CAR PRICE PREDICTION PROJECT

Submitted by:

Ganesh Kumbhar

ACKNOWLEDGMENT

[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.
[5] 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.
[6] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on* (Vol. 2, pp. 682-685). IEEE.
[7] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764.
[8] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, *167*(9), 27-31.
[9] Auto pijaca BiH. (n.d.), Retrieved from: https://www.autopijaca.ba. [accessed August 10, 2018].
[10] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].
[11] Ho, T. K. (1995, August). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on* (Vol. 1, pp. 278-282). IEEE.
[12] Russell, S. (2015). *Artificial Intelligence: A Modern Approach* (3rd edition). PE.

[13] Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journal of machine learning research*, *2*(Dec), 125-137.
[14] Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, *25*, 821-837.
[15] 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [accessed: August 30, 2018].
[16] Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-cars-database. [accessed: June 04, 2018].

[17] OLX. (n.d.), Retrieved from: https://olx.ba. [accessed August 05,2018].

INTRODUCTION

• Business Problem Framing

Car price prediction is somehow interesting and popular problem. As per information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from which 84% of them are cars for personal usage [1]. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase in future. This adds additional significance to the problem of the car price prediction.
Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent changes in the price of a fuel. Different features like exterior colour, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used car price prediction.

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model. We are about to deploy an ML model for car selling price prediction and analysis. This kind of system becomes handy for many people.

Conceptual Background of the Domain Problem

- Machine Learning is a field of technology developing with immense abilities and applications in automating tasks, where neither human intervention is needed nor explicit programming.

- The power of ML is such great that we can see its applications trending almost everywhere in our day-to-day lives. ML has solved many problems that existed earlier and have made businesses in the world progress to a great extent.

- Machine Learning models generally aim to be a solution to an existing problem or problems. And at some point in your life, you must have thought that how would your model be a solution and how would people use this? Indeed, people can't use your notebooks and code directly, and that's where you need to deploy your model.

- You can either deploy your, model, like API or a web service. Here we are using the Flask micro-framework. Flask defines a set of constraints for the web app to send and receive data.

- # Review of Literature

This project contains two phase:-

1. Data Collection Phase –

You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. More the data better the model. In this section you need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.)

You need web scraping for this. You have to fetch data for different locations. The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilo meters, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data. Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.

2. Model Building Phase-

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model. Follow the complete life cycle of data science. Include all the steps like:-

1. Data Cleaning
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building
5. Model Evaluation
6. Selecting the best model

Libraries used for Web Scraping: As we know, Python is has various applications and there are different libraries for different purposes. In our further demonstration, we will be using the following libraries:

- **Selenium:** Selenium is a web testing library. It is used to automate browser activities.
- **Beautiful Soup:** Beautiful Soup is a Python package for parsing HTML and XML documents. It creates parse trees that is helpful to extract the data easily.
- **Pandas:** Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format.

Predicting price of a used cars has been studied extensively in various researches. Listian discussed, in her paper written for Master thesis [2], that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. This is on the grounds that Support Vector Machine (SVM) is better in dealing with datasets with more dimensions and it is less prone to overfitting and under fitting. The weakness of this research is that a change of

simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation.

Another approach was given by Richardson in his thesis work [3]. His theory was that car producers produce more durable cars. Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than traditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.

Wu et al. [4] conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbour machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it [5].

- # Motivation for the Problem Undertaken

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, this deployed web application 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, kilometres driven, if dealer/individual, and finally if the transmission type is manual/automatic. And that’s a brownie point.

Any kind of modifications can also be later inbuilt in this application. It is only possible to later make a facility to find out buyers. This a good idea for a great project you can try out. You can deploy this as an app like OLA or any e-commerce app. The applications of Machine Learning don’t end here. Similarly, there are infinite possibilities that you can explore. But for the time being, let me help you with building the model for Car Price Prediction and its deployment process.

ANALYTICAL PROBLEM FRAMING

- Mathematical/ Analytical Modeling of the Problem

Web Scraping with Python: Imagine you have to pull a large amount of data from websites and you want to do it as quickly as possible. How would you do it without manually going to each website and getting the data? Well, "Web Scraping" is the answer. Web Scraping just makes this job easier and faster.

Why is Web Scraping Used?

