Predicting Car's Resale Value using Machine Learning Models

Aishwary Sharma aishwary20490@iiitd.ac.in

Sarthak Daksh sarthak20403@iiitd.ac.in

Atyant Sony atyant20039@iiitd.ac.in

Singh Ayush Kumar Satish ayush20133@iiitd.ac.in

Abstract

A car depreciates in value from the moment you buy it, and depreciation progresses over time. The make and model of the car, total kilometers driven, overall condition of the vehicle, and various other factors further affect a car's resale value. This project aligns with our goal of creating a transparent sales process by providing the best resale value estimate for a car that one should expect to shell out, thus preventing any kind of dissatisfaction from either the buyer's side or the seller's side and deploying Machine Learning Models to solve real-world problems.

1. Introduction

The private automobile has changed modern civilization for more than a century by enabling independence and freedom of movement. Due to the growing distances between places like home, work, educational institutions, shopping malls, and recreational facilities, mobility is becoming more and more crucial. Without proper individual mobility, it would frequently be impossible for people to participate in social and economic life, especially the elderly, people with impairments, and those living in rural places.

Before the Pandemic, a large car inventory allowed consumers to find the precise model and color they desired at a significant discount off the advertised price. Automakers have been obliged to reduce output due to the global shortage of semiconductor chips, which are currently predominantly supplied from Southeast Asia, where COVID-19 cases are among the highest in the world. As manufacturers struggle with production, new vehicle prices have reached record highs across the nation, and as long as there are strong demand and inventory constraints caused by the chip shortage, prices for new vehicles are not expected to drop anytime soon. This has resulted in increased demand for used cars, and getting a good deal for a car is still one of the most challenging tasks for both buyers and sellers.

Our project aims at creating a transparent sales process by providing the best resale value estimate for a used car that one should expect to shell out based on the features and characteristics of the car by deploying Machine Learning Models.

2. Literature Survey

The study provides a model based on Big Data Analysis using the optimized BP neural network algorithm used to select the optimal number of hidden neurons in the BP neural network to establish a second-hand car price evaluation model to get the price that best matches the car.[1]

A neural network model was trained using data from vehicle users and sellers to best predict cars suitable to the user. Besides the neural network, the model uses natural language processing (NLP) to produce more personalized recommendations.[2]

The paper provides a working model for used car price prediction using a Supervised Machine learning-based Artificial Neural Network model and a Random Forest Machine Learning model. The model is trained on a prior dataset, and the model's accuracy is quite reliable.[3]

3. Dataset

3.1. Details

The Dataset has 9379 rows and 32 columns. It contains information about the car, such as year, mileage, Model, Resale Price, Engine, Fuel Type, Interior, and Exterior Color, Comfort Rating, Interior Design Rating, etc. The Resale price is in dollars ($).
Furthermore, we cannot apply under-sampling or over-sampling because the target value is continuous. PCA is also not applicable to the dataset as majority of features are weakly correlated.

3.2. Pre-processing of the Data

We dropped a few columns as a part of our feature selection because they were irrelevant to our Resale Price prediction. The dropped columns were Model, VIN, State, Zipcode, DealType, StreetName, Stocks, SellerName, Seller Rating, Seller reviews, and Value for Money Rating. Null values were not present in the data, but some values were not specified, such as: a) "Not Priced" value in the Price Column. b) "-" value in Engine, Transmission, DriveTrain, FuelType, InteriorColor, ExteriorColor. So, we dropped these rows because of their small number. The datatype of Price was changed from object to INT. Some categorical values were

mapped under the general class to better predict our model. That was:

1. Extracted engine cylinder capacity from Engine Feature and added an extra column “isTurboEngine”, denoting whether it is a Turbo Engine or not
2. Transmission: All values were mapped to Manual or Automatic. Further, Manual and Automatic were mapped to 1 and 2, respectively.
3. Drivetrain Mapping:
   Front-wheel Drive = 1
   Rear-wheel Drive = 2
   Four-wheel Drive = 3
4. Fuel Type Mapping:
   Flex Fuel = 1
   Gasoline = 2
   Hybrid = 3
   Electric = 4
5. Interior Color Mapping:
   Black Variants = Black
   White Variants = White
   Gold & Red Variants = Golden
   Others = Cream
6. Exterior Color Mapping:
   Black Variants = Black
   White Variants = White
   Red Variants = Red
   Others = Silver
7. Changed all the Company Certified values to “Certified.” Further Certified were mapped to 2 and the rest with 1.

