WQD7006: Machine Learning for Data Science
Group Project (30%)

Used Car Price Prediction using Machine Learning

1. Introduction

1.1 Background

During the last decade, the production of automobiles has increased dramatically (Venkatasubbu & Ganesh, 2019). This is due to the significant growth in the demand for and use of cars. Cars have become a more crucial part of people's daily lives than people think. With the rising use of cars, the market for used cars has also grown since they are more affordable than new cars (Totakura & Kosuru, 2021). In 2021, the used car industry was worth USD 260 billion; by 2027, it is predicted to be worth USD 460 billion. The COVID-19 epidemic had a massive impact on the automotive sector. New additions to the used automobile inventory were not entering the market due to supply chain disruption and transportation standstill. However, in the latter half of 2020 and 2021, the market began to revert to pre-pandemic levels, presenting new chances for the global used automobile industry (MordorIntelligence, n.d.).

Used cars, as the name suggests, are cars that one or more users previously owned. The used car market is a growing business with a market value of nearly doubled in the previous few years. The development of online portals like CarDheko, Quikr, Carwale, Cars24, Craigslist, Carwow, and others has facilitated the need for both the buyers and the sellers to be more informed about the trends and patterns that define the market value of a used car (Venkatasubbu & Ganesh, 2019).

Buying used cars is commonly an alternative option for car purchasers if their budgets are insufficient to acquire a new car. However, purchasing a vehicle has always been prohibitively expensive for middle-income families. This is also because customers' purchasing power is limited due to the high cost of new cars in these recent years. Thus, used cars are getting increasing interest and demand from buyers. Up to December 2021, the average price of new cars has climbed to $47,077, which has risen to $6220 compared to 2020 (Blanco, 2022).

1.2 Problem statement

Purchasing a used car has become the need for the people who would like to buy a car, especially in the current scenario of a short supply of new cars due to the shortage of chips and supply chain issues, as the result of the Covid-19 pandemic (Boudette, 2021). The shortage of new cars would prompt the car buyers who need a car without a long waiting process to opt for used cars. This will allow the used car sellers to raise the used cars' prices indiscriminately. As a result, the used car's average price hit $45,031 in September 2021, an increase of $4,872 or 12.1% over the previous year (Henry, 2021). Due to the increased demand for used cars, car dealers and internet portals are taking advantage of the situation by listing excessive prices on used cars.

On the other hand, it can be tough to determine whether a used car is worth the advertised price while looking at online ads. This is because several attributes can affect a used car's actual value. This includes model, mileage, year, make, model, and so on (Gajera, Gondaliya, & Kavathiya, 2021). Brand, model, age, horsepower, and mileage are typically essential in determining car price (Gegic, Isakovic, Keco, Masetic, & Kevric, 2019).

As a result, an accurate and dependable used car price prediction model using machine learning is imperative for car buyers and sellers to have a clearer picture of the fair price of the used cars. In addition, they could better grasp what makes an automobile desirable and the most important traits to look for in a used vehicle (Asghar, Mehmood, Yasin, & Khan, 2021).

1.3 Research objectives

We want to conduct a comparative study using several algorithms in this study. It involves Linear Regression, Polynomial Regression, Lasso Regression, and Random Forest to predict the used car price in the United States (US) used car market. The research objectives of this study are as below.

RO1: To identify the prominent features that can be used to predict the used car price.

RO2: To develop machine learning models using different algorithms to predict the used car price.

RO3: To evaluate and determine the best machine learning model to predict the used car price.

1.4 Research structure

Prasanta A/L Sathasivam 17157826
Khor Yi Lun S2018292
Sim Soon Jin S2043577
Sajay Thomas Samuel S2102673

The structure of this research paper is as follows in section 2; a literature review of some past works that have been done similarly. Then, in section 3, a description of the methodology used for this study, including the machine learning algorithms and processes throughout the study. Next, in section 4, evaluation and comparison of the result of the algorithms, as well as a discussion related to the results. Finally, section 5 concludes the whole paper.

2. Literature Review

The research works related to used car price prediction are reviewed in this section. The use of machine learning for predicting used car prices is depicted in Figure 1 as a fascinating research pattern. This can be observed that the usage of machine learning in used car price prediction is getting more concern due to the increasing demand for used cars.

