Car Resale value Prediction Using machine learning

Presented by

BALAVIGNESH L

JEEVETH K

GOPINANTH G

DHANAMJAYASATH D

Team ID	PNT2022TMID27631
Project name	Project-Car resale value prediction

ABSTRACT

With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.

In order to predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithm with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that

can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States

Introduction

In this project we have used different algorithms with different techniques for developing Car resale value prediction systems considering different features of the car. In a nutshell, car resale value prediction helps the user to predict the resale value of the car depending upon various features like kilometers driven, fuel type, etc

In this project, we mainly focus on the analysis of the Vehicle Resale Predict and then predict the results through them using training data. The tradein vehicle market is an always rising industry, which has nearly multiplied its fairly estimated worth over the most recent couple of years. The rise of online entrances like CarDheko, Quikr, Carwale, Cars24, and numerous others has worked with the requirement for both the client and the merchant to be better educated about the patterns and examples that decide the worth of the pre-owned vehicle on the lookout. AI calculations can be utilized to anticipate the retail worth of a vehicle, in light of a specific arrangement of highlights. Various sites have various calculations to create the retail cost of the trade-in vehicles, and subsequently there is certainly not a brought together calculation for deciding the cost. Via preparing measurable models at foreseeing the costs, one can undoubtedly get a good guess of the cost without really entering the subtleties into the ideal site.

Need for the system

This resale value prediction system is made for general purpose to just predict the amount that can be roughly acquired by the user.

We try to predict the amount of resale by best 70% accuracy so the user can get estimated value before he resales the car and doesn't make a deal in loss

PROJECT PURPOSE

The main idea of making a car resale value prediction system is to get hands-on practice for python using Data Science. Car resale value prediction is the system to predict the amount of resale value based on the parameters provided by the user. User enters the details of the car into the form given and accordingly the car resale value is predicted.

PROJECT SCOPE

The system is defined in the python language that predicts the amount of resale value based on the given information. The system works on the trained dataset of the machine learning program that evaluates the precise value of the car. User can enter details only of fields like purchase price of car, kilometers driven, fuel of car, year of purchase.

OBJECTIVE

Car resale value prediction system is made with the purpose of predicting the correct valuation of used cars that helps users to sell the car remotely with perfect valuation and without human intervention in the process to eliminate biased valuation.

Due to limited data, system only takes into account limited features for predicting the resale value of the car. Since this is an online system, current system does not take into account any physical damage to the car body or engine while predicting the resale value

The new system developed by us consists of two parts - Data gathering and Prediction using Machine Learning based algorithms.

We have used web scraping libraries to gather data from the webpages of cars24 website. The script runs and captures data from the HTML div mentioned in the code via URL. URL should be entered by the user. For now, we have captured data by entering URL for Swift Dzire cars for 5 cities.

The second part is the web-based car resale value prediction. We have trained a boosting algorithm-based ML model using data from the previous step after preprocessing and cleaning.

Table -1: Sample Table format

Algorithms implemented	
Model Algorithm	RMSE
Support Vector Regression	56000
Logistic Regression	86000
Random Forest Regression	78000
Gradient Boosting Regression	42000

The trained model is used for prediction. The front-end form asks users to fill values which are required for the ML model to make prediction IE- city, kms driven, year of purchase and fuel type.

Upon form submission, the data is sent to the ML model via Flask API and the model responds with a predicted resale value of the car based on user input.

This prediction is displayed on the web page using a render template. Thus, with minimal information and without human intervention or manual examination, a user can predict the resale value of his car.

MOTIVATION

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately, Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

DATASET

For this project, we are using the dataset on used car sales from all over the United States, available on Kaggle [1]. The features available in this dataset are Mileage, VIN, Make, Model, Year, State and City

#	MAKE	CYLINDER VOLUME (CC)	YEAR	MILEAGE/KM	PRICE (RS)
1	TOYOTA	1300	2007	38000	410000
2	NISSAN	1500	2007	50000	325000
3	HONDA	1500	2005	59000	385000
4	TOYOTA	1000	2007	59000	360000
5	TOYOTA	1300	1989	62665	50000
6	TOYOTA	1500	2008	67000	615000
7	TOYOTA	1500	2008	69000	575000
8	TOYOTA	1490	2006	73000	450000
9	TOYOTA	1600	2006	82000	550000
10	TOYOTA	1000	2006	85000	325000
11	TOYOTA	1500	2000	113000	325000
12	TOYOTA	1500	2000	129000	218000
13	NISSAN	1500	2001	145000	195000

