Data Science Term Project

Final Report



**Team 13**

소프트웨어학과 201735856 윤혜주

소프트웨어학과 201735881 조소영

소프트웨어학과 201835505 이하영

목차

1. **Introduction3**
2. **Data description3**
   1. **Data Source3**
   2. **“Vehicles.csv” Original Data Inspection3**
   3. **“Vehicles (1).csv” Data Inspection4**
3. **Data preprocessing6**
   1. **Handle “Null” Condition6**
   2. **Handle dirty data7**
   3. **Fill missing value9**
   4. **Normalization10**
4. **Algorithms11**
   1. **Multiple linear regression11**
   2. **Knn-classification13**
5. **Evaluation15**
   1. **K-Fold cross validation15**
   2. **Ensemble learning – Bagging15**
6. **GUI15**
7. **Role16**
8. **Conclusion16**
9. **Reference 17**
10. **Introduction**

When people buy used cars, it is very important to buy cars in good condition at reasonable prices. However, most of the consumer damage relief measures related to used cars, which were released by the Korea Consumer Affairs Agency from 2016 to June 2019, were “in many cases, the state of the vehicle is different from what was previously announced”.

So, when people buy a used car, they can use we created a program that predicts condition of the used car by entering information such as the mileage, year, manufacturer, etc.자동차, 트럭, 테이블, 앉아있는이(가) 표시된 사진

자동 생성된 설명<Picture1 – KBS new screen capture>[[1]](#endnote-1)

1. **Data description**
   1. **Data Source**

We used 'Used cars dataset' registered in Kaggle.

This data is scraped every few months, it contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories.

실내, 앉아있는, 테이블, 거울이(가) 표시된 사진

자동 생성된 설명우산이(가) 표시된 사진

자동 생성된 설명

* 1. **“Vehicles.csv” Original Data Inspection**

**size**: 25 columns X 539759 row

**columns** = [“id”, “url”, “region”, “region\_url”, “price”, “year”, “manufacturer”, “model”, “condition”, “cylinders”, “fuel”, “odometer”, “title\_status”, “transmission”, “vin”, “drive”, “size”, “type”, “paint\_color”, “image\_url”, “description”, “country”, “state”, “lat”, “long”]

* 1. **“Vehicles (1).csv” Data Inspection**

Select 6 columns to use for data analysis (including target value).

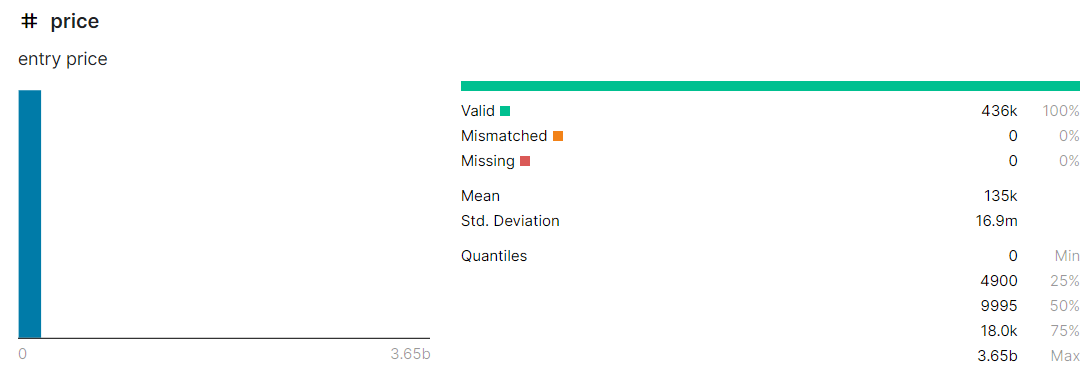
<<<<vehicle(1) 만드는 코드 넣기>>>>

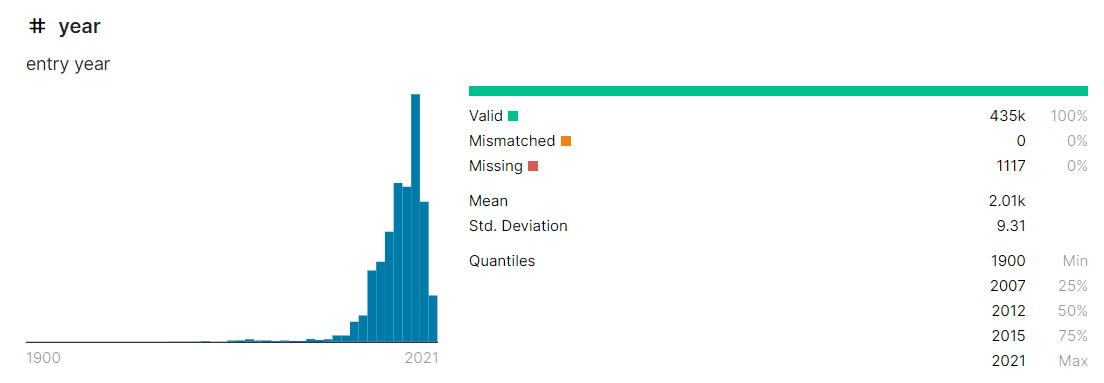
**Used columns** = [“price”, “year”, “manufacturer”, “condition”, “cylinders”]

