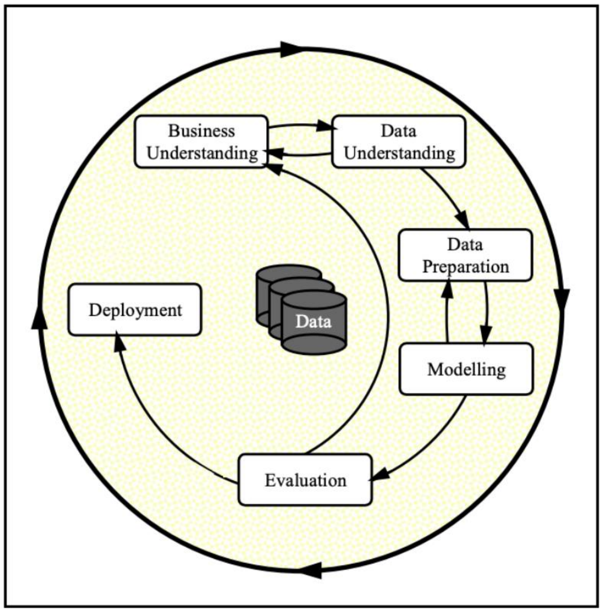
What drives the price of a car?



1 Jupyter Notebook Link

Link to the Jupyter Notebook Link can be found here.

Git Hub Repository link can be found here.

2 Business Understanding

From a business perspective, we are tasked with identifying key drivers for used car prices. In the CRISP-DM overview, we are asked to convert this business framing to a data problem definition. Using a few sentences, reframe the task as a data task with the appropriate technical vocabulary.

2.1 Data Problem Definition

- Used car prices are driven by several factor like usage, car condition, clean title, size, year of manufacturing, manufacturer etc. Based on previous sales, deduce the key parameters that affect the car price and understand customers preferences.

- Based on this understanding, make recommendations to the car dealer on

1. How to estimate the price of used cars?

2. What kind of cars to stock in their inventory?

3 Data Understanding

After considering the business understanding, we want to get familiar with our data. Write down some steps that you would take to get to know the dataset and identify any quality issues within. Take time to get to know the dataset and explore what information it contains and how this could be used to inform your business understanding.

3.1 Exploratory Data Analysis Approach

- Identify the different features in the dataset
- Identify the key features that impact price
- Univariate and Bivariate Analysis
- Correlation Analysis

