**PREDICTIVE ANALYTICS: MODELLING FOR CAR SELLING PRICES**

Authors: Group 2 - Phung Quang Anh, Le Bao Trung, Nguyen Anh Dung, Pham Hoang Long, Ngo Gia Trang

1. **Introduction**

In this supervised learning project, our goal is to predict the selling prices of cars using a diverse dataset encompassing key features such as car name, year of construction, kilometers driven, fuel type, seller type, transmission, owner details, and technical specifications. The project involves a series of tasks, including data cleaning, feature engineering, and the application of regression models such as linear regression, elastic net, regression tree, and random forest. We'll explore essential questions related to model validity, significance of coefficients, overfitting analysis, and model optimization. The final objective is to compare and determine the most effective predictive model for estimating car selling prices, with insights presented in a concise five-page paper along with an accompanying R script. After preporcessing the data, the observation obtained in the training set is 3026 and the test set is 1014. However, one observation of seat (14 seats) in the test set is not included in the training set, so we remove the observation to run models.

1. **Part 2**

**a,** After building our primary model using on the training dataset, we calculate the MSE for both set and obtained the following metrics for linear regression model:

- Training MSE: 58,877,005,073

- Test MSE: 121,030,201,753

Comparing these values, the test MSE is significantly higher than the training MSE. This indicates that the model is overfitting. The model is too closely fitting the training data, including noise and irrelevant patterns, and is not generalizing well to new data.

A graph showing a line of dots

Description automatically generated with medium confidence

**b,** F-test of the model:

H0: Variance of the residuals is equal to the variance of the population.

H1: Variance of the residuals is not equal to the variance of the population.

Conducting summary of the linear regression model, we obtain F-statistic: 186.6 at P-value: < 2.2e-16.

At the P-value threshold of 5%, we reject the null hypothesis, F-statistic is also very high, which suggests that the model is valid and overall significant.

**c,** T-Test of the model

H0: The coefficient is not significantly different from zero (β = 0).

H1: The coefficient is significantly different from zero (β ≠ 0).

From the summary of the model, we can conclude that the quantitative features such as “km\_driven”, “engine”, “max\_power”, “mileage”, “Nm” are significant at 5% level, we reject the null hypothesis and conclude that they have statistically significant on selling prices.

For the qualitative features, we can conclude that the luxury brand name like: Audi, BMW, Jaguar, Jeep, Land, Lexus, Mercedes-Benz, MG, Volvo, the car with seats counts of 4 and the car manufactured in 2018 and 2019 are significant at 5% level and different impact on selling price rather than others in respectively categories.

**d,** Now we build final model with feature “rpm” excluded, repeat step a,b,c and get the following metrics for the linear regression model:

**- Training MSE: 58,935,497,604**

**- Test MSE: 122,005,238,495**

Thus, the model remains overfit.

- F-statistic: 188.9 at p-value: < 2.2e-16 => Final model is valid and overall significant for the data.

- T-test: There are changes in each feature’s t-value and p-value but at 5% threshold, the list of significant features remains the same.

A screen shot of a graph

Description automatically generated

**e,** List of significant features

|  |  |
| --- | --- |
| **1. name**   * Audi: 859,800 * BMW: 1,419,000 * Jaguar: 1,034,000 * Jeep: 475,500 * Land: 1,407,000 * Lexus: 3,594,000 * Mercedes-Benz: 1,075,000 * MG: 651,200 * Volvo: 564,600   **2. year**   * 2018: 610,400 * 2019: 635,400   **3. km\_driven**: -0.6412  **4. fuel**   * Diesel: 100,600 * LPG: 19,150 | **5. seller\_type**   * Individual: -67,770   **6. transmission**   * Manual: -59,930   **7. owner**   * Test Drive Car: 2,526,000 * Second Owner: -30,330 * Fourth & Above Owner: -78,110   **8. mileage**: -5,556  **9. engine**: 85.07  **10. max\_power**: 3,635  **11. Torque (nm)**: 967.6  **12. # Seats**   * 4 seats: 45,7200 |

**f,**

Analyzing the coefficient of the intercept, of one quantitative feature and one qualitative feature on the target feature in the final model

- Intercept, representing the baseline selling price, had an estimated value of -361,400, denotes the expected selling price when all other features are zero.

