Price prediction of used cars

CIND 820: Capstone Project

Prepared by- Proma Anwar

Student ID- 501206536

Supervisor- Dr. Ceni Babaoglu

Date of Submission: 11/27/2023

Ryerson University

Table of Contents

Abstract

The automotive industry is a key industry that plays an important role in the development of countries with its capital-intensive structure and the volume of employment it creates. A detailed study of automobile industry can greatly reflect the socio-economic situation of a country. The dataset that is chosen for this project is related to automobile industry. This is about used car listings and it is extracted from the popular automotive marketplace website. Using supervised machine learning algorithms, the price of a used car can be predicted. In this study three known predictive modeling techniques in Python were used: Linear regression, k-nearest neighbor regression, random forest regression. Results from the modeling indicates that random forest regression model has the best performance with an R^2 value of 0.440 when modeled using train-test split and 0.556 using K-Fold cross-validation.

*** Link to access the dataset: https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset

*** Link to Github Repository: https://github.com/Proma2023/DS.project.git

Introduction

The trade-in vehicle market is a consistently rising industry, which has nearly multiplied its reasonable worth over the most recent couple of years. Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. The development of online data entrances has worked with the requirement for both the client and the dealer to be better educated about the patterns and examples that decide the worth of the trade-in vehicle on the lookout. Machine learning (ML) is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to make useful inferences from data. AI calculations can

be utilized to foresee the retail worth of a vehicle, in light of a specific arrangement of highlights. Accurately predicting used car prices requires expert knowledge due to the nature of their dependence on a variety of factors and features. Various sites have various calculations to create the retail cost of the trade-in vehicles, and thus there is certainly not a brought together calculation for deciding the cost. Via preparing factual models at foreseeing the costs, one can without much of a stretch get an unpleasant gauge of the cost without really entering the subtleties into the ideal site.

The principal objective of this paper is to utilize three distinct regression models to foresee the retail cost of a utilized vehicle and think about their degrees of exactness. The informational index utilized for this study is collected from Kaggle.com. A detailed study of this dataset can be useful for the automobile industries around the world. It can be a great resource for automotive enthusiasts, buyers and researchers who are interested in analyzing market trend for used cars.

Literature Review

For any data science project to be successfully executed, it's very important to have knowledge about the existing work that have been done. It not only helps us in understanding the problem better but also to look at the problem from different perspectives. The dataset that is chosen to study is about the pricing of used cars. Several relevant study is found that used machine learning techniques to make a prediction. Our objective is to apply different combination of methods to predict the price and a comparative study between the methods. Some descriptive study is also very important to understand the whole scenario.

[1] Prashant Gajera, Akshay Gondaliya, and Jenish Kavathiya (March, 2021) used five different supervised machine learning techniques to predict the used car price based on a dataset with 92386 records. Several variables were observed including mileage, make, model, year, km driven**.** Based on the correlations between the variables some variables were dropped. The five mothods that were applied in this case are - **K** Nearest Neighbors (KNN) Regressor,

Random Forest Regressor, Linear Regression, XG Boost Regressor and Decision Tree Regressor. The performance of these five methods were compared based on the root mean square error (RMSE). Among all the methods Random Forest performed best with highest accuracy 93.11% and a minimum RMSE value 3702.34.

[2] Pattabiraman Venkatasubbu, Mukkesh Ganesh used Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees to determine the price of used cars and a comparative study was also conducted. They proposed the model with a hypothesis that, Multiple and Lasso Regressions are better at predicting price than the Regression Tree. From the accuracy values of the three models it was difficult to say if the hypothesis was correct. So an iterative ANNOVA was performed to check the error rates for the three models. Finally a Tukey's test (Tukey's Honest Significant Difference Test) was done to find out significantly different means of the groups. Combining all, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regression and lasso regression models.

[3] **Eesha Pandit, Hitanshu Parekh, Pritam Pashte, Aakash Natani** applied various Regression algorithms to produce a continuous value rather than a classified value as an output. As a result, rather than predicting a car's price range, it will be feasible to estimate its real price. A user interface was also created that takes input from any user and shows the price of a car based on the inputs. Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Decision Tree Regression, Random Forest Regression, XG Boost Regression, and Gradient Boosting Regression were used to predict price of the cars. After executing various regression algorithms on the model, it was concluded that the Decision Tree Algorithm was the top performer, with the greatest R^2 score of 0.95, implying that it provided the most accurate predictions.

