DS 5220

Final Project Report

Car Price Prediction Using Machine Learning Techniques

Pengli Shao

Deadline: 07/31/2024, 11:59 PM

1 Introduction

Background Information and Objectives

In the used car market, many websites or companies provide users with estimated prices for cars that are for sale and preparing to sell. These price estimators greatly help buyers and sellers in the used car market to estimate the value of a vehicle.
The value of each used car is hard to determine based on the unique condition of each car. The goal of this research is to find the best fit model for used car prediction.
There are five models used for this project, MLR(Multiple Linear Regression), Lasso regression, Ridge regression, decision tree regressor, and random forest regressor.

2 Literature Review

There were many research conducted that are related to used car price prediction. A research conducted by Venkatasubbu and Ganesh [1] predicted the used car price by using Lasso Regression, Multiple Regression, and Regression Tree Model. Their data is from 2005 Central Edition of the Kelly Blue Book and contains 804 rows about GM cars. Their research provide a comprehensive methodology for price prediction and measurement of modle performance, the error of all three models they used is below 5%. However, the small data sample size may cause this research not repeatable for the entire used car market. Moreover, the techniques that Venkatasubbu and Ganesh didn't use may lead to better results.
Another research conducted by Hankar and Birjali [2] used 5 models including Multiple Linear Regression, KNN regressor, random forest regressor, gradient boosting regressor, and artificial neural network based regressor to predict the used car price in Morocco. Their data contains 8,000 rows of features mileage, fuel type, production year, mark, model, and fiscal power. The model with best performance in this research is gradient boosting regressor with R^2 of 0.80 and RMSE of 44,516.20. They were not satisfied with the result and willing to further improve the model in the future. In my opinion, their data contains insufficient independent variables to capture the pattern of the used car price.
Both researches suggests that using multiple model and comparison analysis is usually in predicting used car price, the task is complicated because the price can be greatly

affected by multiple potential reasons. And, both researches used MSE and R^2 for the measurements of models' performance.

3 Methodology

3.1 Data Preview

This data is from Kaggle:
https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data
It contains 16 columns:

- year
- make
- model
- trim
- body
- transmission
- vin
- state
- condition
- odometer
- color
- interior
- seller
- mmr
- sellingprice
- saledate

There are 558,838 rows of data before cleaning.

3.2 Data Cleaning

- Dropped `seller, saledate, state, vin` columns
- Dropped all rows that contains NA values
- Dropped rows with selling price equal to 1
- Dropped rows with abnormal odometer
  Odometer greater than 800,000 and lower than 100 are considered abnormal. Most of the odometer value higher than 800,000 exactly 999,999. And, car production factories usually do quality tests on their new cars, dealerships also drive new cars for transportation purposes. Therefore, cars that odometer lower than 100 do not considered as used cars.

- Dropped all rows that exterior color and interior color is number or symbol
- Checked for duplicates

After cleaning, the data contains 437,640 rows.

