Use Car Price Prediction

Ujjawal Sonkambale, Suraj Kudale, Ojas Pawar, Tushar Salunke

Department of Computer Engineering, Zeal College of Engineering and Research Pune, Maharashtra, India

ABSTRACT

In this project, we investigate different machine learning techniques to predict the price of used cars on the dataset taken from Kaggle website. Considerable number of distinct attributes is examined for the reliable and accurate prediction. The predictions are based on historical data collected from car dekho website. To build a model for predicting the price of used cars in India, we applied four machine learning techniques. The predictions are then evaluated and compared in order to find those which provide the best performances and accuracy. The final prediction model is integrated into Web application. In the future, we intend to use more sophisticated algorithms to make the predictions.

Index Terms—Ceramics, Coaxial Resonators, Delay Filters, Delay-Lines, Power Amplifiers.

I. INTRODUCTION

In the era that we tend to board nowadays, technology features a massive impact on our lives. AI, data engineering, Machine learning, Deep learning, linguistic communication process square measure rising technologies that plays an important half within the leading papers of today's world. AI is a region or branch that aims or emphasizes on making machine that works showing intelligence and their reactions is analogous thereto of human. In AI, Machine learning is a necessary and core half providing the flexibility of learning and up by itself. the main focus of this method is on creation of programs which might choose the info and learn from it by itself. Determining whether or not the listed worth of a second hand automotive may be a difficult task, thanks to the numerous factors that drive a second hand vehicle's worth on the market. the main focus of this project is developing machine learning models which will accurately predict the value of a second hand automotive supported its options, so as to form enlightened purchases. we tend to implement and value numerous learning strategies on a dataset consisting of the sale costs of various makes and models. we are going to compare the performance of assorted machine learning algorithms like rectilinear regression, Ridge Regression, Lasso Regression, Elastic internet, call Tree Regressor and select the most effective out of it. reckoning on numerous parameters we are going to confirm the value of the automotive.

II. LITERATURE SURVEY

The first paper is Predicting the value of Used automotive exploitation Machine Learning

Techniques.In this paper, they investigate the appliance of supervised machine learning tech- niques to predict the value of used cars in Mauritius. The predictions are supported historical knowledge collected from daily newspapers. totally different techniques like multiple regression analysis, k-nearest neighbors', naıve mathematician and call trees are wont to create the predictions. The Second paper is automotive value Prediction exploitation Machine Learning Techniques. Goodly range of distinct attributes are examined for the reliable and correct prediction. to make a model for predicting the value of used cars in European coun- try and Herzegovina, they need applied 3 machine learning techniques (Artificial Neural Network, Support Vector Ma- chine and Random Forest). The Third paper is value analysis model in used automotive system supported BP neural net- works. during this paper, the value analysis model supported massive knowledge analysis is planned, that takes advantage of wide circulated vehicle knowledge and an oversized range of auto dealings knowledge to research the value knowledge for every kind of vehicles by exploitation the optimized BP neural network algorithmic rule. It aims to ascertain a second- hand automotive value analysis model to induce the value that best matches the automotive.

In [4], authors explained various indicators used in GDP measurement and explained various prediction techniques used for GDP measurements.

III. METHODOLOGY

We utilized several classic and state of the art methods, including ensemble learning techniques, with a 80% - 20% split for the training and test data. Linear Regression, Lasso Regression, Random Forest and Decision tree were our base- line methods. For model implementations, the open-source Scikit-Learn package was used.

A. Linear Regression

Linear Regression was chosen as the first algorithm due to its simplicity and comparatively small training and testing time. The features, with none feature mapping, were used directly because the feature vectors. We are also uses regularization technique for the better accuracy.

B. Lasso Regression

Lasso regression is a regularization technique which is em- ployed over regression methods for a more accurate prediction. Basically it's used after linear regression. The lasso procedure encourages simple and sparse models like models with fewer parameters. This particular type of regression is compatible for models showing high levels of multi-collinearity or once you want to automate certain parts of model selection, like variable selection or parameter elimination.

C. Random Forest Regressor

Random Forest regressor is an ensemble learning regressor model which is used for better accuracy and mostly it is used on large datasets. Because of this, we are using this regressor technique for our model as it gives more accuracy compared to other algorithms. It creates a forest to evaluate results. Random Forest builds multiple decision trees by picking the 'K' number of knowledge points from the dataset and merges them to urge a more accurate and stable prediction. The training and testing time is more for the random forest as compared to linear and lasso regression. We got the maximum accuracy for random forest, so we finalize this algorithm.

