JUSTYNA JAKUBOWSKA

Used Cars Sales Analytics Report

2023

Table of Contents

01

Overview

For this assignment I assumed that the report I am making is for an imaginary national used car dealership (such as Vroom or Carvana) that ships the cars to all states.

Since this is a national dealer I concentrated on types of cars. Anyone buying a car has a particular type of car in mind depending on their needs. A large family might buy a minivan, a contractor might go for a pickup, a student might go for a coup, while a dog owner might want to have a wagon. All these cars have different features that are attractive to a buyer. I also separated cars that are 25+ years old because they are considered antiques by definition. Hence I decided to divide the data into groups by type of the car:

- Sedan
- SUV
- Pickup
- Truck
- Coupe
- Hatchback
- Convertible
- Van
- Wagon
- Minivan
- Bus
- Offroad
- Antiques
- Convertible
- Van
- Wagon
- Minivan
- Bus
- Offroad
- Antiques

Since the dealer has a national database and ships the cars to all locations the state and region column/feature will not be considered in my ML model. *

- If at some point the dealer would decide to open distribution centers, then new ML models should be created to help the dealer decide what cars should be stocked in each distribution center. At that point the state/ region features should be taken into the consideration to create new ML models.

01

Overview

THE PROCESS OF CREATING BEST ML MODEL

Data Understanding & Preparation

After performing EDA I checked for

- missing values,
- outliers,
- duplicates,
- consistency,
- accuracy,
- completeness.

In the process I :

- saw anomalies in pricing and odometer,
- decided that adding 'age' column is more beneficial than having 'year',
- noticed a lot of missing data can be filled based on 'VIN' column value decoding,
- realized some missing values are in the wrong column,
- dropped id column because it didn't contribute to building a model,
- dropped VIN column because values are all unique and don't affect outcomes (after using it to fill many missing values),
- dropped the 'model' column, because it had 30,000 unique values and would just slow down processing time and wouldn't have valuable input,
- dropped the "size" column because more than 70% of data is missing (also "type" column indicates well the size of the car)

Once the data was relatively well cleaned I split it into smaller datasets by type and continued cleaning the smaller datasets in new Jupyter Notebooks.

Individual Jupyter Notebooks for Each Car Type

After checking the distribution of the odometer data and since it wasn't symmetrical, I used median to replace NaN values.
Since each car type have very specific values for its type, I finely felt comfortable filling the missing values for each feature proportionately.
Now there were no missing values.
I also dropped the type feature since

01

Overview

values were all the same.
I applied ordinal encoding to condition feature, all other non numerical features were treated with One Hot Encoding.
After creating histograms of price column and the logarithm of the price column to decide which one is more suitable for the machine learning model, I chose the logarithm for each ML model for each car type.

Selecting the best ML model.
Here are the steps I took for each car type:

- split the data frame with a Logarythmic target: price,
- Scaled the data with a StandardScaler to:

* avoid bias towards features with larger values
* improve numerical stability:
* reduce the impact of outliers
* enable efficient use of resources

- set a baseline to compare my models to,
- created models with best parameters for:

* Linear Regression with Polynomial Features
* Lasso Regression with Polynomial Features
* Ridge Regression with Polynomial Features

- Ran best models through SFS,
- Ran best models through RFE*
- Selected the best model with lowest testing MSE

Selecting features that influence the price of each car type.
After selecting the best ML model I ran that model through permutation Feature selection and got a list of the most influential features that affect the ML model.
To make the results more readable I created a chart of the features and their importance.

Results
All results for each car type are listed in the pages below.

I also created a separate document for the "client" that is not familiar with the ML jargon. The document is posted on Github.

