Price Prediction of Used Cars Using Machine Learning

Chuyang Jin
University of Sydney, Sydney NSW 2006, Australia
cjin4945@alumni.sydney.edu.au

2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT) | 978-1-6654-3531-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESIT53460.2021.9696839

***Abstract*—This paper aims to build a model to predict used cars' reasonable prices based on multiple aspects, including vehicle mileage, year of manufacturing, fuel consumption, transmission, road tax, fuel type, and engine size. This model can benefit sellers, buyers, and car manufacturers in the used cars market. Upon completion, it can output a relatively accurate price prediction based on the information that users input. The model building process involves machine learning and data science. The dataset used was scraped from listings of used cars. Various regression methods, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression, were applied in the research to achieve the highest accuracy. Before the actual start of model-building, this project visualized the data to understand the dataset better. The dataset was divided and modified to fit the regression, thus ensure the performance of the regression. To evaluate the performance of each regression, R-square was calculated. Among all regressions in this project, random forest achieved the highest R-square of 0.90416. Compared to previous research, the resulting model includes more aspects of used cars while also having a higher prediction accuracy.**

Keywords—price prediction, machine learning, random forest

I. Introduction

A. Background Information

According to Gareth Roberts of FleetNews UK[1] , it is undeniable that cars have become increasingly significant in modern life. Although Henry Ford had significantly brought down new cars' sticker prices with an assembly line in 1913, purchasing a car was never easy for many. With people's rising demand, the concept of used, second-hand, or pre-owned cars have emerged, denoting a vehicle with at least one owner in the past. Presently, with less money, people can purchase used vehicles with a higher original MSRP.

Unlike purchasing new cars from dealerships, purchasing used ones can be tricky. Firstly, used cars can be bought from auctions and private sellers aside from dealerships. They are likely to be priced differently depending on seller type. Secondly, there is no MSRP for used cars, as sellers primarily decide the price. Sellers publish a used car's condition to the public with a corresponding price tag, then the purchasers either buy it directly or negotiate the price before purchasing. Thirdly, used vehicles' conditions are relatively complicated. Considering mileage, transmission type, engine size, and several other aspects, the prices of used cars may vary. Even two identical model vehicles can be priced differently, especially if they have different mileages.

Conventionally, buyers expend much time and effort on the market to assess used vehicles. They check the vehicle's model, year of production, condition, and mileage, among others. Afterwards, they negotiate the price with the sellers. Eventually, after all of those steps, they can make their purchase. As such, they can even purchase overpriced vehicles because they may not be familiar with the reasonable price range of a specific used car. Indeed, such a process can be disorganised and tricky. Overall, it will benefit both buyers and sellers if there is a model for predicting a reasonable salvage value of used cars.

B. Purpose

This project aims to provide a model that can predict a used vehicle's price using machine learning. With this model, vehicle purchasers will determine a particular used car's reasonable price given certain conditions. In this sense, people are less likely to purchase an overpriced car. Meanwhile, a reasonable price can also be beneficial for sellers. Based on the model's prediction, sellers can set prices either higher if they are not in much hurry to sell the car or lower if they need funds right away. Consequently, both sellers and buyers can save much time and effort selling or searching second-hand vehicles in the market.

Furthermore, the proposed model can depict used vehicles' depreciation over the years. It helps consumers decide which model to purchase should they want to sell it sometime in the future. Moreover, car manufacturers like Mercedes-Benz, Toyota, and Honda can learn which model should be produced more if they intend to be competitive in the used cars market.

C. Excepted Result

Upon completion of this project, a model will be produced. This model will predict used cars' value considering their year of production, mileage, tax, mpg, and engine size. With this model, users can input the relevant information of a particular car, after which the model will output a corresponding price after calculation. An R-square score will be applied to evaluate the model's accuracy; a score above 0.9 will be ideal for ensuring a feasible model.

II. Related works (literature review)

The Danh Phan had a study entitled *Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia*. [2] In the paper, machine learning techniques analysed the historical data of house transactions in Australia, after which a model was built to predict property price. Consequently, the author found a significant discrepancy in house prices between the most expensive and most affordable suburbs in Melbourne. The topic was far different from property price; however, the regression, models, and methodology could be utilised as references. The author analysed various property-based aspects including location and condition, then constructed a prediction model

978-1-6654-3531-4/21/$31.00 ©2021 IEEE

based on these features. In this research, some aspects of used cars can be applied as features or indicators to build the model.

K. Samruddhia and Dr R.Ashok Kumar proposed *Used Car Price Prediction using K-Nearest Neighbor Based.* [3] The paper obtained historical data from the internet and analysed it using the K nearest neighbour-based model. The authors attempted to split the train and validation data into different ratios to create different outcomes. Their best outcome had a prediction accuracy of 85%. Finally, they validate the model with 5 and 10 folds using K Fold Method. They also applied linear regression, garnering a result of 71%. While this result is not relatively low, other more suitable models may exist. Despite KNN and linear regression, other models are worth trying.

