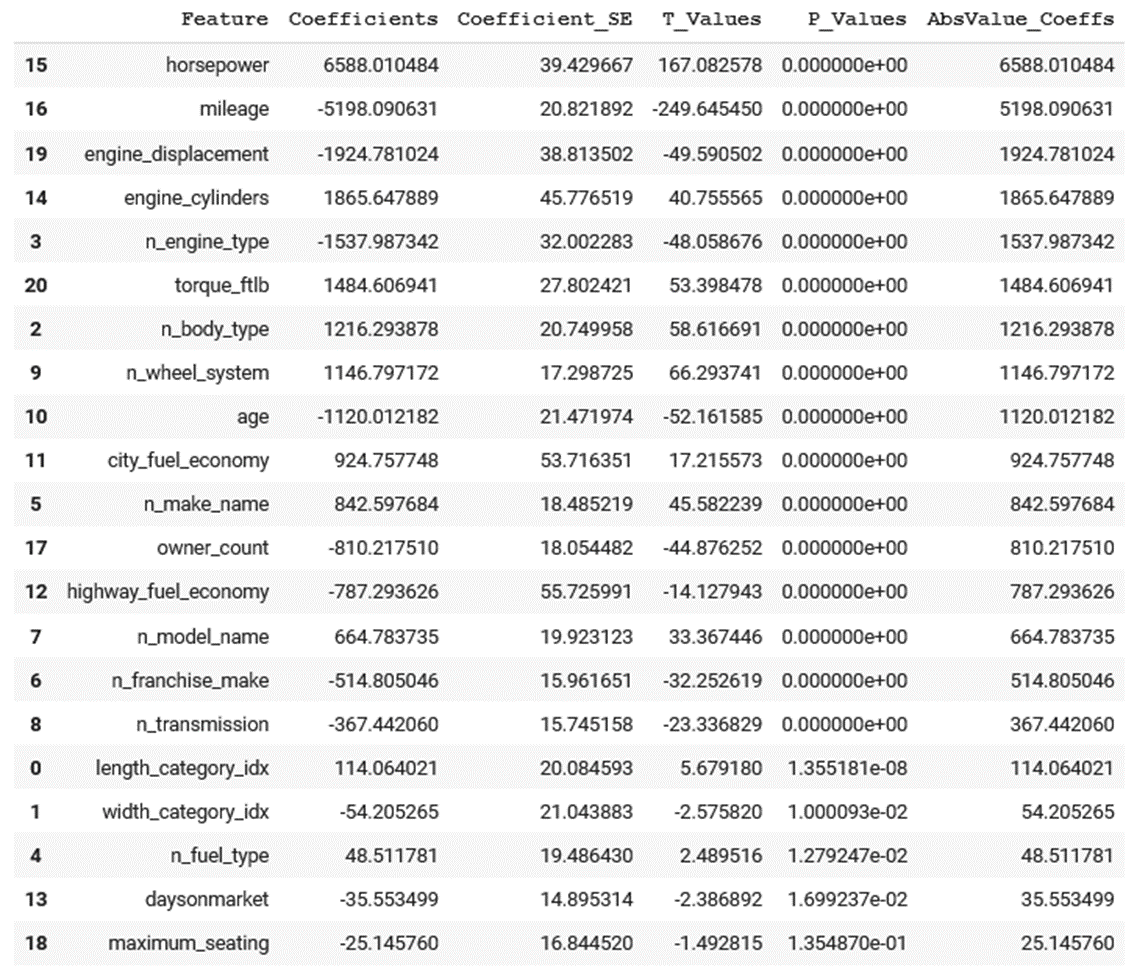
Model1 was build using a grid search technique and 3 folds cross validation. The parameters and the values searched through were regularization parameters of 0.0, 0.01, 0.02, 0.03, 0.4, and 0.7; and elastic net parameters of 0.0, 0.2, 0.4, 0.5, 0.7. The regression evaluator used the Mean Square Error (MSE) for evaluation of each model built fit during the process and after a 3-folds cross validation process the model fit with the lowest MSE had parameter values of 0.4 for elastic net, and 0.03 for regularization. Armed with these values, the data was then split 70% train, 30% test, and linear regression estimator was loaded with the tuned elastic net and regularization parameters and fit to the training data.