테이블이(가) 표시된 사진

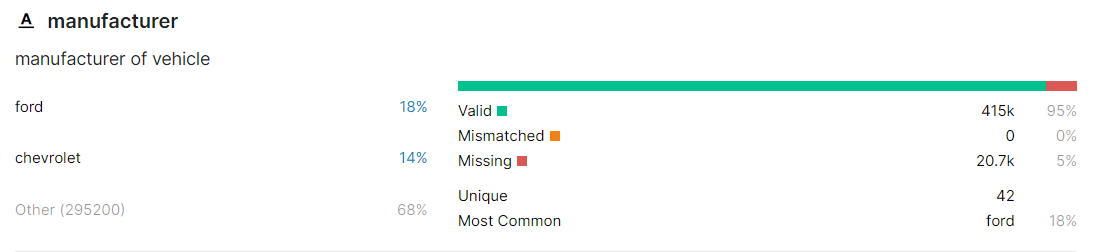
자동 생성된 설명 <- Result of column extraction

**“Price” columns**

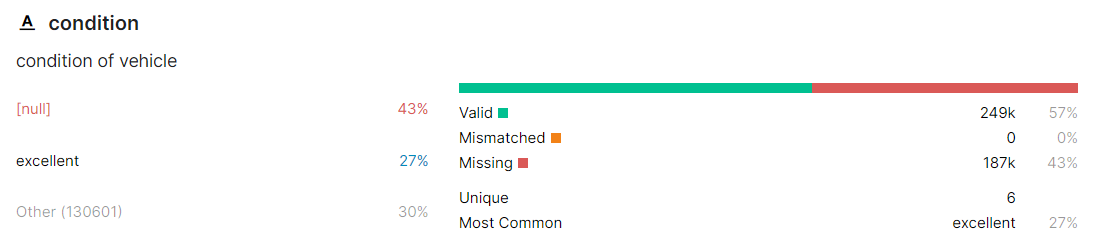


**“year” columns**

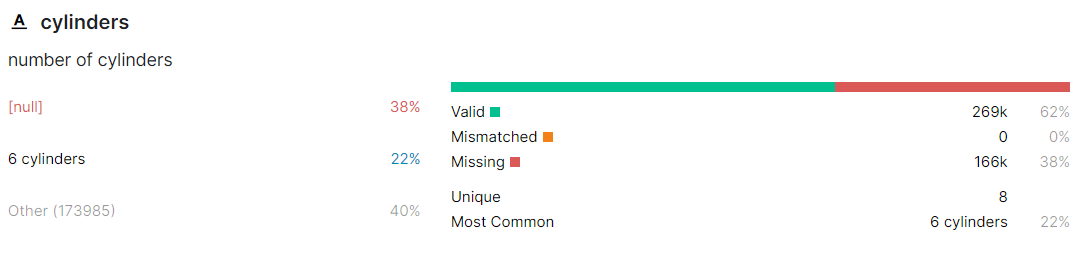
**“manufacturer” columns**



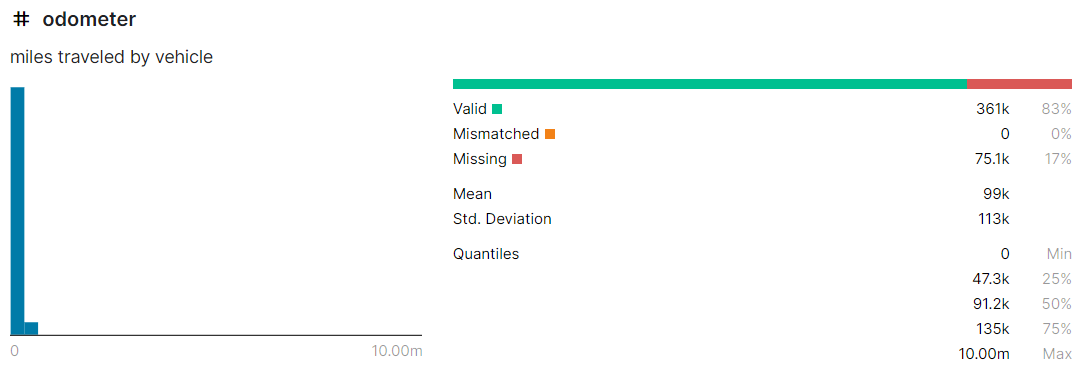
**“condition” columns (target)**



**“cylinders” columns**



**“odometer” columns**

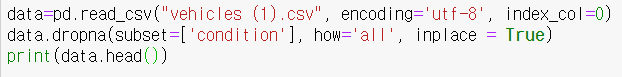


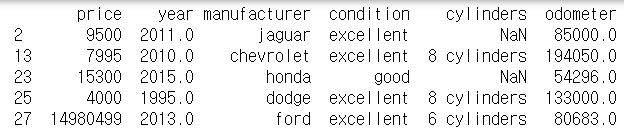
1. **Data preprocessing**
   1. **Handle “Null” Condition**

There were 44 percent missing value in the target feature ‘Condition’. If we start the analysis in this state, or use half the remaining data to fill 44% of the missing values, we thought there would be confusion in the subsequent data analysis.

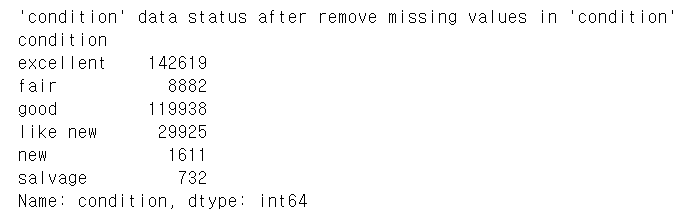
