IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning For Predicting Cost of Pre-Owned Vehicles

[1]B Vysageetha, Assistant Professor, Department of Computer Science and Engineering, Dr.B.R.Ambedkar University, College of Engineering, Srikakulam.

[2]Sravani Chintada, Assistant Professor, Department of Electronics and Communication Engineering, Dr.B.R.Ambedkar University, College of Engineering, Srikakulam.

[3]Tottadi Bhavani, Assistant Professor, Department of Electrical and Electronics Engineering, Dr.B.R.Ambedkar University, College of Engineering, Srikakulam.

[4]Ch Vijayabharathi, Assistant Professor, Department of Civil Engineering, Dr.B.R.Ambedkar University, College of Engineering, Srikakulam.

ABSTRACT:

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to 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. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

Keywords: Car prediction, machine learning, ANN, SVM,Random Forest.

INTRODUCTION:

Brief Information about the Project: Determining whether the listed price of a used car is a challenging task,due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models. We will compare the performance of various machine learning algorithms like Linear Regression, Random Forest regression, Elastic Net, Decision Tree Regress or and choose the best out of it. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs. Kilometerstravelled – We know that the number of kilometers travelled by a

vehicle has a huge role to play while putting the vehicle up for sale. The more the vehicle has travelled, the older it is. Fiscal power – It is the power output of the vehicle. Moreoutput yields better value out of a vehicle. Fuel Type – There were two types of fuel types present in thedataset that we had. Gas and Diesel. It was relatively less dominant.

Objective of the paper: To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes. The system that is being built must be feature based i.e. feature wise prediction must be possible. Providing graphical comparisons to provide a better view.

Features:

- There will be majorly two features provided in the project notethat this will be not
- Re-sale platform: A centralized platform for car resale that will predict prices.
- Feature selection: Feature-based search and prediction.

SYSTEM DESIGN:

Input Design plays a vital role in the life cycle of software development. The attention of developers is required to collect the information about vehicles. The most accurate data must be entered in the input design. The design of input is more important in minimizing the errors that has been given by the user. By the rules of software engineering concepts, the validation control must be defined over the input limit in the input forms orscreens. The validation control must take care of other input related errors. The input screens have been included in almost all the modules. The alert message will be displayed whenever user did any mistakes while giving input. And also some messages will be provided in order to guide the user in correct way. By this we can achieve to get only valid details. The user created input has been converted in to computer related format. The input design is based on data entry logical. The main goal of input design is to make the form as free from errors. The input design will control the errors in the input form. The created application should be user-friendly manner. Wherever the cursor is placed in processing the input must be entered in that same place. By this way the form has been designed. There might be several options for a single input so that the user has to select suited input to get the best result. Each entered data must be validated accordingly. The error message must be displayed whenever the user enters any wrong data or irrelevant data as input. Even the user is in last page of input if he did not get the result properly then he can go to the first page and he can change the input given already. The primary output form has been created in order to get communication between the administrator and the clients. The VPN system produces output in the form of managing clients by the project leaders, in a way such as creating new clients, allotting new projects to them, have a look over table in which to get the details about project status, and the same will be accessed by each clients. A new project will be assigned to every client when he completes his old. At every initial stage of the new project, the user authentication should be maintained. A user registration can be done either by the administrator or the user can do by himself. But only the administrator must have the authority to assign the projects to each user. When the application is executed it starts running. The used browser is internet explorer and the server will start its process. The project will run on the local area network so the server machine will serve as the administrator while the otherconnected systems can act as the clients.

System Architecture:

Approachforcarpricepredictionpro posedinthispaperiscomposedofseveralstep s showinfig.

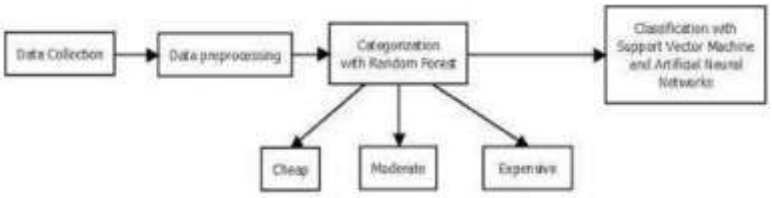


Figure3.1:Blockdiagramoftheoverallclassificatio nprocess

Data is collected from a local web portal for selling and buying cars autopijaca.ba [9], during winter season, as time interval itself has high impact on the price of the cars in Bosnia and Herzegovina. The following attributes were captured for each car: brand, model, car