### *4.1.3 Model 1 Interpretation*

The fit model was significant and resulted in an R-squared value of 0.703. This means that just over 70% of the variation in a used vehicle’s price is explained by the predictors included in the model. When used for predictions, the fit model produced an MSE of 45,872,696 on unseen test data. However, additional analysis was conducted to identify the t-values and p-values for each coefficient used. PySpark documentation states that doing so requires changing the elastic net parameter back 0.0, and setting the solver parameter to “normal” and re-fitting the model. Doing this resulted in a model with the same R-squared value and, interestingly, a slight decrease in the MSE of 45,872,591 on the test data. Fitting the model in this way not only provided the coefficient values, but it also allowed us to extract the t-values and p-values from the fitted model and examine significance levels of each predictor.

### *4.1.4 Model 1 Inference and Key Findings*

Table 1: Linear Regression Summary Table

The best linear regression model included 21 statistically significant features at the 0.05 alpha level and explained just over 72% of the variation in price. The table to the right is ordered by the absolute value of the coefficients to help illustrate which features are most important with respect to price. We can see that the top three features are horsepower, milage, and engine displacement. We also see from the values of the coefficients that horsepower has a positive impact on price while milage and engine displacement have a negative impact on price. Interestingly we also see that as the highway gas milage increases, price decreases, while conversely, if city miles per gallon increase, price does as well.

## 4.2 Model 2: Price Prediction with Linear Regression

Like Model 1, Model 2 was designed to help answer research question number one concerning vehicle price setting and price expectations. However, rather than focus on inference, Model 2 was built to improve our ability to predict the target feature price. Model 2 was built using the same process as Model 1 but for this instance, the model was fitted using PCs as the independent variables. By using PCA it was expected that we could capture more of the variance than we did in the inference model (Model 1), and as a result, make better predictions.

### *4.2.1 Model 2 Data Transformations*

Model 2 used 35 PCs (accounting for over 90% of variation) from the pre-processed data that were scaled using a standard scale transformer with default parameters. No additional transformations were conducted.

### *4.2.2 Model 2 Evaluation*

Following the procedure conducted in Model 1, we applied the same gid search and 3-fold cross validation and arrived at the lowest MSE with parameter values of 0.5 for elastic net, and 0.04 for regularization. Armed with these values, the data was again split 70% train, 30% test, and linear regression estimator was loaded with the tuned elastic net and regularization parameters and fit to the training data.

### *4.2.3 Model 2 Interpretation*

The fit model was significant and resulted in an R-squared value of 0.715. This means that nearly 72% of the variation in a used vehicles price is explained by the 35 PCs. When used for predictions, the fit model produced an MSE of 44,321,213 on unseen test data.

### *4.2.4 Model 2 Inference and Key Findings*

Model 2 saw an improvement of 2% with respect to R-squared, and a reduction of MSE by nearly 1.5 million when compared to model 1. Although, Model 1 provides more intuitive inference, Model 2 proves to be much better at predicting the price of a used vehicle.

## 4.3 Model 3: Feature Importance with Random Forest

A different algorithmic approach to address the same hypotheses noted in Model 1, Model 3 utilizes a random forest regressor to estimate price. The random forest model used for this is a decision tree regressor, which utilizes the bagging method and philosophy of wisdom of the masses asserted by ensemble methods to improve predictive performance while maintaining generalizability. Decision trees utilize information gain to measure the effectiveness of possible splits build the tree based on which splits yield most information gain.

### *4.3.1 Model 3 Data Transformations*

Model 3 predicts price utilizes the same preprocessed features as Model 1 and share the same random state to ascertain how the training and test data are to be constructed.

### *4.3.2 Model 3 Evaluation*

After performing a grid search to reduce the MSE of the model, we find that the best performing tree available given the compute and memory limitations used a maximum split depth of 11 and 100 estimators. This specific model achieved and MSE of 14,704,843, or, 3834.69 RMSE. Additionally, this model was able to achieve a 90.5% R-squared value, which is based on the same 70% to 30% split of data between training and testing samples respectively.

