Yipei Zhao

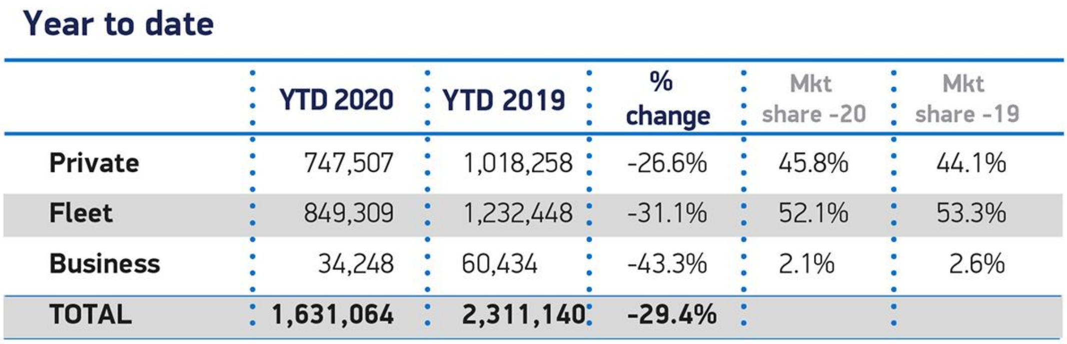
Acceptability of vehicle models prediction using classification techniques

AM40UD

TABLE OF CONTENTS

LITERATURE INTRODUCTION

In this report, we will discuss how different aspect of a car will affect the satisfaction rate of a customer.



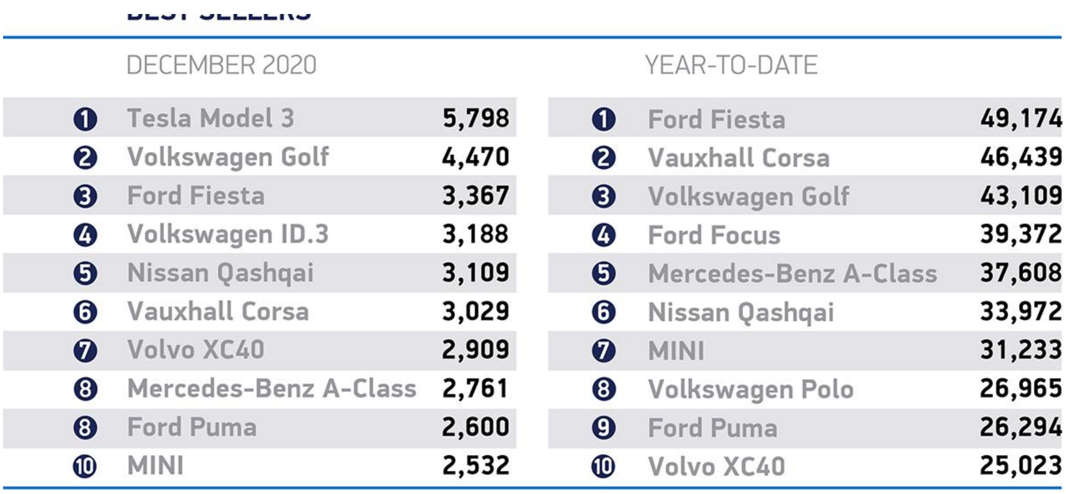
Year to date

	YTD 2020	YTD 2019	% change	Mkt share -20	Mkt share -19
Private	747,507	1,018,258	-26.6%	45.8%	44.1%
Fleet	849,309	1,232,448	-31.1%	52.1%	53.3%
Business	34,248	60,434	-43.3%	2.1%	2.6%
TOTAL	1,631,064	2,311,140	-29.4%		

Figure 1: Number of cars registered in UK in 2019 and 2020. [1]

Even though number of cars registered in 2020 have decreased by 30% compare to 2019, travel and commuting are still important for everyone. Thus, purchase an ideal car is essential.

There are many different features of a vehicle, causing different models produced by different company. A single model of vehicle will not satisfy the need of different group of users but a variation of different specifications. Hence, a key of vehicle designing is a successful combination of different features.



BEST SELLERS

	DECEMBER 2020			YEAR-TO-DATE	
1	Tesla Model 3	5,798	1	Ford Fiesta	49,174
2	Volkswagen Golf	4,470	2	Vauxhall Corsa	46,439
3	Ford Fiesta	3,367	3	Volkswagen Golf	43,109
4	Volkswagen ID.3	3,188	4	Ford Focus	39,372
5	Nissan Qashqai	3,109	5	Mercedes-Benz A-Class	37,608
6	Vauxhall Corsa	3,029	6	Nissan Qashqai	33,972
7	Volvo XC40	2,909	7	MINI	31,233
8	Mercedes-Benz A-Class	2,761	8	Volkswagen Polo	26,965
8	Ford Puma	2,600	9	Ford Puma	26,294
10	MINI	2,532	10	Volvo XC40	25,023

Figure 2: Top 10 models in UK. [2]

These listed models can be considered as great combinations of different features, for example, size, safety measurement and price etc. Our goal is to find a combination of vehicle's characteristic to build a successful model like the listed models. We also want to build models to predict the acceptability of a vehicle model. A dataset has been found and used to try to answer this proposal [3].

DATA PREPARATION

Data Description

The data contains 1727 entries. The data contains no null values or NaN.

The data consists of 6 features and 4 classes:

Variable	Definition	Description
buyingPrice	Purchase price of a vehicle	v-high (very high), high, med, low
maintenanceCost	Cost to maintain a vehicle	v-high (very high), high, med, low
doors	Number of doors of a vehicle	Discrete, 5-more are recognised as 5
capacity	Capacity in terms of maximum number of people the vehicle can carry	Discrete, denote more if capacity exceed 4
luggageSpace	Size of the luggage boot	big, med, small
safety	Estimated safety of a car	high, med, low
acceptability	Target classes, how acceptable for a customer	unacc(unacceptable), acc(acceptable), good, very good

Generally, we can conclude all six features into three categories:

- Cost

General cost of the vehicle, including buying price and maintenance cost. For an ideal model, it doesn’t necessary need to be cheap, but cost efficient.

- Size

Size of a car, including number of doors, capacity and size of the luggage boot. Different sizes of vehicles satisfied different needed. For example, a small vehicle is good for commuting. But a larger vehicle is better for family uses. And usually, a smaller car cost less.

- Safety

Estimated safety of a car. It is also an important factor of vehicle purchasing.

