A Methodology To Predict Selling Price Of Used Car Using Machine Learning Algorithm

PROG 8430: Data Analytics Mathematics and Algorithm

Jemima Paulraj (8890669)

Delaxshana Santhiravathanan (8932185)

Geedhu Kizhakepura Velayudhan (8899510)

***Abstract*—The increasing cost of new cars and the lack of funds for consumers have led to a global rise in used car sales. To ensure a fair deal when buying or selling, it's important to know the accurate market value. Unfortunately, many customers fall prey to dealers who charge unreasonable prices for secondhand vehicles. This highlights the need for a reliable price prediction system that can evaluate a car's value based on various factors. Machine learning algorithms can be employed for this purpose, accurately predicting the selling price of used cars. Here we propose to use machine learning algorithms to predict the selling price of the used vehicles. We employed the Linear Regression model, Support Vector Regression Model, and Random Forest Regression Model and have selected the best model for used car price prediction.**

Keywords—Price Prediction, Regression, Machine Learning

I. INTRODUCTION

As the digital age advances, the automotive industry has encountered a growing demand for accurate price prediction in the used car market. It is vital for buyers, sellers, and industry experts to accurately determine the value of a used vehicle. In this study, we explore the potential of machine learning algorithms in predicting used vehicle prices.

We begin by gathering a comprehensive dataset consisting of various vehicle attributes such as name, year, selling price, kilometer driven, type of fuel, type of seller, transmission type, and details about the ownership of the car. The dataset is collected from a trusted source named ‘Kaggle’. We shall see how the different regression and classification models perform in our data and we choose the best model among them to predict the selling price.

To accurately predict the prices of used cars, we used an open dataset to train our model. We used the 'vehicles data' dataset from Kaggle [1]. This dataset contains information about used vehicles listed on Car Dekho. Dataset Name: CAR DETAILS FROM CAR DEKHO.

To predict the continuous outcome variable, regression models like Multi-Linear Regression Model, Support Vector Regression Model, and Random Forest Regression Model are utilized.

In the forthcoming sessions, we will delve into the details of the data used and its characteristics. We will then discuss the data cleaning process and the reasons for removing any outliers if necessary. We will also cover the importance of predictor variables and how to select and extract relevant features. Lastly, we will explore various regression models and compare them to determine the best model.

II. LITRETURE SURVEY

Transmer, Murphy et.al.2020 [2] implemented multilinear regression. They have did backward and stepwise feature selection to select the features.

Listiani et.al.2009[3] implemented Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. They have used RMSE to show the performance of the model.

Kumbar, Gadre et.al[4] implemented six machine learning algorithms like linear regression, gradient boost, Random forest, XGBoost, Kmeans, Linear Regression Ensemble, Light GBM to predict price of used vehicle.

Pal, Arora et.al[5] implemented random forest and got 83.62% accuracy for test data and 95% accuracy for train-data.

III. METHODOLOGY AND EXPERIMENTAL RESULTS

A. Overview of data – Visual interpretation

Over 4,340 used cars from 20 different brands were sold on the website, and their pricing and characteristics are included in the dataset. We would exclude irrelevant columns, such as vehicle names, from our research because they have little to no impact on a car's pricing. Our dataset comprises 9 distinct properties of a vehicle that is being sold and they are the following:

- Name: the brand names of the cars' manufacturers and models.
- Year: The model's release year or edition.
- Km-driven: The total number of miles the previous owner(s) have driven the car(s).
- Gasoline: The kind of gasoline that is utilized by vehicles.
- Seller type: The seller may be a person or a dealer.
- Transmission: The kind of gearbox that cars employ.
- Owner: Whether the ownership is new, used, or any other type.
- Selling price: The vehicles' current selling price.
- Present price: The car's original purchasing cost.

B. Data Cleaning

- Unnecessary Column: The vehicle name is removed from the dataset since it does not contribute much to the prediction.