condition, fuel, year of manufacturing, power in kilowatts, transmission type, millage, color, city, state, number of doors, four wheel drive (yes/no), damaged (yes/no), navigation (yes/no), leather seats (yes/no), alarm (yes/no), aluminum rims (yes/no), digital air condition (yes/no), parking sensors (yes/no), xenon lights (yes/no), remote unlock (yes/no), electric rear mirrors (yes/no), seat heat (yes/no).Since manual data collection is time consuming task, especially when there are numerous records to process, a "web scraper" as a part of this research is created to get this job done automatically and reduce the time for data gathering. Web scraping is well known technique to extract information from websites and save data into local file or database. Manual data extraction is time consuming and therefore web scrapers are used to do this job in a fraction of time. Web scrapers are programmed for specific websites and can mimic regular users from Page | 11 website's point of view.

METHODOLOGY:

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built indata structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit- test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead,when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makesthis simple approach very effective. Clearly, Python is a popular and in-demand skill to learn. But what is python programming used for? We've already briefly touched on some of the areas it can be applied to, and we've expanded on these and more Python examples below. Python can be used for: 4.2. AI and machine learning: Because Python is such a stable, flexible, and simple programming language, it's perfect for various machine learning (ML)and artificial intelligence (AI) projects.In fact, Python is among the favourite languages among data scientists, and there are many Python machine learning and AI libraries and packages available. If you're interested in this application of Python, our Deep Learning and Python Programming for AI with Microsoft Azure Expert Track can help you develop your skills in these areas. You can discover the uses of Python and deep learning while boosting your career in AI.

Data analytics:

Much like AI and machine learning, data analytics is another rapidly developing field that utilises Python programming. At a time when we're creating more data than ever before, there is a need for those who can collect, manipulate and organize the information. Python for data science and analytics makes sense. The language is easy-to-learn, flexible, and well-supported, meaning it's relatively quick and easy to use for analyzing data. When working with large amount so find formation ,it's useful for manipulating data and carrying out repetitive tasks. 4.4 Data visualization: Data visualization is another popular and developing area of interest. Again, it plays into many of the strengths of Python. As well as its flexibility and the fact it's open-source, Python provides a variety of graphing libraries with all kinds of features. Whether you're looking to create a simple graphical representation ora more interactive plot, you can find a library to match your needs. Examples include Pandas Visualization and plotly. The possibilities are vast, allowing you to transform data into meaningful in sights. If data visualization with Python sounds appealing,

check out our 12- weekExpert Track on the subject. You'll learn how to leverage Python libraries to interpret and analyze datasets.

Programming applications:

You can program all kinds of applications using Python. The general-purpose language can be used to read and create filed directories, create GUIs and APIs, and more. Whether it's block chain applications, audio and video apps, or machine learning applications, you can build them all with Python. We also have an Expert Track on programming applications with Python, which can help to kick-start your programming career. Over the course of 12 weeks, you'll gain an introduction on how to use Python, and start programming your own applications using it.

Machine Learning Introduction:

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work withmachine learning methodologies, or analyze the impact of machine learning processes. In this tutorial, we'll look into the common machine learning methods ofsupervised and unsupervised learning, and common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We'll explore which programming languages are most used in machine learning, providing you with some of the positive and negative attributes of each. Additionally, we'll discuss biases that are perpetuated by machine learning algorithms, and consider what can be kept in mind to prevent these biases when building algorithms.

Machine Learning Methods:

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed. Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labelled by humans, and unsupervised learning which provides the algorithm with no labelled data in order to allow it to find structure within its input data. Let's explore these methods in more detail.

Supervised Learning:

In supervised learning, the computer is provided with example inputs that are labelled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predictlabel values on additional unlabelled data. A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. supervised learning, tagged photos of dogs can be used as input data to classify untagged photos ofdogs.

Unsupervised Learning:

In unsupervised learning, data is unlabelled, so the learning algorithmis left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly valuable. The goal of unsupervised learning may be as

straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data. Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Random Forest :

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like Scikit-learn).

Features of a Random Forest Algorithm:

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of over fitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point. Decision trees are the building blocks of a random forest algorithm.

A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further. The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. The following diagram shows the three types of nodes in a decision tree.

