Indian Institute of Technology Kharagpur

Regression and Time Series Modelling (MA60056)

Pre-Owned Car Price Prediction

Prof. Buddhananda Banerjee

Course Project
PGDBA 2023-25

Name	Roll Number
AMBUJ GAHOI	23BM6JP06
KAPIL BHANWARILAL	23BM6JP20
NIKUNJ GUPTA	23BM6JP30
PARAS NIGAM	23BM6JP34
PREMJEET CHAUBEY	23BM6JP38
SAURABH CHAUDHARY	23BM6JP50

Contents

Pre-Owned Car Price Prediction

1 Problem Statement

Used/pre-owned car market is one of the fastest growing markets with billions of dollars in revenue. Until recently, the market of used cars was fragmented and unorganized. However, in recent years, big conglomerates have entered the market along with the introduction of new players in the online segment. Indian used car market is expected to touch the USD 100 billion-mark in the next ten years on the back of rising disposable incomes and a growing middle class. The anticipated growth of the Indian used car market is set to occur at a 15 per cent CAGR, increasing from USD 25 billion in 2023 to USD 100 billion by 2034.

There are a lot of factors/characteristics that are considered while deciding the price of used cars like model type, year of the first purchase, mileage, engine etc. Therefore, it is necessary to model used car prices in terms of these features. In this project, we did analysis on used cars data and finally predict price of the car using linear regression models. The steps followed were: Exploratory Data Analysis, Feature Engineering, Multiple Linear Regression on the whole data, Feature Selection through forward selection and backward elimination, Regularization, Residual analysis and Final Prediction on the test set.

2 Data Description

- The dataset of used car prices is taken from Kaggle (link given in the reference).
- There are total 4,009 data points available which contain features of the car sold and the corresponding selling price.
- There are a total of 11 columns in which the last column is the selling price of the car which is the response variable in our project. All the remaining columns form explanatory variables whose impact on the dependent variable will be studied. The explanatory/independent variables are:
    1. Brand: The name of the brand under which the car is sold by the manufacturer. There are a total of 56 unique brands given in the data.
    2. Model: The name and model of the car. There are 1898 unique models given in the data.
    3. Model year: The year in which the car was bought (We can calculate the age using this).
    4. Mileage: The number of mi the car has run.
    5. Fuel Type: The type of the fuel on which the car functions. Fuel types given are gasoline, hybrid, flex fuel, diesel etc.
    6. Engine: Type of the engine with the number of cylinders, capacity of the cylinder and power of the engine.
    7. Transmission: The type of transmission employed by the car (automatic or manual).
    8. Ext Col: External colour of the car.
    9. Int Col: Internal colour of the car.
    10. Accident: Reports whether there was any reported accident of the car or not.
    11. Clean Title: It indicates whether the car was involved in serious damage, flood etc. or not.

Before performing exploratory data analysis, we did web scrapping and added two features – brand country and brand continent since we thought it would affect the price of the car sold. Apart from this, one of the columns in the original dataset ‘Engine’ had information about the number of cylinders, capacity of each cylinder, the type of the engine and its power. Therefore, we segregated the column into 4 columns to extract maximum information and find car price dependence on them. This is useful information that could help us understand better the relation between the engine and the price. Therefore, the added features are:

12. Horsepower: The power of the engine.
13. Litres: The capacity of each cylinder.
14. Cylinders: The number of cylinders used in the engine.
15. Brand Country: Country of origin of the brand.
16. Brand Continent: Continent of origin of the brand.

3 Exploratory Data Analysis

At first glance, we can see that there are very high prices that are possibly outliers. Among these high prices, we can find the Quattroporte Base from the Maserati brand, with a price of nearly 3 million dollars. It is likely that these cases affect our model and make noise instead of a good generalization. If we reduce our dataset to prices up to 200000 dollars, we have a better sample of vehicles with a distribution very similar to a normal one. Doing this reduction, we still have 98% of our original data.

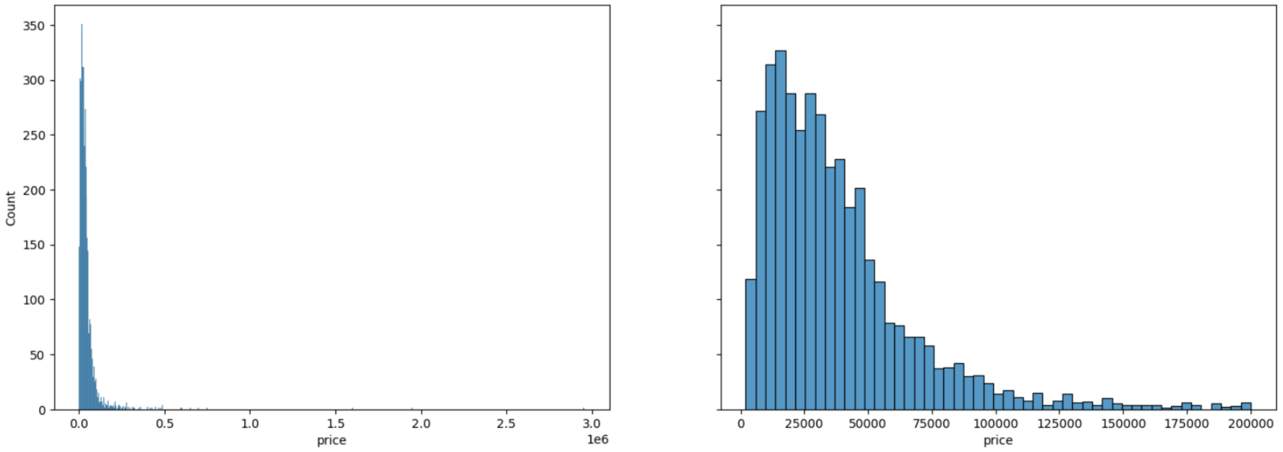
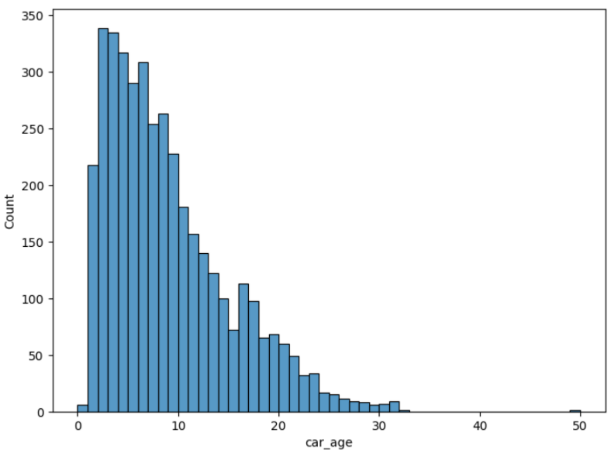
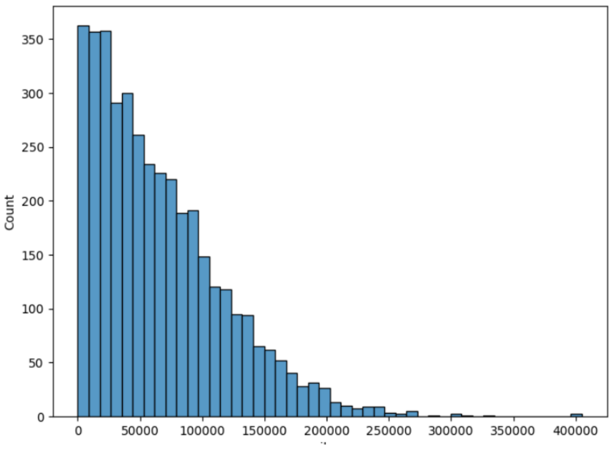