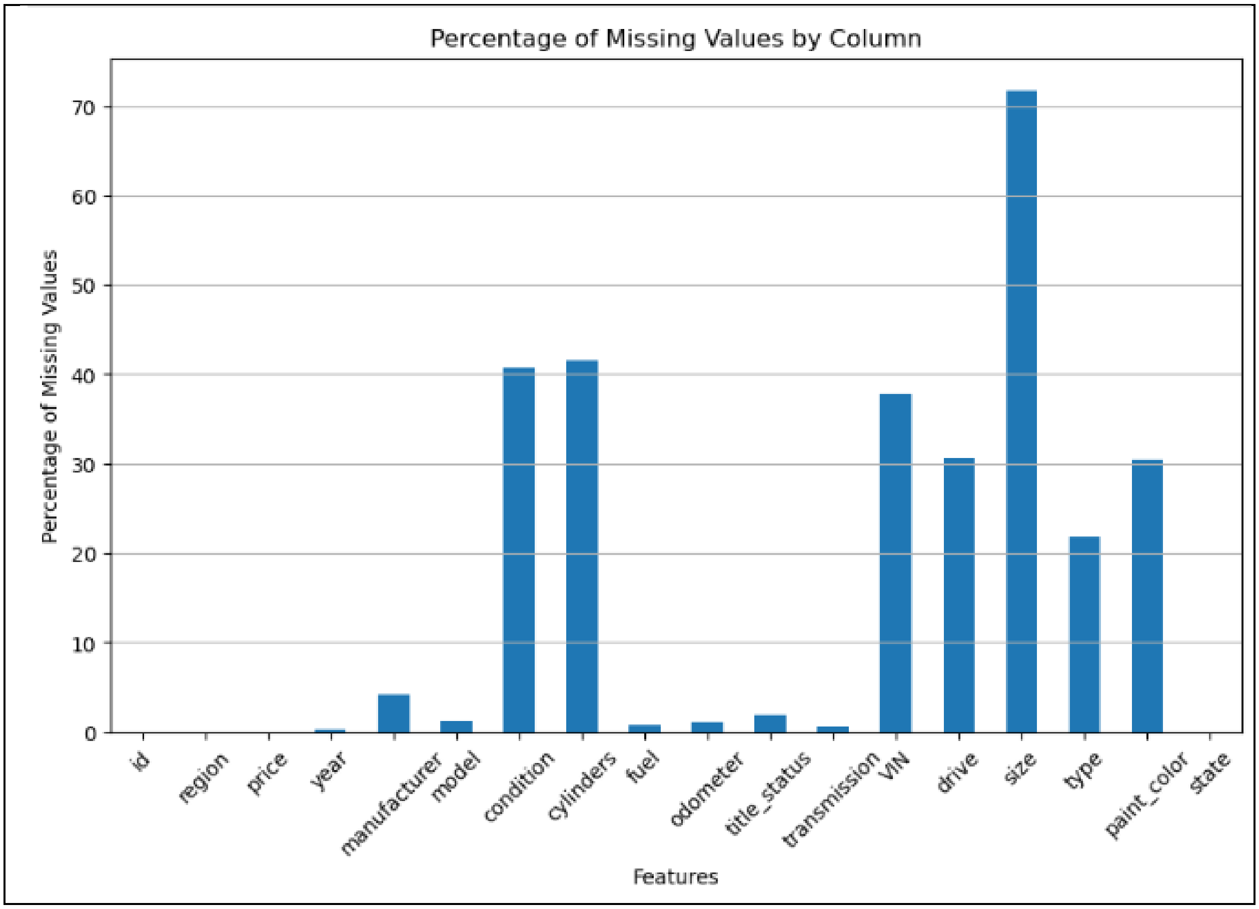
3.2 Data Analysis

- There are 17 features that could potentially determine the price of the used car
- There are 4 columns that have numeric data type: id, price, year of manufacturing and odometer reading
- There is data missing for different features
- There is invalid data like price value of 0, which is not practical

- The features id and VIN, will not influence the cost of the price and will not be included in the model
- Features condition, cylinders, VIN, drive, size, type, paint_color, have more than 30% of the data missing.

3.3 Missing Data Analysis

- Plot below describes the percentage of missing data for all the columns.
- For the price column and other features with less than 10% of missing data; we can remove the missing values and still have enough data for analysis



3.4 Categorical Data Exploration

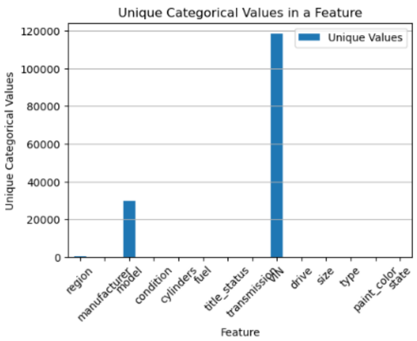
- Estimate the number of unique values within each categorical column.
- Based on this, we can consider if we want to include these features within the analysis.

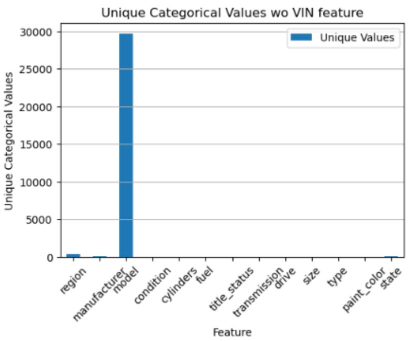
- We can eliminate `VIN`, `model`, and `region` from the analysis, since there are too many categories within the feature

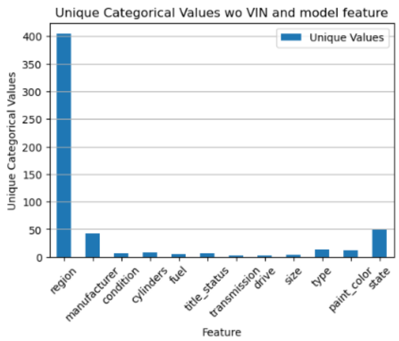
3.5 Numerical Data Exploration

3.5.1 Correlation Analysis

- As odometer reading increases, price decreases
- As the age of the car increases, price decreases







