Prediction of used car prices using linear regression and hypothesis testing *

ID:811150752
Likitha Guthikonda
lguthiko@kent.edu

ABSTRACT:

As we all know, a large number of cars are brought and sold. The project's main goal is to forecast used cars based on a variety of factors such as vehicle mileage, year of manufacture, fuel consumption, transmission, road tax, fuel type, and engine size. In the used car market, this model can benefit sellers, buyers, and car manufacturers. Upon completion, it can generate a reasonably accurate price prediction based on the information provided by the user. Machine learning and data science are used in the model-building process. The dataset was obtained from used car listings. To achieve the highest accuracy, the research used a variety of regression methods, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression. This project visualized the data before beginning model-building to better understand the dataset. The dataset was divided and modified to fit the regression, ensuring the regression's performance. R-square was calculated to assess the performance of each regression. Random forest had the highest R-square of any regression in this project.

BACKGROUND OF PROBLEM:

The cost of all the cars from the factory are fixed by the industry and there will be some extra cost added to the total price of the car which is of taxes which is incurred by the governments. This makes customers buying new car to be worthy of their money. But in recent times the prices of new cars have been rapidly increasing due to global price increase and demand in the raw materials used for the manufacturing of the car. Due to this, customers have been investing funds on used cars rather than new cars. Predicting the values of old cars is a fascinating and urgent topic to solve. Customers can be exploited in large numbers by setting inflated prices for old cars, and many people fall into this trap. As a result, a used car price prediction system is necessary to effectively analyze the automobile's worthiness utilizing a range of factors. Because of the high cost of automobiles and the migratory tendency of people in industrialized countries, most cars are purchased on a lease basis, with a contract between the buyer and seller. After the agreement is completed, these cars are resold. As a result, resale has become an important aspect of today's society.

The prediction of used automobiles is not an easy assignment given the description of used cars. A car's age, make, origin (the manufacturer's original nation), mileage (the number of kilometers it has traveled), and horsepower are all factors to consider. Fuel economy is also important due to rising fuel prices. Other considerations include fuel type, style, braking system, cylinder volume (measured in cc), acceleration, door count, safety index, size, weight, height, paint color, consumer evaluations, and significant distinctions earned by the automobile manufacturer. Sound system, air conditioner, power steering, cosmic wheels, and GPS navigator effect the price.

As a result, we offer an approach for predicting the pricing of secondhand automobiles based on attributes using a Machine Learning models. The accuracy of these models on the used car's data set is described in detail in further sections.

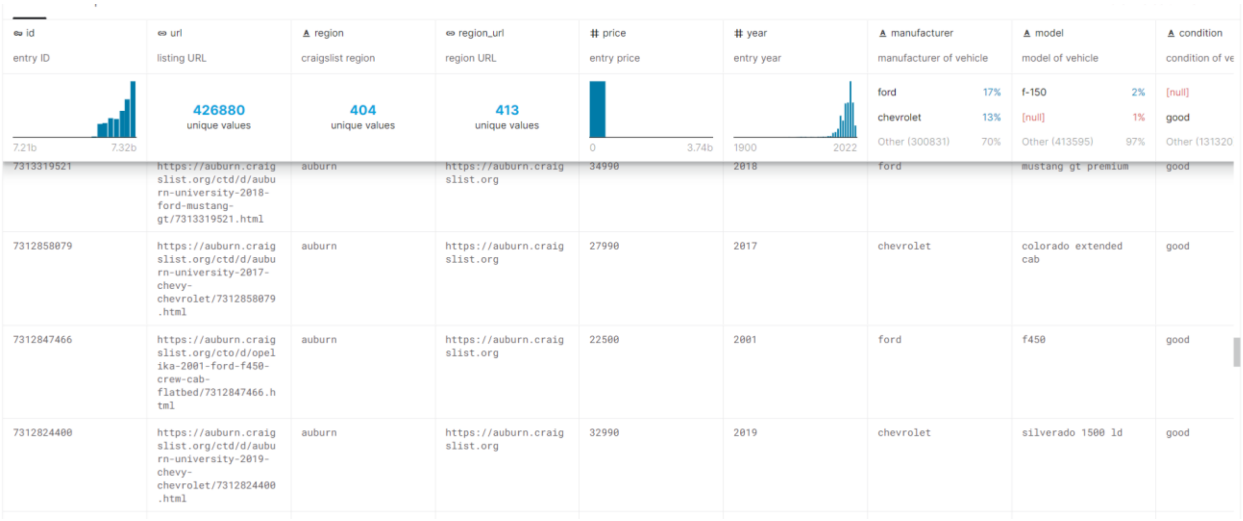
ML TASKS AND EVALUATIONS:

- First, we gather information on used cars and determine key factors that influence the price.
- Second, we preprocess entries with NA values and eliminate them. Remove any features that aren't related to price prediction.
- Third, we use all the models with features as inputs and price as outputs on the preprocessed dataset.
- Finally, using all this it predicts the price of a used car.

DATASET DESCRIPTION:

https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data?select=vehicles.csv

Sample view of dataset:



id entry ID	url listing URL	region craigslist region	region_url region URL	price entry price	year entry year	manufacturer manufacturer of vehicle	model model of vehicle	condition condition of ve
7.21b 7.32b	426880 unique values	404 unique values	413 unique values	0 3.74b	1900 2022	ford 17% chevrolet 13% Other (300831) 70%	f-150 2% [null] 1% Other (413595) 97%	[null] good Other (131320
7313319521	https://auburn.craigslist.org/ctd/d/auburn-university-2018-ford-mustang-gt/7313319521.html	auburn	https://auburn.craigslist.org	34990	2018	ford	mustang gt premium	good
7312858079	https://auburn.craigslist.org/ctd/d/auburn-university-2017-chevy-chevrolet/7312858079.html	auburn	https://auburn.craigslist.org	27990	2017	chevrolet	colorado extended cab	good
7312847466	https://auburn.craigslist.org/cto/d/opelika-2001-ford-f450-crew-cab-flatbed/7312847466.html	auburn	https://auburn.craigslist.org	22500	2001	ford	f450	good
7312824400	https://auburn.craigslist.org/ctd/d/auburn-university-2019-chevy-chevrolet/7312824400.html	auburn	https://auburn.craigslist.org	32990	2019	chevrolet	silverado 1500 ld	good

For used car price prediction, i used the Kaggle dataset. The dataset includes several characteristics from this study that are needed to forecast and categorize the range of used car prices. Few studies in the literature show that academics have employed a similar data collection or related dataset for price prediction.

[1] This patent outlines a general engine platform for determining an asset's value. For asset price prediction, this platform provides a price computation matrix. This platform may use a linear regression model that defines a collection of input variables to calculate car prices. However, it does not specify which features may be utilized for certain types of cars in order to make such predictions. We utilized four models and in that random forest model accuracy is more to forecast the price of secondhand cars based on key criteria.

[2] Zhang et al utilize the Kaggle data set to anticipate used vehicle prices. To examine the performance, the author evaluates the performance of numerous classification approaches (logistic regression, SVM, decision tree, Extra Trees, AdaBoost, and random forest). The random forest classifier outperforms all the others when it comes to prediction. After removing extraneous characteristics and outliers from the dataset, this study employs five features (brand, powerPS, kilometer, sellingTime, VehicleAge) to complete the classification job, yielding an accuracy of 83.08 percent on the test data. The difference is in the addition of a few more important parameters in the prediction model, such as the car's price and vehicleType. These two characteristics are essential in forecasting the price of a used automobile, but they appear to be undervalued in the research [2]. Furthermore, the range of characteristics year of registration, PowerPS, and price appears to have been limited down in work [2], resulting in less accuracy in the test dataset compared to what we analyze by widening the range of the parameters.