### *4.3.3 Model 3 Interpretation*

Given the performance metrics provided, we can interpret our average error as a value of $3,834.69. Leaning on the interpretability of feature importance, we can see that the select order of features yielded this result, inferring that 90.5% of the price can be accounted for when utilizing a random forest and achieves a prediction with about $3,834.69. This allows us to view the importance of the features in the table provided, with some certainty, how important these features are when considering price.

Table 2: Feature Importance Table

### *4.3.4 Model 3 Key Findings*

We can see in the feature importance table some counter-intuitive, yet insightful findings. For example, days on market seems to be the highest of value, which may simply be an artifact of hidden aspects of the vehicles condition which are not disclosed within the data such as rust or other conditions. Other variables such as wheels system, engine type and fuel economy, as anticipated are strong considerations in predicting vehicle price. Oddly, what was not expected was mileage, owner count and age being lower on the list of priorities for assessing vehicle price. This is not well-anticipated, but considering the performance of the algorithm, this should be taken with the consideration that there may be some hidden aspects or correlations nested within these other variables which were simply phased out by the bagging procedure.

4.4 Model 4: Price Prediction with Random Forest

Similar to Model 3, Model 4 is another random forest regressor trained on the same principal components as Model 2. The top 35 principal components are used to grow the decision trees within the random forest regressor in attempts to create a better model in terms of price prediction performance while sacrificing the inferential properties of the real-world data.

*4.4.1 Model 4 Data Transformations*

The principal components utilized in constructing this model are 35 of the 47 principal components which explain virtually 100% of the variance within the dataset. No additional transformations were conducted on these principal components.

*4.4.2 Model 4 Evaluation*

Performing a grid search of the numTrees and maxDepth parameters, controlling the number of trees and maximum number of splits per tree, we get a minimum MSE of 36,321,581.34, which equates to an RMSE of 6026.74. Additionally, the R-squared value of this model was 76.5%. The training and testing data were split similarly to Model 3, with 70% of data being used for training and 30% being used for testing.

*4.4.3 Model 4 Interpretation*

As we can see within the performance metrics, an R-squared value of 76.5% denotes that the random forest regressor was able to explain 76.5% of the variance in price with the given features. Additionally, with a RMSE of 6026.74, which can infer an average error of $6024.74 across the dataset.

*4.4.4 Model 4 Key Findings*

Random forests appear to perform poorly when using principal components as their input. In fact, there seems to be an anomalous pattern when optimizing parameters with principal components where each combination of random forest parameters tested yields the same MSE of 36,321,581.34. This is likely due to the fact that principal components are an artificial variable composed of linear combinations of different features. Because decision tree based random forests don’t utilize any direct derivation of variance, using principal components on their own yields only maximal gains achievable by calculation of the information gain across the number of principal components the model incorporates. Further, the not-so-dazzling R-squared seems to tightly correspond to the proportion of principal components utilized; being 76.5% while the percentage of principal components used was 74%.

## 4.5 Model 5: Vehicle Time on Market Inference Logistic Regression

Model 5 was designed to address our third and fourth research questions concerning the probability of a listed vehicle to stay on the market for a long time: By how much do price and having history of accidents effect a vehicle’s probability to stay on the market for a long time? Model 5 is a binary logistic regression model. Following a series of tests, *60* days on the market was selected as the cut-off value to determine whether a vehicle is listed for a ‘long time’ or not. Using this value, the business sense is maintained as both categories are well represented.

