CS982 – Big Data Technologies

Data Analysis in Used Car Sale Advertisement in Ukraine

Presented by

Chavinpat Naimee (201976778)

MSc Machine Learning and Deep Learning

The study was supervised by

Dr Yashar Moshfeghi

University of Strathclyde

Submitted date

11 November 2019

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Currently, the worldwide used-car market is predicted to grow at a compound annual growth rate of 12.81% from 2018 to 2024 (Mordor Intelligence, 2019). This market is growing rapidly across the globe because many customers are unafforded to buy a new car owing to financial limitation (Statista, 2019). Thus, it would be a good signal for used-car business to trap this opportunity by using “Price” which one type of marketing mix strategy to classify product to serve with appropriate customer group such as high price focus selling on the wealthy customer.

Furthermore, when products are categorized in a suitable target group, it will be easier to do marketing promotion and select market channel. This statistic presents car prices in a different brand, year, mileage, model etc. which useful to second-hand car trader to sell cars or build the right marketing plan with appropriate customers.

Firstly, this report will discuss problems and key challenges to the data. Then using statistic from data analysis and summarise the main points. After that, unsupervised will be applied to cluster the data into the appropriate group using hierarchical clustering and k-means techniques. Finally, price prediction model will be built by supervised using linear and logistic regression.

Chapter 2 Dataset - Car Sale Advertisements

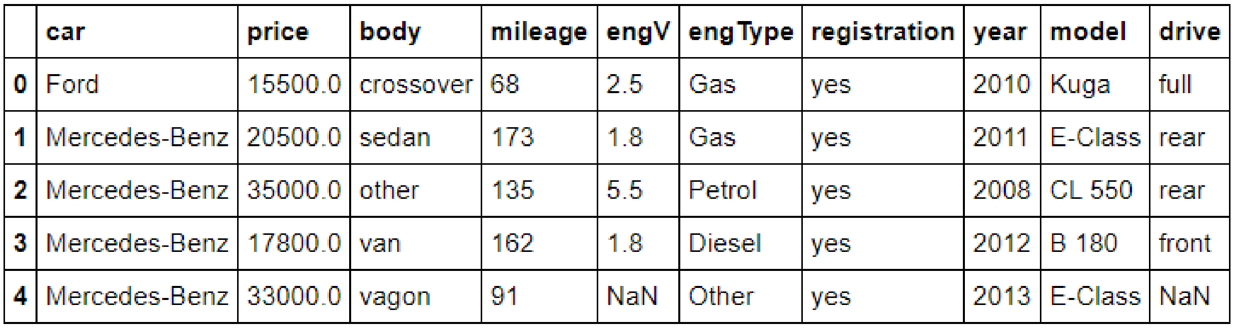
2.1 Key challenges and problem

Firstly, the dataset contains many unwanted data, called noise. So, data pre-processing needs to be processed before doing the task. This process might have to spend some time to deal with it because sometimes we did not know whether noise contains. Secondly, there are many exciting things in statistics for the data. For instance, which car brand is famous in Ukraine or How much should the price of a used car be? Furthermore, other surprise information that could be found. All information can be used by a car company to create a marketing campaign and select market channel. Besides, there are advantages to the customer, who decide to buy a second-hand car. Lastly, applying supervised and unsupervised is challenging because this study will attempt to train the model to predict the car in supervised, and try to group data

2.2 Introduction to a dataset

This dataset was collected by Anton Bobanov in 2016 and is freely download on the Kaggle website. Data collected from private car sales in Ukraine, which most cars are used cars, so there are many features involved in car operation. The full raw dataset consists of 9576 observations and ten attributes related to car operation. The samples of data shown in table 2-1.

Table 2-1 First five samples in the dataset



	car	price	body	mileage	engV	engType	registration	year	model	drive
0	Ford	15500.0	crossover	68	2.5	Gas	yes	2010	Kuga	full
1	Mercedes-Benz	20500.0	sedan	173	1.8	Gas	yes	2011	E-Class	rear
2	Mercedes-Benz	35000.0	other	135	5.5	Petrol	yes	2008	CL 550	rear
3	Mercedes-Benz	17800.0	van	162	1.8	Diesel	yes	2012	B 180	front
4	Mercedes-Benz	33000.0	vagon	91	NaN	Other	yes	2013	E-Class	NaN

The meaning of each variable will be explained below:

car:	manufacturer brand
price:	seller’s price in an advertisement (in USD)

body:	car body type
mileage:	as mentioned in an advertisement (‘000 Km)
engV:	rounded engine volume (‘000 cubic cm)
engType:	type of fuel (“Other” in this case should be treated as NA)
registration:	whether car registered in Ukraine or not
year:	year of production
model:	specific model name
drive:	drive type

2.3 Data Pre-processing

The “Car Sale Advertisements” dataset contains many noises in the data such as ‘NAN’ or car price is equal to zero, which noise is any undesirable data that makes more difficult to learn. Therefore, pre-processing data need to be done before analysing.

A data that car price and mileage lower or equal to zero will be deleted from the dataset because car price lower than or equal to zero is impossible and mileage values also cannot be lower than zero. Mileage equal to zero was ignored because this analysis will focus on used cars only. All data contain ‘NAN’ will be eliminated since the missing values will be the disturbance to the results. Hence, the number of rows or car data that will be used in this analysis have 8,213 used cars.