PREVIOUS WORK

As indicated by author Sameer Chand, they have done the forecasts of vehicle cost from the chronicled information that has been gathered from every day papers. They have utilized the administered AI strategies for foreseeing the cost of vehicles. Numerous different calculations like various straight relapse, k-closest neighbour calculations, gullible based, and some choice tree calculations additionally been utilized. Every one of the four calculations are looked at and tracked down the best calculation for forecast. They have confronted a few challenges in looking at the calculations, by one way or another they have overseen. As indicated by creators Pattabiraman, this paper is more focused on the connection among vender and

purchaser. To foresee the cost of four wheelers, more highlights are required like previously given value, mileage, make, model, trim, type, chamber, litre, entryways, voyage, sound, cowhide. Utilizing these highlights the cost of vehicle has been anticipated with the assistance of factual investigation framework for exploratory information examination. As per creators EnisGegic et al, in this paper the chiefly focus on gathering different information from web entryway by utilizing web scrap methods. Furthermore, those have been contrasted and the assistance of various AI calculations to foresee the vehicle cost in simple way. They arranged the value as per various scopes of value that is as of now given. Fake neural organization, support vector machine, arbitrary timberland calculations were utilized on various datasets to construct classifiers model. Another methodology was given by Richardson in his postulation work. In his hypothesis it states more strong vehicles will be delivered by vehicle maker. He looked at the crossover vehicles and conventional vehicles in scraper it really holds their incentive for longer time utilizing numerous relapse procedures. This works on the natural conditions, and furthermore it assists with giving colossal effectiveness of utilizing energizes. Wu et al, in this paper they have utilized neuro fluffy information based framework to exhibit vehicle value forecast. By considering the accompanying ascribes like brand, year of creation and sort of motor they anticipated a model which has comparative outcomes as the basic relapse model. Additionally, they made a specialist framework named ODAV (Optimal Distribution of Auction Vehicles) as there is a popularity for selling the by vehicles toward the finish of the renting year by vehicle vendors. This framework gives experiences into the best costs

for vehicles, just as the area where all that cost can be acquired. To anticipate a cost of vehicles, the K – closest neighbor AI calculation has been utilized which depends on relapse models. More number of vehicles has been traded through this framework so this specific framework is all the more effectively oversaw.

PREDICTION APPORACH

For accurate prediction and better model training, huge dataset of resale cars of Swift Dezire of 5 cities is gathered via web scraping cars24 website. This dataset contains data of 5 main features i.e., fuel type, kms driven, city, car purchase year and resale value. Here resale value becomes our target column whereas other columns served as features for our model

Data scraped consists of many unwanted characters like comma, whitespaces etc. which has to be removed as model can only understand numbers. Moreover, fuel type was converted into numerical codes via one-hot encoding.

A one hot encoding is a representation of categorical variables as binary vectors. This requires that the categorical values be mapped to integer values. After data preprocessing, all 5 files, each representing each city has to be merged for model training.

Various different machine learning algorithms were implemented on the dataset along with hyperparameter tuning using GRID SEARCH CV

Reason behind GBR's good performance is because of its mathematical working.

The reason why GBR could outcome all other regression algorithms is the mathematics behind it.

Gradient boosting involves three elements :

1. A loss function to be optimized.
2. A weak learner to make predictions.
3. An additive model to add weak learners to minimize the loss function.

1.LOSS FUNCTION

The loss function used depends on the type of problem being solved.

It must be differentiable, but many standard loss functions are supported and you can define your own. For example, regression may use a squared error and classification may use logarithmic loss A. benefit of the gradient boosting framework is that a new boosting algorithm does not have to be derived for each loss function that may want to be used, instead, it is a generic enough framework that any differentiable loss function can be used.

2.WEAK LEANER

Decision trees are used as the weak learner in gradient boosting.

Specifically, regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and “correct” the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss. It is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, splits or leaf nodes. This is to ensure that the

learners remain weak, but can still be constructed in a greedy manner.

3.ADDITIVE MODEL

Trees are added one at a time, and existing trees in the model are not changed.

A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e., follow the gradient). We do this by parameterizing the tree, then modify the parameters of the tree and move in the right direction by (reducing the residual loss.

ALGORITHMS

We used to algorithms in this paper

1. Linear Regression
2. Random Forest Regression
3. Decision Tree Regression

Linear Regression

It is an AI calculation dependent on administered learning. It plays out a relapse task. It is utilized to assess genuine qualities (cost of houses, number of calls, absolute deals and so forth) in view of nonstop variable(s). Here, we set up connection among free and ward factors by fitting a best line. This best fit line is known as relapse line and spoke to by a straight condition Y= a *X + b. Prior to understanding what direct relapse is, let us get ourselves acclimated with relapse. Relapse is a strategy for demonstrating an objective worth dependent on free indicators. This strategy is generally utilized for spreading and discovering circumstances and logical results connection between factors. Relapse methods generally vary dependent on the quantity of autonomous factors.

In figure1 it is very clear how the linear regression algorithm will work.

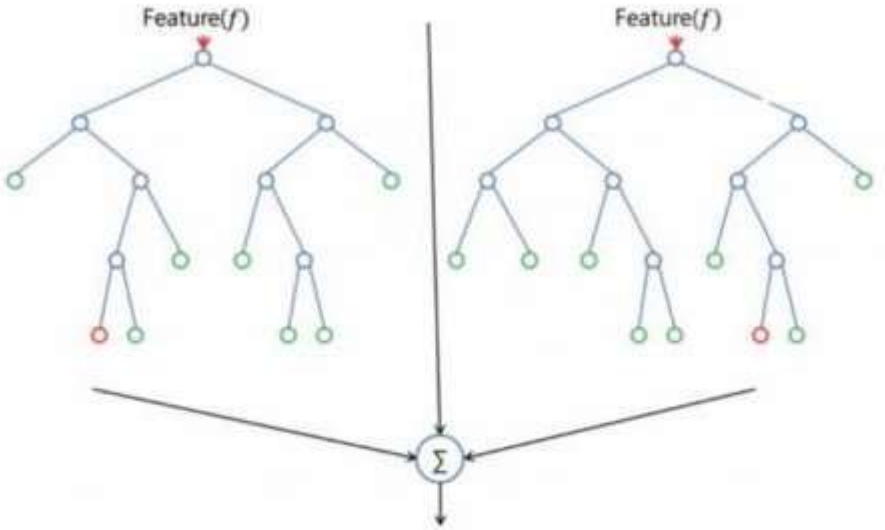


Fig 1. Random Forest

COST FUNCTION

The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value[5]