### *4.5.1 Model 5 Data Transformations*

Model 5 used the scaled data features. The Boolean columns were recast to integer type to convert true and false values to ones and zeros, respectively. In addition, a vehicle’s length and width features were binned into a categorical, ordinal features. Model 5 used 20 features in total. The following features were not included:

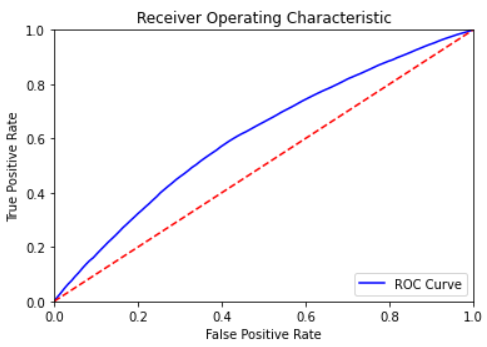
* Bed, cabin, franchise\_make: Did not pertain to all vehicles
* sp\_id, dealer\_zip, listed\_date: Captured similar information compared to other features
* daysonmarket: Derivative of target feature.
* In addition, features which suffered misrepresentation (contained at least one non-mergeable category with less than 5% frequency) were not included in the model.

### *4.5.2 Model 5 Evaluation*

Model 5 utilized a grid search to tune the regularization and elastic net parameters. The grid utilized regularization parameters of 0, 0.01, and 0.02 and elastic net parameters of 0, 0.1, and 0.2. The trained model variations were evaluated using the area under the Receiver Operating Curve (ROC) metric and the validation split. The best model utilized a regularization parameter of 0.01 and an elastic net parameter of 0.1.

### *4.5.3 Model 5 Interpretation*

Figure 4. ROC Curve

The best logistic regression model was able to achieve a ROC score of 0.61 which is nearly twice greater than a no skill classifier with an AUC ROC of 0.31 (see plot below). This illustrates the model enhances a buyer’s potential to predict whether a vehicle is listed for more than 60 days on the market.

### *4.5.4 Model 5 Inference and Key Findings*

The feature coefficients were extracted from the model and sorted based on absolute value to provide insight on feature importance. The top ten feature coefficients are provided below in Table 3

Table 3: Fleet Use Logistic Regression Feature Importance

|  |  |  |
| --- | --- | --- |
| **feature** | **weight** | **abs\_weight** |
| franchiseDealer | -0.272319 | 0.272319 |
| isCabNew | 0.090214 | 0.090214 |
| n\_body\_type\_bin | -0.074962 | 0.074962 |
| price | -0.067508 | 0.067508 |
| width\_category\_idx | 0.064292 | 0.064292 |
| engine\_displacement | -0.061376 | 0.061376 |
| HasAccidents | 0.052054 | 0.052054 |
| n\_engine\_type\_bin | -0.046449 | 0.046449 |
| torque\_ftlb | 0.046355 | 0.046355 |
| n\_wheel\_system | 0.035242 | 0.035242 |

The model’s most important feature is the Boolean of whether a vehicle is promoted by a franchise dealer or not; being promoted by a dealer reduces the probability of staying on the market for more than 60 days. This is a reasonable result since car dealers have dedicated time and resources to expend on selling the vehicle. To answer our business questions, the features of price and having history of accidents do appear on the top ten most important features, however its importance level is by far lower than being promoted by a franchise dealer.

## 4.6 Model 6: Fleet Use Logistic Regression

Model 6 was designed to address our fifth research question concerning vehicle history and usage patterns: How can we determine if a vehicle was previously part of a commercial fleet or a taxi to help buyers avoid heavy use vehicles? Model 6 is a logistic regression model.

### *4.6.1 Model 6 Data Transformations*

Model 6 used the preprocessed dataset from the PCA analysis. The Boolean columns were recast to integer type to convert true and false values to ones and zeros, respectively. Model 6 used 46 of the available features. The following features were not included

* Bed, cabin, franchise\_make: Did not pertain to all vehicles
* sp\_id, dealer\_zip, listed\_date: Captured similar information compared to other features
* is\_cab, fleet: Derivative of target feature, and target feature.

The resulting data frame was split using 60% for training, 30% for validation, and 10% for testing. The transformed features were scaled using a standard scaler transformer to enable feature importance inference following model training.