- Quantitative features represent the change in the selling price for a one-unit increase in the corresponding quantitative feature, holding other variables constant. For example, if “max\_power” increases by 1 unit, the selling price will increase by 3635 units.

- Qualitative features reflected the impact on the selling price compared to the reference category. “Owner” has 5 types, with First Owner being the baseline, the price will be increased by 2,560,000 if it’s a Test Drive car with the same other features, decrease by 30,300 if owenership is second and by 78110 if ownership is fourth and above.

1. **Part 3**

**a,**

After using a 5-fold cross-validation technique, the best value of the l1-l2 allocation 𝜃=0.1 and quantity of regularization 𝜆=0.0139.

**b,** We have the following metrics for the elastic net model:

**Training MSE: 59,043,631,448**

**Test MSE: 112,174,286,392**

The Mean Squared Error (MSE) on the training set is notably lower compared to the MSE on the test set. This notable discrepancy indicates that the model exhibits superior performance on the training data in comparison to the testing data. This scenario is indicative of overfitting.

A graph with a red line and black dots

Description automatically generated

**c,**

We keep factors with absolute values greater than 0.1.

|  |  |
| --- | --- |
| **1. Name**   * Audi: 1.6307 * BMW: 2.6489 * Datsun: -0.254 * Fiat: -0.1354 * Force: -0.3079 * Isuzu: 0.6997 * Jaguar: 1.9224 * Jeep: 0.9408 * Kia: 0.4745 * Land: 2.6363 * Lexus: 6.5506 * Manhindra: -0.1645 * Mercedes-Benz: 2.0084 * MG: 1.1763 * Opel: -0.2022 * Renault: -0.1045 * Tata: -0.2696 * Toyota: 0.4979 * Volkswagen: -0.1963 * Volvo: 1.0673   **2. Year**   * 1995: -0.4985 * 1996: -0.518 * 1997: -0.4329 * 1998: -0.2402 * 1999: -0.3885 * 2000: -0.2213 * 2001: -0.5212 * 2002: -0.4795 * 2003: -0.5013 * 2004: -0.4644 * 2005: -0.305 * 2006: -0.4021 * 2007: -0.4914 | * 2008: -0.3451 * 2009: -0.3304 * 2010: -0.2802 * 2011: -0.2858 * 2012: -0.1931 * 2013: -0.122 * 2015: 0.1464 * 2016: 0.248 * 2017: 0.3701 * 2018: 0.5634 * 2019: 0.6079 * 2020: 0.4815   **3. Fuel**   * LPG: 0.2379   **4. Seller\_type**   * Individual: -0.1259 * Trustmark Dealer: -0.1028   **5. Transmission**   * Manual: -0.1316   **6. Owner**   * Test Drive Car: 4.586 * Fourth & Above Owner: -0.1525   **7. Mileage**: 0.2337  **8. Torque (nm)**: 0.1576  **12. # Seats**   * 4 seats: 0.3951 * 6 seats: -0.1076 |

1. **Part 4**

**a,** We have the following metrics for tree model:

**Training MSE: 52,860,447,352.**

**Test MSE: 178,821,142,250.**

The MSE on the training set is significantly lower than the MSE on the test set. This significant difference shows that the model performs better on training data than on testing data. Such a situation is a manifestation of overfitting.

A screen shot of a graph

Description automatically generated

**b,**

A diagram of a family tree

Description automatically generatedBased on the above graph, there are 13 nodes. Therefore, the depth of the tree is 13-1=12. And the variables used in the splits process are: "max\_power", "year", "Nm", "name", "km\_driven", and "mileage".