Ö Çelik, UÖ Osmanoğlu proposed *Prediction of The Prices of Second-Hand Cars* in 2019. [4] They utilised car data from the internet to establish a model. They mainly employed linear regression and split the data at different ratios (70–30% and 80–20%). Their best R-square outcome was 89.1%, which is relatively high. They applied R-square to indicate the model's performance, which seemed better than prediction accuracy since it could use more information of the model. However, their research included only three factors: price, model, and year of production. In this case, there is some room for improvement; other aspects like mileages and transmission can also be considered.

III. Material and Methods

A. Material and dataset

The dataset used was from Kaggle user Aditya [5], a CSV dataset containing 100,000 UK used cars' scraped data. Each row had a specific used vehicle listed on the market. Each column stood for one aspect of the vehicle, including model, year, selling price, transmission, mileage, fuel type, tax, miles per gallon (mpg), and engine size. The vehicle data were split into car manufacturers, including Audi, BMW, Ford, Hyundai, Mercedes, Toyota, Vauxhall, and Volkswagen. This project used the Mercedes part, containing 13,120 rows of Mercedes vehicles information. When constructing the network, this project intended to consider all nine factors. A commonly used and straightforward method of processing this form of dataset is linear regression and decision tree. The LSTM (Long short-term memory) is also applicable since the dataset is rich in texts.

B. Data Preprocessing

1) Data Visualization

It is advisable to have a big picture of the kind of data that this project deals with and the approximate correlations between data. This step can prevent from making major mistakes when generating the network.

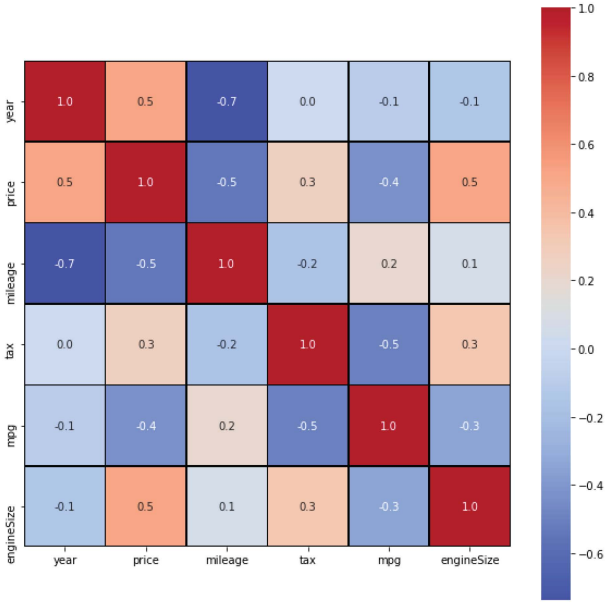

Figure 1. Correlation map.

Figure 1 is a correlation map. In each cell, there is a number between -1 and 1. If the number is positive, the relationship between the two factors is positive; the more it is close to 1.0, the stronger the connection is. The same theory can be applied to negative numbers, but on the opposite side. Take a year of manufacture and price as an example; the value in the year-price cell is 0.5, entailing that these two factors have a positive relationship. When there are no other factors, newer cars have higher price tags.

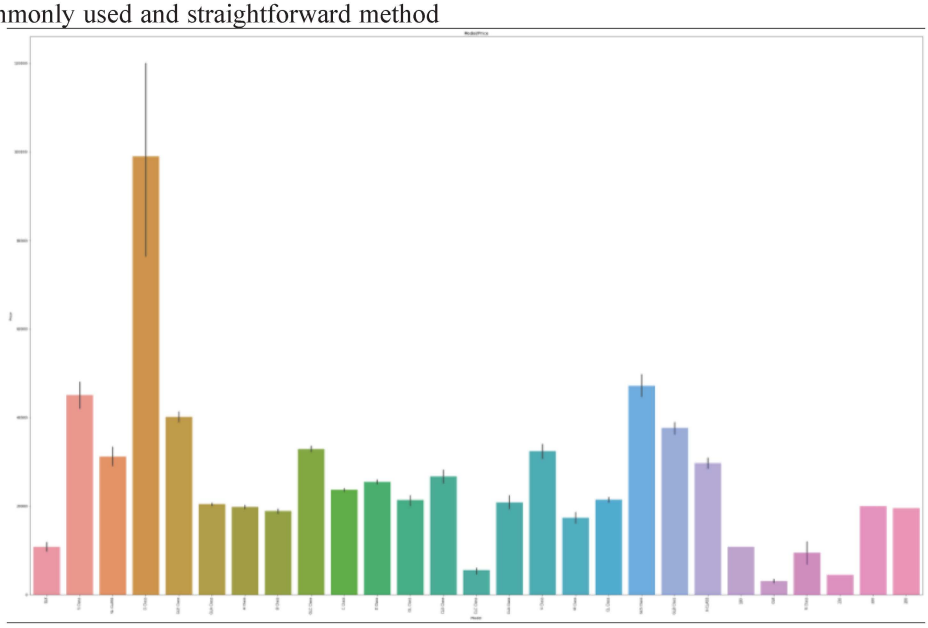

Figure 2. Data visualization.