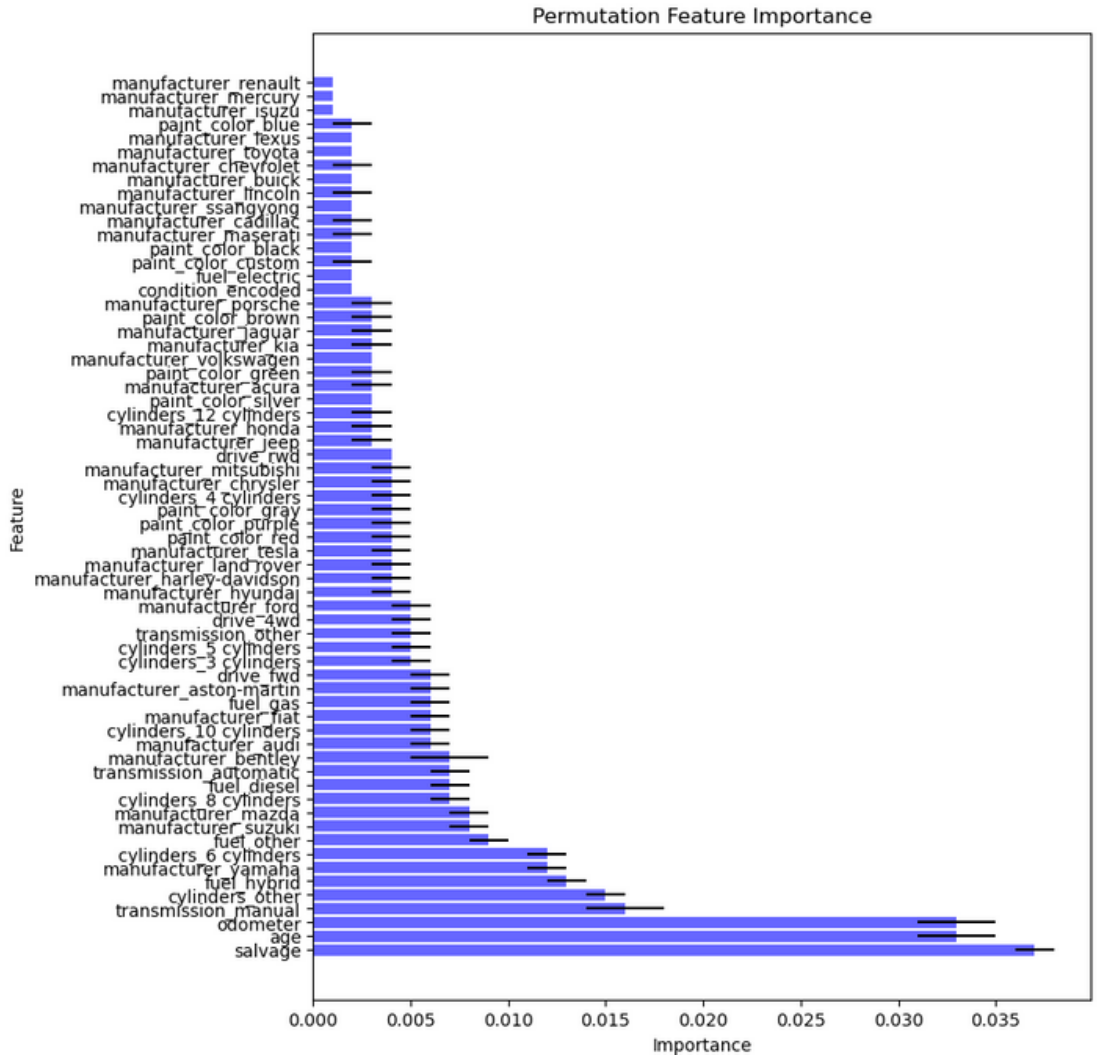
- Due to the limitation of my computer I wasn't able to run a lot of models through RFE

02

Sedan

The best model for predicting sedan price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 6.7328502727423984
Testing MSE: 6.8427712896917186

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a sedan. The R2 score of the model is 0.092, which suggests that the model explains only a small portion of the variance in the target variable: sedan's price. The top 5 most important features (in descending order) are:
1) salvage,
2) age,
3) odometer,
4) transmission_manual, and
5) cylinders_other.
The remaining features have relatively low importance scores, which means that they have little impact on the model's predictive performance.

