Automotive Gender Consumption Differences

Atharva Joshi

Submitted for the Degree of Master of Science in

Data Science and Analytics

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

June 16, 2014

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 12416

Student Name: Atharva Joshi

Date of Submission: 05/12/2021

Signature: Atharva Joshi

Abstract

Automobile industry plays crucial role in counties GDP. Recent studies are mainly focused on to identify which factors generally drive a person to buy automobile and on car features affects the sales of particular car or company. Automobile Companies are also interested in knowing their consumer and what they prefer, since male and female have different qualities and demands when it comes to purchasing a car. Main purpose of this project is to identify car sales difference between women and men, and find pattern in cars features and study about male and female gender preferences. How cars are gendered accordingly to preferences and what factors differentiate them. This paper will take considerable factors which can differentiate between male and female car purchasing due to car features such as price and power etc.

This study focuses on private car registered data in UK, where female and male have different number of car sales. This paper explains how to extract the necessary data from internet for required processing and analysing. Moreover, we will look how various regression and algorithm techniques gives predictions for car model to see whether it can be classified into gender or not.

According to conclusion our paper suggests car models can be categorized into male and female, and how car features play important role in finding out the difference between male and female preferred cars. Moreover, how these factors can change according to the data available. It also gives insight about how we can predict gender for specific car and which factors are playing role to specify gender of car. Some categorized gender models are also given in example.

This study can also help automobile sector to grow and understand differences between male and female gender. There are few studies which are involved in finding this difference to understand the differences and attract people, build car model according to necessities of particular gender.
Language: Python, Jupyter notebook.

Contents

1 Introduction

Automotive Industry has evolved in these upcoming years with new variations and features in Cars. [1] With these and new models in market, Automobiles have always made success in market in spite of economic issues. They have proven to be most important contributing factors in GDP of countries. This Industry plays very crucial role in UK Sales because 'Sales' is an important element in Financial Markets. As
Consumer demands for Automobiles, Sales can express the changes in economy accordingly. Car manufacturers are focused on attracting customers with their new car models and features. They are also focused on making developments in car according to people requirements, which leads us to gender requirements.

People tend to buy car to make their day-to-day commute easy, flexibility of going anywhere, save time etc. Moreover, driving is perceived to be more pleasurable than public transport. But whenever people are buying cars, there are lot of factors which plays huge role such as:
Their age, gender, preferences, style and car features. Car Manufacturers are always interested in finding patterns in Sales of cars. There are a lot of preferences when it comes to Cars according to features such as Engine, Power, Fuel, Body type, Car company, Price, Maintenance, Safety features. [2] Women and men have distinct needs, interests, and different psychological influence that drives their decision to buy an automobile for everyday activities that matches their needs criteria, which are automotive characteristics, style, and size. Women and men have chosen various automobile characteristics; therefore, it is expected that there would be a considerable difference in car class, style, characteristics, and size when it comes to purchasing a car.

[2] The findings of this study are valuable for science in identifying numerous elements influencing women and men's automobile choices, as well as providing information for future scientific growth and psychological area where we can study differences in men and women. Also, this study might help women and men who are thinking about buying a varied car standard selection. Because of many factors that affect women and men in selecting different types of vehicles before buying a car, the research findings are also valuable for car manufacturers or vehicle dealers. This project and idea will be valuable for vehicle manufacturers to improve their quality kind of automobile that consumers like, so that they can increase sales volume in the upcoming years. This study could also help vehicle dealers figure out how to persuade customers in a non-boring, yet effective, and well-targeted way.

In this project our main goal is to analyse, how UK car consumption has changed over period respect to gender, furthermore we will examine how different genders play role in cars sales. Can we use gender to identify which car they will buy? Can cars be categorized into Male or Female preferred by comparing the sales

value between them? We can find the Sales pattern for gender, so we can also find the difference of preferred car model or company.

We can use different machine learning techniques to achieve this task. Classification and regression techniques such as linear, logistic, ridge, KNN, random forest etc. This machine learning techniques are used to fin relation between various data to predict or analyse the trend or pattern. So, we will be using these techniques to compare our data's numerical value and find the pattern for male and female and their preference in car models or companies. Hence, we will be using different regression models for training the data and making prediction on car as gendered car. We will be using python language with scikit-learn library to apply all inbuilt regression and classification models. All the code will be done in Jupyter notebook.

2 Background Information

Modern articles suggests that there is difference between car preference for male and female. According [3] to SMMT driving industry there is spike in female car sales in a decade which is 21%. This was analysed and it was observed that most of these cars with High Sales were for Family vehicles with moderate vehicle features, High Safety, low Maintenance (Ford Fiesta, Vauxhall Corsa). We have Data available for 1994 to 2013 which shows most of the vehicles, from last decade all cars were of fuel or gasoline. Very few cars which runs on Batteries were purchased. Compared to women, men are still dominating overall in car sales. Men has 10% increase in new car purchase, then last year. [2] "As an example, a similar field study carried out by Manski and Sherman (1980), multinomial logit models have been developed to investigate how many cars have purchased and how the way to select such purchased car".

2.1 Factors influencing Sales

As discussed, earlier age and gender plays very crucial role in car purchases. People with lesser age are more focused on new models, car company, car type, power, whereas middle aged people tend to pick well known company over car, safety features of the car, capacity. Moreover, according to our data we need to categorize male and female preferences. Therefore, we will be considering factors to categorize gender by car features such as Engine capacity, Power, Price, Company, Car type, Fuel type which has the huge effect in car purchasing.

2.1.1 Male and Female Preferences:

[2] The article identifies and limits behavioural factors impacting car selection to women and men before purchasing a vehicle, and presents a comprehensive picture of behavioural differences prior to purchasing a vehicle. Both men and women have some factors, such as car characteristics, class, and size, price, power which influence car choosing requirements. [4] According to analysis made by Automotive research & analysis, it showed most of the High-end vehicles such as (BMW, Mercedes, Land-Rover, Audi) registered as Male. Moreover, it has shown that men usually tend to buy vehicles based on its Power, Engine CC, Company, Style and Power. On the other Hand, women were registered for Mini Vehicles Such as (Ford, Honda, Mini). Women usually go for comfort of the vehicle, vehicle capacity, price and maintenance value.

[5] Team of car manufacturers did theoretical research on car preferences according to gender. It helped them understand what type of Cars different gender prefers. This information assisted them in analysing pattern in Sales value of cars according to genders. Companies such as Audi, BMW, IBM are developing such machine learning models to identify patterns in car sales and gender preferences to gain information on which cars do gender prefer. In recent survey, [6] it was shown that men and women have different preferences when it comes to buying a Car. It was also found that men usually prefer cars with power and Engine whereas for

women it showed that they usually prefer safety features, price, car capacity, and maintenance. That's why male gender prefers high sports cars such as BMW, Ferrari, Porsche, Nissan etc. compared to female.

2.1.2 Sales analysis:

The main purpose of [7]'"Auto Car Sales Prediction: A Statistical Study Using Functional Data Analysis and Time Series," paper was to conduct analysis on which factors affect the car sales value, and which factors(variables) makes huge impact on car sales. This study proved we can predict sales value for cars; however, we cannot predict it upcoming future because of other phenomena's which cannot be assured. Although we can get good prediction for present data, data collected for car analysis can be used to predict Sales value for specific cars and prediction of cars models. Linear regression and other regression can be used to predict effects of features(variables) on outputs.

Similar research was done by [1] "Abolfazl (Kouros) Mohammadian from the University of Illinois, Chicago, USA, 2004". This type of research is a descriptive study, and was carried out by doing testing hypotheses. This study is a survey aimed at testing the hypothesis about the various factors that affect women and men choosing the type of vehicle and brands like (Ford, Mitsubishi, BMW, Mercedes Benz). In this paper they proved that Males and Females have different preferences over vehicle. According to their study Woman does not require luxury vehicles but economical vehicles. Similarly, men car buyers do not consider car characteristics detailed but its efficiency, but always consider car class in the car purchase selection process, man sometimes get influenced by class and style and size.

[7] this research paper carried out car selection preference, through information collected through both genders. In this paper, they have used hypothesis and maximum likelihood approach to find which gender have selected which specific car type. They also used regression method to find patterns, significant outputs and analysis on purchasing factors. They used to perform OLS model for this research to find R2 and estimated r2 squared using linear regression to find car type. When it comes to UK car consumption according to gender, we have to consider factors according to country.

So according to previous papers and preferences for car features we will use car features such as price, fuel type, power, engine cc, manufacturer, transmission for predicting gender of the car. Then we will also try to find pattern for male and female preferred cars. We will also be looking at how these factors play various roles in car sales value. Hence, by using regression and classification problems we will look which car can be considered as male and female according to the factors. We will try different models to see the model fitting and accuracy and will analyse results.

3 Methodology

For performing regression and classification of data we can use multiple algorithms which are provided by 'scikit-learn' library. Our main goal is to study about relation in data and which factors affect the sales for specific gender. First, we will classify the dependent and independent variable for our data. We have car features, and its sales value with particular gender. So, we will observe the relation between variables to perform algorithms which are suitable for data.
We will apply various regression algorithms(models), where it actually gives significant output for following data. Regression is mostly used in prediction, weather forecasting and sales value. We can estimate the relationship between dependent and independent variables. Then we can check the normal distribution of each variable to see if data is normally distributed to process it further.

3.1 Analysis types:

We will see different analysis types for our data. As we have both multiple classes as well as gender, more preferred method will be to use logistic regression or classification. However, since we have to predict different class more efficient and better method to get accuracy will be Classification technique.

3.2 Regression

As we know Regression is used to obtain the relation of variables in data, it can be used to obtain sales value for specific car as well. With help of features such as Price, Fuel, Engine etc. we can find the correlation of variables with each other. By this, we will be able to get their relation with each other. It is most common method used in machine learning for finding relation between variables. It can be used to identify which categorical model belongs to data points or output.
Advantage of regression is that, we can use it crunch the numerical data to make better decision on future. Regression analysis is used for predictive analysis, operation efficiency and get new insights from data.

3.3 Classification:

As per [8] Classification modelling is a task of mapping function from variables or features to discrete labels. Mapping function is used to predict certain type of class or given labels. A classification problem requires that examples be classified into one of two or more classes. Hence, when we are predicting both Gender for car or car models our labels are going to be categorical and multiple. So, we will need classification type of problem.

[8] A classification can have real-valued or discrete input values. We can have two classes called as two-class or binary classification problem and more than two

classes called as multi-class classification problem. Multiple classes are called a multi-label classification problem.

3.3.1 Logistic Regression:

Definition: [9] Logistic Regression is method used when want to predict categorical variables instead of binary variables. It can be performed on continuous variable as well. It uses log odds ratio (hypothesis) for the specific outcome.
E.g., coin toss will be head (0) or tail (1).
[10] This method is mostly appropriate for non-normally distributed data or when data have unequal covariance.
Formula:

$$P = \frac{e^{a+bX}}{1 + e^{-(a+bX)}}$$

[11] Logistic Regression

P – is probability of 1 of Dependent Variable.
e - is the base of Natural logarithm (2.718)
a, b – are parameters of models (features)
X – is used to find changes in a, b when b is 0 (P=1) as b adjusts changing value of X.

[12] Logistic method uses probability and odds estimation method (Sigmoid function).
E.g.: The probability that an event will occur is the fraction of times you expect that event to see in trials. Probability of heads to tails occurring in total 3 tries.
Output measured from Hypothesis is the estimated probability for that event to be true. This can be used to show how predicted value and actual values are close.

Maximum likelihood estimation:

Logistic regression uses inbuilt MLE, [13] "Maximizing the likelihood function determines the parameters that are most likely to produce the observed data. From a statistical point of view, MLE sets the mean and variance as parameters in determining the specific parametric values for a given model."

