2021
Sem. 2

INMT5526
Business Intelligence
Team Assignment

Price Prediction of Used BMW 1 Series & 3 Series

TUESDAY 4PM – BMW #1
WARREN DU (22166202)
ISAAC HUANG (23019722)
BIYING WANG (22400062)
RACHEL XU (23023155)

Table of Contents

1. INTRODUCTION

We are a small family business specialising in dealing second-hand cars in London. As a second-hand car seller, we are not only focusing on selling skill and service quality compared to those brand-new car sellers. To predict the price of those second-hand cars accurately is also necessary regarding our daily operation. It ensures we can purchase the cars at a reasonable price and sell them at a competitive price to maximise our profit. However, based on the financial report last year, our operational expense was too high and one of the highest costs was hiring the automotive mechanics to evaluate second-hand cars. In the UK, an experienced automotive mechanic's average salary is reaching £35,000 a year (UK National Careers Service, 2021).

Therefore, we decided to introduce a price predicting system to reduce the workload of our automotive mechanics, so that we can reduce the expense through hiring fewer automotive mechanics. BMW 1 Series and 3 Series are our best sellers and have the most data. Hence, we decided to use these data to build our initial machine learning model.

2. SUMMARY OF THE DATA SET

2.1 Dependent & Independent Variables

Our independent variables include year, mpg, mileage, fuelType, tax, engineSize, model, transmission. The dependent variable is the price range, where category 1 is price between £0 and £10,000, category 2 is price between £10,001 and £20,000, category 3 is price between £20,001 and £30,000, and category 4 is price greater than £30,000.

Price Range Category	
1	Price between £0 - £10,000
2	Price between £10,001 - £20,000
3	Price between £20,001 - £30,000
4	Price above £30,000

Table 1. Price range categories (dependent variable).

2.2 Data Preprocessing

2.2.1 Missing or Zero-Value Data

In our dataset, there is no missing data. However, in the data of engine size, there are a few instances filled with zero, whereas engine size cannot be zero. Therefore, we replaced zero with the mean value, 2.

2.2.2 Noisy Data

We removed "tax" from our dataset for the two reasons below:

1. By using the correlation function in Excel, =correl(A, B), there is a weak correlation between "tax" and the dependent variable, which is only 0.327.

2. As shown in figure 1, we can see the maximum value of tax (555), and the minimum (0) are both on the same line (price range = 1), which does not seem to be beneficial to our model building either. A lot of reasons can cause the tax value to be ununiformed, like tax discount or false claim on the tax form.

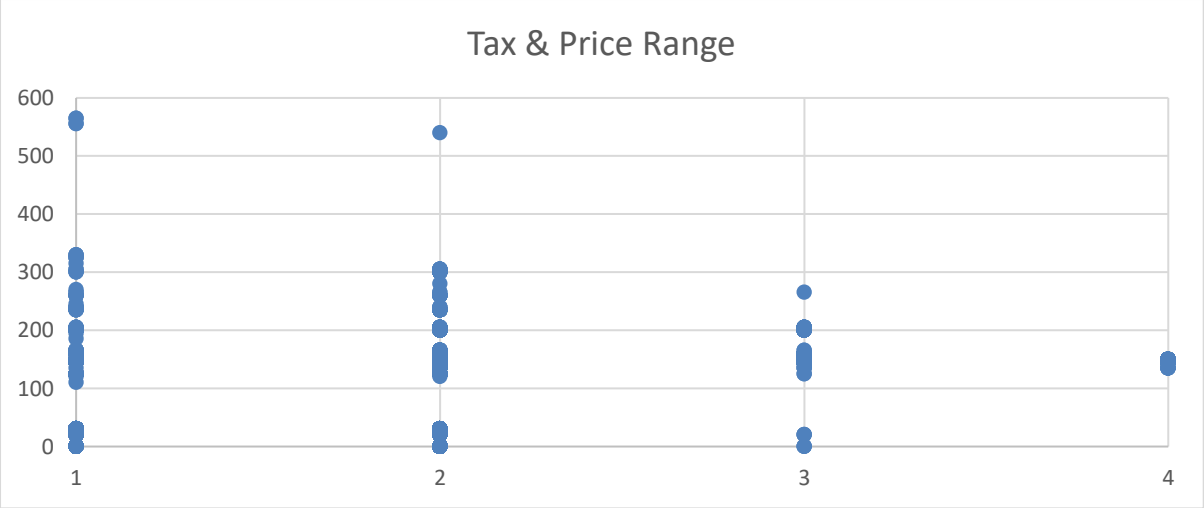

Figure 1. Scatter plot of tax and price range.

