Predicting Car Prices through the use of Machine Learning Models

Team Orion
Eric Cheng, Hung Le, Babu Rajendran, Archana Shokeen, Tingjia Zhang

San Jose State University
Fall 2021

Abstract

Purchasing a new car is a significant conundrum for many people who do not know what to look for in selecting their ideal vehicle. From factors like price to model to depreciation rate, a car's value is changing constantly which makes investing in the right car a major priority. Using Machine Learning models to predict car prices based on variables such as age, price, and mileage, potential buyers can decide which car best suits their needs.

Although many studies have already developed ways to predict car prices based on those various factors using machine learning, not many have looked into how to incorporate customer needs into their car recommendations. By using a diverse range of classification and regression models such as Linear Regression, Random Forest, SVM, K Nearest Neighbors, this study can give a better understanding of the different machine learning models and their effectiveness in conducting car price prediction and car recommendations.

Introduction

Car purchases are a significant part of an individual's life, and choosing the right car can have a significant impact on a person's quality of life. Since a vehicle is a major ordeal, variables such as car price, mileage, brand, model, and year are all important factors for customers looking to choose a car. Since a vehicle's value can depreciate over time and that different models have different rates of depreciation, customers must also have good knowledge of car prices in order to identify the cars which will be of most value to them.

With the advancement of technology especially in the field of machine learning and data mining, humans can now be complemented with machine knowledge to make the best decisions.

In this project, the team has used machine learning algorithms to identify and predict the car prices for used cars so that potential customers can get a better idea of which cars to purchase to guarantee their satisfaction.

Related Work

Since car prices is a widely available source that can be easily accessed on the internet, there exists some studies from data analysts done on car price prediction. In a study conducted by Ning, Sun et al in their article “Price Evaluation Model in Second-hand Car System based on BP Neural Network Theory”, the researchers used BP Neural Network to do prediction on used cars. BP, which stands for backwards propagation is used to train the weights in the previous nodes to get a closer approximation for the expected output. Their results showed a relative drop in relative error, from around 0.78% to 0.58%, which improved the accuracy of their model.

Similarly in a study conducted by Nitis Monburinon et al, the researchers detailed their use of gradient boosting models to predict German car prices in their paper “Prediction of Prices for Used Cars by Using Regression Models”. For their study, they compare multiple linear regression, random forest regression, and gradient boosted regression trees to find out which models are the best when it will be used to predict car prices. Their results showed that for gradient boosted regression, the best performance has a MSE score of MSE =0.28. Comparatively, the MSE for random forest regression is MSE = 0.35. Finally, multiple linear regression resulted in a MSE of 0.55 when compared with the others. Since the lower the MSE, the better the accuracy of the model, gradient boosted regression has the best result, followed by random forest regression and finally multiple linear regression.

These studies are similar to the prediction analysis that is done in this study for car price prediction by looking at regression algorithms and testing their accuracy; however, compared to their studies which only looked at the accuracy of the models through MSE and relative error optimization, the study done in this paper also uses the results to give car recommendations for used cars and identify the vehicles that give the best bang for the buck.

Preprocessing

To conduct this project, a dataset is used as a basis of study. For this project, the dataset is directly scraped from TrueCars using BeautifulSoup, which provides a data frame consisting of various information about cars in the market. After checking the shape, information, and attributes of the dataset, the next step is to check whether or not there are any rows with null values. Since null values can affect the results of the modeling, they must be removed either through dropping the row, or filling them with the mean values of that specific column. After checking that there are no more null values, the next step is to conduct feature extraction to get more information about the cars. The main information that needs to be derived is the depreciation rate, which can then give the original and current price based on the number of years that has passed. The current age of the car is also important as it is an important metric in measuring the depreciation of the vehicle's value.