Figure 2 presents the data visualization. This project set the x-axis as Mercedes models and the y-axis as the sum of model prices. The highest sum price was from the G-class, which was close to 100,000. Meanwhile, the lowest sum price came from the CLK class.

2) Preprocessing

Since the network is based on the dataset, the dataset's quality is significant. Before jumping into the model, several tweaks must be performed to improve data quality.

Firstly, the IsNull function verified whether the dataset contained invalid values; if yes, a data cleaning process would be required. Eventually, an outcome of 0 was obtained in each column, indicating that all values were valid.

Secondly, the data distribution of the prices was validated.

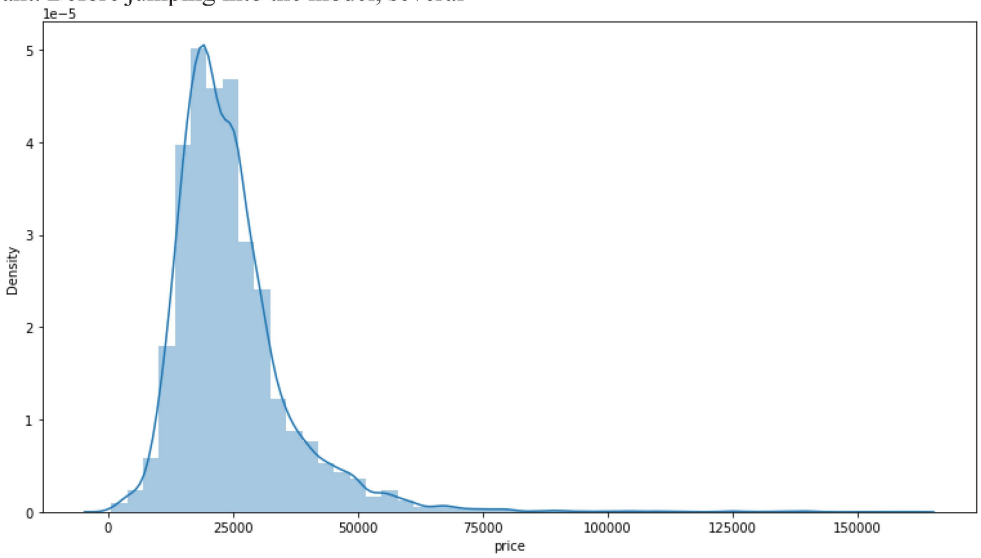

Figure 3. Price distribution.

From Figure 3, the overwhelming majority of price falls in the 0–75,000 range. When the price went higher than 75,000, the sample number dropped significantly. This distribution could interfere with the model's accuracy because some extreme samples would be unnecessary to consider. There were not enough samples in the category of used cars priced over 75,000 to build an effective model based on it. These data are considered outliers, which can affect the model's accuracy. Similar phenomena can be found in other used cars' aspects, including year, mileage, tax, miles per gallon (mpg), and engine size. For example, if only several 1970 used cars for sale are in the dataset, it is far from sufficient to build a model. This problem would not occur regarding car model, fuel type, and transmission because these factors only had a few classes. For instance, there are only three classes for transmission, including automatic, manual, and semi-auto. The same theory applies to the used car model and fuel type. In this case, outlier data were excluded to ensure that the model would be accurate and usable.

The most efficient distribution would be a curve close to the normal distribution. In statistics, the normal distribution is the most prevalent probability distribution since it can suit various realistic situations [6]. One per cent of the margin data was removed from the dataset to improve the model's prediction accuracy. As shown in Figure 4, the distribution would look much better and balanced without the outlier, which might distort the model. After the optimization, the dataset contained 12,988 rows of data.