Therefore, we proceeded with the deletion of rows containing 'missing value' in 'condition' with the aim of having accurate data and giving the user the appropriate predicted results.

<Open vehicles (1).csv & Drop rows with missing values in ‘Condition’>





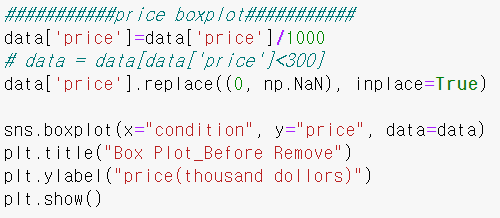
<'condition' data status after the above process>



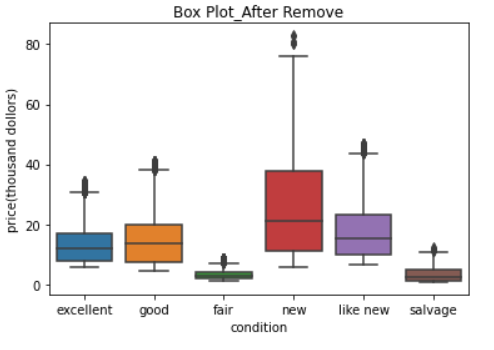
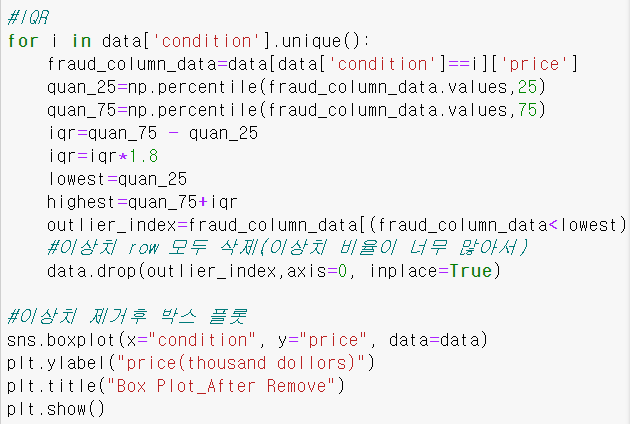
* 1. **Handle dirty data**

'Box plot' was used to view large amounts of data at a glance. We used this to check the outlier, and there were quite a few outliers. Therefore, it was determined that converting 'dirty data' to 'null' and using existing data to populating it than deleting it later would be more helpful for future predictions.

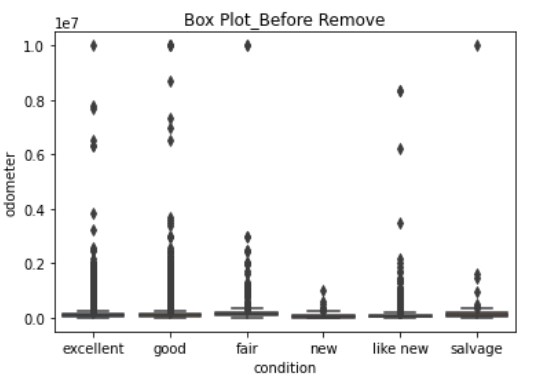
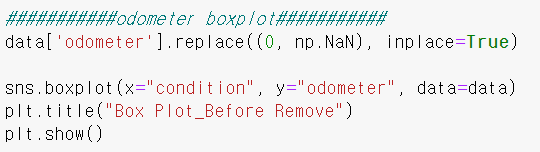
The following is the code and result is expressing the data of the 'price' column using a box plot