Data Pre-processing

The original dataset is in a '.data' format, and it was converted into a csv files, which can be easier to monitor and manage. After loading the dataset using the python pandas package:

	vhigh	vhigh.1	2	2.1	small	low	unacc
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

Table 1: First 5 entries of the raw data.

Not only the column's names are messed, but also the features are in string; they are not good for calculations. It is necessary to encode non-integers features to integers for future calculatios and evaluations.

After processes:

	buyingPrice	maintainanceCost	Doors	Capacity	luggageSpace	Safety	Acceptability
0	3	3	2	2	0	1	0
1	3	3	2	2	0	2	0
2	3	3	2	2	1	0	0
3	3	3	2	2	1	1	0
4	3	3	2	2	1	2	0

Table 2: First 5 entries of the processed data.

All observations' features have been converted to an integer number. This is necessary for mathematical calculation and neural network implementation.

Variable	Before transforming	After Transforming
buyingPrice	v-high (very high), high, med, low	3,2,1,0
maintenanceCost	v-high (very high), high, med, low	3,2,1,0
doors	2,3,4,5-more	2,3,4,5
capacity	2,4, more	2,4,5
luggageSpace	big, med, small	2,1,0
safety	high, med, low	2,1,0
acceptability	unacc(unacceptable), acc(acceptable), good, very good	3,2,1,0 (Lower value means better)

Table 3: Variables before transforming and after transforming

Data Distribution

Understanding the distribution of features can help us to have a clearer view on the data, and justify the fairness of the dataset. For example, we don't want a dataset has 99% of the entries belong to class 1. In this case, the learner models cannot make a good prediction. An even distribution is better for prediction.

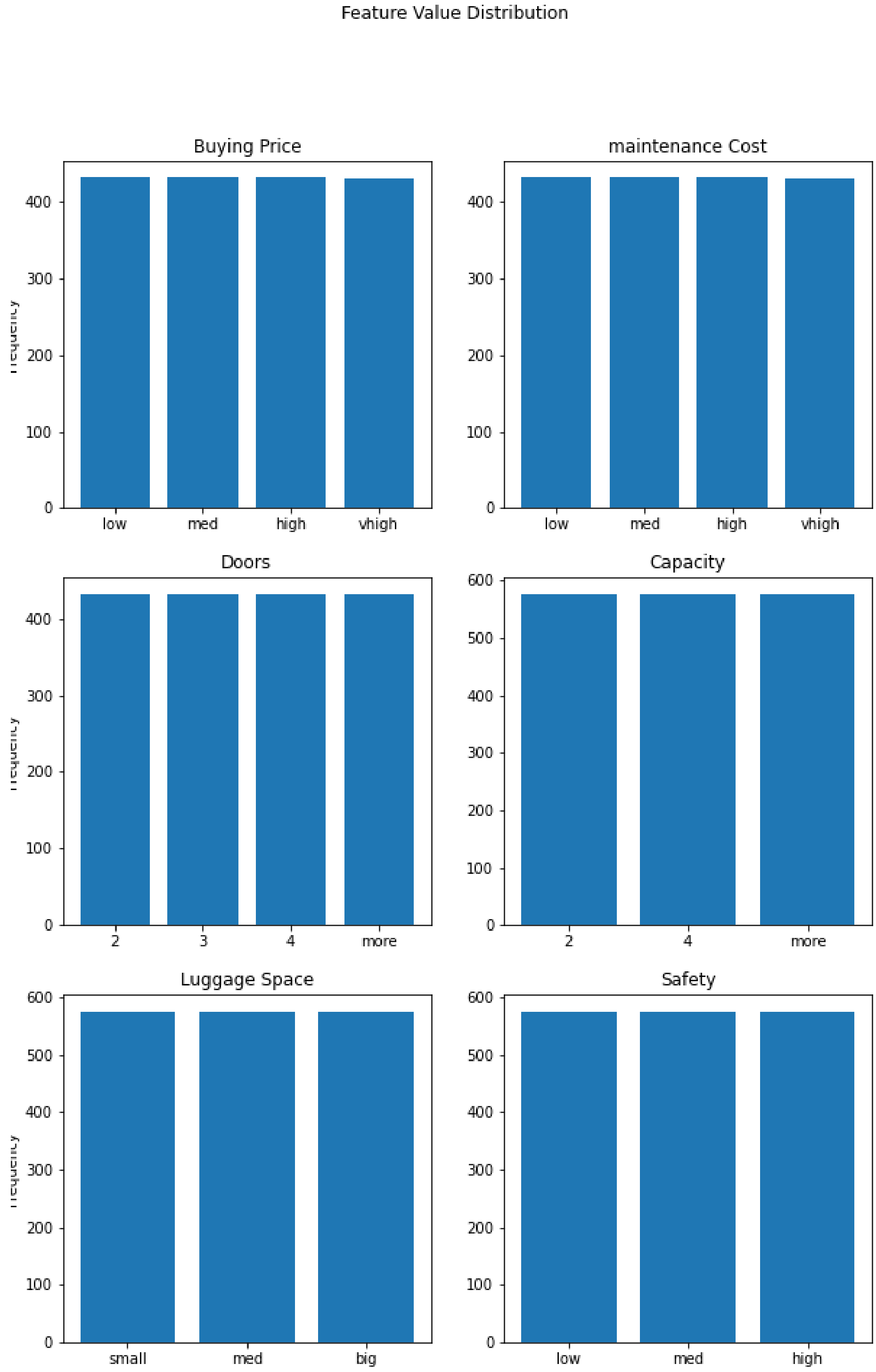
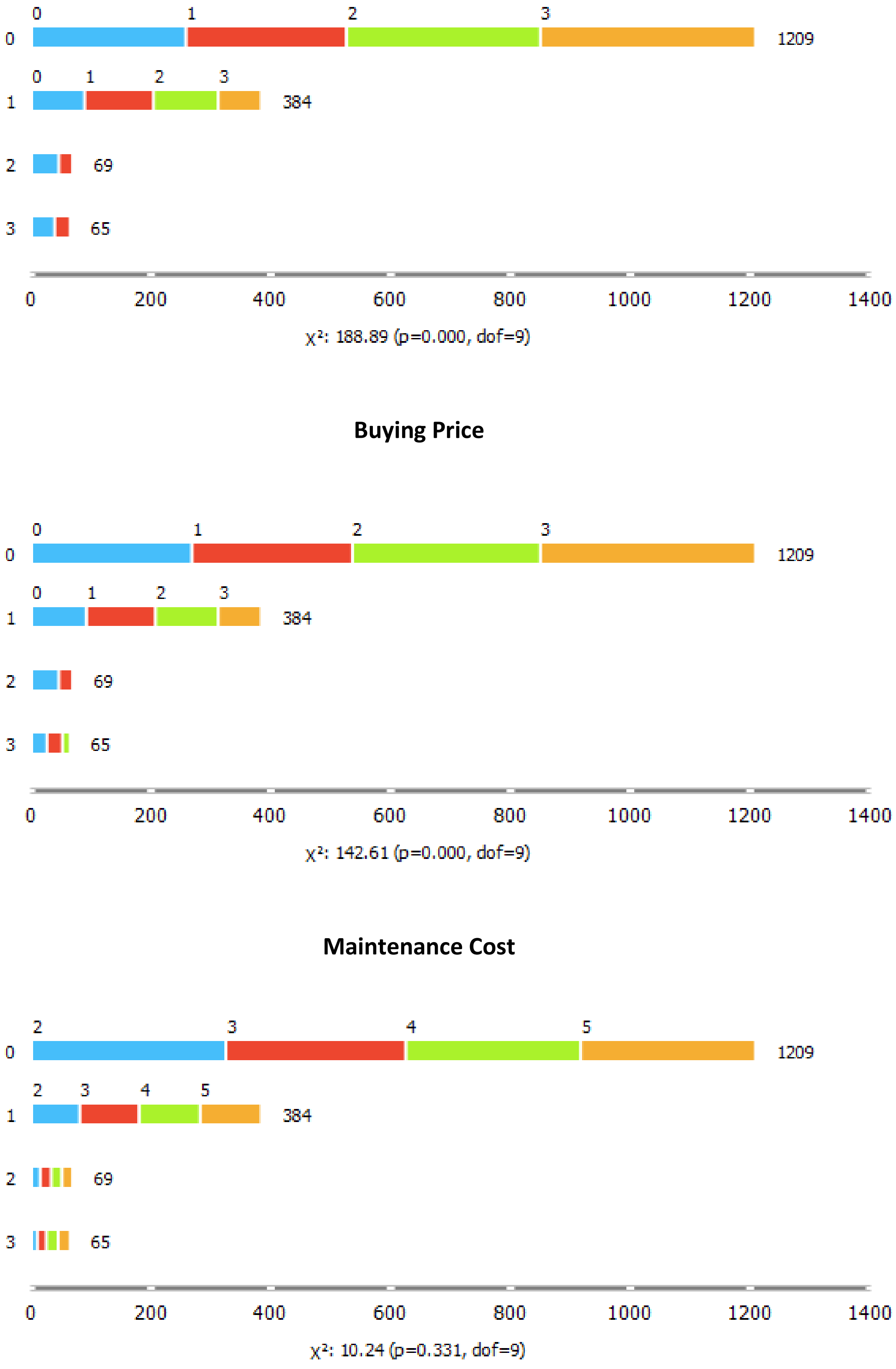


