**CAR RESALE VALUE PREDICTION WITH IBM WATSON MACHINE LEARNING**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No** | **TOPIC** | **PAGE** **No** |
| **1** | **ABSTRACT** | **2** |
| **2** | **INTRODUCTION** | **3** |
| **3** | **LITERATURE SURVEY** | **4** |
| **4** | **THEORITICAL ANALYSIS** | **5-6** |
| **5** | **EXPERIMENTAL INVESTIGATIONS** | **7** |
| **6** | **FLOWCHART** | **7** |
| **7** | **ADVANTAGES AND DISADVANTAGES** | **8** |
| **8** | **APPLICATION** | **9** |
| **9** | **RESULT AND CONCLUSION** | **9** |
| **10** | **FUTURE SCOPE** | **10** |
| **11** | **APPENDIX** | **10-13** |

**1.ABSTRACT**

Determining the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price in 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 across cities in many states.

In order to predict the resale value of a car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithm with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product.

**2. INTRODUCTION**

Predicting the price of used cars is both an important and interesting problem.

According to data obtained from the National Transport Authority, the number of cars registered between 2003 and 2013 has witnessed a spectacular increase . With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright.

After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the installments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the installments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well.

Predicting the resale value of a car is not a simple task. It is trite knowledge that

the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, weight of the car, consumer reviews, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, power steering, GPS navigator all may influence the price as well. The look and feel of the car certainly contributes a lot to the price. As we can see, the price depends on a large number of factors.

**3. LITERATURE SURVEY**

**3.1 Existing problem** ; The prices of new cars in the industry are 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 price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling.

**3.2 Proposed solution:** In order to predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithm with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product. In this project, random forest is been used. The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

**4. THEORITICAL ANALYSIS**

**4.1 BLOCK DIAGRAM**

**4.2 Hardware / Software designing:**

**Software Requirements:**

Python 3 - We have used Python which is a statistical mathematical programming language like R instead of MATLAB due to the following reasons:

1. Python code is more compact and readable than MATLAB

2. The python data structure is superior to MATLAB

3. It is an open source and also provides more graphic packages and data sets Keras (with TensorFlow backend 2.3.0 version) - Keras is a neural network API consisting of TensorFlow, CNTk, Theano etc. Python packages like Numpy, Matplotlib, Pandas for mathematical computation and plotting graphs, SimpleITK for reading the images which were in .mha format and Mahotas for feature extraction of GLCM Kaggle was used to obtain the online dataset. GitHub and Stackoverflow was used for reference in case of programming syntax errors. OpenCV (Open Source Computer Vision) is a library of programming functions aimed at real time computer vision i.e. used for image processing and any operations relating to image like reading and writing images, modifying image quality, removing noise by using Gaussian Blur, performing binary thresholding on images, converting the original image consisting of pixel values into an array, changing the image from RGB to grayscale etc. It is free to use, simple to learn and supports C++, Java, C, Python. Its popular application lies in CamScanner or Instagram, GitHub or a web-based control repository. Google Colaboratory (open-source Jupyter Notebook interface with high GPU facility) - Google Colab /Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely on cloud. With Colab, one can write and execute code, save and share analyses, access powerful computing resources, all for free from browser.[Jupyter Notebook is a powerful way to iterate and write on your Python code for data analysis. Rather than writing and rewriting an entire code, one can write lines of code and run them at a time. It is built off of iPython which ©RCCIIT, DEPT. OF EE Page 20 is an interactive way of running Python code. It allows Jupyter notebook to support multiple languages as well as storing the code and writing own markdown.]

**Hardware Requirements:**

Processor: Intel® Core™ i3-2350M CPU @ 2.30GHz

Installed memory (RAM):4.00GB

System Type: 64-bit Operating System

**5. EXPERIMENTAL INVESTIGATIONS**

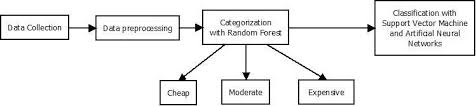
There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms can be chosen according to the objective. As the dataset which we are using is a regression dataset so you can use the following algorithms

* Multi Linear Regression
* Random Forest Regression / Classification
* Decision Tree Regression / Classification
* K-Nearest Neighbors
* Support Vector Machine

The model which is been used here is random forest.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

**6. FLOW CHART**



**7. ADVANTAGES & DISADVANTAGES**

**Advantages:**

Can be able to predict used cars market value can help both buyers and sellers.