2.3 Dashboard User Guide

In the dashboard, we provide a visual graph that can intuitively predict the price range of second-hand cars. When customers use them, provide the second-hand car-related parameters that they need to predict, such as mpg, year, model, etc., so that they can intuitively see the existence of their own vehicles. Price range. When a salesperson uses the model to estimate, he can choose to estimate the relevant parameters of the vehicle, such as model, year, mpg, fueltype, etc., and visually check the number of the same model in the price range, so as to have an estimate of the price range of the vehicle.

When in use, the user can select the known parameters of the vehicle and use the slicer to set it. After the setting is completed, the change trend of the price range can be seen through the line graph, and the bar graph can be used to know the difference between the different series. Comparison of the number of sections and the total number of vehicles in the section, in order to carry out a simple price assessment of the vehicle price.

Of course, the evaluation parameters of used car prices should be from more perspectives. Here, we only use existing parameters to evaluate the range of second-hand cars. After determining the range, you can use other parameters such as appearance and car condition to give a reasonable value in the range.

Through the allocation of intervals, the chart can be displayed more clearly and concisely, and customers can choose according to the price range they want, thereby improving the search efficiency. According to the chart, we can accurately see the degree of correlation between used cars and various independent variables. Year and mpg are highly correlated with used car prices, while engineSize, model, and transmission are relatively low.

When second-hand car dealers use the dashboard, we tend to let second-hand car dealers be newer to the year, and vehicles with high mpg values provide higher price

predictions in the corresponding price range, because we can see in model1 Output year and mpg have a greater impact on the correlation of prices.

3. MODEL SELECTION

3.1 Testing Data & Training Data

We used the function RAND() to create a column of random numbers between 0 and 1, then sorted out these numbers to randomly distribute the order of our data. We kept the first 20% of data, which is the first 882 randomly selected instances, as the testing data, and used the other 80% of instances as our training data.

1	model	year	price	transmi	mileage	fuelTyp	tax	mpg	engineS	Price Ra	Rand()
2	3 Series	2019	23980	Manual	2926	Diesel	145	56.5	2	3	1.576E-05
3	3 Series	2012	6995	Manual	81000	Diesel	30	62.8	2	1	0.0002757
4	3 Series	2015	10999	Manual	39245	Diesel	20	68.9	2	2	0.0005977
5	3 Series	2019	33995	Semi-Auto	4953	Hybrid	135	8.8	2	4	0.000755
6	1 Series	2017	14397	Manual	9243	Petrol	145	52.3	1.5	2	0.0007759
7	1 Series	2017	20481	Manual	25000	Petrol	145	36.2	3	3	0.0010354
8	3 Series	2019	22869	Semi-Auto	11824	Diesel	145	54.3	2	3	0.0010661
9	3 Series	2017	13495	Semi-Auto	44247	Diesel	145	67.3	2	2	0.0012063
10	1 Series	2013	7295	Manual	71000	Diesel	30	64.2	2	1	0.0014404
11	1 Series	2017	15792	Semi-Auto	8789	Petrol	30	55.5	1.5	2	0.0020763
12	3 Series	2019	21450	Semi-Auto	11706	Diesel	150	55.4	2	3	0.0021039
13	3 Series	2015	15100	Semi-Auto	34500	Petrol	160	46.3	2	2	0.0021468
14	3 Series	2017	22480	Semi-Auto	31421	Diesel	150	53.3	3	3	0.0022014
15	1 Series	2016	14998	Manual	23879	Diesel	30	62.8	2	2	0.0025006
16	1 Series	2018	14012	Manual	17143	Petrol	150	38.7	1.5	2	0.0027449
17	1 Series	2013	9700	Manual	58689	Petrol	145	47.9	1.6	1	0.0028104
18	3 Series	2016	17950	Automatic	40337	Diesel	145	53.3	3	2	0.0035903
19	1 Series	2014	9650	Manual	54314	Diesel	30	64.2	2	1	0.0036819
20	3 Series	2018	21498	Semi-Auto	20369	Diesel	145	60.1	2	3	0.0039626

Figure 56. Performing the function RAND() to randomly distribute our data.