3.6 Feature Selection

1. Id: No; similar to index

2. Region: No; 404 different regions; Consider using state instead

3. Year: Yes

4. Manufacturer: Yes; Some Manufacturers have higher car resale value. Even though there are many, it is worth including it in the model

5. Model: No; 29650 different models

6. Condition: Yes

7. Cylinders: No; Too many features to consider, this does not seem critical

8. Fuel: Yes

9. Odometer: Yes

10. Title Status: No: Most of the data has the same Title Status: clean. Not much variance in the data to include this feature

11. Transmission: Yes

12. VIN: No

13. drive: Yes

14. size: No; Use Type instead; 70% of missing values

15. Type: Yes

16. Paint Color: No; Too many features to consider, this does not seem critical

17. State: No; Too many features to consider

4 Data Preparation

After our initial exploration and fine-tuning of the business understanding, it is time to construct our final dataset prior to modeling. Here, we want to make sure to handle any integrity issues and cleaning, the engineering of new features, any transformations that we believe should happen (scaling, logarithms, normalization, etc.), and general preparation for modeling with `sklearn`.

4.1 Data Analysis Approach

- Find critical parameters based on intuition and real-world understanding of the problem.
- Numerical data:
  - Analyze data distribution using box plots and the `describe` function. Remove outliers if any. Adjust the lower and upper bounds based on common sense. For instance, 100K USD for a used Toyota is not realistic.
  - Look for missing values using the `isna` function. Estimate the percentage of missing values. If the percentage of missing data is small, remove the rows with missing values.
  - Look for duplicates using the `duplicated` function. Remove the duplicates using `drop_duplicates`.
- Non-numerical data:
  - Remove categories with many features.
  - Based on intuition, choose some of the non-numerical data that will be critical for price det For eg, ID will not influence the car price estimation.
  - Identify the unique values within the feature.
  - Remove Duplicates and Nan values

4.2 Dataset Cleaning Procedure

4.2.1 Price Data Clean up

- Remove outliers,
- Upper bound: Q3 + 1.5 * IQR

- Lower Bound of Q1 - 1.5 * IQR results in negative results
- Lower Bound is set to 2000, which is below the 15% percentile. A car value of 0 USD is not logical. For the car dealer a valure below 2000 USD is not a very profitable deal too.
- Remove Missing Values: This is done in the next section after feature selection
- Remove Duplicate Values: This is done in the next section after feature selection

4.2.2 Odometer Reading Data Clean Up

- Number of data points above 0.3e7 reading is 115.
- It is a small percentage compared to the entire data set. Removing odometer values above the value 0.3e7 in the analysis.

5 Modeling

5.1 Modeling Approach #1:

- The critical parameters that will influence the cost of the car are chosen as described in Table 1
- Clean Dataset: Remove outliers, invalid data, missing data and duplicates
- Determine the polynomial order for optimum rmse for the following models
- Linear Regression
- Ridge Regression: determine the optimum alpha parameter
- Lasso Regression: determine the optimum alpha parameter

Table 1: Feature selection for different models

Model#	**Features**	**RMSE (USD)**
Model #1	Year, Odometer	8600
Model #2	Year, Odometer, Manufacturer	6582
Model #3	Year, Odometer, Condition	6747
Model #4	Year, Odometer, Fuel	7051
Model #5	Year, Odometer, Drive	6415
Model #6	Year, Odometer, Type	6283
Model #7	Year, Odometer, Transmission	7410
Model #8	Year, Odometer, Transmission, Condition, Type	5552
Model #9	Year, Odometer, Transmission, Condition, Type, Manufacturer, Fuel, State	7356
Model #10	Year, Odometer, Transmission, Condition, Type, Manufacturer	4763