Figure 3: Empirical distribution of the different features' values.

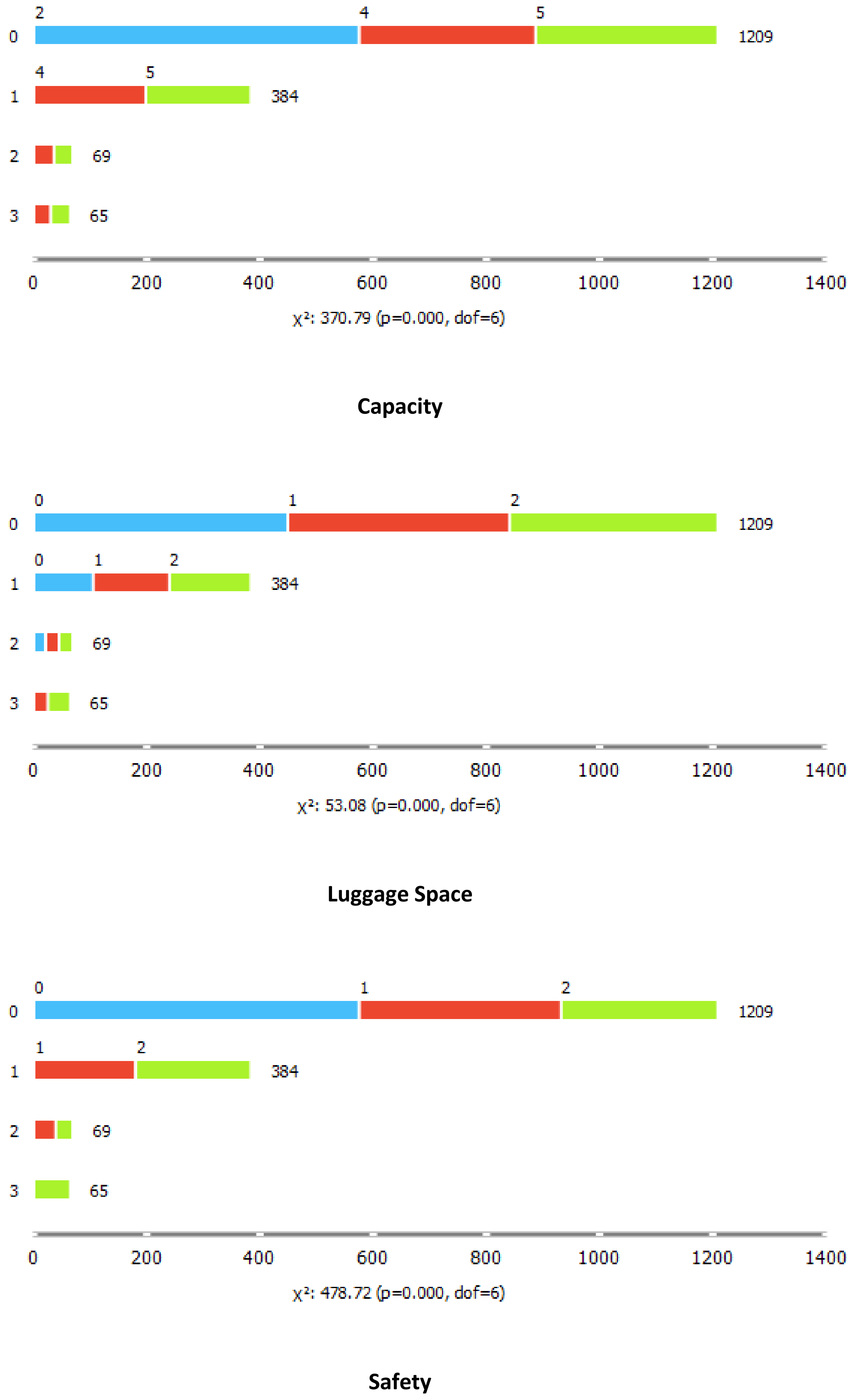


Buying Price

Maintenance Cost

Doors

Figure 4: Feature distribution by classes.