3.2 Model 1 & Model 2

We are going to use **MLF neural networks** to train and compare our models. Firstly, we used all the variables except “tax” to train our first model. As illustrated in Fig 1, we can see “engineSize”, “model”, and “transmission” has the least impact to the model, less than 4%. In our second model, we removed those 3 variables to reduce the complexity of the model. We are going to compare our first model (with all variables except “tax”) and our second model (with only the top 4 variables in the Relative Variable Impacts Chart).

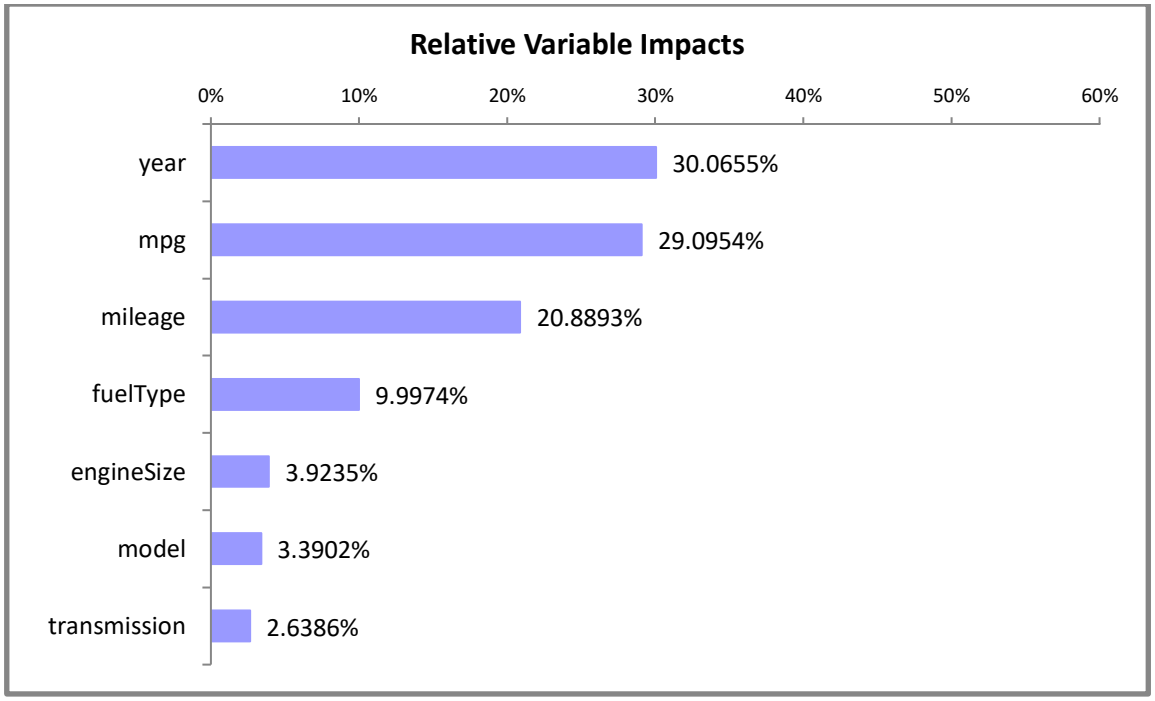

Figure 57. Relative Variable Impact chart of first model.