Web scraping is used to collect large information from websites. But why does someone have to collect such large data from websites? To know about this, let's look at the applications of web scraping:

- **Price Comparison:** Services such as ParseHub use web scraping to collect data from online shopping websites and use it to compare the prices of products.
- **Email address gathering:** Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.
- **Social Media Scraping:** Web scraping is used to collect data from Social Media websites such as Twitter to find out what's trending.
- **Research and Development:** Web scraping is used to collect a large set of data (Statistics, General Information, Temperature, etc.) from websites, which are analysed and used to carry out Surveys or for R&D.
- **Job listings:** Details regarding job openings, interviews are collected from different websites and then listed in one place so that it is easily accessible to the user.

What is Web Scraping?

- Web scraping is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured. Web scraping helps collect these unstructured data and store it in a structured form. There are different ways to scrape websites such as online Services, APIs or writing your own code. In this article, we'll see how to implement web scraping with python.

Is Web Scraping Legal?

- Talking about whether web scraping is legal or not, some websites allow web scraping and some don't. To know whether a website allows web scraping or not, you can look at the website's "robots.txt" file. You can find this file by appending "/robots.txt" to the URL that you want to scrape. For this example, I am scraping Flipkart website. So, to see the "robots.txt" file, the URL is www.flipkart.com/robots.txt.

- Data Sources and their formats

- Approach for car price prediction proposed in this paper is composed of several steps, shown in Fig. 1.

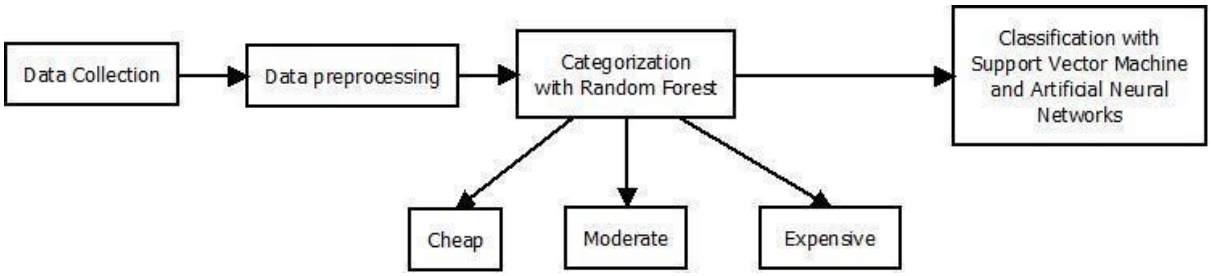


Fig-1

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, colour, 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), panorama roof (yes/no), cruise control (yes/no), abs (yes/no), esp (yes/no), asr (yes/no) and price expressed in BAM (Bosnian Mark).

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 programed for specific websites and can mimic regular users from website’s point of view.

After raw data has been collected and stored to local database, data preprocessing step was applied. Many of the attributes were sparse and they do not contain useful information for prediction. Hence, it is decided to remove them from the dataset. The attributes “state”, “city”, and “damaged” were completely removed.

- # Data Preprocessing Done

- The data that we are going to use in this example is about cars. Specifically containing various information data points about the used cars, like their price, colour, etc. Here we need to understand that simply collecting data isn’t enough. Raw data isn’t useful. Here data analysis plays a vital role in unlocking the information that we require and to gain new insights into this raw data.
- When you run the code for web scraping, a request is sent to the URL that you have mentioned. As a response to the request, the server sends the data and allows you to read the HTML or XML page. The code then, parses the HTML or XML page, finds the data and extracts it. To extract data using web scraping with python, you need to follow these basic steps:

1. Find the URL that you want to scrape
2. Inspecting the Page
3. Find the data you want to extract

4. Write the code
5. Run the code and extract the data
6. Store the data in the required format

- # State the set of assumptions (if any) related to the problem under consideration

- Consider this scenario, our friend, Otis, wants to sell his car. But he doesn't know how much should he sell his car for! He wants to maximize the profit but he also wants it to be sold for a reasonable price for someone who would want to own it. So here, us, being a data scientist, we can help our friend Otis.

  Process to convert .data file to .csv:

  a. open MS Excel
  b. Go to DATA
  c. Select From text
  d. Check box tick on comas(only)
  e. Save as .csv to your desired location on your pc

- # Hardware and Software Requirements and Tools Used

Modules needed:

- **pandas:** Pandas is an open source library that allows you to perform data manipulation in Python. Pandas provide an easy way to create, manipulate and wrangle the data.
- **numpy:** Numpy is the fundamental package for scientific computing with Python. numpy can be used as an efficient multi-dimensional container of generic data.
- **matplotlib:** Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of formats.
- **seaborn:** Seaborn is a Python data-visualization library that is based on matplotlib. Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.
- **scipy:** Scipy is a Python-based ecosystem of open-source software for mathematics, science, and engineering.