Capacity

Luggage Space

Safety

Figure 4 continue: Feature distribution by classes.

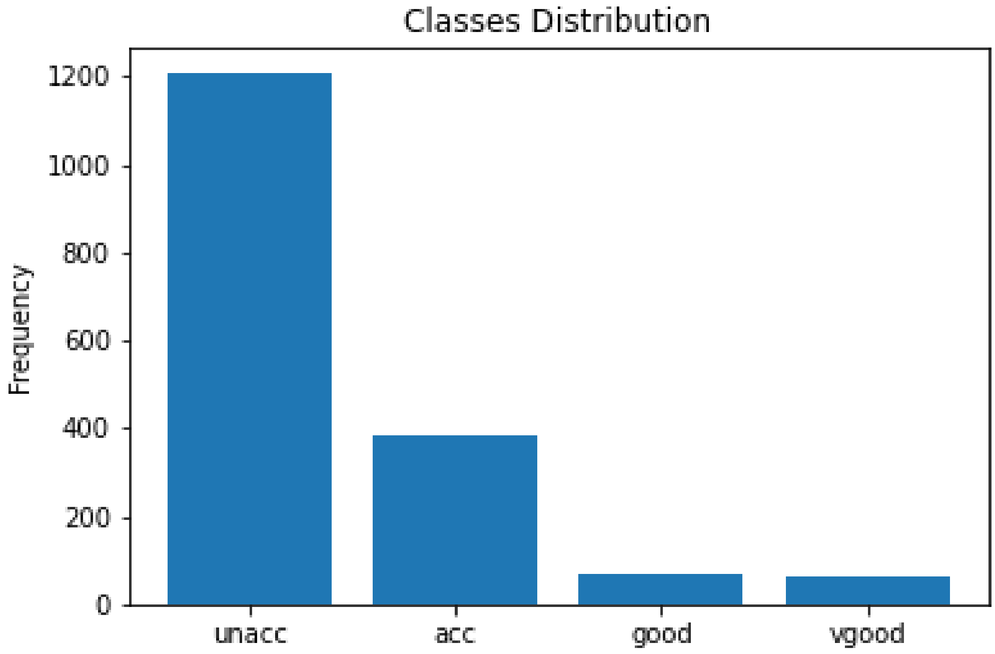


Figure 5: Empirical distribution of classes

Even though all variables distributed evenly, but the distribution of classes are not even. Most vehicles are unacceptable for customers. More vehicles are unacceptable than acceptable, good and very good combined. Our goal is to find the combination of features to build a vehicle that is at least acceptable for customers.

From figure 4, we can have a rough understanding on the relationship between each independent variable and the dependent variable.

- Buying Price:

To build a good or very good vehicle model, the cost is an important factor. As we can observe from the figure, to be considered as good or very good, the price must be low or med. Otherwise, it is not in consideration.

- Maintenance Cost:

Similar to buying price, the maintenance cost also need to be low or med to be considered as good or very good model. However, there are models considered to be very good with high maintenance cost. A reason can be maintenance cost isn’t as intuitive as a buying price. Maintenance is charged during the life time of the vehicle, instead of a buying price, which you will pay in one go.

- Doors:

The variables distributed very evenly, meaning that doors may not be an important factor.

- Capacity:

To be mentioned, all 2 seaters car are unacceptable. This is what to avoid when designing the model.

- Luggage Space:

Generally, we consider this to be larger the better.

- Safety

All low safety vehicles are unacceptable for customers. No one will risk their life. To build a good model, it should have at least medium estimated safety.

DATA MODELLING

Methods

To answer our proposal, we can treat the dataset interpretation as a classification problem. As we learnt, there are 4 classes in this dataset, the job is to train a model that can classify a new entry whether it will be unacceptable, acceptable, good or very good for customers.

Four models will be train to perform the classification:

- Logistic regression

- KNN (K-Nearest Neighbours)

- Random Forest

- Neural Network

Logistic Regression

Logistic regression is a statistical model to model binary/multiclass dependent variables. It is used to find the relationship between independent variables and dependent variable. The main idea is finding a correlation between each independent variable and dependent variable and interception, and build a function that can classify new entries.

Advantages:

- Relatively simple to implement. It is a simple method to classify data yet efficient.
- Model coefficients and interceptions are calculated. Coefficients and interceptions can be used to justify the relationship between independent and dependent variables.

- Good accuracy for simple data. If the dataset is simple, the logistic regression can maintain a good accuracy.

Disadvantages:

- Assuming linearly. While building, and fitting the model, we are assuming there is a linear relation between the independent and dependent variables. If this is not the case, this model cannot be used.
- Difficult to obtain complex relationship. If the independent and dependent variables have a complex relationship, the prediction can be inaccurate or useless.

KNN

A density estimation method that can be used for classification. The main idea is creating a hypersphere centred at a data point with pre-defined parameter K, the hypersphere should contain precisely K data points. The new data point is classified as most votes from its K neighbours.

Advantages:

- Easy to implement. Relatively easy to implement.
- No training period. Making the algorithm much faster than other algorithms.

Disadvantages:

- Doesn’t work well against large dataset. This process will be computationally costly calculating distance between each neighbour and the data point.
- Noise can affect model fitting. An outlier can cause the model to miss-predict.

Random Forest

Random forest method can be used on regression and classification problems. It constructs a decision tree that split a node in two or more sub-nodes. Each node takes a decision to pass the input data to the further sub-nodes based on a condition given.

Advantages:

- Easy to understand, implement and interpret. No professional mathematical knowledge needed to interpret data as it is straightforward.
- Can handle numerical and categorical data. The model has a good compatibility.

Disadvantages:

- Noise can affect the model fitting. There is no good way to avoid noise beside data cleaning before fitting.
- High variance. The model can get unstable.

