Predicting The Price Range Of Used Cars

John Ian Castaneda II

501083068

Supervisor:

Sedef Akinli Kocak, Ph.D.

Ryerson University

Table of Contents

Abstract

Second-hand goods have lower prices, right? Not always. Right now, prices of pre-owned cars have risen, and they have even been selling for more money than new cars. This is mainly due to the shortages in new cars which is driven by a global semiconductor shortage. People have been less willing to suffer through the long wait times to get a new car and it has been difficult to find the models they like in the car lot. Thus, the demand for used cars have steadily increased ergo the price. But what are major factors affecting the selling price of used cars and how significant is their impact on the price? Can we predict the price of a used car? This project aims to produce an algorithm that will do just that using machine learning. Specifically, this project will train a classification model using pre-pandemic data to predict the price range of a used car. That model will then be used to predict the price of used cars during the pandemic.

Classification models learn how to assign labels to an observation or to categorize that observation into the correct class. It can be used in predictive modeling where the class label is predicted given certain inputs. This project will use three classification techniques to build a reliable model: random forest, XGboost, and neural networks. Random forest is an ensemble method for regression by constructing a large number of trees during training. The model then returns the mean or average prediction of individual trees. Random forests generally outperform decision trees but has a lower accuracy than gradient boosted trees. XGboost is a popular machine learning algorithm and is an implementation of gradient boosted decision trees. Its main benefits are its speed and performance. Gradient boosting works by building weak models sequentially. Each model attempts to predict the errors left by the prior model. Neural networks are a set of algorithms that are modelled after the human brain. It is trained to recognize patterns. It consists of input layers, hidden layers, and output layers.

To train the model, the dataset to be used is the Used Cars Dataset from Kaggle.com. This dataset was first scraped in 2018 from Craigslist, one of the largest classified advertisements

websites in the world. There is a huge number of listings regarding used cars for sale on this site. This data is scraped every few months and was last updated in May 2021. Exploratory data analysis will be performed to discover any obvious relationships that may inform and guide the modelling process. Feature selection will also be done to streamline the model and improve its performance.

Kaggle – Used Cars Dataset:

https://www.kaggle.com/austinreese/craigslist-carstrucks-data

This project will be done using Python to take advantage of the various machine learning packages available.

Literature Review

Predicting the price of used cars is a very practical application of machine learning especially today because it can benefit both buyers and sellers of second-hand cars. For the buyers, they can use this model to gauge whether a second-hand car they are interested in is fairly priced. For sellers, this model helps them determine how much they can sell their car for and maximize the residual value they can get back. Thus, there have been a few published studies that applied machine learning to the prediction of used car prices. These studies used different datasets and used various models from the most common algorithms to more complex ones.

In his thesis, Richardson used multiple regression analysis and showed that hybrid cars, those that have both combustion and an electric motor, maintain higher resale values than regular combustion engine vehicles (Richardson, 2009). This is likely to be truer today as climate change has taken center stage in public discussion and more people are choosing products that are better for the environment. With more brands coming out with their own hybrid and electric models, we have more data today to verify this assertion. All the independent variables considered in the

study namely age, mileage, make, condition, fuel efficiency, safety ratings and market segment played important roles in the resale value of cars.

Listiani on the other hand used Support Vector Machines (SVM) to predict the price of a cars in her Master thesis (Listiani, 2009). The SVM model had a higher accuracy when compared to simple multiple linear regression or multivariate regression. This is due to SVM's ability to better handle data with high dimensionality. It is also less prone to overfitting and underfitting. However, this advantage of SVM over simple regression was not quantified in common statistical metrics.

Pudaruth studied this problem in Mauritius by using various machine learning algorithms namely multiple linear regression, k-nearest neighbours, decision trees and Naïve Bayes (Pudaruth, 2014). The main variable used in his study were make, cylinder volume, year, and mileage. All four models had comparable performance. However, this study had a limited number of samples as they were manually taken from newspaper listings. More samples would help train the models better which could give better price prediction capabilities. The number of variables were also limited.

Gegic and his team did something similar. They tried to predict the price of used cars using various algorithms as well and compared the results. However, they chose artificial neural networks, support vector machine, and random forest. They found that using a single algorithm yielded only less than 50% accuracy. Thus, they decided to ultimately use an ensemble of multiple machine learning algorithms and this significantly improved the accuracy. They did note that this gain in accuracy comes at price as ensemble machine learning algorithms consume a lot of computational resources (Gegic et al, 2019). However, similar to Pudaruth, Gegic's model could possibly benefit from having more samples to train the model.