WEBSITES

Cars24

Cars24 is a web platform where seller can sell their used car. It is an Indian Start-up with a simplified user interface which asks seller parameters like car model, kilometers traveled, year of registration and vehicle type (petrol, diesel) [1]. These allow the web model to run certain algorithms on given parameters and predict the price.

Get Vehicle Price

Get Vehicle Price is an android app which works on similar parameters as of Cars24. This app predicts vehicle prices on various parameter like Fiscal power, horsepower, kilometers traveled. This app uses a machine learning approach to predict the price of a car, bike, electric vehicle and hybrid vehicle. This app can predict the price of any vehicle because of the smartly optimized algorithm

Car Wale

CarWale app is one of the top-rated car apps in India for new and used car research. It provides accurate on-road prices of cars, genuine user and expert reviews. It can also compare different cars with the car comparison tool. this app also helps you to connect with your nearest car dealers for the best offers available.

Car Trade

CarTrade is web and Android platform where user can research New Cars in India by exploring Car Prices, Car Specs, Images, Mileage, Reviews, and Car Comparisons. On this app one can Sell Used Car to genuine buyers with ease. One can list their used car for sale along with the details like image, model, and year of purchase and kilometers so that it is displayed to lakhs of interested car buyers in their city. User can read user reviews and expert car reviews with images that help in finalizing a new car buying.

PROBLEM DEFINITION

The costs of new vehicles in the business are fixed by the Government with additional costs ln the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds. So there increase in used cars sales are on a global scale. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling. The purpose of this study is to understand and evaluate used car prices, and to develop a strategy and use machine learning models to predict used car prices.

DATA PREPROCESSING

The data used in this project was downloaded from Kaggle We have the dataset contains the prices and attributes of over 370,000 used cars sold on the website across 40 brands. Our dataset contains 20 unique attributes of a car being sold, out of which we removed a few irrelevant columns that have little to no impact on a car's price from our analysis. First thing is to be preprocessing of the data set helping us narrowing down the feature to consider.

1. Remove datasets that 'url','image URL', 'lat', 'long', 'city_url', 'desc', 'city', 'VIN' features were dropped totally.
2. Extreme values were dropped because they inhibit prediction power of the model. These values were also dropped from dataset because these prices are noise for the data.
3. Cars that have odometer values are extreme and whose values are very low were dropped.
4. Cars from earlier than 1985 were dropped. For our analysis, these data points can be considered as outliers.
5. Missing values were filled with appropriate values like Average odometer of all 'condition' sub-categories were calculated and missing condition values of the car.
6. Filter out all cars listed as unavailable and Filter out invalid registration dates.
7. Filter out all data with value as which are unrealistic or null values.

EXPLORATORY DATA ANALYSIS

While exploring the data, we will look at the different combinations of features with the help of visuals. This will help us to understand our data better and give us some clue about pattern in data for pre-processing, to guide us in the process of exploratory data analysis. Based on the information and parameters available we have created 6 different types of hypotheses. They are

1. Condition Hypotheses
2. Location Hypotheses
3. Category Hypotheses
4. Accessories Hypotheses
5. Time Hypotheses
6. Sales Channel Hypotheses

Based on some criteria, such as information already available in the dataset and data that would be available at the beginning of the collection process. we created a final list of hypotheses that, in a possible second cycle, can increase when we include another item from the previously made list. They are,

1. Used cars with high Mileage should be cheaper
2. Used cars with better Appearance should be expensive.
3. Used cars in west or northeast regions should cost more.
4. Used cars in rich States should cost more.
5. Used cars which come from big manufacturer should be cost more.
6. Used cars with 4wd drive should be cost more.
7. Used cars with electric fuel should be expensive.
8. Used cars with SUV, pickup or truck type should cost more
9. Used cars with automatic transmission should be cost more.
10. Used cars with less age should cost less.