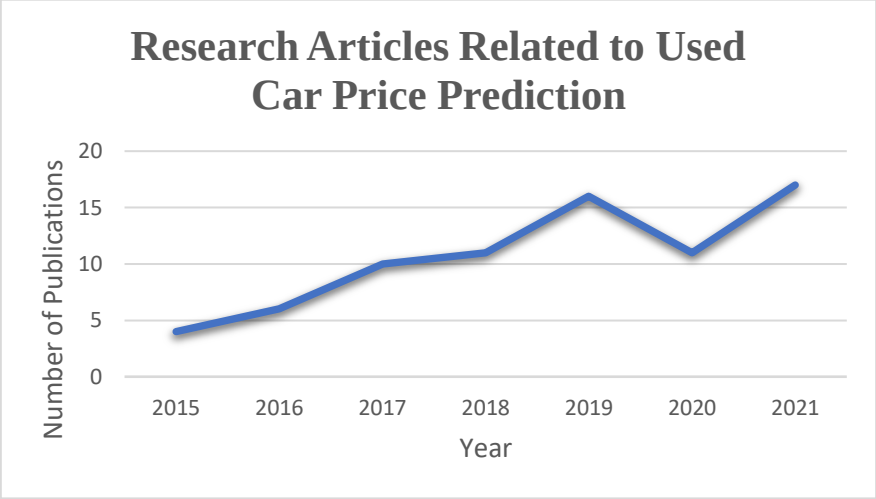

Figure 1 Research Articles Related to Used Car Price Prediction on Web of Science

2.1 Single Machine Learning Algorithms

According to the past research on used car prediction, several single machine learning algorithms were applied to predict the used car price. In the research of Gegic et al. (2019), single machine learning algorithms such as support vector machine (SVM) and ANN were used. This study collected the dataset from the web portal autopijaca.ba using a web scraper. The error rate of SVM was 10.53%, while ANN was 7.05%. Besides, the single machine learning algorithms in the study of Gajera et al. (2021) were k-nearest neighbours (KNN) regressor, linear regression, and decision tree regressor. These single algorithms were employed in the training data, preprocessed by removing the irrelevant variables that may act as outliers. This removal of attributes was examined by the correlation test using heatmap.

As a result, the decision tree regressor had the lowest RMSE and highest R-squared value among the single algorithms. Another study by Asghar et al. (2021) applied only a single machine learning algorithm, linear regression. The dataset was collected from Kaggle. The irrelevant features were removed, and the most crucial and robust correlation features remained. Recursive feature elimination (RFE) was adopted for optimal attribute extraction. The performance of linear regression was 90% for the R-squared value. In addition, a comparative analysis of car sales using supervised algorithms which adopted the single machine learning algorithms such as linear regression and decision tree had the result that linear regression provided the highest accuracy, even compared to ensemble methods such as random forest (Gupta, Kumar, Kumar, & Singh, 2021). This study performed feature engineering and data normalization for better data quality and model performance. There were other single machine learning algorithms, such as multiple linear regression applied in the research of Venkatasubbu and Ganesh (2019) and Monburinon et al. (2018), where the datasets were collected from the 2005 Central Edition of the Kelly Blue Book and Kaggle, respectively. However, the performance of these single machine learning algorithms was less desirable in predicting used car prices if compared to algorithms with regularization and even ensemble machine learning algorithms.

2.2 Single Machine Learning Algorithms with Regularization

In the research of Venkatasubbu and Ganesh (2019) and Amik, Lanard, Ismat, and Momen (2021), the L1 regularization technique, Lasso regression, was used. Feature selection using a Pearson correlation coefficient with a threshold of 85% and data scaling to transform the data to range from 0 to 1 was applied in the study (Amik et al., 2021). As a result, Lasso regression performed better than single machine learning algorithms such as regression tree and linear regression but worse than the ensemble method. However, some single algorithms, such as multiple linear regression and decision tree, performed better than Lasso regression in these two studies. Therefore, we could observe that the regularization method does not improve the model's performance but depends on the training dataset.

2.3 Ensemble Machine Learning Algorithms

Ensemble machine learning algorithms were applied in the research to compare with those single machine learning algorithms, including the bagging and boosting method. The most common bagging method would be random forest, which is the ensemble of multiple decision trees. There were several past studies which applied the bagging method. For instance, the study of Gegic et al. (2019), Longani, Prasad Potharaju, and Deore (2021) (the dataset was collected on ScienceDirect), Gajera et al. (2021), and Gupta et al. (2021). These studies adopted random forest, and most performed better than single algorithms. Random forest had the lowest RMSE and highest R-squared value among the algorithms, including single algorithms and even ensemble machine learning algorithm with boosting method, XGBoost regressor, in the dataset collected from previous consumers (Gajera et al., 2021). However, some research had the random forest performing worse than others. For instance, linear regression performed better than random forest in terms of accuracy in the study of Gupta et al. (2021), in which the dataset was collected from Kaggle. Plus, boosting method such as XGBoost and gradient boosted method had better performance with error rate in both studies of Longani et al. (2021) and Monburinon et al. (2018), which the datasets were collected from ScienceDirect and Kaggle, respectively, From the past research result, we could perceive that the usage of algorithms to optimize the used car price prediction is dependent on the dataset.