Pal and his team used random forest after careful exploratory data analysis. Using data from a German subsidiary of eBay, they came up with a model with 500 decisions trees with a testing

accuracy of 83.62% (Pal et al, 2018). They asserted that the most relevant features in their price prediction model were mileage, brand, and vehicle type. However, they did not present how they narrowed down their features to 10 from the original number of 20. Feature selection must be done carefully to reduce the risk of eliminating variables that may help improve the accuracy of the model.

Monburinon and his team used data from the same e-commerce site and used multiple linear regression, random forest and gradient boosted regression trees (Monburinon et al, 2018). These models were compared using the mean absolute error of their results. Gradient boosted regression trees yielded the highest performance, followed by random forest, with multiple linear regression having the lowest performance. Unlike Pal, Monburinon's work did not have age of the car as one of the features. Adding this feature in the model is important because the value of cars depreciate over time. Thus, age of the car is a major factor affecting its price.

These related works show that brand, vehicle type, mileage and age the common features that affect the price of used cars. Our study will verify if this still holds true during the pandemic or if there are other features that have become important factors to determining the price of cars. This study is also unique because it will try to determine whether models that are trained using pre-pandemic data can accurately predict the price of used cars using data during the pandemic. We will study if there is a dataset shift affecting model performance and how we can adapt our model to any dataset shift.

Dataset

The dataset to be used was taken from Kaggle.com. The used cars dataset was first scraped in 2018 from Craigslist. It is updated every few months with the latest data. The last update was in May 2021. There have been 10 versions of this dataset so far. It has been decided that versions 8 and 10 will be used as the pandemic dataset since these versions contain the column "posting_date". This column gives us a clear indication that this data covers used cars that were put up for sale during the pandemic. All the other versions lack this column. Version 10 contains data of used cars for sale during May 2021. Version 8 contains data of used cars for sale during November and December 2020.

For the pre-pandemic data, versions 2 and 3 will be used. Since the dataset was first mined in 2018, we can be reasonably assured that these versions contain data of used cars up for sale before the pandemic started in March 2020. We decided to not use version 1 since it had more missing columns than version 2 and 3 when compared to the latest versions.

The following are the features present in all 4 versions of the dataset:

No.	Feature	Type
1	url	object
2	region	object
3	region_url	object
4	price	int64
5	year	int32
6	manufacturer	object
7	model	object
8	condition	object
9	cylinders	object
10	fuel	object
11	odometer	float64
12	title_status	object
13	transmission	object
14	VIN	object
15	drive	object
16	size	object
17	type	object

18	paint_color	object
19	image_url	object
20	description	object
21	lat	object
22	long	object
23	posting date	object

Figure 1: Variables and the data type of each variable.

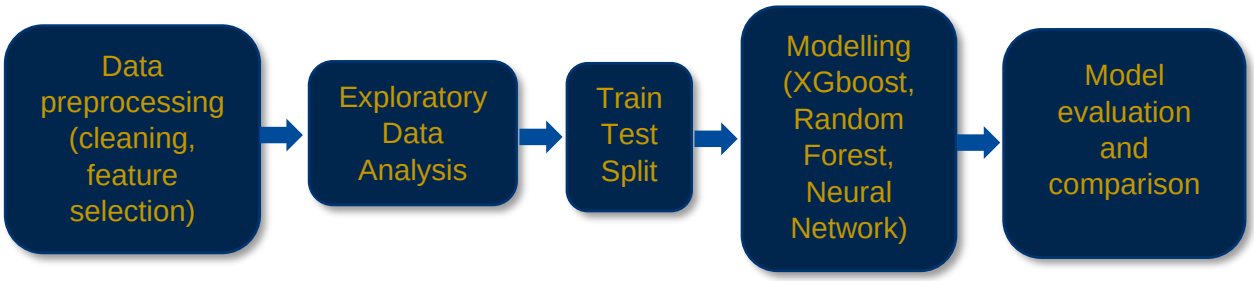
Price, year and odometer (mileage) are the only quantitative variables in the dataset. The table below gives some relevant statistics:

	price	year	odometer
count	1961218	1961218	1692151
mean	68799	2005	101178
std	10856876	103	1583663
min	0	0	0
25%	4399	2007	44372
50%	10000	2012	91500
75%	19995	2016	136770
max	3787876818	2022	2043755555

Figure 2: Summary of statistics pertaining to the data frame.