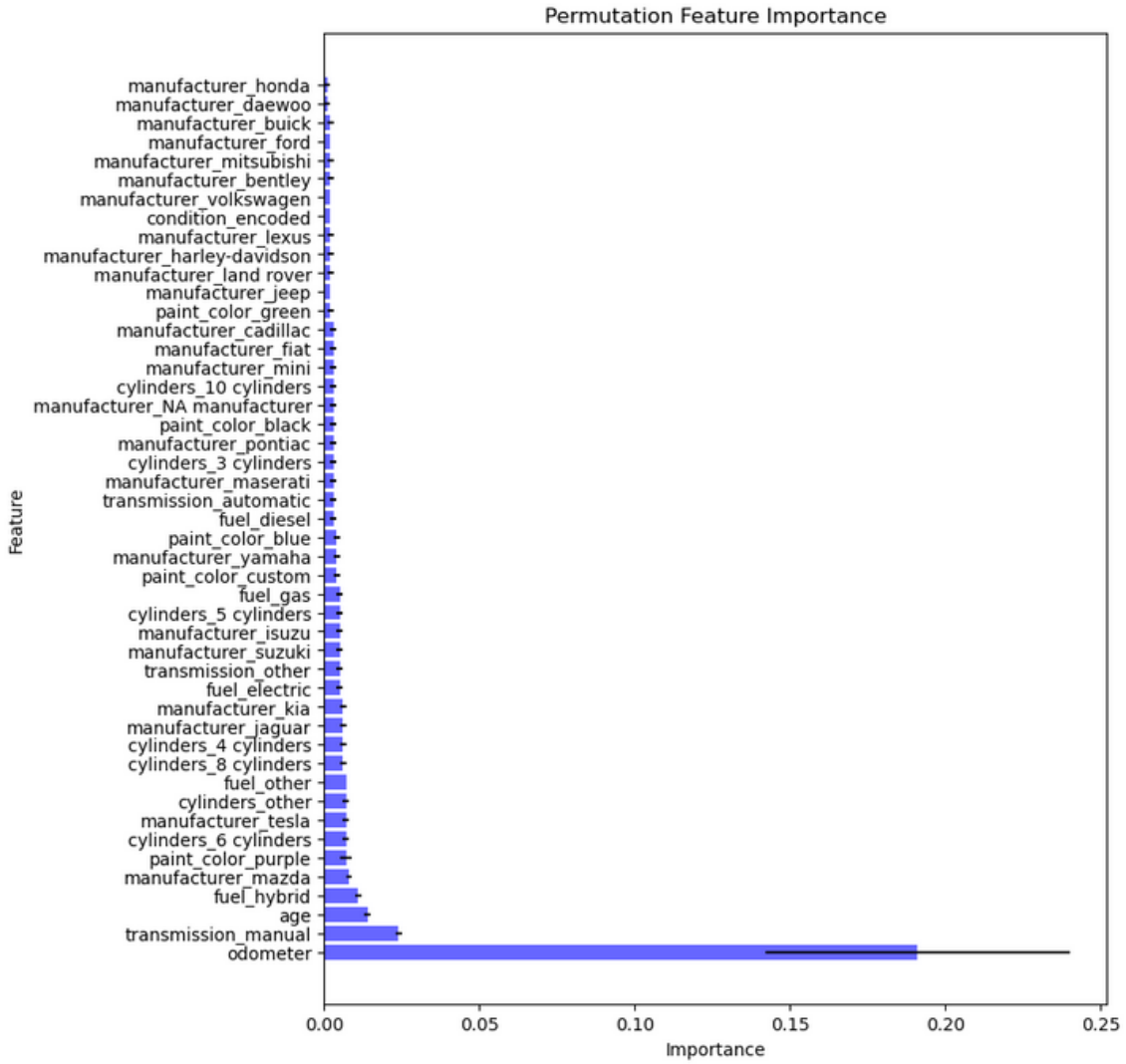


03

SUV

The best model for predicting SUV price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 7.243354372795317
Testing MSE: 7.709987522920182

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of an SUV. The R2 score is 0.040, which hints that the model has poor performance in explaining the variance of the target variable: SUV's price. The feature "odometer" has the highest importance score of 0.191, indicating that it is the most important feature in the model. The "transmission_manual" feature has the second-highest importance score of 0.024, meaning that it also has a significant impact on the model's performance. The other features have lower importance scores, 0.001 to 0.014, indicating that they have a relatively small impact on the model's performance. These features include "age", "fuel_hybrid", "manufacturer_mazda", "paint_color_purple"...

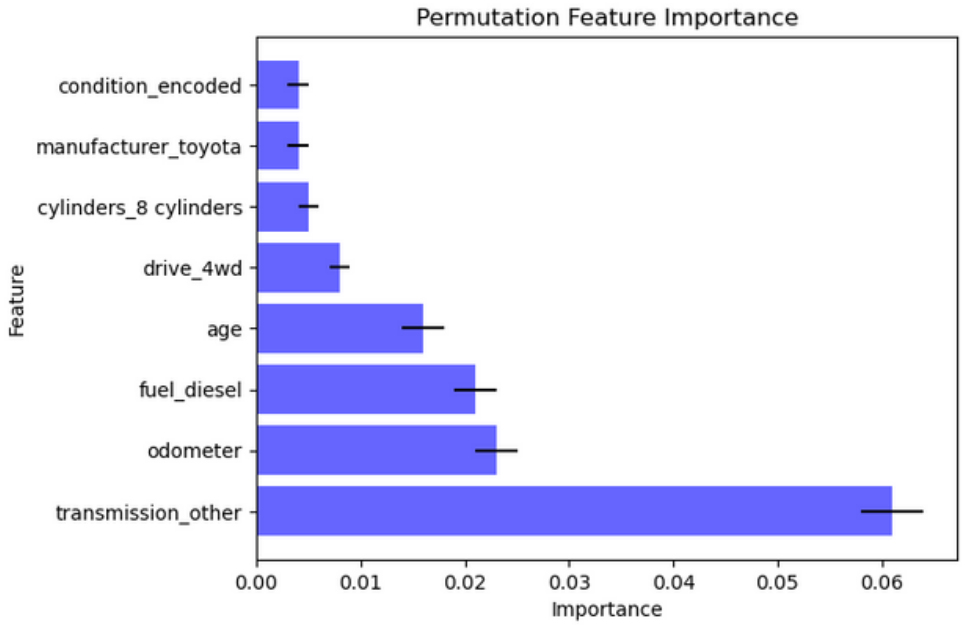


04

Pickup

The best model for predicting pickup price is L2 regularization for Linear Regression with Polynomial Features degree 2, alpha 100, 8 features
Training MSE: 4.9798
Testing MSE: 5.0702

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a pickup. The variable "transmission_other" has the highest R2 score of 0.061, with a standard deviation of 0.003. This means that this variable has a relatively strong relationship with the pickup price in the regression model. The variable "odometer" has a lower R2 score of 0.023, with a standard deviation of 0.002, indicating a weaker relationship with the price. The standard deviation for this variable is smaller, which hints that the relationship is more consistent across the data. The variables "fuel_diesel" and "age" have R2 scores of 0.021 and 0.016, respectively, indicating a relatively weak relationship with the price in the regression model. The standard deviations for these variables are also small, indicating that the relationship is more consistent across the data. The variables "drive_4wd", "cylinders_8 cylinders", "manufacturer_toyota", and "condition_encoded" have the lowest R2 scores of 0.008, 0.005, 0.004, and 0.004, respectively, indicating a very weak relationship with the price. The standard deviations for these variables are also small, indicating that the relationship is more consistent across the data.