[4] **Marcus Collard** in 2022 conducted a study to compare the performance of Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regression ML algorithms in predicting the price of used cars. The features that were used for the study are- ID, Price sold,

Year Sold, ZIP Code, Mileage, Maker, Model, Year, Trim, Engine, Body Type, Num Cylinders, Drive Type. The results showed that out of the four models tested, Random Forest Regression provided the highest accuracy in all of the metrics used and highest overall accuracy.

[5] **Kshitij Kumbar, Pranav Gadre, Varun Nayak** used a similar type of data set to predict the car prices for the cities of United States. The features used in that dataset were Mileage, VIN, Make, Model, Year, State and City. They utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. Linear Regression, Random Forest and Gradient Boost were their baseline methods. The results of the tests were quantified in terms of the *R*2 score. Compared to Linear Regression, most Decision-Tree based methods did not perform comparably well in this case. Linear Regression, the KMeans + Linear Regression Ensemble Learning Method (with K = 3) produced the best *R*2 score on test data without high variance.

[6] Ashutosh Datt Sharma ,Vibhor Sharma,Sahil Mittal,Gautam Jain,Sudha Narang have used several predictive analysis for building the model. They have compared between Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Random Forest Regression, Decision Tree Regression and XGBoost Regression. From the comparison Decision Tree Algorithm is seen to be the most accurate efficient in comparison to the other algorithms. They also created a web application with the use of HTML and CSS. This enables any user to input parameters and accordingly generate the predicted selling price of a used car. The user can input the desired values for parameters such as Year, Initial Price (in Lakhs), Kilometers Driven, Previous Owners and can select values for the parameters like Fuel Type, Transmission Type and Seller Type. After providing the input, user can simply click on the Selling Price button and a final value would be displayed that defines the selling price of used car for which the input has been given

Research Questions and Scope of the study

The regression algorithms that is chosen here is a unique combination, slightly different than the existing studies. The aim of this study was to answer several questions, like-

- Which factors are highly correlated with car price?
- Which regression algorithms to use?
- How to choose best attributes for regression and form the regression model?
- From the comparison of the chosen models, which algorithm gives the best result?

To answer all these questions some exploratory data analysis (EDA) will be done in Python and all the mentioned regression algorithms will be applied and evaluated in terms of accuracy or rate of error. From this comparative study the most efficient system will be chosen.

Data Description

The dataset that we are using in this study shows information about used car listings. It has 4009 instances and 10 columns that give valuable insights about automobile industry. The features that are highlighted in this dataset are- Brand, Model, Model year, Millage, Fuel type, Engine, Transmission type, Exterior color, Interior color, Accident history, Clean title and Price. The dataset is extensively examined in terms of null value, missing value, data types, outliers and so on.

Attribute	Data Types	Unique Values	Missing Values
Brand	Object	4009	0
Model	Object	4009	0
Model_year	Numeric	4009	0
Milage	Numeric	4009	0
Fuel_type	Object	3839	170
Engine	Object	4009	0
Transmission	Object	4009	0

Attribute	Data Types	Unique Values	Missing Values
Ext_col	Object	4009	0
Int_col	Object	4009	0
Accident	Boolean	3896	1013
Clean_title	Boolean	3413	596
Price	Numeric	4009	0

Data Preprocessing

After data collection the dataset was pre-processed to remove samples that have missing value, and remove non-numerical part from numerical attributes, converting categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that doesn't affect the price evaluations if needed to reduce the complexity of the model.

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

The dataset contained some missing values. Fuel_type, Accident and Clean_title- these three features showed some missing values. To fill in for those, some exploratory data analysis was done. For example 'Gasoline' was inserted as fuel type for the missing values of the column. It was observed that majority of the cars listed used Gasoline as fuel, so it was logical to fill in the missing places with 'Gasoline'. Another problem that needed fixing was unit signs added with the numeric values in 'Milage' and 'Price ' column. A new column 'age ' was added later to

the given dataset using the feature ‘model_year’, because knowing for how many years the vehicles were used seemed more logical in predicting the price.

To solve these issues some data preprocessing steps were followed

- After some exploratory data analysis the empty places were filled in with logically appropriate values.
- Removal of unnecessary signs and text.
- Data normalization
- Changing attribute types where necessary.

Exploratory Data Analysis (EDA)

Observing how the variables are correlated and the distribution type of the variables helps us in understanding the dataset better. It saves time and effort by indicating to the most relevant and important features. We have numeric, Object and Boolean type attributes in this dataset. Here some exploratory Data Analysis of data is done which is statistical graphics and other visualization methods to summarize the main characteristics of data. Various graphs and charts are plotted to get a better understanding of the dataset as well as the relationship of features in dataset.