Why is Python Good for Web Scraping?

- **Ease of Use:** Python is simple to code. You do not have to add semi-colons ";" or curly-braces "{}" anywhere. This makes it less messy and easy to use.
- **Large Collection of Libraries:** Python has a huge collection of libraries such as Numpy, Matlplotlib, Pandas etc., which provides methods and services for various purposes. Hence, it is suitable for web scraping and for further manipulation of extracted data.
- **Dynamically typed:** In Python, you don't have to define data types for variables, you can directly use the variables wherever required. This saves time and makes your job faster.

- **Easily Understandable Syntax:** Python syntax is easily understandable mainly because reading a Python code is very similar to reading a statement in English. It is expressive and easily readable, and the indentation used in Python also helps the user to differentiate between different scope/blocks in the code.
- **Small code, large task:** Web scraping is used to save time. But what's the use if you spend more time writing the code? Well, you don't have to. In Python, you can write small codes to do large tasks. Hence, you save time even while writing the code.
- **Community:** What if you get stuck while writing the code? You don't have to worry. Python community has one of the biggest and most active communities, where you can seek help from.

MODEL/S DEVELOPMENT AND EVALUATION

- ## Identification of possible problem-solving approaches (methods)

  We have used the random forest regressor to predict the selling prices since this is a regression problem and that random forest uses multiple decision trees and has shown good results for my model.

```
from sklearn.ensemble import RandomForestRegressor model = RandomForestRegress
or()
```

```
hyp = RandomizedSearchCV(estimator = model, param_distributions=grid,
n_iter=10,
scoring= 'neg_mean_squared_error'
cv=5,verbose = 2,
random_state = 42,n_jobs = 1)
hyp.fit(x_train,y_train)
```

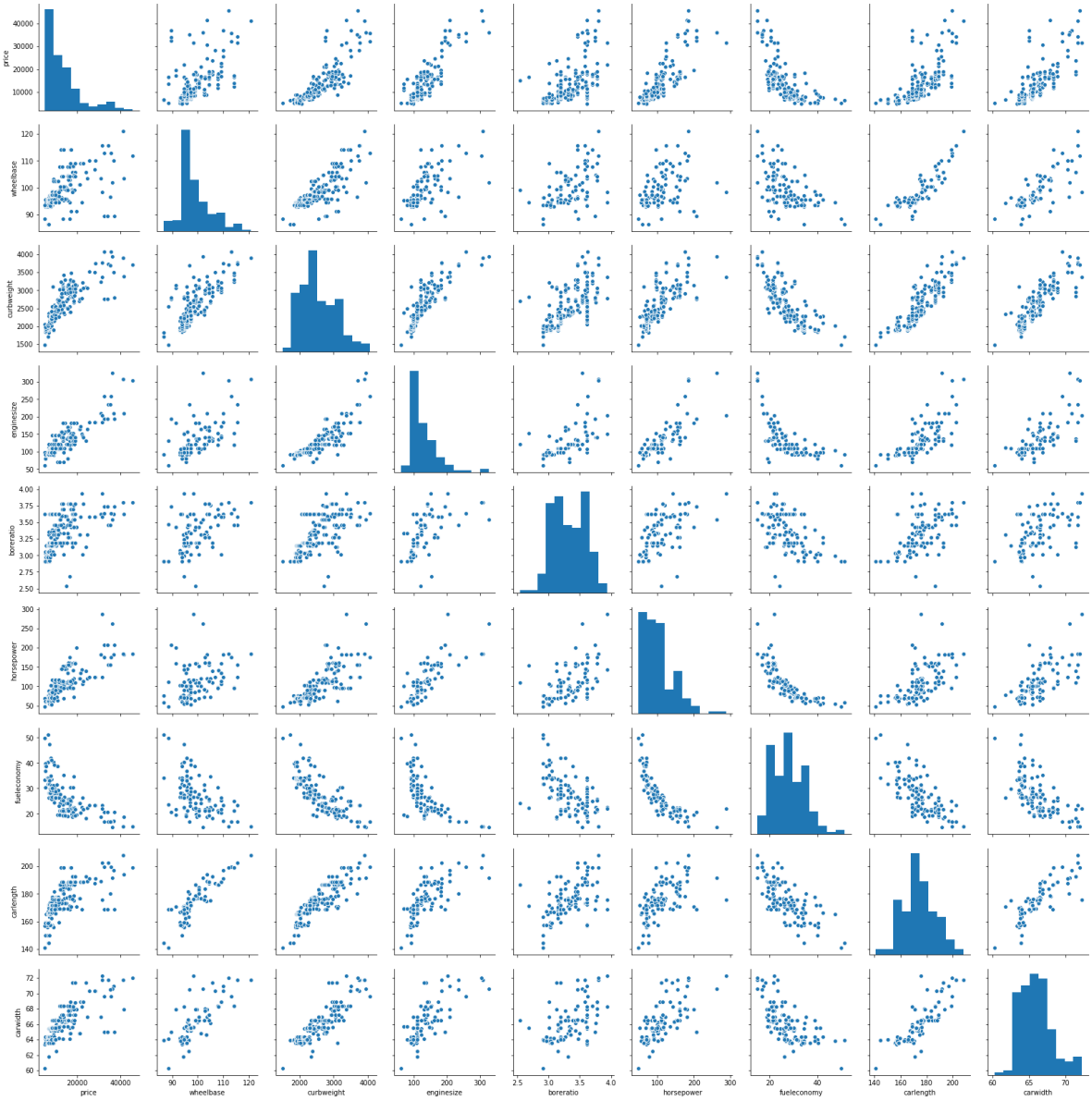
- ## Testing of Identified Approaches (Algorithms)