The combined dataset has 1,961,218 samples with 1,076,152 pre-pandemic data and 885,066 data during the pandemic. The standard deviation and the maximum value for price and odometer indicate that there are outliers that we must investigate and handle during preprocessing.

Methodology



Data Preprocessing

Looking at the combined raw dataset from the four selected versions of the used cars dataset indicates that we have to select the features that will actually affect the price of used cars. Thus, the columns 'url', 'region_url', 'VIN', 'lat', and 'long'. It was initially planned to transform the columns 'image_url' and 'description' into binary variables which would just indicate whether this information is available for a particular record in the dataset. The model would have determined whether having this information would affect the price or not. However, approximately all the rows have this information making these binary variables inconsequential to the model's predictive capabilities.

Missing values were initially handled more conservatively wherein only rows with missing values in various combinations of independent categorical variables were removed. This resulted in a dataset with 861,230 rows. However, due to limitations in time as well as computational power, training the models with this huge amount of data was impractical. Thus, a more aggressive approach to missing values was used wherein all rows that have missing values in any of the independent categorical variables were removed.

Outliers in price and odometer reading were removed using IQR rule. The IQR for price is only $11,800 and the 3rd quartile is $15,900. The upper limit to check for outliers would have been $33,600. Majority

of the used cars in the data would fall into this range. However, it would exclude other car brands, types and models that typically have higher prices such electric vehicles or even SUVs. Thus, the upper limit was extended to 10 times the IQR in order to include higher priced cars in the model. Cars with prices less than $500 were removed as there were a lot of listings with unrealistically low prices such as $10 or even $1.

Professional mechanics say that 12,000 miles per year is an accurate estimate of a car that has not been overdriven. The cars in the dataset have years from 1992 to 2022 which is 31 years. Therefore, the maximum odometer reading expected for the dataset should be 372,000 miles. Rows exceeding this number were removed as outliers.

The data in the column ‘model’ was largely inconsistent. At this stage of preprocessing, the remaining data had 22,577 car models. Sellers sometimes included other information in the field for car model such as the year of the car or other words to entice buyers to purchase their car. To address this large number of car models, rows with car models that have a frequency of less than 150 were removed. There were only 377 models left after this.

This was the same case with the column ‘region’ with the remaining data having 628 regions. Thus, rows with regions that have a frequency of less than 150 were also removed. There were 353 regions remaining.

The price and the odometer reading were then converted to categorical variables to make a classification model that would predict the price range of a used car. There are 28 price ranges from 0 to $135,000 and there are 32 odometer reading ranges from 0 to 372,000 miles.

Independent categorical variables were then one-hot encoded the dependent categorical variable was label encoded.

The final dataset after preprocessing and cleaning contained 194,861 rows. The 14 independent variables were one-hot encoded into 884 independent variables.

Exploratory Data Analysis

In this step, we get an initial understanding of the data by performing investigations on the data. A version of the preprocessed data where the price and odometer columns have not yet been converted to categorical variables was used in this analysis.

The correlation matrix below show that there is a positive correlation between year and price and there is a negative correlation between odometer and price. Year and odometer are not too highly correlated thus we can keep both of them in model.