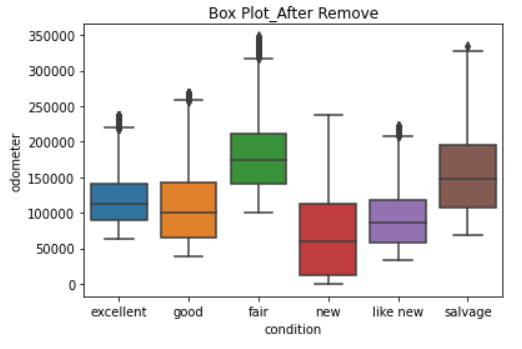
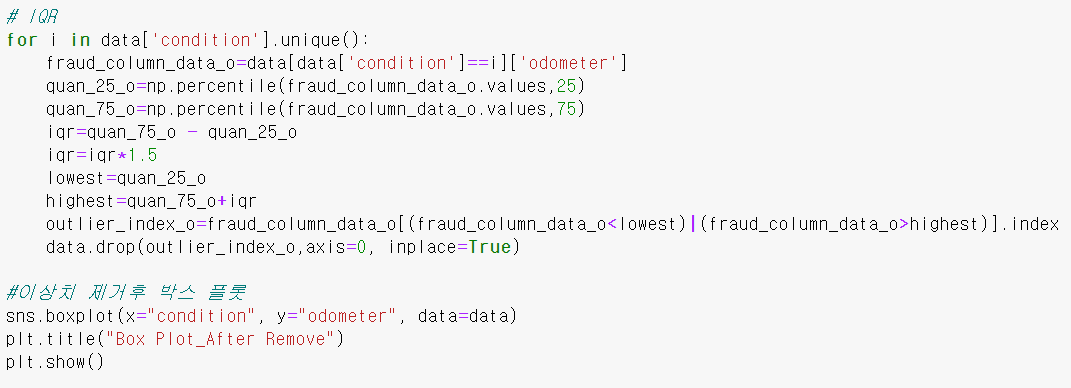
It is the code and result that deleting the outlier exists in 'price' using 'IQR'.

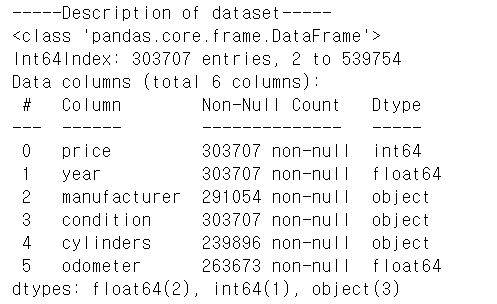
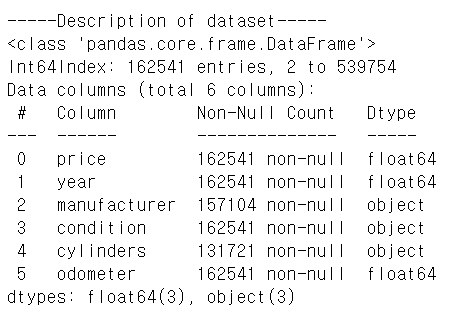


‘The following is the code and result is expressing the data of the ‘odometer’ column using a box plot



It is the code and result that deleting the outlier exists in 'odometer' using 'IQR'.

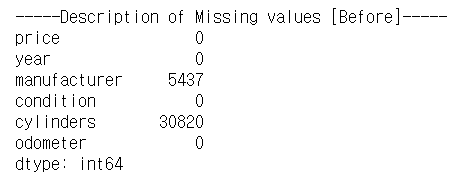
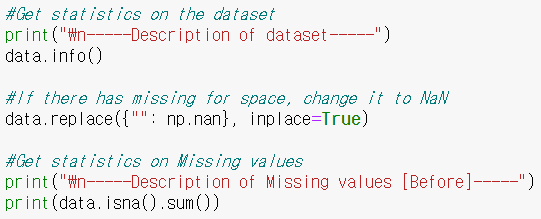


This is the result that has changed since the 'dirty data' processing was completed.