3. Methodology

3.1 Research flow

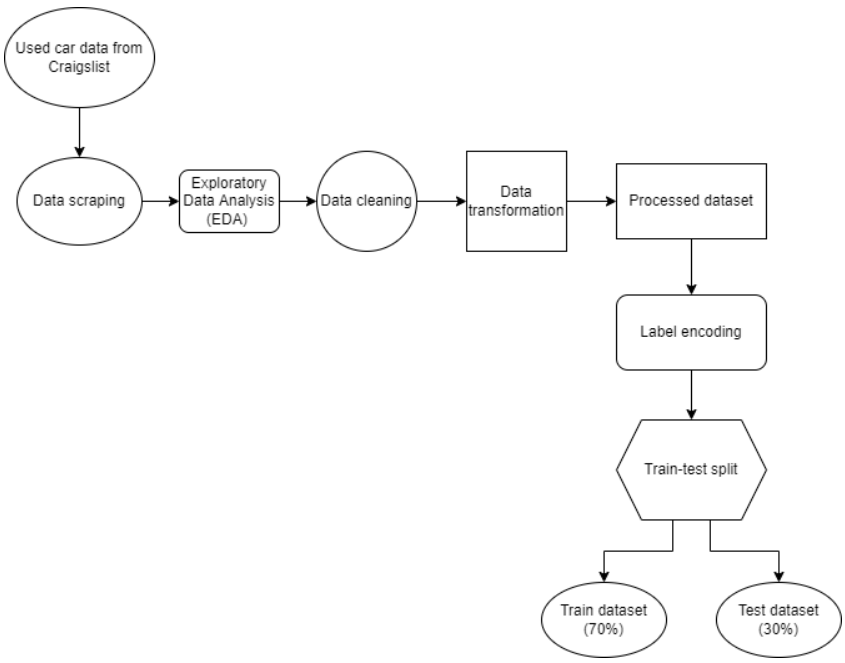

Figure 2 Data preprocessing flowchart

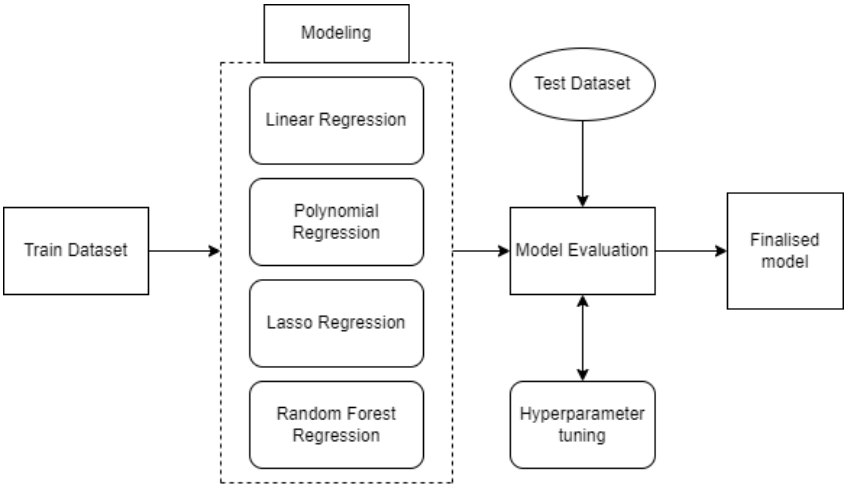

Figure 3 Modelling and model evaluation flowchart

This study is conducted in two stages. Figure 2 is the first stage where data is collected and preprocessed to generate clean data to be trained and assessed. Figure 3 represents the second stage in which data modelling with machine learning algorithms and performance evaluation occur. The next subchapters go through the specifics of each stage.

3.2 Tools used in this study

All the processes described will be conducted utilizing Python. Libraries and packages from Python, such as BeautifulSoup, Pandas, Matplotlib, Seaborn, Sklearn, and so on, are employed in this study to manage the data processes. Python is adopted as the tool for this study since it comprises all the libraries and packages needed for data modelling to build a price prediction model for used cars.

3.3 Data collection

Web scrapping is conducted to collect the attributes of the used cars on the website Craigslist. Used car data in the US market is collected on Craigslist. It is a classified ads website in the US that includes sections for employment, housing, for sale, wanted ads, services, community service, gigs, résumés, discussion forums, and used cars.