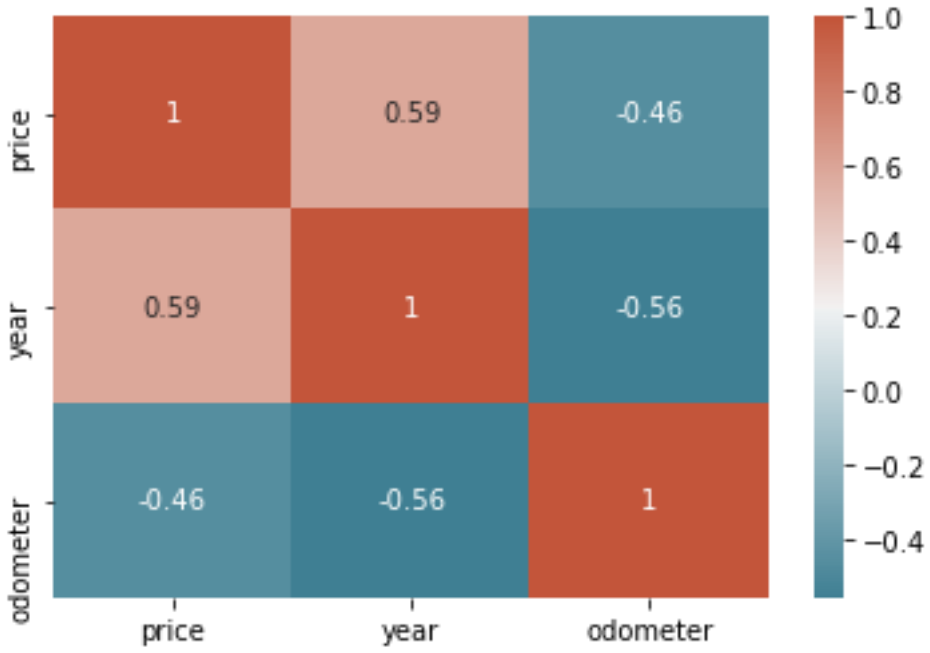


Figure 3: Correlation matrix of price, year and odometer reading.

Scatter plots were also prepared to analyze the relationship between price and the independent variables odometer and year. These plots prove the intuitive idea that cars with higher mileage sell for lower prices and cars manufactured more recently sell for higher prices.

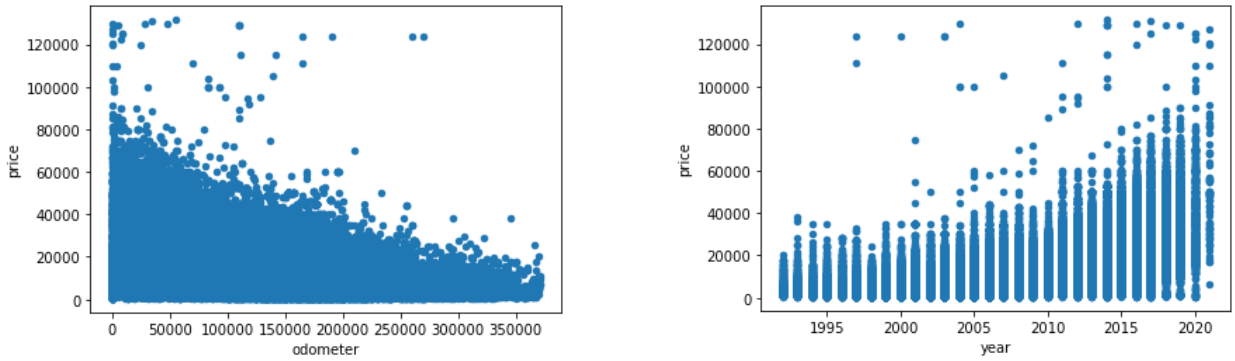


Figure 4: Scatter plot of odometer and price.

*Figure 5: Scatter plot of year and price.*A frequency distribution of the prices was prepared, and it showed that the distribution is skewed to the right with a mean of $12,277 and median of $8,995. Majority of the used cars being sold have prices below $20,000.