Pros [14]**:**

- Easy to implement and interpret, efficient.
- We can extend it for multiple classes to view predictions.
- Provides how appropriate coefficient size is for predictor.
- Good fit for simpler data which is linearly separable.

Cons [14]**:**

- Less number of features may lead to overfitting the model.
- Non-linear problems cannot be solved by this method.
- It requires multicollinearity between independent variable.
- Hard to obtain complex relationship using logistic.

Since we have to predict the models of the car, we have multiple class to predict. So, we will have to use multi classification techniques for this using logistic regression.
E.g.: One-vs-One, One-vs-All

Multicollinearity: [15] It is term when we see independent variables are correlated with each other in some way. So, in regression it may give us some high error and bias rate while training the model. Linear regression is not fitted for this type of problem so we will use other regression and classification problems.

3.3.2 Ridge Regression:

[16] Ridge regression is used when our data has multicollinearity. It is most preferred when we have multiple regression data available, and it is used when we notice more predictor variables in the data which defines dependent variable. Used when independent variables are highly correlated. Moreover, it adds small amount of bias to estimator to reduce standard error. This is great method, to avoid overfitting. This method can also be used for multi class classification as it has inbuilt features with it.
Formula:

$$Rss + \alpha\|w\|^2 = Rss + \alpha \sum_{j=0}^{p-1} w[j]^2$$

[16] Ridge Regression

α – regularization parameter (tuning)
w – weight
j – features

Pros:

- Trades variance for bias if co-linearity in data.
- Prevents overfitting in data.
- Performs well in large multivariate data.

Cons:

- Increases Bias for data.
- Model interpretability becomes low.
- Unable to perform feature selection for data.

3.3.3 K nearest Neighbour:

[17] KNN can be used for both classification and regression predictive problems. The main advantage of this algorithm is it specifically used for multi class

classification problems. Hence if Data consists of multiple labels or binary labels it is preferred method over other models. It is supervised machine learning technique used in regression and classification problems.
[17] It works on principle where the class is classified by its nearest neighbours and categorized accordingly.
If our K value is 2 then it will find distance between 2 nearest neighbours and will classify it into that group accordingly. We can use various distance methods like (Euclidean, Manhattan) to calculate distance between points. [18] Euclidean distance finds the linear distance between n specified number of points.
Formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

[18] Euclidean Distance

x, y – Sample points or values in data

Pros [19]

- Very easy to understand and easy to implement. It reads whole dataset to classify new point.
- It does not make any assumptions.
- Can be used for classification and regression methods.
- Easy for multi class classification.
- Has hyper parameter to select the neighbouring elements to catch.

Cons [19]

- Can affect the speed of algorithm if data is Huge.
- Faces issue of cures of dimensionality with huge number of input variables.
- Does not deal with missing values.
- It is sensitive to outliers.

3.4 Classifiers: ONE VS ALL, ONE VS ONE

3.4.1 One vs One:

[20] This is another method used for multi class classification for binary classification of data. As we have labels for Car models and company, we have data which is in string format. Computer only knows binary language. So, after encoding it with label encoder Logistic regression has this feature where we compare class with other class one by one as name suggests.
It follows

number of classes X (number of classes – 1)/2

Which gives total number of binary classifiers in data. With this each binary classification model, can predict class label. If it predicts class members accurately, class label with most sum score is taken as class label for that observation. However, this method is mostly used in support vector machines.

3.4.2 One vs All:

[20] This is the same method as One vs One, however instead of comparing one class with all other class. It is a method where we fit one class per classifier. We have advantage of interpretability by using this method. This method also uses binary classification and is most used method in logistic regression for multi classification.

3.5 Terms

We will use calculation terms for measuring models' evaluation, error rate and accuracy for its predicting power. So, these terms are used to see how good our model is for the data, and with these values we can change the methods.

3.5.1 MSE: Mean Squared Error

[21] It is used to measure the average of error squares. It is defined by average squared distance between predicted to true values. It is risk function, corresponding to the expected values squared error loss.
The MSE is the second moment of error (around the origin), and it takes into account both the estimator's variance and bias.
If MSE is close to zero the better and if close to 1 worst case.
Formula:

$$MSE = \frac{1}{N}\sum_{i_1=1}^{N}(y_i - \hat{y}_i)^2$$

[21] MSE

$\hat{y}$ – Predicted value of y
y – Actual value of y
N – Number of samples

3.5.2 MAE: Mean Absolute Error

[22] It is used to measure the average of the absolute difference between the true values and predicted values. It also measures the average of the residuals in the dataset.
Formula:

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

[22] MAE

$\hat{y}$ – Predicted value of y
y – Actual value of y
N – Number of samples

3.5.3 RMSE: Root Mean Squared Error

It is used to measured Standard deviation of residuals. Root of MSE.
Formula:

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

[22] RMSE

$\hat{y}$ – Predicted value of y
y – Actual value of y
N – Number of samples

3.5.4 R2 square:

[23] "R squared is used to measure how close our data is fitted to regression line. Also known as coefficient of multiple determination. It can be defined as percentage of response variable variation explained by linear model".
0 means our model is poor.
1 means model is good and being able to predict accurate values.
Formula:

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

[23] R2 Square

$\hat{y}$ – Predicted value of y
$\bar{y}$ – Mean value of y
y – Actual value of y

In our model we will check R2 value to see how good is our model. More the value close to 1 better prediction.

3.5.5 Accuracy Score:

It is a function which matches set of labels predicted exactly to set of labels for that sample. It is used to predict the accuracy of the model for classification models. This function is given in scikit-learn library (accuracy_score).

3.5.6 Precision:

[24] Precision is the fraction of true positive samples for which they were classified as positive in model." Number of true positives divided by number of false positive plus true positive."

$$precision = \frac{tp}{tp + fp}$$

[24] Precision

tp – True positive samples classified
fp – False positive samples classified

3.5.7 Recall:

[24] Recall is fraction on examples classified as positive, among total positive. "Number of true positive upon number of true positive plus false negative".

$$recall = \frac{tp}{tp + fn}$$

[24] Recall

tp – True positive classified samples
fn – False negative classified samples

3.5.8 F-Score:

[24] It is defined by "weighted average of precision and recall values. It takes into account false positive and negative values". F1 is considered more useful than accuracy in case we have uneven class distribution. This methos is used to seek the balance between precision and recall when there is uneven distribution. We can find which classes are negatively classified when f1 score are 0. So, if its zero for all the samples, that means our model in not perfect. Closer the value of f1 to 1 better the results.
Formula:

$$F_1 = 2 \times \frac{PR\,.\,RE}{PR + RE} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

[24] F-Score

tp – True positive classified samples
fp – False positive samples classified
fn – False negative classified samples

These terms are used to check efficiency, error and accuracy of the model to predict certain outcome. Therefore, we will be observing these terms for classification and regression problems. So, in regression we will observe these terms with Roc curve, confusion matrix, scatter plots.

4 DVLA Data Set

DVLA Data consists of privately owned cars licensed in Great Britain for each year (1994-2013). Their sales value is categorized into 2 genders (Male, Female). We have Data where, Total sales for that particular car model is given for each year from 1994-2013. Sales value is also given for specific gender, where it is divided into male, female, neutral(cross). From, this Data we can observe, we have car model and their sales value specified for each gender.

However, our data is not cleaned, so we will have to take a look at data and clean it according to the protocols. We can observe all the (Make/Model) table have some missing car model for specific company. So, we have some missing data present which we need to eliminate. On the other hand, our data also has sales for 1 to 1000 cars, which are poor values according to 1000k Sales.

Some of the cars which we are extracting are not present on website hence, while comparing both the data we will only do analysis on cars data which we have available.

So, our necessary steps will be:

- Clean DVLA Data
- Keeping only necessary Data

Data shows us that Ford models Focus and Fiesta had more total sales in upcoming years. Ford Focus has been registered to high number of Male and Fiesta has been registered for High number of Females. Hence, we can say that we have data where we can classify which car is preferred by Males and Females.

5 Web Scraping

Definition: [25] Web Scraping is the process where we use bots to extract data or valuable information from website. It can be used to extract HTML contents of website as well as Data stored in Database. Scraping tools are bots designed (Programmed) to go through websites and extract data (information). This tool will load the web URLs specified by user and render the entire website.
Automated web scrapper can be helpful while extracting the data because we only have to write the code once to extract the data and then you can reuse that code to extract different websites as many times you want.

[25] Web scraping is also illegal in some cases, when you use it to extract Private confidential Data from Company or Website. Some, sites allow scraper to access the data but some sites may consider it illegal. The site which we will be using to scrap data is safe and allows scraper to access its data. Used for various purposes such as sentiment analysis, personal project, pricing or loading new information

Pros [26]**:**

- It is Cost effective as it is digital technique used to extract information, but also depends on amount of data.
- Easy implementation and can be used for another web pages as well.
- Data extracted is accurate, and helps in effective data management.

Cons [26]**:**

- While accessing sites, sending too many requests can get our IP banned. So, we have to invest in proxies.

5.1 Tool Used in this Project: Beautiful Soup

[27] Beautiful Soup is a python library used for parsing structured data. It is used to pull data out from HTML and XML contents from website in hierarchical manner.
There are few functions which are very useful for extracting data:

- Prettify(): [27] This function is used to prettify the data as function name tells us. It prints data in understandable according to the HTML elements and contents .
- find(): [27] This function can be used to find first instance of that object or tag and return the content inside it.
- find_all(): [27] This function works same as find() but it will return all the elements or tags from webpage.
- Regex: [28] Regex is very useful for filtering out the data. To filter out data we can use match function of regex to match the data we need and discard the rest. We will be using regex functions by importing regex library.

Site for Extraction:

CARS-DATA: We will be using https://www.cars-data.com/ for extracting the data because site is easy to understand, it allows server requests and site is open for public. This site has almost all car models we need, moreover this site has all car

details which are required for Analysis. This site also provides with safety features so in-case of future addition to data we can extract it by adding extension.

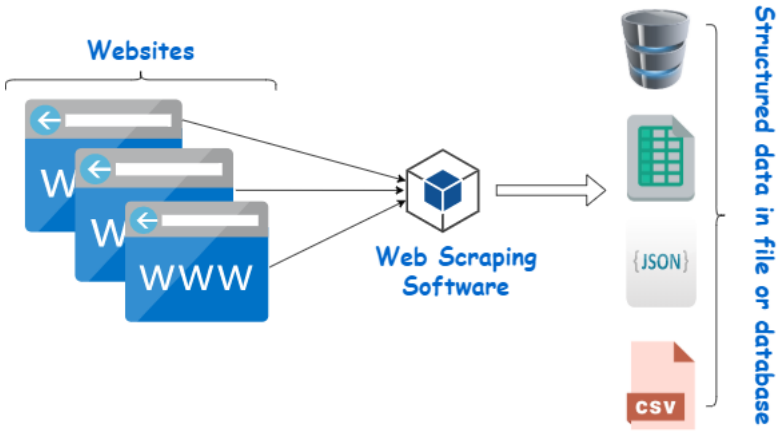


Figure 1

5.2 Web Scrap Code Explanation:

1. **Header Agent to access Data Site:**
   - In weburl we are passing Link of cars-data manufactures site where all car companies and its cars details for models are present in web page which are embedded.
   - To access the site, we used User-Agent to access the site contents. It retrieves and outputs the web content for user through web technology.

2. **Accessing Links:**
   - Now we will fetch the links for car companies which are present on webpage.
   - CarManufact is variable used to fetch class 'col-2-center' and container 'div' where links of each car companies are present by using find_all() method.
   - We pass CarManufact in for loop, passing 'a' and 'href' in findall parameters gives us link which are present in that container.
   - Finally, we append links in one manufacturers list.
   - As we can see we will get outputs for how many manufacturers are present and links for those in a list.