THE EXPERIMENTAL RESULTS:

Price is the feature that we are predicting in this study. To understand what affects change in price of a used car, the relation between features available in the data sat will be examined by using inferential statistic methods. We will test each hypothesis and check whether it true by analyzing it through visual exploration to gather insights about the model that can be applied to the data, understand the diversity in the data and the range of every field. We plotted each hypothesis as shown below.

H1. Used cars with high Mileage should be cheaper

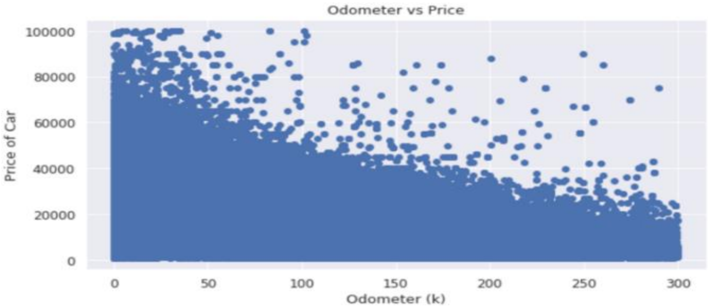


Figure 1: odometer vs price
From figure 1 we can see that price of the car decreases as the odometer value is increased. This information can be used to analyze the price of the vehicle based on odometer reading.

H2: Used cars with better Appearance should be expensive.

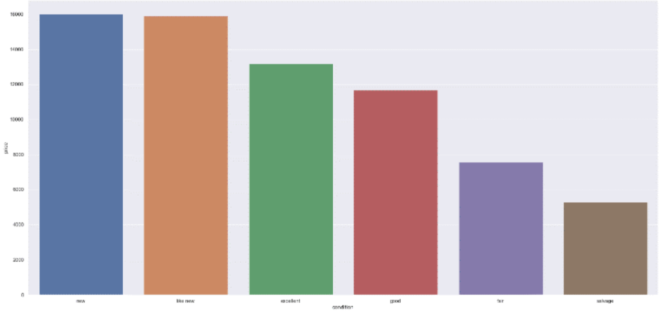


Figure 2: Appearance vs Price
From figure 2 the bar graph shows that appearance of car matters while estimating the price of the car. The cars that are new, look new costs more and the cost of the cars will decrease gradually.

H3: Used cars in west or northeast regions should cost more

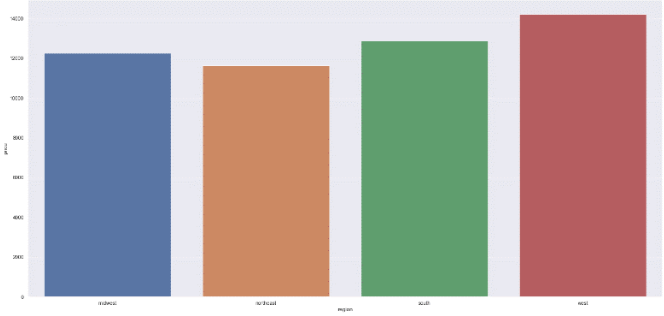


Figure 3: Region vs Price
Form figure 3 the bar graph shows that in west region the cars are costly and northeast region costs the lowest out of the four regions.

H4: Used cars in rich States should cost more

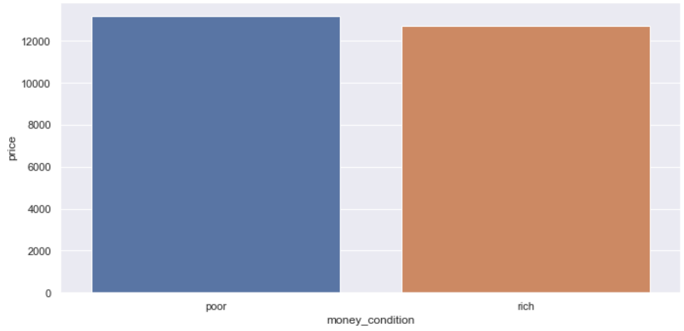


Figure 4: States vs price
Form the figure 4 the bar graphs shows that the cost of the cars is little more in poor states compared with the rich states.