6 Evaluation

6.1 Model with numerical features analysis; Model #1

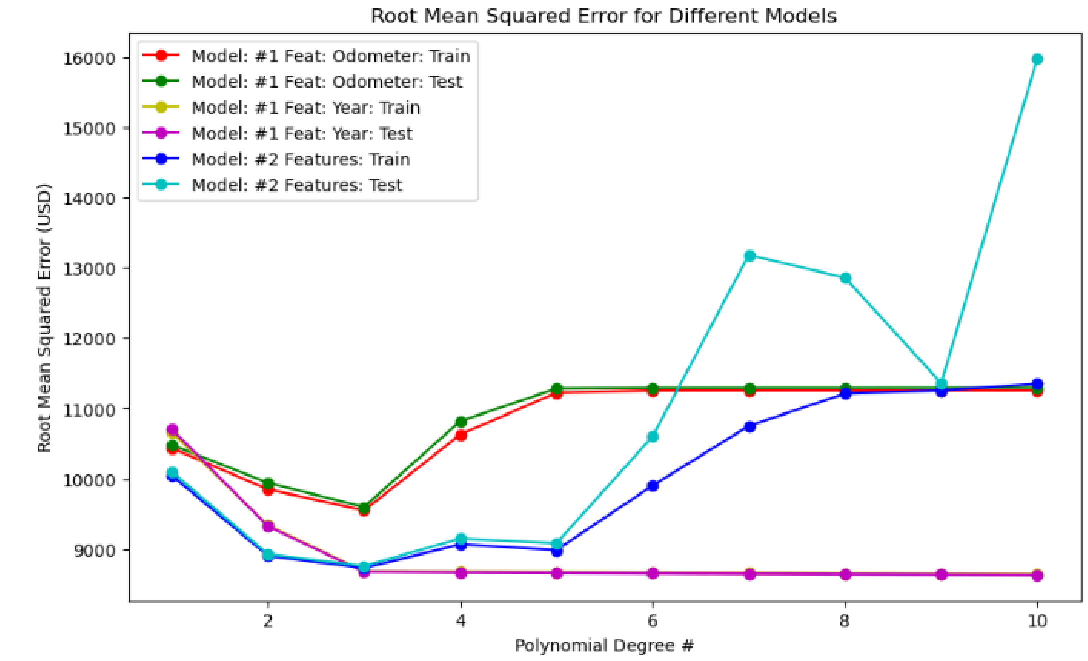


Figure 1: RMSE for Linear Regression model with numerical features

6.2 Key Takeaway

- Model with single feature "year" with polynomial degree 3 and higher has the least RMSE compared to the model with single feature odometer
- Model with two features "year and odometer" with polynomial degree 3 is comparable to the model with single feature year with polynomial order 3. We can see overfitting as the order increases further.
- The median price of the price is ~12K USD. The RMSE is ~8.6K (Figure 1). The error is too high to make accurate prediction
- The R^2 (Figure 2) and MAE (Figure 3) shows the same conclusion
- Ridge (Figure 4) and Lasso (Figure 5) regression did not yield better RMSE