For Interior Color, Exterior Color, and Seller Type, One Hot Encoding is used,so that the model doesn’t get skewed towards a particular color or one type of seller of the car.

3.3. Data Visualization

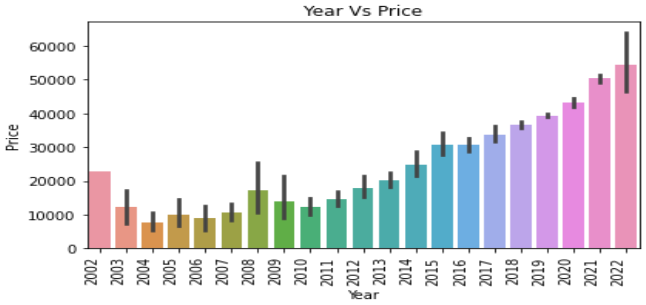

Figure 1 - Year Vs Price

It can be observed in Figure 1 that the Resale Price of a Car decreases with an increase in its age.

It can be observed in Figures 2 & 3 from the above graphs that the resale price of cars increases with better Drivetrains and Complex Fueling systems.

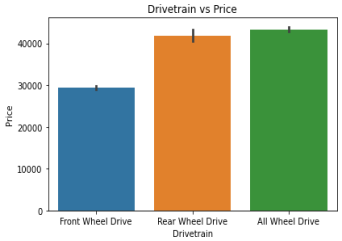

Figure 2

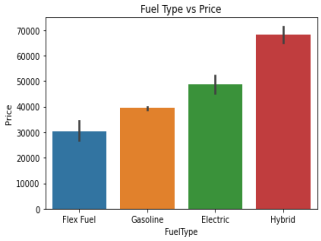

Figure 3

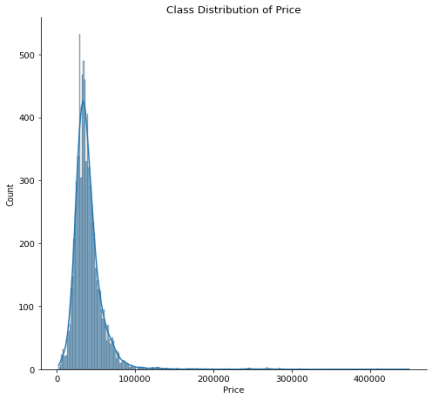

Figure 4 - Class Distribution of Price

It can be observed in Figure 4 that the majority of cars are resold at prices less than $100K