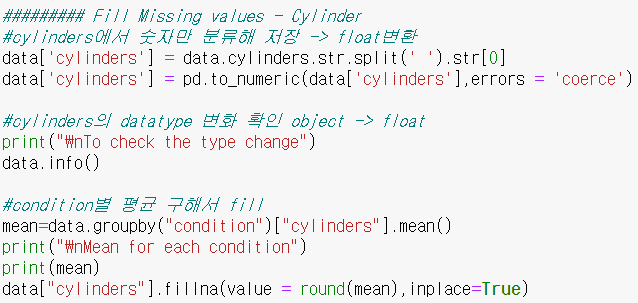
* 1. **Fill missing value**

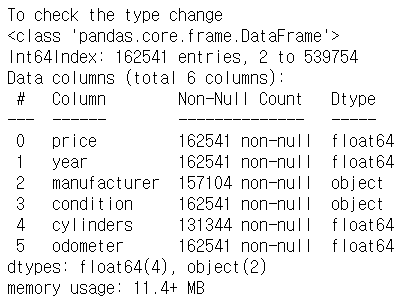
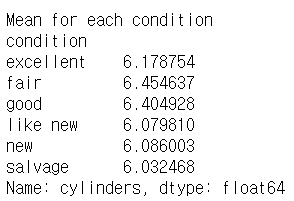
Missing value processing is essential for accurate target value prediction. After finding the missing value in the dataset, based on the target condition value, we divided it by condition and replaced the categorical feature with the mode value, and set limit=2 to drop the rest. The columns other than the categorical feature were replaced by the mean value for each condition.

First, change the empty data to missing value and check how many missing values are in each column.

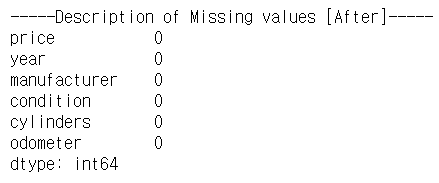
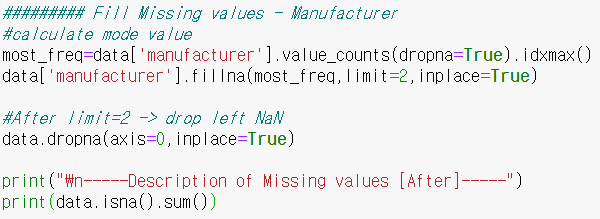


Next, change the cylinder column which has missing value from '8 cylinder' to '8 ' so that the data type of column can be calculated. Then, obtain the mean value for each condition and replace the missing value.





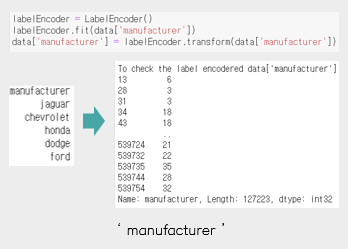
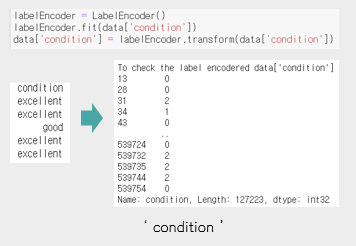
Finally, obtain the mode value for each condition and replace the missing value at the manufacturer column which has missing value. Also, set limit=2 and drop the rest.

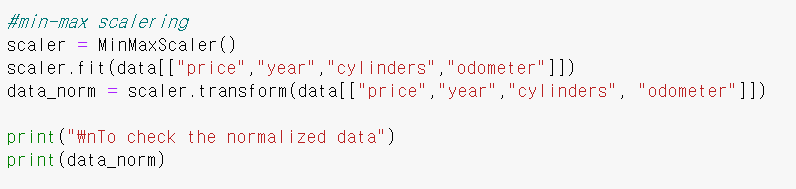


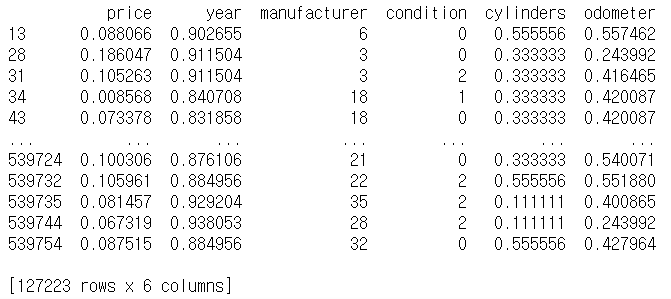
* 1. **Normalization**

After processing ‘Dirty data' and 'Missing value', we standardized using 'min-max normalization' to make 'features' scale equal. At this time, the value of 'categorical value' consisting of characters was quantified using 'Label Encoder'.

Then, we proceeded min-max normalization and completed data preprocessing by converting the data into a value between 0 and 1.