Figure 1: Data Cleaning (Outlier removal)



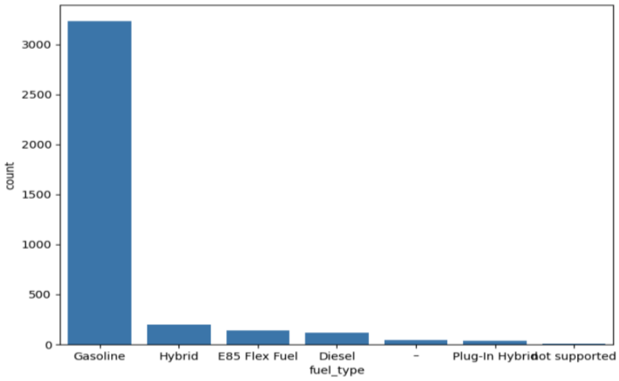
As evident from the left graph, in our dataset, a great proportion of cars are relatively new models, the average being the year 2015.

If we contrast the year of the model with the price, we can see that a positive correlation exists between these two features. In general, the higher the year of the model, the higher the price will be.



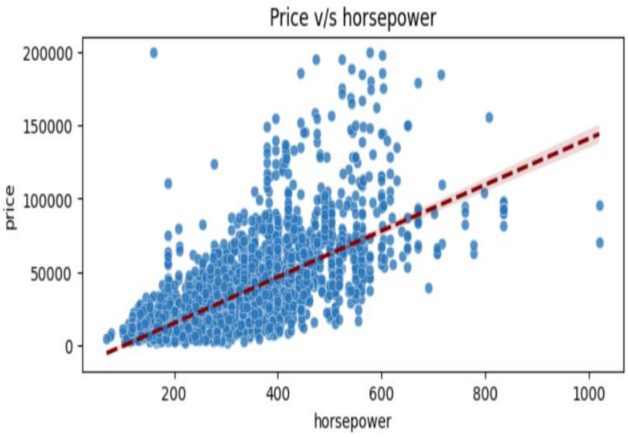
As evident from the left graph, in our dataset, 75% of cars have a value less than 94100 mi. High milage cars are less frequent, probably because these vehicles are ineligible for sale.

We can see that a higher mileage means a lower price. Naturally, a car with high mileage probably does not have good efficiency or in the best condition, so the price must be lower to be more attractive for sale.



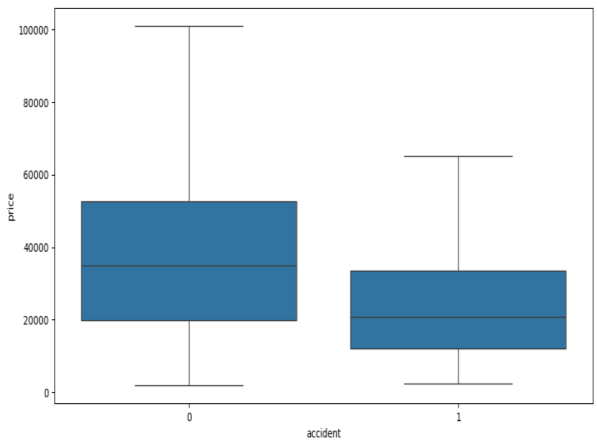
As evident from the left graph, in our dataset, the majority of cars are gasoline-fueled. There are few cars that are hybrid or even electric.

Proportion of cars of diesel type is very low as compared to what we see in middle countries.



As evident from the left graph, in our dataset, from all the engine features, we see that horsepower holds the best positive correlation with the price.

The other features don't give us much insight in relation to price. Therefore, we'll need to find a way of filling the NaNs that are present in this column (20%).



As expected, the cars which have had accident(s) in the past, on an average are priced lower than cars which did not face an accident in the past.

Therefore, we can infer that there will be a correlation between the price and whether the has had an accident or not.

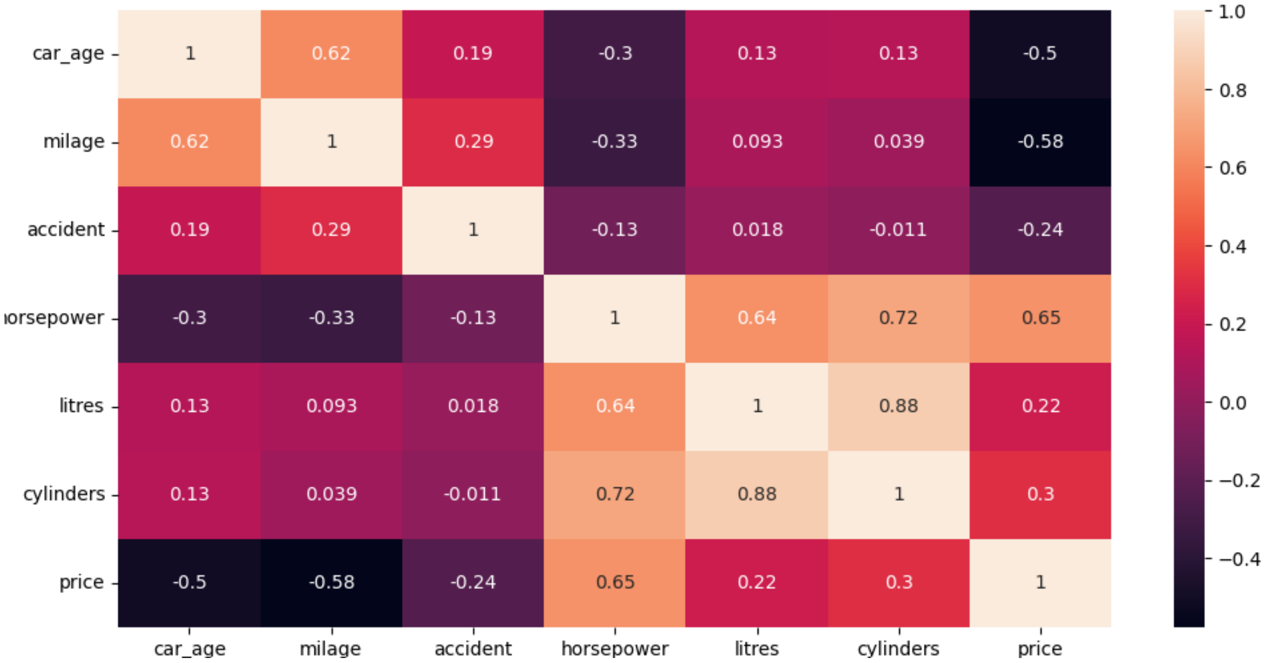


Figure 2: Correlation Heatmap

Through the correlation heatmap, we can observe the following:

1. Car age and mileage are negatively related with the price.
2. Car horsepower has a significant positive correlation with the price.
3. As expected, some engine features like number of cylinders, horsepower and capacity of the cylinder have significant correlation. Therefore, some of these features have to be removed to ensure that there is no multi-collinearity.