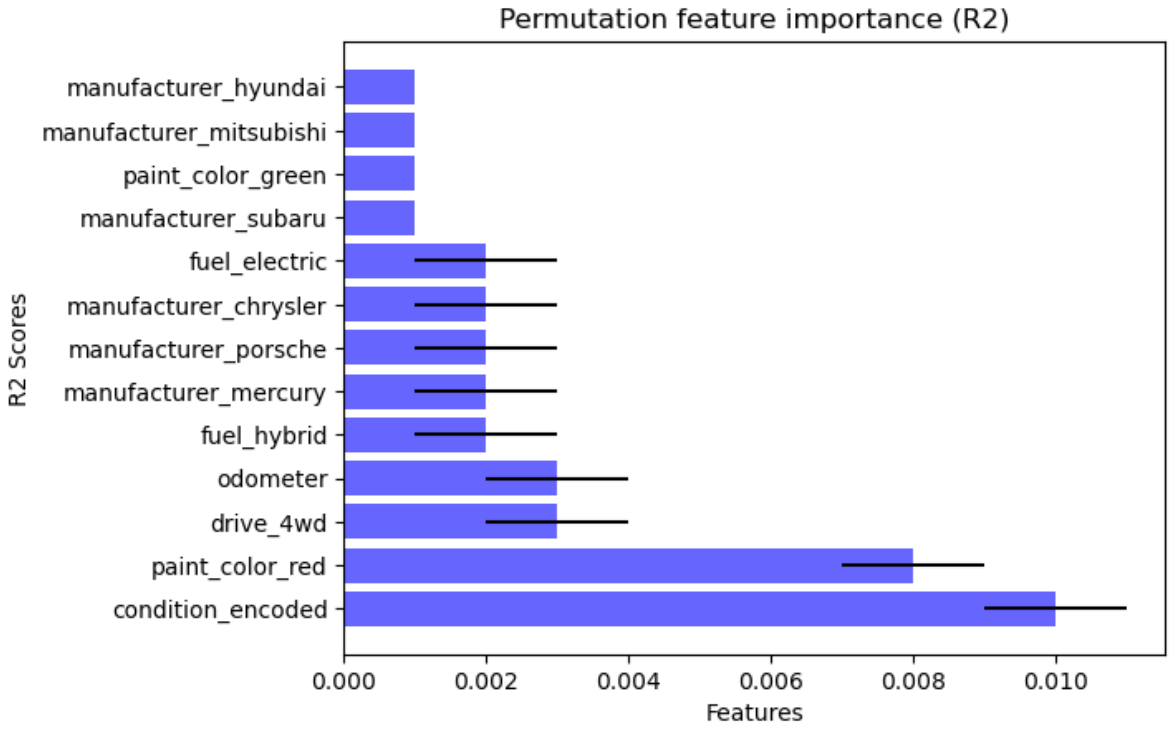


05

Truck

The best model for predicting truck price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 11.536384190802762
Testing MSE: 11.600499719040648

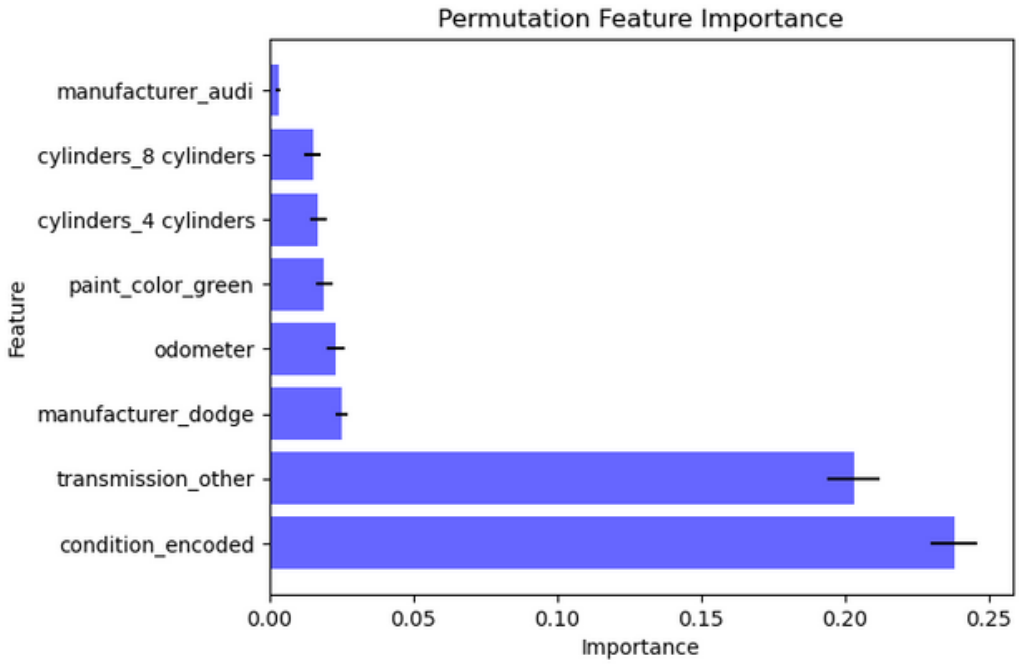
The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a truck. The R2 score of 0.023 suggests that the model is not a good fit for the data and is not performing well in making predictions. The "condition_encoded" feature has the highest importance score with a value of 0.01 +/- 0.001, which means that when this feature is shuffled, the model's performance drops by 0.01 on average. The "paint_color_red" and "drive_4wd" features are the next most important features with scores of 0.008 +/- 0.001 and 0.003 +/- 0.001, respectively. The remaining features all have importance scores less than or equal to 0.002 +/- 0.001, indicating that they have very little impact on the model's performance.