< Results using 'Label Encoder' >  


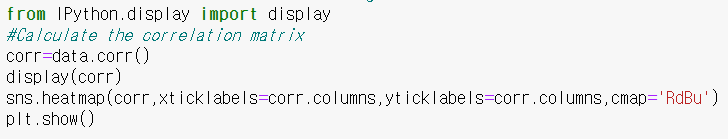


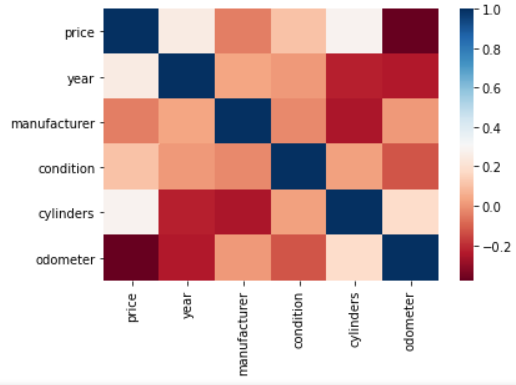
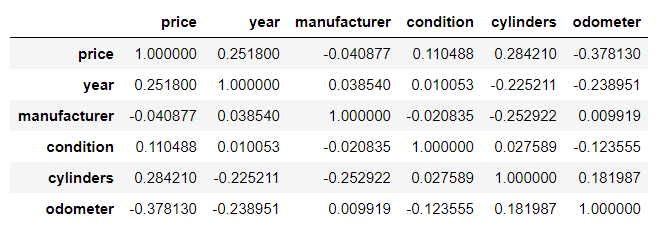
<Result standardized data>

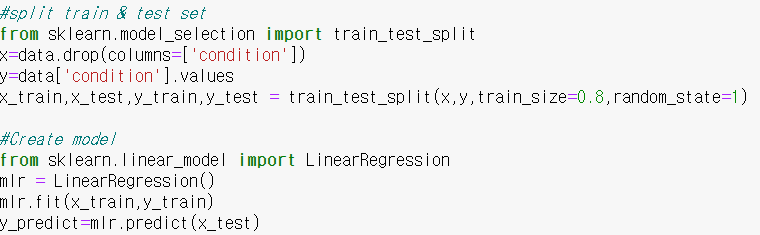
1. **Algorithms**
   1. **Multiple linear regression**

We use multiple linear regression to predict the target value condition because the used car dataset have to predict with multiple columns. Therefore, after dividing the columns except condition into train and test, the results for multiple linear regression were printed as a table for the summary by comparing the predicted values with the condition column.

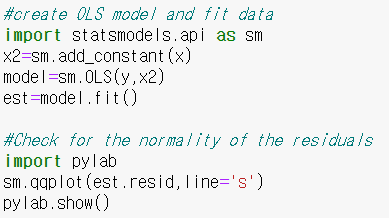
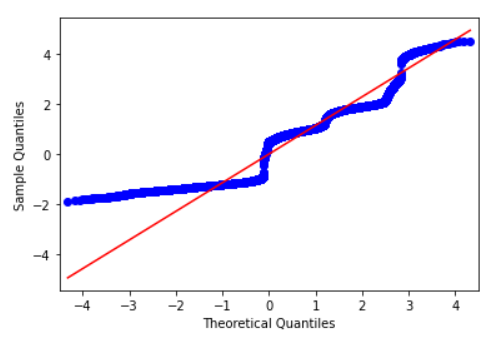
First, Represent the raw data as a matrix by correlation matrix.



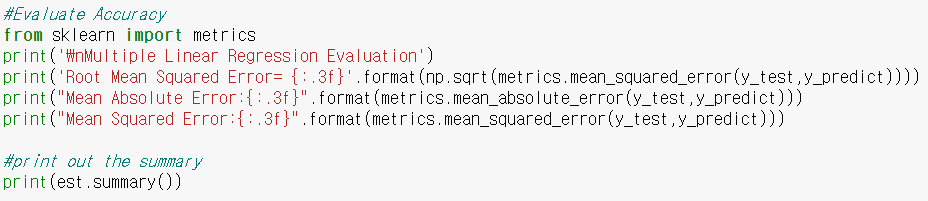


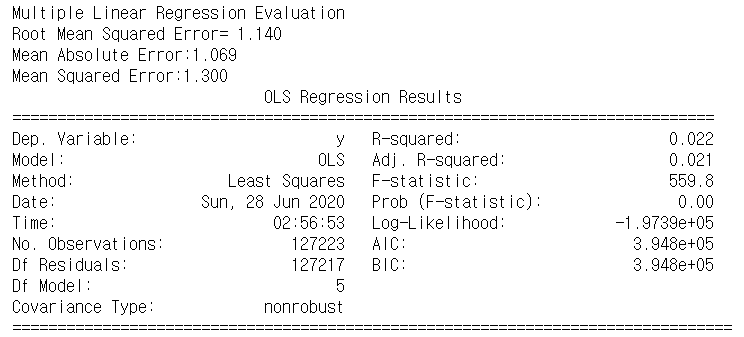
Next, the condition column is placed in y and the rest of the column in x, and the two are divided into train and test, and then learned through linear model. 