4 Feature Engineering

A. Removing skewness from continuous valued features:

We found that in our data, 8 of the features were significantly skewed (all positively). Therefore, we used Box-Cox transformation on these features, so that they closely resemble a normal distribution.

B. We dropped 9 features from the list due to the following reasons:

Dropped Column	Reason
Hp_mean	Intermediate variable used for imputation
Hp_mean_litres	Intermediate variable used for imputation
Hp_mean_brand	Intermediate variable used for imputation
Clean Title	Data was not reliable
Fuel Type	Already considered in fuel variable
External Colour, Internal Colour	Converted to Black, White and Others categories
Engine	Converted to 4 columns as described above
Model	Categorical features with ~1800 categories. It will create a sparse matrix if we do one-hot encoding

C. **One-hot encoding**: All categorical variables (Nominal variables) in the dataset were converted into numerical data using one-hot encoding.

D. **KNN Imputation**: Two continuous valued column Litres and Cylinders have missing values. Therefore, we have used KNN imputation technique to impute these missing values.

E. **Finding multicollinearity using Variance Inflation Factor (VIF):** We calculated the VIF of continuous-valued features. For most of the features, it was higher than the standard range of 6-10 indicating strong multi-collinearity among the features. We will address this issue later in feature selection.

F. **Train-Test Split:** Complete dataset was split into Train and Test dataset for fitting the regression and predicting the values. The train-test split was 80%-20%.

5 Regression Models Used (Using Ordinary Least Squares)

A. **Multiple Linear Regression**

Ordinary Least Squares (OLS) regression is a linear regression technique used to find the best-fit line through a set of data points. In OLS regression, the goal is to minimize the sum of the squared differences between the actual values of the dependent variable and the predicted values from the independent variable(s).

The OLS regression method assumes that there is a linear relationship between the independent and dependent variables. The equation of the best-fit line is given by:

y = b0 + b1x1 + b2x2 + ... + bnxn + e

where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are the coefficients (also known as slopes), and e is the error term. The coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other independent variables constant.

The OLS regression method calculates the coefficients that minimize the sum of the squared errors (SSE) between the actual and predicted values of the dependent variable. The SSE is given by:

SSE = Σ(yi - ŷi)2

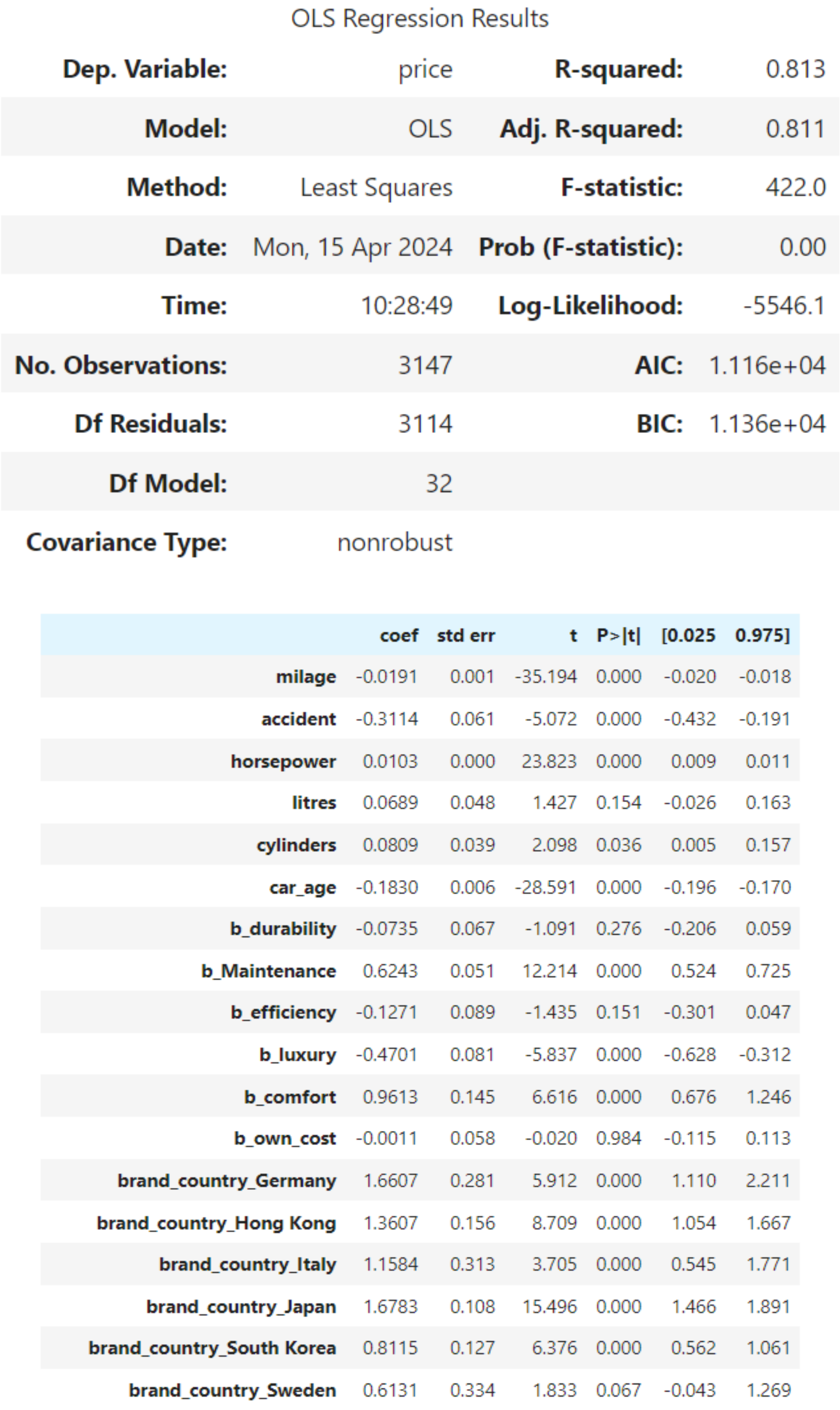
Where yi is the actual value of the dependent variable, and ŷi is the predicted value of the dependent variable based on the independent variables.

The base line model was implemented with all the columns in the dataset.

● Based on results of this, further analysis is done to deal with outliers, multicollinearity, influential points, etc.

● The coefficient values, their standard error, t-statistic value and p-value are attached along with code.