3. **Accessing models links:**
   - After getting all links of Car Manufacturers we will access those links we extracted to access next page.
   - From them to access all car models links which are present on next page. That means we will iterate link through link to fetch all required links present in 'div' container with class 'col-4'.

- Using regex we are ignoring unwanted links with help of re.compile method where we are specifying extract only cars-data.com links.
- These extracted links are then appended in models list.

4. **Model different versions:**
   - After getting all links for models we will need to fetch links for model versions.
   - We apply same code as described above only this time we pass regex function to extract links with digits as all car models with different version have links ending with numbers.
   - Also, we import sleep in between so we will send http requests with some delay.
   - We will store all these car models link in one list.

5. **Car models Data:**
   - Performing same steps, we will iterate through all 15535 models link, to fetch car details which are present in 'td' and class 'col-6 grey'.
   - This step will extract all information which is present in that class.
   - We are extracting data in 1000 of cars as we don't want to get our IP address blocked by user for sending to many server requests.
   - We can see we have list of details for every car now and now we will save these details to csv.

5.3 Code Chart for Web Scrap Data Extraction:

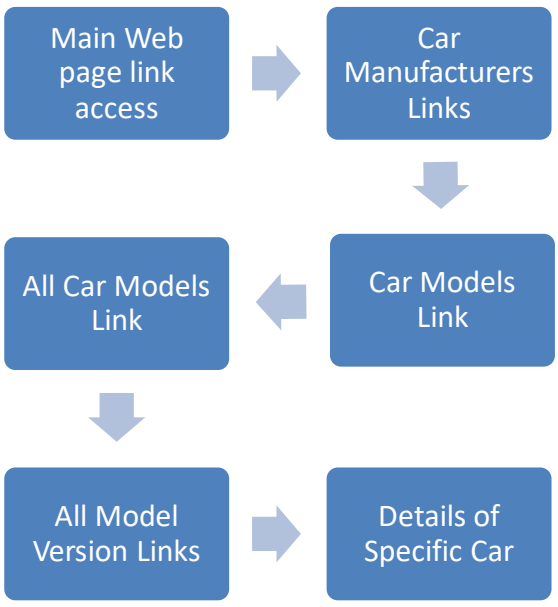
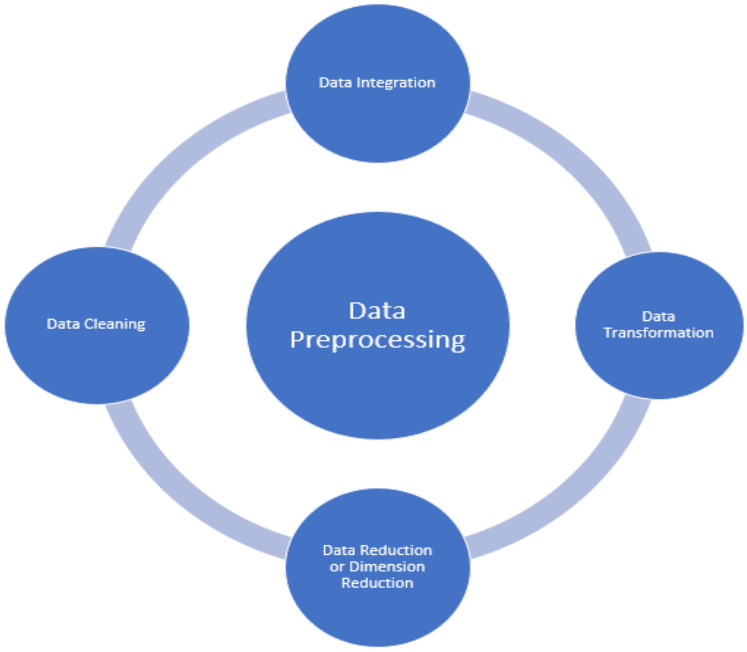


Figure 2

6 Data Pre-processing

Pre-processing:

[29] To analyse the data, computer needs to understand the data, so we need to bring the data into format where it will be easy to process.



[29] **Figure 3: Data Pre-processing Steps**

Data Cleaning:

[30] Data Cleaning: Data Cleaning is a process used for fixing incorrect data or removing these data. Data consists of duplicate, incorrect format, missing values. So, this is the step where we clean Data to bring it into one format so that algorithm will work on this efficiently. Cleaning Data steps may vary from dataset to dataset. Hence it is necessary that you look do data visualization to clean the data.

According to the data we have available we have to clean data to process it:

6.1 Outliers:

[29] An outlier is the observation or sample in data which is very far from the other observation. To remove outliers is necessary because it can cause low accuracy on model and affects statistical analysis.

In our Data value for Sales (Male, Female, Unknown) value are given in range of 1,000k up to 1. So, we are going to exclude data where sales value is smaller than 1000 or 5000. We have only 1 sale value for more

cars compared to ratio 2000k. It can affect the analysis so we will be excluding cars which has sales value lesser than 1000.

6.2 Missing Values:

[29] Missing values also affect statistical power of data. It treats it like lost data and can give bias in model. We need to be sure to replace missing values.

Our data have Cars in which companies' car model is not given. We cannot replace it with NAN or any other value because it will be considered as wrong data classification as we do not know which car model was actually sold. So, just ignoring those cars with missing model name will be satisfactory in our case.

6.3 Extracted Data Cleaning and Merging:

For Data we have extracted we will be considering only those factors which actually affects the male and female genders car selection preferences. Our Data has lot of useless information where it has factors for car CO2 emission, chassis, weight. Excluding these factors won't affect out analysis that much. So will drop those columns with drop method in Dataframe. We still have missing values in Extracted data. So we will replace them str.replace() function or we can use regex as well.
Now we will combine two csv files, extracted data and cleaned data into one by checking if the car and model exist in our data. As mentioned earlier we have some car details which are not necessary. In this process, we have used str comparison and if it matches the string of car of same name on both files it will calculate mean of all numerical variables of that car, and then it will increase the iterating value to next. After going through each model's car link, it will append details in list. Finally, we pass these lists to Data-frame and create CSV file for this data.

6.4 Data Transformation:

[29] Bringing Data into one format is crucial and we can achieve this by changing data type, smoothing and normalization. Transformed data helps humans and computer to understand easily. Properly formatted data helps improve data quality and removing duplicate and incompatible format.
Converting string format and € to integer datatype for price. All the data type will be changed from string to integer or float using astype() or to_numeric() function by pandas. (Except make and model).

6.5 Time complexity: For Web Scrap

We have 3 nested for loops for extracting links which means we have $O(n^3)$ complexity for the algorithm.

Now, we have our data ready for each unique car model for each year (2013-1994) with (mean of same car model features). While comparing we have removed car models data for which we did not have any car features available or the car which was not available on website.

	Year	Make	Model	Make/Model	Price	Transmission	Power	Engine CC	Fuel	Male	Female	Unknown	Total
0	2013	Ford	Focus	Ford Focus	30.619322	5.966102	94.033898	1497.169492	gasoline	814172	422731	56487	1293390
1	2013	Ford	Fiesta	Ford Fiesta	18.532143	5.714286	68.571429	1166.142857	gasoline	554879	631666	54057	1240602
2	2013	Volkswagen	Golf	Volkswagen Golf	31.242154	6.164835	89.461538	1537.406593	gasoline	483216	310604	47563	841383
3	2013	Renault	Clio	Renault Clio	22.100000	5.615385	75.576923	1219.653846	gasoline	241287	312556	28004	581847
4	2013	BMW	320i	BMW 320i	47.848370	6.444444	126.111111	1995.777778	gasoline	408016	115843	29125	552984
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6087	1994	Land-Rover	Defender	Land-Rover Defender	108.747195	7.853659	207.609756	2304.975610	diesel	1012	150	80	1242
6088	1994	Toyota	RAV4	Toyota RAV4	43.548516	1.354839	137.774194	2261.193548	gasoline	670	482	66	1218
6089	1994	Alfa-Romeo	Spider	Alfa-Romeo Spider	55.200000	6.000000	163.500000	2696.500000	gasoline	790	247	81	1118
6090	1994	Honda	Shuttle	Honda Shuttle	30.081000	4.000000	110.000000	2254.000000	gasoline	639	416	49	1104
6091	1994	Mitsubishi	Space	Mitsubishi Space	23.165158	3.947368	82.157895	1817.315789	gasoline	721	251	40	1012

Figure 4

7 Model Implementation

In this section, we will perform feature scaling on data, and will see the correlation between the variables. Moreover, we will observe how data looks like and then we will perform some regression and classification analysis for data.

Independent variable: Independent variables are considered as features of the data. Where we use these variables to predict the output and see if they are correlated to the variables. For our data car features are independent variable.
E.g.: Price, Transmission, Power, Engine CC, Fuel, Male, Female, Make, Model.

Dependent variable: Dependent variables are considered as labels of the data. They are the data we are interested to predict by comparing the independent variable. In our data we have car model or company as dependent variable, which we will be predicting by car features for gender.
E.g.: Gender and Car.

Before we start, we will perform data transformation to bring it in one form.

9	10	11	12
8,14,172	4,22,731	56,487	12,93,390
5,54,879	6,31,666	54,057	12,40,602
4,83,216	3,10,604	47,563	8,41,383
2,41,287	3,12,556	28,004	5,81,847
4,08,016	1,15,843	29,125	5,52,984
...	...	...	...
1,012	150	80	1,242
670	482	66	1,218
790	247	81	1,118
639	416	49	1,104
721	251	40	1,012

Figure 5

As you can see some our numerical data is in string format (Male, Female, Unknown, Total). So, to convert it we will perform:

- Remove the commas by using str.replace() with blank space.
- Change the data type to numeric by using pd.to_numeric() function.

This will bring our Data in correct data type.

7.1 UK Car consumption according to Gender:

In this section, we will observe how each gender play's role in UK car consumption. According to data available we can see male as a dominant factor for car Consumption. However, lets see if it is true by graphical presentation.
For Gender (Male, Female, Unknown):

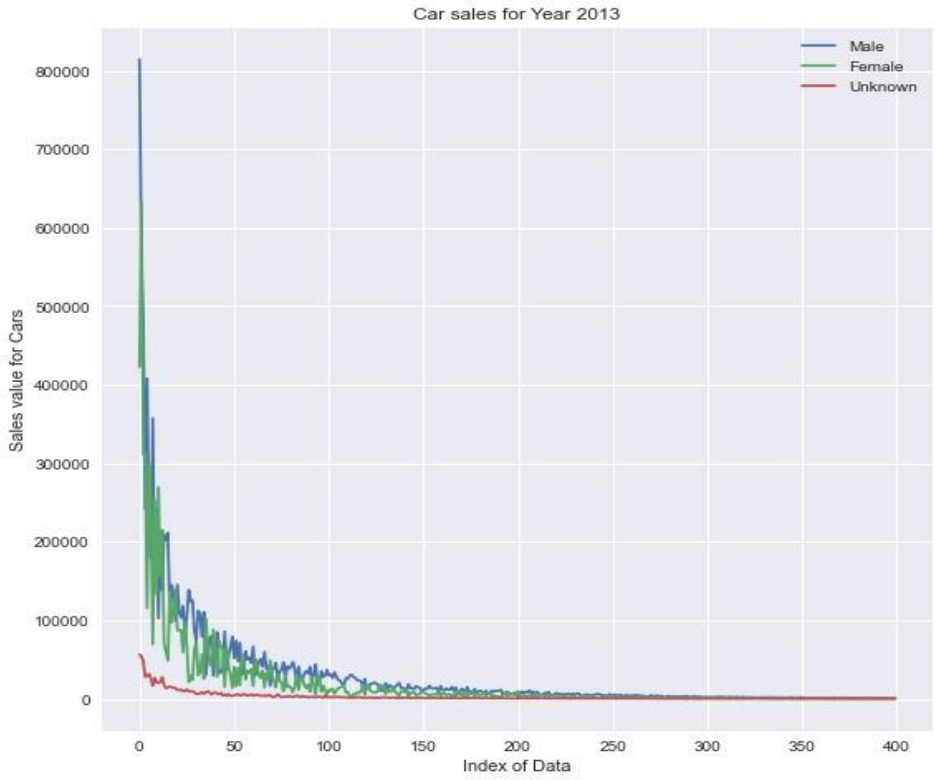


Figure 6

y-axis = Sales value of Car in millions.
x- axis = Index of Cars in data for different cars for year (2013).
This graph represents that which has higher sales according to gender. Whichever has highest sales will be at top. According to the figure we can see that only male and female gender has more car registered in UK compared to Unknown. So, we will focus mostly on female and male gender according to car models. Green colour shows Female has more sales and Blue is for Male.