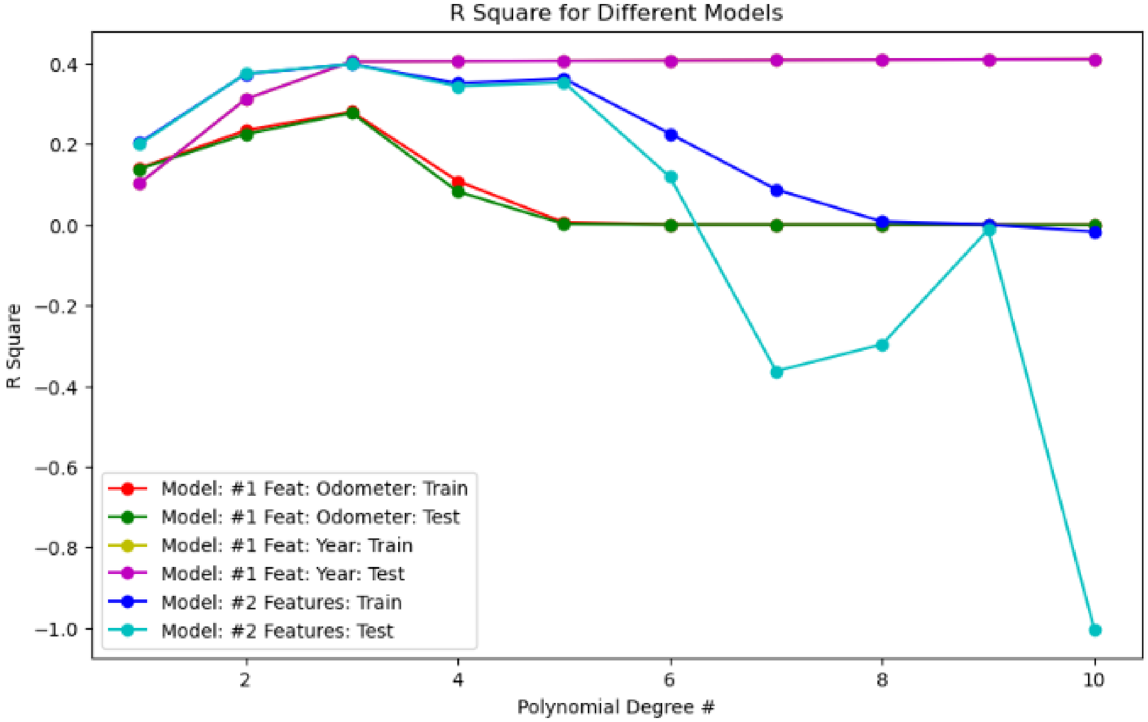


Figure 2: R^2 for Linear Regression model with numerical features

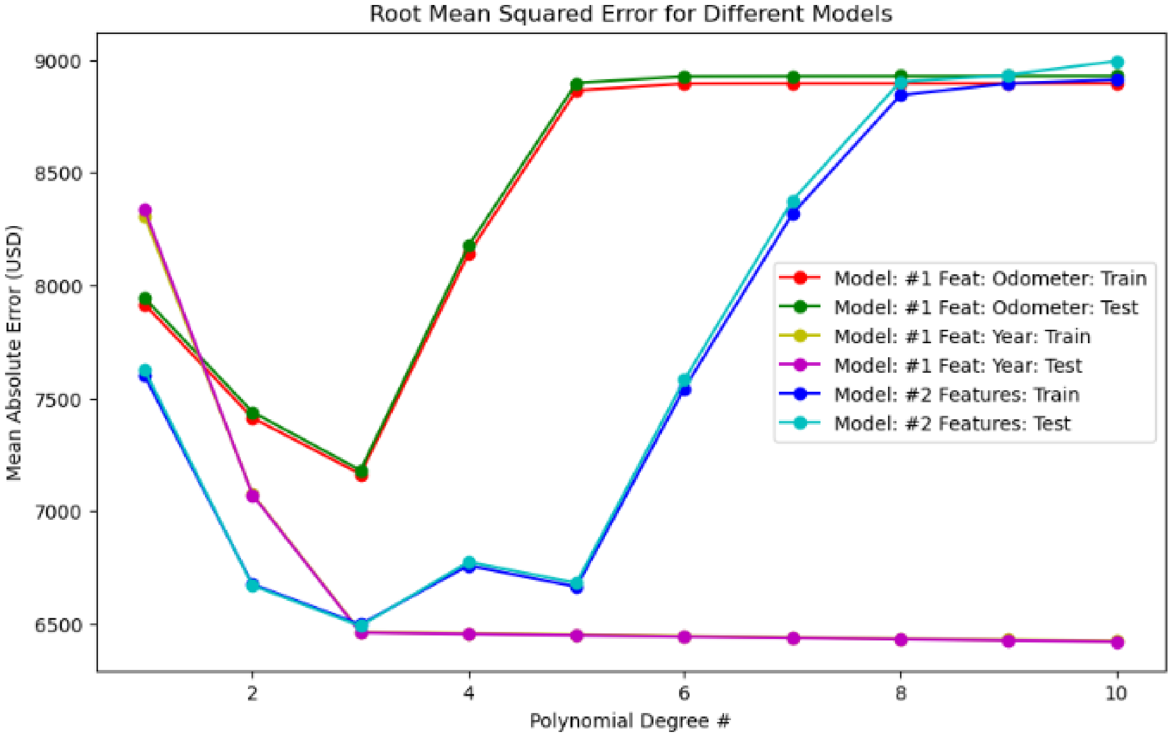


Figure 3: Mean Absolute Error for Linear Regression with numerical features