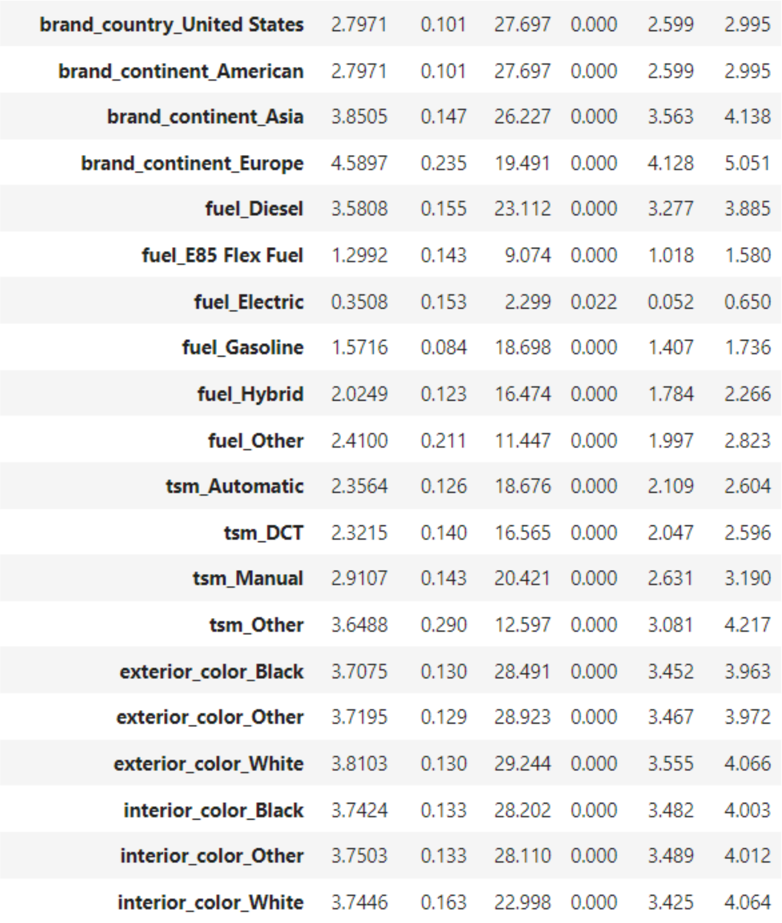
● The key OLS Regression results (R-Squared, F-statistic) are given as below:



OLS Regression Results

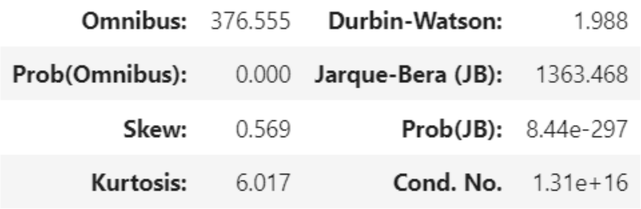
Dep. Variable:	price	**R-squared:**	0.813
Model:	OLS	**Adj. R-squared:**	0.811
Method:	Least Squares	**F-statistic:**	422.0
Date:	Mon, 15 Apr 2024	**Prob (F-statistic):**	0.00
Time:	10:28:49	**Log-Likelihood:**	-5546.1
No. Observations:	3147	**AIC:**	1.116e+04
Df Residuals:	3114	**BIC:**	1.136e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
milage	-0.0191	0.001	-35.194	0.000	-0.020	-0.018
accident	-0.3114	0.061	-5.072	0.000	-0.432	-0.191
horsepower	0.0103	0.000	23.823	0.000	0.009	0.011
litres	0.0689	0.048	1.427	0.154	-0.026	0.163
cylinders	0.0809	0.039	2.098	0.036	0.005	0.157
car_age	-0.1830	0.006	-28.591	0.000	-0.196	-0.170
b_durability	-0.0735	0.067	-1.091	0.276	-0.206	0.059
b_Maintenance	0.6243	0.051	12.214	0.000	0.524	0.725
b_efficiency	-0.1271	0.089	-1.435	0.151	-0.301	0.047
b_luxury	-0.4701	0.081	-5.837	0.000	-0.628	-0.312
b_comfort	0.9613	0.145	6.616	0.000	0.676	1.246
b_own_cost	-0.0011	0.058	-0.020	0.984	-0.115	0.113
brand_country_Germany	1.6607	0.281	5.912	0.000	1.110	2.211
brand_country_Hong Kong	1.3607	0.156	8.709	0.000	1.054	1.667
brand_country_Italy	1.1584	0.313	3.705	0.000	0.545	1.771
brand_country_Japan	1.6783	0.108	15.496	0.000	1.466	1.891
brand_country_South Korea	0.8115	0.127	6.376	0.000	0.562	1.061
brand_country_Sweden	0.6131	0.334	1.833	0.067	-0.043	1.269



brand_country_United States	2.7971	0.101	27.697	0.000	2.599	2.995
brand_continent_American	2.7971	0.101	27.697	0.000	2.599	2.995
brand_continent_Asia	3.8505	0.147	26.227	0.000	3.563	4.138
brand_continent_Europe	4.5897	0.235	19.491	0.000	4.128	5.051
fuel_Diesel	3.5808	0.155	23.112	0.000	3.277	3.885
fuel_E85 Flex Fuel	1.2992	0.143	9.074	0.000	1.018	1.580
fuel_Electric	0.3508	0.153	2.299	0.022	0.052	0.650
fuel_Gasoline	1.5716	0.084	18.698	0.000	1.407	1.736
fuel_Hybrid	2.0249	0.123	16.474	0.000	1.784	2.266
fuel_Other	2.4100	0.211	11.447	0.000	1.997	2.823
tsm_Automatic	2.3564	0.126	18.676	0.000	2.109	2.604
tsm_DCT	2.3215	0.140	16.565	0.000	2.047	2.596
tsm_Manual	2.9107	0.143	20.421	0.000	2.631	3.190
tsm_Other	3.6488	0.290	12.597	0.000	3.081	4.217
exterior_color_Black	3.7075	0.130	28.491	0.000	3.452	3.963
exterior_color_Other	3.7195	0.129	28.923	0.000	3.467	3.972
exterior_color_White	3.8103	0.130	29.244	0.000	3.555	4.066
interior_color_Black	3.7424	0.133	28.202	0.000	3.482	4.003
interior_color_Other	3.7503	0.133	28.110	0.000	3.489	4.012
interior_color_White	3.7446	0.163	22.998	0.000	3.425	4.064

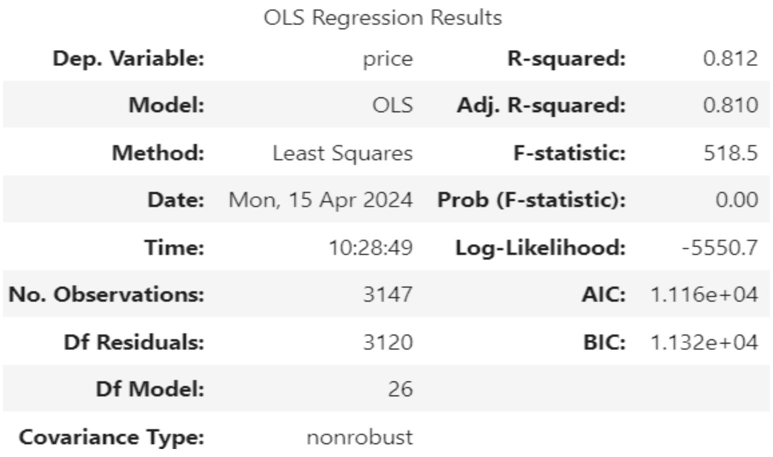
The table below shows a high condition number (1.31e+16) indicating very high multi-collinearity among features. The smallest eigen value is 2.73e-24. This also indicates that there are strong multi-collinearity problems.



Omnibus:	376.555	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1363.468
Skew:	0.569	**Prob(JB):**	8.44e-297
Kurtosis:	6.017	**Cond. No.**	1.31e+16

.

We dropped the features iteratively with non-significant p-values (p>0.05) and fit the model repeatedly with the remaining features. The summary table is as shown below. However, the possibility of multi-collinearity still exists as shown by condition number (1.31e+16) and smallest eigen value (2.72e-24).