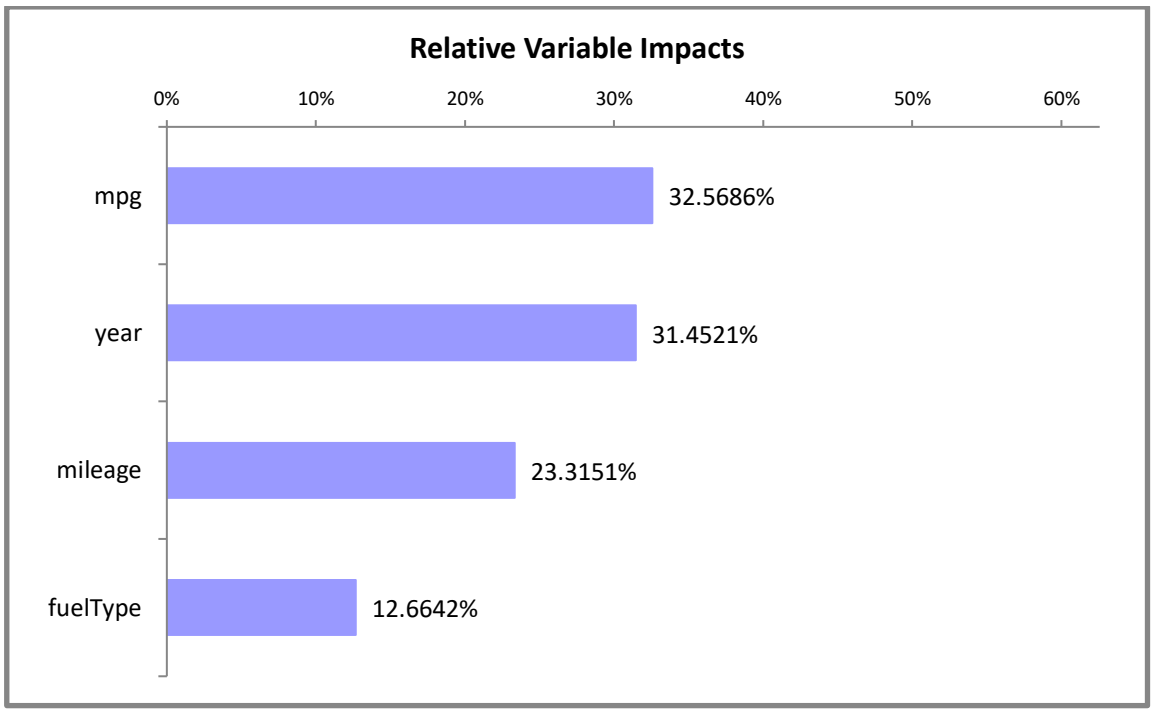

Figure 58. Relative Variable Impact chart of second model.

With complexity reduction in our second model, apart from avoiding overfitting and redundancy, it also leads to better human interpretations and less computational cost (Ghasemi et al., 2018). However, we still want to get a robust model with higher accuracy rate for our second-hand car valuation; thus, we are going to do K-fold Cross-Validation on both models to see which model has better fitness and accuracy (Gunasegaran & Cheah, 2017).

3.3 Model Validation

With k-fold cross validation, first we had to decide what our k is. There is a bias-variance trade-off associated with the choice of k. When the variance decreases, the bias of the model becomes smaller, and vice versa. There are no set rules for what k is (Kuhn Johnson, 2013). However, the choice of k is usually 5 or 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance (James et al., 2013). Plus, since we don't particularly have a big dataset, we chose k = 5 here for our cross validation.

We split our training data into 5 equal folds. Each time we used 1 fold of data to test the model and used the rest to train the model. We repeated this process until each fold of the 5 folds had been used as the testing set.

Table 2 illustrates both training and testing accuracy with 5-fold cross validation of model 1 being higher than model 2, and they are also all very close to the training accuracy before cross validation, 88.53%. Furthermore, as shown in Figure 59 and Figure 60, we can see a smoother line from model 1 than model 2, which means the accuracy rates of each 5-fold cross validation on model 1 are more uniform than model 2, indicating a more robust model.