- Factor: The categorical data like fuel_type, seller_type, transmission, owner is converted into factors.
- Unique: The 763 redundant data are removed from the dataset.
  Original Dataset: 4340 After Removing Duplicate: 3577

C. Visual Interpretation and Remove Outliers if Required

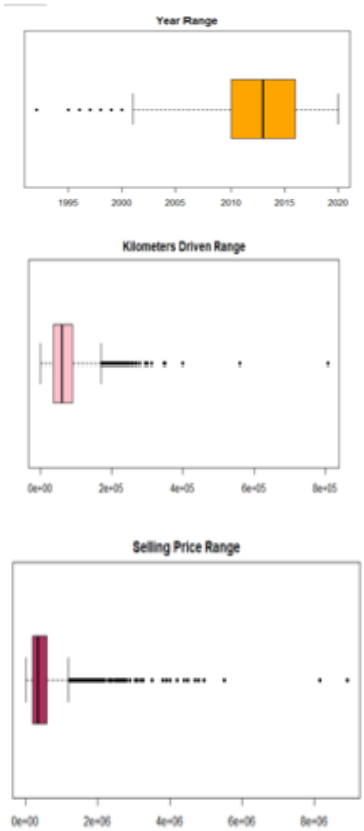
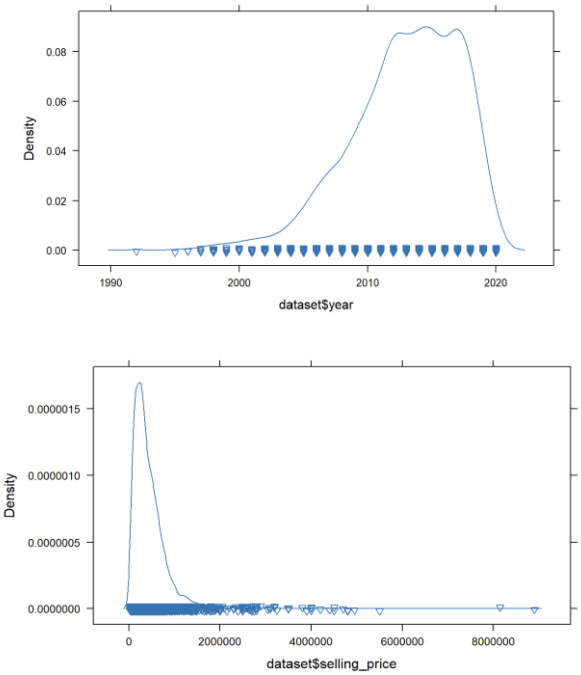


Fig. 1. Box Plots for Numerical Data



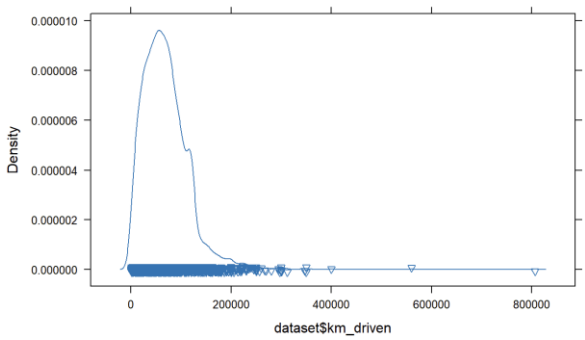


Fig. 2. Density Plots for Numerical Data

From the data shown in Figures 1 and 2, it seems that some values in the numerical data are outliers. However, it is crucial to have valid justifications before removing any outliers as it can impact the accuracy of the model. The selling price can vary depending on market demand while the kilometer driven can differ based on the user's driving preferences and years of usage. Additionally, the dataset comprises cars that were purchased before the year 2000, and these models may no longer be in use due to advancements in technology and comfort. Therefore, we have decided to exclude data with year values less than 2000.