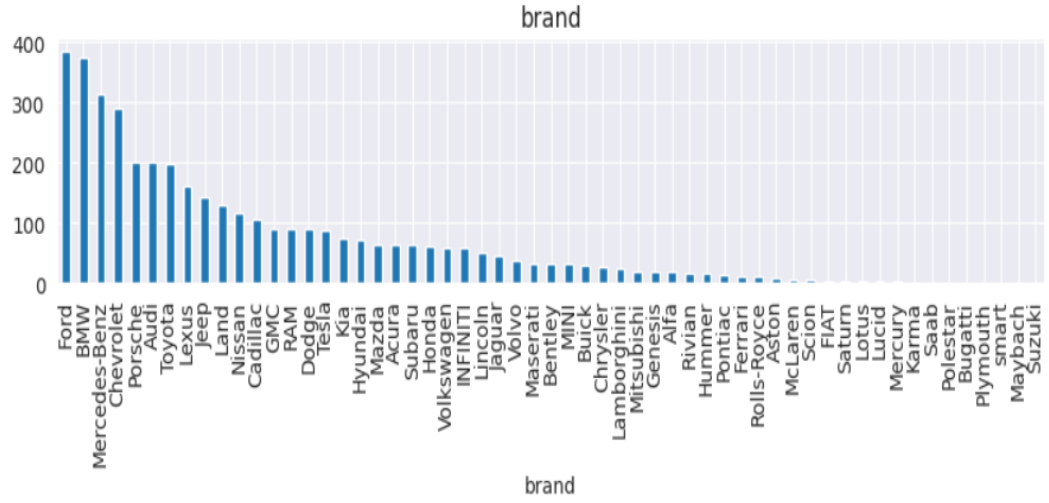


Fig: Car brands in the dataset

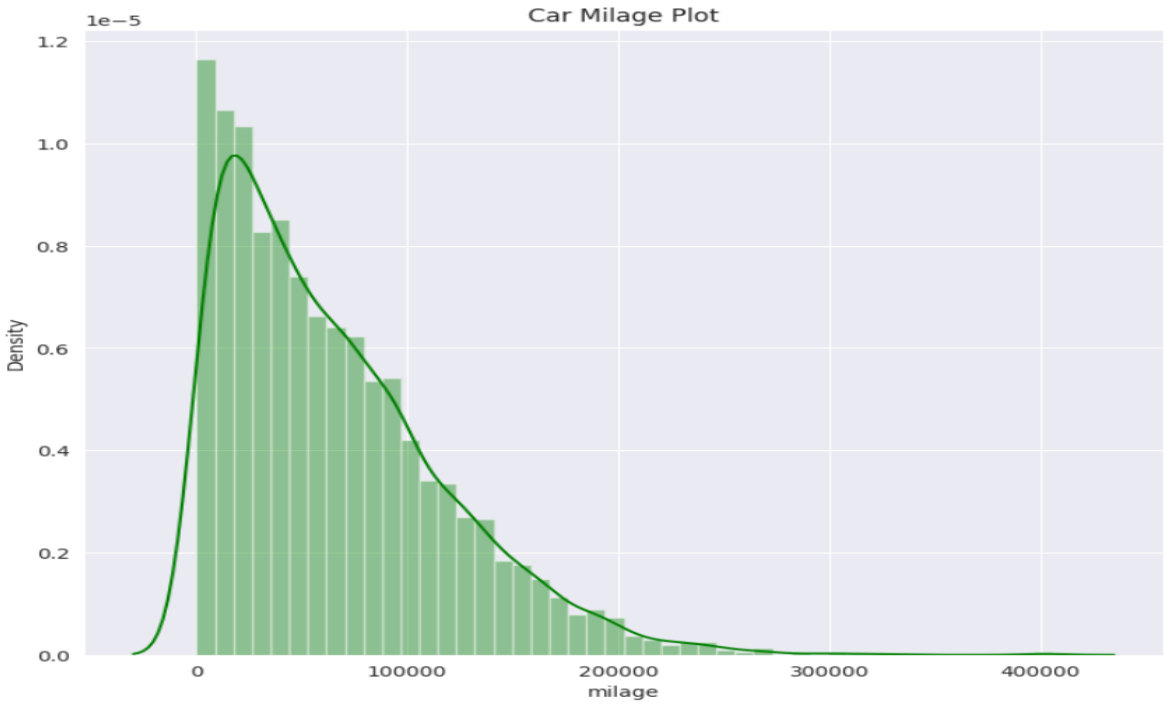


Fig: Distribution of car mileage