As seen above, Unknown variable has fewer sales compared to male and female. So, we will create one table for gender (Sex) which specifies which has higher sales according to gender.

- Y variable = Male Sales value per car / Female Sales value per Car
- Storing this Y into one column specifying 1 as male and 0 as female.

So, after converting we will get new column in data which will consist of binary number representing male (1) and female (0).

Car Manufacturers data observation with Features:
This table contains all the Manufacturers we have after removing outliers of sales and car companies which are not present on cars-data website. So, we are comparing only those car companies which has higher sales in cars data.

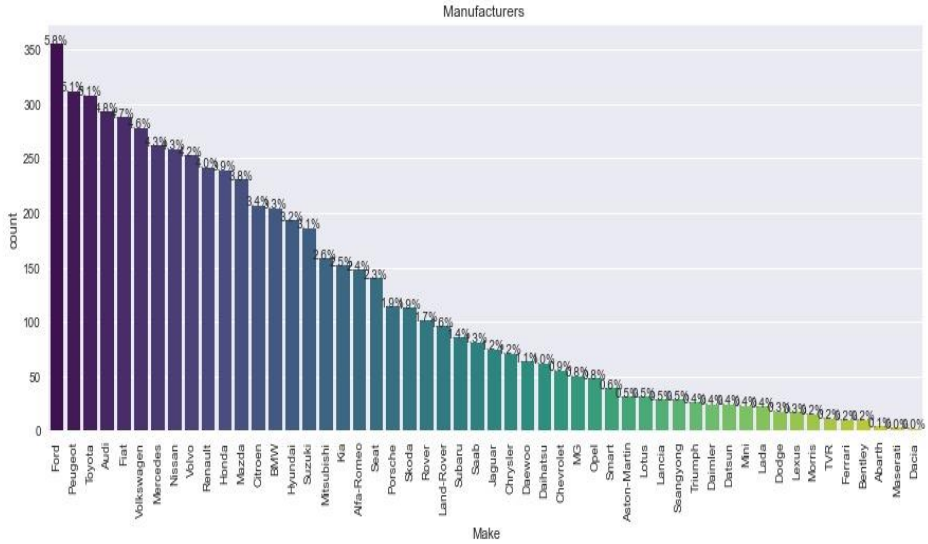


Figure 7

We can observe only some companies had higher sales which were Ford, Peugeot, Volkswagen, Audi, Toyota, Mercedes, some companies had very low sales which were Ferrari, Bentley, Abarth, Ferrari. So, we can compare which car has high end sales and low-end sales according to gender. We will observe this in upcoming report.

Transmission:

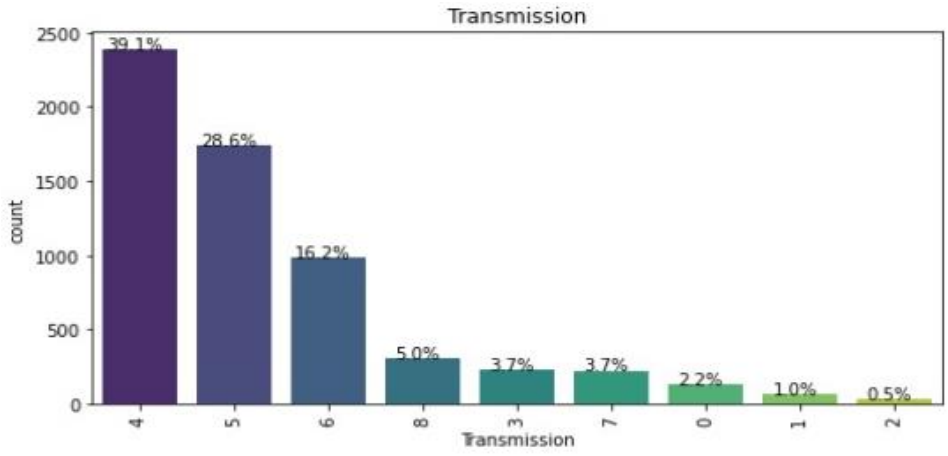


Figure 8

This graph represents transmission count for all car sales. We can observe 4 has highest transmission than others. 0 and 1, 2 has very low count where 0 represents automatic car.

Fuel Data:
In this data we can see fuel type for cars. Almost all cars registered were for gasoline, and diesel. In this graph we can see automatic/fuel/diesel consist of 0.6% where the cars fuel type was not mentioned.
This graph will be helpful for analysis part to find relation between variables.

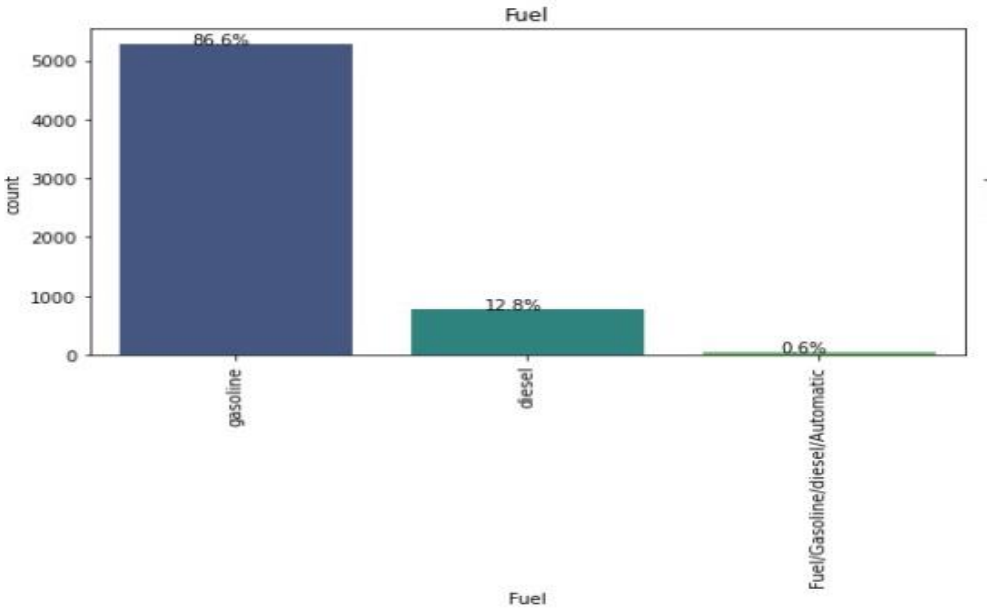


Figure 9

7.2 Feature Scaling:

It is a technique used to transform the values which are in different range to same scale. This technique helps machine in producing better and optimal results for model.
Available Data has very high values for price and Sales where price ranges from (10000 - 40000) and sales from (100000 - 800000). Using standardization outliers in data does not affect the distribution. So, we will apply various feature scaling methods to bring data in one format so machine will be able to understand the data.

Advantages:

- Easy for program to work on model efficiently.
- Converts all data into one format.
- Easy for understanding data with plots.

7.2.1 Normalizing Data:

Data Normalization is a process where we rescale the data between 0 and 1. 1 being the highest observation and 0 being the lowest observation in data.

Minmax Scaler:

We have min max scaler from scikit-learn to perform normalization on data. It scales the data to [0,1] using this scaler method.

Formula:

$$\mathbf{x_{norm}} = \frac{\mathbf{x} - \mathbf{x_{min}}}{\mathbf{X_{max}} - \mathbf{x_{min}}}$$

[31] Minmax Normalization

x – sample value

x-min – minimum sample value in data

x-max – maximum sample value

X(i) is the data which we are normalizing. We use this method on train and fit it on test by same transform method. Hence, we can make predictions using these values on data.

7.2.2 Standardizing the Data:

Standardization is scaling technique where all the values are centred around the mean with a unit standard deviation. It means that the mean of the attribute becomes 0 and the distribution has 1 standard deviation.

Formula:

$$\boldsymbol{X'} = (\boldsymbol{X} - \mu)/\boldsymbol{\sigma}$$

[31] Standard Deviation

μ - Mean of Feature values

$\boldsymbol{\sigma}$ – Standard deviation of feature values

Library: Standard Scaler from Scikit-learn

[31] Rescaling the data values to mean of values and standard deviation of 1. By importing Standard Scaler provided by pre-processing library we can use it to fit and transform the data. E.g.:

X_variable = scaler.fit_transform(X_variable)

As we can see we have describe() function which gives us look at the data to see, after applying standard scaler we got mean and std of 1 for all variables.

	Year	Price	Transmission	Power	Engine CC	Male	Female	Unknown	Total	Sex
count	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6.092000e+03	6092.000000
mean	-3.311519e-14	1.139017e-16	8.101603e-16	-2.451621e-16	2.758882e-16	4.811209e-18	-4.553336e-17	-8.846976e-17	5.543158e+04	0.703381
std	1.000082e+00	1.000082e+00	1.000082e+00	1.000082e+00	1.000082e+00	1.000082e+00	1.000082e+00	1.000082e+00	1.412886e+05	0.456804
min	-1.935380e+00	-7.379775e-01	-3.818791e+00	-1.417985e+00	-2.425033e+00	-4.016199e-01	-3.301096e-01	-4.218378e-01	1.001000e+03	0.000000

Correlation: First we will need to find correlation between all the features we have. So, for that we will use heatmap to see which features are correlated. This will help us in finding collinearity of all variables.

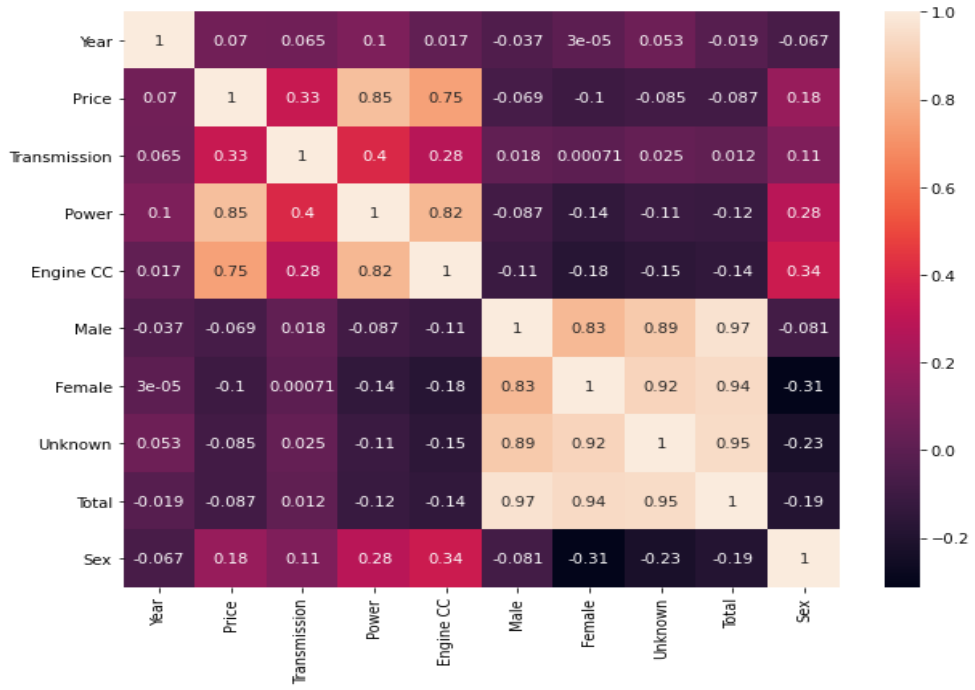


Figure 10 Heatmap for Standardize data

Heatmap observations: Above table gives us correlation between all the features.