The dataset will be divided into 70% of the training set and 30% of the test set, which are 5,749 and 2,464 samples, respectively, in order to train models by using unsupervised and supervised methods.

Before training the unsupervised and supervised models, the category data, which are car, body, engType, registration, model and drive, need to be transformed to numerical data.

2.4 Data Analysis

2.4.1 Overview

Overall of the dataset shown in the scatter plot in figure 2-1. This chart shows the distribution of all numerical variables that can be used to observe any interesting situations. It can be noticed that price versus year and mileage versus year have a similar distribution.

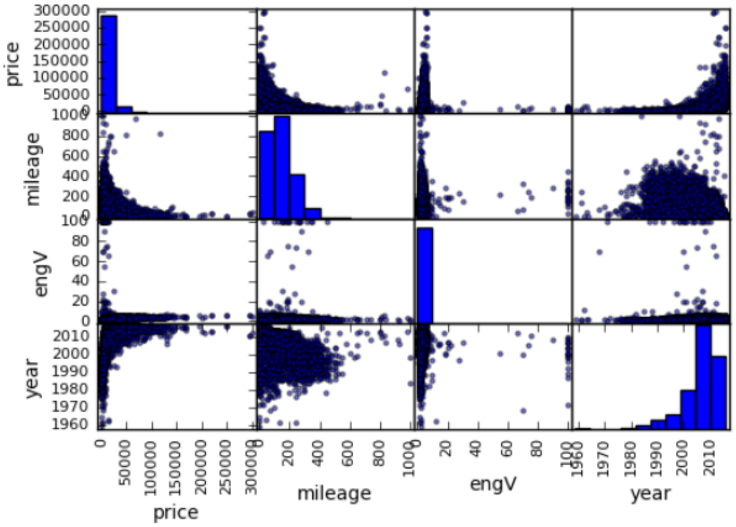


Figure 2-1 Scatter of all numerical data

This heat map in figure 2-2 shows correlations of the dataset. It is noticeable that there is a clear direct correlation between price and year of production (indicate to ages of a car) because it is common that older car should be cheaper than a newer car. There is an extreme inverse correlation between price and mileage and between year and mileage. The reason is older cars should be driven more, which increasing mileage, and then the cost of more lifetime is naturally lower.

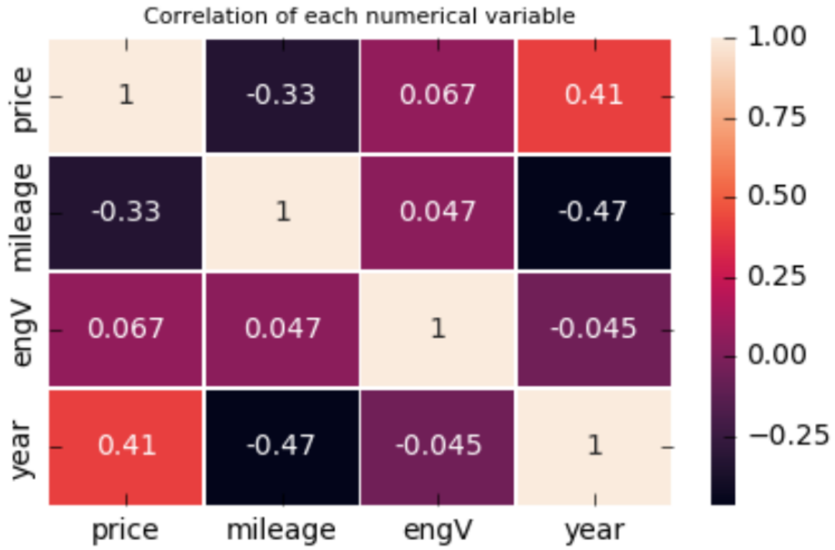


Figure 2-2 numerical variables heat map

2.4.2 Insight Data

Even there is a correlation between price and year of production, it cannot be commented that the price increases as the years increase, but in general, an increasing price had been observed in recent year and it increases dramatically until 2016. See figure 2-3.