	year	make	model	city	state	mileage	price	style	accidents	owners	age	original_Price	depreciation	percent_Loss
0	2021	Chevrolet	Suburban	San Antonio	TX	8404	67994	SUV	0	1	0	408679.188580	340685.188580	0.166375
1	2017	Cadillac	CTS	Phoenix	AZ	29342	37000	Sedan	0	1	4	222389.181067	185389.181067	0.166375
2	2019	Kia	Forte	Bensalem	PA	17593	22395	Sedan	0	2	2	134605.559730	112210.559730	0.166375
3	2018	Ford	F-150	Hudson	WI	195205	20995	Pickup Truck	0	1	3	126190.833959	105195.833959	0.166375

Fig 1. Scraped TrueCars Dataset

One of the most important operations after extracting the features is to find the Gini score to capture the most important features. By running a regression model over the training and testing sets, one can figure out which of the attributes have the most effect in determining whether or not purchasing that vehicle is the right decision. Using the Gini score can give better insight into the dataset by giving better understanding of which specific features have the most impact on the machine learning models when doing training and testing.

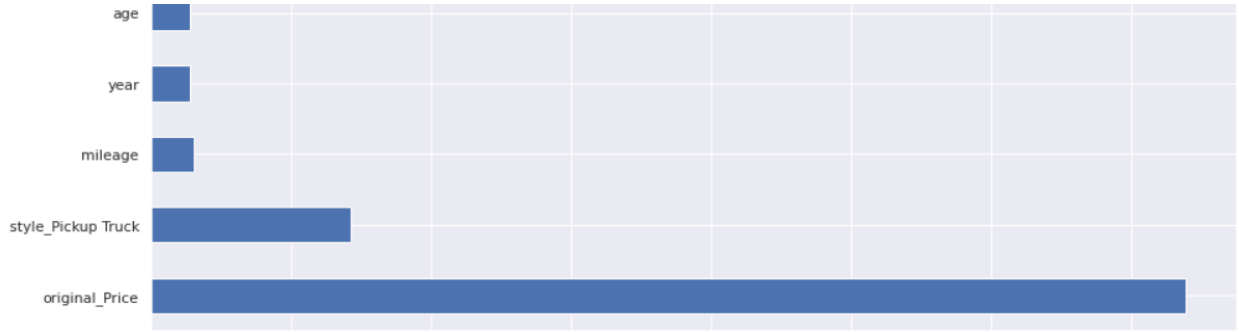

Fig 2. Extracted Most important features

For this specific dataset, original price were the significant factors in the prediction of price, followed by mileage, year and age of the car. After completing the preprocessing process, the next step is to execute the machine learning models to do the predictions.

Models

To best understand the variation of the data that is being worked with, it is imperative to use KMeans clustering to find the golden cluster, or the cluster of datapoints of cars that would give the best purchase value. By finding the SSE and the silhouette score, the number of clusters K that determines the number of clusters can be derived, and that is used to run the KMeans algorithm. After running KMeans and getting the best cluster performance for the car value prediction column, the process is repeated two more times for a total of three KMeans iterations.