Craigslist

BeautifulSoup is applied for web scrapping used car data on the Craigslist website. First, the cities for scrapping the used car data are chosen beforehand: Dallas, Chicago, New York, SF Bay, Los Angeles, Houston, Phoenix, Philadelphia, San Antonio, Washington Dc, Boston, Nashville, Atlanta, and Miami, Seattle. After that, a list containing all the links for each city's page is collected and saved to a CSV file. These links are then looped over to retrieve each used car attribute. The acquired attributes are price, VIN, date time, city, geo coordinates, post body, post ID, condition, cylinders, drive, fuel, odometer, paint colour, size, title status, transmission, type, and year make model. The data is then saved into a CSV file after doing some simple cleaning to the data, such as providing the name to the unlabelled attribute, which is "year make model," cleaning the same year in it, and creating a new column "year_c make model" for it.

3.4 Data Preprocessing

The data preprocessing stage is critical for ensuring that machine learning algorithms can be appropriately employed. Therefore, we examine and explore the data to get more information from each attribute before working on data cleaning.

3.4.1 Exploratory Data Analysis (EDA)

Before going for data cleaning, EDA is conducted to explore the features of the Craigslist dataset. There are 21455 rows and 21 columns ("Column1", "VIN", "city", "condition", "cylinders", "date-time", "drive", "fuel", "lat", "long", "odometer", "PID", "paint colour", "post body", "price", "size", "title status", "transmission", "type", "year make model", "year_c make model") in total.

Firstly, we check the date-time columns to ensure they are in the correct format. The range of used car listings is also confirmed, over a month, from 13th April 2022 to 12th May 2022. Next, we check the percentage of missing values of each attribute.

Columns	Missing Values	Percentage	Columns	Missing Values	Percentage
Column1	0	0.00	pID	12	0.06
VIN	18124	84.47	paint color	5915	27.57
city	0	0.00	post body	218	1.02
condition	3637	16.95	price	0	0.00
cylinders	5285	24.63	size	10668	49.72
date time	0	0.00	title status	0	0.00
drive	6853	31.94	transmission	0	0.00
fuel	0	0.00	type	7311	34.08
lat	12	0.06	year make model	0	0.00
long	12	0.06	year_c make model	0	0.00
odometer	0	0.00			

Table 1 Count and percentage of missing values of each column

Since VIN is missing more than 80% of the data, it is then dropped out of the Craigslist dataset.

After that, it is clear from the figure below that, except for Boston and San Antonio, most cities have over 1500 used car listings, while there are less than one thousand postings in Philadelphia and Nashville.

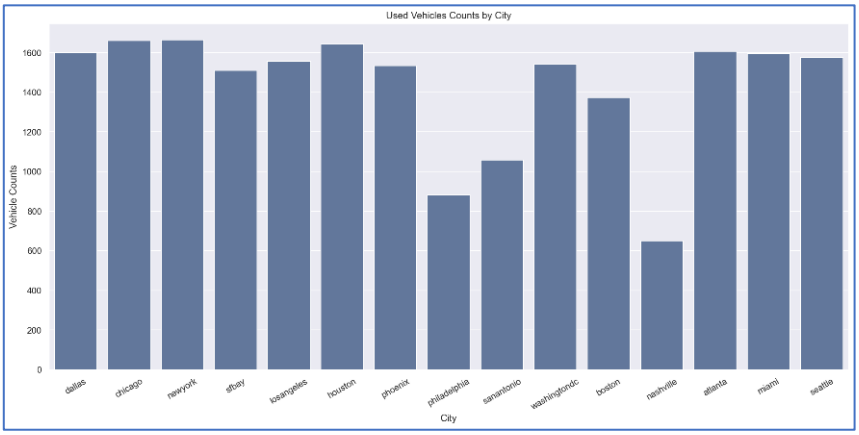

Figure 4 Used vehicles counted by the city

Besides, the number of posts each day peaks on the 9th - 11th days (22 - 24 April 2022), which are Friday – Sunday, as in Figure 5. However, after specific dates, the number of posts declines substantially.

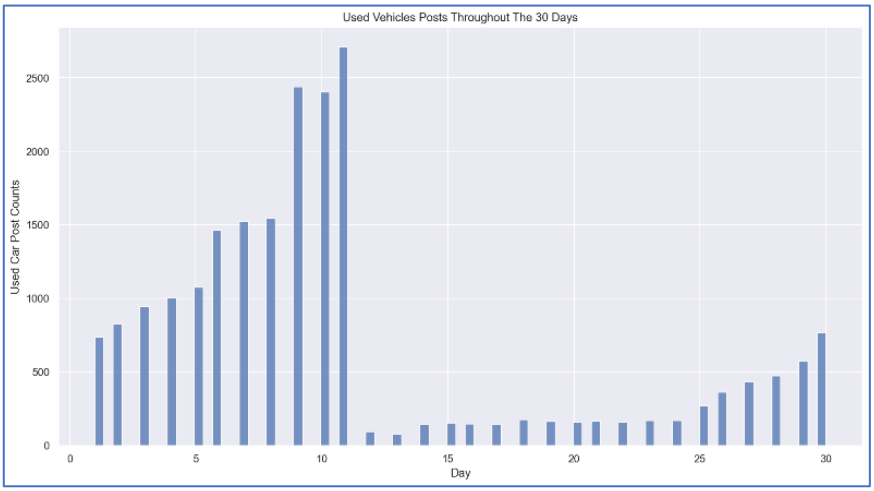

Figure 5 Used vehicles post count throughout 30 days

The most expensive category is “offroad,” followed by “other” and “truck” if we refer to the graph below.