### *4.6.2 Model 6 Evaluation*

Model 6 utilized a grid search to tune the regularization and elastic net parameters. The grid utilized regularization parameters of 0, 0.01, and 0.02, and elastic net parameters of 0, 0.1, and 0.3. The trained model variations were evaluated using the area under the Precision/Recall curve (AUC PR) metric and the validation split. The PR metric was used because of the class imbalance with only 22% of observations with a positive fleet condition. The PR metric is only concerned with prediction of the positive scenario. The best model utilized a regularization parameter of 0.01 and an elastic net parameter of 0.1.

### *4.6.3 Model 6 Interpretation*

The best logistic regression model was able to achieve an AUC PR score of 0.38 which is almost two times greater than a no skill classifier with an AUC PR of 0.22. This demonstrates the model enhances a buyer’s potential to predict if a vehicle was used in a commercial fleet compared to a guess based on population statistics.

### *4.6.4 Model 6 Inference and Key Findings*

The feature coefficients were extracted from the model and sorted based on absolute value to provide insight on feature importance. The feature coefficients are provided below in Table 4.

Table 4: Fleet Use Logistic Regression Feature Importance

|  |  |  |
| --- | --- | --- |
| **feature** | **weight** | **abs\_weight** |
| torque\_grade | -0.709929 | 0.709929 |
| salvage | -0.494289 | 0.494289 |
| frame\_damaged | 0.337285 | 0.337285 |
| is\_cpo | -0.248896 | 0.248896 |
| theft\_title | 0.210891 | 0.210891 |
| owner\_count | 0.175646 | 0.175646 |
| franchise\_dealer | -0.162414 | 0.162414 |
| n\_wheel\_system | -0.125713 | 0.125713 |
| maximum\_seating | 0.115323 | 0.115323 |
| has\_accidents | -0.063154 | 0.063154 |

The model’s most important feature is torque grade rating which is the max torque in horsepower divided by the required engine RPM. These results may indicate that many fleet vehicles utilize lower performing powertrains compared to other vehicles of the same model and trim level. This makes sense as commercial fleets may prioritize economy and limiting capital investment in lieu of driving satisfaction through increased performance.

## 4.7 Model 7: Fleet Use Random Forest

Model 7 was also designed to address our fifth research question of determining if a vehicle was previously part of a commercial fleet or a taxi. Model 7 served as an alternative to the previous logistic regression in an attempt to improve model performance. Model 7 is a random forest classifier model.

### *4.7.1 Model 7 Data Transformations*

Model 7 used the same transformed data as Model 6 and the resulting training, testing, and validation splits to enable meaningful comparison of the best models. The random forest classifier experienced Java runtime errors in the Google Colab environment when trying to include categorical features with many categories. The “sp\_name” and “interior\_color” feature columns were removed from the random forest model which reduced the required max bin size from 24,607 to 4,482 and eliminated the runtime errors. Model 7 did not include a standard scaler transformer and relied on the built-in feature importance function of the random forest classifier model.

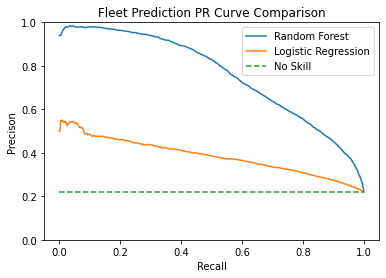
### *4.7.2 Model 7 Evaluation*

Model 7utilized a grid search to tune the maximum tree depth and number of trees hyperparameters. The grid utilized max depth values of 5, 7, 9, and 20, 30, and 40 trees. The trained model variations were also evaluated using the area under the PR curve metric using the validation split. The best model utilized 40 trees with a maximum depth of 9.

### *4.7.3 Model 7 Interpretation*

The best random forest model was able to achieve an AUC PR score of 0.76 outperforming both the Model 6 logistic regression at 0.38 and a no-skill classifier at 0.22. A graph displaying the PR curves for the logistic regression and random forest models can be seen in Figure 5.

Figure 5: Fleet Prediction PR Curve Comparison



### *4.7.4 Model 7 Inference and Key Findings*

The feature importance scores were extracted from the random forest model using the built-in feature importance method available in Pyspark. The sorted feature importance scores are provided below in Table 5.