  1. from sklearn.metrics

  import r2_score

  O/P- 0.8614595209022039

- Visualizations

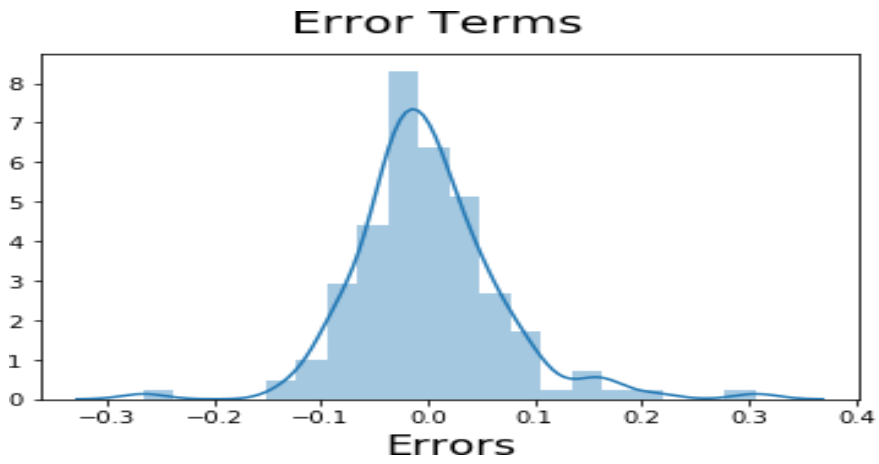


Pairplot of all the attributes of the datasets

- Interpretation of the Results

Random forest (RF) also known as random decision forest belongs to the category of ensemble methods. RF can be used for classification and regression problems. The algorithm was developed by Ho as an improvement for overfitting of the decision tree algorithms [11]. Artificial Neural Networks is the machine learning model that tries to solve problems in the same way as the human brain does. Instead of neurons, the ANN is using artificial neurons also known as perceptron. In the human brain, neurons are connected with axons while in ANN the weighted matrices are used for connections between artificial neurons. Information travels through neurons using connections between them, from one neuron information travels to all the neurons connected to it. Adjusting the weights between neurons system can be trained from input examples [12]. Support Vector Machine can be used for solving classification and regression problems. For input data set, the SVM can make a binary decision and decide in which among the two categories the input sample belongs. The SVM algorithm

is trained to label input data into two categories that are divided by the widest area possible between categories [12]. In cases when input data is not labelled, SVM algorithm cann't be applied. For unlabelled data, it is necessary to apply unsupervised learning method and SVM has its implementation called Support Vector Clustering (SVC) [13][14].



CONCLUSION

- ## Key Findings and Conclusions of the Study

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major 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 of 92.38%. 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.

Although, this system has achieved astonishing performance in car price prediction problem our aim for the future research is to test this system to work successfully with various data sets. We will extend our test data with eBay [16] and OLX [17] used cars data sets and validate the proposed approach.

Inferences:

- *R-squared and Adjusted R-squared (extent of fit)* - 0.899 and 0.896 - 90% variance explained.
- *F-stats and Prob (F-stats) (overall model fit)* - 308.0 and 1.04e-67(approx. 0.0) - Model fir is significant and explained 90% variance is just not by chance.
- *P-values* - P-values for all the coefficients seem to be less than the significance level of 0.05. - meaning that all the predictors are statistically significant.