LITERATURE SURVEY

PROJECT TITLE: CAR RESALE VALUE PREDICTION

Team Leader: Nivetha M

Team Member1: Ramguhan R T

Team Member2: Rithin A

Team Member3: Ruthika R

ABSTRACT:

Cars of a particular make, model, year, and set of features start out with a price set by the manufacturer. As they age and are resold as used, they are subject to supply-and-demand pricing for their particular set of features, in addition to their unique history. The more this sets them apart from comparable cars, the harder they become to evaluate with traditional methods. Using Machine Learning algorithms to better utilize data on all the less common features of a car can more accurately assess the value of a vehicle. This study compares the performance of Linear Regression, Ridge Regression, Lasso Regression, and Random Forest used cars. An important qualification of a price prediction tool is that depreciation can be represented to better utilize past data for current price prediction. The study has been conducted with a large public dataset of used cars. The results show that Random Forest Regression demonstrates the highest price prediction performance across all metrics used.

Keywords: Machine Learning, Price Prediction, Used Cars, Regression Analysis, Depreciation.

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.

PROBLEM STATEMENT:

The used car market is a large and strategically important market for car manufacturers. Imagine a situation where you have an old car and want to sell it. You may of course approach an agent for this and find the market price, but later may have to pay pocket money for his service in selling your car. But what if you can know your car selling price without the intervention of an agent. Or if you are an agent, definitely this will make your work easier. Yes, this system has already learned about previous selling prices over years of various cars.So, to be clear, we will provide you will the approximate selling price for your car based on the fuel type, years of service, showroom price, the number of previous owners, kilometers driven, if dealer/individual, and finally if the transmission type is manual/automatic.

OBJECTIVES:

The main objective of this this project are:

1.To develop an efficient and effective model which predicts the price of a used car according to user's inputs.

2.To achieve good accuracy.

3.To develop a User Interface (UI) which is user-friendly and takes input from the user and predicts the price.

SCOPE:

The purpose of this project is to evaluate several different machine learning models for used car price prediction and draw conclusions about how they behave. This will deepen the knowledge of machine learning applied to car valuations and other similar price prediction problems.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.

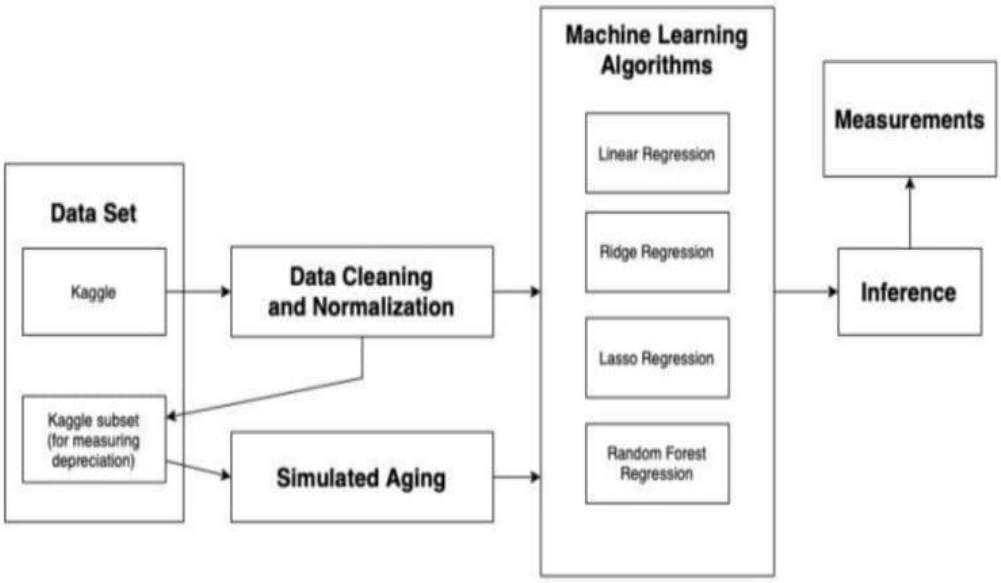
•Useful for predicting the accurate price for Indian Car type.

•System can be deployed on the webservers as a feature.