Neural Network

A modern learning model. The idea is to build a network with multiple layers. Each layer can function differently, an input layer and an output layer is compulsory. Each layer has a pre-set number of perceptron. Each perceptron has a ‘weight’ to model the input data. The weight is updated every time until the model can predict result confidently.

Advantage:

- Neural network can be built to model any data. Including numerical, categorical, simple relationship or complex relationship.
- A single failure of neuron will not affect the whole system. The model can still predict accurately.

Disadvantages:

- Difficult to implement. Implement an optimal neural network model requires a lot of knowledge and repetition experiment of parameters.
- Requires large amount of data. A neural network model usually requires a large amount of data. Otherwise, the model can be inaccurate.
- Black box. We can observe the output but not the process. In another word, we will not have an idea of how this output is produced, causing feature interpreting to be difficult.

Features evaluation using logistic regression model

As stated in the advantages of logistic regression model, the coefficient of each feature can be used to model the importance of each feature. This is an important step to achieve our goal as our goal is to find out the relationship between each feature and the acceptability of a vehicle model.

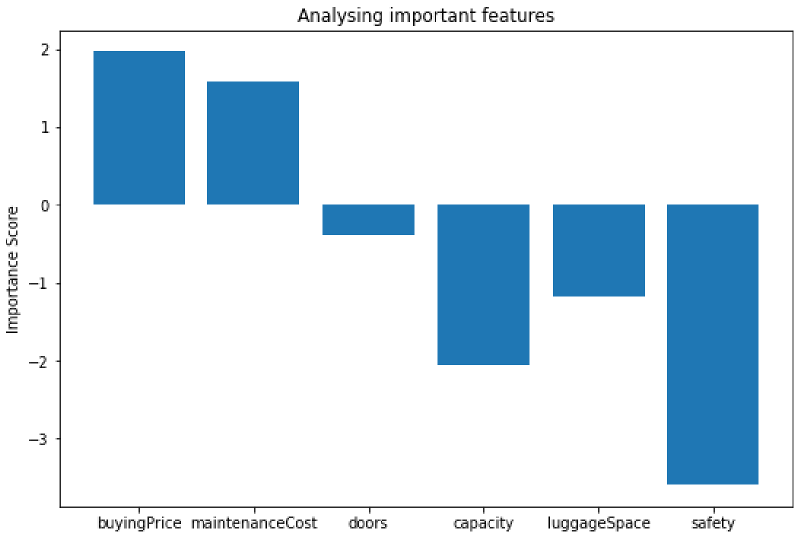


Figure 6: The relationship between each feature and the output.

From this figure, we can observe that safety is the most important feature when the customer evaluates a vehicle model. The safety feature is in a negative correlation, reinforcing that customers generally think that safer cars are better. Both costs are in positive correlations, meaning the more expensive vehicles are less acceptable. And as suggested, number of doors doesn't make a lot of impact on the decision of customers. But capacity and luggage space do. We discussed about the capacity impact on the acceptability, and drew a conclusion that a 2-seater vehicle is unacceptable, that's why the correlation between capacity and acceptability is negative (more seats are better).

Models Implementation

To model the data, Orange is used. Here is the workflow to fit the model and obtain the result (Input data is input after cleaning). All models are fitted using random sampling. 80% of the data are used to train the model, 20% of the data are used to obtain the accuracy of the data.

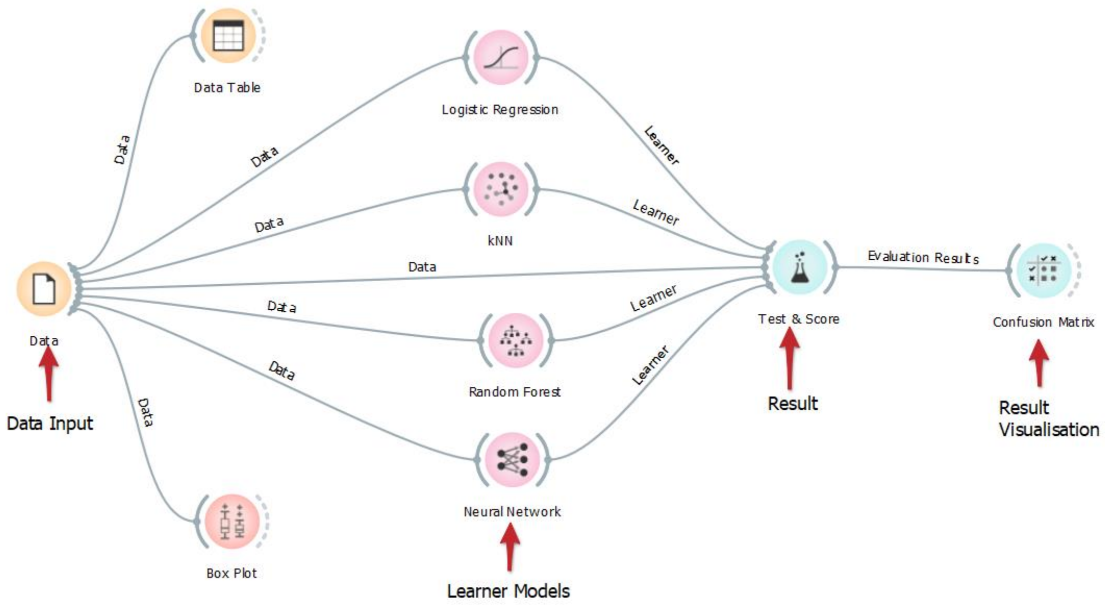


Figure 7: Workflow of the Orange data mining program.

Four models are implemented and trained. Parameters are used as followed:

- Logistic regression: No parameters needed.
- KNN: K=5
- Random Forest: Number of trees=6
- Neural Network: 2 ReLu layers with 50 neurons each, 50 iterations, complied with Adam solver.

Models Precision

Model	Precision
Logistic Regression	0.917
KNN	0.875
Random Forest	0.926
Neural Network	0.981

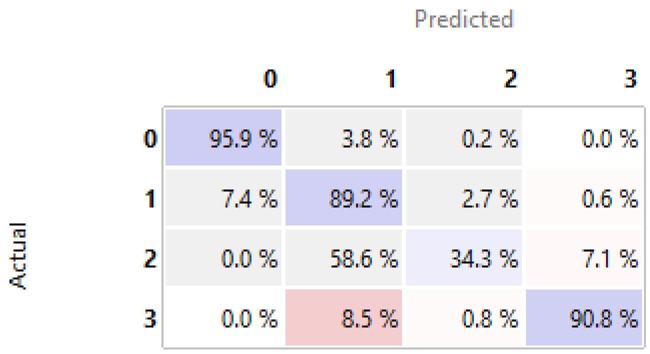
Table 4: Models' prediction precision.

