INM430 Coursework - A Tiny Data Science Project

Karunakar Kotagaram, City University London, UK

Project Title: *London cars in UK*

***Abstract* —** The aim of this project to provide an introduction and different ways to predict the used car prices in London area based on historical data by using different analytical like linear regression, random forest and K-mean clustering techniques. The data is collected from Auto Trader group and from various sources. The results are cross-validated to find which method provides the best performance. Though, some of the predicted used car prices are slightly different than predicted prices so different structure and network methods required to predict better results.

1. *Introduction:*

Predicting the used car prices and depreciation cost is challenging task and important problem. According to the data obtained from BBC news, the number of used car sold in the UK has crossed more than 4.8 million vehicles in the first half of the year 2015, this was an 8% increases on second-hand car sales. *"Mike Hawes, chief executive of the SMMT, said: "The UK's used car market is at its strongest ever."* Owners who bought the new cars also wanted to know the resale price of the used car in forthcoming years because this will help them to get an idea how much will be the car cost in the market. It is always good practice to predict the car prices and check the old depreciation value before we buy the car. There are several factors to be measured to predict used cars prices, here we will emphasis on the factors such as model, year, manufacture, mileage, body style, doors, engine size, horsepower, fuel usage per gallon in kilometres, weight, age, emissions, colour and interior and exterior design, Sat Nav GPS, extra facilities like rear cameras etc. Also, there are other factors to be considered like servicing history, seriousness of road accidents, and is there any damages in the car. Therefore, predicting the used cars price is a very admirable for buyers, dealers, and private owners. In this project, we will emphasise how accurately machine learning algorithm techniques can be used to predict the car price and compare results with other methods.

Research question: Does the car price decrease with the increase of mileage over the period and predict the variation between actual and predicted car price?

We can see the below figure(1) and figure(2) how the used car price distributing based on mileage over the period.

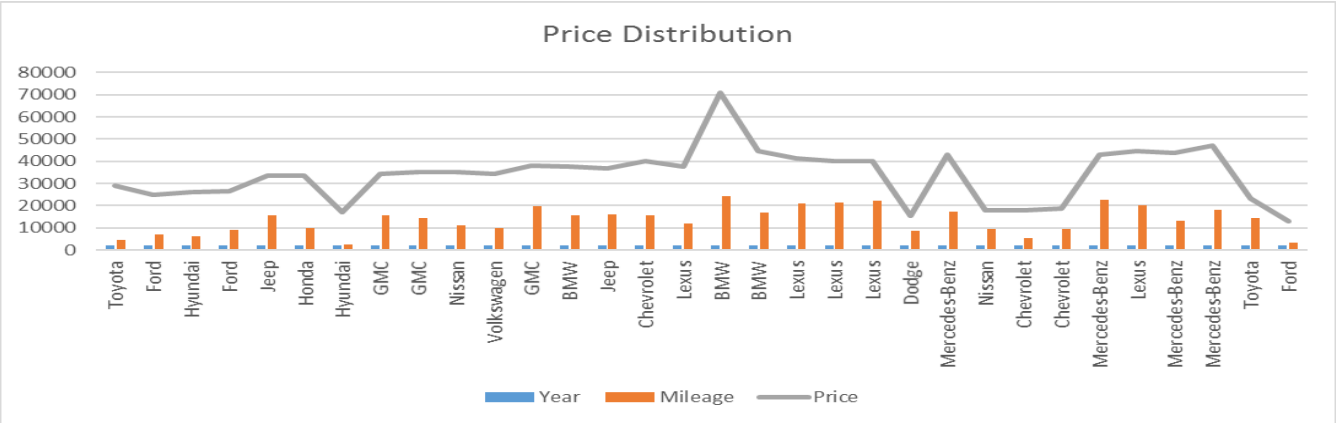


Figure (1) Average Car Price distribution based on make, year, mileage.