H5. Used cars which come from big manufacturer should be cost more.

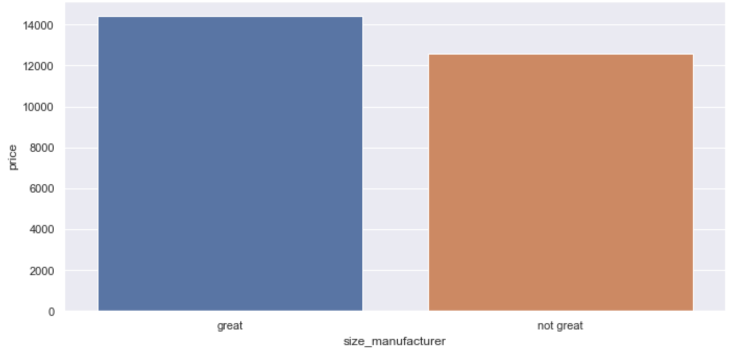


Figure 5: manufacturer vs price
Form the figure 5 the bar graph represents two bars with great as premium vehicles and not great as non-premium economy cars. For the graph we can say validate that premium car are more costly.

H6. Used cars with 4wd drive should be cost more.

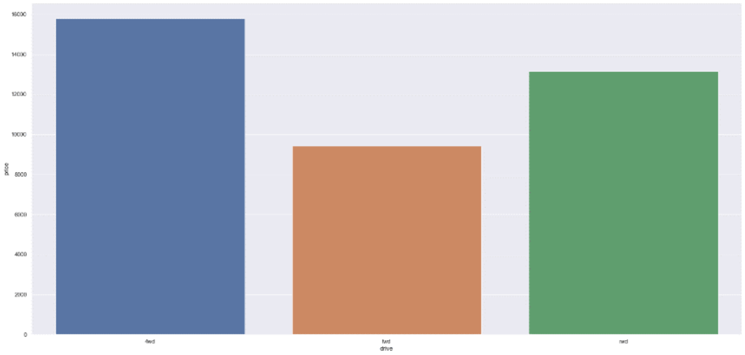


Figure 6: Car type vs price
From the Figure 6 we can see three bar graphs with 4-wheel drive, front wheel drive and rare wheel drive respectively. From the bar graph We can the conclude that 4-wheel cars are more costly compared with the other two types.

H7. Used cars with electric fuel should be expensive.

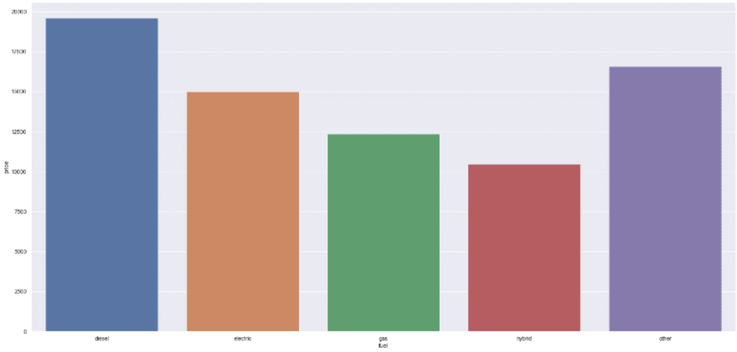


Figure 7: Fuel type vs price

From the Figure 7 the bar graph shows that diesel cars cost the most and hybrid cars are least cost of them with petrol cars at second and electric cars at third place.

