Predicting Used Car Prices using Machine Learning and Deep Neural Networks

1st Guryash Singh Dhall
Faculty of Computer Science
Dalhousie University
Halifax, Canada
guryash.dhall@dal.ca

2nd Smriti Mishra
Faculty of Computer Science
Dalhousie University
Halifax, Canada
sm689498@dal.ca

3rd Chanpreet Singh
Faculty of Computer Science
Dalhousie University
Halifax, Canada
c.singh@dal.ca

4th Vinay Vilas Patil
Faculty of Computer Science
Dalhousie University
Halifax, Canada
vinay.patil@dal.ca

***Abstract*—The widespread increase in the car market industry makes it difficult for newly manufactured cars to reach buyers for several reasons. Several factors include heavy car costs for both used and newly manufactured cars, factors such as a lack of supply, finances, and others. As a result, the used automobile industry is growing rapidly all over the world, But in the outside world, the used car industry is majorly managed by the private sector. This creates the potential for deception when purchasing a used automobile. Thus, a very accurate model that can predict the cost of a used car without favoring either the consumer or the merchandiser, or any other entity is a must. This project aims in building a Machine Learning model using both a supervised modeling approach and a deep neural network (DNN) approach. We have developed a working model using several supervised algorithms as well as a DNN algorithm and obtained the models with the least error value, after which, we have compared these models to find the best optimal model to predict the used car prices. For this project, we have used Craigslist Dataset which consists of data related to used cars. A DNN model is built using the dataset as well as several supervised algorithms like Random Forest, XGBoost, Linear Regression, and Light GBM. The results showed that out of all Random Forest performed the best on the dataset with an R2 Score value of 0.90, as well as the Mean Absolute Percentage Error on the test data, which was around 17 percent. Additionally, as the dataset gets updated the model can perform more effective results and help in devising the used car sales startegy.**

***Index Terms*—Used Cars Price Prediction, DNN, Machine Learning, Linear Regression, Random Forest, Light Gradient Boosting Machine (GBM), XGBoost, Regression**

I. Introduction

The Craigslist dataset is the world's largest collection of used cars which are for sale. This dataset is built using a web scrapper, and it includes all used car records in the United States. This dataset is refreshed using that scrapper every few months. It includes columns like price, condition, manufacturer, latitude/longitude, and 18 other categories, and it comprises the majority of the essential information that Craigslist provides on auto transactions [1].

Talking about the features of the data. The data contains the appropriate features to be categorized as **big data**. **Volume** is one of the features of this dataset, as the size is 1.45 Gigabytes. Another feature is **Variety**, as the data consists of values relating to all segments of the automotive industry, which are vital in deciding the price of a car, such as model, make, colour, year of manufacturing, engine, etc., along with this the data is collected and merged from different regions in the United States. Lastly, the data gets refreshed every few months, which gives an idea about the **Velocity** present in the data.

The aim of using this dataset is to get a better idea of the car market in United States. Along with this we can understand that the perspective from both the customer and seller. Seller can make better sales strategy and customer can make better purchase decisions. Using the model and analysis we can answer the common questions a customer may ask.

II. Previous Work (Literature Review)

Numerous disciplines, including corporate intelligence, environmental modelling, and financial forecasting, frequently use regression analysis techniques.

[2] Mustafa used Avito to collect data as well he and his team wrote a python script that would scrape data of used cars from different websites using Beautiful Soup. Initially they applied simple Linear regression and then they moved to multiple linear regression models like K nearest neighbor regressor, gradient boosting regressor, Random Forest regressor and also applied Artificial Neural Networks on the data. They evaluated the results using r2 score and root mean squared error and they found that for the algorithm Gradient Boosting Regressor the score was the highest. In terms of hyper parameters in GBR the loss was 'ls' and the max depth was set to 6. The worst performing algorithm for them was MLR which gave a R2 score of 0.57.

[3] Janke, Jahnavi and Dr Laxmi used Deep Neural Network, Lasso Regressor, Linear Regression, Ridge Regression and Random Forest Regressor on their used car data set. They used R squared error and mean absolute error as their evaluation metrics. Their best performing algorithm was Random Forest Regressor which gave a R2 score of 0.97 and their worst performing algorithm was the Ridge Regressor which gave a R2 score of 0.80 and a MAE of 1.143.

III. Data Description

A. Description of the features in the data set.

The total size of the dataset is around 1.34 GB. The dataset has a total of 4,26,880 rows and 26 columns. The dataset

has the following columns id, URL, region, region URL, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title status, transmission, VIN, drive, size, type, paint color, image URL, description, county, state, lat, long, posting date. The Kaggle website from where the dataset is being taken, says that the dataset is updated every few months.

Price: This column gives information about the listed price of the used car on the website. The price column has no null values. The maximum value in the dataset is 3736928711 and the lowest value in the dataset is 0.

Year: This column gives information about the year in which the car was manufactured. The oldest car is manufactured in 1900 and the latest is in 2022 manufactured. The number of null values in this column is 1205.

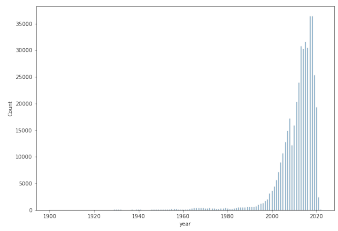

Fig. 1. The figure shows the count plot of cars based on the year

Manufacturer: This column gives information about the manufacturing company of the car. The column has 17645 null values which account for 0.04 percent of the entire data set with the highest number of cars from the manufacturer named 'ford'.

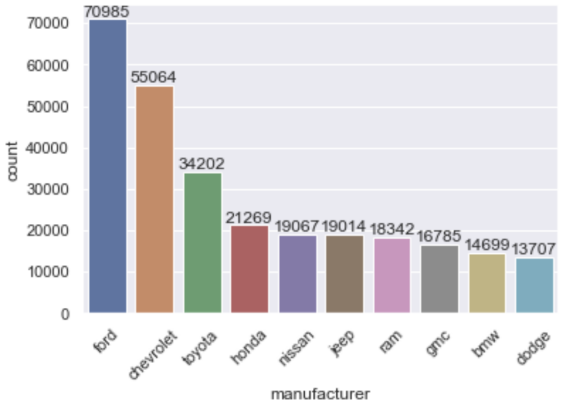

Fig. 2. The figure shows the count plot of Manufactures

Condition: This column gives information about the condition of the car, whether it is good, fair, new, or excellent in condition. The column has 174104 null values.