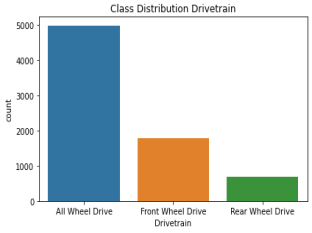

Figure 5
Class Distribution of Drivetrain

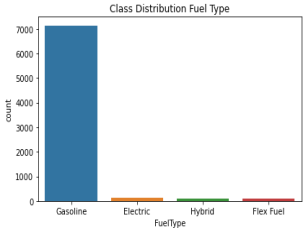

Figure 6
Class Distribution of Fuel Type

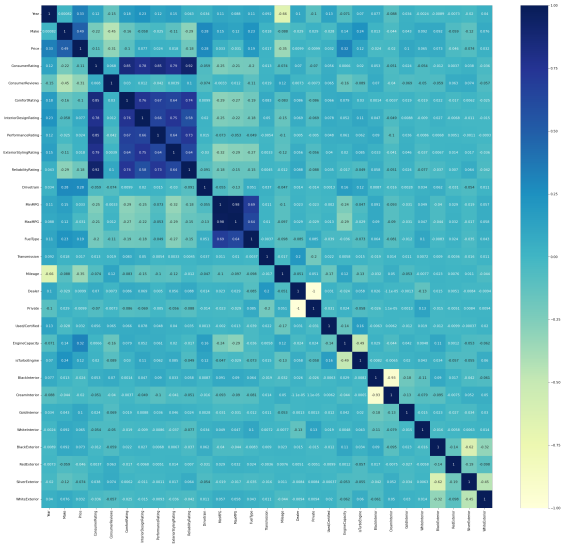

Figure 7: Heatmap

As observed in Figure 7, Consumer Rating, Comfort Rating, Interior Design Rating, Performance Rating, Exterior Styling Rating, Reliability Rating are highly correlated with each other. MinMPG, MaxMPG, and FuelType also show a high correlation. Rest of the features are moderately or weakly correlated.

4. Methodology

To predict the prices of a car we used Regression Models. For this, The dataset above was split into Train:Validation:Test of ratio 60:20:20. GridSearchCV was used for Hyperparameter tuning using 4 fold cross validation. We used different supervised machine learning models for the task; a total of 5 models. The R2 score, Absolute error, Mean Squared Error and Root Mean Squared Error where used as performance metrics..

4.1 Linear Regression

Applied Linear Regression on the DataSet with 60-20-20 train-validation-test split. This model is used as the base model for the Performance Metrics.

4.2 Lasso Regression

Applied Lasso Regression on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter (Alpha) tuning. The value of alpha was changed from 0 to 50 with a step of 0.1 to find the best value for the parameter corresponding to best RMSE Score. The plot for R2 score and RMSE with respect to each value of alpha can be seen in fig.8 and fig. 9 for both training and validation.

Figure 8

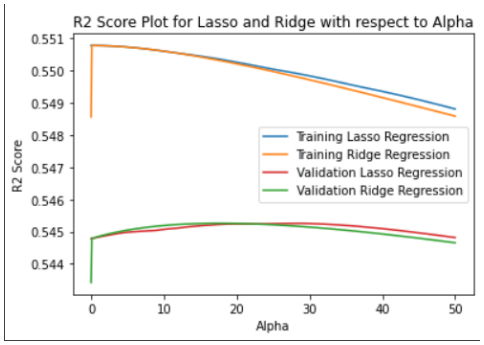

Figure 9

4.3 Ridge Regression

Applied Ridge Regression on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter (Alpha) tuning. The value of alpha was changed from 0 to 50 with a step of 0.1 to find the best value for the parameter corresponding to best RMSE Score. The plot for R2 score and RMSE with respect to each value of alpha can be seen in fig.8 and fig. 9 for both training and validation.

4.4 Decision Tree Regressor

Applied Decision Tree Regressor on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter (criterion and max_depth) tuning. Figure 10 shows the plot for RMSE Score and max_depth taken for a tree for both Training and Validation where as figure 11 Shows graph of R2 Score vs Max-depth.

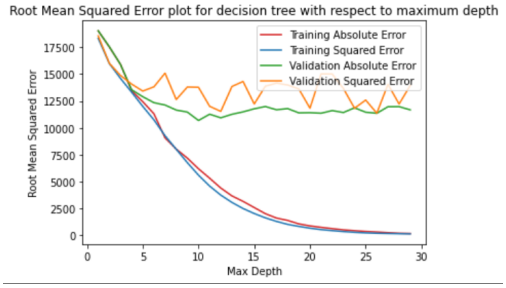

Figure 10

Figure 11

4.5 Decision Tree Regressor with ADA boosting

Applied Boosting technique on the Decision Tree Regressor to boost the performance of the Machine Learning Model. The best parameters obtained in the previous section were used for the Regressor and also used Grid Search for Hyperparameter(n_estimator and loss) tuning. Figure 12 and 13 shows the plot of RMSE score and R2 score with respect to the number of estimators for both training and validation test.