D. Decision Tree

The decision tree regressor is a predictive machine learning model. This is the type of supervised machine learning. The output is very easy to interpret. You can teach someone to interpret it in a few minutes. They make no assumptions about the errors. They are excellent at finding interactions that exist in

just a part of the info . Decision tree regression observes features of an object and trains a model within the structure of a tree to predict data within the future to supply meaningful continuous output. This regressor technique also takes more time for testing and training like a random forest.

Algorithms	R² Score
1. Linear Regression	60.65 %
2. Lasso Regression	59.06 %
3. Random Forest Regressor	92.19 %
4. Decision Tree Regressor	82.86 %

Fig. 1. Algorithm Comparison

IV. TECHNOLOGY USED

A. NumPy

Here, in this our project we use python module called NumPy for manipulating the data of our dataset and do some mathematical operation to calculate desired output from that using python NumPy module. Numpy is a general- purpose array-processing package. It provides a high- performance multidimensional array object, and tools for working with these arrays. It's the abecedarian package for scientific com- puting with Python. Besides its egregious scientific uses, Numpy can also be used as an effectivemulti-dimensional vessel of general data. Array in Numpy is a table of rudiments (generally figures), all of the same type, listed by a tuple of positive integers. In Numpy, number of confines of the array is called rank of thearray.A tuple of integers giving the size of the array along each dimension is known as shape of the array. An array class in Numpy is called as ndarray. Rudiments in Numpy arrays are penetrated by using square classes and can be initialized by using nested Python Lists.

B. Panda

We have used Panda module in this project to manipulate huge dataset with the functionality of pandas module effec- tively and we get desired dataset in form in which we want by manipulating raw dataset with the help of pandas. Pandas is an open- source library that's erected on top of NumPy library. It's a Python package that offers colorful data structures and operations for manipulating numerical data and time series. It's substantially popular for importing and assaying data much easier. Pandas is fast and it has high- performance productivity for druggies. Advantage Of Panda are as follow :

1) Fast and effective for manipulating and assaying data.
2) Data from different train objects can be loaded.
3) Easy running of missing data (represented as NaN) in floating point as well asnon-floating point data.

C. Matplotlib

It is a python data visualization library or module. We have used this module in our project to visualize table format data into graph format, so that if there is any kind of error or inaccuracy so that we can easily get it out by data visualization as compared to data in table format and we able correct it immediately, so in directly it will increase our speed. Matplotlib is easy to use and an amazing imaging library in Python. It's erected on NumPy arrays and designed to work with the broader SciPy mound and consists of several plots like line, bar, smatter, histogram,etc. Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is amulti-platform data visualization library erected on NumPy arrays and designed to work with the broader SciPy mound. It was introduced by John Hunter in the time 2002. One of the topmost benefits of visualization is that it allows us visual access to huge quantities of data in fluently digestible illustrations. Matplotlib consists of several plots like line, bar, smatter, histogram etc.

D. Random Forest

We have used this machine learning algorithm in our project to calculate estimated car price. There are many such algorithm to find estimated price such as

Lasso, Linear Regression, Decision Tree, But we go with this algorithm "Random Forest" Because it gave us highest accuracy among all above mentioned. Random Forest is a popular machine literacy algorithm that belongs to the supervised literacy fashion. It can be used for both Bracket and Retrogression problems in ML. It's grounded on the conception of ensemble literacy, which is a process of combining multiple classifiers to break a complex problem and to ameliorate the performance of the model. As the name suggests,"Random Forest is a classifier that contains a number of decision trees on colorful subsets of the given dataset and takes the average to ameliorate the prophetic delicacy of that dataset." Rather of counting on one decision tree, the arbitrary timber takes the vatication from each tree and grounded on the maturity votes of prognostications, and it predicts the final affair. The lesser number of trees in the timber leads to advanced delicacy and prevents the problem of overfitting. The below diagram explains the working of the Random Forest algorithm:
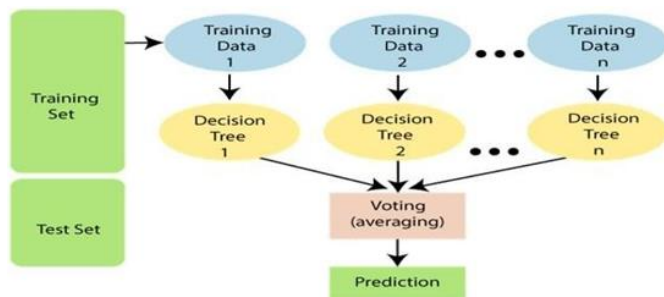