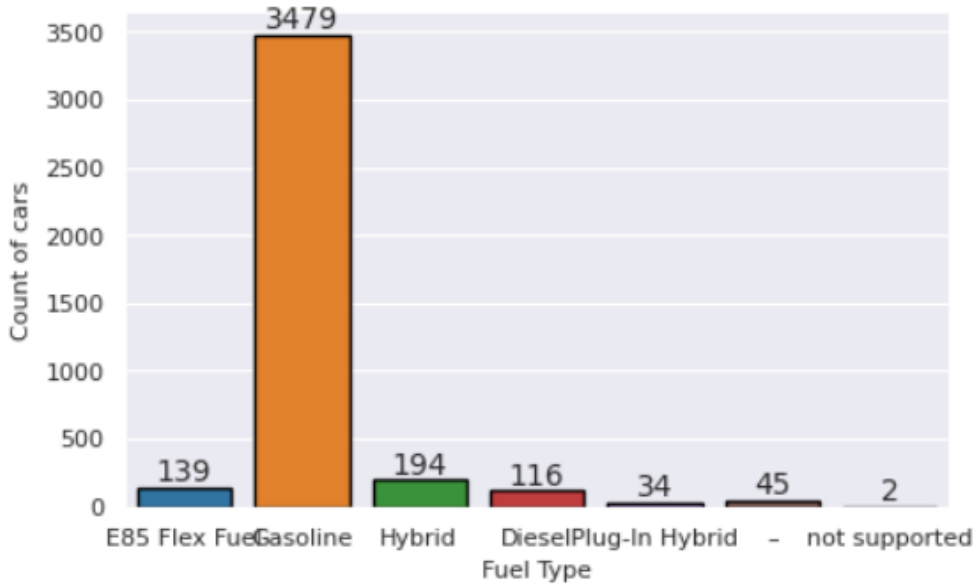


Fig: Types of fuels used

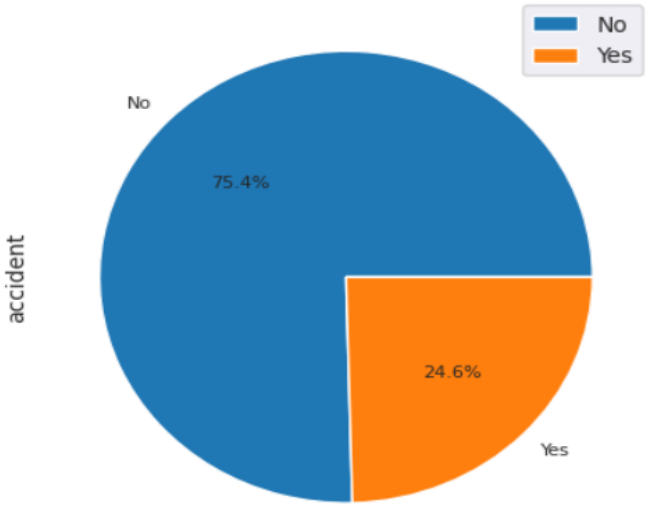


Fig: Pie chart of accident history