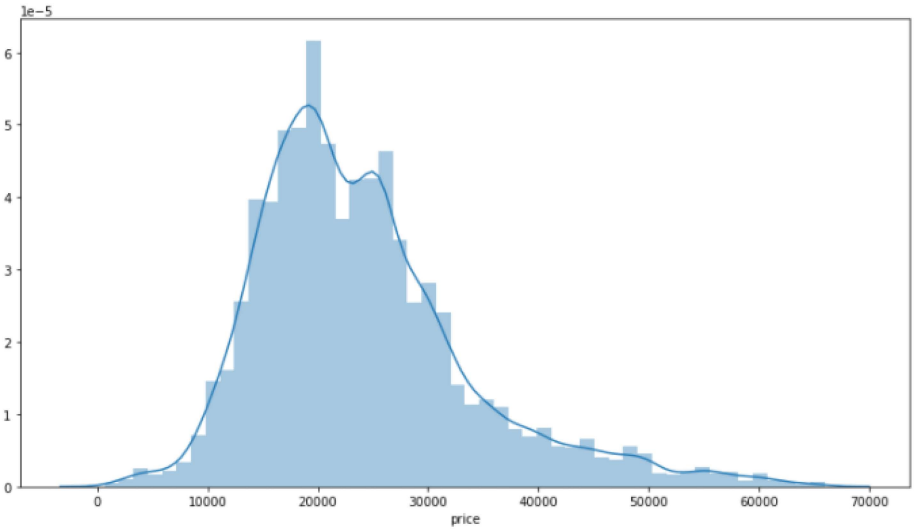

Figure 4. Price distribution after optimization.

Some indicators of the dataset were calculated to analyze the impact of removing 1% of the data.

Data Before Optimization						
	Year	Price	Mileage	Tax	MPG	Engine size
count	12988.000000	12988.000000	12988.000000	12988.000000	12988.000000	12988.000000
Mean	2017.281876	24074.926933	22132.741146	129.689714	55.437142	2.050901
Std	2.228515	9866.224575	21196.776401	65.183076	15.025999	0.532596
Min	1970.000000	650.000000	1.000000	0.000000	1.100000	0.000000
25%	2016.000000	17357.500000	6322.000000	125.000000	45.600000	1.675000
50%	2018.000000	22299.000000	15369.500000	145.000000	56.500000	2.000000
75%	2019.000000	28706.000000	31982.250000	145.000000	64.200000	2.100000
max	2020.000000	65990.000000	259000.000000	580.000000	217.300000	6.200000

Figure 5. Data before optimization.

Data After Optimization						
	Year	Price	Mileage	Tax	MPG	Engine size
count	13119.000000	13119.000000	13119.000000	13119.000000	13119.000000	13119.000000
mean	2017.296288	24698.596920	21949.559037	129.972178	55.155843	2.071530
std	2.224709	11842.675542	21176.512267	65.260286	15.220082	0.572426
min	1970.000000	650.000000	1.000000	0.000000	1.100000	0.000000
25%	2016.000000	17450.000000	6097.500000	125.000000	45.600000	1.800000
50%	2018.000000	22480.000000	15189.000000	145.000000	56.500000	2.000000
75%	2019.000000	28980.000000	31779.500000	145.000000	64.200000	2.100000
max	2020.000000	159999.000000	259000.000000	580.000000	217.300000	6.200000

Figure 6. Data after optimization.

According to the figure 5 and figure 6 above, the dataset had not dramatically changed after the removal. For some critical indicators like standard deviation and count, the changes were hardly noticeable. In this project, the negative impact of removing 1% margin data was minimal, which could make the data distribution more balanced.

This research found that the mean price value of used cars from 1970 was abnormally high when exploring the data. It reached 24999.000000, while the mean value of 1997 while only 9995.000000. It could result from lacking adequate data from 1970, thus potentially affecting the model's accuracy. This research calculated the average price by year and applied these numbers to solve the distorts. After the change, the mean price value of used cars from 1970 dropped to 9995.000000, which was much more sensible than the original value of 24999.000000.

C. Model and Methodology

First, this project used the train_test_split function from sklearn to split the dataset. The split dataset was labelled as training dataset and testing dataset. The training dataset was used the train the model and improve the prediction result, while the testing dataset was used to validate the model's accuracy.

This project attempted several regressions to determine which one had the best outcome, specifically which regression would the dataset the best. The R-square was calculated for each regression to evaluate the model's performance. Also known as the coefficient of determination, R-square is commonly used in statistics. In most cases, it is a number between 0 and 1; the larger number, the better. It provides a straightforward measure of how well a model learns the data. After obtaining each regression's R-square, the one with the highest score was chosen.