2. *Project Domain and Data sources:*

The London cars data set is collected from UCI machine learning repository and Autotrader, UK. This CSV file contains 206 rows and 31 columns, 19 numeric. The dataset contains all the used cars for sale in Autotrader.co.uk from year 1996 to 2015. The main aim of this project is to forecast car prices based on selected attribute set to analyse how accurateness the results. Also, I want to check the dissimilarity between forecast car prices and actual car price. There are so many factors to be considered to predict the car prices for example if we take auto trader statistics says diesel car prices drastically fallen since October 2015 and they are now cheaper than petrol car. In this project, I consider some factors like make, model, year, mileage for forecasting the future prices.

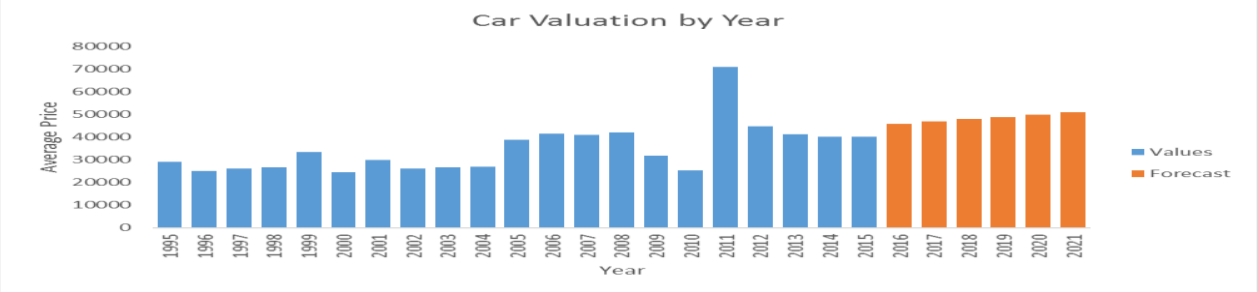
Through this project, we will explore the dataset using Python, Tableau, MATLAB. Here are the few libraries which will be used for data analysis in Python such as NumPy, SciPy, Pandas, Matplotlib.

Exploratory analysis: Load the Londoncars.csv to dataframe using pandas, identify attributes data types, missing values and select the right data for analysis.

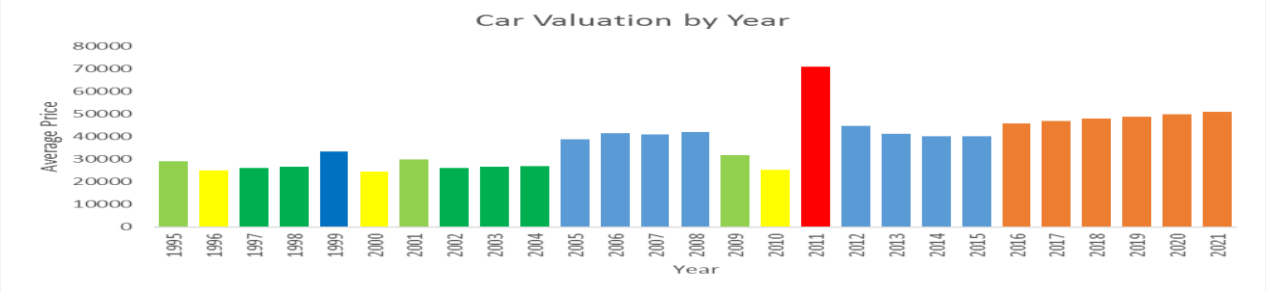
Model build: Build predictive modelling like regression model using python code. Also, I used Seaborn, Matplotlib libraries for statistical data visualisations.

Visual Analytics: Using Tableau and MATLAB software to display different visualizations.

We start the initial analysis with Tableau and MATLAB. In the figure (2) below average predicting car prices over the years is plotted. Here, we can observe the year 2014 with high price rate so it indicates about outliers in the data. This plot shows the prediction cars formally based on the year in which car was manufactured. We will be doing more analysis in further section for other factor like mileage, and depreciation.