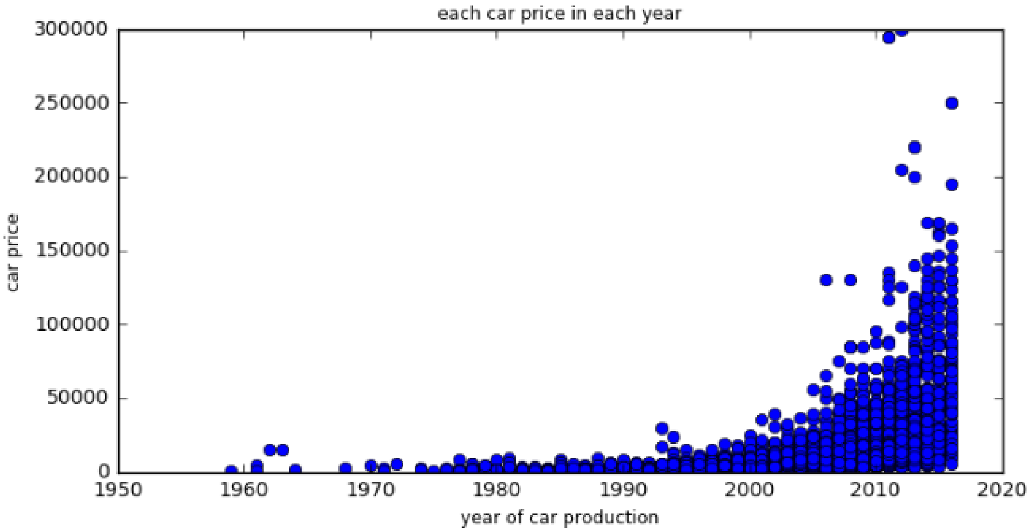


Figure 2-3 car price in each production year

The number of car sales in each year shows in figure 2-4. In this figure, It can be seen that, after 2000, there was an increasing in used car on-sale could mean citizens would like to buy a new car, so it made used car market boom. However, since the global financial crisis in 2008, the number of vehicles on-sale dropped dramatically, then it recovered again in 2011.

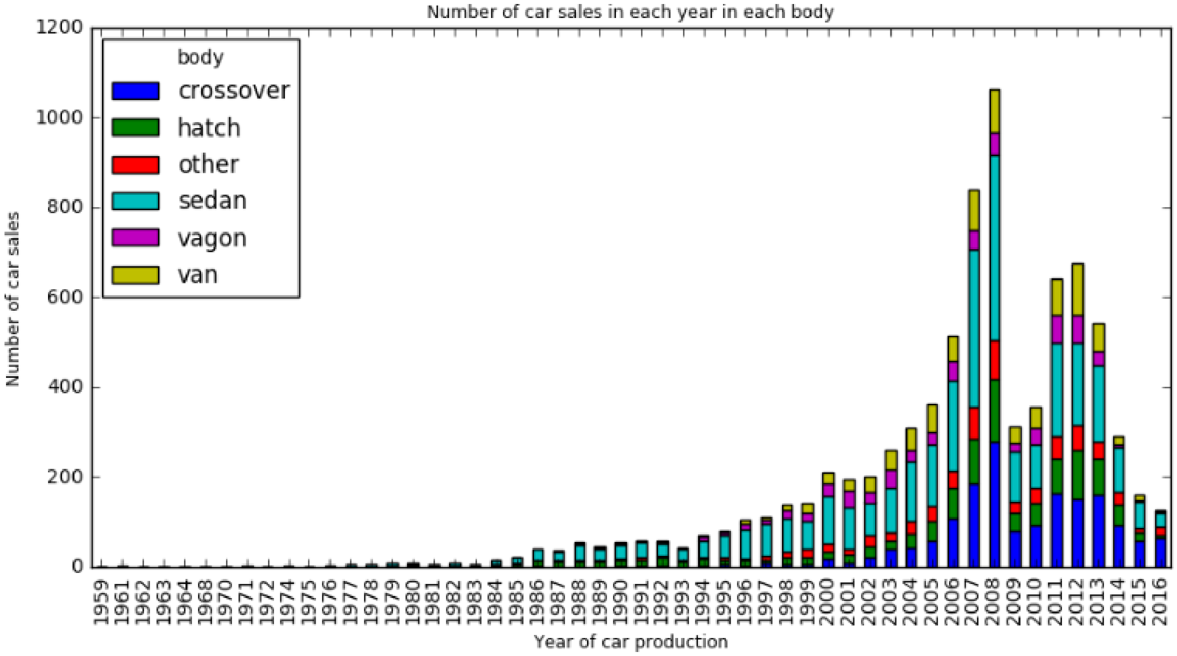


Figure 2-4 Number of car sale in each production year in each body

Volkswagen, Mercedes Benz and BMW, which are German brands, are the top three car brands that were selling in the market. There are 810, 720 and 594 cars on-sale of Volkswagen, Mercedes Benz and BMW, respectively. These three companies are approximately 25% of whole cars on-sale. See figure 2-5. This chart can also tell that the German brand car is the most popular in Ukraine. It is interesting that fourth place is Toyota, which is the Japanese brand.