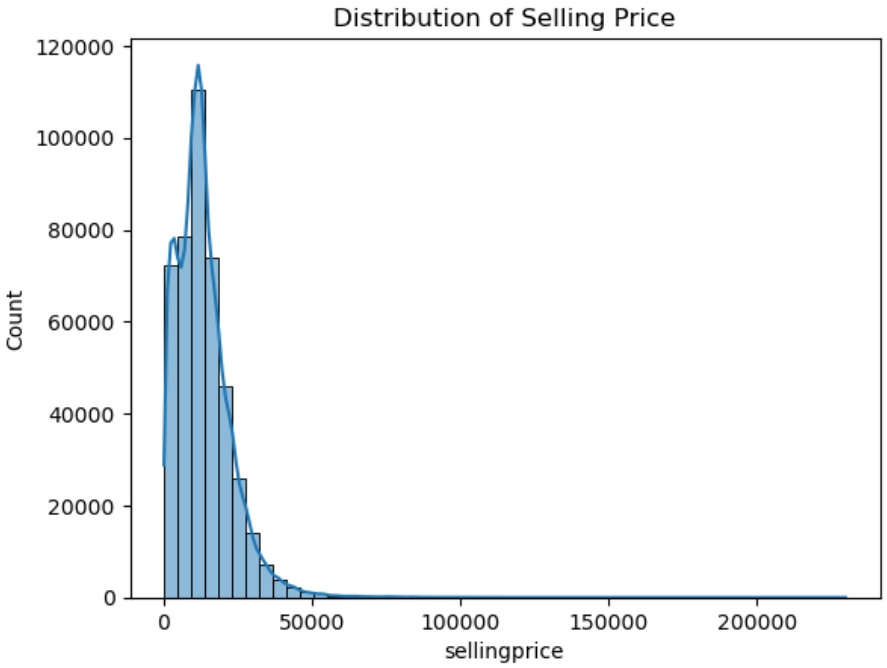

Figure 1: Selling Price Distribution

From figure 1, it is clear that the selling price is extremely right-skewed, representing the selling price of majority of the cars in the dataset are between the range 1 to 40,000. The outliers representing the unusual expensive luxury or high performance vehicles.

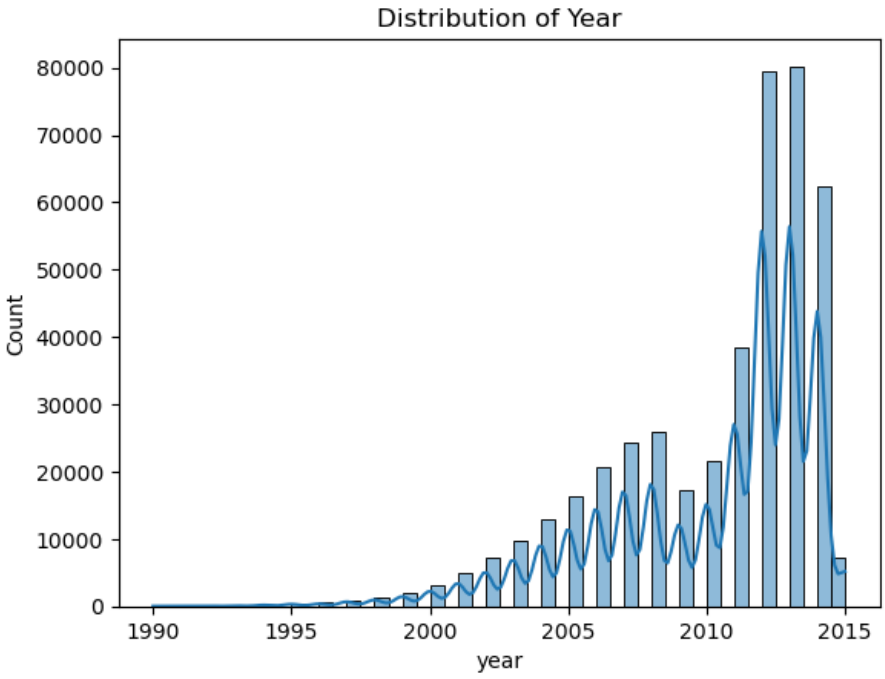

Figure 2: Distribution of Production Year

In figure 2, the distribution of production year suggests that the majority of the cars were produced after 2010.

3.3 Model Performance Measurements

- Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where:

  - n is the number of samples.
  - y_i is the actual value for the i-th sample.
  - $\hat{y}_i$ is the predicted value for the i-th sample.

MAE is the average of the absolute differences between the predicted value and the actual value

- Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

  - n is the number of samples.
  - y_i is the actual value for the i-th sample.

- $\hat{y}_i$ is the predicted value for the i-th sample.

MSE is the average of the squared differences between predicted value and actual value

- R-squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

- y_i is the actual value for the i-th sample.
- $\hat{y}_i$ is the predicted value for the i-th sample.
- $\bar{y}$ is the mean of the actual values.

R^2 indicates the proportion of the target variable that can be explained by independent variables

- Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of samples.
- y_i is the actual value for the i-th sample.
- $\hat{y}_i$ is the predicted value for the i-th sample.

RMSE is the root of MSE

3.4 Models

3.4.1 Multiple Linear Regression with Numerical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Where:

- y is the target variable selling price
- β_0 is the intercept
- $\beta_1, \beta_2, \beta_3$ are the regression coefficients for x_1, x_2, x_3
- x_1, x_2, x_3 are numerical independent variables representing year, odometer, and condition
- ϵ is the error term.