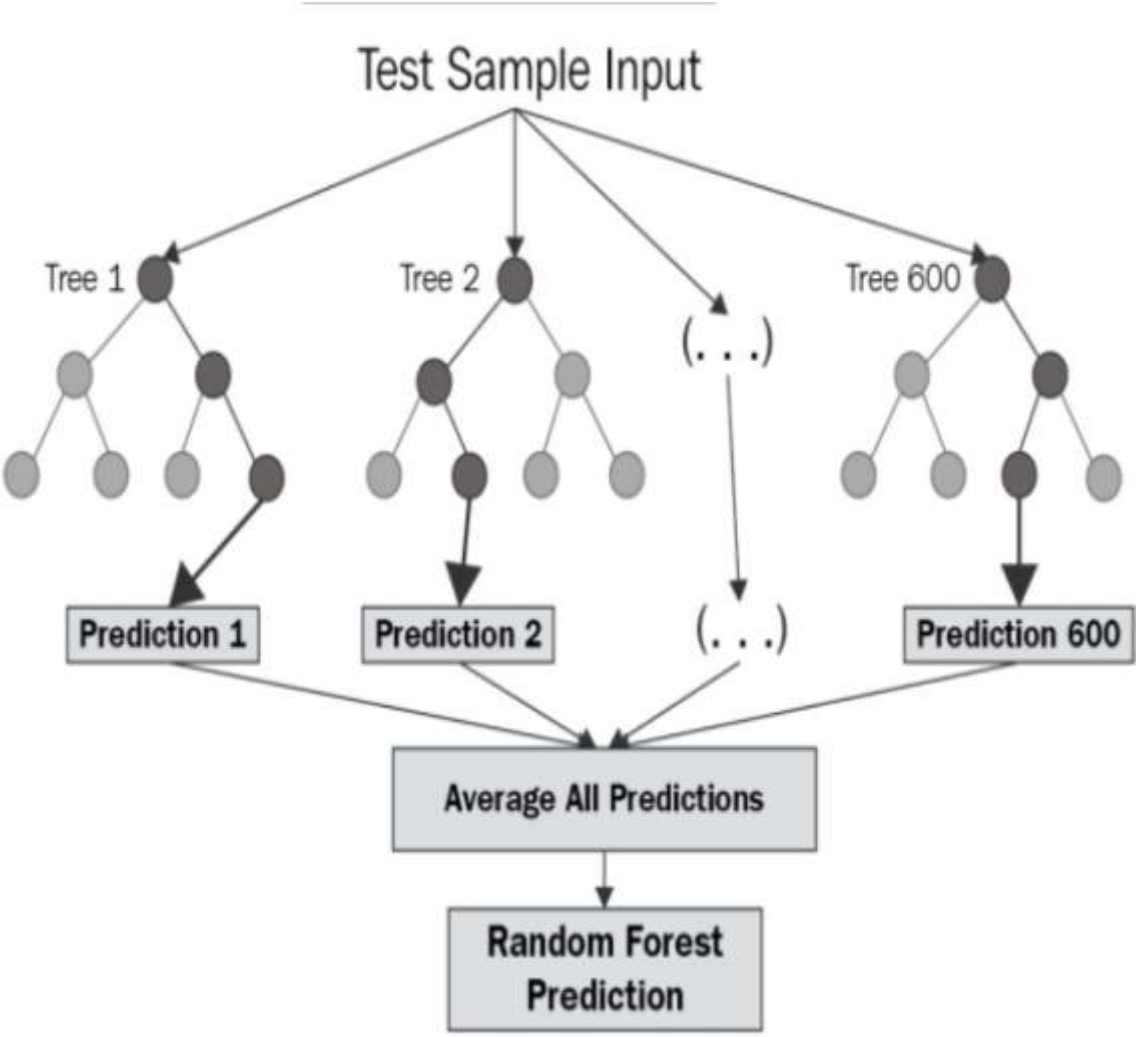
$$\text{Minimize } \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$\mathbf{J}= \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (prediction) and true y value (y).

RANDOM FOREST ALGORITHM

Random Forest is an adaptable, simple to utilize AI calculation that produces, even without hyperboundary tuning, an incredible outcome more often than not. It is likewise perhaps the most utilized calculations, in view of its effortlessness and variety (it very well may be utilized for both order and relapse errands). In this post we'll figure out how the arbitrary woodland calculation functions, how it contrasts from different calculations and how to utilize it



DATA ANALYSIS

This section performs the Selling price prediction using a dataset consisting of **8,128** used car details. This dataset is prepared by Cardekho.com and available on Kaggle.