06

Coupe

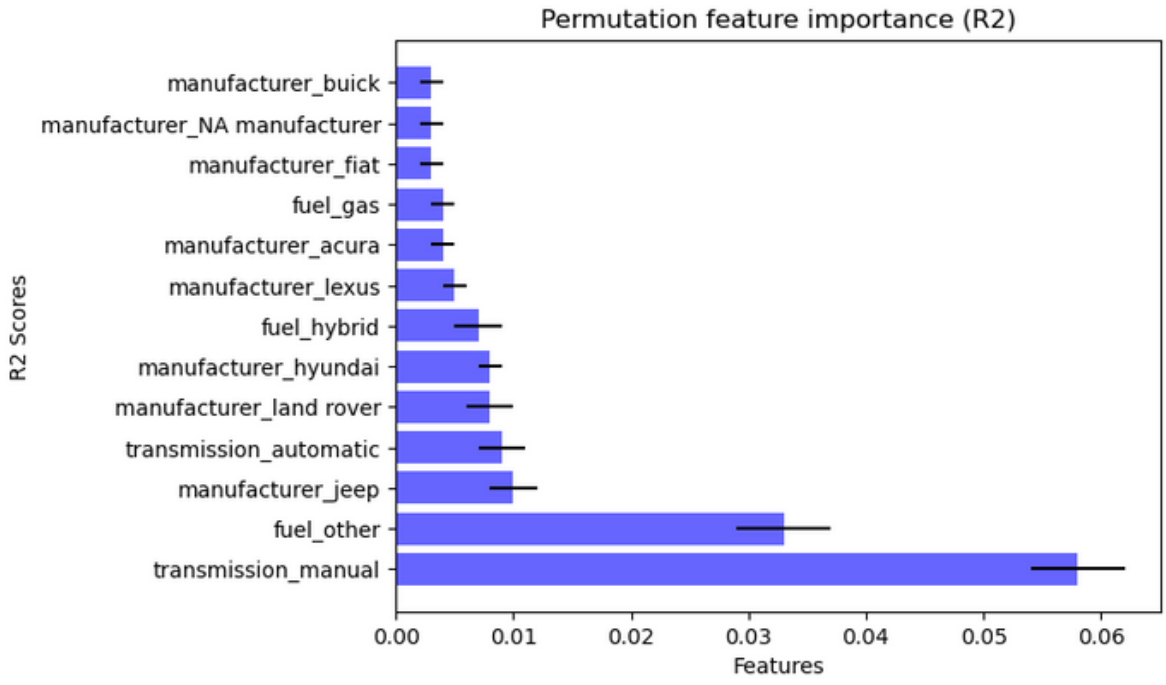
The best model for predicting coupe price is L2 regularization for Linear Regression with Polynomial Features degree 2, alpha 100, 8 features
Training MSE: 5.4912
Testing MSE: 5.2487
The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a coupe. The most important feature for predicting the price has been condition with an importance score of 0.238, and the model is fairly confident in this score, given that the standard deviation is only 0.008. Similarly, transmission_other has an importance score of 0.203, which is the second-highest among all features, and a standard deviation of 0.009, indicating a high level of confidence in the importance score. On the other hand, manufacturer_audi has the lowest importance score among all the features, with a value of only 0.003, and a standard deviation of 0.001. This means that the model does not consider this feature to be very important in making its predictions.



07

Hatchback

The best model for predicting hatchback price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 3.881785168860403
Testing MSE: 3.74148052390479
The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a hatchback. The feature "transmission_manual" has the highest importance among all the features, with an R2 score of 0.058 and a standard deviation of 0.004. This means that the "transmission_manual" has the highest impact on the hatchback's price. Similarly, the feature "fuel_other" has the second-highest importance, with an R2 score of 0.033 and a standard deviation of 0.004. The feature "manufacturer_jeep" has the third-highest importance, with an R2 score of 0.010 and a standard deviation of 0.002.nOn the other hand, the features "manufacturer_fiat", "manufacturer_NA manufacturer", and "manufacturer_buick" have the lowest importance, with R2 scores of only 0.003 and standard deviations of 0.001.

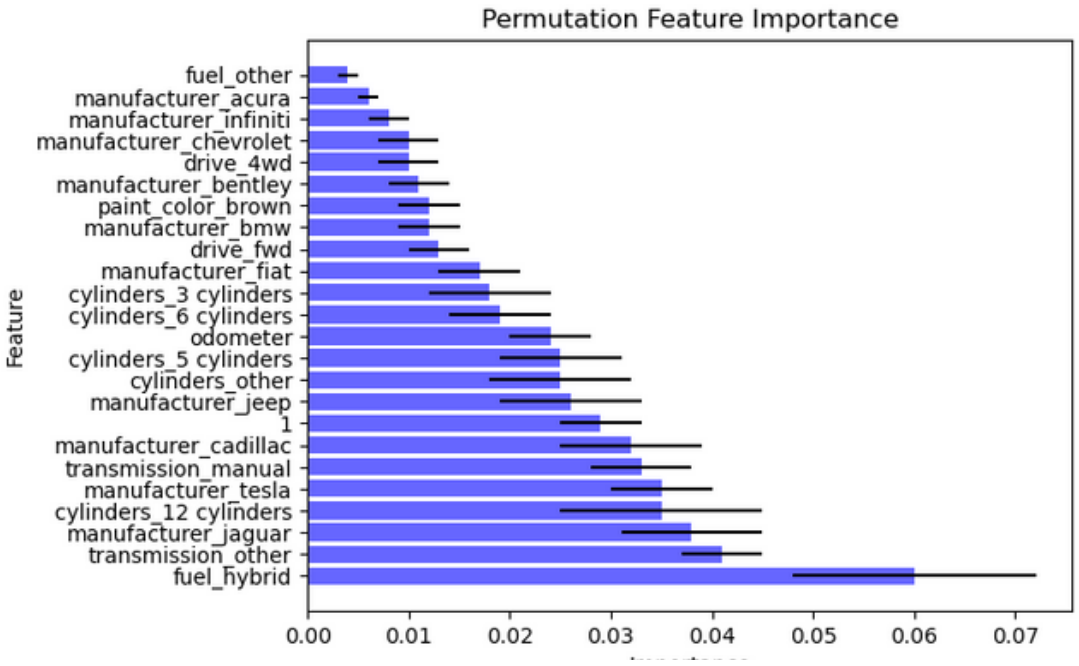


08

Convertible

The best model for predicting convertible price is L1 Regularization (Lasso regression)for Linear Regression with Polynomial Features degree 2, alpha 0.1
Training MSE: 6.19896820073899
Testing MSE: 5.831518584641779