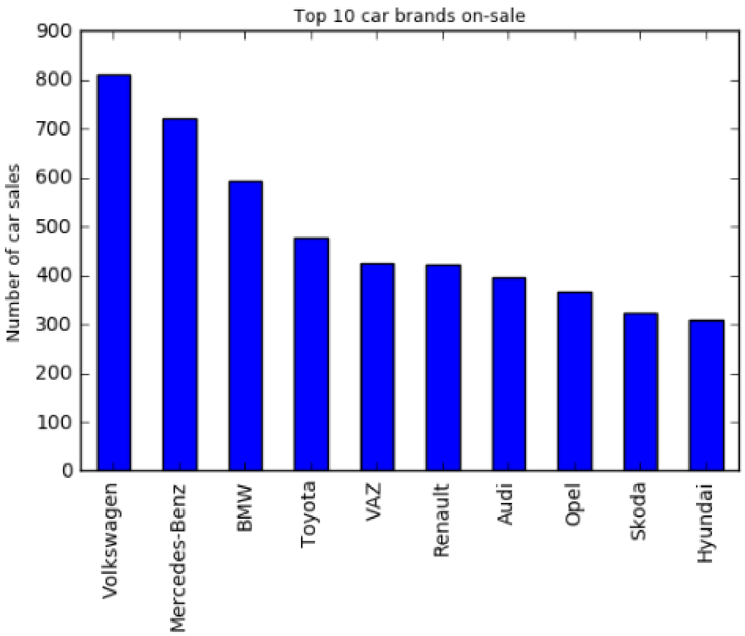
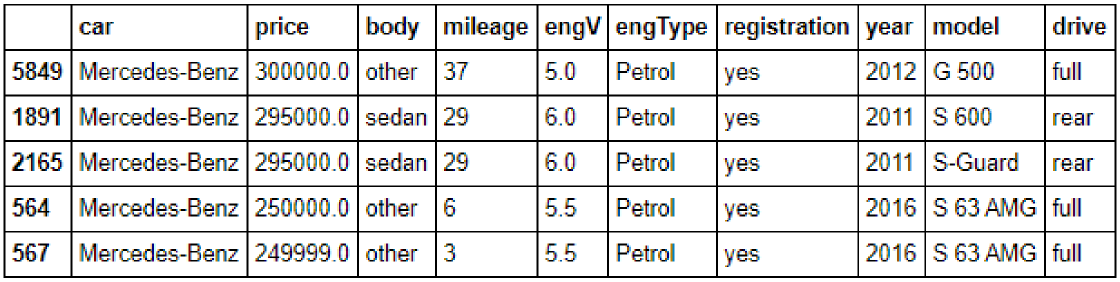


Figure 2-5 Top 10 car brand on-sale in Ukrine

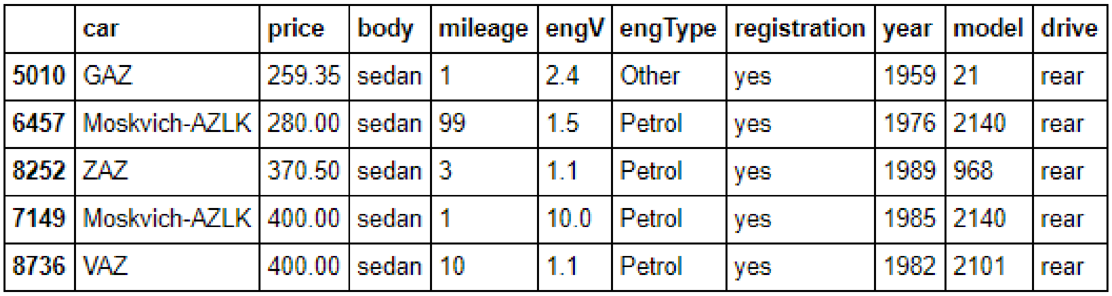
This data is also helpful for selling car company that can build marketing campaigns to meet sales goal. Table 2-2 shows the five most expensive cars that could be used to create marketing campaigns to high salary group of people. On the other hand, Table 2-3 illustrates the five cheapest cars that would target low income people.

Table 2-2 Top 5 most expensive cars on-sale



	car	price	body	mileage	engV	engType	registration	year	model	drive
5849	Mercedes-Benz	300000.0	other	37	5.0	Petrol	yes	2012	G 500	full
1891	Mercedes-Benz	295000.0	sedan	29	6.0	Petrol	yes	2011	S 600	rear
2165	Mercedes-Benz	295000.0	sedan	29	6.0	Petrol	yes	2011	S-Guard	rear
564	Mercedes-Benz	250000.0	other	6	5.5	Petrol	yes	2016	S 63 AMG	full
567	Mercedes-Benz	249999.0	other	3	5.5	Petrol	yes	2016	S 63 AMG	full

Table 2-3 Top 5 cheapest car on-sale



	car	price	body	mileage	engV	engType	registration	year	model	drive
5010	GAZ	259.35	sedan	1	2.4	Other	yes	1959	21	rear
6457	Moskvich-AZLK	280.00	sedan	99	1.5	Petrol	yes	1976	2140	rear
8252	ZAZ	370.50	sedan	3	1.1	Petrol	yes	1989	968	rear
7149	Moskvich-AZLK	400.00	sedan	1	10.0	Petrol	yes	1985	2140	rear
8736	VAZ	400.00	sedan	10	1.1	Petrol	yes	1982	2101	rear

Chapter 3 Unsupervised Analysis by Clustering

Unsupervised will be used when a dataset lack supervisor or training data, which means there is unlabelled data in the dataset. Clustering is an unsupervised method that users can solve problems with little or no idea that how the result will look like and then the hidden structure of the data will be found by clustering the collected data from a relationship of the variables of the data (EMC Education Services, 2015).

The purpose of this chapter is to cluster the body of cars into n clusters that the algorithm will generate itself. All columns will be used to train the model because every column has an impact on a buyer or seller to make the decision when they want to buy or sell a car.

3.1 Agglomerative Clustering

The agglomerative clustering is a hierarchical clustering that each data is initially put in its group (Price. *et al*., 2019). Then the group will be joined by the most similar attributes using a bottom up approach. The distance of data is calculated by the distance between observations of pairs of groups. In Scikit Learn, the linkage criterion is used to determine the gap between data. In this study 'Ward linkage' and 'Euclidean affinity' will be used as parameters. 'Ward' is a variance-minimising approach to minimises the sum of squared distance within all clusters and 'Euclidean' is a distance between a pair of points that suitable for numerical data (Scikit-learn Machine Learning in Python, 2019).