Model: This column gives information about the model's name of the car.

Cylinders: This column gives information about the number of cylinders present in the dataset. The column has 177678 null values.

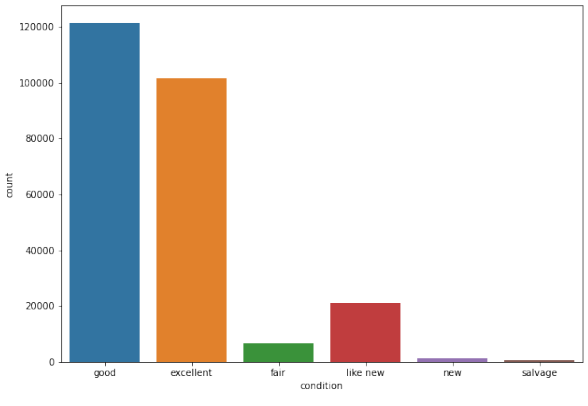

Fig. 3. The figure shows the Count Plot of Condition of the cars.

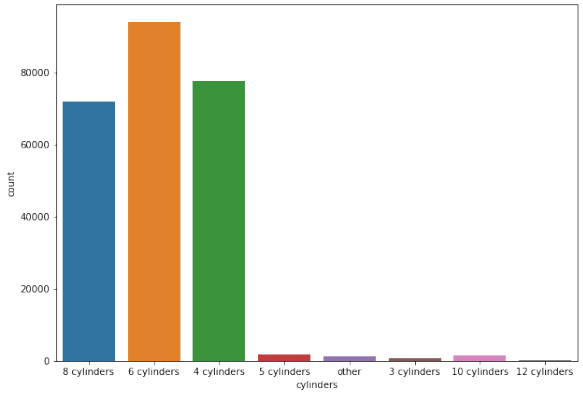

Fig. 4. Count plot of Number of Cylinders in Cars.

Drive: This column gives information on whether the car is a 4-wheel drive, front-wheel drive, or rear-wheel drive. This column has 130567 null values.

Odometer: This column gives information about the total distance traveled by car. The column has 4400 null values.

Fuel: This column indicates the fuel type of the car if the car is a gas, diesel, electric, or hybrid car.

State: The state is a political region that is briefly depicted in the data collection. like how Florida's state abbreviation is "fl."

Latitude and Longitude: The location of the car's sale may be determined.

Posting Date: This column gives details about the date on which it was posted on the website.

Paint Color: The column gives information about the color of the car.

B. Data Visualization

For our study, we worked to comprehend each column's pattern and values, as well as to answer a few questions that

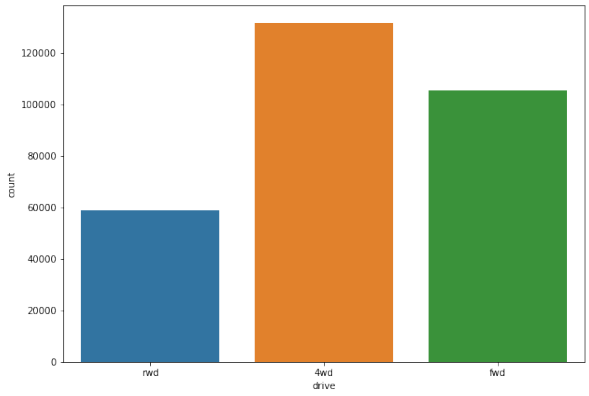

Fig. 5. Count of the car drive type.

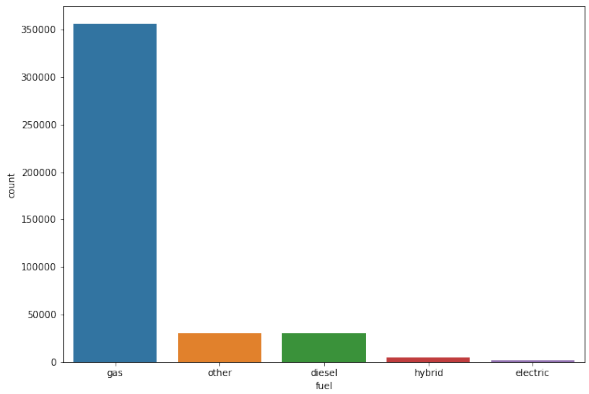

Fig. 6. Number of different fuel types of Cars.

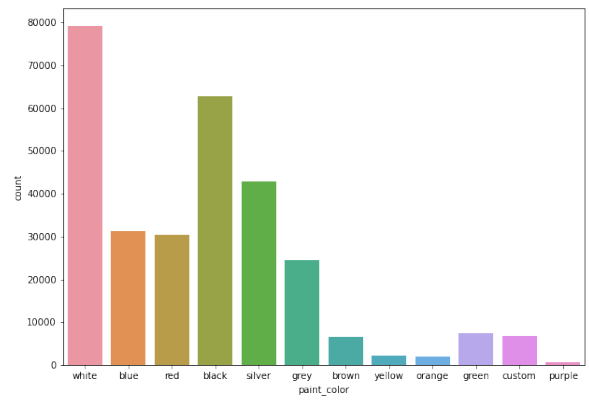

Fig. 7. Count plot of the available cars in different colors.

would assist us in providing a strong understanding of our project. We started by classifying the columns into categorical and continuous data [4].

Continuous data: id, price, year, odometer, country, lat, long

Categorical data: url, region, region url, manufacturer, model, condition, cylinder, fuel, title status, transmission, VIN, drive, size, type, paint color, image url, description, state, posting date.

Before getting to the continuous and categorical data, let's have a look at the number of vehicles that were produced from the year 1900 to 2022 as our research is focused on the car data set. To depict the pattern of how many cars were developed in one set of years and how that trend has changed in another set of years, we have divided the year into two phases [5].