**c,**

A line graph on a white background

Description automatically generated

We will trim the tree from our original tree to obtain the ideal tree size of three since, as we can see, the tree sizes of six have the lowest cross-validation error.

**d,** We have the following metrics for pruned tree:

**Training MSE: 82,289,936,488.**

**Test MSE: 177,579,554,291.**

This indicates a result of overfitting.

A screen shot of a graph

Description automatically generated

A diagram of a diagram

Description automatically generated with medium confidenceBased on the above graph, there are 6 nodes. Therefore, the depth of the tree is 6-1=5. And the variables used in the splits process are: "max\_power", "year", and “km\_driven".

**e,**

Based on the Mean Squared Error (MSE) metric, we can compare the performance between an unpruned tree and a pruned decision tree., then:

If we wish to minimize MSE to achieve good performance, it appears that the unpruned decision tree shows a lower MSE on the training set and slightly higher MSE on the test set than the trimmed decision tree. cut.

Therefore, based on MSE, it can be concluded that the pruned decision tree seems to yield better results on the test set.

1. **Part 5**

**a,** We have the following metrics for the random forest:

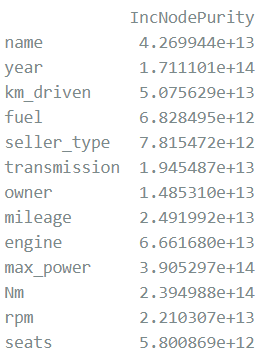
**Training MSE: 9,284,965,576**

**Test MSE: 149,687,000,000**

A graph with black dots and red line

Description automatically generatedThe substaindicateference between the MSE values on the training and test sets indicates that the model performs significantly better on the training data than on the test data. This discrepancy is a clear sign of overfitting, where the model excessively fits the training data and thus shows poorer performance when dealing with unseen data.

**b,**

The importance of features of the model:

The listed coefficients for various features indicate their respective impact or relevance in predicting the outcome. The variable ‘max\_power’ seems to have the highest significance, suggesting that it has a substantial influence on the predictions compared to other features. Next, the variables ‘Nm’ and ‘year’ are also two important variables in the model with high importance. Besides, the ‘seats’ variable seems to be the least important variable compared to the remaining variables with low importance.

**c,**

Using a 5-fold cross-validation technique and a range of feature numbers spanning from 1 to 13, RMSE was employed to identify the optimal model, aiming for the smallest value. Eventually, the chosen parameter value for the model was mtry = 13.

Using 13 features, the optimal number of trees was determined to be 50, resulting in a mean RMSE value of 160939.2. To assess for potential enhancements, a random forest model was constructed with mtry = 13 and ntree = 50 to explore any potential improvements

A graph with a red line and black dots

Description automatically generated

**d,** We have the following metrics for the random forest:

**Training MSE: 5,618,531,004**

**Test MSE: 143,724,584,834**

Despite the lower RMSE and improvement in both the training and test sets, the RMSE in the test set improved insignificantly and still remains significantly higher than that in the training set. This indicates that the model is still overfitting even after tuning the model parameters.

A screenshot of a computer

Description automatically generatedThe importance of features of the model:

In the new model, the importance of the 'max\_power' variable remains the highest. Next in importance are the 'years' and 'km\_driven' variables (the latter being less important in the old model). The importance of the 'Nm' variable has dropped to fourth in the list. Surprisingly, the 'fuel' variable has become the least important instead of the 'seats' variable in the old model.

**e,**

After adjusting the parameters of the Random Forest model, there was a noticeable but insignificant improvement. The MSE on the training set decreased from 9,284,965,576 to 5,618,531,004, which represents a reduction of approximately 39%. On the test set, the MSE decreased from 149,687,000,000 to 143,724,584,834, indicating a decrease of only about 3.98%.

Despite this improvement, the relatively small reduction in MSE on the test set compared to the training set suggests that the model is still overfitting. The substantial difference between the MSE of the training and test sets implies that the model struggles to generalize well when faced with new data.

