Car Resale Value Prediction System

Dhwani Nimbark[1], Akshat Patel[2], Sejal Thakkar[3]

[1-2]Student, Department of Computer Engineering, Indus University, Gujarat, India
[3]Assistant Professor, Department of Computer Engineering, Indus University, Gujarat, India

Abstract - *Used car resale market in India was marked at 24.2 billion US dollars in 2019. Due to the huge requirement of used cars and lack of experts who can determine the correct valuation, there is an utmost need of bridging this gap between sellers and buyers. This project focuses on building a system that can accurately predict a resale value of the car based on minimal features like kms driven, year of purchase etc. without manual or human interference and hence it remains unbiased.*

***Key Words*: UML, ML, GBR, API, RMSE, MAE**

1.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.

1.1 Need for the System

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

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

1.2 PROJECT PURPOSE

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

1.3 PROJECT SCOPE

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

2. OBJECTIVE

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

Table -1: Sample Table format

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

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

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

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

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

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

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

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

3. PREDICTION APPROACH

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

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

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

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

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

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

Gradient boosting involves three elements:

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

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

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

3. Additive Model
Trees are added one at a time, and existing trees in the model are not changed.
A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e., follow the gradient). We do this by parameterizing the tree, then modify the parameters of the tree and move in the right direction by (reducing the residual loss.

4. 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.

5. FUTURE ENHANCEMENT

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

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

Currently, only few features are used to predict resale value of the car. This can be extended to more features. One can also implement CNN to determine physical condition of the car from images like identifying dents, scratches etc. and thus predicting more relevant resale value of a car.

6. CONCLUSION

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

7. ACKNOWLEDGEMENT

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

8. 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