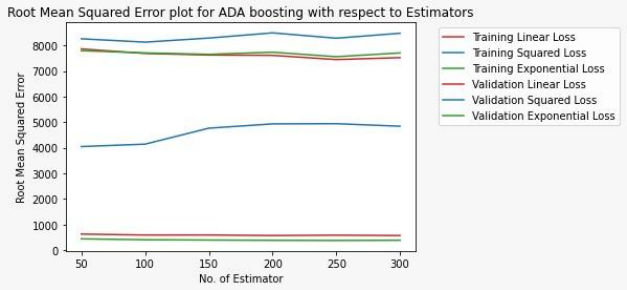

Figure 12

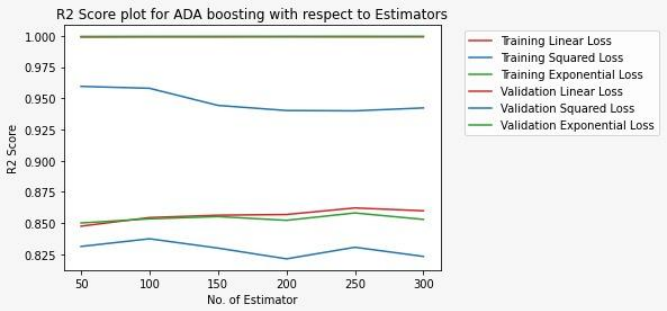

Figure 13

4.6 K-Nearest Neighbor

Applied K-NN on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter (Neighbors and weights) tuning. Figure 14 and 15 shows the graph of RMSE vs N-Neighbors and R2 score vs N-Neighbors for both Training and Validation respectively.

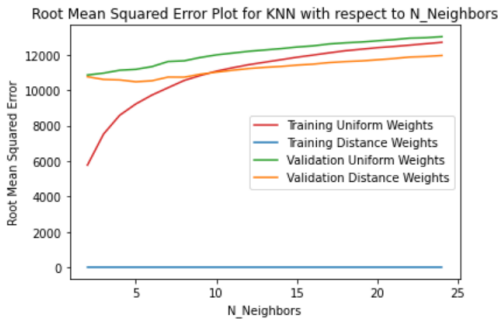

Figure 14

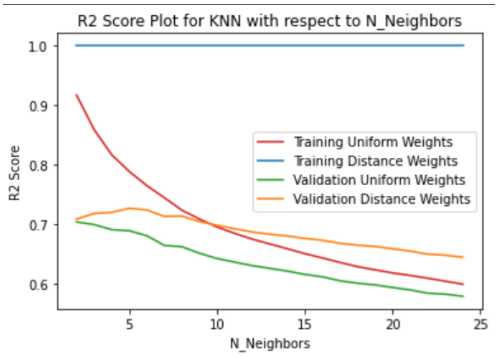

Figure 15

4.7 SVR

Applied Support Vector Regressor on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter tuning. The Two categories were epsilon and Linear

4.8 Random Forest

Applied Support Vector Regressor on the DataSet with 60-20-20 train-validation-test split. Used Grid Search for Hyperparameter tuning. The plot for RMSE score and R2 with respect to each value of alpha can be seen in fig.16 and fig. 17 for both training and validation.