1. **Part 6**

Among all four models, the Elastic net is the best model to predict the selling price of cars in this situation.

The reason is that Elastic net model has the lowest MSE. After applying the linear regression, elastic net, decision tree and random forest models, we conduct the following tables about the MSE and RMSE (which is the squared root of the MSE) in the test set:

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **RMSE** |
| Linear regression | 122,005,238,495 | 349,292 |
| Elastic net | 112,174,286,392 | 334,924 |
| Decision tree | 177,579,554,291 | 421,402 |
| Random forest | 143,724,584,834 | 379,110 |

As is presented in the table above, the MSE of Elastic net and linear regression are much lower than the MSE of Decision tree and Random forest, even though all the models are overfiting (as all MSE are too high). The reason could be that the dataset includes many qualitative variables, higher than the number of quantitative variables, and this affect the Decision tree and Random forest to take more chance to overfit the data rather than Linear regression and Elastic net. Between the Linear regression and Elastic net, the Elastic net provides further regularization (including Ridge and Lasso) for the model to reduce the coefficients of both insignificant dummy and quantitative variables to zero. Thus, it helps to reduce the number of categories for each qualitative variable and remove the quantitative variables that do not have effect on the target features. Also, the linear regression is too simple model and it does not fit the plot of target feature (selling price of cars). Thus, the Elastic net model helps to reduce the overfitting rather than using linear regression.

However, as we can see in the RMSE column, the RMSE of all four models are too large. It means that in the Elastic net model, the average predicted selling price differs from the actual selling price of 334,924. The figures for other models are larger. This can be interpreted that the predicting error costs a car in the selling price range of 350,000. This does not make sense and all four models applied are still not the best models. The reason can be that this dataset includes many qualitative variables, which are higher than the number of quantitative variables. Also, the quanlitative variables, “name”, “year” and “seats” have more than 10 categories, which is tough and difficult for all models to predict the price. For example, there are about 25 categories for “name”, 20 categories for “year”, 10 categories for “seats”. Then if we run the linear regression and Elastic net, there are too many dummuies included in the models, then it is meaningless to interpret and explain. Because when each car brand is represented by a separate dummy variable, multicollinearity can occur when there is a strong correlation between the dummy variables, increasing the standard deviation and uncertainty when estimating parameters. The increased number of dummy variables also increases the complexity of the model and increases the number of parameters that need to be estimated, reducing the performance of the model. Furthermore, due to the various categories, the models experience dramatically heavy overfitting.

Thus, we recommend that the dataset should be transformed, and the variables should be selected at the beginning of running models to have better predictions.

First of all, depending on the nature of data, we should group up the categories in the qualitative variables to have better control of overfitting. As we can see, the car in the luxury brand will have higher sellng prices than others, or the car which is manufactured recently in 2018 to 2020 will have higher selling prices than older versions. Thus, based on the selling price, we should group up the categories in these variables.

Next, outliers of the dataset should be considered to be removed or replaced. Many existing outliers can be a reason for havy overfitting as it manipulates the tendency of the target feature. Then, the single factor analysis should be conducted to assess the effect and the sign of the independent variables on the dependent variable. Then if the sign is not as expected and the effect is insignificant, we should remove the variables. Then, the correlation matrix between variables should be conducted to make sure that multicollinearity does not exist in the models.

Finally, after modifying the dataset to have better interpretation and explanation that matches the reality, models and machine learning can be applied to the dataset to predict. As can be seen from previous plot of actual selling price and predicted selling price, it may follow Michaelis-Menten model which has the following graph:

A graph of a curve

Description automatically generated

Thus, we recommend this type of model to have better predictions and low MSE. Other machine learning can be used to predict based on modified dataset.

1. **Conclusion**

In conclusion, according to this project, we can conclude that the Elastic net model is the best one among those four models to predict selling price. However, the dataset should be modify to match with the nature of the data and the reality, and other models should be applied to have better predictions and low MSE.