3.4.2 Multiple Regression with Lasso and Ridge, K-fold cross validation

I implemented Lasso and Ridge regression for numerical variables only to avoid overfitting by lower the coefficient of each independent variable with objective function of lasso and ridge. Then, I implemented K-fold Cross Validation to fully deny the possibility of overfitting. The K-fold divid the datasets into 5 groups randomly, and choose one arbitrary subset for test set. The cross validation examine the performance of the result from each training subset and compare them. If the performance is stable, then the possibility of overfitting is low.

3.4.3 Multiple Linear Regression with One-Hot Encoding

One-Hot Encoding is a technique to transform categorical variables into numerical variables. This method allows me to add more features for my model, with more features, the probability that the model can capture the pattern of the selling price may increase.

3.4.4 Decision Tree Regressor

Decision tree regressor is a tree type structure supervised machine learning algorithm that input the date from root node, split the data at each internal nodes, and output the result from leaf node. Decision Tree regressor can capture non-linear relationship, but overfitting is a inevitable problem.

3.4.5Random Forest Regressor

random forest regressor generates multiple decision trees and take the average of the results of all decision trees to reduce the possibility of overfitting. I set the `n_estimators` equal to 40 to generate 40 decision trees in my random forest model.

4 Results

Multiple Linear Regression

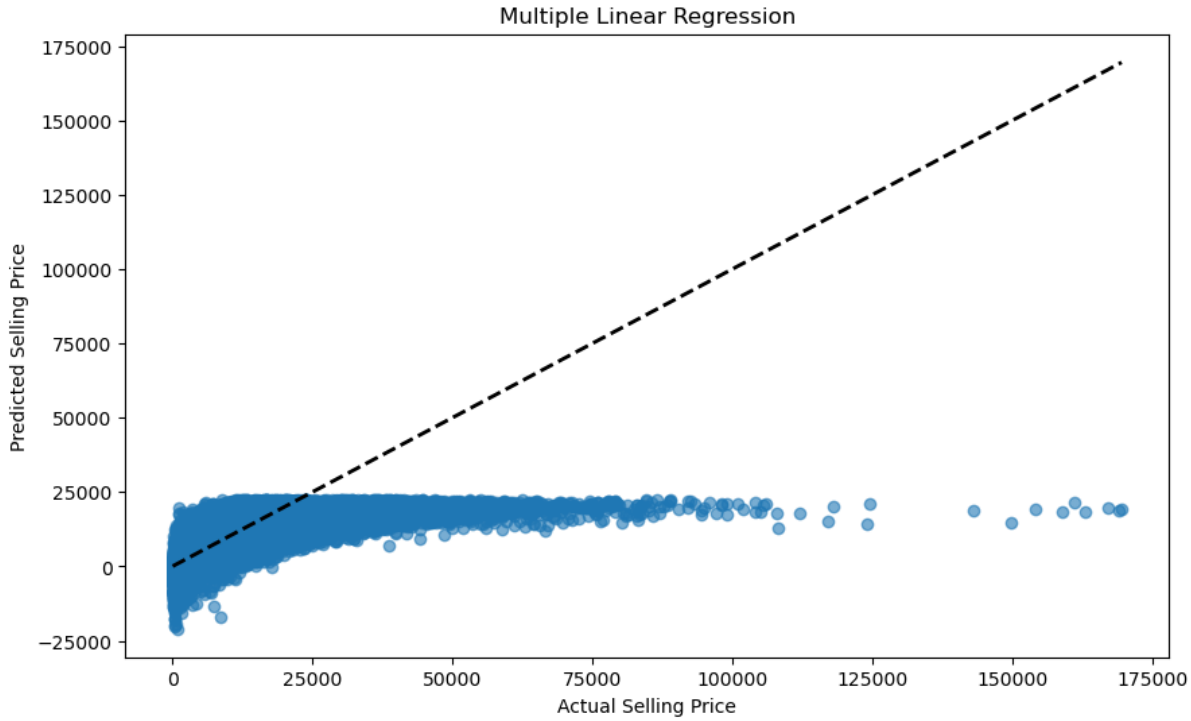

Figure 3: MLR

From figure 3, the highest selling price that can be predicted by multiple linear regression is about 25,000. Further more, the majority of the data points are below the ideal fit line, indicating the predicted value is much lower than the actual value.