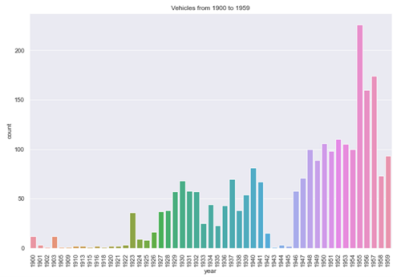

Fig. 8. Number of cars from year 1900 to 1959.

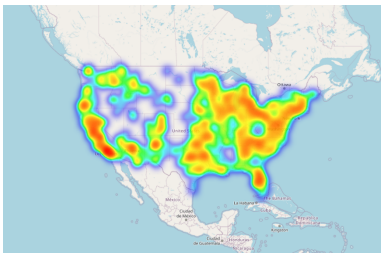

Fig. 9. Histogram of cars from year 1900 to 1959 in US.

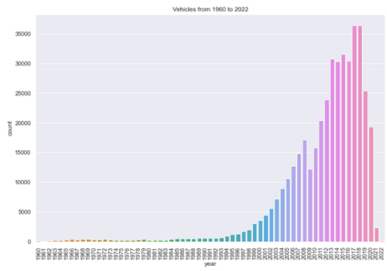

Fig. 10. Number of cars from year 1960 to 2022 in US.

The figures Figure 8,9,10 and Figure 11 above demonstrates how the number of cars has risen from set 1 to set 2. The lowest number of cars were produced in Set 1 during the years of 1900 and 1922. Gradually, this number rose and it peaked to the maximum in 1955. Similarly set 2 between 2015 and 2018 saw the largest peaks in car counts. However, it is also noted that this pattern abruptly changed in the years 2021 and 2022.

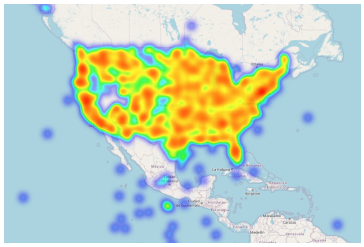

Fig. 11. Histogram of cars from year 1960 to 2022 in US.

After understanding the growth trend of the vehicles, we can now discuss about the continuous data and correlation among those data.

Continuous Data In order to examine the maximum and minimum values, any outliers that might be found, and the pattern of distribution of the values in each column, we have built histograms for the continuous data.

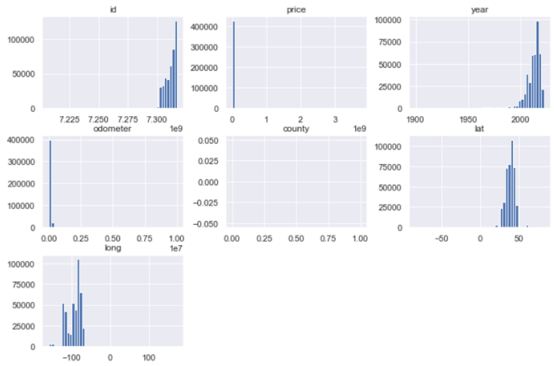

Fig. 12. Histogram of continuous data.

According to the preceding figure 3, ID ranges from 7.30 1e9 to 7.32 1e9. The incremental trend is displayed in the id column. The year column, which spans from 1900 to 2022, is located after that. The frequency of this year-specific column is progressively rising and will peak around 2020. An odometer is a device used to measure a vehicle's travel distance. It falls between 0 and 1 e7. When the odometer is close to zero, which indicates the automobile hasn't been driven before, the frequency is higher. Lat and long are the two columns that are more crucial since they make it easier to link the coordinates of various locations throughout the world. Latitude extends from 25 to almost 70, and longitude from -150 to -70. There are some columns that are empty and will be removed during the cleaning process.

To understand the correlation and relationship of the discussed column we have designed the Pearson correlation

Categorical Data

While working on the categorical data we wanted to answer few questions like:

- What are the top 10 regions from where the cars are manufactured?.
- What is the mean car price for each manufacturer?.

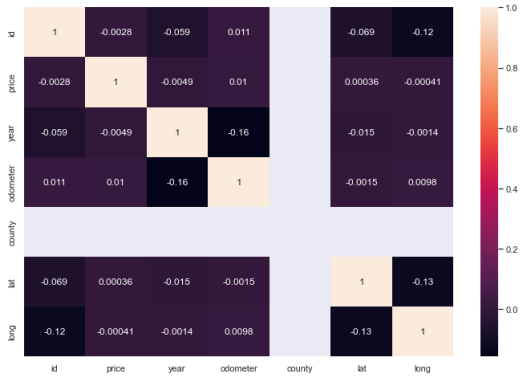

Fig. 13. Pearson Correlation of different continuous columns.

- Which condition of the car has the highest price in the market?.
- What is the relationship between the car's condition and the odometer?
- How many types of cylinders we have, and which one is the costliest ones?

These are some set of questions which we have targeted when studying through the categorical data. This helped us to understand the major and more impacting features which will help us later in the flow creation.

What are the top 10 regions from where the cars are manufactured?.

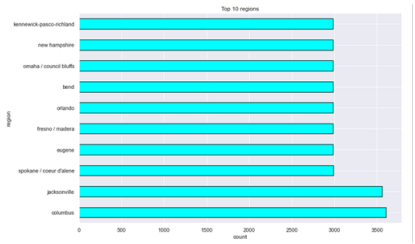

Fig. 14. Top 10 regions.

What is the mean car price for each manufacturer?.

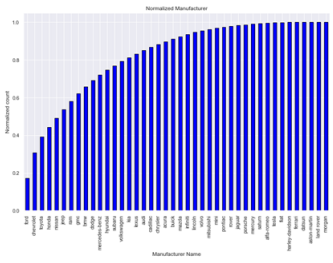

Fig. 15. Normalized Manufacturer.

Which condition of the car has the highest price in the market?.

What is the relationship between the car's condition and the odometer?

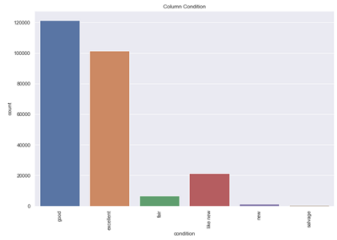

Fig. 16. Count of cars with various conditions.