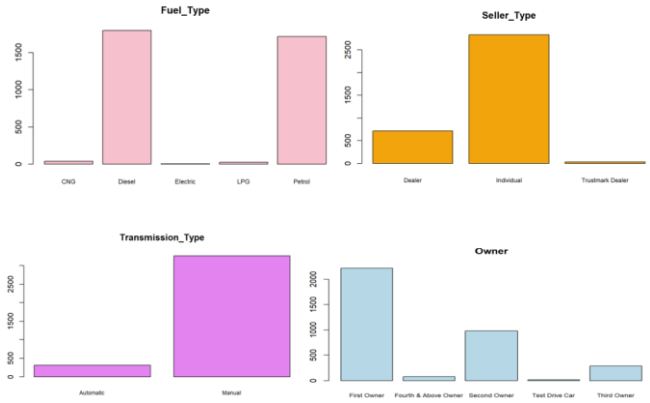
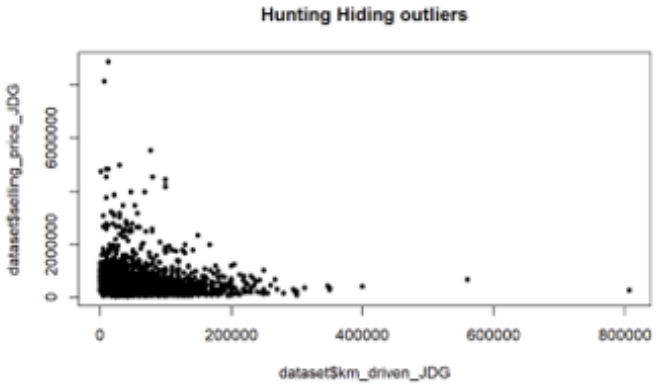


Fig. 3. Distribution of Categorical Data

In Fig.3, the frequency of values belonging to various categorical data, such as Fuel type, Seller type, Transmission type, and Owner category, is depicted.



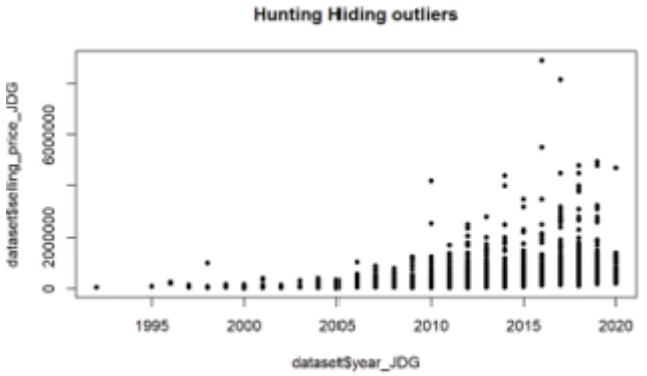


Fig. 4. Relationship between the target 'selling price' with two attributes Kilometer driven and Year.

Based on the data in Figure 4, it is evident that the selling price of a car declines with an increase in its age and a greater number of kilometers driven. The majority of the data is concentrated within the range of 200000 kilometers, indicating that buyers tend to avoid purchasing cars with over 200000 kilometers on the odometer. Additionally, buyers generally prefer cars purchased after 2010.

Feature Selection:

Feature selection is a technique used in machine learning to improve the accuracy of the data by focusing on the most important factors and eliminating the redundant and irrelevant ones, it also improves the algorithms' ability to predict outcomes.

The following are the three advantages of feature selection:

- avoids over-fitting - Less duplicate data decreases the likelihood that judgments will be based on noise.
- Improves Accuracy - Better modeling accuracy outcomes from fewer misleading data.
- Reduces training time - Algorithms run quicker with less data.

A. Filter Feature Selection:

Before creating a predictive model, relevant features are chosen from a dataset using the filter feature selection process in machine learning. This strategy, features are removed based on predetermined standards, such as ranking systems or statistical metrics, without using the target variable or a machine learning algorithm itself. By focusing on the most insightful and non-redundant features, filter feature selection aims to enhance model performance and lower the risk of overfitting.