Fig. 2. Random Forest Algorithm

E.   Flask

we have used flask python module in our project, the purpose of this module is to connect backend machine learning module to frontend. The interface will be shown on website with the help of HTML, CSS and Bootstrap and flask module will help to connect ML backend module to frontend to calculate estimated price. So here used need to give some basic information related to car purchasing and this information or input will pre-process in our machine and output or estimated car price will shown on

website with the help of flask. Flask is an API of Python that allows us to build up web- applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine.

V.   SOFTWARE TESTING

Software testing can be stated as the process of verifying and validating that software or application is bug-free, meets the technical requirements as guided by its design and development, and meets the user requirements effectively and efficiently with handling all the exceptional and boundary cases. The process of software testing aims not only at finding faults in the existing software but also at finding measures to improve the software in terms of efficiency, accuracy, and usability. It mainly aims at measuring the specification, functionality, and performance of a software program or application. Software testing can be divided into two steps:

1)   Verification: it refers to the set of tasks that ensure that software correctly implements a specific function.
     Verification: "Are we building the product right?"

2)   Validation: it refers to a different set of tasks that ensure that the software that has been built is traceable to customer requirements. Validation: "Are we building the right product?"
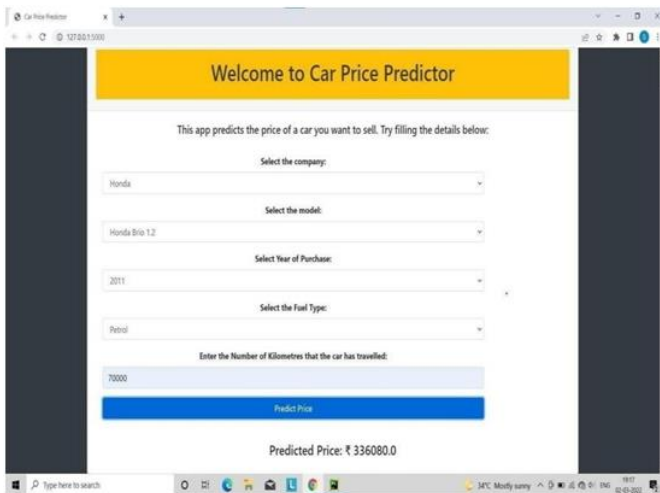
VI. RESULT

Fig. 3. Output

VII. ACKNOWLEDGEMENT

VIII. FUTURE WORK

1) As a part of future work, we aim at the variable choices over the algorithms that were used in the project.
2) In future this machine learning model may bind with various website which can provide real time data for price prediction.
3) Also we may add large historical data of car price which can help to improve accuracy of the machine learning model.
4) We plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

IX. CONCLUSION

Looking at the other researchers that are done in the similar field, our project stand apart all is there is no website to input the data and test it manually for the normal users. We are overcoming this disadvantage. The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features.

X. REFERENCES

[1]. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric ,Car Price Prediction using Machine Learning Technique , TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019.

[2]. Sameerchand Pudaruth , Predicting the Price of Used Cars using Machine Learning Techniques , International Journal of Information Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764.

[3]. Ravi Shastri1, Dr. A Rengarajan2 , Prediction of Car Price using Linear Regression, International Journal of Trend in Scientific Research and Development (IJTSRD) Volume 5 Issue 4, May-June 2021.

[4]. Aparna V.Mote, Utakarsha Musmade,Nikita Deshmukh,"Gross Domestic Product (GDP) Prediction: A Review", 2019 JETIR May 2019, Volume 6, Issue 5

[5]. Panwar Abhash Anil , Used Car Price Prediction using Machine Learning, Aug 3,

2020, https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2

[6]. Kajal Kumari, Car Price Prediction – Machine Learning vs Deep Learning, July 25, 2021, https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/