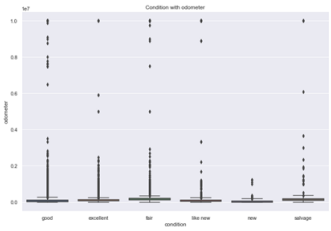

Fig. 17. Condition and odometer.

How many types of cylinders we have, and which one is the costliest ones?

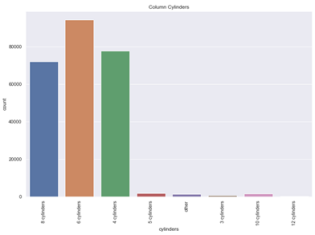

Fig. 18. Car counts on different cylinder types.

IV. Data Processing

As a part of data processing for this dataset we have performed data cleaning finding out null values, duplicate records, and then removed the outlier from the target feature. In case of cleaning the null values, features with more than 40% of null values were dropped and the imputation for odometer was done. For rest all null values, records were dropped.

A. Data Cleaning

Data cleaning is a technique used in this dataset to ensure that the data we are using for analysis is fully prepared, i.e., that it is free of duplicates and missing values, that it is in the correct format, that it is not corrupted, and that it is thus ready to be used for analysis.

First, we have converted the dataset from its object notation to string/float. Then we checked the null values in the data. Out of the total 26 features, 20 features have null values, but many features have more than 40% null values, so we will not be using them as they can distort the analysis and the model. The total features that we have after removing the features having more than 40% null values are 22 features. After that, we checked for duplicate records in the dataset and found none.

	null	percent
county	426880	1.00
size	306361	0.72
cylinders	177678	0.42
condition	174104	0.41
VIN	161042	0.38
drive	130567	0.31
paint_color	130203	0.30
type	92858	0.22
manufacturer	17646	0.04
title_status	8242	0.02
lat	6549	0.01
long	6549	0.01
model	5277	0.01
odometer	4400	0.01
fuel	3013	0.01
transmission	2556	0.01
year	1205	0.00
description	70	0.00

Fig. 19. Null values and their percentages

In the following dataset we checked for target feature ‘price’ and found out that there are many outliers present in the dataset. The ‘price’ feature not only has zero values as the minimum but also the highest values go beyond 1000k which is pretty much unfeasible. Looking at the distribution of the data and numerous outliers present in the price feature we observed that the dataset is highly unbalanced. Hence, we followed an approach to get the data from 25% to 75% of the IQR (Inter-Quartile Range). Below is the box plot showing the data after applying the above filter [21].

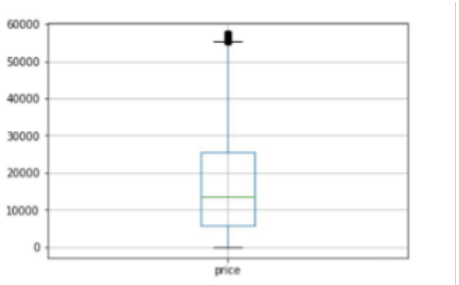

Fig. 20. Box plot of Prices after filtration

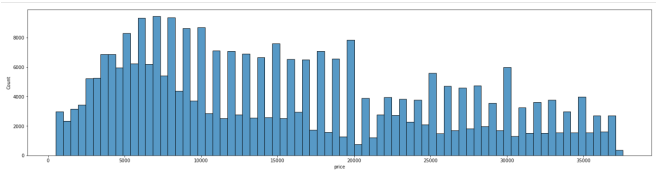

Fig. 21. Histogram of Price in Analytical Base Table

After this another cleaning operation that we performed was based on the feature ’odometer’. Odometer is very important factor to be considered in case of price prediction as it helps to understand how the condition of the car is and what

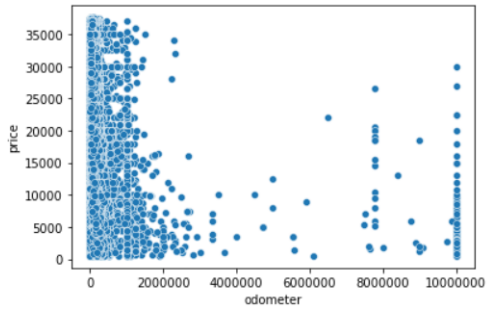

Fig. 22. Odometer scatter plot with raw data

should be its deprecated value according to the current market scenario. Hence, we checked the null values in the odometer and replaced it by the mean value from each segment derived from 'condition'.

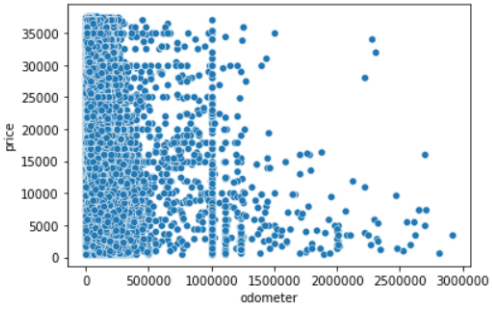

Fig. 23. Odometer scatter plot after cleaning

So, we are diving the data set as per condition and then taking the mean of each division. Mean of the condition=good is used to impute the null values where the condition is good. Using this we are able to fill out the null values in this feature.

For rest of the 19 features, if they have some null values, those records are deleted. So that it doesn't affect the model. After performing all this cleaning, we have our final Analytics Base Table (ABT) ready which has 129447 records and 22 features. We will be performing further preprocessing and feed this to our model.

B. Pre processing

Our data set includes text or category values in features. And for the model to work need numerical inputs. Only a few algorithms, such as catboost and decision-trees, can effectively handle categorical data, but the majority of algorithms rely on numerical values to produce cutting-edge outcomes.

As a result, our key difficulty is to transform textual or category input into numerical data while still creating an algorithm or model to make sense of it. Deep learning's foundational technology, neural networks, assumes that input values will be numerical.

We are using **label encoding** to perform this task of conversion from categorical column to numerical column. This strategy is relatively straightforward and entails turning each value in a column into a number. Label encoding, which relies on number sequencing, can provide a new issue depending on the data values and type. The issue with employing numbers is that they introduce comparison and relationship between them.