- *Identifying Low Variance:*

	year_JDG	selling_price_JDG	km_driven_JDG	fuel_JDG	seller_type_JDG
nbr.val	3552.000000000	3552.000000	3552.0000000	NA	NA
nbr.null	0.000000000	0.000000	0.0000000	NA	NA
nbr.na	0.000000000	0.000000	0.0000000	NA	NA
min	2000.000000000	20000.000000	1.0000000	NA	NA
max	2020.000000000	8900000.000000	806599.0000000	NA	NA
range	20.000000000	8880000.000000	806598.0000000	NA	NA
sum	7150424.000000000	1692160163.000000	245907182.0000000	NA	NA
median	2014.000000000	350000.000000	60000.0000000	NA	NA
mean	2013.069819820	476396.442286	69230.6255631	NA	NA
SE.mean	0.068238286	8557.009782	799.9354858	NA	NA
CI.mean.0.95	0.133790186	16777.149476	1568.3793240	NA	NA
var	16.539759134	260086023087.571442	2272913367.8766541	NA	NA
std.dev	4.066910269	509986.296961	47675.0812047	NA	NA
coef.var	0.002020253	1.070508	0.6886415	NA	NA

Year_JDG has the lowest coef.var. Although the coefficient of variation is low, since it belongs to year category. We have decided to include the year as a valid attribute in predicting the selling price. Therefore, we will not be removing it from our analysis.

- *Identify High Correlation:*

	year_JDG	selling_price_JDG	km_driven_JDG
year_JDG	1.0000000	0.4277128	-0.4367229
selling_price_JDG	0.4277128	1.0000000	-0.1872496
km_driven_JDG	-0.4367229	-0.1872496	1.0000000

According to the finding, there is no high correlation found among the features.

B. Wrapper Feature Selection:

The model is trained to utilize a subset of features using wrapper approaches. We have chosen to either include or remove specific qualities from the subgroup since we have all the information necessary for predicting the price of used cars.

Here we have used the standard wrapper methods, forward feature selection, backward feature elimination, and recursive feature elimination.

Forward Selection:

The model has no features at first, and iterations move forward. Add the feature that enhances the model the most each time until adding a new variable has no effect on how well it performs.

Forward Propagation

```
Start:  AIC=93362.76
selling_price_JDG ~ 1

Step:  AIC=92402.51
selling_price_JDG ~ transmission_JDG

Step:  AIC=91697.84
selling_price_JDG ~ transmission_JDG + year_JDG

Step:  AIC=91402.95
selling_price_JDG ~ transmission_JDG + year_JDG + fuel_JDG

Step:  AIC=91374.12
selling_price_JDG ~ transmission_JDG + year_JDG + fuel_JDG +
    km_driven_JDG

Step:  AIC=91354.7
selling_price_JDG ~ transmission_JDG + year_JDG + fuel_JDG +
    km_driven_JDG + seller_type_JDG

Step:  AIC=91350.19
selling_price_JDG ~ transmission_JDG + year_JDG + fuel_JDG +
    km_driven_JDG + seller_type_JDG + owner_JDG
```

Backward Elimination:

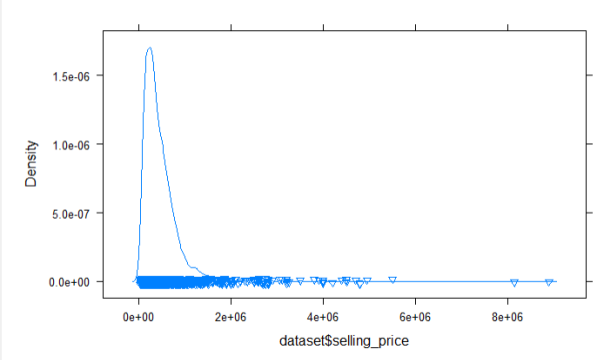
Backward elimination, which starts with all of the features and eliminates the least important feature after each iteration, enhances the performance of the model. Up till no progress is seen, this practice is repeated.

Backward Propagation

```
Start:  AIC=91350.19
selling_price_JDG ~ year_JDG + km_driven_JDG + fuel_JDG + seller_type_JDG +
    transmission_JDG + owner_JDG
```

These features gives the lowest AIC Value, which indicates better fitting model.

Data is not normally distributed. We have processed scaling to make data uniform.