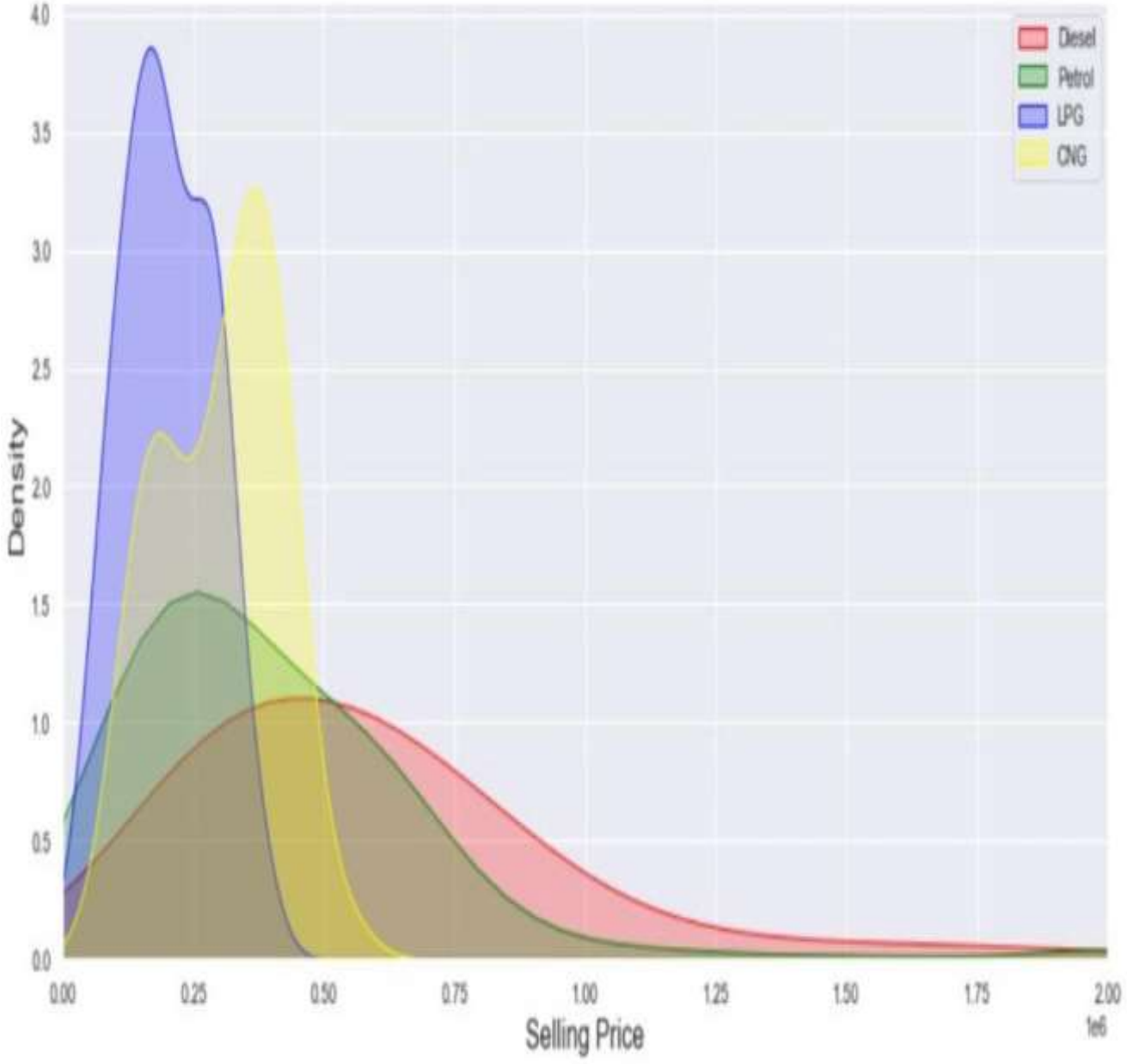
```
import pandas as pd
cars = pd.read_csv("car_data.csv")
cars.head(5)
```

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4	1248	74	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14	1498	103.52	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7	1497	78	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0	1396	90	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1	1298	88.2	5.0

PRICE DISTRIBUTION FOR DIFFERENT FUEL TYPE CARS

Let's visualize the price of cars versus the car fuel type.