C. Feature Selection

Currently we have 22 features in the Analytical Base Table (ABT) and before using all the features in the model, we need to identify which features have high importance and greater influence on the model.

For this we ran a feature selection method: Random Forest Regressor [6]. One of the most widely used machine learning techniques is random forests. They are highly effective because they often have good predictive accuracy, little overfitting, and are simple to interpret. The fact that it is simple to determine the significance of each variable on the tree decision contributes to this interpretability. In other words, it is simple to calculate the percentage of the decision that each variable contributes.

Using a random forest to choose features falls under the heading of embedded approaches. Filter and wrapper techniques' advantages are combined in embedded methods. They are put into practice by algorithms with built-in feature selection techniques. Finally, we selected 16 best features out of 22, to be used for the model.

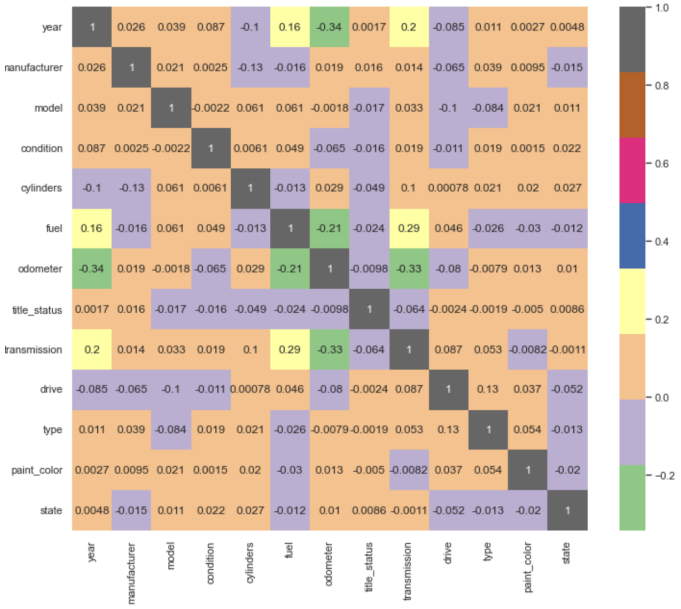

Fig. 24. Pearson Correlation

Along with the random forest regressor, we used pearson correlation to find out if the features in the dataset are highly correlated. A value between -1 and 1 known as a Pearson correlation describes the degree to which two variables are linearly connected. The "product moment correlation coefficient" (PMCC) or simply "correlation" are other names for the Pearson correlation. Only metric variables are appropriate for Pearson correlations. Values for the correlation coefficient

range from -1 to 1. A number that is nearer 0 indicates a weaker association (exact 0 implying no correlation). Closer to 1 value indicate a stronger positive association. A value nearer to -1 denotes a more significant negative correlation. We kept the threshold as 0.7 and found out that none of the features in the dataset are highly correlated.

V. Solution and Model Building

Based on the research we did on the previous works we were able to find that majority of the research word done on similar datasets for used cars was performed using both supervised and neural networks. The earlier researchers used mainly Random Forest and Deep Neural Network in building the model for the data set.

In our innovative approach we thought of using other types of models to understand how it performs on our current dataset. We used Linear Regression, Light GBM and XG-Boost.

We used R squared error to check the performance of the models as well as it formed our basis of model comparison and statistical significance test.

A. Solution and Data Flow Diagram

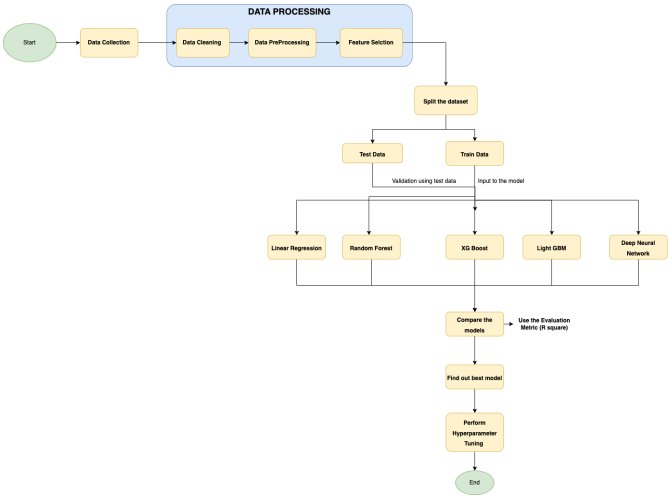

Fig. 25. Data Flow Diagram [9]

The data that we used is for used car in United States is refreshed every few months and then after we read that data. The next step performed is Data Processing, this step involved Data cleaning, Data pre-processing (in which we converted the categorical data into numerical data) and Feature Selection. We used Random Forest regressor to select the optimum 12 features that has the highest influence on our target feature 'price'. Next step that we performed is splitting the dataset into test (30%) and train data (70%). Then we used the train data as an input to our models. We have used 5 models: Random Forest, XG Boost, Light GBM, Linear Regression and Deep Neural Network. After which we checked the performance of models via checking the predicted output on test data, using the evaluation metric R-square. This helps us find out the best model for this dataset. Then at the end we perform hyperparameter tuning to get the best parameters for this model.

B. Solution based on literature review

Random Forest Regressor

Random Forest works both as a classifier and a regressor, that uses a variety of decision trees on different subsets of the input data. For the most precise forecasts, the information from various trees is then combined [10]. While an autonomous decision tree only offers a single conclusion and a small number of groups, the forest ensures a more accurate result with a greater variety of groups and decisions. Using many models that have been trained on the same data and averaging their findings to provide a more accurate prediction or classification is known as ensemble learning which is used in Random Forest.

For this dataset we applied random forest on the cleaned data set using default values. For the test data we were able to get a R2 Score of 0.90, root mean squared error of 2922.14 and mean absolute percentage error of around 0.16.

TABLE I
RANDON FOREST METRICS

R2 Score	MAPE	MAE	RMSE
0.90	0.16	1676.90	2922.14

Deep Neural Networks

A layered association of neurons connected to other neurons is a deep neural network. These neurons transmit signals or messages to other neurons depending on received information and a network structure that uses a feedback mechanism. The first layer received input, which the neurons in it used to produce output that is sent to the next layers based on the activation function employed. Each layer has one or more neurons, each of which has an activation function that determines whether the neuron should be turned on or off by computing a weighted aggregate and then adding bias to it [11].