- We can notice that there is strong relation between variables Engine CC, Power, Transmission, Price.
- We can see there is no strong correlation between all the variables and male, female and unknown value.
- Sex has some relation between Power, Engine CC and Price. So, we can identify by features which is male and female.

Train Test split:

Train test split is a technique used to estimate the performance on machine learning model. It is used to divide data in two sets one for train and another for test. Train data is used to fit the model and train it on algorithm, whereas Test is used to evaluate the algorithm fit of the model. It also provides parameter, where we decide the split size for data. Most preferred split size is 80:20 ratio and 75:25 ratio. We will train our model for X and Y with specific size of 0.2 which means it train data size is 80% and for test it is 20%.

Dummy variable:

Now before fitting the data, we have categorical data in training. So, we need to make those values to 0 and 1 for true samples. So, in case if sample is true, it will

consider it as 1 and other as 0. Hence, we need to make a dummy variable for X categorical features. E.g.: Fuel type has Gasoline, Diesel and other. So, it will make 3 columns for these values and will pass 1 if Gasoline is true and 0 for others (Diesel and other). It will create like similarly for all data similarly. Refer below figure:

Fuel_Fuel/Gasoline/diesel/Automatic	Fuel_diesel	Fuel_gasoline
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	1	0
0	0	1
0	1	0
0	0	1

After creating dummy variables, we will train this data on Regression model. We are using Ridge regression in built method.
We will just pass this model in object and will use this object to fit Training dataset.

7.3 Encoders:

One Hot Encoder
Many machine learning models cannot deal with categorical data for input. So whenever there is categorical data for input we need to transform it into one hot encoder or create its dummies. One hot encoder represents categorical data into binary vectors. A binary vector is created with all zero values, except for integer index whose value is denoted by 1, which means that observation is true.
E.g.:

- ['Red', 'Green', 'Green']
- [0,1,1]

It will be represented for 2 colours such as:

- [1, 0]
  [0, 1]
  [0, 1]

It acts same as dummy variable and it creates automatic labels according to input.
We can use 'OneHotEncoder' library which is provided by scikit-learn pre-processing or we can use pd method to get dummies.

Label Encoder
Label Encoder is same method which is used to convert string data to numeric. It is used for multi classification problems when you need to pass array of labels in 1d.

So, this method passes specific integer value to string data. Rather than converting it into binary, it converts it into numeric format.
E.g.:

- ['USA', 'UK', 'INDIA ', 'AFRICA']

It will convert it:

- [4, 3, 2, 1]

We can use this method when we need to classify the data for multiple labels and predict them.

7.4 Logistic Regression for Gender Classification:

In this process we will fit logistic regression for Male and Female gender prediction according to the car features. So, in this model we will be able to understand Male and Female preferred car models and their prediction probability. We have already classified gender in Gender table (Male = 1, Female = 0).
Now, we will train this data on features and labels such as:

Features: (Year, Make, Model, Price, Engine CC, Power, Fuel, Transmission)
Labels: (Gender)

As we see we have features in categorical value so we will convert them into 0's and 1's for true value by using get_dummies() feature provided by pandas.

After this procedure we will, split data into train and test. In train we fit the model for train data and will predict results and accuracy on test data.

Working of Logistic Regression:

As we have 0 for female and 1 for male, we are using logistic function with one vs all classifier to predict the probability of each samples label. So, it will use maximum likelihood of that label occurring in that specific sample. Now, we will use inbuilt methods for solver and iteration ('lbgfs',1000).
As we are classifying gender, we can use any one of the classifier methods. After fitting One vs rest method for logistic regression, we will estimate the log likelihood probability of each gender for each sample.

Predict:

We will now use predict function to predict the outcome of test data. While predicting labels it will consider most favoured label for that specific sample.

Accuracy:

We will use accuracy score function to calculate accuracy of our model for gender prediction based on Car Features. This will match predicted labels with real labels and calculate its score according to labels which are correctly classified. So according to our model, models' accuracy for predicting correct gender values is: 95%.

As we have seen our model is able to classify gender with respect to car features, that means our model is able to predict which car is preferred by male and female. So, we can assume that whichever car is considered for male gender that car has more preference of Male gender rather than female hence vice versa.

To predict Car (Model):
In this section we will fit our data for model classification instead of gender. So, we will see that our car features and gender is able to predict correct car model according to the observation.

So, we have Categorical and continuous data, where we have to predict Car models which is Categorical. So, we will convert it first into numeric value for specific car models. For this, we will use label encoder to convert categorical values into numerical label.

7.5 Ridge Regression:

X = features (Year, Make, Price, Engine CC, Power, Fuel, Transmission, Gender)
Y = labels (Car Model)
This will be our X (features) and y(labels) for the data. Now we will follow these steps:

- Creating dummies for X variable as it contains categorical data. The shape will change for X after using this.
- Creating encoder values for Car models label to predict.
- Now we will split data into train and test samples using train_test_split library.
- We will fit these X and y train values for Ridge regression model. We are using ridge regression because it can perform (L2) regularization and it can be used for classification problems.
- We will fit this available Ridge function on x and y train.
- We will use predict method to predict labels for test sample.

7.6 Logistic Regression with One vs Rest:

Our variables will consist of same data for X and Y throughout Code.
X = features (Year, Make, Price, Engine CC, Power, Fuel, Transmission, Gender)
Y = labels (Car Model)
This will be our X (features) and y(labels) for the data. Now we will follow these steps:

- Creating dummies for X variable as it contains categorical data. The shape will change for X after using this.
- Creating encoder values for Car models label to predict.
- Now we will split data into train and test samples using train_test_split library.

- Fitting the logistic regression Function provided by scikit-learn library. We will import some additional inbuilt functions as we need to add multi classifier for data. So, we will use OVR (One vs Rest Classifier) in this step.
- We will compare the dependent variables values with each other by passing it to function and comparing the independent variables relation. Hence by using sigmoid function and log likelihood probability we will be able to get estimated probability for that dependent variable (each sample).
- We will fit this available Logistic function on x and y train.
- We will use predict method to predict labels for test sample.
- Compare the results with Confusion matrix and see accuracy for this regression method.

7.7 K Nearest Neighbours Classifier:

X = features (Year, Make, Price, Engine CC, Power, Fuel, Transmission, Gender)
Y = labels (Car Model)

This will be our X (features) and y(labels) for the data. Now we will follow these steps for KNN model:

- Creating dummies for X variable as it contains categorical data. The shape will change for X after using this.
- Creating encoder values for Car models label to predict.
- Now we will split data into train and test samples using train_test_split library.
- Fitting the KNN Function provided by scikit-learn library. We will have to specify the number of neighbours to compare distance between them and sample.
- In this step we will pass test sample in which we will provide number of neighbours to compare. When we pass the parameter, it finds the distance between its neighbours. It uses any distance formula such as Manhattan, Euclidean etc. After comparing it returns all nearest neighbours and predicts label according to maximum occurrence of sample in nearest class.
- We will fit this available KNN function on x and y train.
- We will use predict method to predict labels for test sample.
- Compare the results with Confusion matrix and see accuracy for this method.
- Compare roc curve, f score for this method with all labels.

8 Analysis

In this part we will see some key observations and plots which will give us details on cars data. In first part we observed that some car manufacturing company has high car sales for certain models. Those who had high sales were Ford, Volkswagen, Renault, BMW, Toyota, Mercedes etc.

Let's now observe which car had highest sales and lowest sales throughout all years.

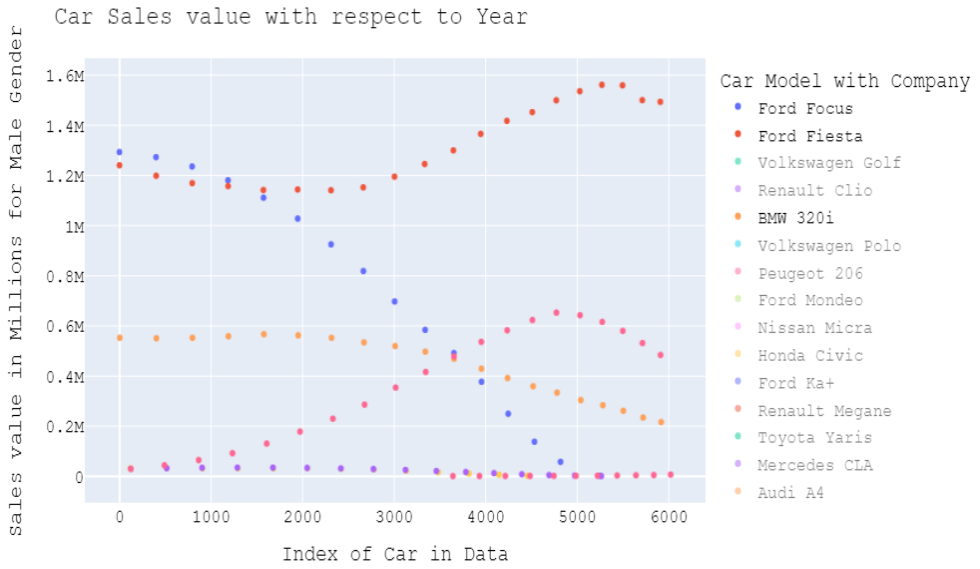


Figure 11

In this figure we have Car sales value on Y-axis, and on X axis we have all Car models through year 1994-2013 (Each dot starting from Left to right is represents year for that car and its sale). We can observe Ford Focus, fiesta are models which has highest sales in UK which range from 1.4 to 1 million. Some cars had low sales which were Land-Rover, BMW, Volvo. There are some cars which have had higher sales at the beginning but later decreased. You can observe this data and trend for specific Cars by double clicking on model in Code file.

We can also see ford escort had high number of sales but decreased exponentially from 1998 to 2013. According to the graph below, we observe Fiesta had more sales in total but then its sales value decreased linearly. As compared to Fiesta, Focus gained popularity around 2004 and has highest sales overall in 2013.

Now let's compare Sales difference between Male and female in Car models sales. We will try to notice which car had higher sales for each gender. We will compare this for only 15 cars. We will try to observe if there are any major changes in those gender.