Metric	**Value**
MSE	54608852.22
MAE	5086.46
R^2	0.40
RMSE	7389.78

Table 1: Error Metrics

From table 1, we can conclude that the performance of the MLR model is not satisfactory, the error is unacceptable and only 40% of the dependent variable can be explained by independent variables.

4.1 Lasso and Ridge Regression

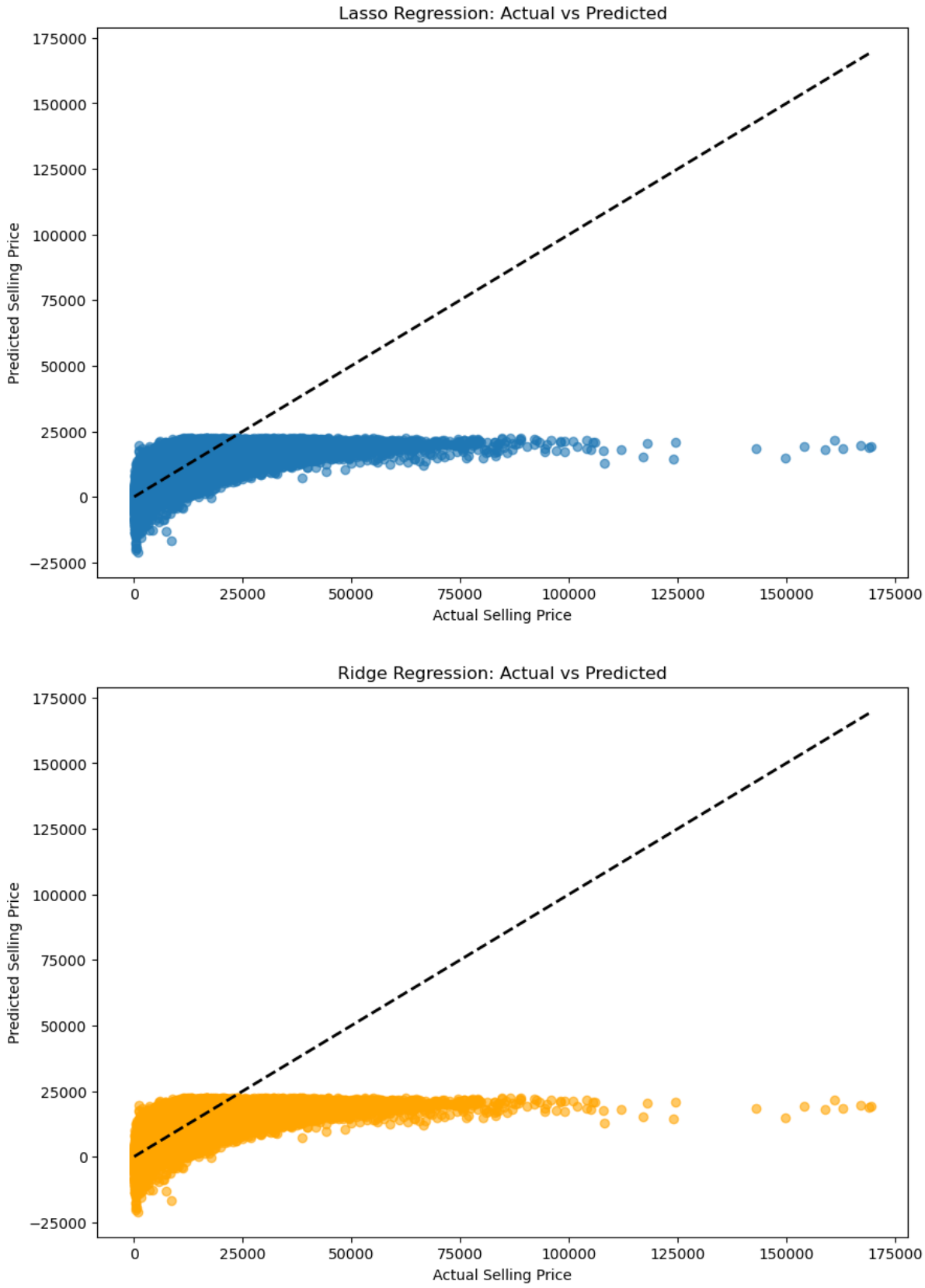

Figure 4: Lasso and Ridge

Metric	Lasso	Ridge
MSE	54608853.38	54608852.24
MAE	5086.46	5086.46
R^2	0.40	0.40
RMSE	7389.78	7389.78

Table 2: Error Metrics for Lasso and Ridge Regression

Metric	Value
Cross-validated R^2 scores	[0.40, 0.40, 0.40, 0.39, 0.39]
Average R^2	0.40
Average MSE	54403560.07
Average MAE	5084.93
Average RMSE	7375.57

Table 3: Cross-Validated Metrics for Linear Regression