Below is the architecture of the neural network used in building the model Below are the Training and Validation Losses with Respect to MAE during the DNN training process.

We were able to achieve a R2 Score of -0.37 and root mean squared error of 7929.94

TABLE II
DNN METRICS

R2 Score	MAPE	MAE	RMSE
-0.37	0.64	5607.32	7929.94

C. Our Approach

In our approach we planned to use Boosting and Decision Tree based algorithms which would function similar to approaches used in the literature review.

XGBoost

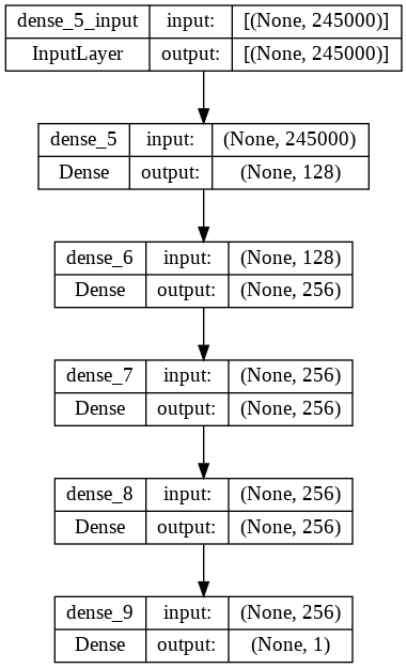

Fig. 26. Model Architecture

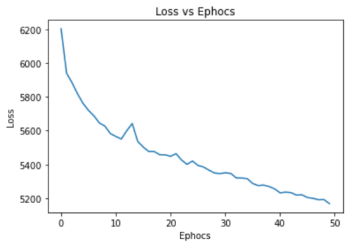

Fig. 27. Training Loss vs Ephocs

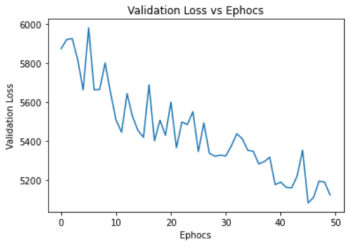

Fig. 28. Validation Loss vs Ephocs

The gradient boosting framework is used by the decision-tree-based ensemble machine learning method known as XGBoost. Both XGBoost and Gradient Boosting Machines (GBMs), ensemble tree approaches, use the gradient descent architecture to boost weak learners (CARTs in general). But XGBoost enhances the fundamental GBM architecture with system optimization and algorithmic improvements [12].

For this dataset we applied XGBoost on the cleaned data set using default values. For the test data we were able to get a R2 Score of 0.88, root mean squared error of 2125.08 and mean absolute percentage error of around 0.20.

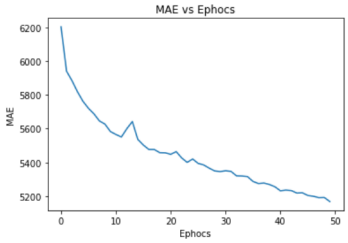

Fig. 29. MAE vs Ephocs

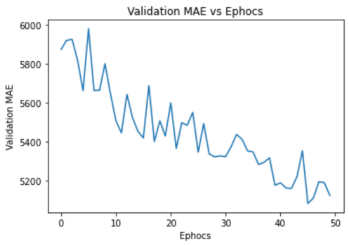

Fig. 30. Validation MAE vs Ephocs

TABLE III
XGBOOST METRICS

R2 Score	MAPE	MAE	RMSE
0.88	0.20	2125.08	3166.86

Linear Regression

Linear regression is a supervised machine learning model that identifies the linear connection between the dependent and independent variables by determining the best fit linear line between them [13].

For this dataset we applied XGBoost on the cleaned data set using default values. For the test data we were able to get a R2 Score of -0.32 and root mean squared error of 7553.77 and mean absolute percentage error of around 0.94.

TABLE IV
LINEAR REGRESSION METRICS

R2 Score	MAPE	MAE	RMSE
-0.04	2.16	5110.43	6852.00

Light GBM

LightGBM is a gradient boosting framework that improves model performance while using less memory.

The primary technique employed in all GBDT (Gradient Boosting Decision Tree) frameworks, histogram-based, is replaced by gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which addresses its drawbacks [14].

TABLE V
LIGHT GBM METRICS

R2 Score	MAPE	MAE	RMSE
-0.04	2.16	5607.32	7929.94

Random Forest with Randomized Search CV, Hyper Parameter Tuning

Out of the all the models built we observed that we were able to get the highest R2 score for Random Forest Regressor, so we performed Random Forest along with Randomized Search CV to get the best hyper parameters [15].

On performing Hyper Parameter tuning on the data we were able to find the following best parameters

- N Estimators: 50
- Min Sample Splits: 6
- Min Samples Leaf: 1
- Max Features: SQRT
- Max Depth: 40
- Bootstrap: False

For this data set we applied RF with Hyper Parameters on the cleaned data set. For the test data we were able to get a R2 Score of 0.82 and root mean squared error of 3625.31 and mean absolute percentage error of around 0.17.

D. Solution Using Sub Sample of Data

In order build the model on sample data, we sub sampled around 30 percent of data from the cleaned dataset which we obtained. Based on the research and discussion we sorted the values of the model based on the year in which the car the manufactured. We then divided the model in the ratio of 85:15 where the training size of the sample is 30000 records, and the testing size is 5000 records. We trained the model using Random Forest Regressor and tested the model on the test data and we were able to obtain a R squared score of 0.37 and a Mean Absolute Error of 3693.49. The model was able to closely predict around 83 percent of the values.

Metrics For: **Sampled Random Forest**
Mean Absolute Percentage Error: 0.17232569659473007
Mean Absolute Error: 3693.4893836978363
Mean Squared Error: 26259583.157638967
Root Mean Squared Error : 5124.410518063416
R2 Score: 0.37525936828513917

VI. Model Evaluation, Analysis and Comparison

A. Accuracy Metrics

Since this is a regression problem, we have used R squared error, mean absolute error, mean absolute percentage error, mean square error and root mean square error as our evaluation metrics for this model [7].