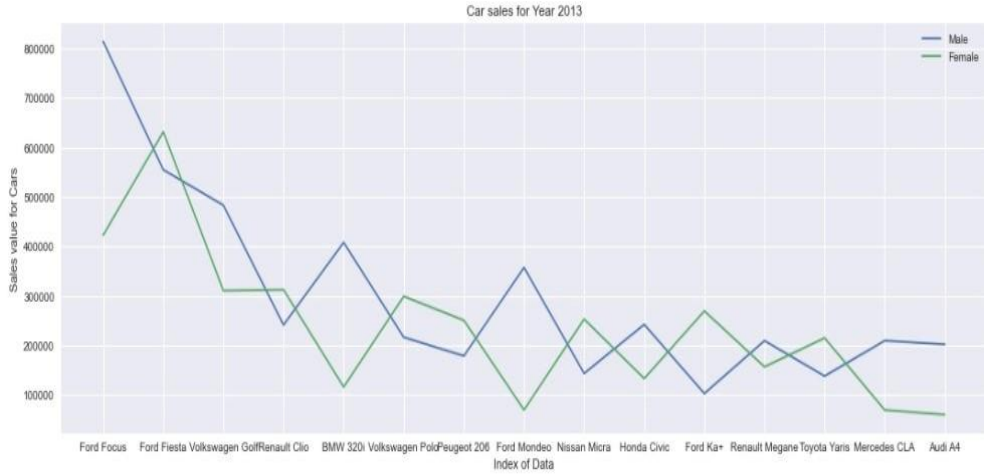


Figure 12

We can see some car models have fixed higher number of sales for specific gender. Which means we can see that gender prefers to buy that car in particular. As for ford focus male has higher preference and for fiesta female. We can observe male car has higher sales in BMW, Audi, Mercedes. So definitely there is some relation for particular car buying according to gender.

As per our data can make predictions on Car models or Company, where we can identify whether car model can be described as gender. So, if our model is able to predict which gender has purchased that car, we might be able to predict Car as Gendered Car according to number of sales.

8.1 Logistic Regression for Gender:

As mentioned in 6.4 we fitted the logistic model (multinomial) to predict gender. As we were able to predict the model with 95% accuracy let's see the results. We can observe that through confusion matrix true values and false values.

Confusion matrix:
[32] "It is tabular summary of the number of correct and incorrect predictions made by a classifier. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score".

```
Accuracy: 96.47251845775226
[[ 160   23]
 [  20 1016]]
```

: <AxesSubplot:>

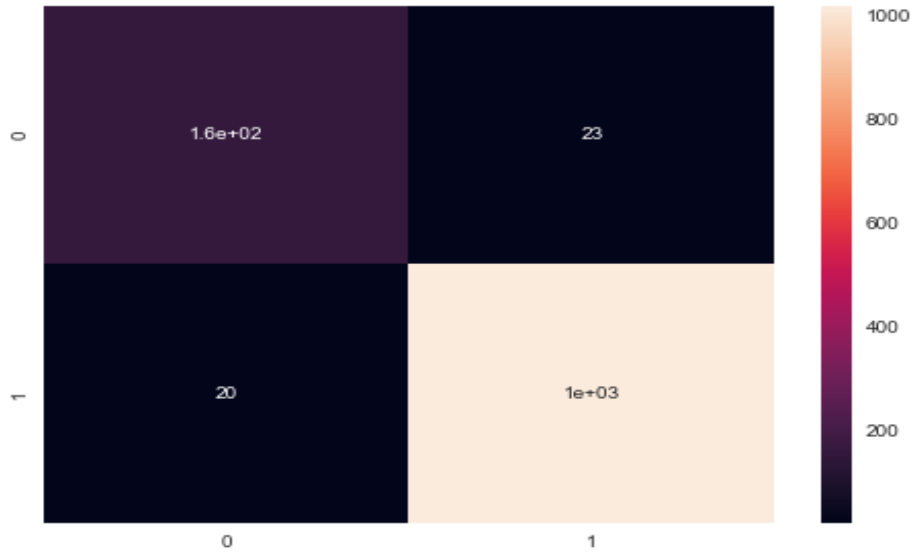


Figure 13

As you can see our model was able to predict almost 95% values correctly for which gender according to car features.

- true positive: for correctly predicted Female values (160).
- false positive: for incorrectly predicted Female values (23).
- true negative: for correctly predicted Male values (1016)
- false negative: for incorrectly predicted Male values (20).

As we can observe most of the predictions were accurate and our model was able to predict values with 95% accuracy. In this graph we are also noticing there are very few models registered to women then men. So, we can say that male gender has highest preference for car selection in test set.
For same our F-score was also 96% which is very good. Closer our value to 1 the better the results for our model.

Now let's see roc curve for out matrix.

Roc curve: [33] "The ROC curve shows the trade-off between sensitivity (TP) and specificity (1 – FP). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FP = TP). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test".

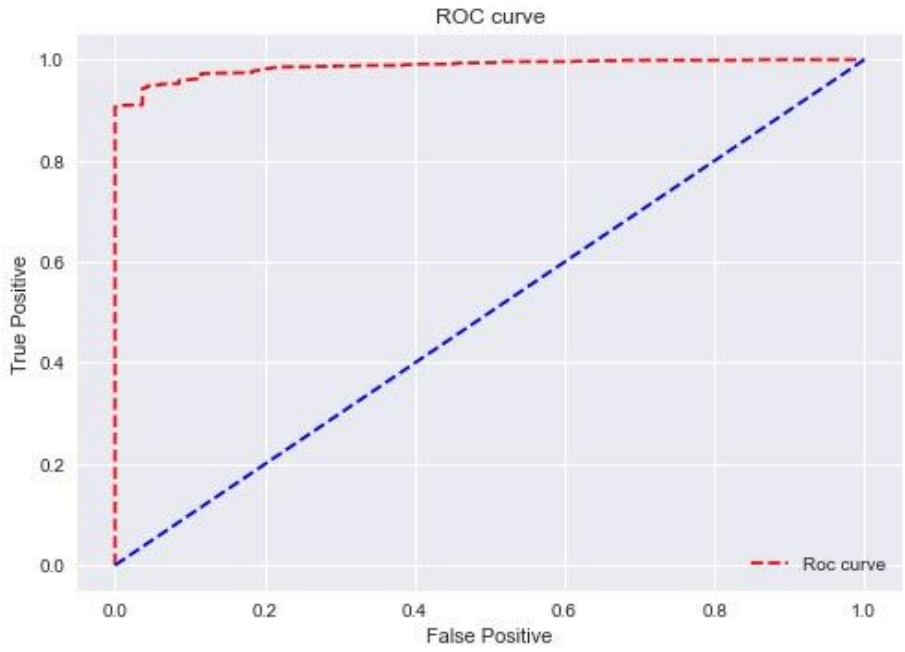


Figure 14

According to Gender classification our model was able to predict 95% values correctly for Car features. As well as our graph is also accurate for roc curve. Hence, we can definitely predict which gender will buy which specific car.
Now let's compare at predictions of gender.

Output Predicted Values

```
Actual values for gender: [1 1 0 1 1 1 1 0 1 1 1 1 1 0 1]
Predicted values for gender: [1 1 0 1 1 1 1 0 1 1 1 1 1 1 1]
```

As you can see our model is able to differentiate between male and female accurately and model is not overfitting as well. It was able to mis predict only one value in this 15 data length. So, now we will try it for Car model prediction.

Classifying models with respect to gender:

We will try to predict car model with car features and gender. With this we will be able to understand which gender actually prefers which car model and we will be able to classify that car as gendered car.

First, we will see Ridge regression method as using regularization technique in it. As this regression can be used for multi class problems we will see accuracy with this model and how it fits our data.

Accuracy: After fitting our model we were able to see our accuracy was only 33% for the model. Which shows that our predicted values were far away form the true values.

Plot:

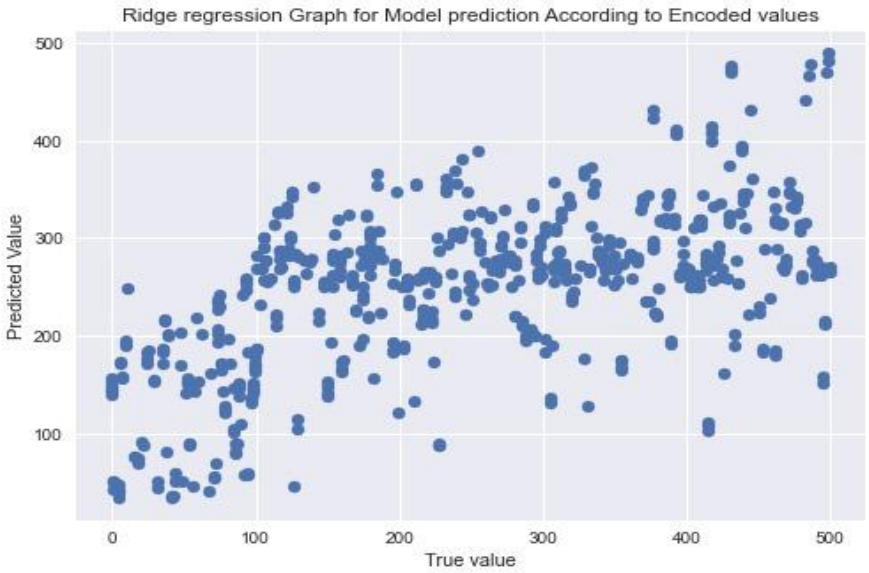


Figure 15

As you can notice in Figure 15 we have specific model number for each models and our accuracy for model was 33 percent. Notice, we can see scattered plot for true to predicted values, as all the values are far from each other. That means our model was not able to classify the model for gender according to the car features. This model is worst for our model with low accuracy and high error rate as well.

8.2 Logistic OVR:

In this model we are using multi classifier for comparing models with each other. As discussed in Logistic regression OVR we will use one vs rest classifier with labels to predict the output for all the labels. So, in this algorithm we were comparing one labels probability to all other labels.
Our model was able to predict only 52% values accurately which means we improved the accuracy of model by this method. Although we still have less accuracy and f score for this model. Our model was not able to classify all models

accurately. Some of the labels which were predicted were not in sample. So our model was not able to classify them all accurately.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
2	0.00	0.00	0.00	2
3	0.00	0.00	0.00	2
4	0.38	1.00	0.55	3
5	0.00	0.00	0.00	1
6	0.00	0.00	0.00	4
7	0.00	0.00	0.00	3
8	1.00	1.00	1.00	1
10	0.55	1.00	0.71	6
11	0.33	0.50	0.40	2
13	0.00	0.00	0.00	1
15	0.00	0.00	0.00	5
16	0.11	0.50	0.18	2
17	0.25	0.25	0.25	4
18	1.00	0.50	0.67	2
19	0.00	0.00	0.00	5
20	1.00	1.00	1.00	1

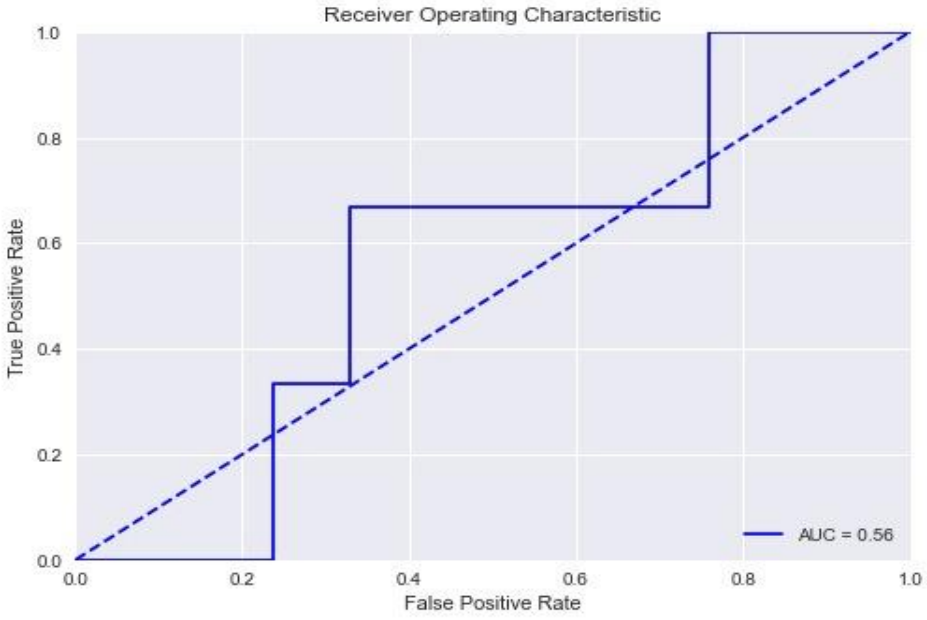


Figure 16

We can see the classification report for the labels predicted and their score. In the image below you can see precision and recall and f-score values for the labels which are given on left side.