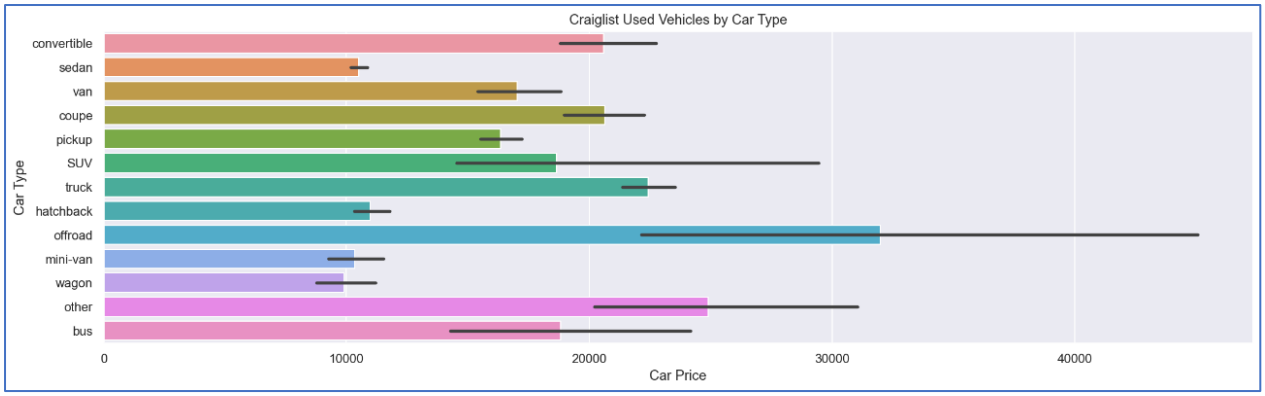

Figure 6 Used car prices on Craigslist by car types

There appears to be a significant outlier in the SUV price shown in Figure 7, which we may need to remove later in the data cleaning process.

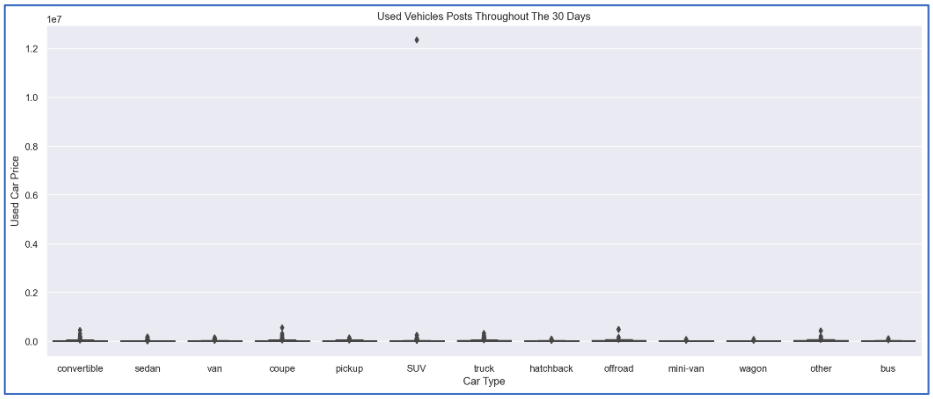

Figure 7 Boxplot of used vehicles price

From the attribute year make model, it can be observed that most cars built between 2006 and 2013 had the most Craigslist car postings. On Craigslist, the mode age of used cars is twelve. The popularity of the vehicle brand is then shown in the graph below.