Table 5: Fleet Use Random Forest Feature Importance

|  |  |
| --- | --- |
| **column** | **weight** |
| year | 0.409421 |
| mileage | 0.10942 |
| owner\_count | 0.048645 |
| n\_make\_name | 0.037899 |
| price | 0.033223 |
| length | 0.024229 |
| wheelbase | 0.023218 |
| latitude | 0.021761 |
| n\_model\_name | 0.021318 |
| torque\_rpm | 0.020103 |

The model’s most import feature is model year, indicating the model may distinguishing blocks of vehicles that are entering the market as fleets refresh their inventory. Using the model as-is may lead to drift over time as model years change as fleets continue to refresh their vehicles with newer models thus reducing model performance. This could be combated through periodic retraining of the model with current data. Future work could include performing feature engineering to add a vehicle age column to replace model year and reevaluate model performance.

# **5. CONCLUSION**

Linear regression and random forest converge to provide very informative inference on what features help explain price. However, Random Forest was superior with respect to predicting price. Both the linear regression and the random forest used the same features with the same treatment, yet the Random forest resulted in a model with an MSE on unseen test data of 14,704,843. This represents a reduction of nearly 68% in MSE. Additionally, the Random Forest model was able to achieve an R-squared value of 90.5% representing an increase of over 17% more variation in price explained than that of the highest performing Linear Regression model using 35 PCs. In summary, Random Forest was the model of choice for predicting use vehicle prices.

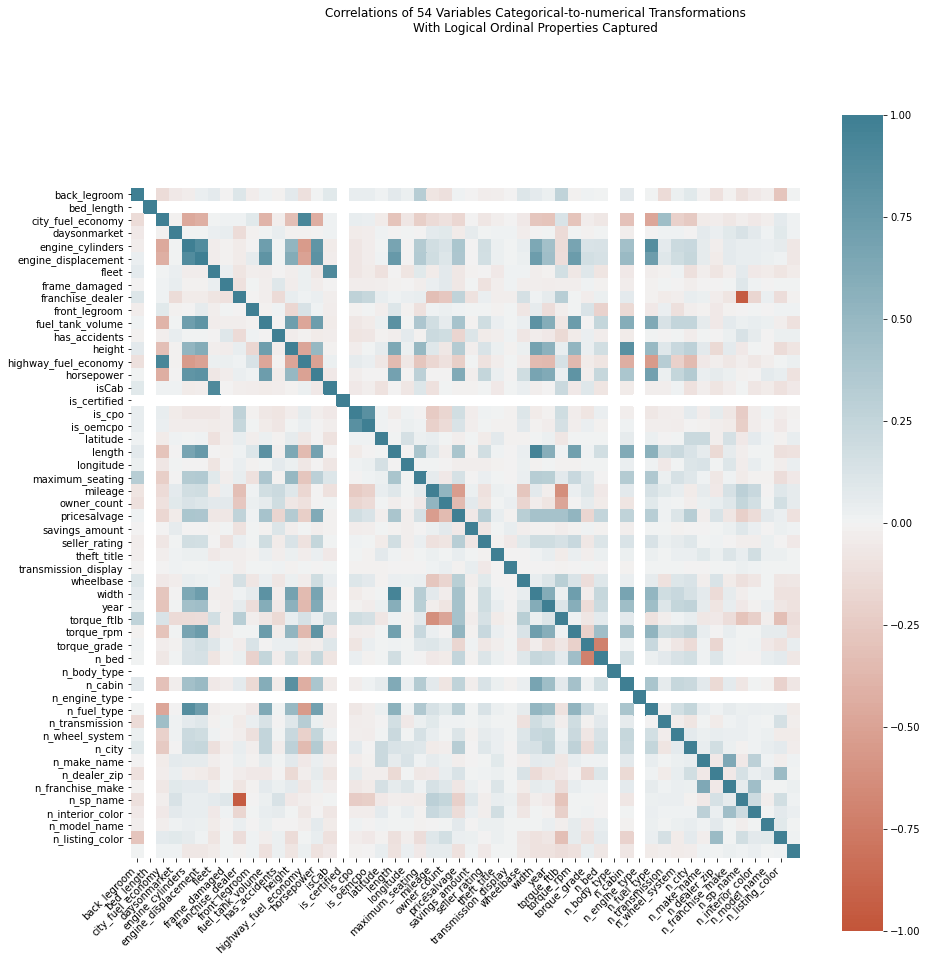
The binary logistic regression model answered our two research questions regarding a vehicles’s probability to remain on the market for a time greater than 60 days. Based on the feature importance findings, both price and having history of accidents are somewhat impactful when determining whether a vehicle to be listed on the market for more than 60 days or not. However, the involvement of a franchise dealer in the process of selling the vehicle appears to be by far the most important feature. Having a franchise dealer assisting in selling a vehicle will result at higher probability for that vehicle to be on the market for less than 60 days.