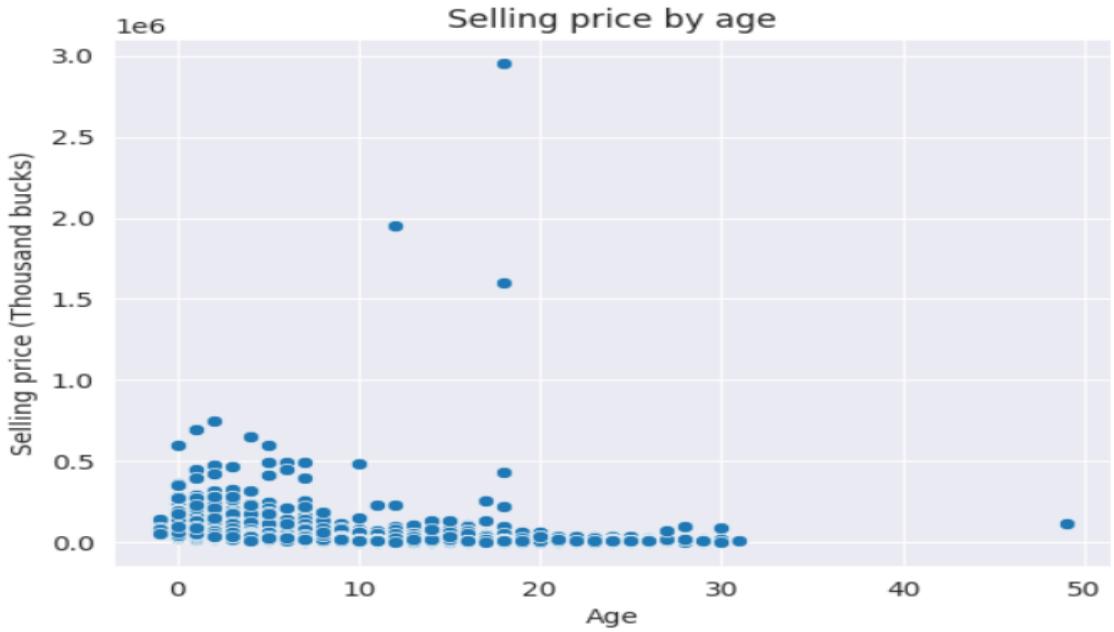


Fig: Car price v/s car age plot

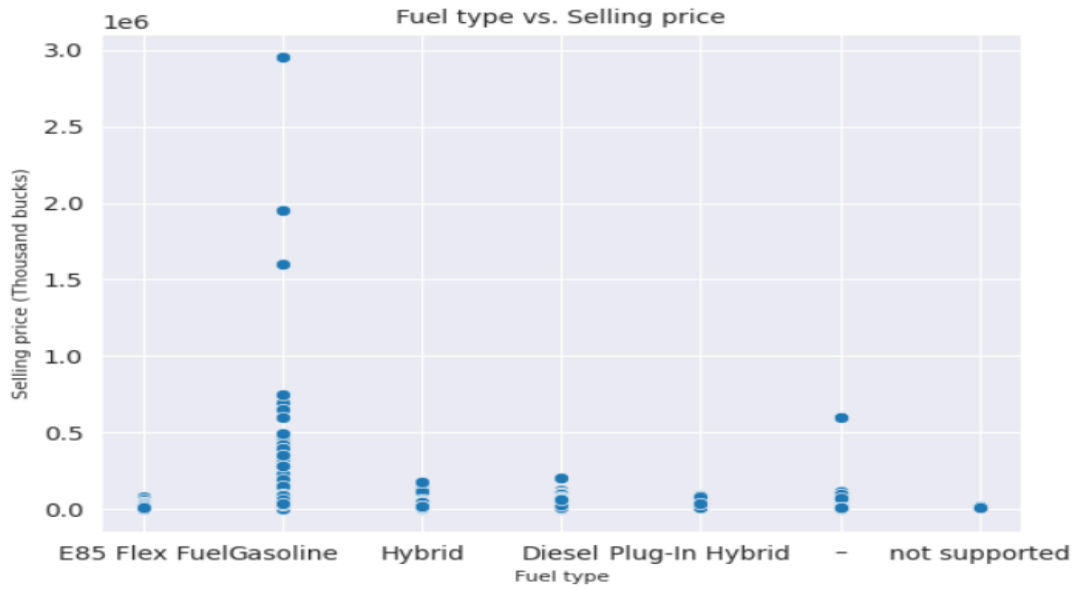
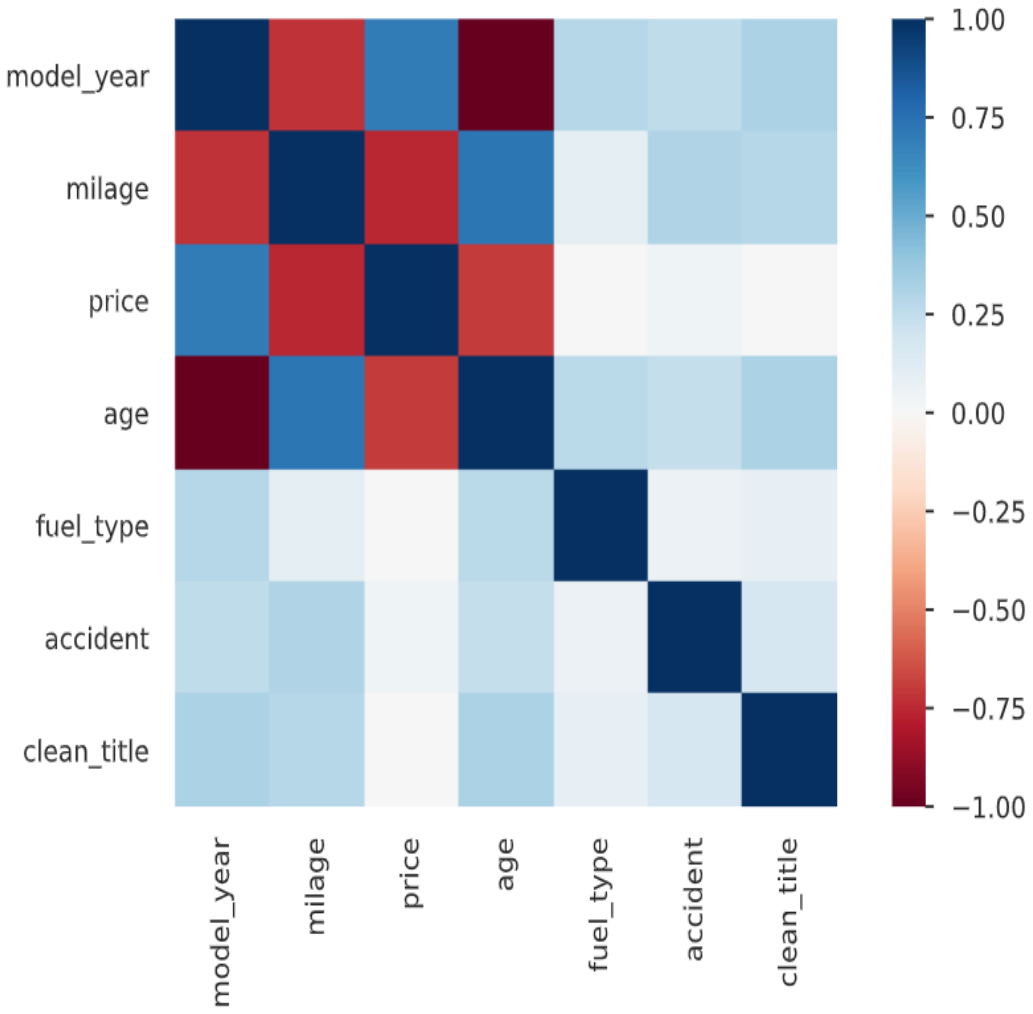