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.812
Model:	OLS	**Adj. R-squared:**	0.810
Method:	Least Squares	**F-statistic:**	518.5
Date:	Mon, 15 Apr 2024	**Prob (F-statistic):**	0.00
Time:	10:28:49	**Log-Likelihood:**	-5550.7
No. Observations:	3147	**AIC:**	1.116e+04
Df Residuals:	3120	**BIC:**	1.132e+04
Df Model:	26		
Covariance Type:	nonrobust		

6 Feature Selection (Forward Selection and Backward Elimination)

A. Backward Elimination

Backward OLS (Ordinary Least Squares) regression is a type of stepwise regression that starts with a model that includes all independent variables and gradually eliminates variables that are not statistically significant. The goal of backward OLS regression is to find the best subset of predictors that explain the variation in the dependent variable with the fewest number of predictors. The advantages of backward OLS regression include that it is a systematic approach to variable selection and can help to identify the most important predictors for a given outcome.

After applying backward elimination exhaustively, we find the best combination of features on the basis of cross-validation R-squared value. 21 features were selected out of 40 features after this process. The OLS regression R-squared value improved greatly to 0.995. However, the condition number was found to be high at 5.03e+18.

B. Forward Selection

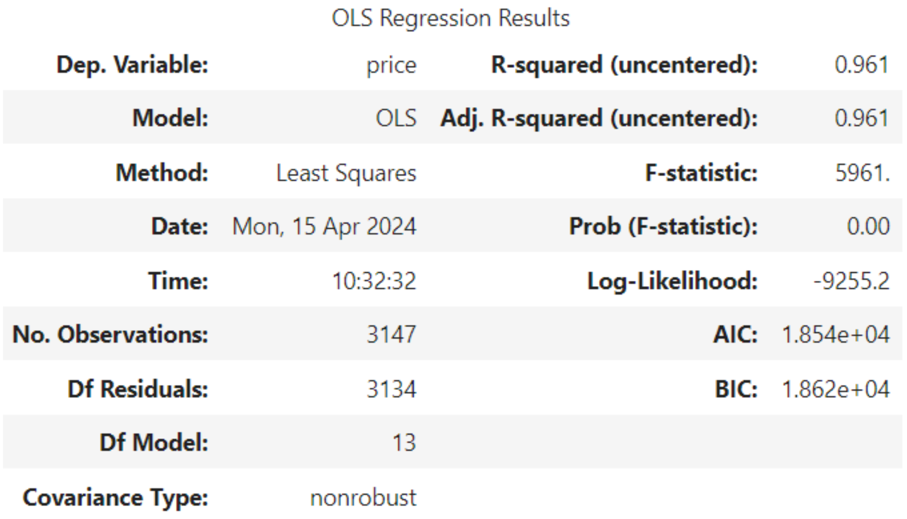
Forward OLS (Ordinary Least Squares) regression is a type of stepwise regression that starts with a model that includes no variables and gradually adds variables that are statistically significant. The goal of forward OLS regression is to find the best subset of predictors that explain the variation in the dependent variable with the fewest number of predictors. The advantages of forward OLS regression include that it is a systematic approach to variable selection and can help to identify the most important predictors for a given outcome.

After applying forward selection exhaustively, we find the best combination of features on the basis of cross-validation R-squared value. 23 features were selected out of 40 features after this process. The OLS regression R-squared value improved greatly to 0.995. The condition number also improved and was observed to be 6.79e+03.

Model	R-squared	RMSE
Forward Selection	0.73	1.64
Backward Elimination	0.73	1.66

C. Handling Multicollinearity through VIF

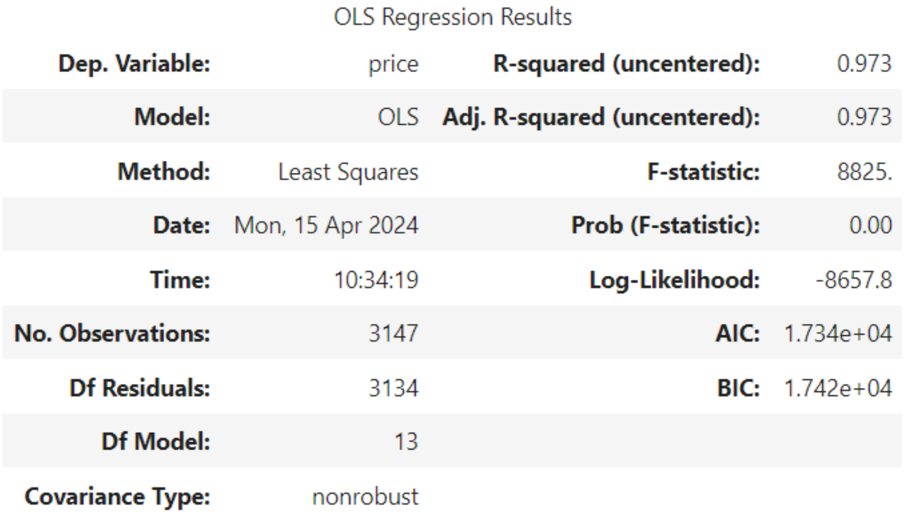
Even after selecting a subset of features from the above techniques, namely forward selection and backward elimination, the condition number of the model is sufficiently high for both of them (even though it has greatly improved from the baseline model for forward selection– from 1.31e+16 to 6.79e+03) and therefore, it still indicates a possibility of multi-collinearity. We use VIF scores again on the selected subset of features by both the techniques and sequentially eliminated features with highest VIF scores (>10). The results of the OLS after performing B and C are displayed below. The R-squared value improved to 0.961.



OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.961
Model:	OLS	**Adj. R-squared (uncentered):**	0.961
Method:	Least Squares	**F-statistic:**	5961.
Date:	Mon, 15 Apr 2024	**Prob (F-statistic):**	0.00
Time:	10:32:32	**Log-Likelihood:**	-9255.2
No. Observations:	3147	**AIC:**	1.854e+04
Df Residuals:	3134	**BIC:**	1.862e+04
Df Model:	13		
Covariance Type:	nonrobust		

The results of the OLS after performing A and C are displayed below. The R-squared value improved to 0.973.



OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.973
Model:	OLS	**Adj. R-squared (uncentered):**	0.973
Method:	Least Squares	**F-statistic:**	8825.
Date:	Mon, 15 Apr 2024	**Prob (F-statistic):**	0.00
Time:	10:34:19	**Log-Likelihood:**	-8657.8
No. Observations:	3147	**AIC:**	1.734e+04
Df Residuals:	3134	**BIC:**	1.742e+04
Df Model:	13		
Covariance Type:	nonrobust		

Therefore, without compromising greatly on the R-squared value, we were able to address multi-collinearity by eliminating features through VIF score. The condition number after applying B and C reduced greatly after filtering on the basis of VIF score from 5.03e+18 to 2.91e+03.

Note: We can observe that the condition number is still sufficiently high. There can be two reasons for that: multi-collinearity or any other numerical problems. Since we have addressed the multi-collinearity issue through VIF score, the high condition number can be attributed to large number of features/sparse matrix (caused by large number of categorical variables and their subsequent one-hot encoding).

D. **Evaluating the model performance on test data**

The model was evaluated on the test data. The results are shown below in the table.