H8. Used cars with SUV, pickup or truck type should cost more

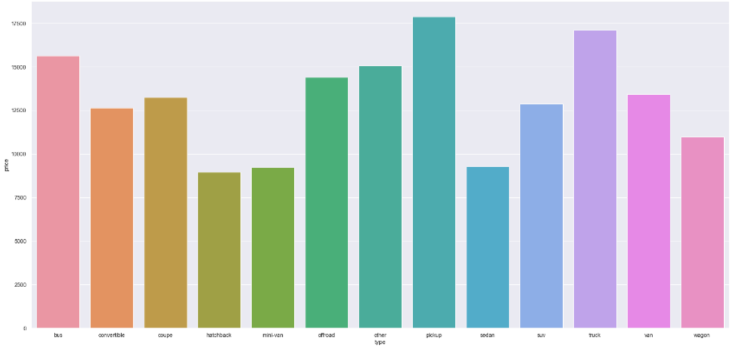


Figure 8: Vehicle type vs price

From the Figure 8 the bar shows the x axis with different types of vehicles and with their costs in y axis.

H9. Used cars with automatic transmission should be cost more.

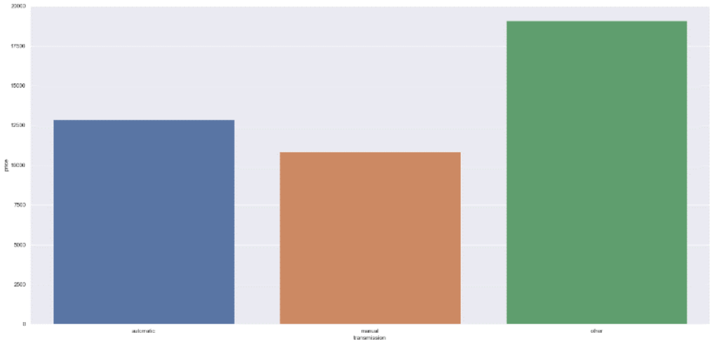


Figure 9: Gare type vs price

From the Figure 9 the bar graphs we can say that automatic cars cost more than the non-automatic cars.

H10: Used cars with less age should cost less.

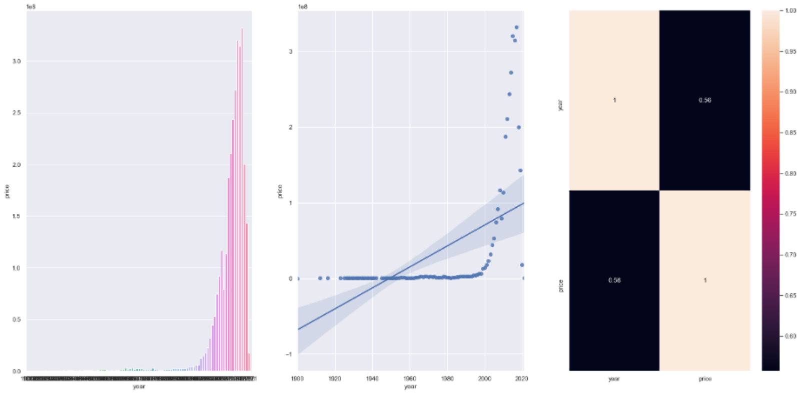


Figure 10: Age vs price
From the figures we can say that cars with the less age costs more. This information can be helpful analyzing the cost of the vehicle.

After data transformation, it is time to check for relationships among attributes. A correlation matrix was used to determine whether relationships exist

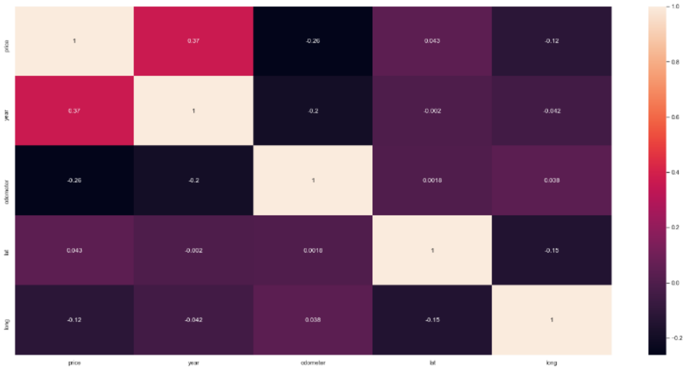


Figure 11: Comparison matrix
For the figure 11 It is evident from this table that the attributes that odometer, year are negatively correlated with the target class. Hence, negatively affecting the price

ML/DL METHODS:

We utilized several methods, including ensemble learning techniques, with a 90 - 10 split for the training and test data. Linear Regression, Random Forest, Lasso Linear Regression and XGBoost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package [4] was used.