Fig: Fuel type v/s selling price plot

Correlation Matrix

	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price	age
brand	1.000000	-0.070170	0.001970	-0.012389	0.033300	-0.066116	-0.005099	-0.002001	0.008545	-0.023373	0.013011	0.030957	-0.001970
model	-0.070170	1.000000	0.028237	0.031513	0.004079	-0.037443	-0.024244	-0.008342	0.040801	0.000537	-0.039634	-0.033313	-0.028237
model_year	0.001970	0.028237	1.000000	-0.617720	-0.075843	0.148065	0.064506	-0.036169	0.035141	-0.194561	-0.264272	0.199496	-1.000000
milage	-0.012389	0.031513	-0.617720	1.000000	-0.096195	-0.227913	-0.043796	0.000891	-0.051394	0.301174	0.253614	-0.305528	0.617720
fuel_type	0.033300	0.004079	-0.075843	-0.096195	1.000000	0.080890	0.094140	-0.010056	0.013986	-0.038539	-0.004947	0.008496	0.075843
engine	-0.066116	-0.037443	0.148065	-0.227913	0.080890	1.000000	-0.011988	-0.037665	0.023628	-0.098442	0.024433	0.285172	-0.148065
transmission	-0.005099	-0.024244	0.064506	-0.043796	0.094140	-0.011988	1.000000	0.001548	-0.030224	0.021412	-0.038643	0.036943	-0.064506
ext_col	-0.002001	-0.008342	-0.036169	0.000891	-0.010056	-0.037665	0.001548	1.000000	0.085077	-0.004037	0.014161	0.004035	0.036169
int_col	0.008545	0.040801	0.035141	-0.051394	0.013986	0.023628	-0.030224	0.085077	1.000000	-0.009041	-0.090435	0.064821	-0.035141
accident	-0.023373	0.000537	-0.194561	0.301174	-0.038539	-0.098442	0.021412	-0.004037	-0.009041	1.000000	0.171904	-0.114088	0.194561
clean_title	0.013011	-0.039634	-0.264272	0.253614	-0.004947	0.024433	-0.038643	0.014161	-0.090435	0.171904	1.000000	-0.085710	0.264272
price	0.030957	-0.033313	0.199496	-0.305528	0.008496	0.285172	0.036943	0.004035	0.064821	-0.114088	-0.085710	1.000000	-0.199496
age	-0.001970	-0.028237	-1.000000	0.617720	0.075843	-0.148065	-0.064506	0.036169	-0.035141	0.194561	0.264272	-0.199496	1.000000

Correlation heat map



The correlation matrix and the correlation heat map indicates that mileage and age are negatively correlated with car price to some degree and other object type features are very poorly correlated with price.