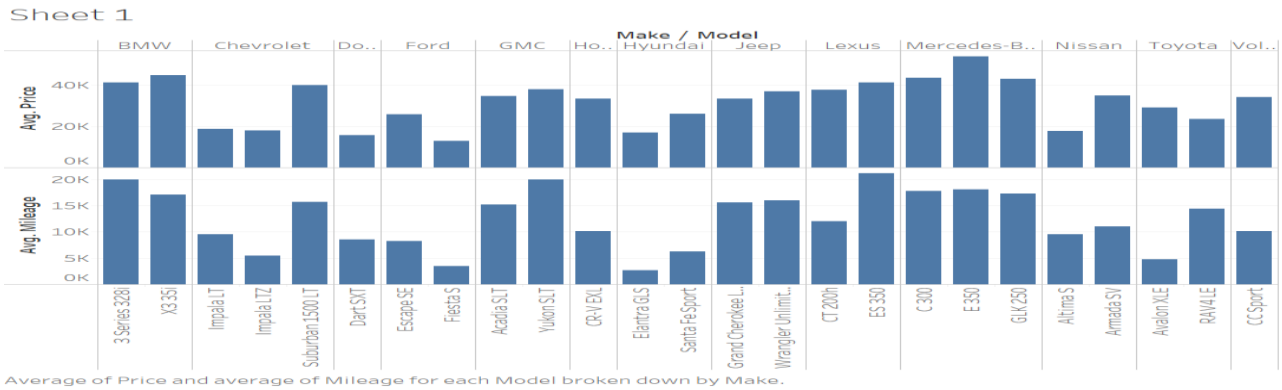


Figure(2) represents the forcasting car price by year.



Figure(3) indicates about outlier in red colour in the dataset.

For initially data visualisation and data discovery I have used the Tableau software as it is easy and very quickly to visualise the data features, in the figure (4) shows the average car price is plotted for each model and make.



Figure(4) Average of price for each model broken down by manufacturer

3. Developing an Analysis Strategy:

a) Cleaning data:
Originally there are 206 rows in the in the dataset with 31 attributes. In the data, some of them are numeric values and other categorical values. Data is never perfect and is segregated, it is important to find the missing data to avoid getting wrong results after analysis. I verified all the 31 columns names are correctly labelled by using python. Firstly, we will fill the missing values with zero of that specific column and then we will fill with mean values of each column, comparing the results. We can observe that replacing with mean values to be fitting better comparing to other data points.

b) Outlier detection: Outlier analysis is to identify the unusual occurrences in prices trend at certain years or mileage. There are several ways to check the outliers, here comparative analysis has been used to find the unusual trends. I have used scatter plot for plotting the data points in two ways i.e. 1D and 2D outlier. We can measure the points that are on the outside boundaries of the distribution. To detect these outliers, I applied the high dimensional Mahalanobis distance method. Upon investigation, unusual higher price was observed. In the below figure (6), the dark green spot is our outlier because of its unusual highest price.