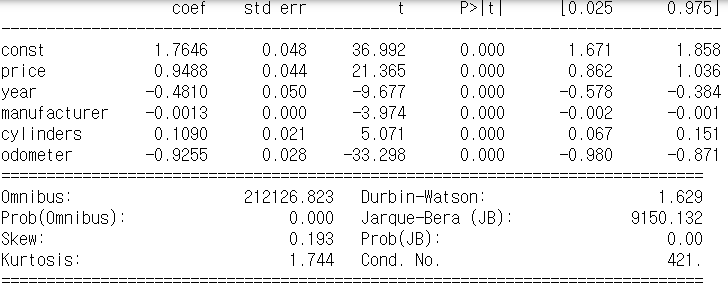
Create an OLS model and fit data to show the results of the multiple linear regression model in detail. And check for the normality of the residuals with the graph in below.

Finally, to evaluate the accuracy of the model, RMSE, mean absolute error and mean squared error is calculated, and the results of the OLS model are output in summary.





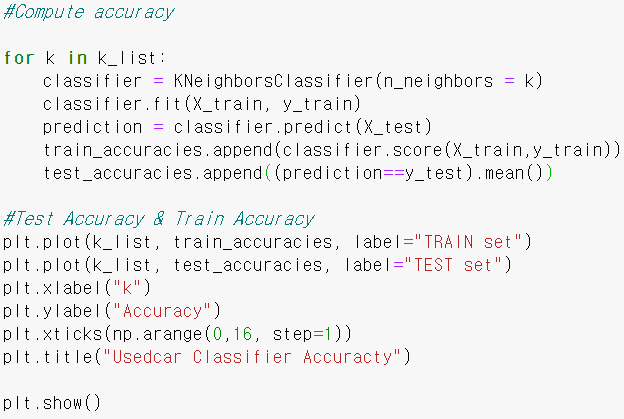


The accuracy of the multiple linear regression was low because, as you can see from the graph above, the accuracy of the blue dots in the front part was low because they were far from the red line so it shows the dataset contain anomaly value. Also, because the target value, condition column is a categorical feature, the accuracy of the multiple linear regression was low.

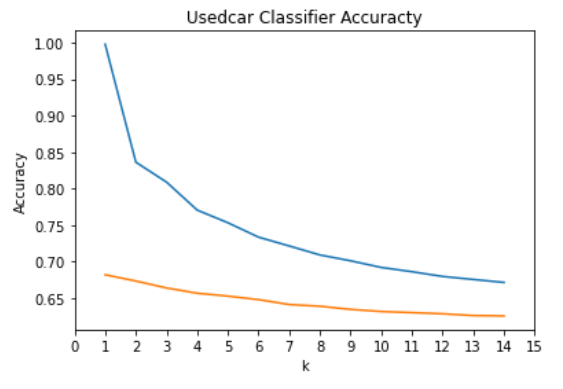
* 1. **Knn-classification**

'used car dataset' is the data that the target already has a class as 'catalogical value'. So we decided to use classification, of which we used Knn-classification.

The accuracy of Knn classification depends on the k value. So, to find the optimal k value, we calculated the accuracy of the case where k is 1 to 15 and plotted it on a graph.

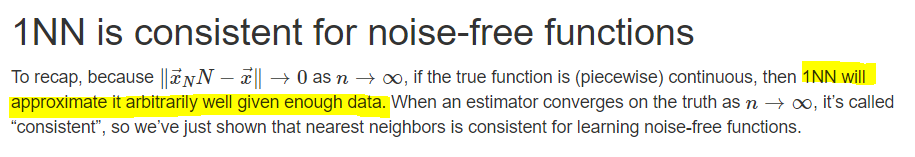


< Code for finding optimal k values >

< optimal k accuracy result graph >

The graph shows that the smaller the k value, the higher the accuracy.

However, when k=1, it is said that the best case is very special. So We looked at why these results came out.

[[2]](#endnote-2)

After looking at the various data, We saw an article saying, "If sufficient data exists, the accuracy of 1NN can be good.

Based on this article, the first result is as follows. "Because of the large amount of data sets we used." In fact, if you look at the data, "excellent", "good", which accounts for most of the total condition values, there were many more than other condition values. So when you look for close neighbor values, the above two targets are more likely to be predicted and already more distributed, so the greater the k, the more confusion the greater the accuracy is, and the higher the k=1, the higher the accuracy.

Our predictions may be wrong. But through this project, I think I could learn a lot by thinking about this.

Anyway, the result of setting k as 1 and applying KNN classification was 68% accuracy.