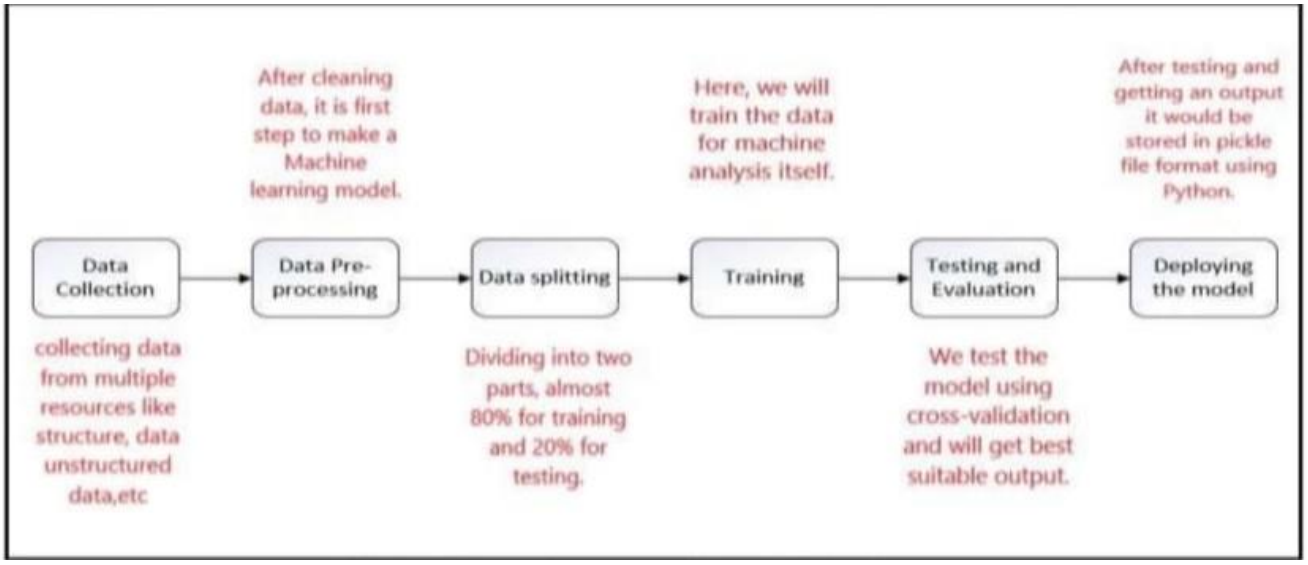
•System can be integrated into chatbots

•Standalone system by merging the data from different sources can be implemented

IMPLEMENTATION/DESIGN:



PROPOSED MODEL:



USED ALGORITHMS:

Following are some regression algorithms that can be used for predicting the selling price.

- Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

- Decision Tree Regressor
  Decision trees tend to be the method of choice for predictive modeling because they are relatively easy to understand and are also very effective. The basic goal of a decision tree is to split a population of data into smaller segments. There are two stages to prediction. The first stage is training the model—this is where the tree is built, tested, and optimized by using an existing collection of data. In the second stage, you actually use the model to predict an unknown outcome. We'll explain this more in-depth later in this post. It is important to note that there are different kinds of decision trees, depending on what you are trying to predict.

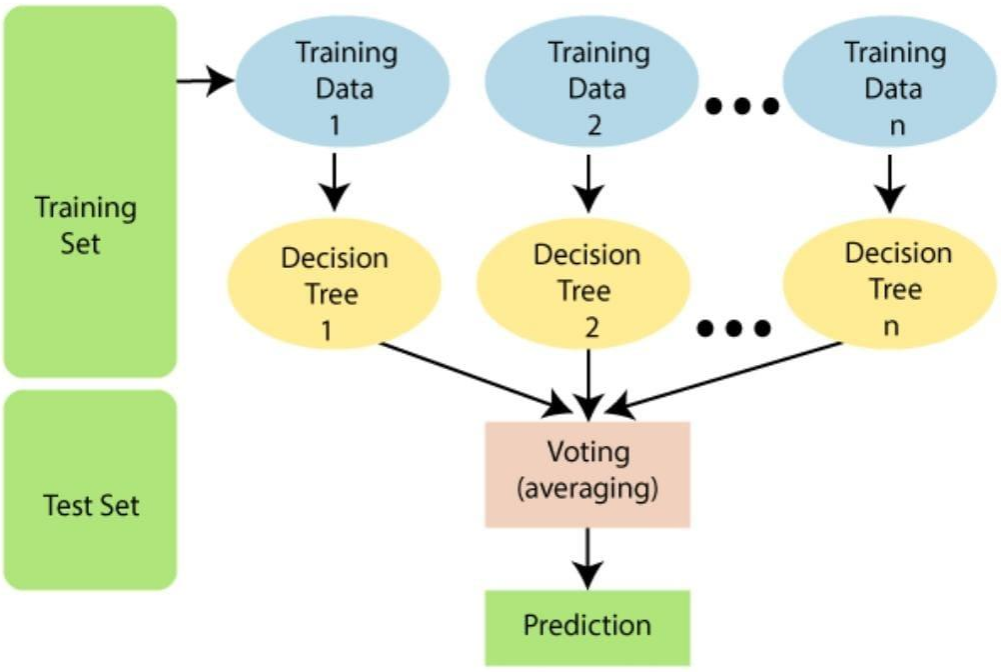
- Support Vector Regressor
  Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

- KNN Regressor
  In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive

integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

- Random Forest Regressor

  Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
  As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



TEST CASES:

• Missing values

The trained ML model requires 4 feature inputs for predicting the output. Failing which, the model throws invalid Input error. All the fields in the html form have

been marked required using CSS and thus user must input all fields.