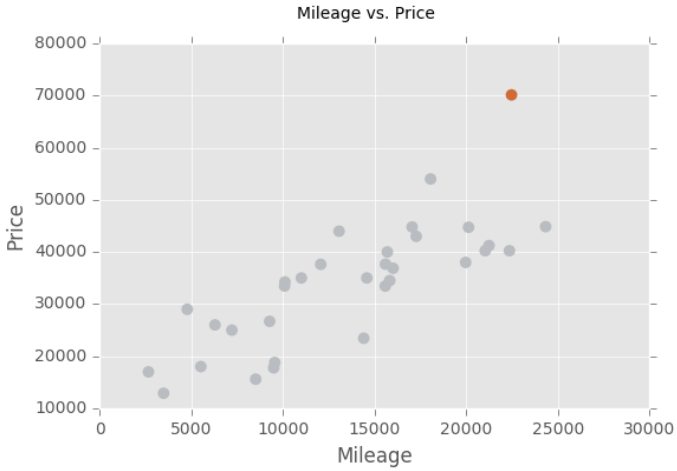


Figure (5) average car price with 1D outlier

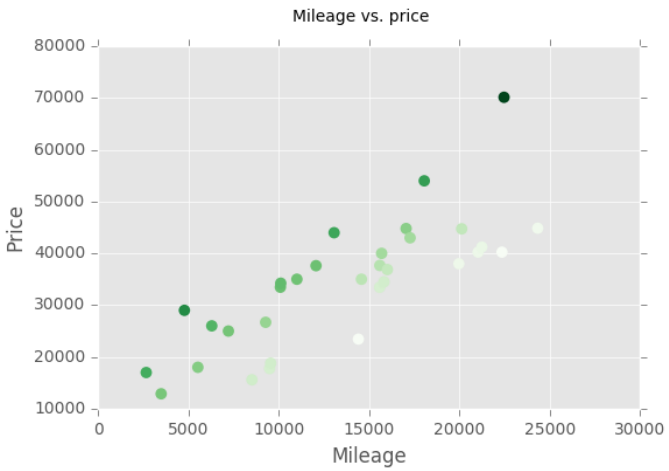


Figure (6) Green colour mapping outlier using Mahalanobis distance method

c) Feature Selection: Predictive performance will be reduced if too many features are used. From the dataset, there are 31 attributes variables, we shall pick only highly predictive variables like model, make, year, mileage, price, and doors. This feature selection method will improve the predictive performance and computation cost. Extracted the 3-dementainal features by using DictVectorizer from sklearn.feature_extraction function. The below sample dataset collected around 206 records with different features for used cars as shown in the table.

Sample snapshot for London Cars data

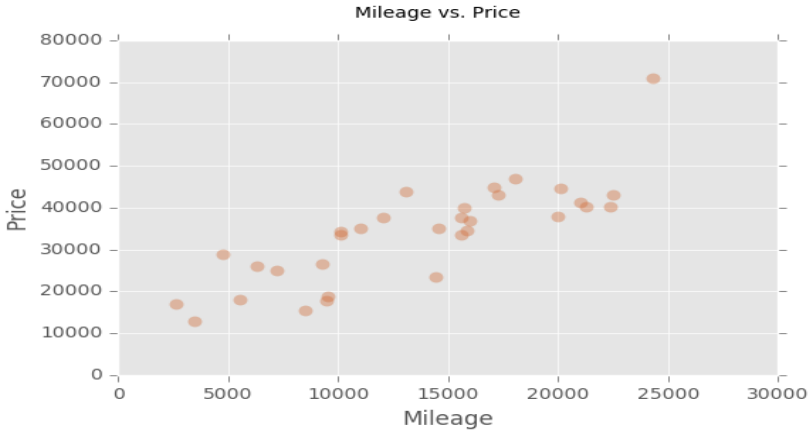
Class label	Make	Model	Year	Mileage	Price	Doors
1	Jeep	Grand Cherokee Laredo	2009	17165	48900	4
2	Lexus	ES 350	2009	17168	25093	4
3	Honda	CR-V EXL	2009	17173	17986	4
4	Volkswagen	Tiguan SEL	2009	17185	32500	4
5	Nissan	Altima S	2009	17186	46991	4

4. Performing the Analysis

Primarily, there are 206 rows were collected and after applying data wrangling techniques, we kept only the 5 attributes and removed all the uncertainty data so finally we have only 33 records for our analysis.

4.1 Correlation Analysis:

The correlation coefficient can be used to find the relation between Price and Mileage in kilometres. Here, identified the Mileage and Price are two variables to perform the correlation coefficient analysis. The first step in observing the relationship between Mileage and Price variables is to draw the scatter plot and verify the linearity. It is not possible to do correlation coefficient analysis if the relationship is not linear. To explore the dataset to best fit there are two different ways to analyse linearity. Spearmen computes on rank so it depicts monotonic relations and Pearson is true values so it depicts linear relationships. The range of coefficient values should be between -1 and +1. Correlation coefficient values show as 0.858870967742 which means there is very strong relationship between two variables.



Figure(7) car price decreases with increase in mileage.

4.2 ***Random Forest:*** We used this method to do the predictions based on severable independent variables. Here make, mileage and MPG (mileage per gallon) are response variables and price is predicter variable. I have used the MATLAB to apply the random forest method to predict the results. The accuracy value obtained was 84.4%.