The best model to predict the input data is the neural network model. In fact, if we alternate the model parameters, for example, give the model 200 maximum iterations, the model can actually predict the result with 100% accuracy. However, this is not possible for model such as logistic regression because there isn't any parameter to alternate and improve precision.

Nevertheless, all four models are very good as they can predict the input data with at least 87.5% precision.

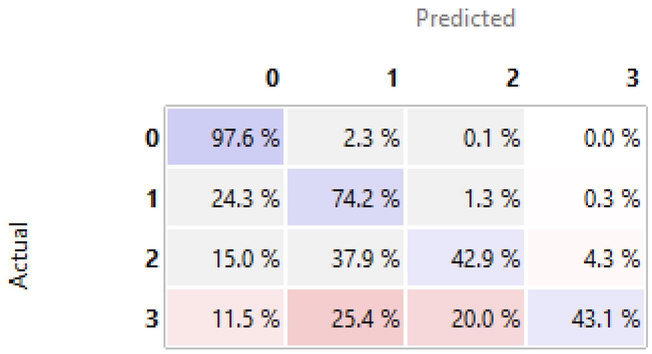
Precision Visualisation

Even though we drew the conclusion that prediction models are accurate, we also want to visualise how models predict data correct and incorrect. By doing this, we can find out which classes are most confusing for the models to predict.


Predicted
Actual

	0	1	2	3
0	95.9 %	3.8 %	0.2 %	0.0 %
1	7.4 %	89.2 %	2.7 %	0.6 %
2	0.0 %	58.6 %	34.3 %	7.1 %
3	0.0 %	8.5 %	0.8 %	90.8 %

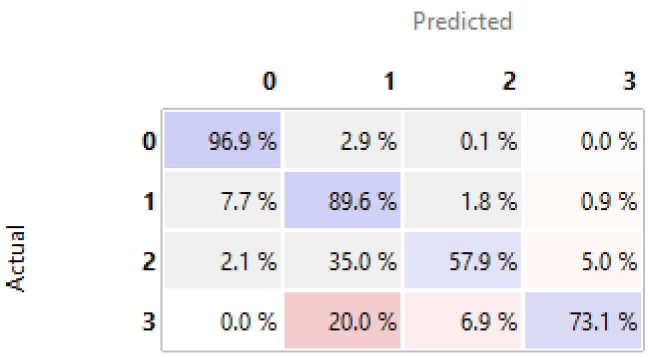
Logistic Regression


Predicted
Actual

	0	1	2	3
0	97.6 %	2.3 %	0.1 %	0.0 %
1	24.3 %	74.2 %	1.3 %	0.3 %
2	15.0 %	37.9 %	42.9 %	4.3 %
3	11.5 %	25.4 %	20.0 %	43.1 %

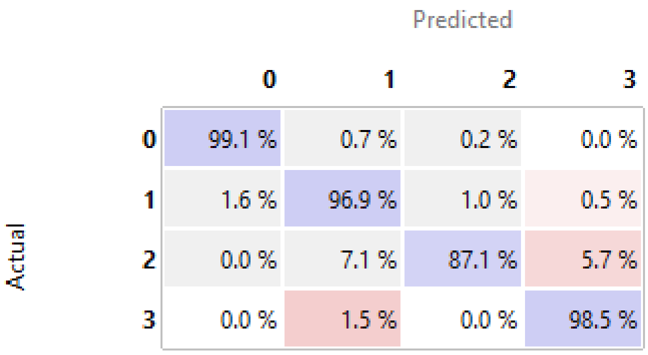
KNN

Figure 8: Confusion matrix, each entry represents the proportion of actual



		Predicted			
		0	1	2	3
Actual	0	96.9 %	2.9 %	0.1 %	0.0 %
	1	7.7 %	89.6 %	1.8 %	0.9 %
	2	2.1 %	35.0 %	57.9 %	5.0 %
	3	0.0 %	20.0 %	6.9 %	73.1 %

Random Forest



		Predicted			
		0	1	2	3
Actual	0	99.1 %	0.7 %	0.2 %	0.0 %
	1	1.6 %	96.9 %	1.0 %	0.5 %
	2	0.0 %	7.1 %	87.1 %	5.7 %
	3	0.0 %	1.5 %	0.0 %	98.5 %

Neural Network

Figure 8 (Continue): Confusion matrix, each entry represents the proportion of actual

For all models except neural network, the most difficult classes to classify is class 1 and 2 (acceptable and good). A large proportion of class 2 data has been predicted to be class 1. Especially for logistic regression model, more incorrect prediction than correct prediction is not acceptable.

Additionally, for KNN model, predicting class 3 (very good) seems to be difficult as well. 56.9% of the labels are predicted incorrect. Also for random forest model, 20% of the observations have been classified as class 1 instead of class 3. As we suggested in the disadvantages of these two models, KNN and random forest can be easily impacted by noise, which can be the cause of inaccuracy.

To draw a conclusion, the neural network model is the best among the four. We can use this model to predict the acceptability of a new vehicle model.

DIMENSION REDUCTION AND PREDICTION

Data Transform

As we suggested, even though there are 6 features in this dataset, we can categorise 6 features into 3 boarder categories:

- Cost
1. Buying Price
2. Maintenance Cost
- Size
1. Capacity
2. Luggage Space
3. Doors
- Safety

Therefore, we can transform this 6-features data into a 3-features data using the following dictionary (size feature is transformed and combined in a similar approach):

Buying Price	Maintenance cost	Cost
low	low	0
low	med	1
low	high	2
low	vhigh	3
med	low	4
med	med	5
med	high	6
med	vhigh	7
high	low	8
high	med	9
high	high	10
high	vhigh	11
vhigh	low	12
vhigh	med	13
vhigh	high	14
vhigh	vhigh	15

Table 5: Cost transformation dictionary

A reason for doing this is to save space and computational cost. It might not be very efficient in this case, but if the data contains millions of entries and hundreds of columns, we can save space and cost by reduce the dimensionality. But we also want to know, can we still predict the acceptability correctly?