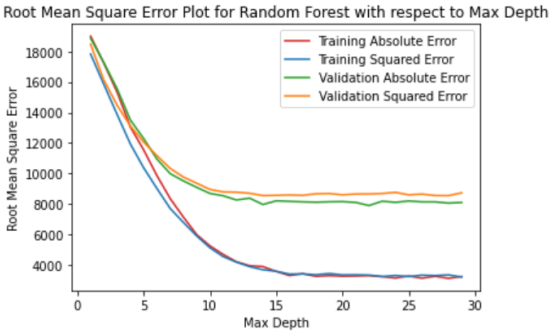

Figure 16

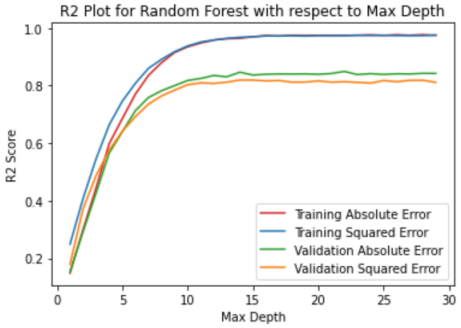

Figure 17

5. Results

We have tested different models and optimized it using hyperparameter tuning. We will now analyze the results obtained on different models and we will also analyze the R2-Score, Mean Absolute Error, Mean Squared Error, Root Mean, Square Error for both Training Set and Testing Set corresponding to every Model.

Testing Set

Model	R2 Score	Absolute Error	Mean Squared Error	Root mean Square error
Linear	0.5047	7290.1	207909	14419.

Regression	594243059601	92246735668	323.58681455	061120156699
Lasso Regression	0.50547699231994	7253.69477310440	207608077.914029	14408.6112416856
Ridge Regression	0.50372372495319	7272.58715567909	208344125.504187	14434.1305766640
Decision Tree	0.68933521767122	4725.25855130784	130421673.679912	11420.2308943345
ADA boosting	0.86946584829189	3336.00207804944	61609962.1316838	7849.20136903646
KNN	0.84081035571020	4423.42105624578	78582295.8893053	8864.66558248563
SVR (linear)	0.45425006699348	8137.46125405514	242839659.885238	15583.3135078916
SVR (epsilon)	0.43779440403628	8137.46125405514	242839659.885238	15583.3135078916
Random Forest	0.87243212012750	3515.09423205902	52371376.2413207	7236.80704740155

Model	**Best Parameter**
Linear Regression	No parameters
Lasso Regression	Alpha : 19.138276553106213
Ridge Regression	Alpha : 16.032064128256515
Decision Tree	criterion : absolute_error, max_depth : 10
ADA boosting on Decision Tree	Loss : linear, n_estimators : 250
KNN	N_neighbors: 5, weights: distance
Random Forest	Criterion: absolute_error, max_depth: 22
SVR (epsilon)	C: 10.0, epsilon: 0.01, kernel: linear, gamma: scale
SVR(Linear)	Loss: squared_epsilon_insensitive, C: 0.1

6. Conclusions

The methodology we used involved cleaning and pre-processing of the dataset. From the pre-processing of the data highly correlated features were removed resulting in feature reduction. Feature selection was done to remove the unnecessary data.The dataset was further split to training data and test data set on which the models were trained and tested. A total of 8 models were implemented namely Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, ADA boosting on Decision Tree Regressor, SVR, KNN, and Random Forest. K fold validation was also implemented along the models to get the best model. The Random Forest model had the lower RMSE among all the models with the highest R2 score among them during the testing phase. Therefore, The best model to us currently among these 8 models is Random Forest to predict the resale values of the car.

6.2 Contribution

Data preprocessing and Data Visualization - Atyant Sony and Aishwary Sharma

Creating Model ,Model Training and Testing - Sarthak Daksh, Singh Ayush Kumar Satish

The project slides and report were contributed equally by everyone.

7. References

[1] N. Sun, H. Bai, Y. Geng, and H. Shi, “Price evaluation model in second-hand car systems based on BP neural network theory,” in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

[2] P. Boteju and L. Munasinghe, “Vehicle Recommendation System using Hybrid Recommender Algorithm and Natural Language Processing Approach,” in 2020 2nd International Conference on Advancements in Computing (ICAC).

[3] J. Varshitha, K. Jahnavi, and C. Lakshmi, “Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning,” in 2022 International Conference on Computer Communication and Informatics (ICCCI).