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a convertible. The R2 score is 0.111 means that the model can explain 11.1% of the variance in the target variable. The most important features for the model are 'fuel_hybrid', 'transmission_other', 'manufacturer_jaguar', 'cylinders_12 cylinders', and 'manufacturer_tesla', which have the highest R2 scores. On the other hand, the least important features are 'manufacturer_acura' and 'fuel_other', which have the lowest R2 scores. The standard deviation of the R2 scores for each feature is relatively small, which indicates that the feature importance estimates are reliable.at the model is not highly dependent on any single feature for making predictions.

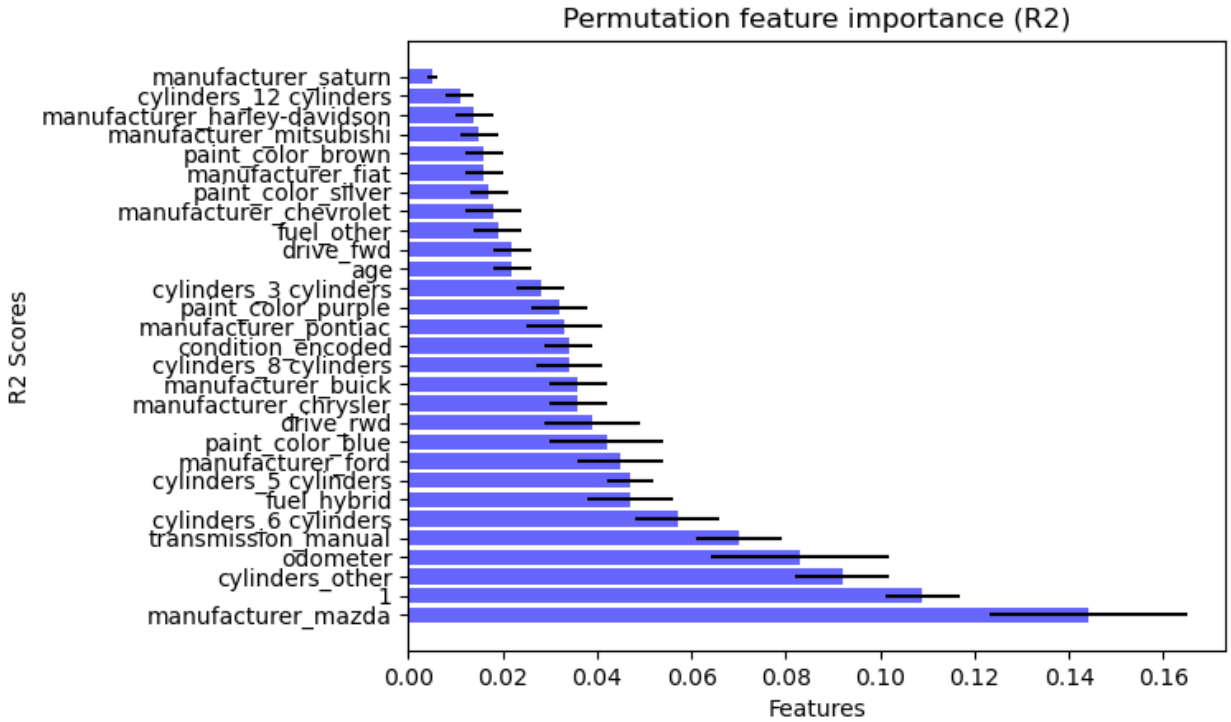


09

Van

The best model for predicting van price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 5.737965567823798
Testing MSE: 6.495383561525091

The R2 score of 0.076 suggests that the model explains only a small proportion of the variance in the target variable: van price. TNevertheless the top five most important features for predicting the van price, in order of importance, are:
* manufacturer_mazda
 * 1 (fair condition)
* cylinders_other
* odometer
* transmission_manual
These features have the highest mean importance values and relatively low standard deviations, meaning that their importance estimates are relatively certain. The other features have lower importance values and relatively higher standard deviations, suggesting that their importance estimates are less certain. However, they may still be important predictors of the van price and should not be ignored.