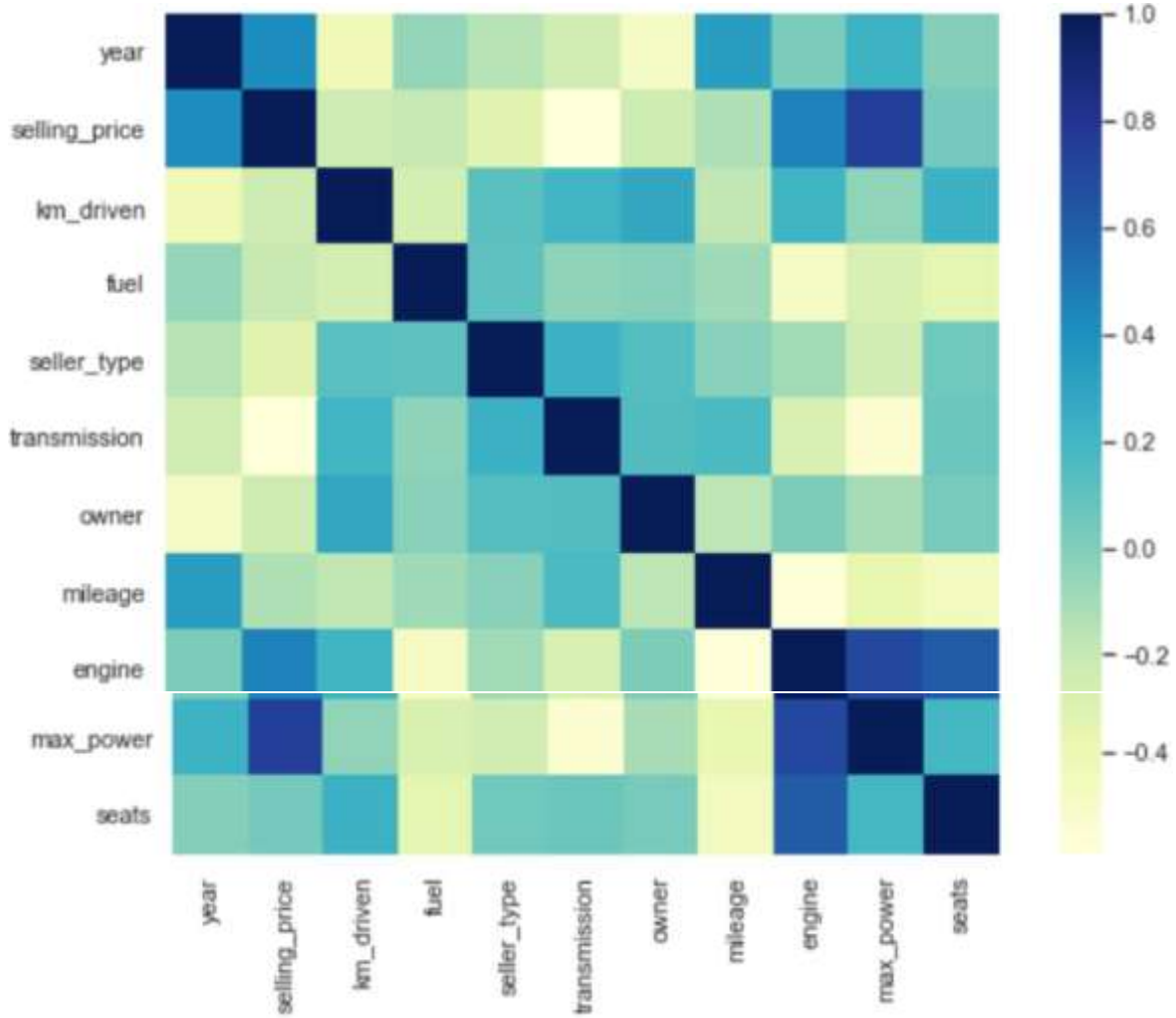
```
sns.kdeplot(cars.loc[(cars['fuel']==label), 'selling_price'],
            color=clr, shade=True, label=label)
```


Desel
Petrol
LPG
CNG
Density
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e6
Selling Price

CORRELATION MATRIX

Visualizing the correlations is an effective way of determining the dependencies. In the given plot, the selling price has a high correlation with the manufacturing year, engine, max power, and transmission. The Engine and Manufacturing year has the same approximate correlation, so we can select any one of them in the final set of features.