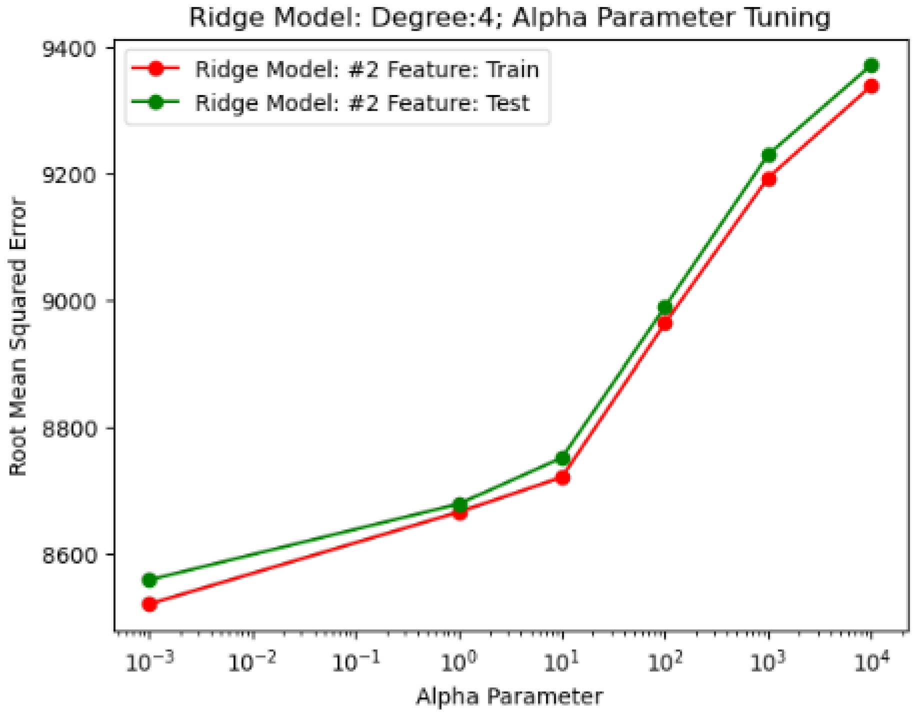


Figure 4: Ridge Model RMSE for Linear Regression with numerical features

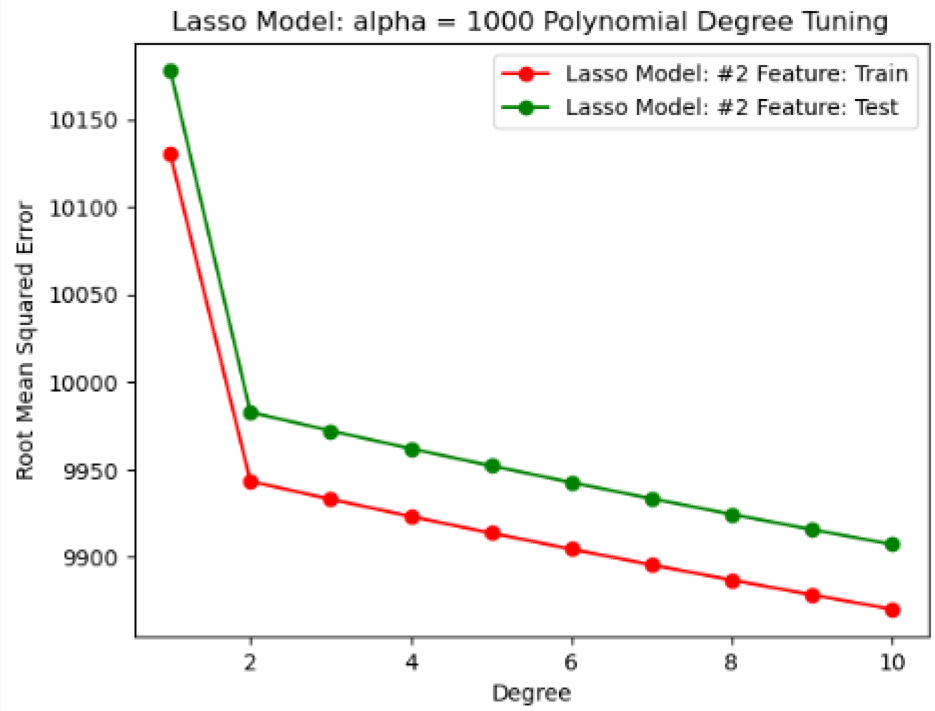


Figure 5: Lasso Model RMSE for Linear Regression with numerical features