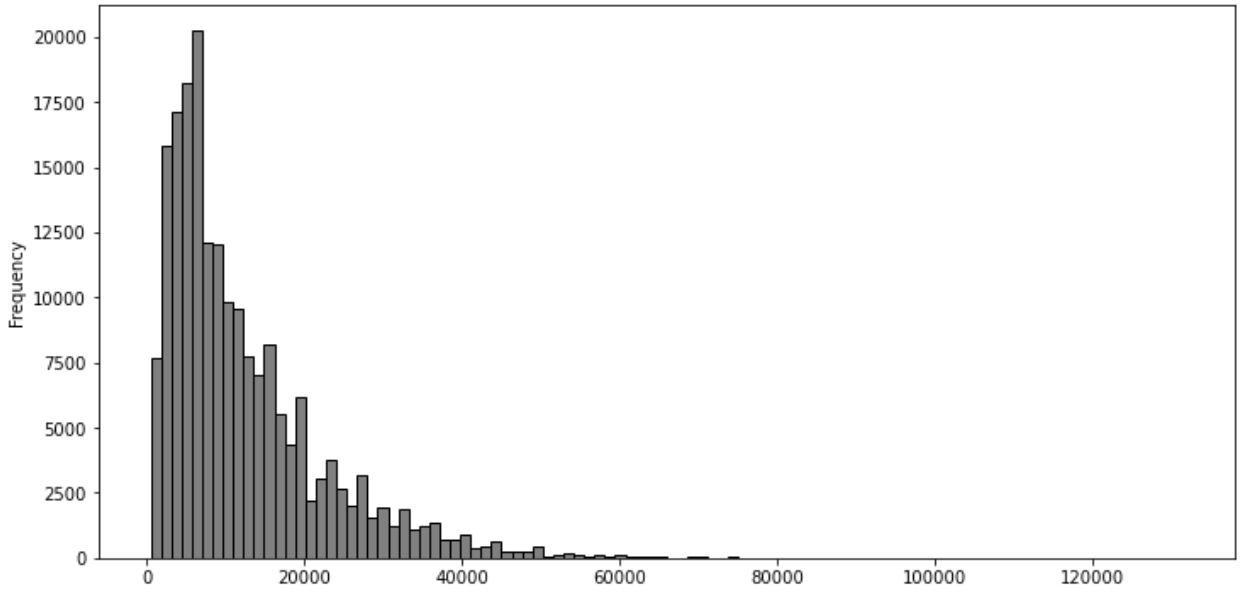


Figure 6: Distribution of price.

Train Test Split

The dataset was first divided into two subsets: the pre-pandemic dataset and the pandemic dataset. This was done through the column ‘posting_date’. Pre-pandemic data had null values in this column but pandemic data had this information indicating when the cars were listed for sale on Craigslist. The pre-pandemic dataset will be used to train the model and evaluate its performance to make sure it is an effective predictor of the price range of used cars. The pandemic dataset will be used to check whether the model trained using pre-pandemic data can correctly predict the price ranges of cars in the pandemic data. A significant decrease in the performance of the model using the pandemic data will indicate that there is a dataset shift.

The pre-pandemic dataset was then divided into two subsets: train dataset and test dataset. The split used was 70-30 with 70% as the train dataset and 30% as the test dataset. The train dataset will be used to build the model. The test dataset will then be used to test the accuracy of the model by comparing the predicted values generated by the model with the expected values.

Modelling

Three models were created to predict the price of used cars. The scikit-learn machine learning library was used to create a random forest model, the XGBoost library was used to create the XGBoost model, and the Keras library was used to create the neural network model. The three models were trained using the same training dataset. K-fold cross validation was performed to evaluate the stability of the models.

Model Evaluation, Comparison

The performance of the models was evaluated using a test dataset to see how the models would perform predicting the price ranges of cars not seen during training. The results were then compared using accuracy, precision, recall, F1 score, and Cohen’s Kappa as criteria. Accuracy is the ratio of correctly predicted observations to the total observations. Precision is the ratio of

correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the total actual positive observations. F1-score is the weighted average of precision and recall. Cohen's kappa measures the proximity of the predicted classes to the actual classes when compared to a random classification. It is one of the best metrics for evaluating multi-class classifiers trained on unbalance datasets. The out put is normalized from 0 to 1. The closer the score is to 1, the better the classifier.

Evaluation focused mainly on F1 score and Cohen's Kappa as they are more reliable measures of multi-class classification problems.

Results

The figure below shows the performance of the three models. The first table show the runtime of the k-fold cross validation and fitting of the model. XGBoost took the longest time during k-fold cross validation with a runtime. of 2hours and 14minutes. Neural networks took the longest time during fitting with a runtime of 26minutes. Random forest was the fastest in both k-fold cross validation and fitting with a runtime of 17minutes and 2minutes, respectively.

	Runtime (minutes)	
	5-fold Cross Validation	**Fitting**
Random Forest	16.91	1.50
XGBoost	194.46	16.64
Neural Networks	126.63	25.99

Figure 7: Training runtimes in minutes.

The figure below show how well the models performed in predicting the price range of the used cars in the test dataset. Neural networks had the highest score across all the metrics with a 76.60% F1-score and Cohen's kappa coefficient of 0.7064. Random forest was not far behind in performance with an F1-score of 76.13% and Cohen's kappa coefficient of 0.6993. XGboost had the lowest scores with and F1-score of 67.60% and Cohen's kappa coefficient of 0.5931.This

means that the neural networks model was the best at predicting the price range of a used car. However, random forest is also an acceptable model to use to predict the price range of a used car given its efficiency in training and close performance to the neural networks model.