The information theory can provide more information on how decision trees work. Entropy and information gain are the building blocks of decision trees. An overview of these fundamental concepts will improve our understanding of how decision trees are built. Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables. The information gain concept involves using independent variables (features) to gain information about a target variable (class). The entropy of the target variable (Y) and the conditional entropy of Y (given X) are used to estimate the information gain. In this case, the conditional entropy is subtracted from the entropy of Y. Information gain is used in the training of decision trees. It helps in reducing uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) has been removed. Entropy and information gain are important in splitting branches, which is an important activity in the construction of decision trees. Let's take a simple example of how a decision tree works. Suppose we want to predict if a customer will purchase a mobile phone or not. The features of the phone form the basis of his decision. This analysis can be presented in a decision tree diagram. The root node and decision nodes of the

decision represent the features of the phone mentioned above. The leaf node represents the final output, either buying or not buying. The main features that determine the choice include the price, internal storage, and Random Access Memory (RAM). The decision tree will appear as follows.The main difference between the decision tree algorithm and the random forest algorithm is that establishing root nodes and segregating nodes is done randomly in the latter. The random forest employs the bagging method to generate the required prediction. Bagging involves using different samples of data (training data) rather than just one sample. A training dataset comprises observations and features that are used for making predictions. The decision trees produce different outputs, depending on the training data fed to the random forest algorithm. These outputs will be ranked, and the highest will be selected as the final output. Our first example can still be used to explain how random forests work. Instead of having a single decision tree, the random forest will have many decision trees. Let's assume we have only four decision trees. In this case, the training data comprising the phone's observations and features will be divided into four root nodes. The root nodes could represent four features that could influence the customer's choice (price, internal storage, camera, and RAM). The random forest will split the nodes by selecting features randomly. The final prediction will be selected based on the outcome of the four trees. The outcome chosen by most decision trees will be the final choice. If three trees predict buying, and one tree predicts not buying, then the final prediction will be buying.

SVM:

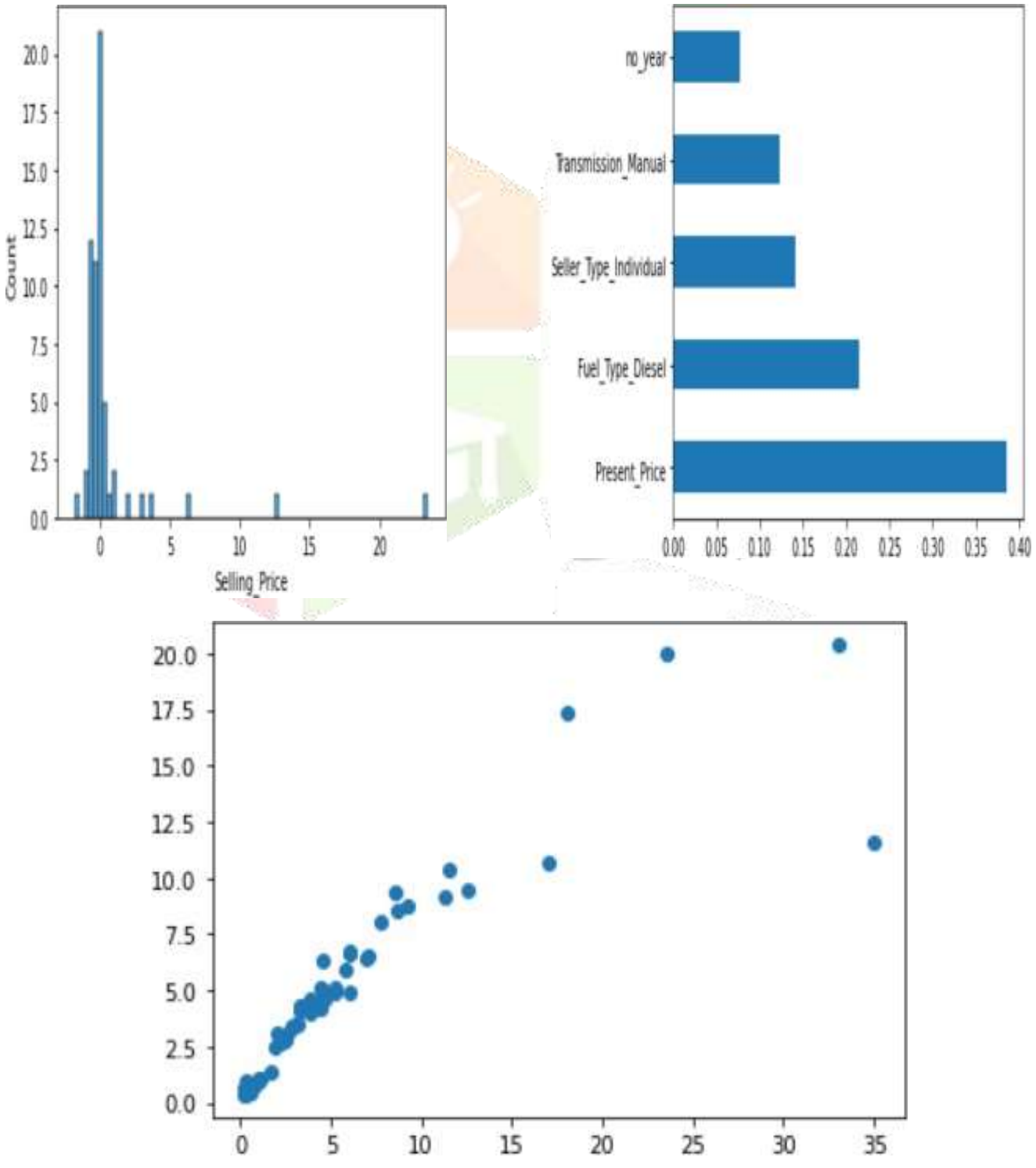
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. 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. More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier. Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2- dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So, by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to use for multi-class problems.