1) Linear Regression

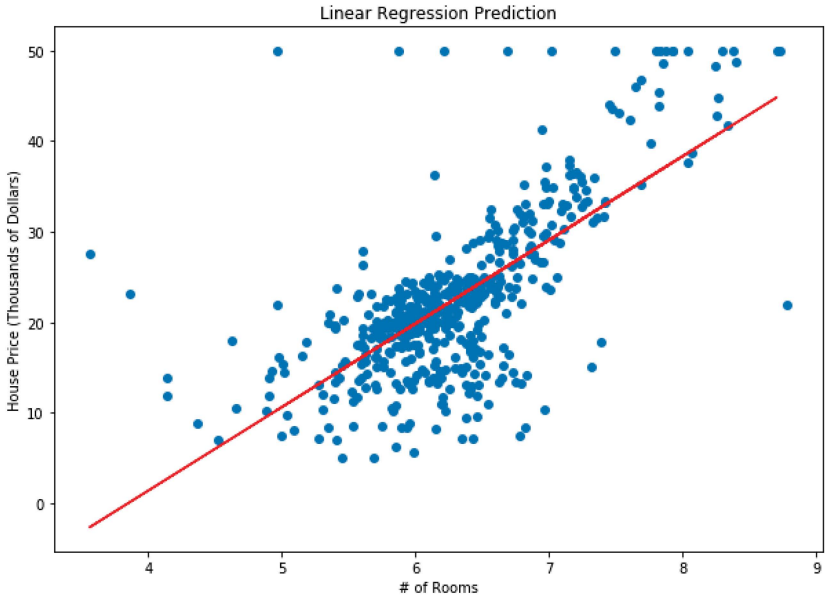

Figure 7. Example of linear regression [7]

Linear regression is one of the most extensively used regressions and the simplest type of regression [8]. As shown in Figure 7, linear regression is relatively straightforward. It tries to model a linear equation between two variables. In this project, the relations involved price-model, price-mileages, and price-transmission. After calculating each relationship and equation, all equations were combined to generate the model. They performed well when two variables had a linear relationship. However, linear regression had its limitations. For instance, it is susceptible to outliers, as an extreme sample can affect the regression. Although this project deleted 1% of margin data in data preprocessing to prevent extreme samples from happening, they remained. Furthermore, linear regression may confuse and perform poorly if two variables have twisted relations with some corners on the relationship curve.

2) Polynomial Regression

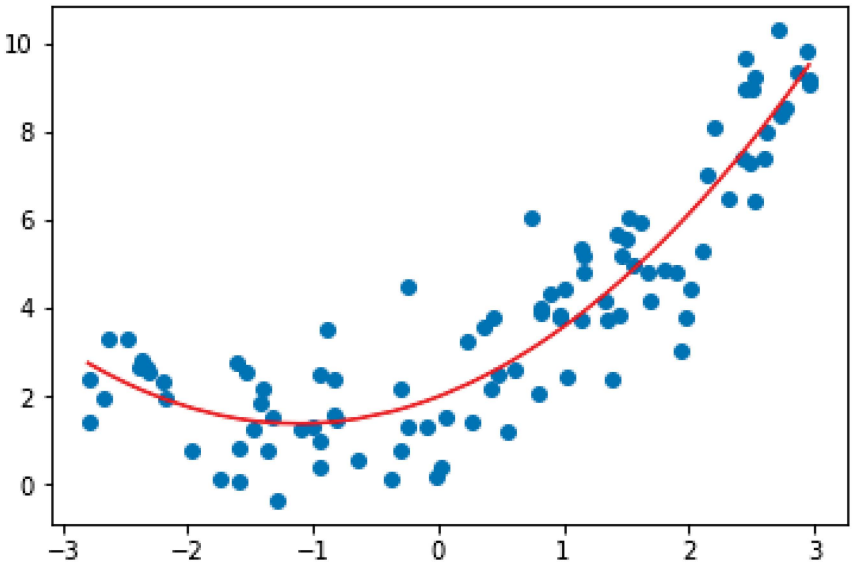

Figure 8. Example of polynomial regression [7]

As opposed to linear regression, polynomial regression has two variables with curved relation equations, as indicated in Figure 8. This regression suits nonlinear cases. One advantage of polynomial regression is that it can fit a vast range of functions. It has a minimal restriction for data distribution and is particularly suitable for providing an approximation of the relationships between the variables. If we properly divide the relationship curve of a polynomial regression, we can determine that each portion of the curve is linear; thus, polynomial regression is also a branch of multiple linear regression. It makes polynomial regression inherit linear regression's disadvantage—that is, very sensitive to outliers. After the deletion of 1% margin data, this error can be mitigated, but not eliminated.

3) Support Vector Regression (SVR)