Model 1	**CV1**	**CV2**	**CV3**	**CV4**	**CV5**	**Mean**	**Before CV**
Training Accuracy	88.77%	88.63%	88.53%	88.39%	88.85%	**88.63%**	**88.53%**
Testing Accuracy	87.39%	88.81%	87.96%	88.09%	87.09%	**87.87%**	
Model 2	**CV1**	**CV2**	**CV3**	**CV4**	**CV5**	**Mean**	**Before CV**
Training Accuracy	81.23%	81.23%	81.98%	80.74%	81.55%	**81.35%**	**81.13%**
Testing Accuracy	80.31%	81.44%	78.90%	83.55%	80.28%	**80.90%**	

Table 2. Accuracy table of 5-fold cross validation.

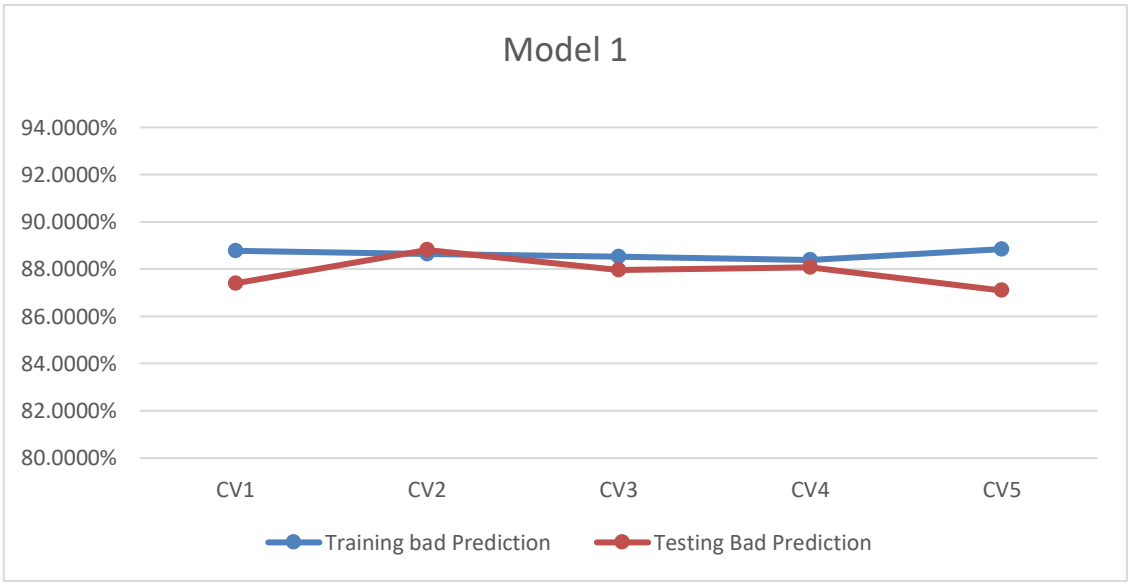

Figure 59. Accuracy rates of 5-fold cross validation on model 1.