This data was applied hierarchical clustering dendrogram to cluster without any idea of the result. The result is shown in figure 3-1, that there are two different colours, which are green and red. However, it was difficult to read because the row of the data is copious. Truncation was used to condense the dendrogram in figure 3-2. The result in figure 3-2 is easier to understand, which shows the number of samples in the chart, but still, do not know the labels of each group. Then Agglomerative Clustering would be applied to the data.

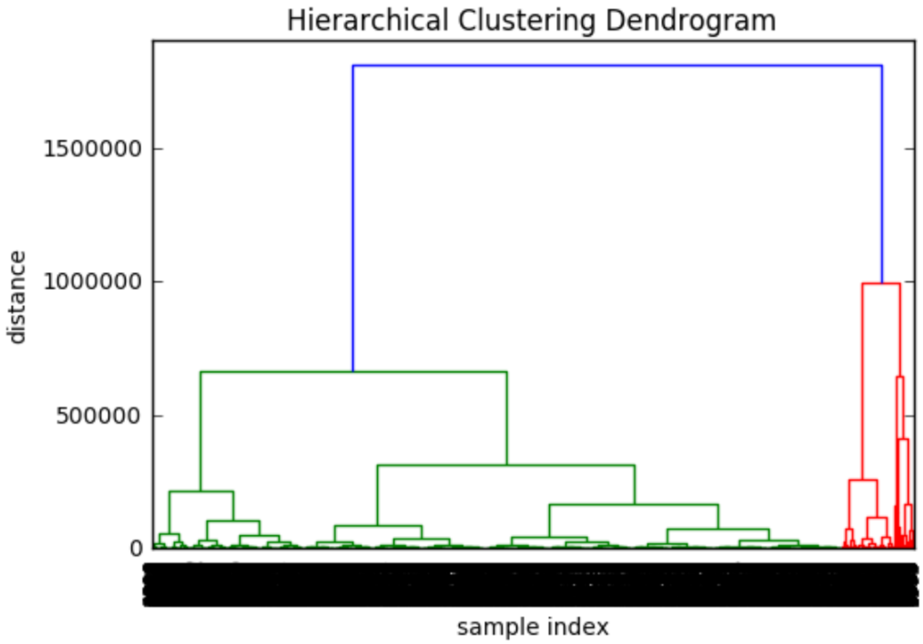


Figure 3-1 Hierarchical dendrogram with truncate_mode = None

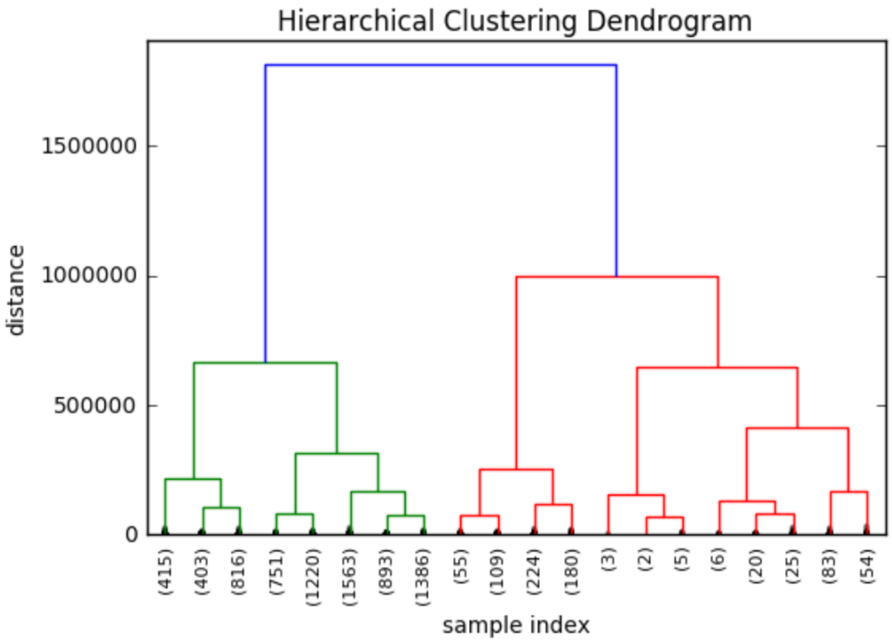


Figure 3-2 Hierarchical dendrogram with truncate_mode = ‘lastp’

Figure 3-3 illustrates the Silhouette score in each number of clusters. The silhouette score decreases when the number of cluster increases. There is the highest score in k equal to 2 (2 clusters) that is consistent with the result of the dendrogram.

Silhouette Score is a measurement of how similar an item is to other objects in its group compared to objects in different groups. Value range -1 to 1, which 1 indicate well match in its cluster (Scikit-learn Machine Learning in Python, 2019).