The performance of Lasso and Ridge regression is also poor, by K-fold cross validation, from table 3, the model is stable in terms of the performance of each subset. Therefore, multiple linear regression may not be able to capture the pattern with only numerical variables.

4.2 Multiple Linear Regression with One-Hot Encoding

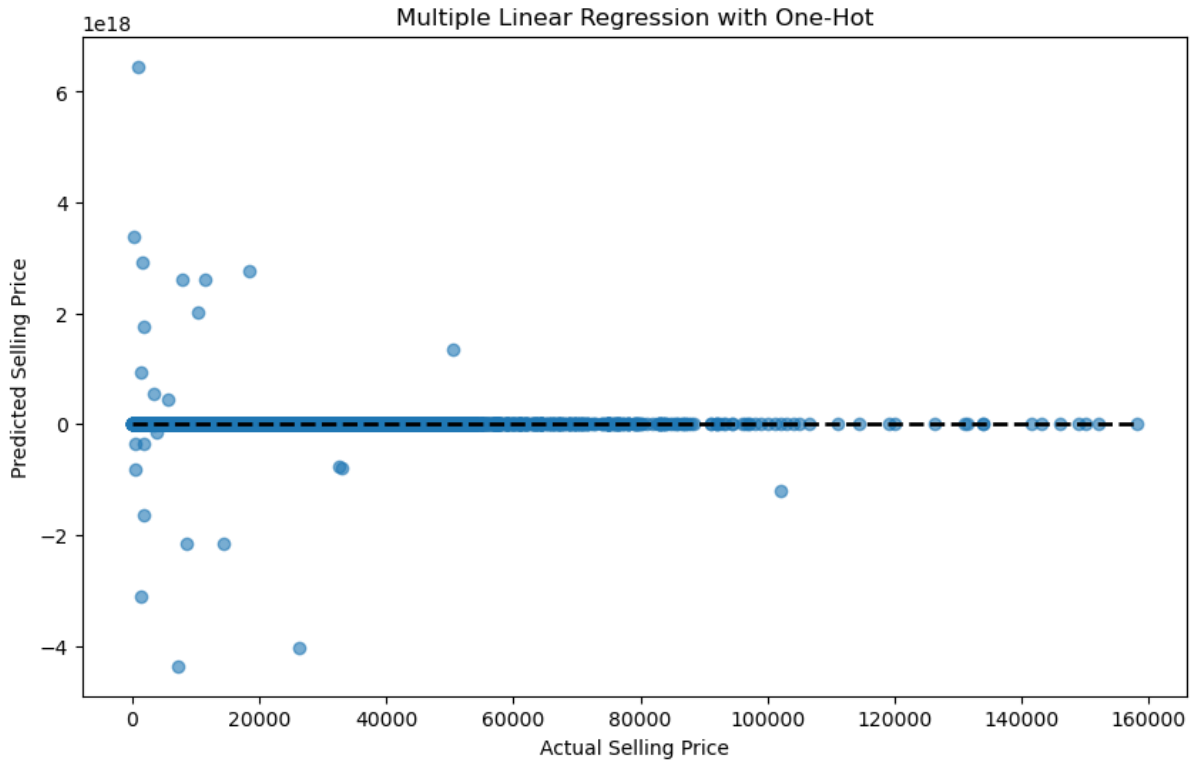

Figure 5: MLR with One-Hot

Metric	Value
MSE	1.76e+33
MAE	5.68e+14
R^2	-1.95e+25
RMSE	4.19e+16

Table 4: Error Metrics

By examine the results of multiple linear regression with all the features, it is clear that the performance of the model became worse. There might be non-linear relationships between independent variables and dependent variable, or the complexity of the model is too high for MLR.

