**Business Understanding:**

The objective of this exercise it to construct a regression model for predicting used car prices, facilitating fair market pricing, and aiding pricing strategies and product development for used car dealerships.

**Data Understanding:**

The original dataset contained information on 3 million used cars. The provided dataset contains information on 426K cars. Given the large nature of the dataset and the restricted resources that I have on my laptop and JupyterLab, I was forced to work with a fraction of the data. In addition, after careful analysis of the data set and market research of the used car industry, and using CRISP-DM methodology, it made sense to fragment the data into more realistic/useful business case. For example, demand is very localized. Demand for cars in CA where the weather is mild and steady and higher gas prices is very different from states like CO or IA. Demand for hybrids, more gas efficient vehicles here in CA, tend to push their demand and value higher. Whereas in CO due to the snow, less expensive gas, and rural nature, 4X4s, pick-up and SUV are on higher demands.

The new dataset, containing data for vehicles only for CA, encompasses initially 51,614 used car listings with 12 attributes, including price, make, model, year, mileage, and condition. Initial analysis unveils a right-skewed price distribution, prevalent makes like Ford and Chevrolet, and varied mileage and condition. Furthermore, 'size' and 'cylinders' are simplified, possibly due to their limited variability or potential correlation with other features. 'Drive' is also removed as it's deemed irrelevant for the analysis based on the geographical scope of the data (focused on California). Handling missing values, especially in the 'cylinders' column, poses a challenge, with the decision to drop rows with missing values potentially leading to loss of information. However, by dropping these rows, it ensures the integrity of the remaining data and prevents biased analysis.

**Data Preparation:**

To refine the dataset, null values, categorical columns, and irrelevant features like 'id' and 'VIN' are removed. The remaining columns were evaluated graphically and/or statically before removing outliers, or missing values replaced (ex: for colors, missing rows was replaced by “unknown”. I filtered the price and odometer columns to more realistic values. Cars that are “salvaged or parts” were removed since the safety implication or non-working nature. Cars older then 25 years were dropped since they are considered “classics”. Typically, “classic” cars value does not have a negative correlation with their age therefore including would complicate the model.

Box-Cox treatment was applied to the price and odometer columns to reduce the variance. The manufacturer and model columns were target encoded to the mean price since a car value is highly tied to the make and model. A new numerical column, “age” is derived from the year of manufacturing. Categorical variables undergo numerical encoding. Subsequently, after all treatment, the dataset is partitioned into training and test sets.

**Feature Engineering:**

Feature engineering entails target encoding for 'make' and 'model', while categorical variables are subjected to one-hot encoding. Sequential feature selection identifies prominent features.

**Modeling:**

The data was modeled using multiple regression models: Features selection and Ridge regression, and the hyperparameter (n-features, alpha) were tuned via grid search cross-validation.

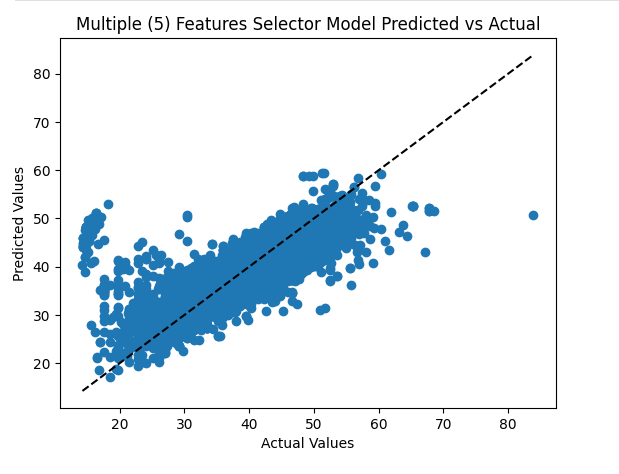
**Evaluation:**

The models were assessed on the test set, achieving a mean squared error (MSE) of 17.81 for the multiple “n-features” selector pipe and 5 features being the optimal features.

Selected Features Coefficient Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Odometer** | **Age** | **Model** | **Condition Fair** | **Fuel Diesel** |
| -0.00371 | -0.554558 | 0.843563 | -6.300303 | 2.69708 |

Predicted Values Vs Actuals:

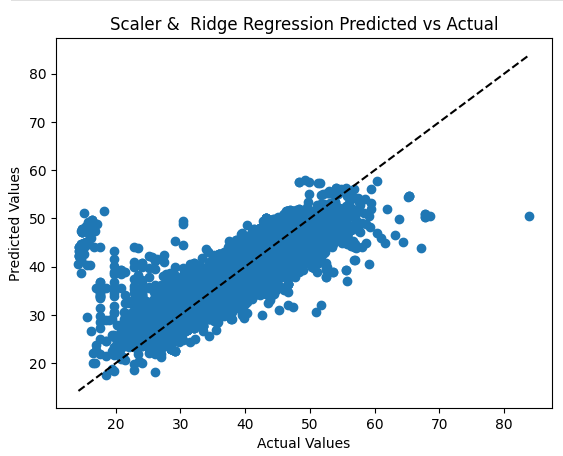


For the optimized “alpha” scaled Ridged regression the MSE was 16.99. Permutation importance analysis gauges feature significance as seen on the table below with year/age, odometer, model, and manufacturer being the highest.

Ridge regression Permutation Importance Table (top 6)

|  |  |  |
| --- | --- | --- |
|  | **Column** | **Permutation Importance** |
| 1 | year | -121.20 |
| 2 | odometer | -203.51 |
| 3 | age | -121.20 |
| 4 | manufacturer | -2.07 |
| 5 | model | -61.27 |
| 6 | Paint color yellow | 6.97 |

Plot of Predicted Vs Actual for the scaled and Ridge Regression Model:



**Deployment:**

The model's coefficients and selected features are extracted, primed for deployment as an API for potential production use.

**Conclusion:**

Leveraging the CRISP-DM methodology, a linear regression model is fine-tuned for used car price prediction, delivering commendable performance, and furnishing insights into pricing determinants.

Next Steps and Recommendations:

Future investigations may explore additional feature inclusion or experiment with alternative regression algorithms to further enhance model efficacy and deepen comprehension of pricing influences.

Findings:

Data preprocessing involves meticulous column curation and data refinement, tailored to cater to localized demand and market dynamics, particularly focusing on California data. Despite challenges such as missing value management and outlier removal, these preprocessing endeavors streamline the dataset, potentially amplifying model performance and interpretability.