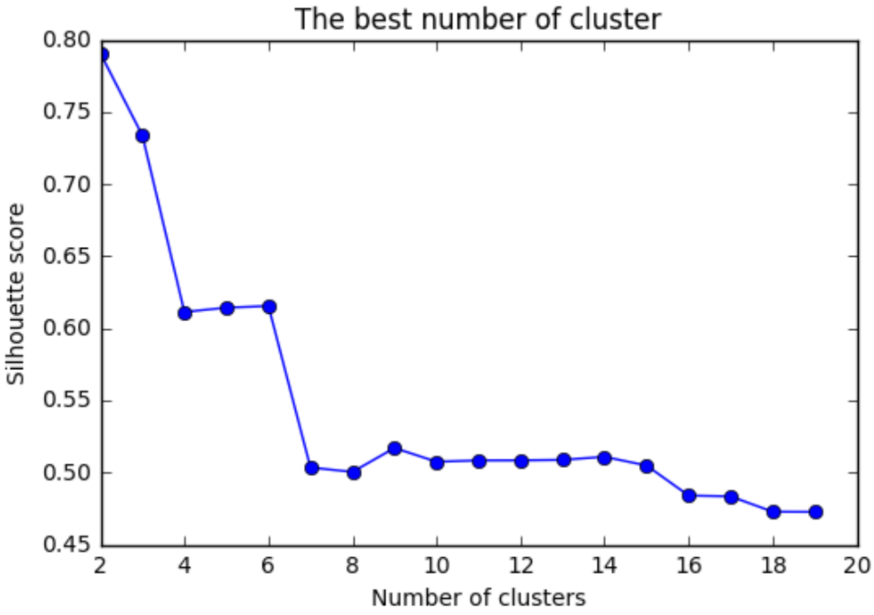


Figure 3-3 Silhouette score depending on the number of clusters

3.2 K-Means Clustering

K-Means is a popular clustering technique that that divide n data points into k clusters by the closest mean, n is a number of data and k is a number of groups. K-means need an input k as an initial number of clusters (EMC Education Services, 2015).

The basic algorithm of K-means is:

1. Choose K points as the initial centroids. Then repeat.
2. From K clusters by assigning each point to its closest centroid.
3. Recalculate the centroid of each cluster until the centroids do not change.

K-means is computed for k equal to 2 to 19 clusters, the score that measures the performance of each number of clusters is shown in figure 3-4. The model performs the best when two clusters, which consist with the hierarchical clustering method. Completeness and Homogeneity score are also computed, which compared to the body's column of the data. In this case, the aim that calculated Completeness and Homogeneity score that is to find the consistency between them. However, both scores are very low in every number of clusters. As a result, the data is not suitable for clustering method. The improvement of the model will be discussed in reflection's chapter.

Completeness Score result satisfies completeness if all data points that are members of a given class are elements of the same cluster. Value range 0 to 1, where 1 indicates all objects are in the right group (Scikit-learn Machine Learning in Python, 2019).

Homogeneity Score result satisfies completeness if data points consist of members in the same cluster. Value range 0 to 1, where 1 indicates all items are the same in its group (Scikit-learn Machine Learning in Python, 2019).

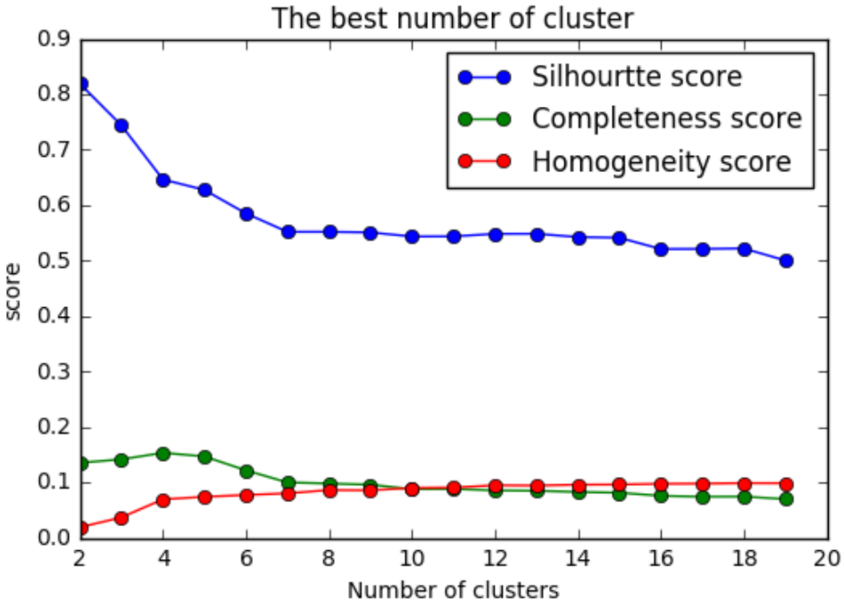


Figure 3-4 Silhouette, Completeness and Homogeneity score in each number of clusters

Chapter 4 Supervised Method

The supervised method can apply what has been learned in the past to new data using labelled examples to predict the future result (Mohammed, 2016). The training data composed of input X and output Y of labels. For example, the learning algorithm obtains the set of inputs X with their actual outputs Y, and then the algorithm can assess its actual output with correct output to find errors to adjust the model correspondingly (Alpaydin, 2014). It attempted to forecast outcomes in a discrete output in a classification problem. In other words, input variables are attempted to map into separate classifications. Generally, it is used in applications that historical data can predict future events.