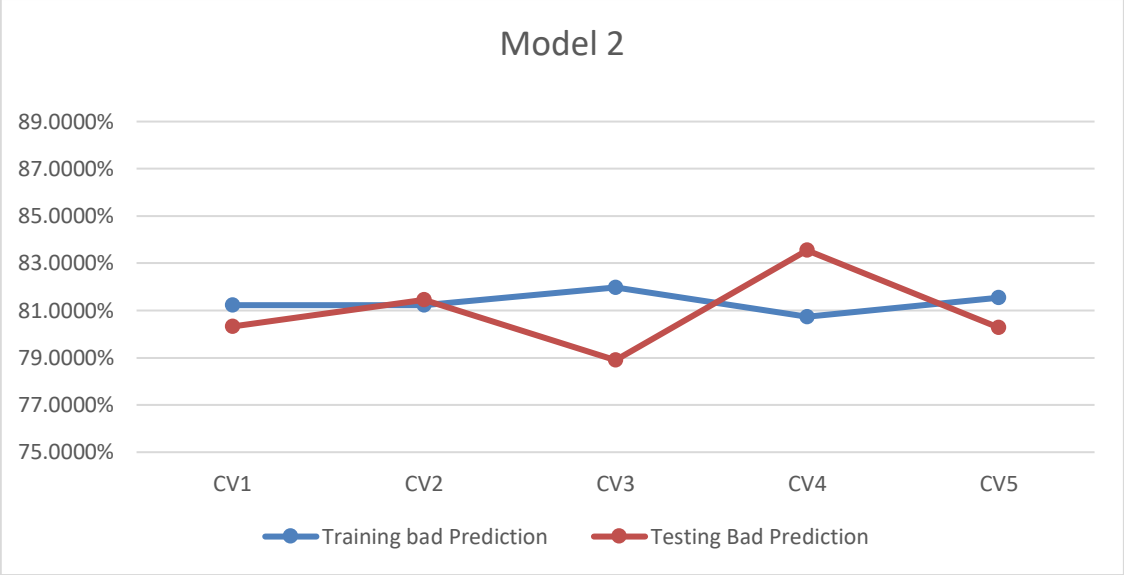

Figure 60. Accuracy rates of 5-fold cross validation on model 2.

3.4 Model Testing

Here, we are going to test our models from the 20% of data we held from the beginning. Since our models have never seen this dataset before, it will help us evaluate the performance of our models.

Table 3 shows the testing accuracy is similar to the training accuracy on both model 1 and model 2, and model 1 still has higher accuracy than model 2.

	Training	***Testing***
Model 1	88.53%	87.42%
Model 2	81.13%	81.18%

Table 3. Training and testing accuracy table of model 1 and model 2.

Table 4 illustrates the confusion matrices of model 1 and model 2. We can see without the contribution of those 3 low-impact variables, “engineSize”, “model”, and “transmission”, the misclassification of price range 4 on model 2 is 100%.

Model 1 Confusion Matrix					
	1	2	3	4	Bad (%)
1	77	16	0	0	17.2043%
2	18	460	25	0	8.5487%
3	0	24	184	13	16.7421%
4	0	0	15	50	23.0769%

Model 2 Confusion Matrix					
	1	2	3	4	Bad (%)
1	71	22	0	0	23.6559%
2	18	446	39	0	11.3320%
3	0	22	199	0	9.9548%
4	0	0	65	0	100.0000%

Table 4. Confusion matrices of model 1 and 2.

3.5 Recommended Model

From the model training, validating, and testing above, they all suggest model 1 is a more robust model with higher accuracy than model 2. Hence, we recommended model 1 as our machine learning model for tackling our problem and performing future predictions. Model summary is shown as Table 5.

Model Summary

Net Information	
Name	**Net Trained on _Model 1**
Configuration	**MLFN Category Predictor (3 nodes)**
Independent Category Variables	**4 (model, year, transmission, fuelType)**
Independent Numeric Variables	**3 (mileage, mpg, engineSize)**
Dependent Variable	**Category Var. (Price Range)**

Table 5. Summary of recommended machine learning model.

4. PREDICTION

We went on https://www.motors.co.uk/, one of the biggest second-hand car websites in the UK, on 6th Oct 2021 and arbitrarily selected 5 instances from BMW 1 Series cars and 5 instances from BMW 3 Series cars, as shown in Table 6. We are going to use our trained MLF neural net to predict the price range of each instance.

model	year	transmission	mileage	fuelType	mpg	engineSize	Price	Prediction
1 Series	2016	Automatic	17000	Diesel	68	1.6	£ 16,995	
1 Series	2020	Semi-Auto	3925	Diesel	62	2	£ 27,805	
1 Series	2012	Automatic	59000	Diesel	62	2	£ 9,150	
1 Series	2018	Semi-Auto	23642	Petrol	47	1.6	£ 25,143	
1 Series	2018	Semi-Auto	22628	Petrol	38	3	£ 28,333	
3 Series	2007	Manual	98916	Petrol	33	2.5	£ 7,995	
3 Series	2019	Semi-Auto	14625	Petrol	50	2	£ 29,000	
3 Series	2015	Semi-Auto	53539	Diesel	67	2	£ 14,500	
3 Series	2018	Semi-Auto	17975	Diesel	51	3	£ 28,321	
3 Series	2018	Semi-Auto	20602	Diesel	44	3	£ 29,976	

Table 6. Real world second-hand car data derived on 6th Oct 2021.