Scaling

Data scaling is the function of modifying a dataset's feature values until they fall within a specific range. This can assist the algorithm work better by minimizing any one attribute from having an excessive impact on distance calculation. We have done the scaling using min-max scaling and Tukeys ladder, and compared between them

Min-max scaling

It is used to scale variables (features) to a predetermined range, typically between 0 and 1. All attributes are intended to be scaled similarly so that certain algorithms will work better and the data will be less sensitive to the scale of specific features.

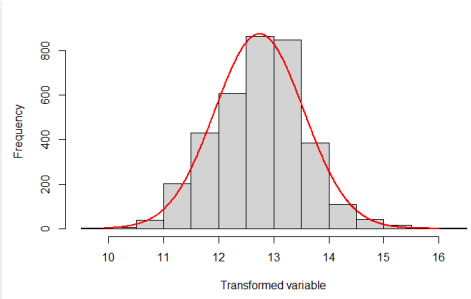
The formula for Min-Max scaling is:

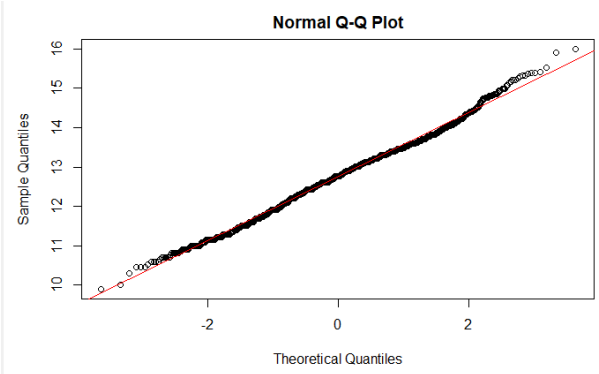
$$XScaled = (X-Xmin) / (Xmax-Xmin)$$

X is the original value of the feature
Xmin is the minimum value of the feature
Xmax is the maximum value of the feature

Tukeys ladder

A skewed distribution can be transformed into a normal or nearly-normal distribution using the Tukey ladder of powers (also known as the Bulging Rule). Heteroscedasticity, the variability of errors, can also be decreased.





Data Splitting:

The dataset is splitted into training and testing set with 80:20 ratio. 80% data for training the data and 20% data is for testing the model.

```
Dataset size:  3552
Number of train set:  2841
Number of test set:  711
```

K Fold Cross Validation:

Cross Validation is one of the most commonly used method for evaluating the accuracy of a model. K in K-fold validation stands for the number of folds to which the data is to be partitioned. Here the data is randomly partitioned into K mutually exclusive subsets, each approximately equal size. At the ith iteration Di is used as test set whereas others are considered as the training set. Also, each sample is used same number of times for training and once for testing.

10-times-K-fold Cross-Validation:

It goes one step further by performing the K-fold cross-validation procedure more than once (in this case, ten times), using various randomly generated data partitions into the K folds. The average of the outcomes from each of the 10 runs serves as the final performance metric.

D. Equations

Following machine learning models have been implemented.

1. Linear Regression model
2. Support vector machine Regression
3. Random Forest Regression

A. Linear Regression:

Linear regression is a linear approach to model the relationship between a dependent and independent variable. Multiple Linear Regression is an extension of simple linear regression where the response variable, y, modeled as a linear function of two or more predictor variables. The regression equation is of linear form given by [1]

$$y = \beta0 + \beta1.x1 + \beta2.x2 + \ldots.. + \beta n.xn(1)$$

Where β0 is the intercept and β1, β2..... βn are the coefficients. The variables x1, x2....xp are the predictor variables.

There are few steps involved in generating the regression beginning with feature selection or extraction, normalizing features, select loss function and hypothesis, set hypothesis parameters, and minimizing the loss function, and finally testing the function of the data.