Model	R-squared	RMSE
Base Model	0.79	1.46
Forward Selection + VIF	0.27	2.74
Backward Elimination + VIF	-0.78	4.29

We can observe that after implementing the features via forward selection/backward elimination and VIF, we get unsatisfactory results. Specifically, the R-squared value comes out to be negative in case of backward elimination and VIF based filtering which indicates that our prediction in case of this method is worse than the trivial estimate i.e., y_mean of test data. Therefore, we have lost some critical information while applying these methods and eliminating features. The methods don't seem to work well because of the presence of a lot of categorical variables in the data. We will therefore restore those features which have been removed during VIF based filtering.

7 Polynomial Regression

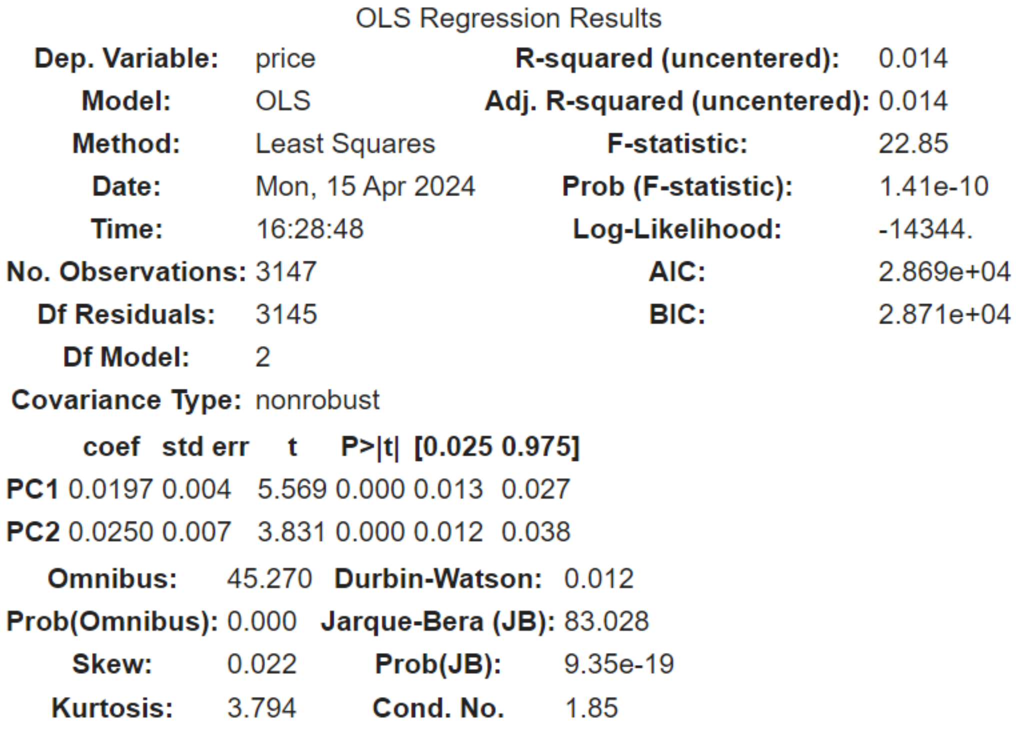
To perform polynomial regression, we introduce polynomial features up to degree 2 in the numerical features of the data. Thereafter, multiple linear regression was performed with OLS on the data and the results are reported below.

Model	R-Squared	R-Squared (Test)	RMSE (Test)
Polynomial	1	0.99	0.06
Poly + Forward	1	0.99	0.15
Poly + Backward	1	0.99	0.11

The results are good on test data. However, these models have some issues like possibility of multi-collinearity indicated by high condition number and possibility of high variance in prediction for out of sample data points. So, these methods are not reliable.

8 Principal Component Regression

Principal component regression is not good when there are a lot of categorical variables, however we are performing for the purpose of experimentation. We found out that only the first 2 principal components explained 99% of the variance present in the data. However, as expected, the R-squared values came out to be very less which reinforces our understanding that Principal component regression is not a good model in presence of lot of categorical variables.



```
                        OLS Regression Results
     Dep. Variable:  price               R-squared (uncentered):  0.014
            Model:   OLS            Adj. R-squared (uncentered):  0.014
           Method:   Least Squares               F-statistic:     22.85
             Date:   Mon, 15 Apr 2024       Prob (F-statistic):   1.41e-10
             Time:   16:28:48                Log-Likelihood:      -14344.
No. Observations:    3147                               AIC:      2.869e+04
    Df Residuals:    3145                               BIC:      2.871e+04
        Df Model:    2
 Covariance Type:    nonrobust
           coef   std err    t    P>|t|  [0.025 0.975]
PC1  0.0197  0.004   5.569  0.000  0.013   0.027
PC2  0.0250  0.007   3.831  0.000  0.012   0.038
      Omnibus:      45.270   Durbin-Watson:    0.012
Prob(Omnibus):      0.000    Jarque-Bera (JB): 83.028
         Skew:      0.022         Prob(JB):    9.35e-19
     Kurtosis:      3.794         Cond. No.    1.85
```

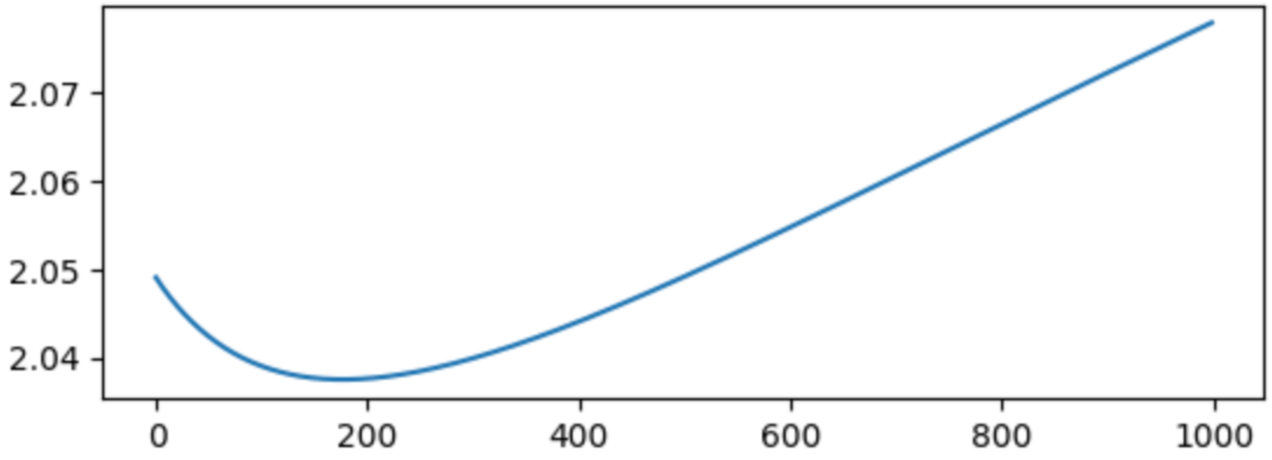
9 Ridge and Lasso Regression

Regularization involves adding a penalty term to the objective function of a model during training, which discourages the model from overfitting by shrinking the parameter estimates towards zero. The penalty term is proportional to the magnitude of the parameters, so that larger parameter estimates are penalized more heavily than smaller ones. Ridge regression is a type of linear regression that is used to deal with multicollinearity (high correlation) among the independent variables. It is a regularization technique that adds a penalty term to the least squares objective function of the linear regression model, in order to shrink the coefficients of the independent variables towards zero.