스크린샷이(가) 표시된 사진

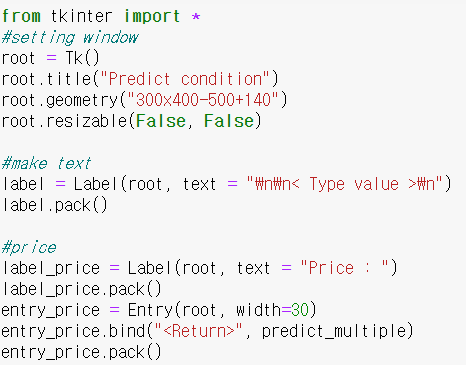
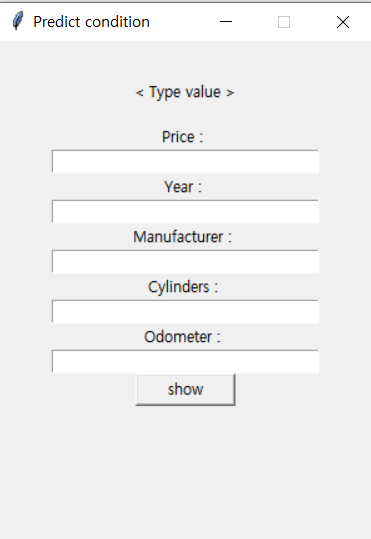
자동 생성된 설명

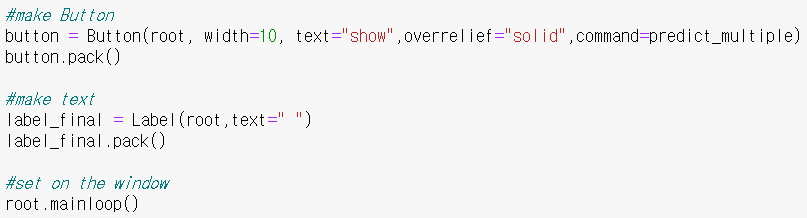
<knn algorithms computing accuracy code and accuracy output>

1. **Evaluation**
   1. **K-Fold cross validation**
   2. **Ensemble learning – Bagging**
2. **GUI**

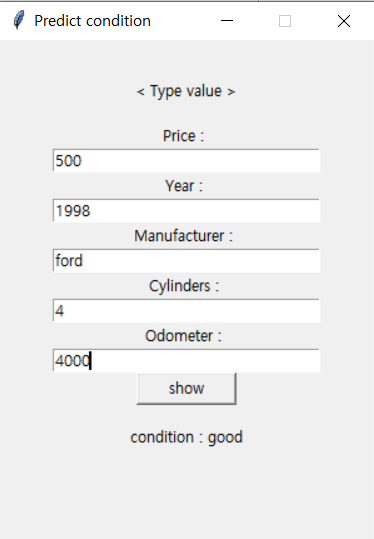
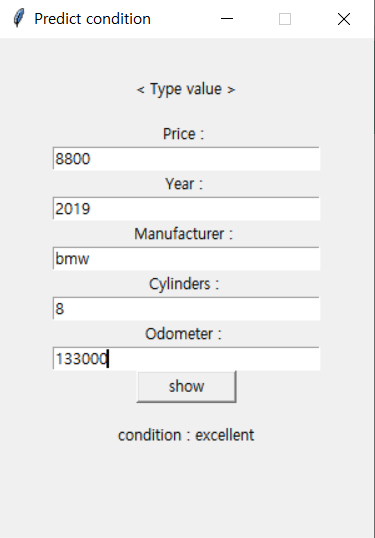
We created a GUI based on the code that we implemented, so that when the user enters it according to the column, we can predict the condition accordingly.

First, set the size and tickle of the window and insert the text, input window, and button to go inside.



Enter a value in the input window entry via the predict\_multiple function and click button to convert it to labelEncoder if it is not a number and show the predicted condition value through the ensemble in gui.

1. **Member role**

|  |  |  |
| --- | --- | --- |
| 윤혜주 | 조소영 | 이하영 |
| Data preprocessing | Data preprocessing | Data preprocessing |
| Data analysis (KNN) | Evaluation | Data analysis  (multiple regression) |
| Final presentation | Ensemble | GUI |
| Final presentation | Proposal PPT | Final PPT |

1. **Conclusion**

Significant difference in prediction accuracy depending on data preprocessing, and realized that preprocessing is very important.

In dealing with quite a lot of data, we experienced an unexpected error, the more systematic and meticulous data pre-processing is necessary.

it is necessary to apply appropriate algorithms according to the data, and all processes are important.

1. **Reference**

1. KBS NEWS, 『 중고차 성능·상태, 실제와 다른 경우 많아…‘주의’』, <https://news.kbs.co.kr/news/view.do?ncd=4261735> [↑](#endnote-ref-1)
2. K-Nearest Neighbors I, <http://www.stat.cmu.edu/~cshalizi/dm/19/lectures/09/lecture-09.html> [↑](#endnote-ref-2)