A random forest classifier was most successful at predicting whether a vehicle was ever in a commercial fleet when compared to a logistic regression model. The resulting random forest model relied on features that will statistically drift over time and reduce model performance without periodic retraining or feature engineering to transform the model year feature to a more relative value like age.

In summary, this analysis was successful at answering our research questions and offers valuable insight with respect to used vehicle transactions. The analysis helps show which vehicle attributes matter most with respect to price, time on market and probability of being a member of a fleet. This analysis also offers an accurate price prediction model with our top performing model being a Random Forest models. More generally we see that Random Forest was the lead performer in answering all of the research questions which suggests this problem is more complex than a linear equation. Together, these tools can be used to help determine the price of a used vehicle and alleviates pain points for both U.S. vehicle buyers and sellers as they try to determine a vehicle’s fair market value.

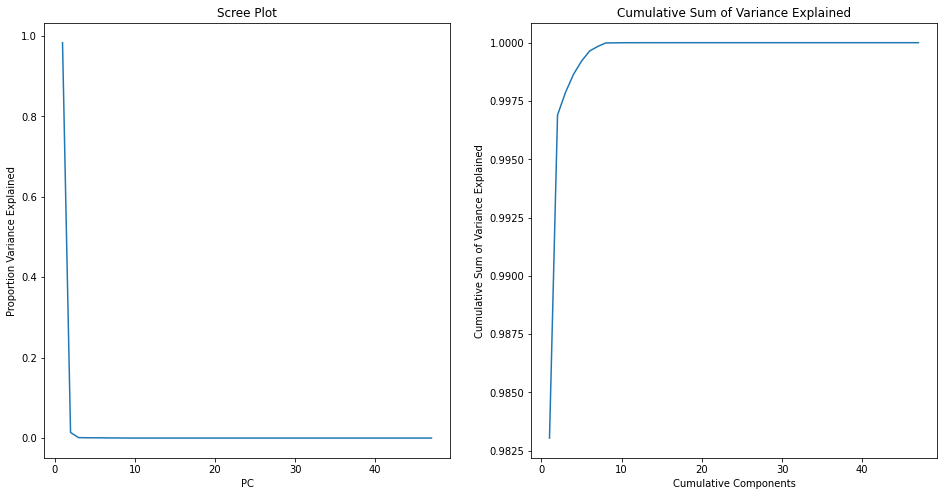
# **ANNEX 1: Detailed Correlation Matrix**

We can see, detailed by the correlation matrix of variables after some categorical encoding transformations, there are multiple columns which belong to one class and are therefore dropped. Additionally, we get a glimpse into what variables may be used as predominant predictors of dependent variables in addition to which variables may be under-valued or usurped during an analysis depending on algorithmic tendencies.



# **ANNEX 2: Detailed PCA Plots**

Through a principal component analysis depicted in the associated plots, we can see that a relatively low number of principal components describe almost all of the variance, not including vehicle price. Additionally, we see that a large proportion of principle components account for virtually no proportion of the variance.



1. Wall Street Journal, “During COVID-19 Pandemic, the Used-Car lot is Hot”, Mike Colias, July 2020 <https://www.wsj.com/articles/during-covid-19-pandemic-the-used-car-lot-is-hot-11593774001> [↑](#footnote-ref-2)