Even though it's not used here, we can also assign a specific weight to each feature to balance the importance. For example:

$$X_{cost} = w_1 * X_{buying\ price} + w_2 * X_{maintenance\ cost}$$

$$X_{size} = w_1 * X_{capacity} + w_2 * X_{doors} + w_3 * X_{luggage\ space}$$

where w_i is the weight to balance feature importance.

Principle Component Analysis(PCA)

In addition to manually combined features together, we can also reduce the dimension of data using PCA. In this study, we have 6 features, meaning we have a 6D space. Using PCA, we can project the data along 3D space with the largest variance, to produce a 3D data. Thus, we can produce data in the same dimension as the manually combined data and compare the prediction accuracies.

However, to be noticed, for the manually transformed data, all features are categorical. But PCA can only output numeric data. It is a slightly unfair to compare them even though they are same dimension, as using numeric features might generate better result.

Comparison

Prediction Result

Workflow and models' parameters are equivalent.

Model	Data	Precision
Logistic Regression	Manually Transformed	0.937
	PCA	0.806
KNN	Manually Transformed	0.760
	PCA	0.986
Random Forest	Manually Transformed	0.893
	PCA	0.978
Neural Network	Manually Transformed	0.991
	PCA	0.840

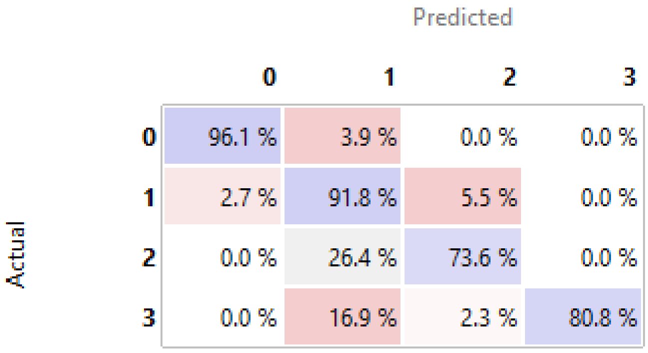
Table 6: Prediction accuracies comparison

All models predict the result well, and we cannot conclude which approach is better. For the manually transformed data, neural network is the best model. It seems neural network will be the best model when it comes to categories prediction. However, for PCA data, KNN is the best model to use.

To conclude, we might use different models here to predict the acceptability of a vehicle based on how we reduce the dimension.

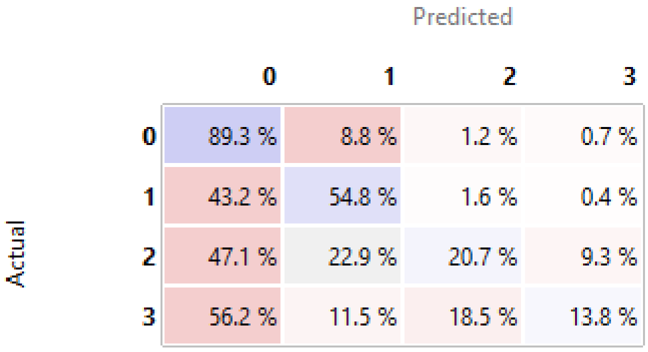
Prediction Visualisation

Transformed Data:



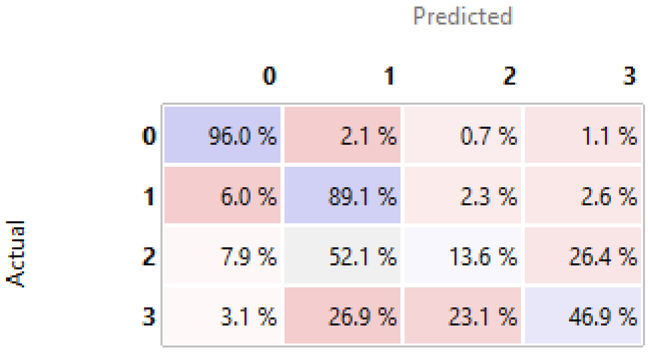
		Predicted		
Actual	0	1	2	3
0	96.1 %	3.9 %	0.0 %	0.0 %
1	2.7 %	91.8 %	5.5 %	0.0 %
2	0.0 %	26.4 %	73.6 %	0.0 %
3	0.0 %	16.9 %	2.3 %	80.8 %

Logistic Regression



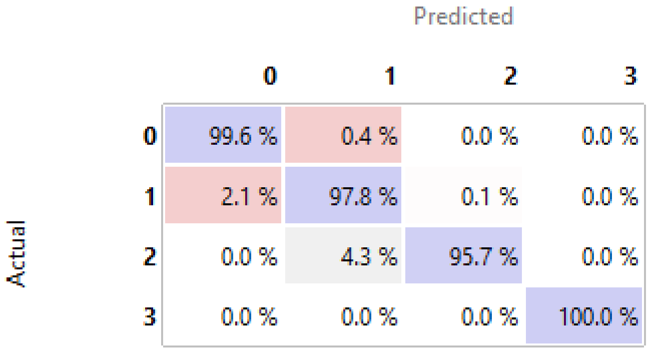
		Predicted		
Actual	0	1	2	3
0	89.3 %	8.8 %	1.2 %	0.7 %
1	43.2 %	54.8 %	1.6 %	0.4 %
2	47.1 %	22.9 %	20.7 %	9.3 %
3	56.2 %	11.5 %	18.5 %	13.8 %

KNN



		Predicted		
Actual	0	1	2	3
0	96.0 %	2.1 %	0.7 %	1.1 %
1	6.0 %	89.1 %	2.3 %	2.6 %
2	7.9 %	52.1 %	13.6 %	26.4 %
3	3.1 %	26.9 %	23.1 %	46.9 %