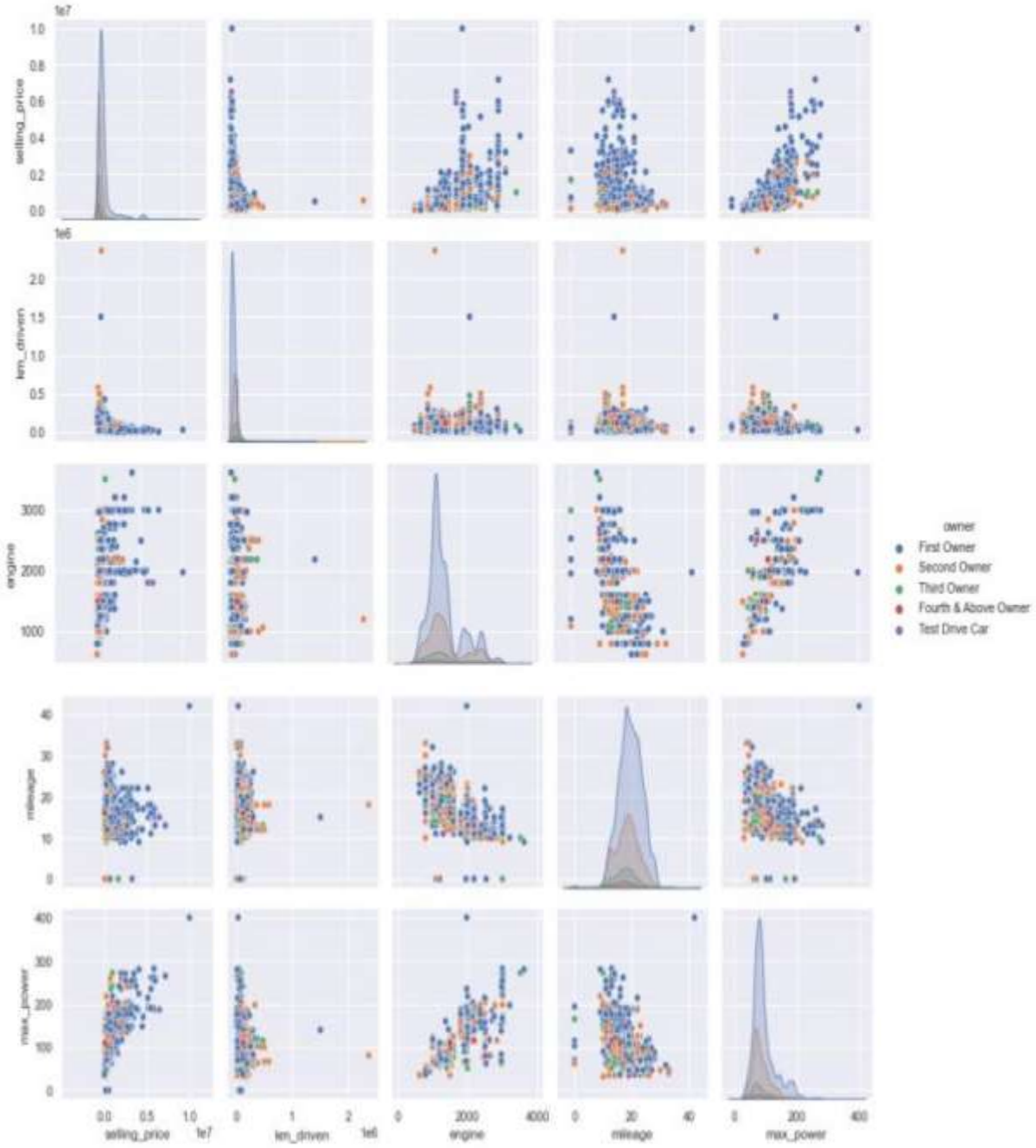
```
sns.heatmap(data = cars.corr(), cmap="YlGnBu", square=True)
```



PAIR PLOT

Pair plots are a great method to identify trends for follow-up analysis and, fortunately, are easily implemented in Python!

```
sns.pairplot(cars[["selling_price", "km_driven", "engine", "mileage",
                   "max_power", "owner"]], hue="owner")
```



DECISION TREE ALGORITHM

Decision Tree calculation has a place with the group of directed learning calculations. In contrast to other managed learning calculations, the choice tree calculation can be utilized for taking care of relapse and order issues as well The objective of utilizing a Decision Tree is to make a preparation model that can use to foresee the class or worth of the objective variable by taking in straightforward choice standards induced from earlier data(training information)

In Decision Trees, for anticipating a class mark for a record we start from the foundation of the tree. We analyze the upsides of the root characteristic with the record's quality. Based on correlation, we follow the branch relating to that worth and leap to the following hub.

Accuracy (e.g. classification accuracy) is a measure for classification, not regression so we can't calculate accuracy for a regression model.

For regression, one of the matrices we've to get the score (ambiguously termed as accuracy) is Rsquared (R2).

You can get the R2 score (i.e accuracy) of your prediction using [12] the score(X, y, sample_weight = None) function R2 score(accuracy) of our project is 76.65341530515258 %

Regression model accuracy calculated in following ways

1.R-Squared

2.Mean Absolute Error

3.Mean Squared Error

1.R-Squared of our project is 76.65341530515258 %

2.mean absolute error of our project is 3412.8022022817163
3.mean square error of our project is 47719623.3816024

REGRESSION	R2SCORE	RMSE	MAE
Linear	0.7665341530515258	6907.939155898986	3412.8022022817163
Decision	0.6693132403821778	8221.389576449885	4398.592078169706
Random	0.8614440569888814	5321.6879374429045	1725.187967350853

Table 1. Accuracy results of linear, decision and random forest algorithms

The above Table 1 contains the results of all three Regression algorithms Linear, Decision and Random for all the above algorithms we are calculating the R2SCORE, RMES and MAE errors.

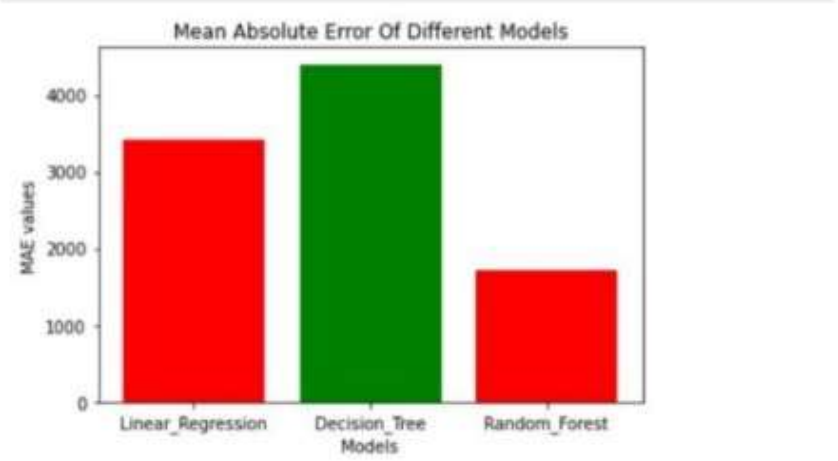


Fig 2. Mean Absolute Error of three models

The above Fig 2 shows the results of the mean absolute error of three models that is linear regression, decision tree and Random Forest models.