Approach

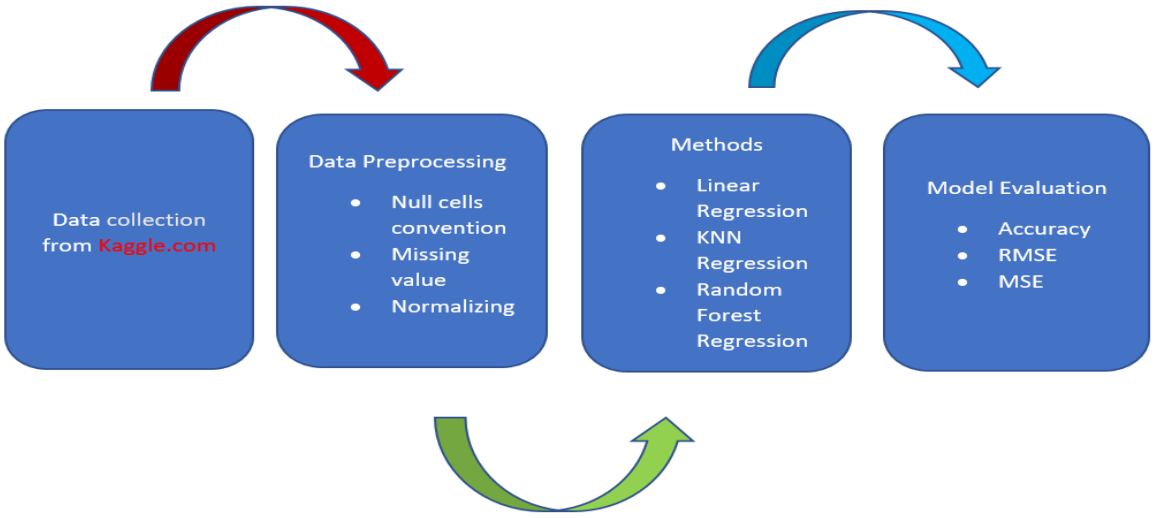


Fig: Flowchart showing working steps

Methodology

Various regression methods will be used for predicting the output of used car price. Every method will be evaluated in terms of precision and accuracy. Finally, we will compare all the accuracies of all the machine learning algorithms and choose the best algorithm for the prediction. The algorithms that will be used are- Linear Regression, Random Forest Regression and KNN- regression.

Linear Regression: It is a linear approach in statistics for modeling the relationships between a scalar response and dependent and independent variables. In linear regression, relationships are modelled using functions such as linear predictor, and unknown model parameters are estimated

from data. Here a linear regression model will be created with price as the dependent variable and other features as independent variables.

Random Forest Regression: Random Forest is a Supervised Learning Algorithm that employs the ensemble learning approach for classification and regression. Random forests are made up of trees that run parallel to each other and have no interaction while they develop. Random Forest is a meta -estimator that aggregates the outcomes of several predictions. It also aggregates numerous decision trees with certain modifications.

KNN regression: KNN regression is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

Label Encoding

The Label coding technique is used to deal with the categorical variables in the dataset. It generates a sparse matrix or a dense array depending in the parameters while creating a binary column for each category or parameter. It is best to use label encoding when there is a large number of categories for the features. The categorical variables in our dataset were: Brand, model, fuel_type, enigne, transmission, ext_column, int_column, acciedent, clean_title. After one hot encoding a binary representation of these variables are generated. For the accident variable 1 and 0 represents one or more accident reported or no accident reported for the vehicle respectively. Same technique is applied to the other categorical variables.

Train-Test Split

Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We used 80% of the data to train our model and 20% to test it.

Cross Validation

Cross validation is applied to the three algorithms that are selected. Using the K-Fold Cross-Validation method, the consistent dataset (the dataset before train-test split) was used to be split into k number of subsets, where k-1 subsets are used to train the models and the last subset is kept for validation to test the models. The scores of each fold are then averaged to evaluate the overall performance of each model. Cross-validation using 10-folds, where 9 folds were used for training and 1 used for testing, returned higher accuracy results in all three algorithms: Linear regression, Random Forest regression and KNN regression

Feature Selection

After observing the correlation matrix and correlation heatmap of the dataset, it is found that majority of the features have low correlation with the target variable. Only few features like-milage, age, accident have significant correlation with price. Also using stepwise regression method in R we see that these features have better correlation with the price variable. So we selected these features and observed the performance of the models again.

Model Evaluation Parameters

The regression model can be evaluated on following parameters:

1. Mean Square Error (MSE):

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors

2. Root Mean Square Error (RMSE):

RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

3. Mean Absolute Error (MAE):

This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset

Tools

Python is used as the primary tool for the study. Different regression algorithm in python is used for prediction and from the comparative study between these algorithms will lead us to a conclusion.