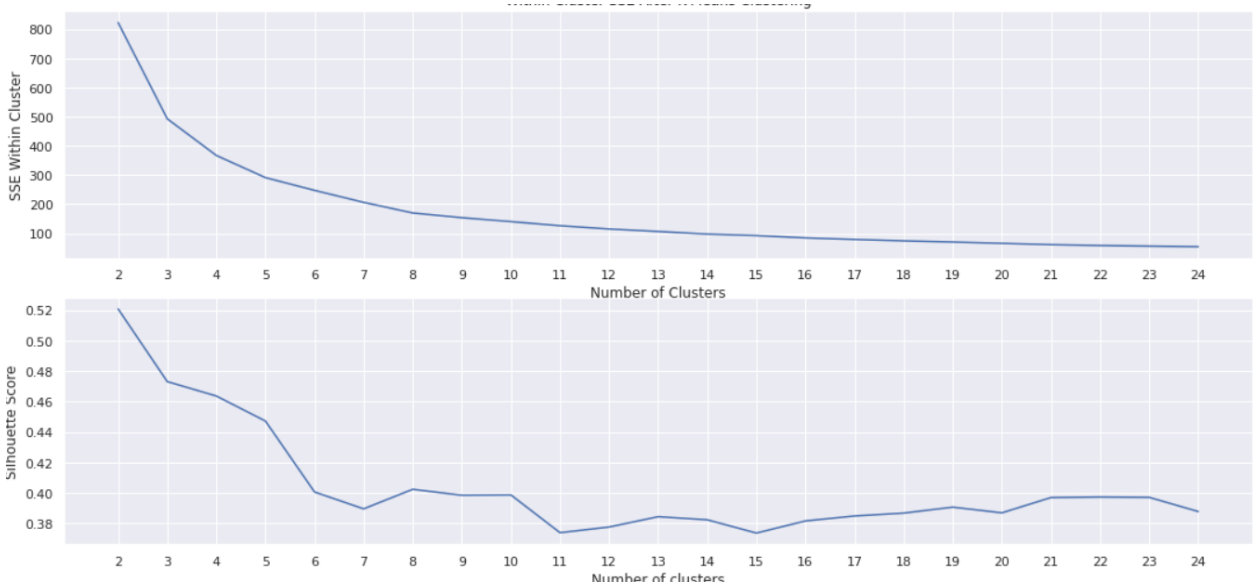

Fig 3. SSE and Silhouette Scores for KMeans

These are similar to the first KMeans process in that they use the SSE and silhouette score to determine the number of clusters that should be applied in the next iteration. Finally in the third KMeans, the best cluster is selected and the make and models for the best cars can be determined.