1. **Linear Regression**

   Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.

2. **Random Forest**

   Random Forest is an ensemble learning based regression model. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behavior is partly ensured using Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.

3. XGBoost

Extreme Gradient Boosting or XGBoost is one of the most popular machine learning models in current times. XGBoost is quite similar at the core to the original gradient boosting algorithm but features many additive features that significantly improve its performance such as built-in support for regularization, parallel processing as well as giving additional hyperparameters to tune such as tree pruning, sub sampling and number of decision trees. A maximum depth of 16 was used and the algorithm was run on all cores in parallel.

4. Lasso Linear Regression

LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization doesn't result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

MODEL DESCRIPTION

The problem at hand is a regression problem. I tried with linear regression, Lasso, XGBoost and random forest regression. After much testing, it was found that random forest regression performed much better as it overcame the overfitting problem. Also, the model evaluation parameters show that random forest regressor gives the best performance compared to other. The accuracy of the regression was less than 75% even in training data.

Random forest is primarily used for classification, but we used it as a regression model by converting the problem into an equivalent regression problem. Random forest comes under the category of ensemble learning methods, which contains a cluster of decision trees, usually hundreds or thousands in number. These trees are individually trained on parts of the dataset and help in learning highly unpredictable patterns by growing very deep. However, this may create an overfitting issue. This is overcome by averaging out the predictions of individual trees with a goal to reduce the variance and ensure consistency.

1. Model Parameters

Random Forest has several parameters to be tuned to which certain parameters have higher importance and are described below:

- Number of Estimators: This is the number of decision trees constituting the forest.

- A maximum number of features: It defines the maximum number of features a single decision tree should be trained.

A Grid Search Algorithm was employed to find the optimum number of trees, and best accuracy was found when decision trees were used to build the forest.

Now, the maximum number of features is chosen to be equal to the number of features in the input data in case of regression problems and the square root of some features in case of classification. Since the problem at hand is a regression problem; we are going to the former.

2. Training and Testing

We split our input data into training, testing data and cross-validation. The splitting was done by picking at random which results in a balance between the training data and testing data amongst the whole datasets. This is done to avoid overfitting and enhance generalization.

3. Performance and Accuracy

The model score is the coefficient of determination R2 of the prediction. The training score was found out to be 94.82%, and the testing score was 83.63%.

The model was tuned in such a way that, only important features are taken, and the rest are discarded. The important features are found using correlation, measuring their importance towards the estimation of the price of a vehicle.

Overall, the random forest model effectively captured the nuances of the data and produced accurate predictions on the price of the vehicle. Below graphs represent the prices predicted based on the city, year, and manufacturer respectively. We can clearly see how the model performs very poorly with cars that have very low sales prices.