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Random Forest	76.26	76.13	76.26	76.13	0.6993
XGBoost	67.87	67.60	67.87	67.60	0.5931
Neural Networks	76.67	76.64	76.67	76.60	0.7064

Figure 8: Model performance in predicting used car price ranges using the test dataset.

The models were then used to predict the price range of cars in the pandemic dataset. Their performance is shown in the table below. Model performance has gone down across the board for all the models. However, it be noted the neural networks still performed the best with an F1-score of 59.39% and Cohen's kappa coefficient of .5032. XGboost performed better than random forest when used on the pandemic data. XGboost had an F1-score of 54.72% compared to random forest which had 53.58%. XGBoost's Cohen's kappa coefficient was 0.4443 while random forest had 0.4317.

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Random Forest	55.31	52.49	55.31	53.58	0.4317
XGBoost	56.20	53.76	56.20	54.72	0.4443
Neural Networks	60.44	58.89	60.44	59.39	0.5032

Figure 8: Model performance in predicting used car price ranges using the pandemic dataset.

The metrics show that there has been a degradation in model performance when the models trained using pre-pandemic data were used on data taken during the pandemic. This degradation is summarized in *Figure 9*. Random forest showed the greatest degradation with a decrease of 22.5% on its F1-score and 0.2676 on its Cohen's kappa coefficient. XGBoost experienced the least loss in performance with a decrease of 12.88% on its F1-score and 0.1488 on its Cohen's kappa coefficient.

	Actual decrease		% decrease	
	F1 score %	Cohen's Kappa	F1 score	Cohen's Kappa
Random Forest	22.55	0.2676	30%	38%
XGBoost	12.88	0.1488	19%	25%
Neural Networks	17.21	0.2032	22%	29%

Figure 9: Degradation in model performance on pandemic dataset.

Model Retraining – Neural Network

The degradation in the performance is a sign of dataset shift. A dataset shift occurs when there is a shift in the distribution of data. There are various ways of addressing this problem but for this paper, we will only employ model retraining. Since neural network had the best classification performance, it was selected the model for retraining. The model was retrained using the combined dataset with not distinction between pandemic and pre-pandemic data. The same split was used to get training set and the test set. The retrained model had an F1-score of 74.71% and Cohen's kappa coefficient of 0.6834. These results are better than that of the model trained solely on pre-pandemic data.

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Neural Networks	74.74	74.78	74.74	74.71	0.6834

Figure 10: Model performance in predicting used car price ranges trained and tested using the combined pre-pandemic and pandemic dataset.

Recommendations and Future Studies

Due to time constraints and computing capabilities, the explorations done in this study was limited. Future studies can consider other algorithms instead of the 3 used in this study. Feature selection after model training can also be done to see if this improves model performance. The columns

‘model’ and ‘region’ had so many unique values because Craigslist does not restrict the input these fields to a standard set of inputs. Future studies could try take more time to clean the data in these columns.

The degradation in the performance of the models showed that there is a dataset shift. The model was retrained to include the pandemic data, and this indeed improved the performance. However, such retraining would not always the most efficient way to solve the problem of dataset shift. Other ways to address this could be explored in future studies.

Conclusion

We were able to predict the price range of used cars using the 3 classification models: random forest, XGBoost, and neural networks. These models were trained using data before the start of the pandemic in March 2020. Out of the 3, neural networks performed the best with the highest F1-score of 76.60% and Cohen’s kappa coefficient of 0.7064. The models’ performance declined when they were use to predict the price range of cars listed during the pandemic. Neural network was retrained with data from both before and during the pandemic. The model’s performance improved when handling pandemic data.

GitHub

https://github.com/jiancii/CIND-820

References

Richardson, M. S. (2009). *Determinants of used car resale value* (Doctoral dissertation, Colorado College.).

Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application. *Unpublished. https://www. ifis. uni-luebeck. de/~ moeller/publist-sts-pw-andm/source/papers/2009/list09. pdf*.

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764.

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), 113.

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How much is my car worth? A methodology for predicting used cars' prices using random forest. In *Future of Information and Communication Conference* (pp. 413-422). Springer, Cham.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018, May). Prediction of prices for used car by using regression models. In *2018 5th International Conference on Business and Industrial Research (ICBIR)* (pp. 115-119). IEEE.