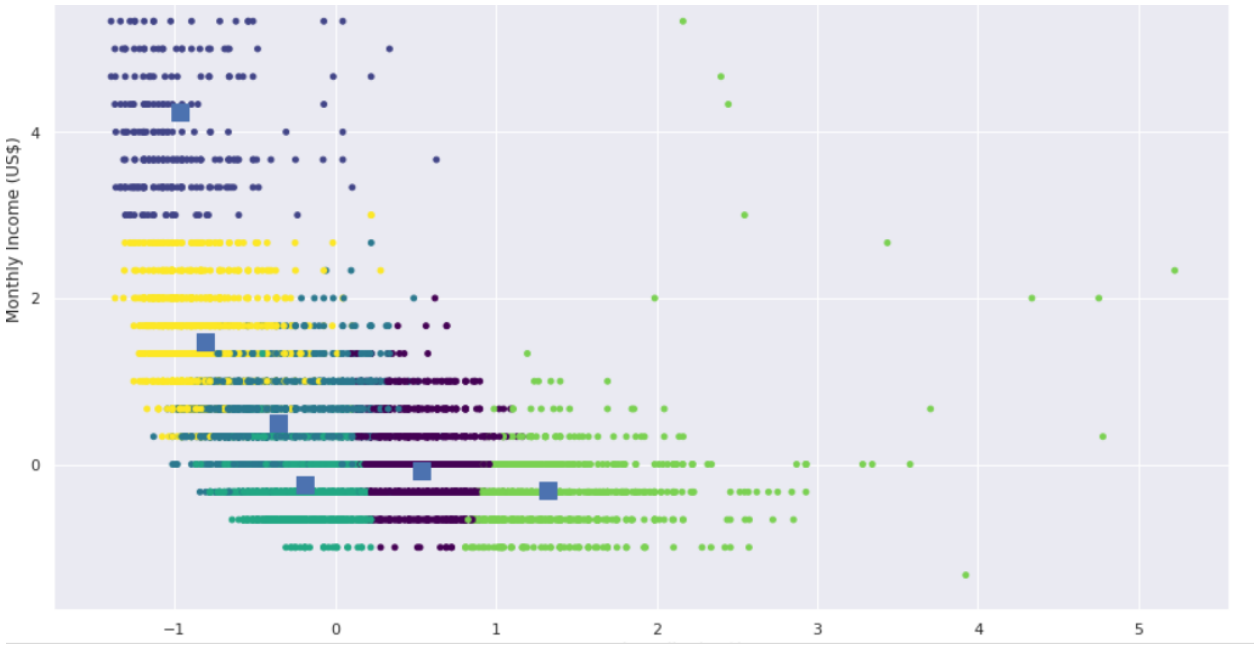

Fig 4. KMeans Cluster and Centroids

The next step is to do the actual modeling on the training and testing set. Splitting the dataset into a training to testing set based on a 80: 20 ratio using the train_test_split() function allows X_train and y_train to be used as training to predict the X_test and y_test. These will be used as the input for the multiple machine learning algorithms. The main classification models used are Nearest Neighbors, Decision Tree, Random Forest, Neural Net, and AdaBoost classifiers since these are the most prominent classification algorithms currently used. These are run through a muller loop and the classification algorithm with the best results for car price prediction is identified. The main regression models used include Linear Regression, MLPRegressor, RandomForestRegressor, KNNRegressor, and XBoost Regressor, which are used to predict the prices of the cars. Similarly to classification algorithms, these algorithms are run through a muller loop and the regression algorithm with the best result is identified.

Results

The results from the regression algorithm showed that for different regression models, there were different accuracy scores; however, some models had better accuracy than others. Linear Regression and MLP Regressor had 100% accuracy, while KNN Regressors, XGboost regressors, and Random Forest Regressors had between 90-100%. The car prices predictions comparing y_test and model predicted showed near identical prices, showing that for the random forest regression, the prediction was successful.

```
R2 SCORE = 1.00,
regressors = Linear Regression, Score (test, accuracy) = 100.00,
R2 SCORE = 1.00,
regressors = MLPRegressor, Score (test, accuracy) = 100.00,
R2 SCORE = 0.89,
regressors = RandomForestRegressor, Score (test, accuracy) = 89.53,
R2 SCORE = 1.00,
regressors = KNNRegressor, Score (test, accuracy) = 99.98,
/usr/local/lib/python3.7/dist-packages/sklearn/ensemble/_gb.py:290: FutureWarning:
  FutureWarning,
R2 SCORE = 1.00,
regressors = XBoost Regressor, Score (test, accuracy) = 99.99,
```

Fig 5. Regression model results

After Running the regression algorithms, the next step was to pickle the data and develop a web based application to make a user interactable prediction interface. Using heroku to deploy the web application and flask API, the user can input variables such as their car mileage, make, model, and year to get a prediction of the value of their vehicle.

Conclusion

Machine Learning algorithms are novel ways in helping humans better understand the world around them, and in this case, provide valuable insights for car purchasing prices and recommendations. By running classification models and regression models on the TrueCars data, a complete and accurate prediction of car prices changes can be identified and used to recommend the cars with the best value. The models that gave the best predictions are XGBoost, Random Forest, and KNN Regressors. Random Forest is then used as part of a flask application that allows users to get their car price with given car information.

Some of the things to improve in future include using more regression models in order to conduct the prediction for more varied model comparison. In addition, when doing the

preprocessing for the dataset in the beginning after scraping the data, it is possible to have a more accurate depreciation rate by scraping the purchase value of the model and submodel of the vehicle. These would also help improve the accuracy of the machine learning models.

References

1. Sun, N., Bai, H., Geng, Y., & Shi, H. (2017). Price evaluation model in second-hand car system based on BP neural network theory. *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. https://doi.org/10.1109/snpd.2017.8022758
2. Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of prices for used car by using regression models. *2018 5th International Conference on Business and Industrial Research (ICBIR)*. https://doi.org/10.1109/icbir.2018.8391177