Random Forest



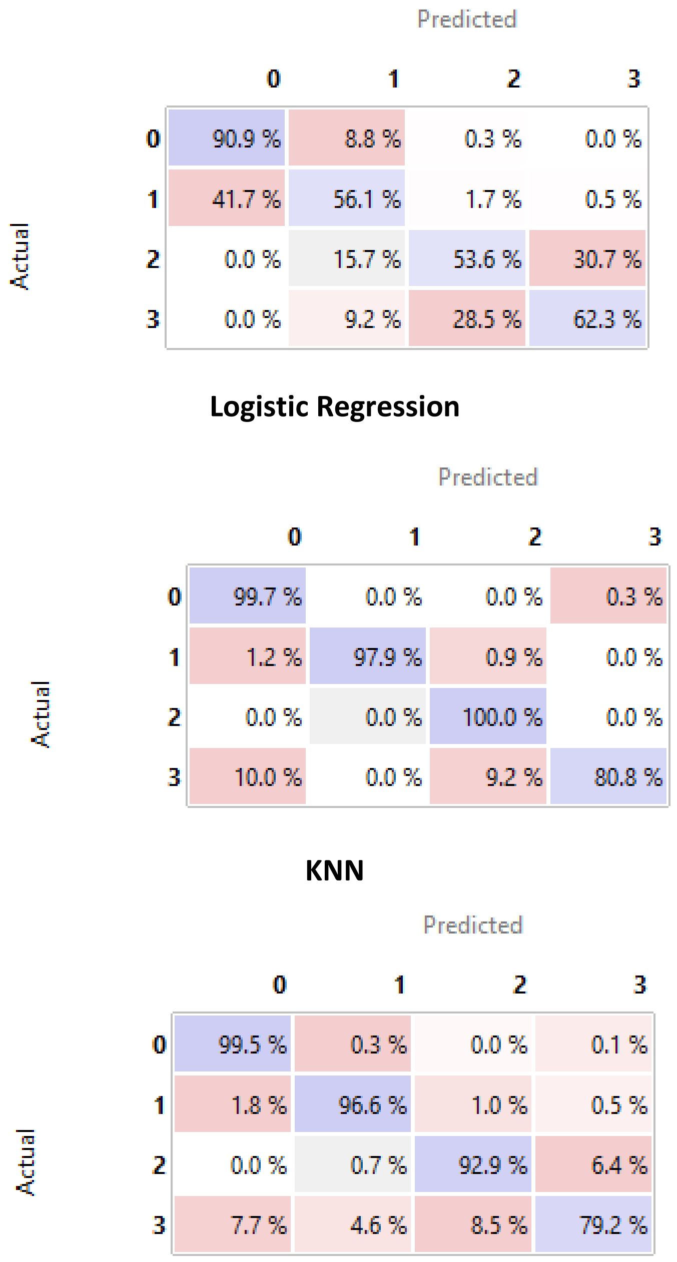
		Predicted		
Actual	0	1	2	3
0	99.6 %	0.4 %	0.0 %	0.0 %
1	2.1 %	97.8 %	0.1 %	0.0 %
2	0.0 %	4.3 %	95.7 %	0.0 %
3	0.0 %	0.0 %	0.0 %	100.0 %

Neural Network

Figure 9 (continue): Confusion matrix of the predictions on the transformed data

Instead of neural network model, all other three models have lost accuracy on prediction. Especially random forest model, which cannot properly classify class 2 and class 3. For the KNN model, a large proportion of data has been classified as class 0 incorrectly. Thus, we can reduce the dimension in this way as using neural network to predict the acceptability is still efficient.

PCA:



Actual \ Predicted	0	1	2	3
0	90.9 %	8.8 %	0.3 %	0.0 %
1	41.7 %	56.1 %	1.7 %	0.5 %
2	0.0 %	15.7 %	53.6 %	30.7 %
3	0.0 %	9.2 %	28.5 %	62.3 %

Logistic Regression

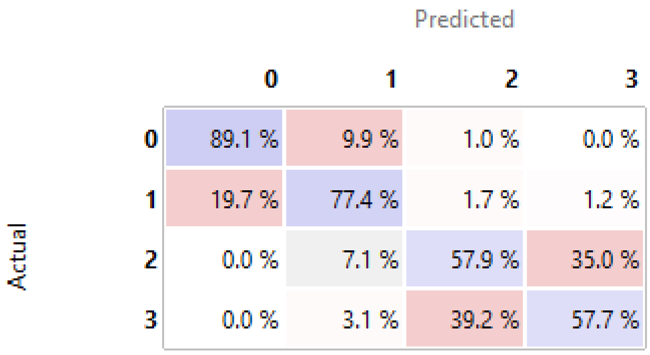
Actual \ Predicted	0	1	2	3
0	99.7 %	0.0 %	0.0 %	0.3 %
1	1.2 %	97.9 %	0.9 %	0.0 %
2	0.0 %	0.0 %	100.0 %	0.0 %
3	10.0 %	0.0 %	9.2 %	80.8 %

KNN

Actual \ Predicted	0	1	2	3
0	99.5 %	0.3 %	0.0 %	0.1 %
1	1.8 %	96.6 %	1.0 %	0.5 %
2	0.0 %	0.7 %	92.9 %	6.4 %
3	7.7 %	4.6 %	8.5 %	79.2 %

Random Forest

Figure 10: Confusion matrix of the predictions on the PCA data



		Predicted			
		0	1	2	3
Actual	0	89.1 %	9.9 %	1.0 %	0.0 %
	1	19.7 %	77.4 %	1.7 %	1.2 %
	2	0.0 %	7.1 %	57.9 %	35.0 %
	3	0.0 %	3.1 %	39.2 %	57.7 %

Neural Network

Figure 10 (continue): Confusion matrix of the predictions on the PCA data

As suggested, KNN is the best model to predict the data transformed using PCA, it can predict every class with at least 80% accuracy. Also, random forest is also a good model to use. However, for both logistic regression and neural network, they classified class 2 and class 3 data with a bad accuracy. Hence, KNN or random forest is the best option to use in this case. By using these 2 models, we didn’t lost a lot of information.

CONCLUSION

In this study, we learnt about the relationship between 6 features proposed and the acceptability of the vehicle model. We investigated the data and found out that the most important feature for customers is the estimated safety, while the general cost of the vehicle is also an important factor to consider.

We also built statistical models to predict the acceptability of customers. When we have a new idea about vehicle designing, we can input the features variables in and get an estimated customers’ acceptability.

REFERENCE

[1]: [https://www.smmt.co.uk/vehicle-data/car-registrations/] [Accessed on 13/01/2021]

[2]: [https://www.smmt.co.uk/wp-content/uploads/sites/2/Car-regs-best-sellers-Dec-2020.png] [Accessed on 13/01/2021]

[3]: [https://archive.ics.uci.edu/ml/datasets/Car+Evaluation] [Accessed on 11/10/2021]

APPENDIX

Appendix A: Coursework1.py

Used to read raw data and data pre-processing. Also includes class distribution and logistic regression coefficient evaluation.

Appendix B: Coursework2.py

Dimension reduction.

Appendix C: classification.ows

Orange file, workflow of processed data.

Appendix D: transformedclassification.ows

Orange file, workflow of manually transformed data.

Appendix E: PCAclassification.ows

Orange file, workflow of PCA data.