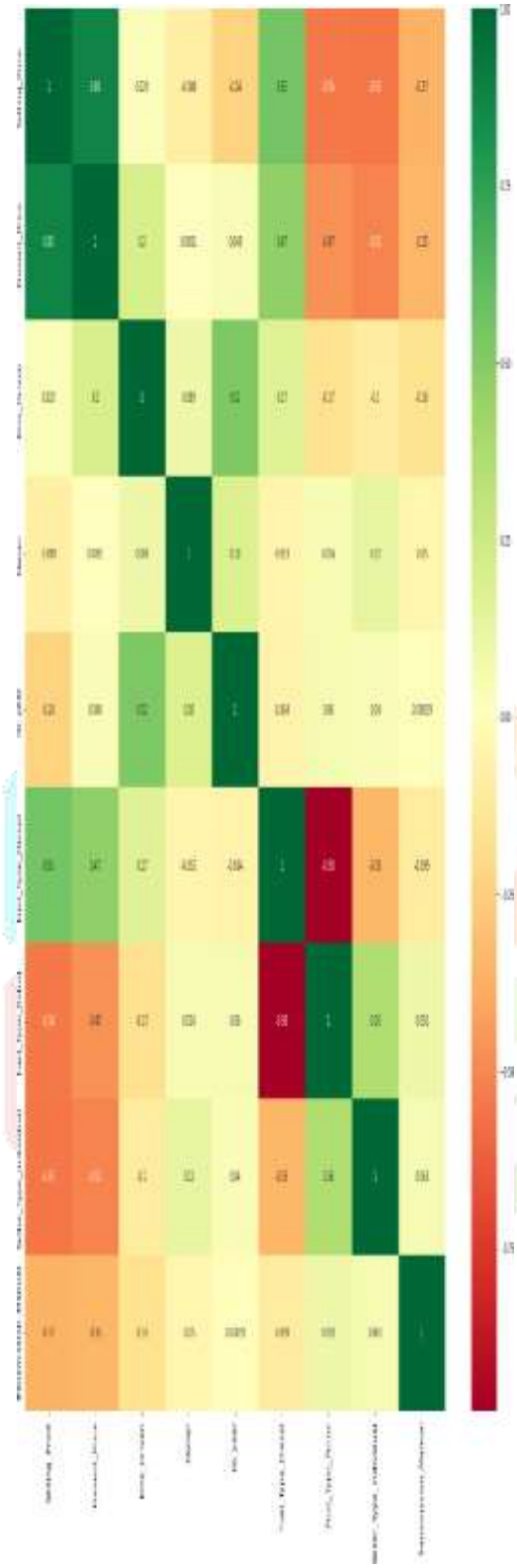
Artificial Neural Network:

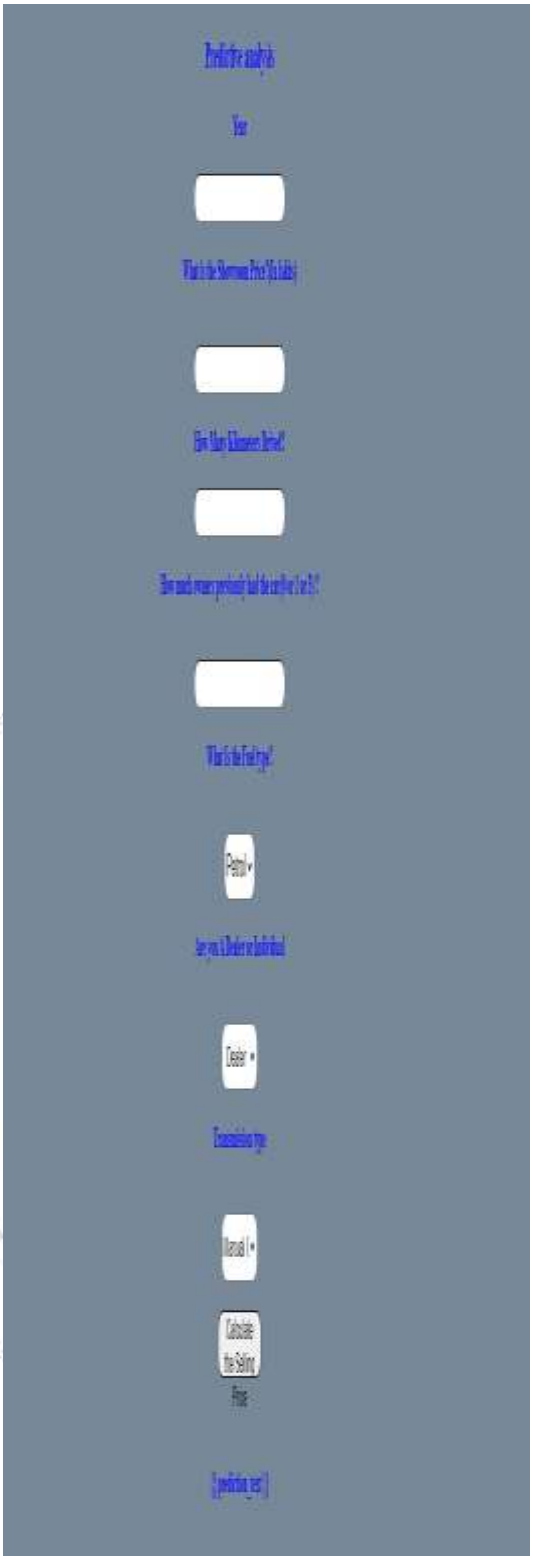
Artificial Neural Network is used for beginners as well as professions. The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes. Artificial neural network tutorial covers all the aspects related to the artificial neural network. In this tutorial, we will discuss ANNs, Adaptive resonance theory, Kohonen self-organizing map, Building blocks, unsupervised learning, Genetic algorithm, etc. What is Artificial Neural Network? The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons

interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes. 1 figure 4.3 the typical diagram of Biological Neural Network. The typical Artificial Neural Network looks something like the given figure. Page | 28 Fig 4.4 ANN Archecture Dendrites from Biological Neural Network represent inputs in Artificial Neural Networks, cell nucleus represents Nodes, synapse represents Weights, and Axon represents Output

SCREENSHOOTS:






Predictive analysis
Year
What is the Showroom Price (Ex-showroom price)?
Kms Driven (Distance Driven)?
Owner (Number of owners previously had the car before it was sold)?
What is the Fuel type?
Petrol
Are you a Dealer or Individual
Dealer
Transmission type
Manual
Calculate the Selling Price
Predict

CONCLUSION:

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. Themajor step in the prediction process is collection and preprocessing of the data. In this research, PHP scripts were built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms. Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets as the one in this research. Applying single machine algorithm on the data set accuracy was less than 50%. Therefore, the ensemble of multiple machine learning algorithms has been proposed and this combination of ML methods gains accuracy of92.38%. This is significant improvement compared to single machine learning method approach. However, the drawback of the proposed system is that it consumes much more computational resources than single machine learning algorithm

FUTURE SCOPE:

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset

REFERENCE:

[1] Agencija za statistiku BiH. (n.d.), retrieved from:http://www.bhas.ba . [accessed July 18, 2018.]

[2] Listiani, M. (2009).Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg- Harburg).

[3] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object / coccc%3A1346 [accessed: August 1, 2018.]

[4] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.Du, J., Xie, L., & Schroeder, S. (2009).

Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

WEB SITES REFERRED:

https://www.kaggle.com
https://www.wikipedia.or