Output: User must input all the fields, failing which, form shows warning message "this field needs to be filled".Thus, there can be no errors in model prediction.

• Invalid Input

The trained ML model requires only numerical input for all 4 features. Thus, if user uses symbols such as comma while input, model may throw error. To overcome the same, preprocessing script is deployed in backend which removes all unwanted characters like comma, whitespaces etc. so that model gets required input.

Output: Due to python preprocessing script, model will get the desired input and thus will give accurate prediction.

• Unseen year of purchase

The model is trained with data from cars purchased since 2011 to 2020. If the user inputs details of car purchased after that i.e., 2021, model may get confused since that data is quite new and unseen to model.

Output: Model has been trained with boosting algorithm and thus it gives quite accurate results with around RMSE 65,000 INR.

DATASET DESCRIPTION:

The data types of each attribute were corrected/converted by performing on each attribute individually. Details and description of dataset is described in the table given below

SNO	Column Name	Datatype	Unique values	Mode/Mean of column
1	Car Name	Object	4009	Pajero 2016
2	Brand	Object	82	Nissan
3	Model	Object	725	Pajero
4	Production year	Int	48	2012
5	Engine size	Float	150	2.00L
6	Gear type	Object	2	Automatic
7	Mileage	Int	20503	146475.39
8	Fuel type	Object	4	Diesel
9	Color	Object	14	White
10	Price	Int	684	61301.55

PROS:

- Good at learning complex and non-linear relationships
- Highly explainable and easy to interpret
- Robust to outliers
- No feature scaling is required
- Can get cars that are no longer available in market

CONS:

- Consumes more time
- Requires high computational power
- Difficult to maintain the engine to be in good condition

CONCLUSION:

Using data mining and machine learning approaches, this project proposed a scalable framework for Dubai based used cars price prediction. Buyanycar.com website was scraped using the Parse Hub scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy followed by Bagging Regressor with 88%. Each experiment was performed in real-time within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab.

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)

3) Gokce, E. (2020, January 10). “Predicting used car prices with machine learning techniques. “

4) Predicting Used Car Prices with Machine Learning Techniques | by Enes Gokce | Towards Data Science

BIELSKI, V., & RAMARATHNAM, S. (2020, July 16). UAE’s used car sales set to surge past

5)1 million mark by 2025. Retrieved from gulfbusiness: https://gulfbusiness.com/uaes-used-car-sales-set-surge-past-1-million-mark-

2025/#:~:text=For%20every%20new%20car%20sold,crossing%20the%201%20million%20mark.

6)Bridge, S. (2020, January 10). Why the value of used cars is rising for the first time in the UAE.Retrieved from arabianbusiness: https://www.arabianbusiness.com/retail/435520-why-the-value-of-used-cars-is-rising-for-the-first-time-in-the-uae

7)Ceriottia, M. (2019). Unsupervised machine learning in atomistic simulations, between predictions and understanding. 150-155.

8)Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019, February). Car Price Prediction using Machine. TEM Journal, 8(1), 113-118. doi:10.18421/TEM81-16

9)Gongqi, S., Yansong, W., & Qiang, Z. (2011). A New Model for Residual Value Prediction of the Used Car Based on BP Neural. Third International Conference on Measuring Technology and Mechatronics Automation (pp. 682-685). Shanghai: IEEE. doi:10.1109/ICMTMA.2011.455

10)Great Learning Team. (2020, August 17). Introduction to Multivariate Regression Analysis. Retrieved from mygreatlearning: https://www.mygreatlearning.com/blog/introduction-to-

multivariateregression/#:~:text=Multivariate%20Regression%20is%20a%20supervised,try%20to%20predict%20the%20output.

11)Jian Da Wu, C.-c. H.-C. (2017). "An expert system of price forecasting for used cars using adaptive. ELSEVEIR, 16, 417-957.K.Samruddhi, & Kumar, D. R. (2020, September). Used Car Price Prediction using K-Nearest

12)Neighbor Based Model. International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), 4(3), 686-689.

13)Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth? Journal of Statistics Education. doi:10.1080/10691898.2008.11889579

14)Listiani, M. (2009). Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg: Hamburg Univesity of Technology .

15)Matthew Botvinick, S. R.-N. (May 2019). Reinforcement Learning, Fast and Slow. Trends in cognitive sciences, 23(5), 408-422.

16)Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of Prices for Used Car by Using Regression Models. 5th International Conference on Business and Industrial Research (ICBIR), (pp. 115-119). Bangkok.

17)Nabarun Pal, P. A. (2018). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Future of Information and Communications Conference (FICC) 2018 , 1-6.Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 27-31.