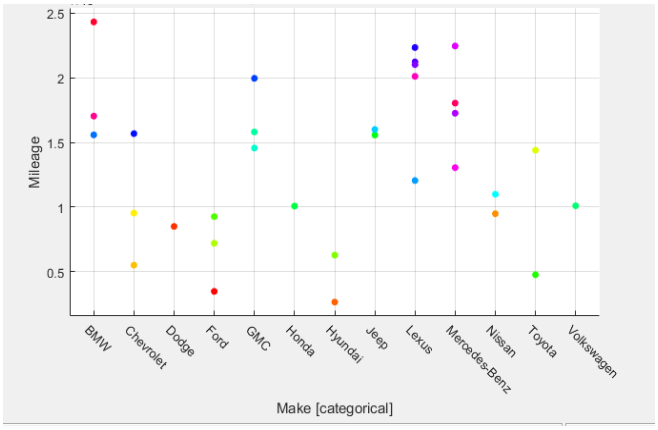


Figure (8) car price prediction for make and mileage

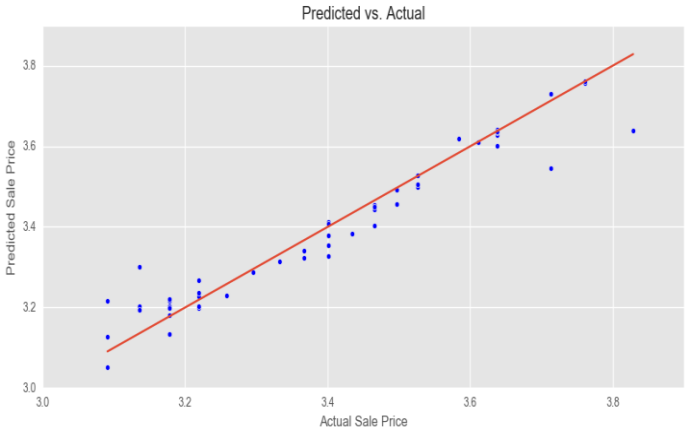


Figure (9) predicted price vs Acual price

4.3 *Regression Analysis:*

Regression analysis is one of the most important statistical method for estimating the relationship between two variables which are price predicted variable and mileage is response variable. In the linear regression model, it will only consider numerical data, in case of any different categorical variables we need to convert them into numerical variables. Here, price is dependent variable(y) and mileage(x) is independent variable. Ordinary least-squares(OLS) regression was used to model a price predicted variable, to estimate the future price, we use the formula y =m*x + b. We first use the scipy library to apply the linear regression to evaluate the values of slope(m), intercept(b) to calculate the car valuation. In figure below, black scattered line represents the y = m * x + b function for each points of values in x-axis. The model which we fitted using polyval function, figure (11) related to 3rd order polynomial and figure (12) belongs to 5th order polynomial. The estimated regression function shows third order polynomial is better fitted data comparing to fifth order polynomial, it looks like over fitted data. Based on R squared value we can determine the model how well it fitted. In our case R-squared value is 0.858, so we can conclude that **85.8%** of the discrepancy in the data.