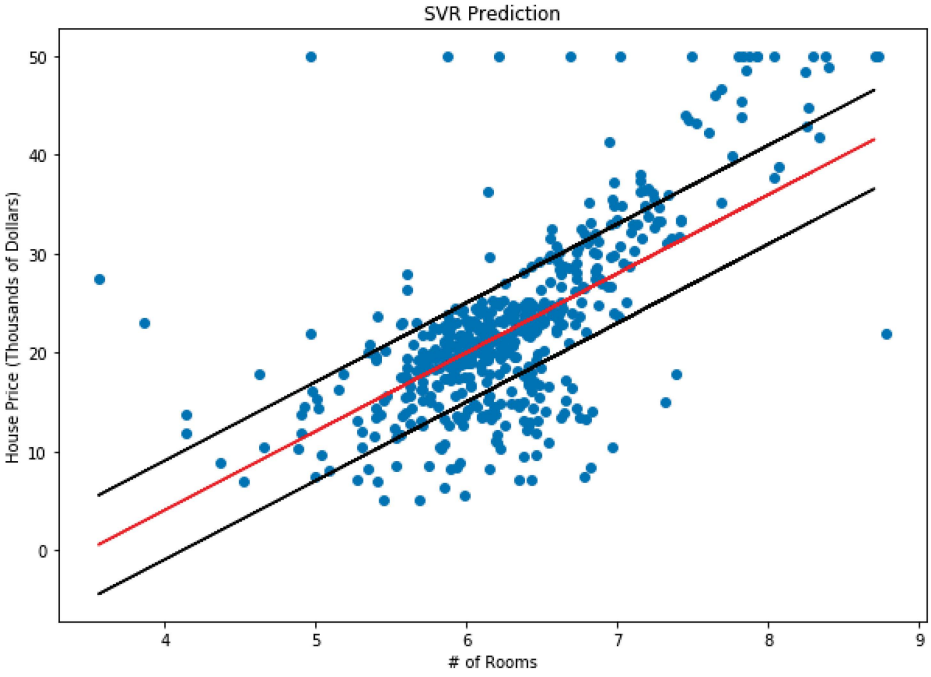

Figure 9. Support vector regression (SVR) [7]

Developed from the support vector machine (SVM) by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman, and Alexander J. Smola in 1996 [9], SVR fits both linear and nonlinear functions. Figure 9 shows aexample of SVR.Unlike linear regression and polynomial regression, SVR can suit errors within pre-set tolerance values. It depends only on some selected proportion of the training data. When training the model, SVR ignores data that are out of margin. As the value of pre-set threshold increases, the model become less sensitive to errors. This feature can solve the disadvantages of linear and polynomial regression since the outliers are considered errors and are not calculated during model training.

4) Decision Tree Regression

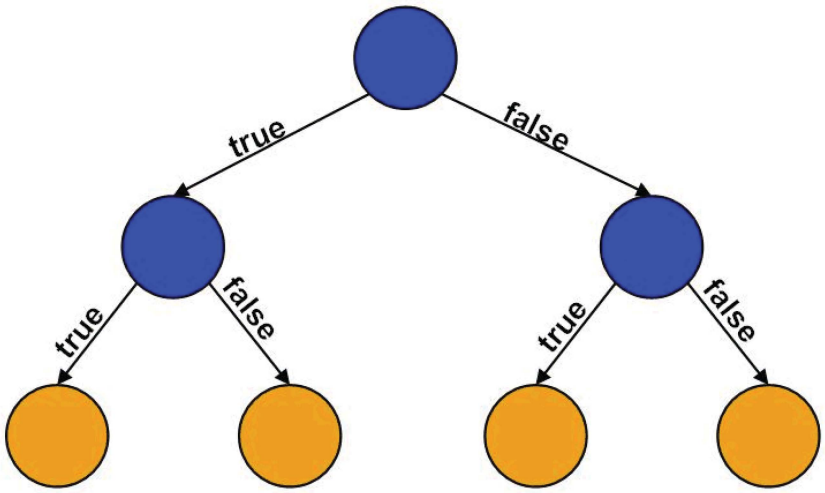

Figure 10. Decision tree.

As the name indicates, decision tree regression builds a model that has a structure similar to a tree. As with Figure 10, it indicates several nodes (or leaves) and branches that reach out from the same start node. The initial point stands for the original dataset. After several levels of splitting, the tree becomes more complicated. Each node stands for one class label, while the branches stand for the combination of nodes.

In this project, decision tree regression has several advantages. It has minimal requirements on data structure and form. Other models must tokenize value and optimize the data distribution in the dataset for training, while the decision tree can take such data easily because it can handle qualitative predictors. It is particularly suitable for large datasets because it splits the data into small packages, enhancing efficiency and accuracy.

However, the decision tree is not perfect. The bottom leaves values' are highly dependent on their predecessor; thus, it is vital to ensure that the data on the top leaves are correct and as accurate as possible. A small error in the top leaves can lead to severe impacts on the final model. Moreover, the regression splits the source dataset into small packages; it is straightforward for the model to have overfitting. The model may perform reasonably well in the training set but perform poorly in actual. This situation can be made worse while dealing with small datasets; when training the data, the model can perceive only a small portion of the data rather than the entire dataset. Thus, the decision tree regression often emphasizes too much on optimal local values while necessitating optimal global values.