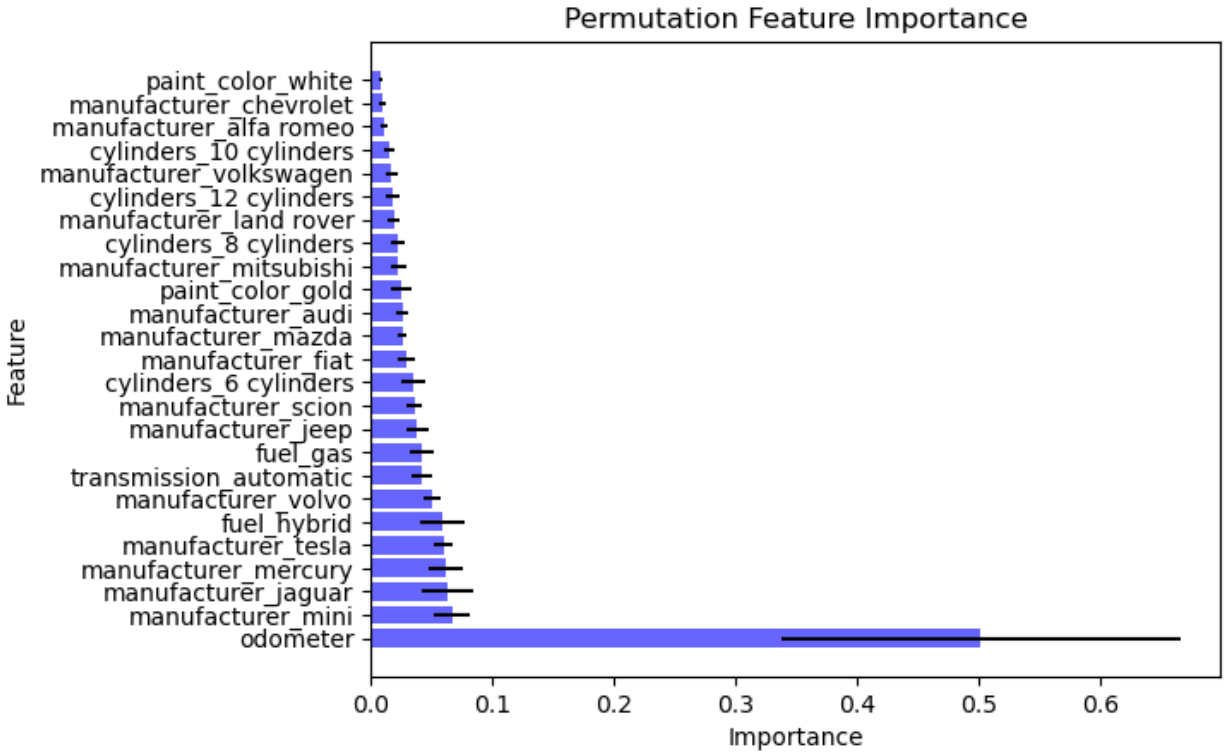


10

Wagon

The best model for predicting wagon price is L2 regularization (Ridge regression) for Linear Regression with Polynomial Features degree 2 and alpha 100
Training MSE: 4.135103866994625
Testing MSE: 5.440542189710208

Since the R2 score can be from -∞ to 1, (1 being a perfect fit between the model and the data), the R2 score of -0.279 might mean that the model performs very poorly and the predicted values are far away from the actual values. Nevertheless the features listed on the chart below are ranked based on their importance. The highest-ranking feature is odometer with a value of 0.501 +/- 0.164. This means that this feature has the most significant impact on the model's performance and therefore wagon's price. The other features have much lower importance values, ranging from 0.011 to 0.067. They have much less impact on the wagon's price. Some of the features with higher importance values include fuel type, transmission type, and number of cylinders, as well as the make and paint color of the car. Given the poor R2 score, it is important to keep in mind that the feature importance values may not be very reliable...

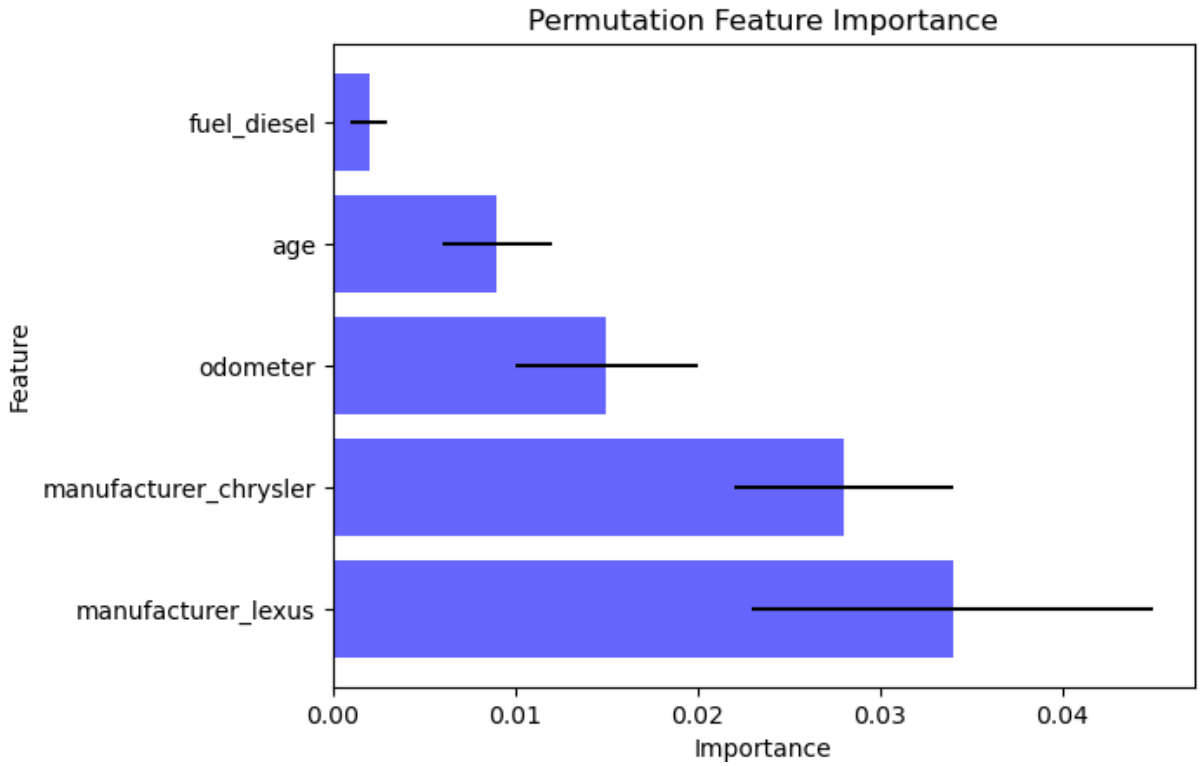


11

Minivan

The best model for predicting minivan price is L1 Regularization (Lasso regression)for Linear Regression with Polynomial Features degree 2 and alpha 0.1
Training MSE: 4.629120453560208
Testing MSE: 5.031026942461045