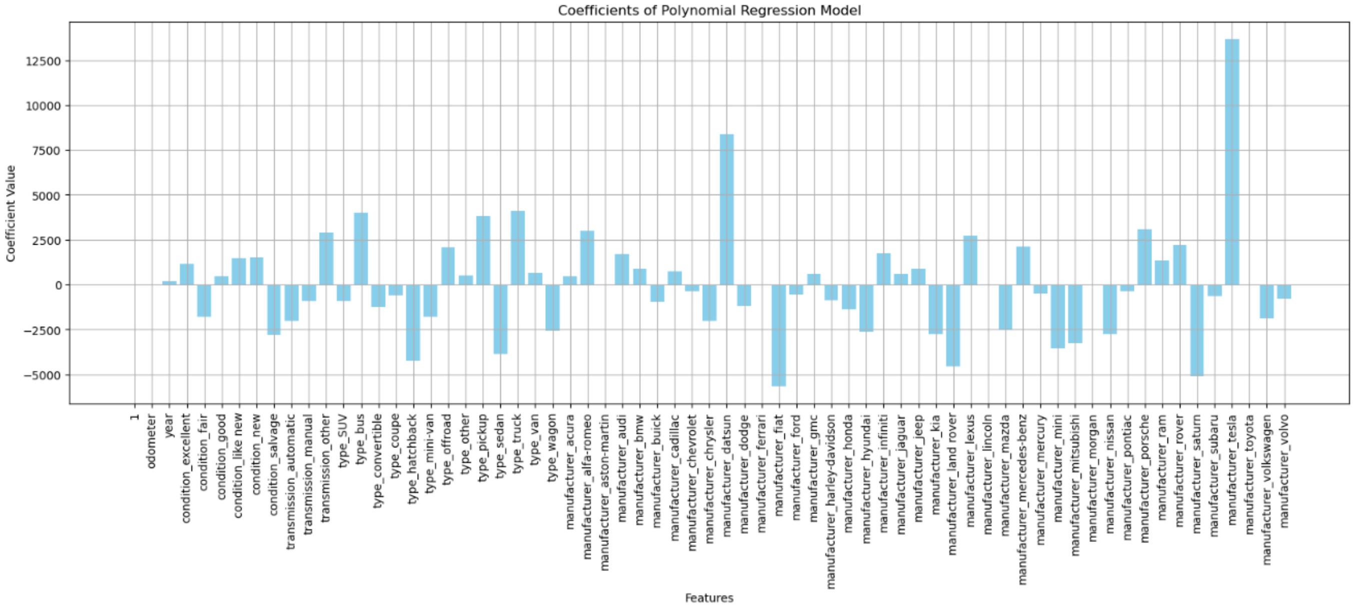


Figure 6: Coefficients of Model#10 with the best RMSE

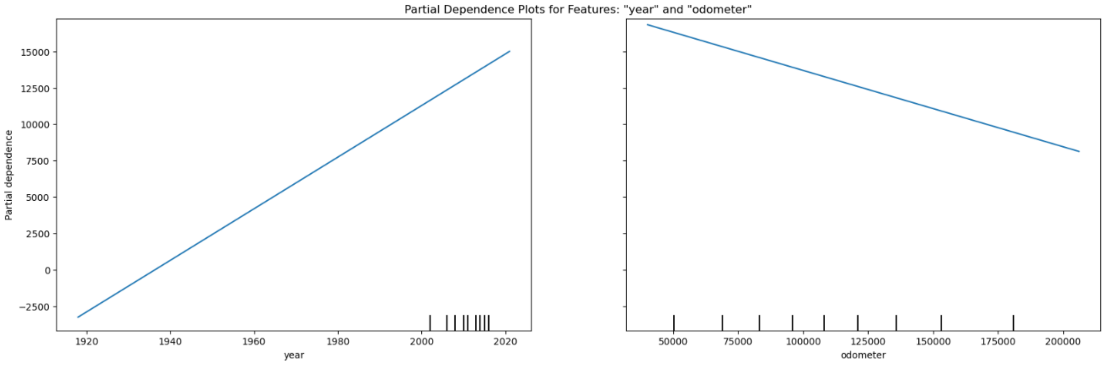


Figure 7: Partial Dependence Plot for the features "year" and "odometer"