5) Random Forest Regression

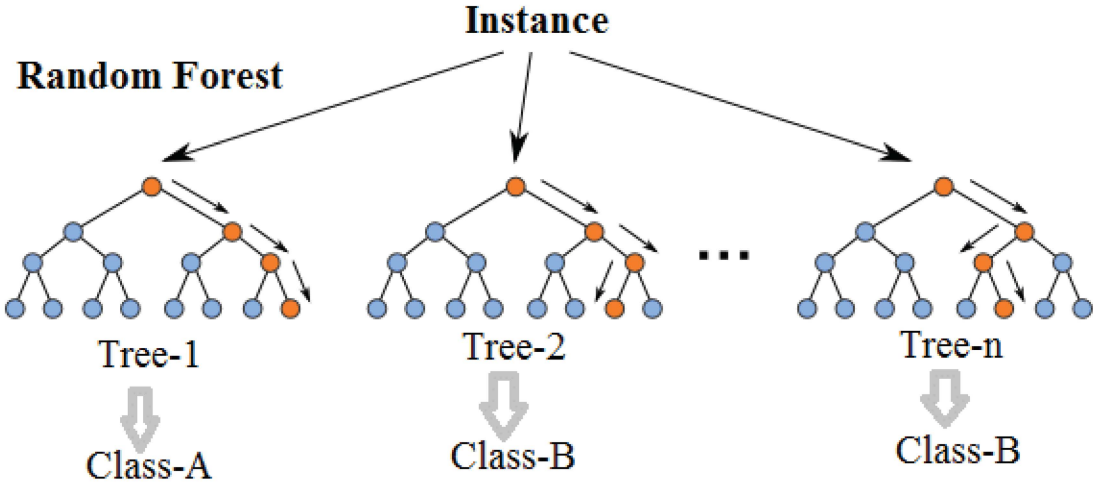

Figure 11. Random forest [10]

Created by Tin Kam Ho [11], random forest regression consists of several individual decision trees. We can tell by the Figure 11 that the regression evolved from decision tree regression; thus, they share many similarities. They both work for linear and nonlinear regression, split the source data into small packages, and are affected by errors.

When making a prediction, the trees in a random forest regression work as an ensemble. Every individual tree in the network produces a prediction based on the input. The prediction with the highest ratio become the model's final prediction. For example, in a model, we have 150 trees; when typing in the input, we have 100 trees that predict the value as 5, while 50 trees predict the value as 10. Then, the model's prediction is 5.

Random forest regression sees the dataset from several perspectives compared to a decision tree that keeps digging the dataset in one direction. This feature eases the central decision tree problem, overfitting the training set [12]. The relationship between trees in the forest is not too strong to affect other trees, but strong enough to make the ensemble outperform any standalone tree. This mechanism makes trees in the forest prevent themselves from individual errors. Even some errors occur in some decision trees, as the majority of other trees tends to ignore the errors. It somehow simulates voting formation in human society. However, the random forest regression is not that interpretable.

This project expects that this regression performs the best because it fits the dataset's requirements and attributes.

IV. Results and Conclusion

A. Result

After fitting the training data to each model, we have the following result.

	Linear Regression	Polynomial Regression	Support Vector Regression (SVR)	Decision Tree regression	Random Forest Regression
R-square	0.72354	0.83127	0.83545	0.85140	0.90416

Figure 12. R-square with an accuracy of 5 decimal points.

Figure 12 shows R-square for five models adopted in the paper. Among all five models in the research, SVR needs some extra work. Scaling is needed before we run the model. The mechanism of SVR is to construct a plane that has the longest distance to the data points. The data point is called the support vector. If features have very different ranges, the more prominent features will dominate other smaller ones when calculating. It is essential to scale them in the same range when constructing the plane to make all features have a similar influence on the model.

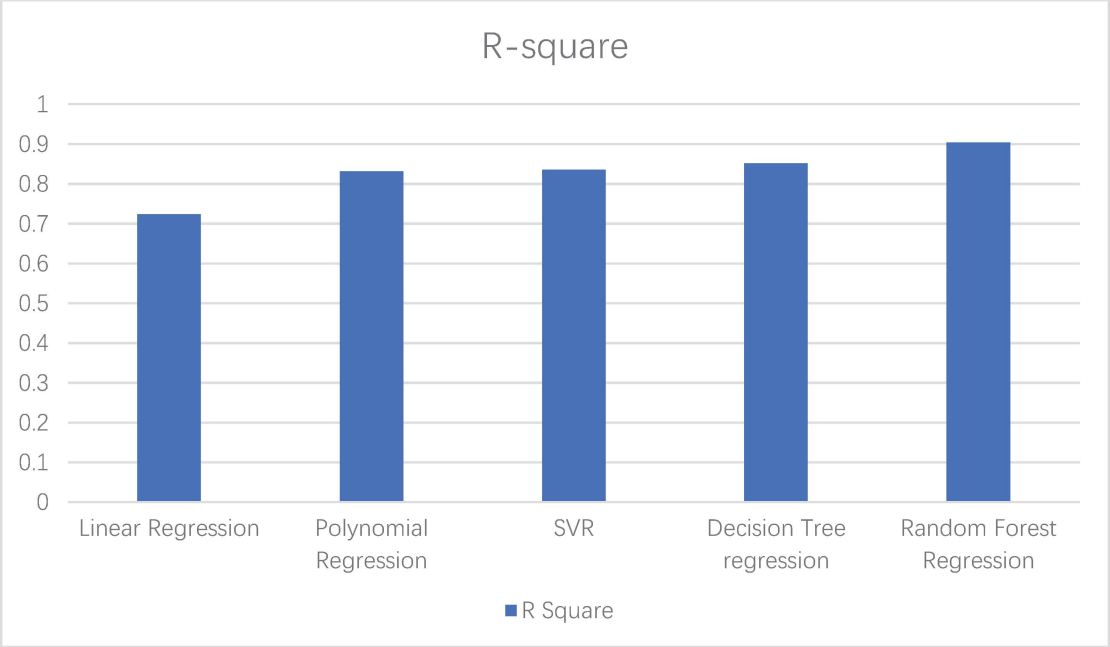

Figure 13. R-square bar chart.