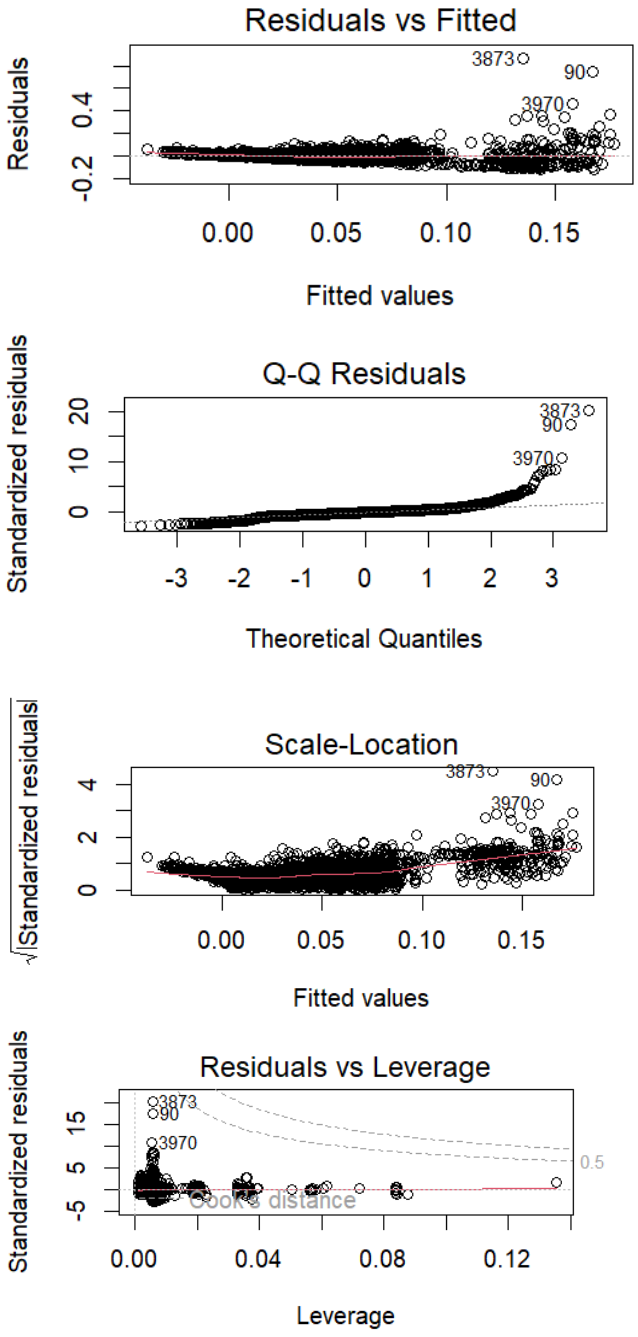


Fig. 5. Multi-Linear Regression Model Results

From Fig. 5. Multi-Linear Regression Model Results, Following are the inferences made:

- Residual vs Fitted Graph: Based on actual vs. predicted scatter plots, We can find how effectively the model is working. The points should be closer to a diagonal line for the ideal model.

- Q-Q Residuals Plot: This plot is used to find the normal distribution of errors. The results for properly distributed data should generally lie in a straight line. The points establish a curve that deviates from a straight line if the data is non-normal, which is a problematic condition. For the ideal case, we require a 45-degree slope.

- Scale-Location: The y-axis in this scatter plot indicates residuals, and the x-axis indicates fitted or predicted values. The model's non-linear behavior uneven error variances and outliers are all detected using this diagram.

- Residual vs Leverage: For the Ideal model, this plot will not show any pattern. But if any pattern is visible such as a curve, or linear shape then it indicates that there is non-linearity in the data set. In this case, if there are any influential data points, this plot may be useful. Not all outliers have an impact on linear regression; in other instances, even if their value is within an acceptable range, it can still have a significant impact. Therefore, removing or eliminating these numbers can significantly change the outcomes. When data points go outside of the Cook's distance, which is shown by having greater Cook's distance scores, the regression results are influenced. The Cook's distance is the dashed line in the graph. If we take these instances out of the regression, the findings will vary.