Results

Linear Regression

	Without cross validation	Cross validation on training set	Cross validation on testing set
R^2	0.319	0.24649	0.33040
MSE	1825516368.79	23903.82	22190.8846
RMSE	42726.0619	154.6086	148.966

Random Forest Regression

	Without cross validation	Cross validation on training set	Cross validation on testing set
R^2	0.440	0.0325	0.5569
MSE	1500185040.333	16790.1474208528	15305.532827160496
RMSE	38732.222	129.576801244871	123.71553187518734

KNN Regression

	Without cross validation	Cross validation on training set	Cross validation on testing set
R^2	-0.1026	0.08251	0.295800
MSE	2956351650.329	22904.2615	21771.140990740743
RMSE	54372.342	151.3415	147.55046930030667

From the above tabular summary, we can conclude that random forest regression out performed all the other regression model with the highest R^2 value 0.44 before cross validation and 0.556 after cross validation. The MSE and RMSE values for random forest regression is also better than the other two. MSE and RMSE both are minimum among the three models. After cross validation these values further improved (MSE: 15305.53 and RMSE: 123.715). But the values are not good in terms of describing the data. Low R^2 value indicate low accuracy of the model.

Model evaluation with feature selection

	Linear Regression	Random Forest Regression	KNN Regression
R^2	0.18699	-0.0111	0.0498
MSE	2179868515.9626217	2711065397.6833167	2548567942.1937656
RMSE	46689.0620	52067.89219551063	50483.34

From the above table we can see that even with feature selection the values of R^2 ,MSE and RMSE are not very different. In fact other than KNN regression, none of the other regression models are showing any better result. So possibly the dataset itself is not a good representative of the automobile industry.

Limitations

As this dataset has many categorical variables, so before fitting into the model we used labels encoding. So all the categorical variables had numeric representation. This can however lead to a loss of information and reduced performance of the various ML models to different degrees.

If the dataset does not include features that are strongly correlated to the price, the ML algorithm might not have access to enough information to accurately infer the price. From the correlation heat map we can observe that there are only three variables that have considerable correlation with the price variable. Which indicates that there are lack of strong predictors. Some of the features show almost no correlation with the target variable, like- Brand, model, Transmission. All these things have made the prediction a little difficult, thus we have obtained some poor R2 and RMSE values for the dataset.

Conclusion

Predicting used car prices is a difficult task due to the large number of features and parameters that must be examined in order to get reliable findings. The first and most important phase is data collection and preprocessing. The model was then defined and built in order to implement algorithms and generate results. After executing various regression algorithms on the model, it was concluded that the random forest regression algorithm was the top performer, with the greatest R2 score of 0.440, implying that it provided the best predictions among all the algorithms. Aside from having the highest r2 score, the Random forest regression also had the lowest Mean Square Error (MSE) and Root Mean Square Error (RMSE) scores, indicating that the errors in predictions were the lowest of all.

The first research question was to determine which of the features are highly correlated to this dataset. After various exploratory data analysis, results show that very few features like mileage, age and clean_title have considerable correlation with the target variable. Most of the variables of this dataset are categorical with low correlation value, which has been a disadvantage.

The second research question was to determine which of the features were best for the prediction and can most accurately predict price. For this part stepwise regression method was used in R environment. Combining forward and backward elimination technique three features were selected. But with the selected features the results did not improve much

The third research question is to determine which model has the highest efficiency in predicting price of this particular subset of used cars. After executing various regression algorithms on the model, it was concluded that the Random Forest algorithm was the top performer, with the greatest r2 score of 0.440, implying that it provided the best R2 values out of the three algorithms. Aside from having the highest r2 score, the Random Forest algorithm also had the lowest Mean Square Error (MSE) and Root Mean Square Error (RMSE) scores, indicating that the errors in predictions were the lowest of all.

Recommendations and future work

In the future, more data will be collected using different web-scraping techniques, and deep learning classifiers will be tested. Algorithms like Quantile Regression, ANN and SVM will be tested.

Afterwards, the intelligent model will be integrated with web and mobile-based applications for public use. 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. https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf
2. https://www.researchgate.net/publication/343878698_Used_Cars_Price_Prediction_using_Supervised_Learning_Techniques
   #https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses
3. https://www.irjet.net/archives/V9/i12/IRJET-V9I1261.pdf
4. https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf
5. https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf
6. https://www.irjmets.com/uploadedfiles/paper/volume3/issue_6_june_2021/12071/1628083486.pdf