4.3 Decision Tree

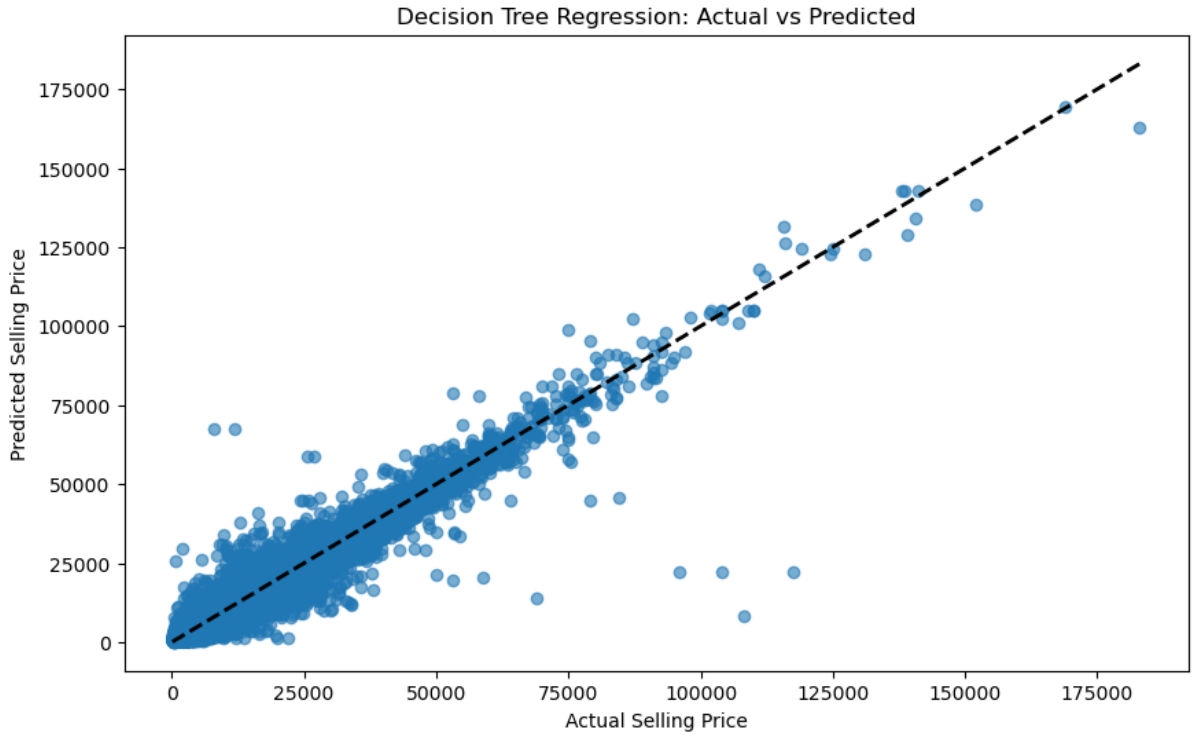

Figure 6: Decision Tree Regressor

Metric	Value
MSE	4254711.18
MAE	1286.08
R^2	0.95
RMSE	2062.70

Table 5: Error Metrics

From figure 6, the data points concentrated near the ideal fit line, few outliers exists but it is within the acceptable range. Table 5 indicates that there is a significant increase in

R^2 compare to MLR and significant decreases in errors.
However, decision tree model can not prevent overfitting and the performance can be further improved by random forest.