B. Support Vector Regression:

When Support Vector Machine is used for Regression data, then the model is called Support Vector Regression. The Hyperplane in an N dimensional space (N-Number of features) and the best fitting margin can be used to differentiate the data. In SVR we try to fit the error within a certain threshold. The equation of a hyperplane is given by

$$y = \beta 0 + \beta 1x \quad (2)$$

The equations for the decision boundary are

$$\beta 0 + \beta 1x = +a \quad (3)$$
$$\beta 0 + \beta 1x = -a \quad (4)$$

Therefore, any hyperplane that satisfies Support Vector Regression should satisfy

$$-a < y - \beta 0 + \beta 1x > +a \quad (5)$$

```
Call:
svm(formula = selling_price_JDG ~ ., data = train_set, type = "eps-regression", na.action = na.omit)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.07142857
    epsilon:  0.1


Number of Support Vectors:  1940
```

C. *Random Forest Regression:*

In general, random decision forests or random forests are used for classification, and regression, where they construct a large number of decision trees during training time and output the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. The tendency of decision trees to overfit their training set is corrected by a random decision forest. Random forests are often more accurate than decision trees, although they fall short of gradient-enhanced trees. However, they are impacted by data properties.
The intercept attribute was used to train the random forest regressor in this project. The model is assessed using score, MSE, RMSE, and MAE errors. From there, the outcomes that follow are achieved.

E. *Results*

K Fold Cross Validation results for linear regression model

```
Linear Regression

3552 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3197, 3197, 3197, 3196, 3194, 3197, ...
Resampling results:

  RMSE        Rsquared   MAE
  0.04236024  0.4565369  0.02330913
```

10 times K Fold Cross Validation results for linear regression model

```
Linear Regression

3552 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 3197, 3197, 3197, 3196, 3194, 3197, ...
Resampling results:

  RMSE        Rsquared   MAE
  0.04241863  0.4546043  0.02329786
```

The regression model can be evaluated on following parameters:

1. *Mean Square Error (MSE):*

The only value that indicates how well a regression line performs is the MSE. The MSE number should be as low as possible because a lower value suggests less mistakes in magnitude.

$$MSE = (1/n) * (Actual - prediction) \wedge 2$$

Here $\sum$ - symbol that means sum,
n – size of the sample
actual: Actual data value
prediction: the predicted data value

1. *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.

$$RMSE = \sqrt{[\ \Sigma(A_i - P_i)^2 / n\ ]}$$

Pi is the predicted value for the i^{th} element in the dataset
Ai is the observed value for the i^{th} element in the dataset

2. *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.

$$\boldsymbol{MAE} = \mathbf{1}\ /n \sum |\boldsymbol{yi} - \boldsymbol{y}|$$

n: Size of the sample

Yi: Predicted value

Y: Actual value

The comparison of the models shown in table Fig 5: Comparison of Models

Model name	RMSE value of Train set	RMSE value of Test set	Model Training time in secs
MLR	0.04	0.04	0.02
SVM	0.04	0.04	1.19
RFR	0.03	0.04	3.744

F. *Conclusion*

After comparing the RMSE values of all three models for both their training and testing data, we have determined that the multi-linear regression model is the superior choice due to its comparatively faster training time.

G. *Future Opportunities*

With this dataset, we will test different regression models to see which one performs the best at forecasting the selling price of used cars. We'll also aim to categorize the selling price into a "High Price" and a "Low Price category," or even more, to make it easier for customers to filter out the vast dataset. Applications based on the web and mobile devices will

incorporate our intelligent model for usage by the general public. Additionally, following the data gathering period, there were severe shortages of semiconductors due to the pandemic, which drove up the cost of cars and had a significant impact on the used car market. Therefore, it is necessary to periodically gather and analyze data; ideally, we would use a real-time processing program.

REFERENCES

[1] https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=car+data.csv

[2] Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M.(2020) Multiple Linear Regression (2nd Edition); Cathie,Marsh Institute Working Paper 2020-01.

[3] MarianaListiani.” Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”.In: March 2009.

[4] Kshitij Kumbar, Pranav Gadre ,Varun Nayak Predicting Used Car Prices,Stanford University, CS 229

[5] Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy.” How much is my car worth? A methodology for predicting used cars prices using Random Forest”.In: Future of Information and Communications Conference (FICC), 2018