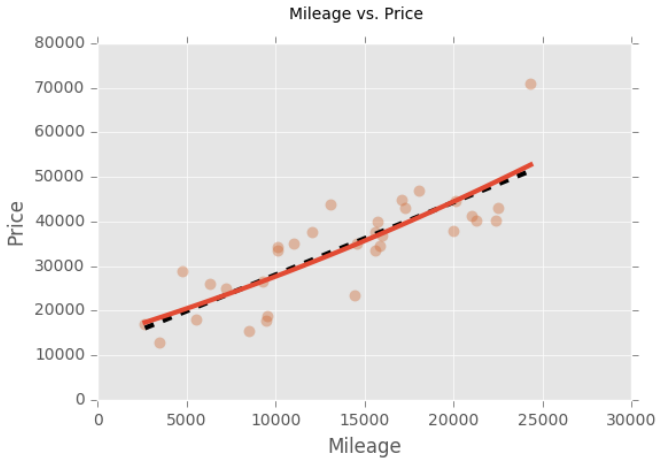


Figure (10) Second order polynomial

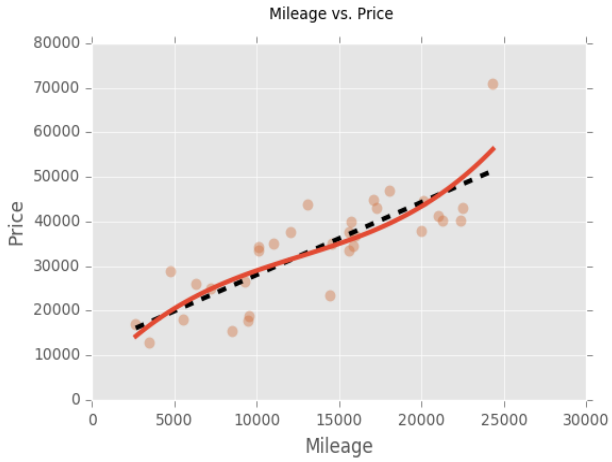


Figure (11) Third order polynomial

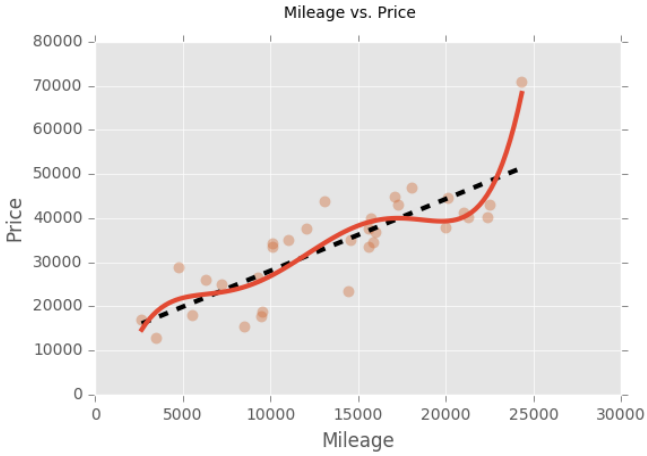


Figure (12) Fifth order polynomial

4.4 *Cross-validation:*

We can measure the exactness of our model by using K-fold cross-validation and validate the output. To generate the sampling indices, I chose to run 5 iterations. We observe the results and pattern in the below scatter plot which shows that the data is not over fitting and compare the estimated values against actual values, the difference always close to zero. Here r_value is good for each run so our regression model fitted correctly. In the test set each data point should have appeared only once but in the case of training set it should be k-1 times.

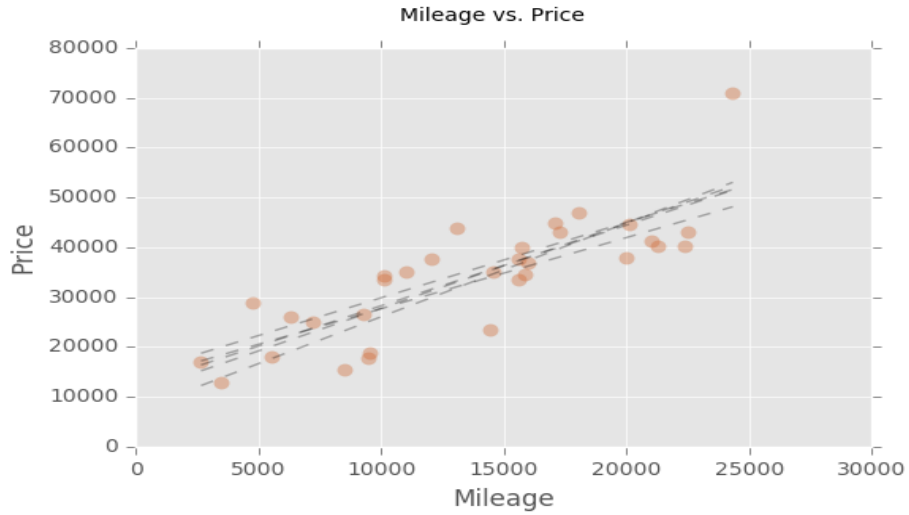


Figure (13) Scatter plot and the overlaid regression lines

4.5 Clustering Analysis:

In this analysis phase, we have used scikit-learn functions for clustering observations. I have specified the number of clusters to be K=3 and K=5 and found different values we get for each time we run the K-means clustering algorithm. To evaluate the results correctly PCA method can be used for the data points and visualize them. The below plot shows results on PC's as three distinct clusters such as blue colour indicates the price, green indicates mileage and red colour indicates year of the manufacture. In the observation, we can conclude that each cluster is separately classified.