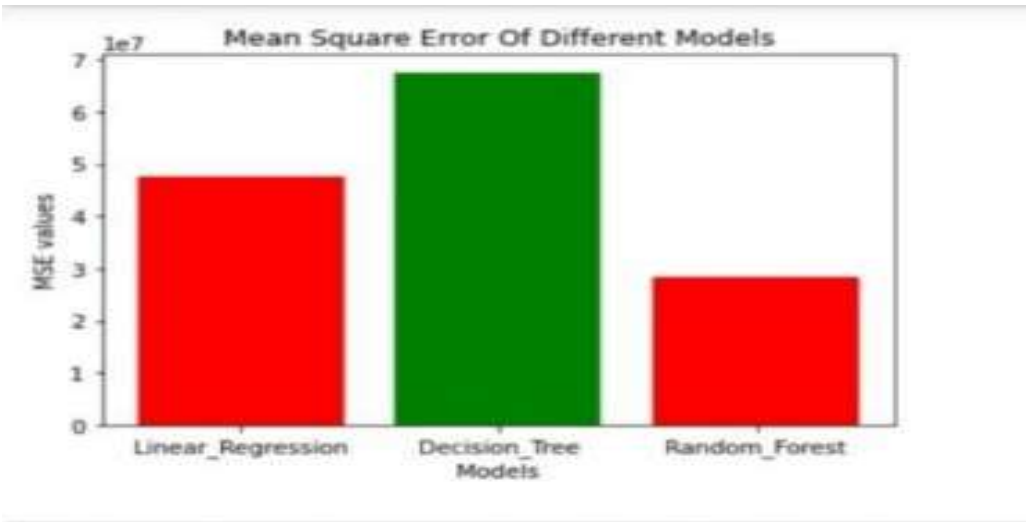
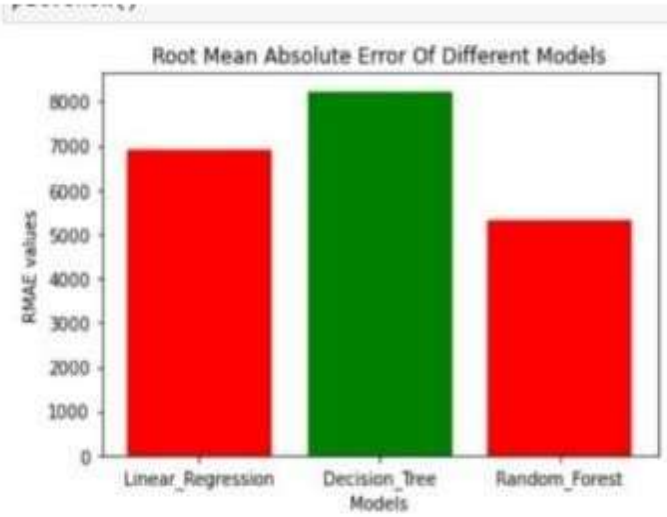


Fig 3. Mean Square Error of three models

The above Fig 3 shows the results of the mean square error of different models that is linear regression, decision tree and random forest algorithm.



PRE PROCESSING

In order to get a better understanding of the data, we plotted a histogram of the data. We noticed that the dataset had many outliers, primarily due to large price sensitivity of used cars. Typically, models that are the latest year and have low mileage sell for a premium, however, there were many data points that did not conform to this. This is because accident history and condition can have a significant effect on the car's price. Since we did not have access to vehicle history and condition, we pruned our dataset to three standard deviations around the mean in order to remove outliers.

We converted the Make, Model and State into one-hot vectors. Since we had over 2000 unique cities in the dataset, we

replaced the string representing the city with a boolean which was set if the population of the city was above a certain threshold i.e. a major city