4.1 Linear regression

Linear regression is a mathematical technique to create a model the relationship between two variables by fitting a linear equation to observed data (Price. *et al*., 2019). In this data, there is a correlation between price and year of production, so linear regression will be used to build a model to predict the car price from the year of production.

The linear regression equation is given below:

$$Y = a + bX + e$$

where **X** is the year of production, **Y** is the car price, **b** is the slope of the data, **a** is the intercept and **e** is the residual (error).

$$Y_i = a + bX_i + \ e_i$$

$$\hat{Y}_i = a + bX_i \quad , \hat{Y}\ is\ estimate\ value.$$

In Linear regression, the performance of the model is evaluated by root mean square error (RMSE) and the coefficient of determination (R^2).

Root mean square error (RMSE) is the square root of the average of sum of the squares of residual, which is defined by

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} {e_i}^2} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}$$

The coefficient of determination (R^2) is how much the total variance of the dependent variable can be reduced by using least square regression. If R^2 equal to 1, the model is a perfect fit or very confident. If R^2 close to 0, the model is a poor fit and it is not convinced. R^2 is defined by

$$varience : Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$R^2 = \frac{explained\ varience\ of\ Y}{total\ varience\ of\ Y}$$

The results show in table 4-1, which RMSE is very large and R^2 is very low in the prediction of both the training set and test set. This means that the model is not fit well in this data. Therefore, linear regression is not suitable to predict car price from a year of production. Even there is a correlation between them, but the relation of these data does not have to be linear. Many outliers appear in the chart since some car could be a limited edition or exclusive vehicle that is very expensive, see figure 4-1. How to improve the model will be explained in the reflection section.

Table 4-1 RMSE and R^2 of the training set and test set

	Training set	**Test set**
RMSE	270024939.365	333243024.706
R^2	0.173	0.162

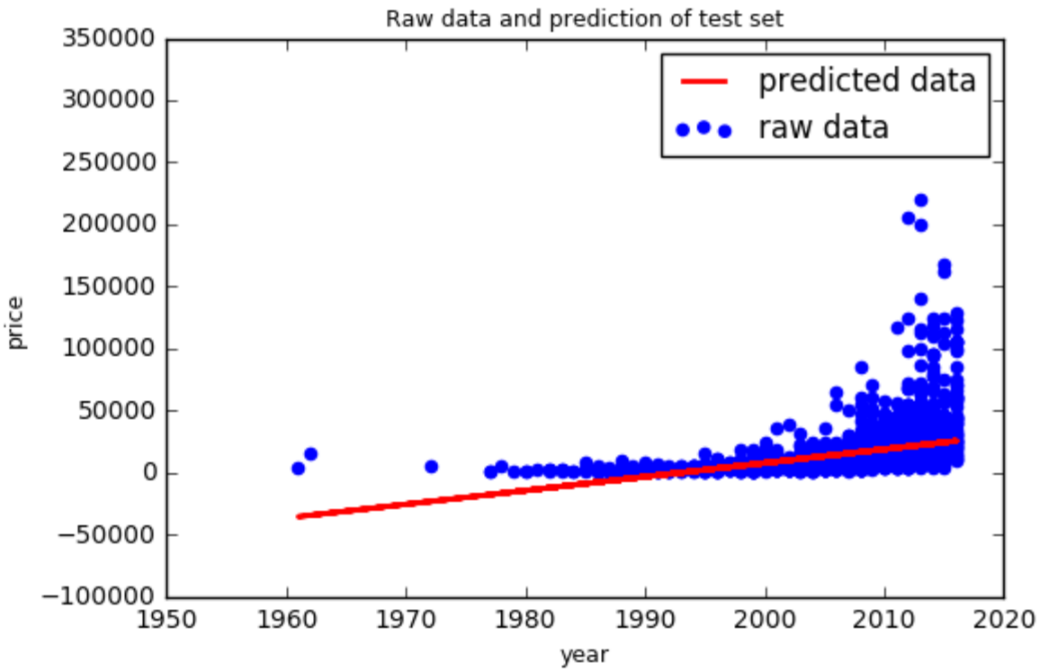


Figure 4-1 Raw data and predicted data of test set

4.2 Logistic regression

Logistic regression is a classification method, which can use to predict an outcome that is a category. It will predict the outcome based on the input variables. Although, logistic regression works well on the binary condition, which is the outcome represents two categories such as yes/no or true/false. However, it can be used in multiclass classification. Logistic regression can be used to predict the missing values if missing values in a category. In this task, the prediction of the body of cars will be illustrated as an example. The score of precision, recall and F1 use to measure the performance of logistic regression.

Precision expresses how often the classifier is right when predicting a given class

Recall expresses how often the classifier notices that a sample of a given class belongs to that class

F1 represents the average of Precision and Recall

The results are shown in table 4-2. Overall, the model predicts correctly approximately 80%, which the total scores are acceptable. There is only a vagon type that model did not once predict accurately, it might because the data of vagon seem similar to another body's type or it did not contain in the test set. As a result, this model can apply to predict data that missing the body of a car in every type of body, except vagon type.