Figure (14) K-means clustering result for K=3 (with seeds =10)

Figure (15) K-means clustering result for K=5(with seeds =10)

4.6 Dimension reduction with PCA:

After investigating on cleaned version of London cars sales data which has 31 columns, we use PCA to build the model to reduce the dimensionality of a dataset by finding the two principle components. Now we try to plot the two axes like mileage and price and identify if there are any dominant pattern for each column data. We notice that column "revolutions per mile" has very high loading (value of 0.995507797244 which is almost equivalent to '1' so we need to start further analysis without this column.

4.7 Multidimensional scaling:

MDS method used to visualize the similarity of individual information of London cars data set. Here multidimensional scaling applied, each red dot represents the car models and manufacturer. Data set contains 31 * 31 columns since the

distance matrix is high dimensional, it is very difficult to interpret the plot. We noticed in Figure (16) and Figure (17) that model e350 and Mercedes-Benz manufacturer is the separated point in our plot.

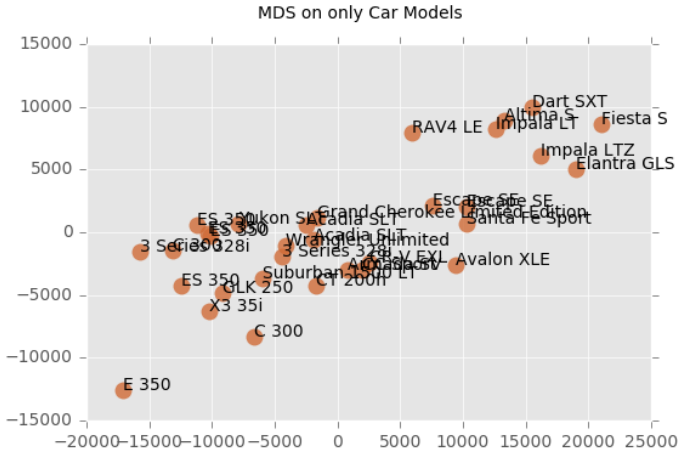


Figure (16) Multi-Dimensional Scaling on Car model

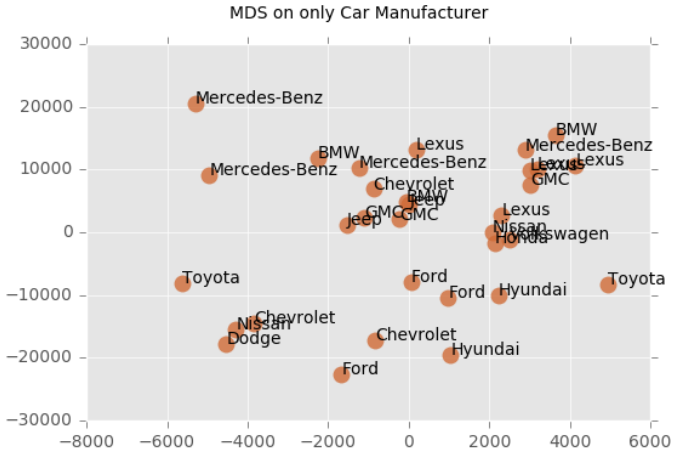


Figure (17) Multi-Dimensional Scaling on Car make

5 *Conclusion and Future Work:*

In this project, different techniques have been used to predict the price of used cars. We are mainly focused on relationship between price and mileage and covered on random forest to check the relation between actual and expected results. In the linear regression method, the Pearson coefficient is 85.8% so our regression model is good fit. In the random forest method, we got the accuracy of classification is 84.4%. All possible experiments have been performed with cross-validation techniques of 5 folds. The price of car, year, make, model, mileage are the main attributes which have been used to predict the price of the car. We found here, linear regression is better predictions comparing to other approaches.