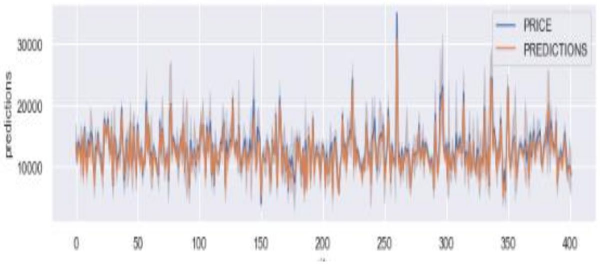


Figure 12: Line plot of price vs. predictions based on city

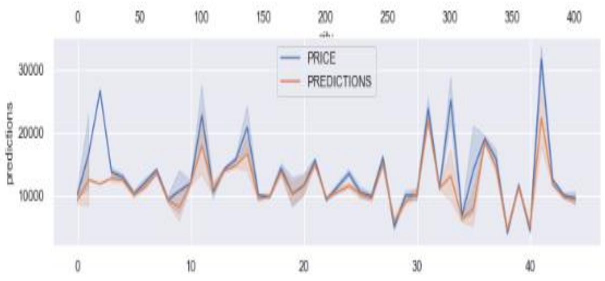


Figure 13: Line plot of price vs. predictions based on year

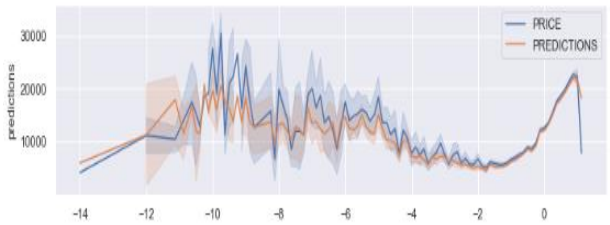


Figure 14: Line plot of price vs. predictions based on manufacturer

The model has its errors when the prices are predicted, and each component yielded different error rates which are plotted as below.

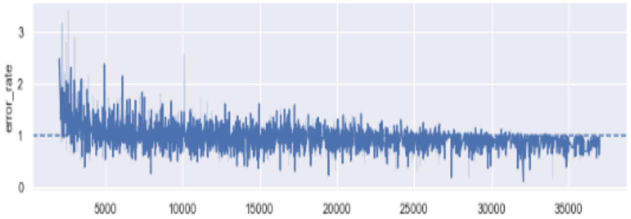


Figure 15: Scatter plot of error rates produced by model

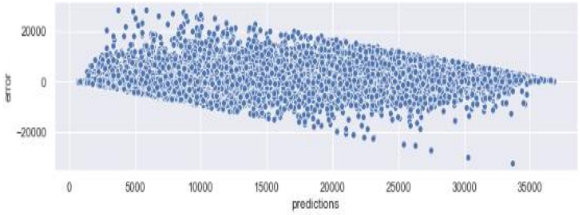


Figure 16: Scatter plot of error produced by the model for predicting prices

CONCLUSION

This paper evaluates used-car price prediction using Kaggle dataset which gives an accuracy of 83.62% for test data and 94.8% for train-data. An efficient machine learning model is built by training, testing, and evaluating four machine learning regressors named Random Forest Regressor, Linear Regression, and Lasso Regression and XG Boost Regressor. The most relevant features used for this prediction are price, kilometer, brand, and vehicle Type by filtering out outliers and irrelevant features of the dataset. Being a sophisticated model, Random Forest gives good accuracy in comparison to prior work using these datasets.

FUTURE WORKS

Keeping the current model as a baseline, we intend to use some advanced techniques like fuzzy logic and genetic algorithms to predict car prices as our future work. We can intend to develop a fully automatic, interactive system that contains a repository of used cars with their prices. This enables a user to know the price of a similar car using a recommendation engine, which we would work in the future.

Moreover, after the data collection phase Semiconductor shortages have incurred after the pandemic which led to an increase in car prices, and greatly affected the secondhand market. Hence having a regular Data collection and analysis is required periodically, ideally, we would be having a real time processing program.

REFERENCES

[1] Strauss, Oliver Thomas, and Morgan Scott Hansen. "Advanced data science systems and methods useful for auction pricing optimization over network." U.S. Patent Application No. 15/213,941.

[2] Xinyuan Zhang, Zhiye Zhang and Changtong Qiu, "Model of Predicting the Price Range of Used Car", 2017

[3] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4.7 (2014): 753-764

[4] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou,"Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119

[5] https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] https://github.com/panambY/Used_Car_Price: Hypothesis Testing, Machine Learning models, comparison and performance for Kaggle Dataset.

[7] Used Vehicle Value Index. (2021, April). Retrieved from manheim: https://publish.manheim.com/en/services/consulting/used-vehicle-value-index.html