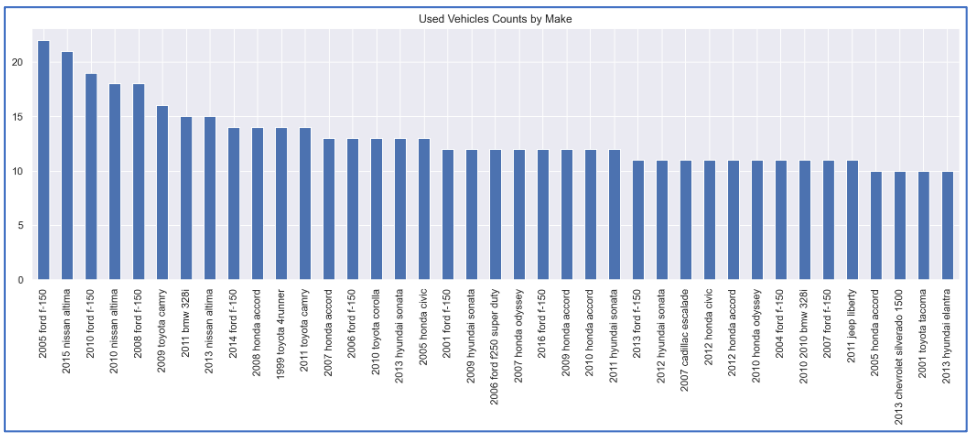

Figure 8 Count of the used vehicle make

Moreover, we can observe that Ford, Nissan, and Toyota are the most popular used car brands. There seem to be many Japanese cars made in the used car listings.

The most popular colour is “white,” followed by “black” and “silver,” as revealed in Figure 9. The top three colours have 8,813 postings over 15,540 posts, accounting for more than 50% of the postings.

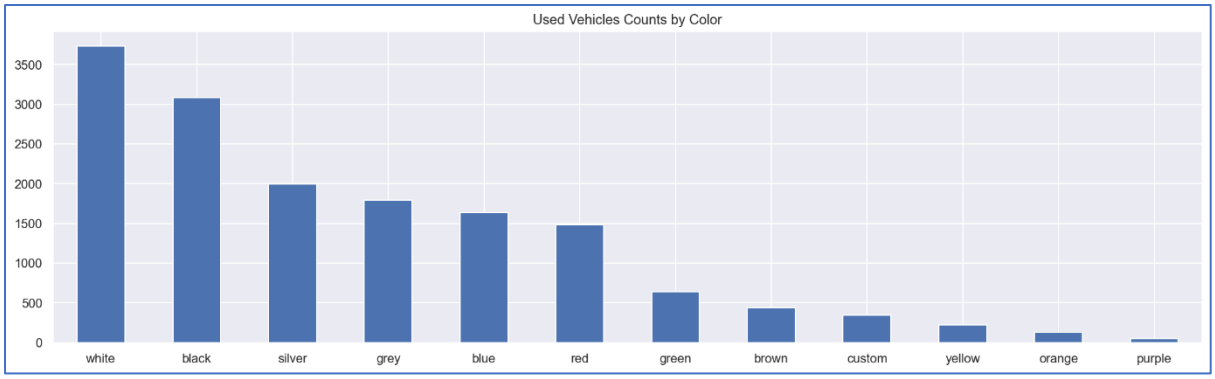

Figure 9 Used vehicle counts by colour

Lastly, most of the used car postings are automatic transmissions, more than 86% of the used car postings in the Craigslist used car dataset.

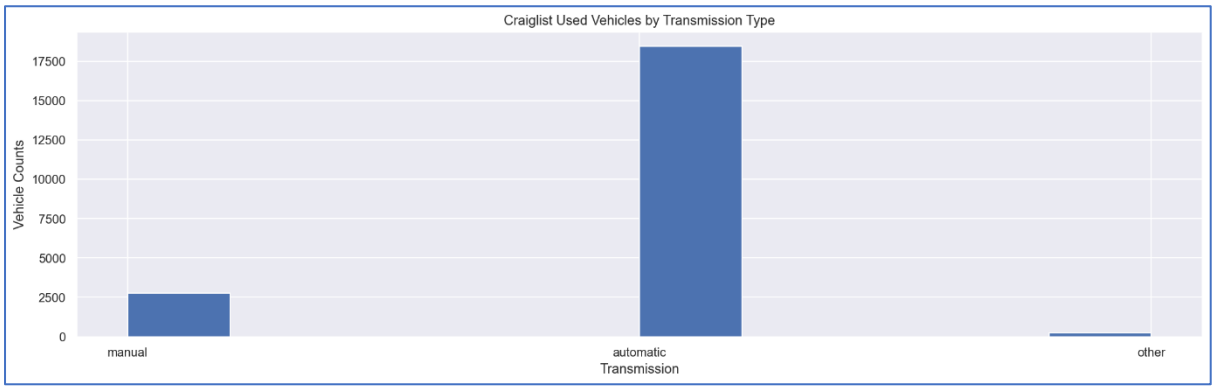

Figure 10 Used vehicle counts by transmission type

3.4.2 Data cleaning

There are several data cleaning steps for the Craigslist used car dataset. First, it drops columns VIN since it has more than 80% missing values and other irrelevant columns such as city, condition, cylinders, date time, drive, lat, long, post_body, size, title status, type, PID, and Column1. Next, the column "year make model" is removed since it is a duplicate column. Next, the column "year_c make model" is split into four columns, which are "year," "brand," "model," as well as "variant," and it is removed afterwards.