R Squared Error

R-squared shows the percentage of variation of the response variable's actual value that the regression model has captured. The coefficient of R2 is a measurement that tells us how well a model fits the data. It is a statistical indicator of how closely the regression line resembles the real data in the context of regression. Therefore, it matters whether a statistical model is employed to make predictions about the future or test hypothesis [16].

Mean Absolute Error (MAE)

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}, \\ = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Fig. 31. R Squared Error Formula [16]

The degree of mistake in your measurements is known as absolute error. It is the difference between the measured value and the "actual" value. The Mean Absolute Error(MAE) is the average of all absolute errors [18].

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Fig. 32. Mean Absolute Error Formula [18]

Mean Squared Error (MSE)

The degree of inaccuracy in statistical models is evaluated by the mean squared error, or MSE. Between the observed and projected values, it evaluates the average squared difference. The MSE is equal to 0 when a model is error-free. Its value grows when model error does as well [17].

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Fig. 33. Mean Squared Error Formula [17]

B. Data Balance / Imbalance

When it comes to data imbalance, having uneven data can cause a lot of problems. When building machine learning models on imbalanced data we see that models tend to predict the values which have a high frequency of occurrence in the dataset. This is a major case when it comes to binary classification problems. Bias in the data always ignores the data which has a low occurrence [8].

When it comes to regression, data imbalance is less visible, so we must analyze the target variable.

We examined the target feature "price" in the given dataset and discovered that it has a significant number of outliers. The minimum and maximum settings for the "price" feature are both zero, which makes it almost impossible to use. We noticed that the dataset is quite uneven by observing the distribution of the data and the large number of outliers included in the price feature. Consequently, we used a method to obtain the data from 25

C. Comparison of Models

On comparing all the models based on the evaluation metrics we were able to find that the Random Forest model performed the best out of all models and gave the lowest error

rate. The mean absolute percentage error for RF was around 17 percent and the R squared error was 0.82 [19].

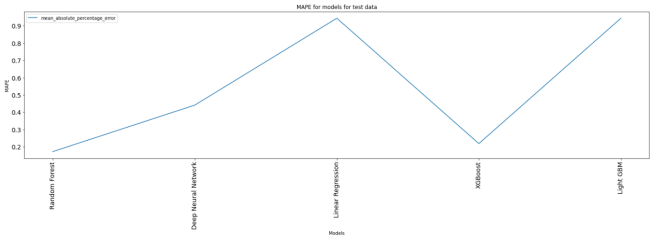

Fig. 34. Mean Absolute Percentage Error Line Graph

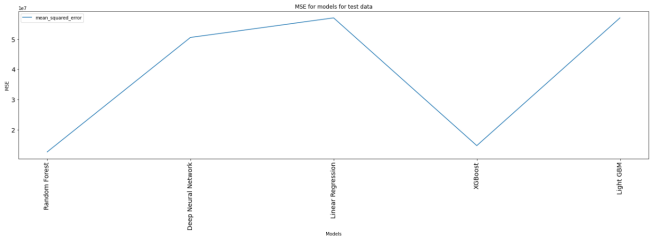

Fig. 35. Mean Squared Error Line Graph

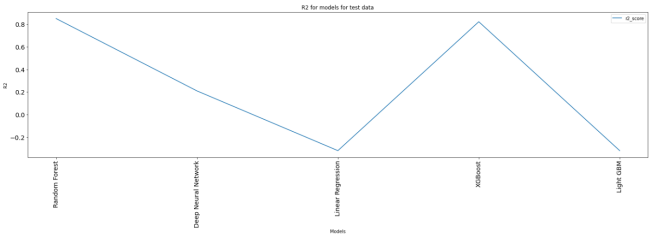

Fig. 36. R Squared Error Line Graph

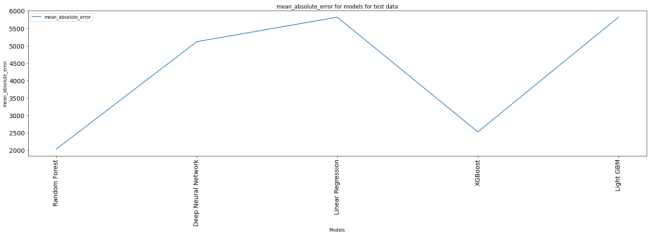

Fig. 37. Mean Absolute Error Line Graph

From the above figure 30,31, 32 and 33 we can see that Random Forest and XGBoost were the best performing models with a slight difference in their metrics, out of which Random Forest perfomed the best among all the algorithms used.

D. Statistical Significance Test

When developing a model on a dataset it is very important to build models on various algorithms. Comparing machine Learning models help in identifying the best optimal algorithm. K fold cross validation is used mainly in comparing

TABLE VI
EVALUATION METRICS FOR MODELS

Model	R2 Score	MAPE	MAE	RMSE
Random Forest	0.90	0.16	1676.90	2922.14
DNN	-0.37	0.64	5607.32	7929.94
XGBoost	0.88	0.20	2125.08	3166.86
Linear Regression	0.88	0.20	2125.08	3166.86
Light BGM	-0.04	2.16	5607.32	7929.94

models but it is sometimes difficult to identify accurate model based on statistics. In order to address this issue and estimate the probability of the skill score samples being seen on the hypothesis that they were taken from the same distribution; statistical significance tests were developed. It is implied that the difference in skill scores is statistically significant if the null hypothesis, or starting point, is rejected [20].

For this dataset we used R squared error as the basis of our hypothesis testing. Based on the scores of R squared error we found that Random Forest Regressor and XGBoost performed the best with scores of 0.89 and 0.87 out of which Random Forest Regressor was the best.