After applying our model to predict the price range, we can see 9 out of 10 instances are correctly classified, as shown in Table 7. The accuracy is 90%, which is close to our training accuracy, 88.53%, and testing accuracy, 87.42%.

As we arbitrarily chose these 10 instances, if all the data on motors.co.uk website are uniformly distributed, which can mean all the car prices are evaluated by the same automotive mechanic, then theoretically we should get a similar outcome on predicting another 10 or 100 instances from that website. Therefore, we only demonstrate 10 predictions here.

model	year	transmission	mileage	fuelType	mpg	engineSize	Price	Prediction
1 Series	2016	Automatic	17000	Diesel	68	1.6	£ 16,995	2
1 Series	2020	Semi-Auto	3925	Diesel	62	2	£ 27,805	3
1 Series	2012	Automatic	59000	Diesel	62	2	£ 9,150	1
1 Series	2018	Semi-Auto	23642	Petrol	47	1.6	£ 25,143	2
1 Series	2018	Semi-Auto	22628	Petrol	38	3	£ 28,333	3
3 Series	2007	Manual	98916	Petrol	33	2.5	£ 7,995	1
3 Series	2019	Semi-Auto	14625	Petrol	50	2	£ 29,000	3
3 Series	2015	Semi-Auto	53539	Diesel	67	2	£ 14,500	2
3 Series	2018	Semi-Auto	17975	Diesel	51	3	£ 28,321	3
3 Series	2018	Semi-Auto	20602	Diesel	44	3	£ 29,976	3

Table 7. Prediction result.

5. RECOMMENDATIONS AND CONCLUSIONS

As our model predicts a price range shown as below. For example, when our predicted outcome is 3, the only thing we would know is the evaluation price is between £20,001 and £30,000, which will not be the final selling price.

Price Range Category	
1	Price between £0 - £10,000
2	Price between £10,001 - £20,000
3	Price between £20,001 - £30,000
4	Price above £30,000

The reason behind how we designed our model is because there are still some other factors that also need to be considered to really reflect the best-selling price. The other factors are:

1. External Cosmetic Value: scratch, sunken, rust or paint peel off on the car surface.
2. Colour: the less popular colour is more difficult to sell, and the price is generally lower.
3. Accessories and Upgrades: the car configuration, stereophony, navigation, etc.
4. Internal Conditions: unpleasant smell, cigarette or pet traces inside the car.

Without the assistance of the automotive mechanics, we can still evaluate these 4 factors above ourselves and get an appropriate evaluation price. The result will be more accurate than only taking age of cars or mechanic conditions into account.

However, there are also some limitations of this prediction method. This model is only trained for BMW 1 Series and 3 Series, and other types of cars are not suitable for using this model. For some unpopular cars in our dataset, their data sizes are too small to build a good predicting model. Therefore, the assistance from the automotive mechanics is still required for selling other models of cars. Finally, the factors that determine the price might not be constant and it will change overtime, such as the consumer preference, the effect of the fashionable style, change in lifestyle, etc.

In conclusion, it is always not an easy thing to operate a company. But by using business intelligence in a proper method, it allows more businesses to operate at a more smart and efficient level.

6. REFERENCES

(UK National Careers Service, 2021)
UK National Careers Service. Motor Mechanic Job Profile. (2021) Retrieved from https://nationalcareers.service.gov.uk/job-profiles/motor-mechanic

(Ghasemi et al., 2018)
Ghasemi, F., Mehridehnavi, A., Perez-Garrido, A., & Perez-Sanchez, H. (2018). Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov. Today, 23(10), 1784-1790.

(Gunasegaran & Cheah, 2017)
Gunasegaran, T., & Cheah, Y. N. (2017, May). Evolutionary cross validation. In 2017 8th International Conference on Information Technology (ICIT) (pp. 89-95). IEEE.

(Kuhn et al., 2013)
Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

(James et al., 2013)
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.