ROC plot: Second label:

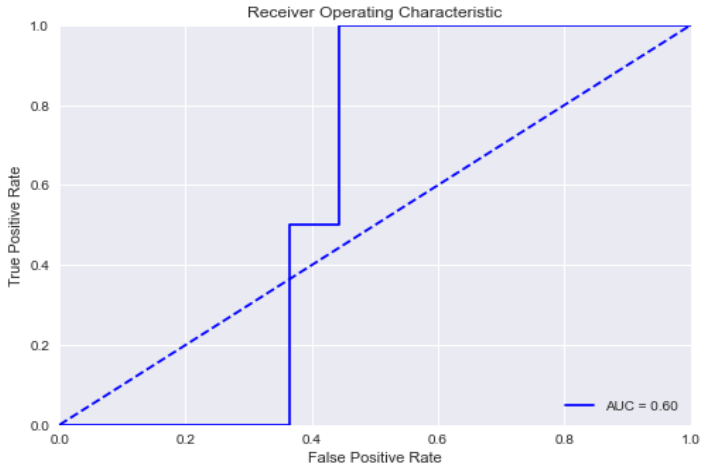


Figure 17

In above Roc plot we can see one of the labels accuracies in out model. Label 4 f score was giving top left curve for the graph. However, this is only for one of the labels from 500 labels. So now let's compare for another label.

```
Gender:  [[0 1 1 1 0 1 1 1 1 1 1 1 1 0]]
Actual car models:
 ['Puma' 'Cerato' 'Corrado' 'Tucson' 'Cinquecento' 'Z3' 'Granada' 'Kangoo'
 'Prisma' 'Nexia' 'Hilux' 'CLK' 'S40' 'Kangoo' 'RAV4']
Predicted car models:
 ['Puma' 'Mentor' 'Scirocco' 'Sonata' 'Cinquecento' 'Z3' 'Maverick'
 'Kangoo' 'Dedra' 'Nexia' 'Hilux' 'CLK' 'S70' 'Laguna' 'RAV4']
```

Now we can see for label 16 roc curve is not perfect. Curve is almost going diagonally for the label. So, it proves that our model is not good fit for some label. Accuracy for this regression was 50% with f score 52%. So, according to previous model we can say this model performs better in our case.

Output Predicted Values:

Now you can observe the predicted values that we got from this model. Our model was able to predict all correct models for female within specified range although

some male models were miss classified. Moreover, our model is able to predict correct company for models, but its just specifying wrong car model. Hence, we will try to fit another model which can fit data perfectly.

As we observe our model is able to differentiate between male and female preferred car, however some of the labels are classified very poorly according to f score. Therefore, we will try new model for this data.

8.3 KNN Classifier

To overcome the model prediction, KNN is best in this case as it is used for both types of problems regression and classification. So, we will see how our model predicts values for car model classification. KNN uses neighbours' comparison to predict output value to belonging class.

```
Accuracy: 68.58080393765381
F1 score: 72.03791469194313
```

As we get accuracy of 69%, we can see our model increased the accuracy. F score is also 72% which is very good compared to other models. Our model is able to predict and generate good prediction for test samples.

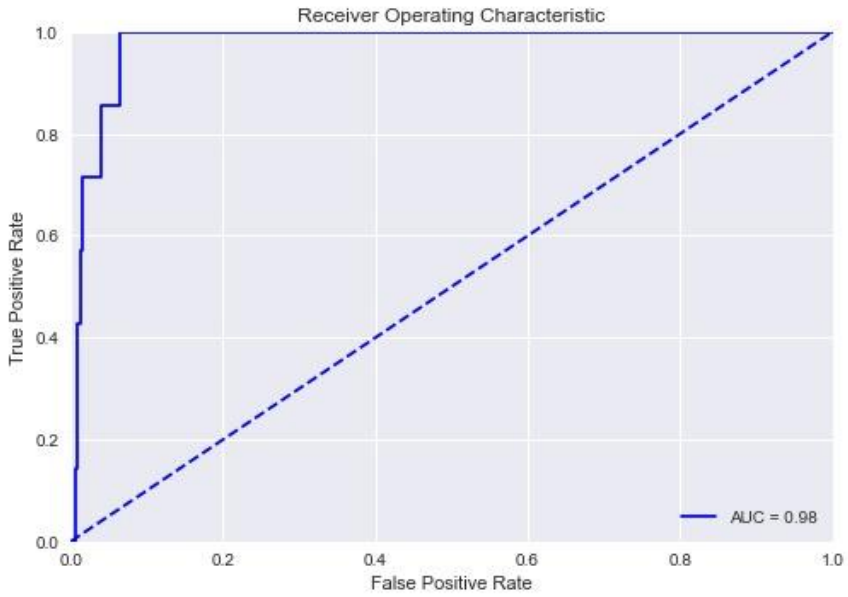


Figure 18

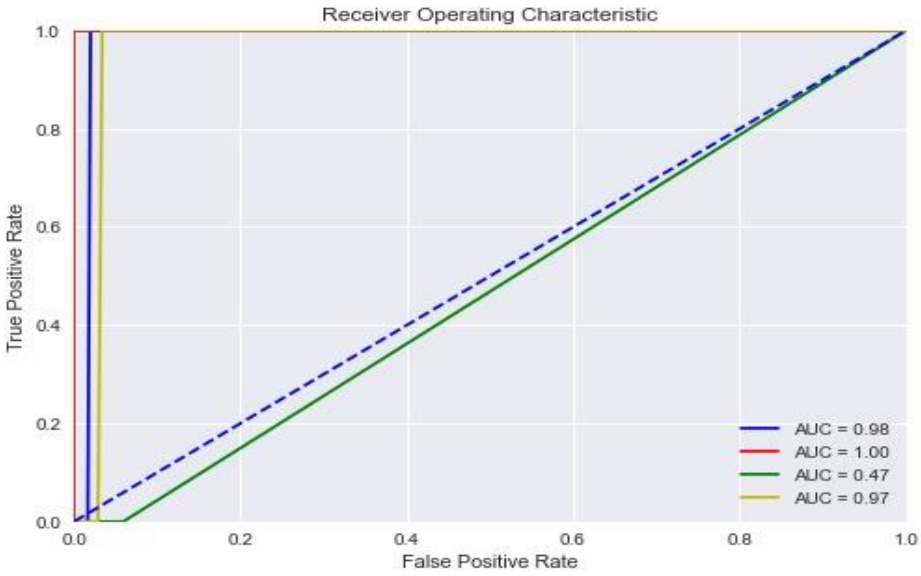


Figure 19

As we see our model prediction for roc curve is also good for samples. Only some samples had curve close to 45%. Some samples, were not predicted as they were not present in test labels as some models may contain in test because of random state. So, output may differ every time. But our model was able to predict accurate values. You can change the roc curve to see output for various labels by changing the label in code. As we have 500 + models it is very hard to visualize and plot those models.

Output Predicted Values:

```
Gender : [[1 1 1 1 0 1 1 1 1 1 1 0 1 1 1]]
Actual car models:
 ['Leon' 'Lanos' 'Golf' '806' 'Silvia' 'MX-3' 'Corrado' 'Lancer' '323' '99'
 'Roadster-coupe' 'Polo' 'Beetle' '166' 'Prius']
Predicted car models:
 ['Alhambra' 'Lanos' 'Golf' '806' 'Silvia' '5' 'Corrado' 'Lancer' 'MX-3'
 '99' 'Roadster-coupe' 'Polo' 'Jetta' '164' 'Prius']
```

As we see, Silvia and Polo are predicted as female gendered car as it is registered more to female rather than male. So, we can predict that car as female, while other cars as male. However, still some cars are misclassified but its fine we don't want to overfit our model for prediction. Our model is able to generate accurate results for this without any major classification errors.

9 Conclusion

From analysis of data and predictions we were able to classify car models as gendered. Moreover, according to observations, we can notice car model considered as female have lower price range compared to others (e.g., Fiesta, Polo, Clio Estate). We observe all these cars have lower price range, less power and engine cc whereas for male categorized car are pretty huge brands of cars such as Audi, BMW, Mercedes, Land rover. All male gendered car has price, high power and high-end engine cc. This can describe that, according to [6] "Male prefer cars which are well known in brands such as Audi, BMW, Mercedes. Moreover, they prefer higher power and engine CC in car. Price factor doesn't affect much for each gender although it depends on persons income. However, female prefer cars which safer or smaller in size which are family cars. Female has tendency of going towards basic car with less power and more durable which can be used for day-to-day purpose. For those cars which has both male and female sales same, we can classify these cars as cross such as (Mazda MX-5).

However, we collected car data from site which might have some errors as well. Hence it depends on website or data as well to predict correct results. If data is incorrect our model won't be able to produce any better understandable results. In our case our model is able to predict true results and we were able to find pattern. Moreover, it's easy to predict car models if we pass car manufacturer in input as well, as model is able to understand the relation between them. We can also apply same method to identify which gender prefers which car companies. However, in our case once we are able to predict correct car model, we are also predicting the car manufactures. For this data KNN, or Random Forest method would be perfect for classifying car models.

However, there are also other machine learning methods which we can apply to predict better outcome such as deep learning techniques, MLP, CNN, SVM. SVM and CNN might be able to give more accurate results than this because those are trained on neural network. It depends on how many algorithms we try on data and see which algorithm is best for specific data.

Hence, we can definitely say that car models can be categorized into gendered cars according to the observation and analysis made. Here are some examples of car models considered as male and female.

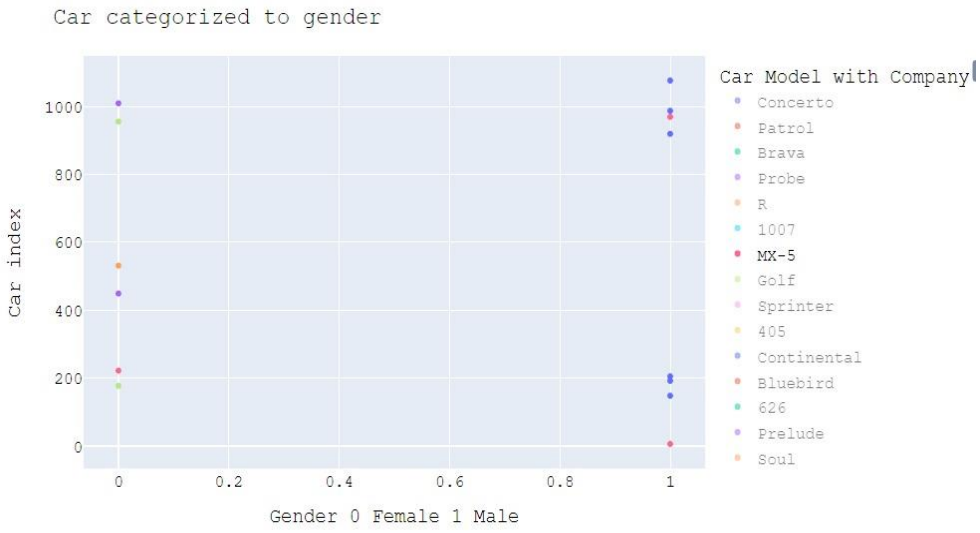


Figure 20

In above graph you will notice, we have categorized cars as gendered 0 for Female and 1 for male. As you can notice Ford focus is classified as male car whereas, Audi A, city coupe is for female. Mx-5 model is present in both so we can count it as cross gendered car. So, we can see prediction by clicking on car model to see its prediction for the car gender.