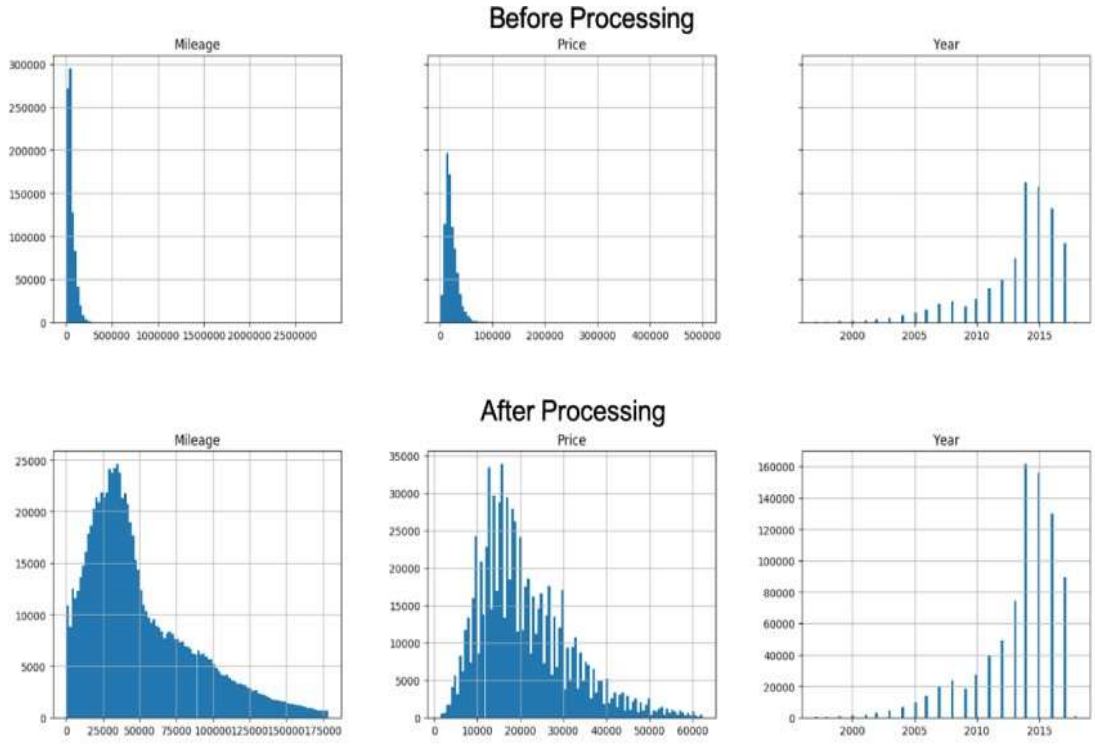


Fig 1: (Top) Histogram for raw data. (Bottom) Histogram after pruning.

Certain features such as VIN numbers dropped during training as these were unique to each vehicle, thus adding no value to training process. Apart from this, while implementing certain Machine Learning frameworks, we realised that certain categorical features had common values which causes problems while using frameworks such as XGBoost which require unique feature names (across all features). For eg: having the string “genesis” in both Make and Model features is not allowed. These common feature names were hence filtered out and renamed to work with these frameworks

ANALYSING LINEARITY IN DATASET

To analyze the degree to which our features are linearly related to price, we plotted the Price against Mileage and Year for a particular Make and Model. There seemed to be a fair degree of linearity for these two features.

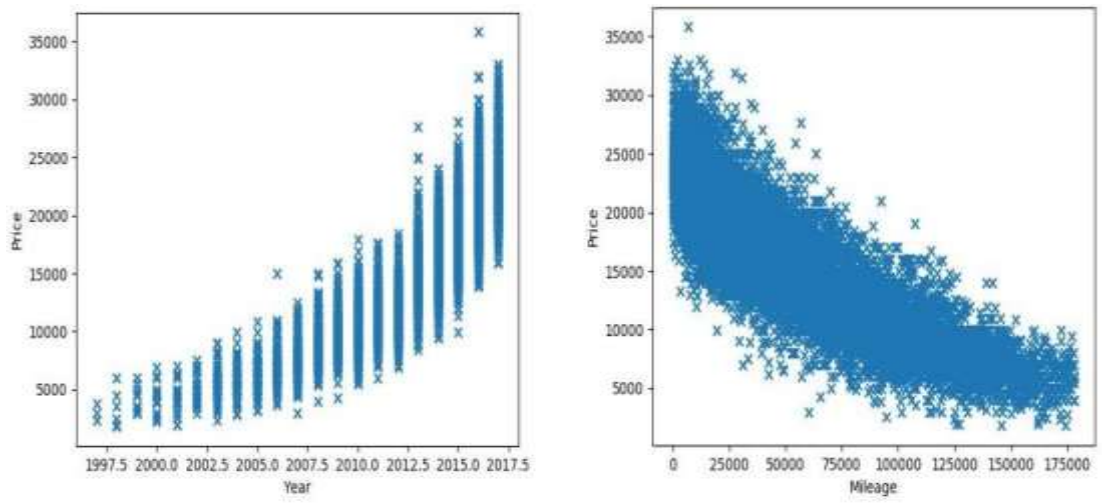


Fig 2: (Left) Price vs Year scatter plot. (Right) Price vs Mileage scatter plot.

FUTURE ENHANCEMENT

Currently, system can only deal with Swift Dzire cars due to lack of data. Also, data has been collected of only 5 cities of India. This can be extended to multiple car models and cities so as to improve accuracy and usability.

Efficient use of deep learning such as LSTM (Long shortterm memory) or RNN (Recurrent Neural networks) can be implemented once enough data is collected. This can improve accuracy and decrease RMSE drastically.

Currently, only few features are used to predict resale value of the car. This can be extended to more features. One can also implement CNN to determine physical condition of the car

from images like identifying dents, scratches etc. and thus predicting more relevant resale value of a car.

ACKNOWLEDGEMENT

I have no other words to express my sincere thanks to faculties of Indus University, Ahmedabad for their kind cooperation and able guidance. Especially to Mrs. Sejal Thakkar, my project guide in college without whom the project could not be executed.

CONCLUSION

However, once more data is collected and various different cars are included in the system, deep learning-based ANN or LSTM would perform better. But currently, GBR based car valuation system can predict resale value of a car with Root Mean Squared Error (RMSE) of 50,000 INR.

REFERENCES

1) Pudaruth, S., 2014. “Predicting the Price of Used Cars using Machine Learning Techniques.” Vol 4, Number 7 (2014), pp. 753-76.

2) ijictv4n7spl_17.pdf (ripublication.com)