4.4 Random Forest

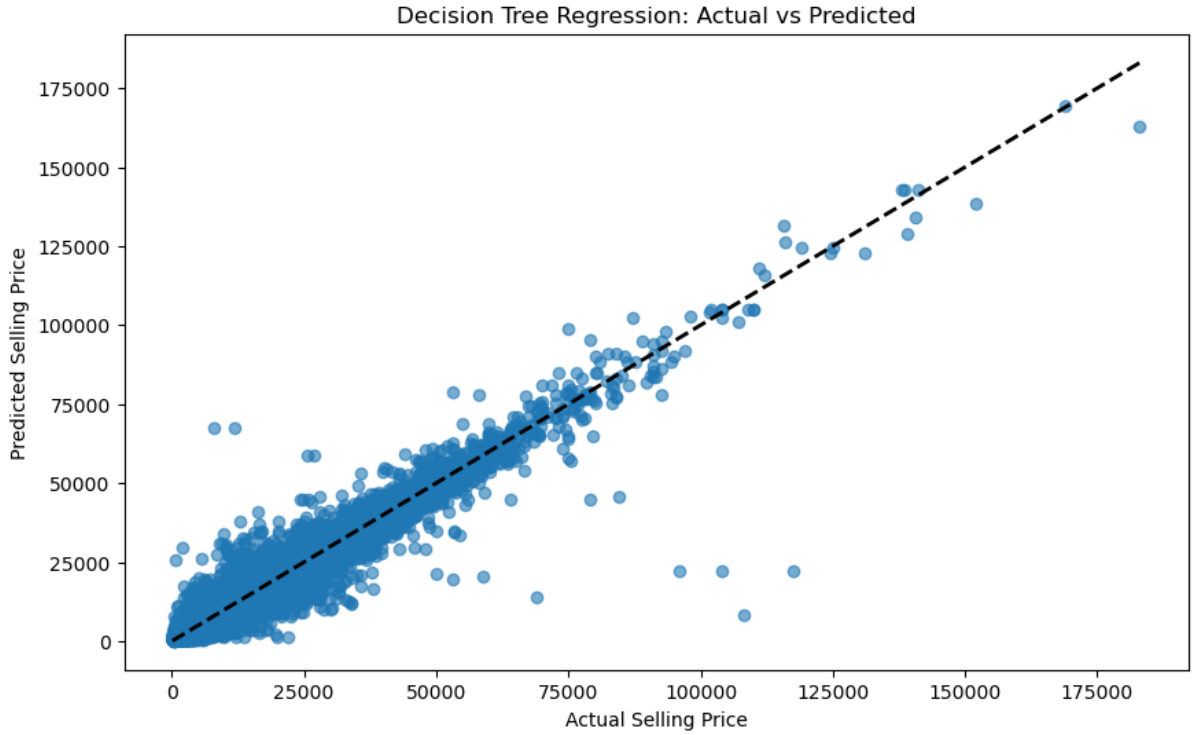

Figure 7: Random Forest

Metric	Value
MSE	2198737.87
MAE	949.57
R^2	0.98
RMSE	1482.81

Table 6: Error Metrics

The results shows a further improvement in error and R^2, a R^2 of 0.98 is an ideal value and the MAE and MSE is completely acceptable.

5 Discussion

One of the possible reason that the performance of MLR is poor after implementing One-Hot encoding is the curse of dimension, one-hot encoding increase the dimension by transform one column into n columns where n represents the unique values in one specific categorical column. Each of the column `make, model, color, interior` contains more than 10 unique values, especially `make` column. This results in increasing in the data dimension and model's complexity.

The well performance of random forest and decision tree regressor might because MLR cannot capture the non-linear relationship between independent variables and dependent variables. For those cars that are produced in recent years, the price decreases greatly as year decreases, but after a certain point, the price of old cars produced decade ago may not be affected greatly by production year.

The result of MLR in this research compare to the researches in literature review is similar, the performance of MLR is poor in both researches in literature review. However, in Hankar's research, the R^2 value of random forest regressor is only 0.74 with RMSE 44516. In this research, the performance of random forest regressor is much better, the R^2 value of 0.98 and RMSE equal to 1482.81.