Table 4-2 the performance of logistic regression model

Body / score	Precision	Recall	F1 - score	Support
0 - crossover	0.95	0.97	0.96	497
1 - hatch	0.92	0.87	0.89	297
2 - other	0.93	0.14	0.24	191
3 - sedan	0.72	1.00	0.84	987
4 - vagon	0.00	0.00	0.00	204
5 - van	1.00	0.97	0.98	288
average	0.78	0.82	0.77	Total = 2464

Chapter 5 Conclusion

This study analyses used car sale advertisements in Ukraine. Overall, this analysis is beneficial to used car trader or seller that can create a marketing campaign and select market channel to the right customer. In another perspective, there are advantages to the customer who is going to buy a used car.

In supervise and unsupervised approach, the results have been inconclusive because models in this study did not work well in this data.

The unsupervised method by clustering provided the result as two clusters without any labels. Therefore, problems in this study cannot be solved by this algorithm.

Linear regression, a supervised technique, do not fit the data well. There are too many outliers in the results.

Another supervised approach, logistic regression seems to be better than other. The accuracy of the model is around 80%, that is acceptable. However, the benefit of the model is that use to predict missing value, which does not answer the problems.

To sum up, the models need to be improved efficiently to conduct with these issues.

Chapter 6 Reflection

I had searched for a week to find the dataset that I am interested in until I found this data. In the beginning, I would like to use a dataset from Thailand, which I am more familiar, and it can also use in my country. Honestly, I could not find any dataset of Thailand that there is enough data and suitable for this task. I realise that this is the data from Ukraine and I also do not know much about that country. However, I think this dataset is still beneficial for me, which I can train different models and approaches with the raw data and apply it to data in my own country as a further step. One thing that makes me surprise is Toyota, a Japanese brand car, is the fourth ranking of the number of vehicles on-sale. My question is why Toyota? In my view, there are many high-class brands in Europe that better than Toyota, but why people use it.

Let’s discuss the outcomes of this study. The most challenging task for me is to find the problems and analyse them. This is the first time that I had to indicate the issues and investigate them by myself. However, after I got the key challenges, then I can deal with it. There are not any issues in the outcomes of the statistic part, and I think it is necessary for a real situation.

On the other hand, in the supervised part, the linear regression model is not suitable for this data. The model could be improved by data transformation techniques such as polynomial expansion or principle component analysis. Logistic regression seems to work well, but utilisation is limited. Clustering, an unsupervised approach, provided two clusters as a result which I still have no idea the label of it. Next step, using other different models to solve these problems will be done.

Bibliography

Alpaydin, E. (2014) *Introduction to machine learning.* 3rd edn. distributor & M.I.T. Press.

Bobanov, A. (2016). *Car Sale Advertisements.* [online]. [Accessed 6 November 2019]. Available from: https://www.kaggle.com/antfarol/car-sale-advertisements

EMC Education Services, (2015) *Data Science & Big Data Analytics*. Available at: https://onlinelibrary-wiley-com.proxy.lib.strath.ac.uk/doi/book/10.1002/9781119183686?fbclid=IwAR1XI7Z8sioW0XXx_FO_yySZOk6AmODLZjDXYqCKenJn1w5FWem9_O1_jqA (Downloaded: 1 November 2019).

John, H., Darren, D., Eric, F., Michael, D. and the Matplotlib development team (2019) *Matplotlib version 3.1.1*. Available at: https://matplotlib.org/3.1.1/index.html?fbclid=IwAR1Zajh1k36cZRi9f3iUQSfJf9ayMoi9UDdbXoym-UyHgEe-tFAdq7ViwvE (Accessed: 5 November 2019).

Mohammed, M. (2016) *Machine learning, algorithms and applications*, Boca Raton, CRC Press.

Mordor Intelligence (2019) *Global Used Car Market– Growth, Trends, And Forecast (2019 – 2024).* Available at: https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024 (Accessed: 5 November 2019).

Nelli, F., (2018) *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Available at: https://link-springer-com.proxy.lib.strath.ac.uk/book/10.1007%2F978-1-4842-3913-1?fbclid=IwAR3Ig64J4CsIq_6nZYXKdSlHrnhnhVNPqghEAfSoUHDQYA1XCrLkzC0A-u4 (Downloaded: 1 November 2019).

Price. *et al*. (2019) *Essentials of Business Analytics: An Introduction to the Methodology and its Applications*. Available at: https://link-springer-com.proxy.lib.strath.ac.uk/book/10.1007%2F978-3-319-68837-4. (Downloaded: 1 November 2019).

Scikit-learn Machine Learning in Python (2019) Available at: https://scikit-learn.org/stable/index.html (Accessed: 7 November 2019).

Statista Research Department (2019) *The UK Used Car Industry - Statistics & Facts.* Available at: https://www.statista.com/topics/2190/the-uk-used-car-industry/ (Accessed: 5 November 2019).

Appendix

Development environment: Anaconda Navigator

Language: Python 3.5

Software versions: Jupyter notebook 4.2.3

Packages used

- numpy
- pandas
    - scatter_matrix
- seaborn
- matplotlib.pyplot
- scatter_matrix
- sklearn
    - model_selection
    - preprocessing
    - linearRegression
    - metrics
    - mean_squared_error
    - r2_score
    - train_test_split
    - PolynomialFeatures
    - LogisticRegression
    - cluster
    - scale
    - PCA
- scipy.cluster.hierarchy
    - dendrogram
    - linkage