There are three distinct values in the "transmission" column which are "automatic," "manual," and "other." The "other" value in the "transmission" column is replaced with mode, which is "automatic" since the "other" value only has 224 entries, which accounts for 1% of the dataset. Besides, the outlier value price in used car type SUV, which is 13800, has also dropped. Rows containing more than two missing values in the columns are also removed since they are considered useless entries. After removing the rows containing more than two missing values, there are 21063 rows left in the dataset. To manage the remaining missing values in columns "paint colour," "model," and "variant," those missing values are all imputed by "Unknown" values since there is no clue to impute the missing values with valid data.

Finally, checking the extreme values in the dataset is also vital to remove unreasonable data. First, column "price" is checked with prices less than $400 and more than $1000000. Since the used cars' prices of less than $400, only a small number of buyers, and contain many noises, these rows are dropped. On the other hand, used car prices over $1000000 have valid data; therefore, they are not removed. We then work on column "odometer" to check the extreme values, which are values equal to 0 and more than three million. After examining the rows with extreme values, those with odometer values equal to 0 are dropped because they are not considered used cars. Upon checking column "year," there is a noise with the value "2006.". After checking the row with the year '2006.", it is found that the data is reasonable. The value "2016." is replaced with "2016". Moreover, there is also the value "2023" in column "year". The rows with the value are then dropped since "2023" is not a valid value for column "year". The total number of rows of the clean data after performing data cleaning is 20729, while there

are nine columns, which are “fuel”, “odometer”, “paint colour”, “transmission”, “year”, “brand”, “model”, “variant”, and “price” as the class variable.

3.4.3 Data Transformation

The values in newly created columns, such as “brand,” “model,” and “variant,” have the same values but in different cases. For instance, in brand, there are some values with “Toyota” and some with “Toyota.” Since Python is case-sensitive, it will interpret this as two various brands. Hence all the values in these columns are changed to lowercase.

3.5 Train-test split

Throughout the machine learning process, it is necessary to partition the dataset into train and test sets. We can evaluate how well a model responds to new data by segmenting the data. Since the train and test data are separated, the model could not see the test data ahead of time. This allows us to accurately analyse the performance of the regressors through the evaluation of the test set. In this study, the processed data is split into a 70:30 ratio. Therefore, we would have 14510 rows in the training dataset and 6219 in the test dataset.

3.6 Label Encoding

Converting categorical attributes into numerical values is referred to as label encoding. In attributes “fuel,” “paint colour,” “transmission,” “brand,” “model,” and “variant,” there are various distinct values which can be encoded to get the numerical representation. LabelEncoder from the sci-kit-learn library is applied for this task. For example, the variable “fuel” has five distinct values, and they are encoded into integers from 0 to 4. This applies to other categorical attributes as well.

3.7 Regression Algorithms

3.7.1 Linear Regression

A linear regression model that assumes a linear relationship between the features (X) and the target class is known as linear regression (y). The target class is determined by the linear combination of the input variables in this model (Tibshirani, 1996). The equation for a linear regression line is $\hat{y} = a + bX$, where $\hat{y}$ = predicted dependent variable, X = independent variable, b = slope of line, and a = intercept. LinearRegression() from sklearn.linear_model library is adopted to fit the training set to train the model.

3.7.2 Polynomial Regression

Polynomial regression is a regression analysis method that fits a non-linear relationship between the independent and dependent variables. This is different from linear regression; it only works best on linear relationships. Polynomial regression overcomes the problem of non-linear data by adding polynomial terms to linear regression and converting it into polynomial regression. The input variables will be converted into polynomial terms before modelling. Then, finding the degree’s optimum value can balance the variance and bias.

3.7.3 Lasso Regression

Lasso Regression is a regularization technique used in feature selection using a Shrinkage method, also referred to as penalized regression. A model using the L1 regularization technique is a lasso regression. LASSO stands for Least Absolute Shrinkage and Selection Operator, which is sometimes used for regularization or model selection.

3.7.4 Random Forest Regression

Random Forest Regression uses an ensemble learning method for regression where an additional layer of randomness is added to bagging. In addition to constructing each tree using a different bootstrap sample of the data, Random Forests change how the classification or regression trees are constructed. In a Random Forest, each node is split using the best among a subset of predictors randomly chosen at that node (Breiman, 2001).