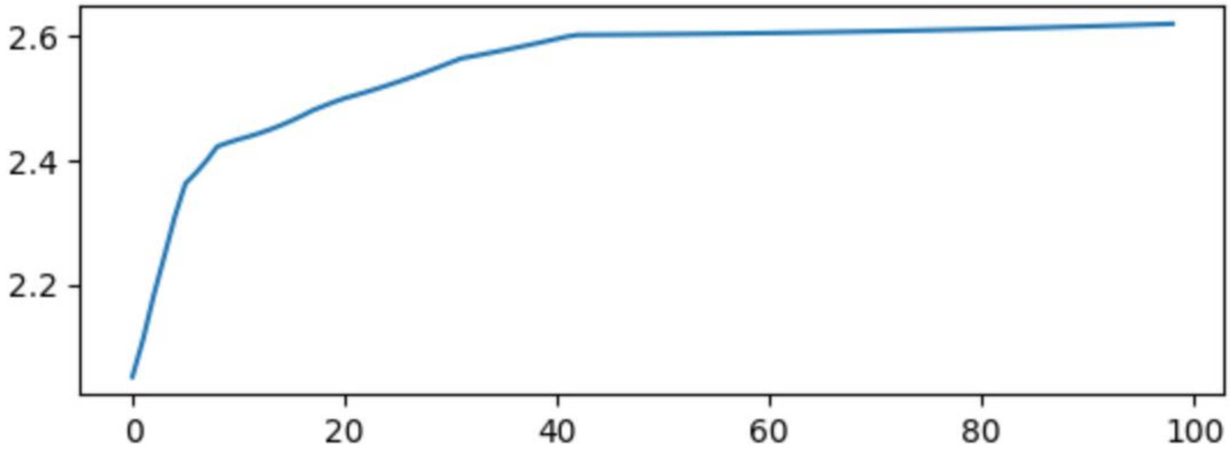
The penalty term, also known as the L2 penalty, is calculated as the square of the sum of the coefficients multiplied by a regularization parameter lambda (λ). This penalty term is then added to the least squares objective function and the resulting model is fit using an optimization algorithm to minimize the sum of squared errors.

The main advantage of ridge regression is that it can improve the stability and accuracy of the linear regression model when dealing with multicollinearity, which can lead to unstable and unreliable coefficient estimates in the traditional linear regression model. Additionally, ridge regression can handle datasets with more predictors than observations.

We applied Ridge regression and Lasso regression on our data, and found the optimal value of the regularization parameter (which minimizes the MSE on the test set) by iterating over it.

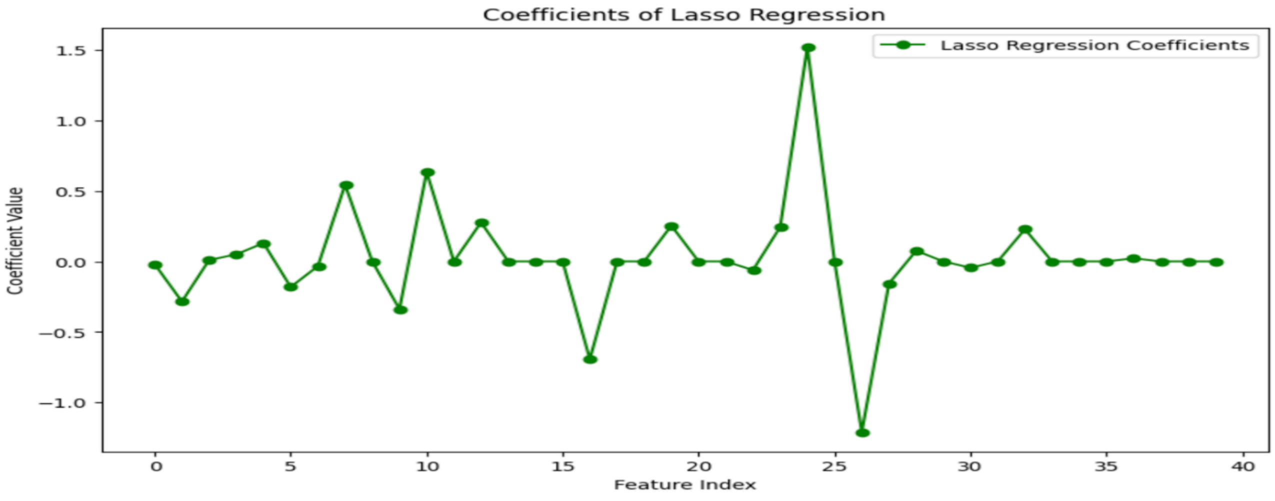


Test set loss over the regularization parameter range for Ridge regression



Test set loss over the regularization parameter range for Lasso regression

It is evident that ridge regression does a better job of minimizing the MSE over the test set than Lasso regression. The ideal regularization parameter was found to be 17.7.



Using lasso regression, we were able to determine the parameters which were irrelevant. (As can be seen from the above plot). 18 out of the 40 features were found to be unimportant for the regression. If these features are removed from the model, then it would generalize better on unseen data.

R-squared value for ridge regression: 0.8007

R-squared value for lasso regression 0.7994

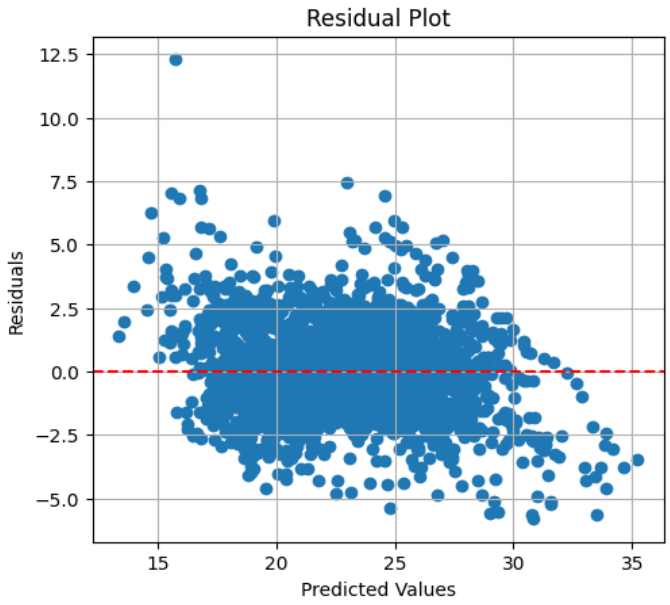
10 Residual Analysis

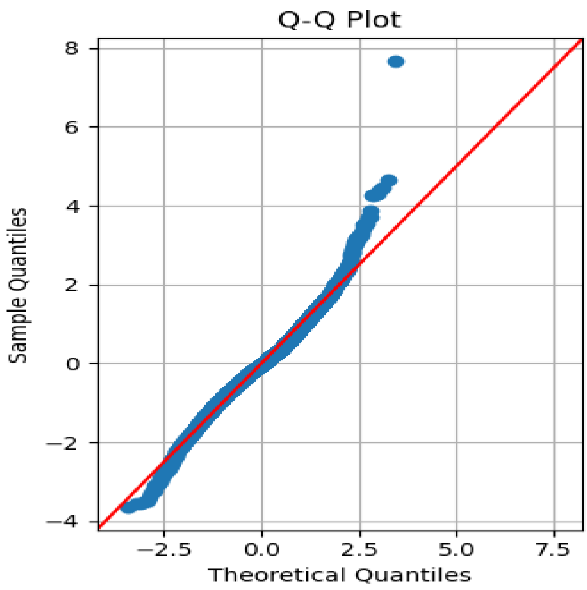
Residual analysis is a statistical technique used to evaluate the quality of a regression model by examining the residual errors between the observed values and the predicted values. Residuals are the differences between the actual values and the predicted values of the dependent variable in a regression model.