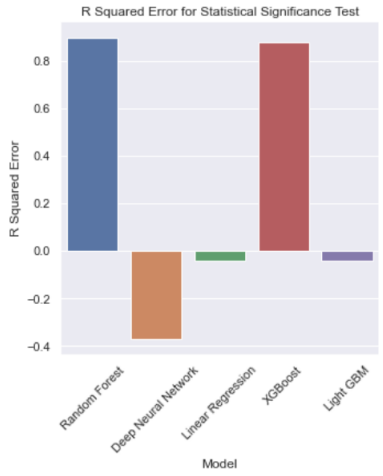

Fig. 38. Statistical Significance Test Statistics

E. Selection of Best Evaluation Metric

R Squared Error

The best evaluation metric selected for this problem is the R Squared Error. The main reason behind using this evaluation metric is that it gives the value of dependency on the feature. Calculating R Square involves dividing the whole sum of the squares representing the prediction error by the total sum of the squares representing the mean replacement. The R Square value ranges from 0 to 1, and a higher number denotes a better match between the forecast and the observed value.

How well the model fits the dependent variables may be assessed using R Square. However, it does not account for the over fitting issue. Because of the model's complexity, a regression model with several independent variables may fit training data quite well yet perform poorly on test data.

VII. Conclusion and Future Scope

To anticipate the price of a used car, this study tested numerous models. Because the random forest model had the highest R-square, which was close to 0.9, it was chosen. There is, however, still great space for development. To see if better outcomes can be attained, other models like Naive Bayes, LSTM, or Gradient Boosting algorithms can be used.

A larger data set is usually preferable. More training data can be provided to the model using a larger dataset. After data preprocessing, the research's dataset contained 12988 samples. A more precise model should be anticipated with a larger dataset. The work that will come after is anticipated to broaden the perspective on used cars, including their appearance and interior quality and engine functioning. Because there is no established grading system for these two elements, taking them into account can be extremely difficult. Finding a somewhat high grade is required.

Although the model developed here is limited to used automobile price prediction, it may also be used to any electric device or home appliance. The model can be linked to real-time websites whose data can be scraped, and it is trained using reinforcement learning on the dynamic dataset. Instead of using a limited dataset for training, the model can be expanded to train on data clusters. Large historical data sets can improve the model's accuracy. Using APIs (Application User Interfaces) like Heroku, REST, Git, etc., the model may be distributed on the web.

References

[1] A. Reese, "Used cars dataset," Kaggle, 06-May-2021. [Online]. Available: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data. [Accessed: 03-Nov-2022].

[2] "IEEE Xplore." [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp. [Accessed: 04-Nov-2022].

[3] "IEEE Xplore." [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp. [Accessed: 04-Nov-2022].

[4] Ismailsefa, "Used Cars Data Analysis and visualization (EDA)," Kaggle, 27-Jun-2021. [Online]. Available: https://www.kaggle.com/code/ismailsefa/used-cars-data-analysis-and-visualization-eda. [Accessed: 10-Nov-2022].

[5] "Used cars dataset analysis and machine learning using random Forrest Regressor," YouTube, 08-May-2022. [Online]. Available: https://www.youtube.com/watch?v=0xChHWhrg-s. [Accessed: 15-Nov-2022].

[6] A. Dubey, "Feature selection using Random Forest," Medium, 15-Dec-2018. [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f. [Accessed: 18-Nov-2022].

[7] S. Wu, "What are the best metrics to evaluate your regression model?," Medium, 05-Jun-2021. [Online]. Available: https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b. [Accessed: 19-Nov-2022].

[8] P. Brus, "Data Imbalance in regression," Medium, 01-Jun-2021. [Online]. Available: https://towardsdatascience.com/data-imbalance-in-regression-e5c98e20a807. [Accessed: 21-Nov-2022].

[9] "Diagrams.net - free flowchart maker and diagrams online," Flowchart Maker &; Online Diagram Software. [Online]. Available: https://app.diagrams.net/. [Accessed: 22-Nov-2022].

[10] N. Beheshti, "Random Forest regression," Medium, 02-Mar-2022. [Online]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d. [Accessed: 22-Nov-2022].

[11] M. AL-Ma'amari, "Deep neural networks for regression problems," Medium, 25-Oct-2018. [Online]. Available: https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33. [Accessed: 25-Nov-2022].

[12] J. Brownlee, "XGBoost for regression," MachineLearningMastery.com, 06-Mar-2021. [Online]. Available: https://machinelearningmastery.com/xgboost-for-regression/. [Accessed: 26-Nov-2022].

[13] Deepanshi, "Linear regression: Introduction to linear regression for data science," Analytics Vidhya, 25-May-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression. [Accessed: 28-Nov-2022].

[14] DataTechNotes, "LIGHTGBM regression example in Python," LightGBM Regression Example in Python, 21-Mar-2022. [Online]. Available: https://www.datatechnotes.com/2022/03/lightgbm-regression-example-in-python.html. [Accessed: 29-Nov-2022].

[15] Arjunprasadsarkhel, "Simple random forest with hyperparameter tuning," Kaggle, 18-Aug-2021. [Online]. Available: https://www.kaggle.com/code/arjunprasadsarkhel/simple-random-forest-with-hyperparameter-tuning. [Accessed: 30-Nov-2022].

[16] "R Squared Error," Numeracy, Maths and statistics - academic skills kit. [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html. [Accessed: 2-Dec-2022].

[17] J. Frost, "Mean squared error (MSE)," Statistics By Jim, 14-Nov-2021. [Online]. Available: https://statisticsbyjim.com/regression/mean-squared-error-mse/. [Accessed: 03-Dec-2022].

[18] Stephanie, "Absolute error &; mean absolute error (MAE)," Statistics How To, 28-Dec-2020. [Online]. Available: https://www.statisticshowto.com/absolute-error/. [Accessed: 05-Dec-2022].

[19] Vbmokin, "Used cars price prediction by 15 models," Kaggle, 05-Dec-2019. [Online]. Available: https://www.kaggle.com/code/vbmokin/used-cars-price-prediction-by-15-models. [Accessed: 06-Dec-2022].

[20] J. Brownlee, "Statistical significance tests for comparing machine learning algorithms," MachineLearningMastery.com, 08-Aug-2019. [Online]. Available: https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/. [Accessed: 07-Dec-2022].

[21] Msagmj, "Data Cleaning + EDA + used cars prediction(86),Kaggle, 01-Jun-2020. [Online]. Available: https://www.kaggle.com/code/msagmj/data-cleaning-eda-used-cars-prediction-86. [Accessed: 07-Dec-2022].