3.8 Hyperparameter tuning

The hyperparameter tuning was only done to Random Forest (RF) model as it had the edge in terms of performance over the other three models. Randomized Search CV method was conducted on the RF algorithm and the best parameters ('n_estimators', 'min_samples_split', 'min_samples_leaf', 'max_depth', 'bootstrap') was determined. The RF model was executed one more time to see the difference in the result. Only small incremental changes

were shown by that method, as the MAE value was 17.94 lower, but the R^2 value rounded off to two decimal places stayed the same.

Best Parameters for RF:

Name	n_estimators	min_samples_split	min_samples_leaf	max_depth	bootstrap
Value	30	2	1	30	True

Table 2 Random Forest parameters after Hyperparameter Tuning

Result:

	MAE	MSE	RMSE	R
Before	8419.74	9349980784.52	96695.3	-3.973
After	7854.93	5624042465.8	74993.62	0.53

Table 3 Performance of Random Forest (Before & After)

4. Results and Discussion

Regression Algorithms	**MAE**	**MSE**	**RMSE**
Linear Regression	11639.06	791074506.9	28126.05
Polynomial Regression	11933.74	838565306.72	28957.99
Lasso Regression	11640.89	791165297.88	28127.66
Random Forest Regressor	7854.93	5624042465.8	74993.62

Table 4 Performance of regression algorithms

R^2 – Random Forest Model = 0.53

The random forest regressor was the better performing model compared to Linear, Polynomial and Lasso regressors as it had the lowest Mean Absolute Error and significant difference in correlation coefficient (R2) value of 0.53 on average 0.4 higher than the other three models.

Even though the RF regressor was the best performing model in this study, the performance metrics of R^2, MAE, MSE, and RMSE indicate that it is still not good enough to apply the developed model in the real-world scenario as the MAE (Mean Absolute Error) is still exceedingly high. Furthermore, the model is not robust enough, as indicated by the model's Correlation Coefficient (R2), which is only 53% robust. Furthermore, the result returned by the base learners indicates that there is no straightforward linear relationship between dependent and independent variables in this study.

5. Conclusion

In conclusion, this study indicates that more interest needs to be given to this domain to have a stable, accurate and universal model to predict used car prices. Furthermore, an ideal model must be market independent and location irrelevant, meaning the best-used car price predictor model must accurately predict old car prices by negating location, currency, regional market value, etc.

Data should be obtained from all places to have diverse car models and brands representing the overall population. Data should be scraped from many heterogeneous sources, especially car websites like Carwow, Carsome, Paultan.org etc. A unified dataset needs to be created for the used car price market.

Secondly, instead of eliminating features to accommodate other datasets, clever methods should be used to feature engineer new variables/features to have a meaningful analysis and substantial accuracy in predicting car prices. PCA (Principal Component Analysis) should also be considered for future study. Outliers and noise should be double or triple-checked (even better!), and proper methodology must be designed to eliminate those.

Lastly, the FAIR principle needs to be applied by all the stakeholders in the domain so that the analysis done by anyone in this domain will be meaningful and be rich with benefits. Furthermore, the lessons learned, information or knowledge gained by the studies within this domain should be applied by all manufacturers in future.

[Links of Jupyter Notebook: https://github.com/prasanta97/WQD7006_Machine_Learning_Group_Assignment]

6. References

Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh. *Information, 12*(12), 514.

Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology, 4*(2), 113-119.

Blanco, S. (2022). New Car Price Keeps Climbing, with Average Now at Almost $47,100. Retrieved from https://www.caranddriver.com/news/a38748092/new-car-average-sale-prices-47100/

Boudette, N. E. (2021). Want to Buy a Car? You Might Have to Get on a Plane to Claim It. Retrieved from https://www.nytimes.com/2021/12/22/business/economy/car-chip-shortage-pandemic.html

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old Car Price Prediction With Machine Learning. *Int. Res. J. Mod. Eng. Technol. Sci, 3*, 284-290.

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

Gupta, P., Kumar, P., Kumar, K., & Singh, N. (2021). Comparative Analysis of Car Sales Using Supervised Algorithms.

Henry, J. (2021). Average New Car Price Tops $45,000, Used Car Price Over $25,000. Retrieved from https://www.forbes.com/wheels/news/new-car-price-tops-45000/

Longani, C., Prasad Potharaju, S., & Deore, S. (2021). Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques. In *Recent Trends in Intensive Computing* (pp. 178-187): IOS Press.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). *Prediction of prices for used car by using regression models.* Paper presented at the 2018 5th International Conference on Business and Industrial Research (ICBIR).

MordorIntelligence. (n.d.). Used Car Market - Growth, Trends, Covid-19 Impact, And Forecast (2022 - 2027). Retrieved from https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Totakura, S. S. G. S. N., & Kosuru, H. (2021). Comparison of Supervised LearningModels for predicting prices of UsedCars. In.

Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. *Int. J. Eng. Adv. Technol.(IJEAT), 9*(1S3).