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a minivan. The most important feature is manufacturer_lexus, with a feature importance score of 0.034 +/- 0.011, meaning that shuffling the values of this feature leads to a decrease in model performance by an average of 0.034 with a standard deviation of 0.011. The second most important feature is manufacturer_chrysler, with a feature importance score of 0.028 +/- 0.006, followed by odometer, age, and fuel_diesel. The smaller feature importance scores for the latter three features suggest that they are less important for the model than the first two features.

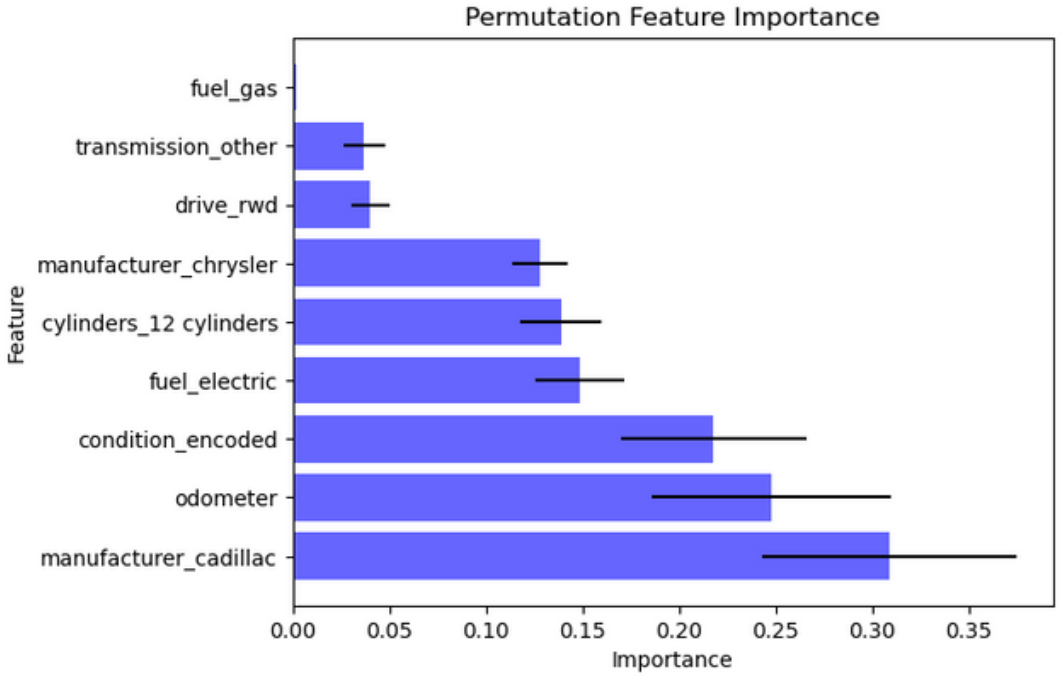


12

Bus

The best model for predicting bus price is L1 Regularization (Lasso regression)for Linear Regression with Polynomial Features degree 2, alpha 0.1
Training MSE: 2.965700698602185
Testing MSE: 3.468088387723858

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of a bus. The R2 score of 0.317 indicates the model is able to explain 31.7% of the variability in the data. The manufacturer_cadillac feature has an importance score of 0.309 with a standard deviation of 0.066, indicating that shuffling the values of this feature had a relatively large impact on the R2 score, and that this feature is relatively important in predicting the price. On the other hand, the fuel_gas feature has a very low importance score of 0.002 +/- 0.000, indicating that shuffling its values had almost no impact on the R2 score, and therefore it is not a significant predictor of the outcome variable.



13

Offroad

The best model for predicting offroad price is L1 Regularization (Lasso regression)for Linear Regression with Polynomial Features degree 3 and alpha 0.1
Training MSE: 2.6626483500226863
Testing MSE: 2.976122056422494

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of an offroad car. A negative R2 score of -0.304, indicates that the regression model is not a good fit for the data and performs worse than a model that simply predicts the mean value of the dependent variable.
The "transmission_other 4.159 +/- 1.111" means that this feature has an importance score of 4.159, with a standard deviation of 1.111. This suggests that the "transmission_other" feature has a relatively high importance in the model's performance. This was the only feature permutation feature importance algorithm returned.

14

Antiques

The best model for predicting antique price is L1 Regularization (Lasso regression)for Linear Regression with Polynomial Features degree 2 and alpha 0.1
* Training MSE: 2.8996873600333326
* Testing MSE: 2.6298832808872454

The results of permutation feature importance indicate the relative importance of each feature in predicting the price of an antique car. The R2 score of the model is 0.175, which means that the model explains only 17.5% of the variability in the target variable: price. The feature "age" has the highest importance, with a score of 0.079, followed by "condition_encoded" with a score of 0.054. The features "cylinders_5 cylinders", "manufacturer_kia", and "cylinders_3 cylinders" have moderate importance, with scores between 0.034 and 0.035. The remaining features have relatively low importance, with scores between 0.006 and 0.021. These scores indicate that the model is not highly dependent on any single feature for making predictions.