A residual plot is a graphical representation of the residuals against the predicted values. The plot can reveal any patterns or trends in the residuals, such as nonlinearity, heteroscedasticity, or outliers, that indicate problems with the model.

From the results of different models, we can observe that the MLR after backward elimination model is the simplest model as it is using the least number of parameters (17). Also, the results are almost comparable to other models, so we will do the error analysis for this model to verify the nature of errors. We could have considered the regularized model as well but there are not a lot of standard functions defined for the analysis of these methods. So, for the sake of simplicity, we are moving ahead with the MLR model after backward elimination method.

As observed from model summary, the R-squared value of this model is decent (0.73). This indicates that the fitted model is good. However, we have to check for multi-collinearity, remove redundant features so that overfitting can be avoided. This will be further confirmed with residual analysis. The figures below show that errors are random and are indeed normally distributed with mean 0 and constant variance. Therefore, the assumptions of linear regression are valid.





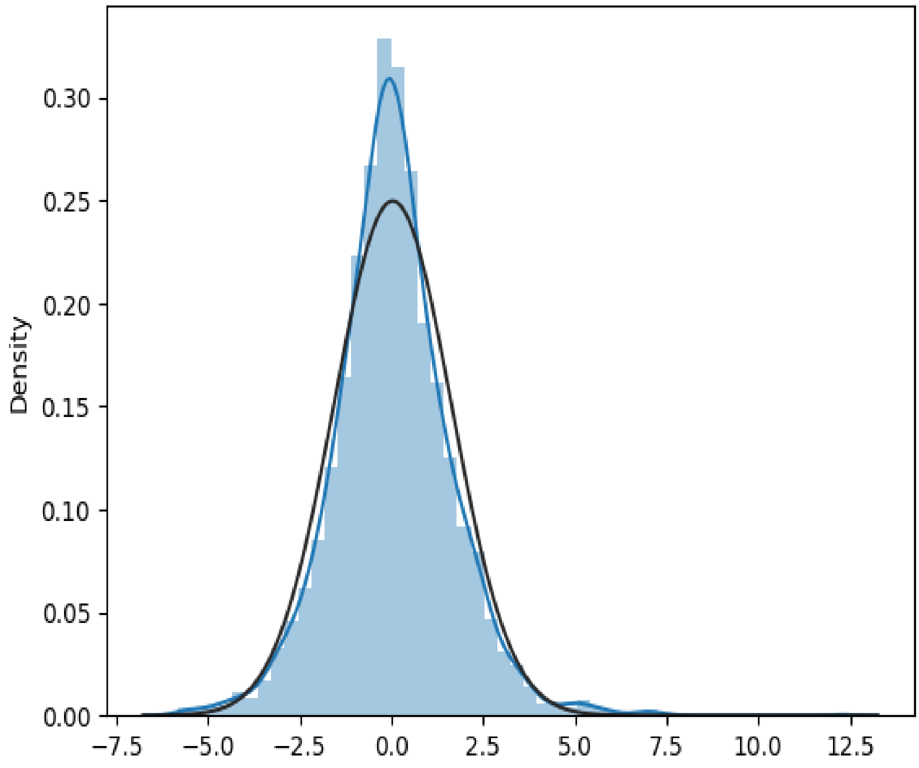
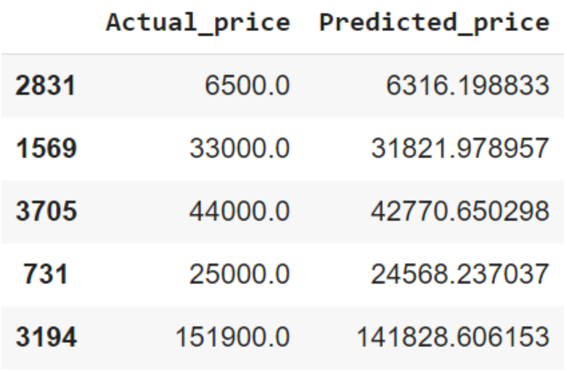


Figure 3: Residual Plots

Further, we use Jack-Knife method to detect outliers and influence points. Cook’s distance, standard residual and studentized residual were used as the metrics to classify a point as outlier on the basis of standard threshold. 72 outlier points were detected and removed.

11 Final Predictions

The final predictions over a sample from the test set after applying inverse box-cox transformations are reported below.

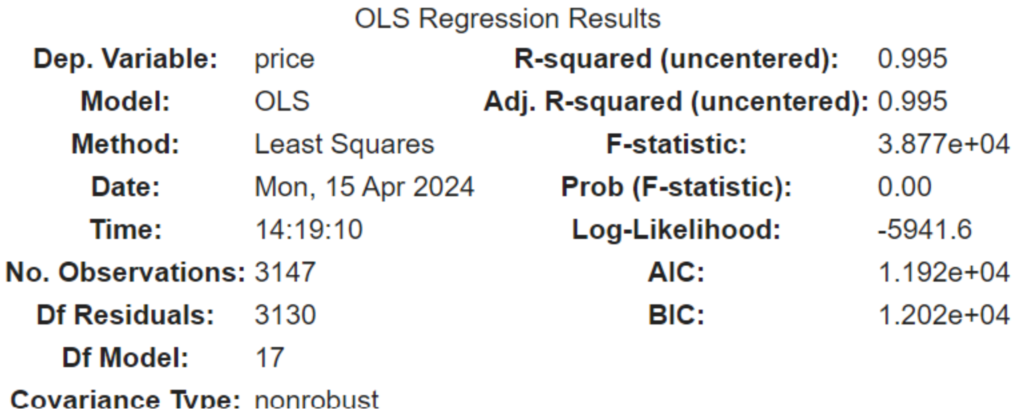


	Actual_price	Predicted_price
2831	6500.0	6316.198833
1569	33000.0	31821.978957
3705	44000.0	42770.650298
731	25000.0	24568.237037
3194	151900.0	141828.606153

12 Conclusion

The regression line using statistical methods with Ordinary Least Squares, Backward OLS, Ridge regression and Principal Component Regression were fitted and results were analysed using R-squared, RMSE, condition number, residual analysis and number of features used. Feature selection was done on the basis of these metrics.

All the metrics were not satisfied simultaneously for any model specially the metrics related to multi-collinearity like condition number. Therefore, out of all the models we implemented, we are selecting the model with least number of features (simple model) without compromising on other metrics. This was the model MLR with backward elimination. This is based on the assumption that it will generalise well on unseen data. The metrics of this model on test data are showed in green in the table above. The model summary is shown below.



OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.995
Model:	OLS	**Adj. R-squared (uncentered):**	0.995
Method:	Least Squares	**F-statistic:**	3.877e+04
Date:	Mon, 15 Apr 2024	**Prob (F-statistic):**	0.00
Time:	14:19:10	**Log-Likelihood:**	-5941.6
No. Observations:	3147	**AIC:**	1.192e+04
Df Residuals:	3130	**BIC:**	1.202e+04
Df Model:	17		
Covariance Type:	nonrobust		