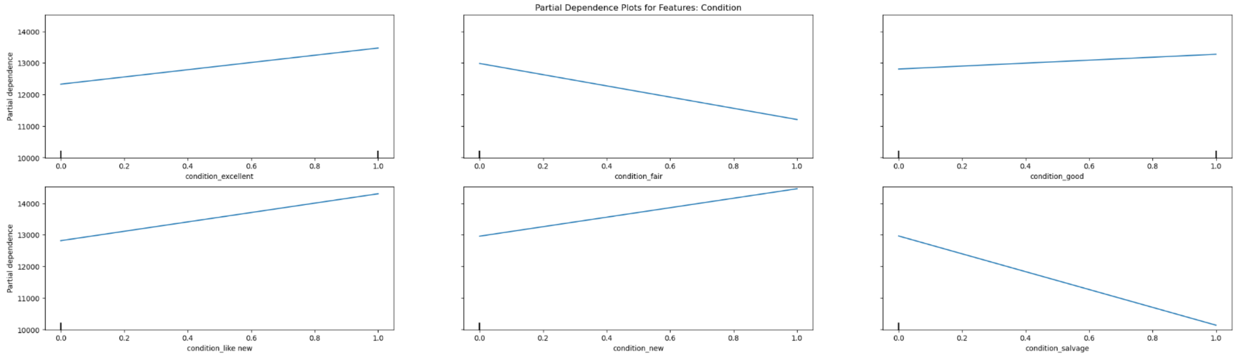


Figure 8: Partial Dependence Plot for the feature "condition"

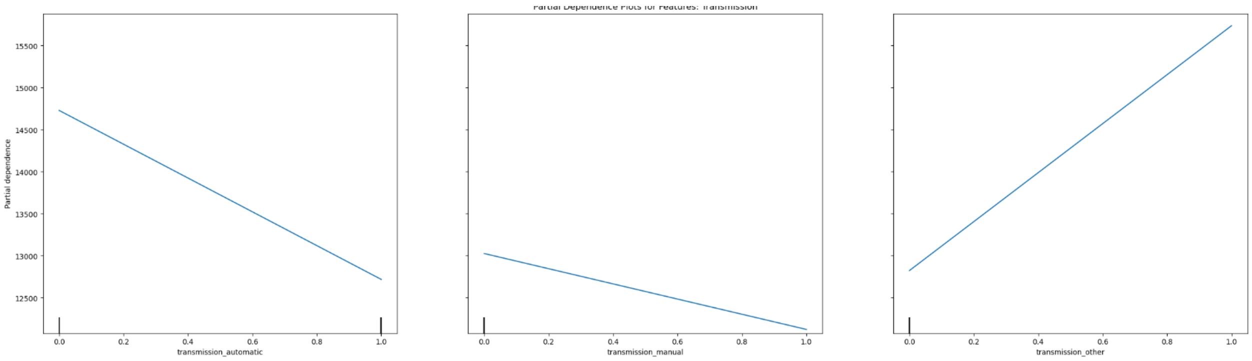


Figure 9: Partial Dependence Plot for the feature "Transmission"

6.3 Model Evaluation Conclusion

- Based on the RMSE obtained from the different models and the delta between RMSE for the test and train data set, model #10 got the least RMSE of 4763 USD
- As odometer value increases car price decreases, as the manufacturing year increases the car value decreases
- As seen the Partial Dependence Display (Figure 7) and from the coefficients, the year had a stronger impact on the car price compared to odometer reading.
- Condition of the car affects the car price (Figure 8). Car in condition good and like new got better price than car with condition fair and salvage condition.

- The effect of transmission was not clear. Since both manual and automatic cars showed decrease in price. The Transmission feature value “transmission_other” does not clearly indicate what kind of transmission it is.
- The effect of type of car and manufacturer is not conclusive.

7 Deployment

7.1 Conclusion

- From the above models, the best model had an error of ~4763 USD. The median price of the car is ~12000 USD.
- This model is not good enough to predict the price of the car with high accuracy.
- There is correlation between car price and year of manufacturing, odometer and condition of the car
- As odometer value increases car price decreases, as the manufacturing year increases the car value decreases
- The year had a stronger impact on the car price compared to odometer reading. Stocking newer cars is recommended
- Condition of the car affects the car price. Car in condition good and like new got better price than car with condition fair and salvage condition.
- The dealers should focus on these factors when they stock the cars in the inventory

7.2 Next Steps

- Use advanced models
- Include sales profit margin information to make appropriate recommendations to the car dealer