6 Conclusion

The research compare the performance of Multiple Linear Regression with different techniques, Decision Tree Regressor and Random Forest Regressor. Random Forest Regressor has the best performance with R^2 value of 0.98 and RMSE equal to 1482.81.
One of the limitation of this research is that the random forest model needs maintenance for current used car market price prediction. As I mentioned in discussion section, there is a non-linear relationship between independent variables such as year and dependent variable, the data used in this research only contains the cars produced in 1990 to 2015, but for a 2024 price prediction task the data might be out-of-date.
Using the random forest regressor also will facing the problem of high computational resources requirements, the random forest regressor is considered computational expensive and it runs very slow. In addition, the interpretability of the results output from random forest is low. It is hard to explain the contribution of each independent variable to dependent variable in random forest regerssor.
And, the values in condition column range from 1 to 50, it is unclear that what each condition represents. If the contribution of condition is great, the change in scale may cause a decrease in R^2 and increase in error.

7 References

References

[1] Venkatasubbu, P., & Ganesh, M. (2019). Used cars price prediction using supervised learning techniques. Int. J. Eng. Adv. Technol.(IJEAT), 9(1S3).

[2] Hankar, M., Birjali, M., & Beni-Hssane, A. (2022, May). Used car price prediction using machine learning: A case study. In 2022 11th International symposium on signal, image, video and communications (ISIVC) (pp. 1-4). IEEE.

8 Appendix: Additional Figures

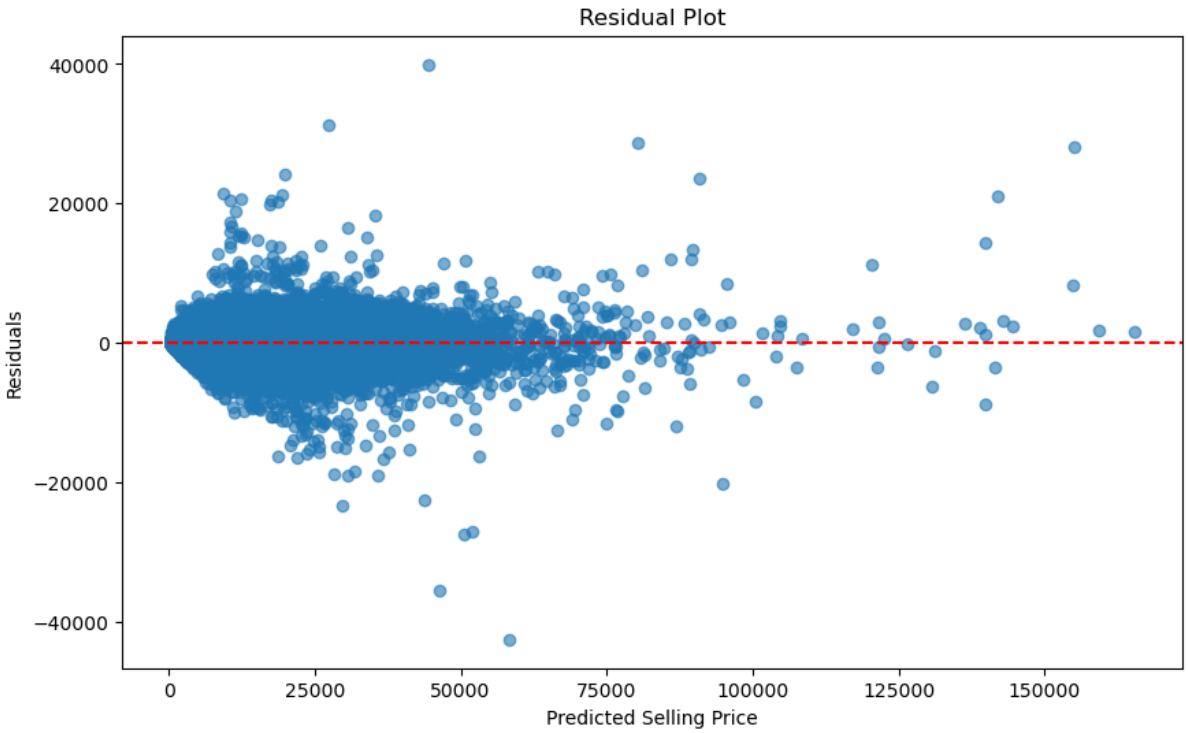

Figure 8: Enter Caption