As Figure 13 bar chart indicates, the random forest regression has the highest R-square of over 0.9. It is the only regression that reaches 0.9.

Linear regression has the R-square of 0.72354, the lowest among the five models. It is reasonable because linear regression is most suited to simple linear data; however, this research is not the case. Polynomial regression performs much better than linear regression as a particular case or upgraded version of linear regression. It has an R-square of 0.83127, more than a 10% improvement. Through the different R-square of linear and polynomial regression, we know that data points are more like curves than straight lines. Support vector regression gets an R-square of 0.83545. It is much higher than linear regression and slightly higher than polynomial regression. A similar story can be found on the decision tree regression. It scores at 0.85140, higher than SVR, but with minimal margin. We can find poly regression, SVR, and decision tree regressions have close R-squares. When we moved to random forest regression, the square jumped to 0.90416. It is much higher than all the other four regressions. As an upgrade of decision tree regression, the result is expected. It means that the random forest regression suits the data structure the best and makes the best prediction.

In the last part of the code, this research added the

function of eliminating with elimination Ols. This step aims to remove the redundancy of the model, hence improving the efficiency of the model. It can help the model to output an even better result.

B. Conclusion

This project generated a used car price prediction model using machine learning and neural network. To construct the model, it tried several regressions, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression. This project calculated the R-square for each regression to indicate their performance. Among all five regressions, random forest regression gets the highest R-square of 0.90416. Thus, random forest regression was selected to construct the used vehicles prediction model.

V. Discussion and Future Work

This research tried several models to predict the used car price. It chose the random forest model because it had the highest R-square of around 0.9. However, there is still much room to improve. Other models like Naive Bayes, LSTM, or Gradient Boosting algorithms can be applied to determine whether better results can be obtained.

A more extensive data set is always better. With a larger dataset, more training data can be fed to the model. The dataset used in the research has 12988 samples after data pre-processing. If a larger dataset were obtained, a more accurate model would be expected.

The following work is expected to expand the considered aspect of used cars, including exterior and interior fineness and engine working status. Considering these two factors can be particularly challenging because there is no universally used grading system for them. It is necessary to find a relatively good grading standard or come up with a customized one.

References

[1] Roberts, G., 2020. "Public transport 'failings' increase car dependency". [online] Fleetnews.co.uk. Available at: <https://www.fleetnews.co.uk/news/fleet-industry-news/2020/02/21/public-transport-failings-increase-car-dependency> [Accessed 27 September 2021].

[2] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[3] K. Samruddhi and R. Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model", International Journal of Innovative Research in Applied Sciences and Engineering, vol. 4, no. 3, pp. 686-689, 2020. Available: 10.29027/ijirase.v4.i3.2020.686-689.

[4] Ö. Çelik and U. Ö. Osmanoğlu , "Prediction of The Prices of Second-Hand Cars", Avrupa Bilim ve Teknoloji Dergisi, no. 16, pp. 77-83, Aug. 2019, doi:10.31590/ejosat.542884

[5] https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

[6] Nadarajah S. "A generalized normal distribution"[J]. Journal of Applied statistics, 2005, 32(7): 685-694.

[7] T. Sharp, "An Introduction to Support Vector Regression (SVR)", towardsdatascience.com, 2020. [Online]. Available: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2. [Accessed: 27- Sep- 2021].

[8] C. Liu, "Linear to Logistic Regression, Explained Step by Step - Velocity Business Solutions Limited", Velocity Business Solutions Limited, 2020. [Online]. Available: https://www.vebuso.com/2020/02/linear-to-logistic-regression-explained-step-by-step/. [Accessed: 27- Sep- 2021].

[9] Drucker H, Burges C J C, Kaufman L, et al. "Support vector regression machines"[J]. Advances in neural information processing systems, 1997, 9: 155-161.

[10] C. Yi, "ML Introduction: Random Forest", Medium.com, 2019. [Online]. Available: https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%83-%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-6afc24871857. [Accessed: 27- Sep- 2021].(in Chinese)

[11] Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

[12] Dietterich T. "Overfitting and undercomputing in machine learning"[J]. ACM computing surveys (CSUR), 1995, 27(3): 326-327.