The data what we have collected is very limited and it is historical data. Using the linear regression model, it is difficult to make conclusion that predicted future car price, as it can't give accurate forecasting prices. For building regression model, we used the past data to do predicting but the purpose of our model to forecast future car prices. In the future for better prediction results, need to have high volume of data and use the other suitable machine learning algorithms (like neural networks) to make the predictions.

6 *Software and Tools used:*

I used Python code for analysis, python includes different packages like NumPy, SciPy, Pandas, Matplotlib, seaborn, matplotlib as these tools were already used in the labs. Scikit-learn library used for performing clustering like K-means clustering, linear regression and for dimensional reduction. IPython Notebook is tool very useful to perform exploratory data analysis and to plot the results. Pandas is a python library which was used in data manipulation and analysis like data cleaning, missing values handling. Matplotlib package library was used for basic plotting and visualising data and for quick results. Scipy library used for estimating the values between two variables, numpy is to make basic data operations. Statsmodels were used to estimate statistical models, and perform statistical tests.

I have also used Tableau, MATLAB as it is easy and quickly to use for creating charts for better visualisations comparing to python code.

7. *Appendix of code and Graphs:*

I can provide python source code which I have used for my analysis and plotting the graphs.

References:

- Decision Trees — scikit-learn 0.18.1 documentation. Available at: http://scikit-learn.org/stable/modules/tree.html .
- Polynomial Regression Examples | STAT 501. Available at: https://onlinecourses.science.psu.edu/stat501/node/325.
- Car depreciation: the cars that hold their value best | Auto Express. Available at: http://www.autoexpress.co.uk/best-cars/87343/car-depreciation-the-cars-that-hold-their-value-best.
- Cross Validation. Available at: https://www.cs.cmu.edu/~schneide/tut5/node42.html.
- Data Driving The Future of Cars: Data Science Innovations in the Automotive Industry | Pivotal. Available at: https://blog.pivotal.io/data-science-pivotal/news/data-driving-the-future-of-cars-data-science-innovations-in-the-automotive-industry.
- Download Fuel Economy Data. Available at: http://fueleconomy.gov/feg/download.shtml
- Journal of Statistics Education, v16n3: Shonda Kuiper. Available at: http://ww2.amstat.org/publications/jse/v16n3/datasets.kuiper.html .
- Python for Data Analysis Part 27: Linear Regression. Available at: http://hamelg.blogspot.co.uk/2015/11/python-for-data-analysis-part-27-linear.html
- Second-hand car sales hit record level in first half - BBC News. Available at: http://www.bbc.co.uk/news/business-37154928
- UCI Machine Learning Repository: Automobile Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Automobile [Acce.
- ŷhat | Detecting Outlier Car Prices on the Web. Available at: http://blog.yhat.com/posts/detecting-outlier-car-prices-on-the-web.html .
- Huang, W. et al., 2015. A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal*, 14(15), pp.1–8. Available at: http://dx.doi.org/10.5334/dsj-2015-015.
- Khaidem, L., Saha, S. & Dey, S.R., 2016. Predicting the direction of stock market prices using random forest. *Applied Mathematical Finance Month*, 0(20), pp.1–20.
- Lundkvist, E., 2014. Decision Tree Classification and Forecasting of Pricing Time Series Data. , (July).
- Peerun, S., Chummun, N.H. & Pudaruth, S., 2015. Predicting the Price of Second-hand Cars using Artificial Neural Networks 1 INTRODUCTION 2 RELATED WORKS. , pp.17–21.
- Pudaruth, S., 2014. Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*, 4(7), pp.753–764.
- Wah, Y.B. et al., 2014. A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal*, 14(7), pp.1994–1999.
- Wah, Y.B., Ismail, N.H. & Fong, S., 2011. Predicting car purchase intent using data mining approach. *Proceedings - 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2011*, 3, pp.1994–1999.