E.g.:

- Ford Focus – Male
- Volkswagen Golf – Male
- Land Rover Range – Male
- Audi A6 – Male
- Nissan Note - Male
- Ford Fiesta - Female
- Mini cooper – Female
- Mercedes A – Female
- City Coupe – Female
- Fiat 500L - Female

These are some classified models from the models which suggests which car is gendered car. However, when there is no difference between some car models, our model was not able to classify it to one gender. So, we can assume that car as Neutral car.

10 Professional Issues:

- While web scraping all the car features from car-data site, faced some issues for server requests. As the code was sending to many requests in short time. So, IP address was getting banned for this.
- Furthermore, as we are scraping whole web site and getting all car details it takes about 2-3 hours to run the specific code. So, it is not time efficient.
- After extracting and completing all the data faced some issues plotting for specific car linear plot for sales.
- Simple linear regression is not good for this data as it was giving some high MSE and very low value for prediction.
- Faced some issues plotting the ROC graphs and matrix because while predicting we had 500 labels for cars. So, tricky part was to plot 500 car models and visualize those points.

10.1 Solution for issues:

- Hence, when we are web scraping, we can ask specific site holder for site access. Or we can import some sleep function in-between the code. So, it will stop for certain seconds and our IP won't be considered as suspicious.
- Time complexity of algorithm is O(n^3) for the last part of code where we extract all data from links. We, can optimize this code for better time complexity.

11 Self-Assessment:

In this section, we will see how I executed my project from the first step, where I will also describe where this project has its weakness and where it can perform better. How this project can also be updated in future case depending upon application we need.

11.1 Strengths:

As I started this project, I developed project plan to keep track of the project activities and their dead line. First main task was to obtain information on how to extract web data as I was new to this web scraping technique. I gathered information through YouTube tutorials, web engines and programming site and learned how to write code for web scraping.

Tools:

My advisor was also very helpful and gave me some advice on how to start my project and keep UpToDate. He also helped me in suggesting me to use beautifulsoup library. As I am experienced in python than R, it gave me advantage to grasp new techniques easily. Moreover, it was easy to use scikit learn library to perform new models testing and its operations to analyse the results. As I learned new techniques this will also help me in my future career as data scientist how to gather data through web. As I was making progress I was keeping in touch with my advisor and also explained him about my issues in code, where he guided me and gave me some references. All this work was updated on one drive storage which helped me to keep in touch with files even without my laptop.

11.2 Weakness:

As when I was scraping the data from website, site was cancelling my server requests because of too many requests. As I had to collect 15000 car models' data from web it was consuming huge amount of time. It almost took 3-4 hours to run 1 part of the code. Moreover, some time after sending too many requests the server stops the request and considers it as suspicious. Although somehow manged to get data for all cars, after that there was no need to re run the program again.
While plotting ROC curve, we have 500 different models so, it's very tricky to see the curve for different labels, so instead I took only 4 labels to see the difference between curves.

11.3 Project opportunities:

Throughout this task I got to learn more about time management and information gathering. Even though I was lacking behind in time on web scraping, I learned that executing tasks in time yield better results as you get more chance to explore the

areas. Moreover, I got to learn in depth about machine learning algorithms and their various applications which can help me in future whenever I am working on a project. Using data visualization techniques such as plots and PCA, helped me understanding how data looks like and how to interpret it.

These are the things that I learned in this project and would definitely improve all my weakness in upcoming years.

12 References

References

[1] M. Prieto and B. Caemmerer, "An exploration of factors influencing car purchasing decisions," *International Journal of Retail & Distribution Management*, vol. 41, *(10)*, pp. 738-764, 2013. Available: https://www.emerald.com/insight/content/doi/10.1108/IJRDM-02-2012-0017/full/html. DOI: 10.1108/IJRDM-02-2012-0017.

[2] S. G. ESENGALIEVNA, "CAR PURCHASING SELECTION OF WOMEN AND MEN: A DIFFERENT BEHAVIOR," vol. 22, *(1)*, 2015. Available: https://doaj.org/article/7e41cefdf6ed4b26a819b5e58cb9dd68. DOI: 10.32421/juri.v22i1.96.

[3] (16 APRIL). *Female car owners up by20 Percent*. Available: https://www.smmt.co.uk/2018/04/female-car-owners-up-20-in-a-decade-reveals-uks-biggest-automotive-motorparc-analysis/.

[4] (). *Linear Regression Model applied in Used Vehicle Market* . Available: https://towardsdatascience.com/car-selection-and-sales-day-prediction-8c4a474f9dca.

[5] (OCTOBER 16,). *Car-buying still has a gender gap*. Available: https://www.cbsnews.com/news/car-buying-still-has-a-gender-gap/.

[6] (October 4,). *Men vs Women Car Buying*. Available: https://www.cjponyparts.com/resources/men-vs-women-car-buying.

[7] Y. Lin, "Auto Car Sales Prediction: A Statistical Study using Functional Data Analysis and Time Series." , 2015.

[8] (December 11,). *Classification vs Regression*. Available: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.

[9] (). *Logistic Regression Analysis*. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis.

[10] (March 15,). *Logistic Regression Detailed Overview*. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

[11] (). *Logistic Regression*. Available: http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html.

[12] (). *Linear logistic regression explained*. Available: https://www.kdnuggets.com/2020/03/linear-logistic-regression-explained.html.

[13] (December 16,). *Logistic Regression*. Available: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

[14] (). *Advantages and Disadvantages of Logistic Regression*. Available: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/.

[15] (February 18,). *Multicollinearity*. Available: https://www.investopedia.com/terms/m/multicollinearity.asp.

[16] (November 12th,). *Regularization: Ridge, Lasso and Elastic Net*. Available: https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net.

[17] (March 26,). *K Nearest Neighbor Algorithm*. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

[18] (). *Distance Metrics in KNN*. Available: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html.

[19] (September 25,). *KNN Pros and Cons* . Available: https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/.

[20] (). *One vs Rest and One multi class classification*. Available: https://www.kdnuggets.com/2020/08/one-vs-rest-one-multi-class-classification.html.

[21] (30 Jun,). *Python Mean Squared Error*. Available: https://www.geeksforgeeks.org/python-mean-squared-error/.

[22] (Dec 8,). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?*. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

[23] Minitab Blog Editor, "Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?" 30 May, 2013.

[24] (). *F score , Precision, Recall and Accuracy*. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score.

[25] (). *Web Scraping attack*. Available: https://www.imperva.com/learn/application-security/web-scraping-attack/.

[26] (). *Why web scraping a full list of advantages and disadvantages*. Available: https://raluca-p.medium.com/why-web-scraping-a-full-list-of-advantages-and-disadvantages-fdbb9e8ed010.

[27] (). *Beautiful Soup*. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[28] (). *Regular Expressions*. Available: https://docs.python.org/3/howto/regex.html.

[29] (). *Data Preprocessing in Data Mining -A Hands On Guide*. Available: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/.

[30] (Nov 19,). *Data Cleaning and Preprocessing*. Available: https://medium.com/analytics-vidhya/data-cleaning-and-preprocessing-a4b751f4066f.

[31] (June 10,). *Standardscaler and minmaxscaler transforms in python*. Available: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/.

[32] (). *Confusion matrix in python*. Available: https://www.educative.io/edpresso/how-to-create-a-confusion-matrix-in-python-using-scikit-learn.

[33] (). *ROC - curve how to interpret it*. Available: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/.

[34] (). *ROC curve plot*. Available: https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python.

13 Program Instructions:

Observations and code were executed with Python language. All the code was executed in Jupyter Notebook. There are 2 folders in file.

1. Code
2. Excel csv files.

There are 4 files in code folder:

- **Scrapped dat**a: Code for extracting data from website.
- **Cleaned data:** Code for cleaned DVLA data for merging purpose with scrapped data.
- **Cleaned scrapped data:** Code for all cars which are required for final data file, through cleaning and finding mean values of all car models.
- **Analysis regression:** Main part of code where all analysis where various models were trained for data.

Instead of running all the web scrap program I would suggest just run analysis regression part, because web scrap program might take some time as you are extracting the data from internet.

To refer CSV files:

There are 4 csv files:

- **Data.csv:** Which contains merged final data with sales value according to gender for analysis.
- **Scrappeddata.csv:** It contains rough data gathered from website, uncleaned data.
- **Cleaned scrapped data**: It contains data for cleaned scraped data, which we will need for analysis.
- **Cars**: It contains cleaned DVLA data set values without missing models and outliers.

Before running any code make sure you have csv file in program locations so that while running the code it won't give any errors.

Libraries:

To execute all the code, computer must have all the libraries installed in python. If libraries are not installed you can install them by using command
--- > pip install (library name).

Libraries:

- Pandas
- Numpy
- Sklearn
- Seaborn
- Plotly
- Math
- Matplotlib
- Statsmodels

You can execute below code for libraries.

```
import pandas as pd
import numpy as np
import csv
import re
import sklearn as sk
import plotly.express as px
import seaborn as sns
from sklearn.metrics import confusion_matrix, accuracy_score
import sklearn.metrics as sm
from sklearn.metrics import accuracy_score
import statsmodels.api as smk
import statsmodels.formula.api as smf
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import Ridge
from sklearn.multiclass import OneVsRestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.metrics import classification_report
from keras.utils.np_utils import to_categorical
from math import sqrt
from sklearn.metrics import roc_curve
from sklearn.metrics import f1_score
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
```

Web Scrap file:

You might encounter some huge amount (2 hours aprox) of time for input from this file, or it may give error after running the code. As we are scraping data from website continuously server request is tracked by web siter handler, therefore if server request is sent without interruption or time delay, website may stop the connection request, and you can get some errors shown below:

- httpsconnectionpool(host='www.cars-data.com', port=443): max retries exceeded with url: /en/lada/priora-2172 (caused by newconnectionerror('<urllib3.connection.httpsconnection object at 0x0000026c44f73160>: failed to establish a new connection: [winerror 10013] an attempt was made to access a socket in a way forbidden by its access permissions'))

As web site is denying the access for us after some time. So, I suggest you can easily see the scrapped data from 'ScrappedData.csv' file.

For executing all these programs, you can open those files in Jupyter notebook and just hit the run button which is provided, or you can click the cell and press (shift+enter) to run the cells. All other program files are executable very easily and does not take time to see the results.

Code from stackoverflow:

Below code was taken from stackoverflow as I was facing some problem with roc curve library. So, I decided to use this specific code to print roc curve for plots.

```
[34] def plot_roc(model, X_test, y_test):
    # calculate the fpr and tpr for all thresholds of the classification
    pred_prob = model2.predict_proba(X_test2)
    predictions = pred_prob[:, 1]
    fpr, tpr, threshold = roc_curve(y_test2, predictions,pos_label=7)

    roc_auc = metrics.auc(fpr, tpr)

    plt.title('Receiver Operating Characteristic')
    plt.plot(fpr, tpr, 'b', label='AUC = %0.2f' % roc_auc)
    plt.legend(loc='lower right')
    plt.plot([0, 1], [0, 1], 'b--')
    plt.xlim([0, 1])
    plt.ylim([0, 1])
    plt.ylabel('True Positive Rate')
    plt.xlabel('False Positive Rate')
    plt.show()
```

All other codes are in Code section with ipynb extension.