**Used car sellers (dealers):** They are one of the biggest target group that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

**Online pricing services:** There are websites that offers an estimate value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car’s market value.

**Individuals:** There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it’s a big corner to pay too much or sell less then it’s market value.

**Disadvantages**: Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven.

**8. APPLICATIONS**

* With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.
* In order to predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle.

**9.RESULT AND CONCLUSION**

* In the given guided project I understood the problem to classify if it is a regression or a classification kind of problem.
* I also came to know how to pre-process the data using different data-preprocessing techniques
* Not only this, I also grasp the knowledge about applying different algorithms according to the dataset
* I also learn about the features of flask application

**10.FUTURE SCOPE**

By using deep learning the system can be made more proficient in predicting performance. Web application which is been made using flask can be improved in order to make it more user-friendly. As a result, people would use the web-application more and get the benefit from it before consuming cars. Apart from this, the feature which are been considered while making a prediction can be enlarge so that the accuracy level would boost-up.

**Appendix**

**DATA PRE-PROCESSING:**

**import matplotlib as plt**

**from sklearn.preprocessing import LabelEncoder**

**import pickle**

**df = pd.read\_csv(r"C:\Users\Sri Priya\Downloads\Data Car Resale Value Prediction\Data\autos.csv", header=0, sep=',', encoding='Latin1',)**

**print(df.seller.value\_counts())**

**df[df.seller != 'gewerblich']**

**df=df.drop('seller',1)**

**print(df.offerType.value\_counts())**

**df[df.offerType !='Gesuch']**

**df=df.drop('offerType',1)**

**print(df.shape)**

**df = df[(df.powerPS > 50) & (df.powerPS < 900)]**

**print(df.shape)**

**df = df[(df.yearOfRegistration >= 1950) & (df.yearOfRegistration < 2017)]**

**print(df.shape)**

**df.drop(['name','abtest','dateCrawled','nrOfPictures','lastSeen','postalCode','dateCreated'], axis='columns', inplace = True)**

**new\_df = df.copy()**

**new\_df = new\_df.drop\_duplicates(['price','vehicleType','yearOfRegistration','gearbox','powerPS','model','kilometer','monthOfRegistration','fuelType','notRepairedDamage'])**

**new\_df.gearbox.replace(('manuell', 'automatik'), ('manual','automatic'), inplace=True)**

**new\_df.fuelType.replace(('benzin','andere','elektro'),('petrol','others','electric'),inplace=True)**

**new\_df.vehicleType.replace(('kleinwagen','cabrio','kombi','andere'),('snall car','convertible','combination','others'),inplace=True)**

**new\_df.notRepairedDamage.replace(('ja','nein'),('Yes','No'),inplace=True)**

**new\_df = new\_df[(new\_df.price >= 100) & (new\_df.price <=150000)]**

**new\_df['notRepairedDamage'].fillna(value='not-declared', inplace=True)**

**new\_df['fuelType'].fillna(value='not-declared', inplace=True)**

**new\_df['gearbox'].fillna(value='not-declared', inplace=True)**

**new\_df['vehicleType'].fillna(value='not-declared', inplace=True)**

**new\_df['model'].fillna(value='not-declared', inplace=True)**

**new\_df.to\_csv("autos\_preprocessed.csv")**

**labels = ['gearbox','notRepairedDamage','model','brand','fuelType','vehicleType']**

**mapper = {}**

**for i in labels:**

**mapper[i] = LabelEncoder()**

**mapper[i].fit(new\_df[i])**

**tr = mapper[i].transform(new\_df[i])**

**np.save(str('classes'+i+'.npy'),mapper[i].classes\_)**

**print(i,":",mapper[i])**

**new\_df.loc[:,i + '\_labels'] = pd.Series(tr, index=new\_df.index)**

**labeled = new\_df[['price','yearOfRegistration','powerPS','kilometer','monthOfRegistration'] + [x+"\_labels" for x in labels]]**

**print(labeled.columns)**

**Y = labeled.iloc[:,0].values**

**X = labeled.iloc[:,1:].values**

**Y = Y.reshape(-1,1)**

**from sklearn.model\_selection import cross\_val\_score, train\_test\_split**

**X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state = 3)**

**from sklearn.ensemble import RandomForestRegressor**

**from sklearn.metrics import r2\_score**

**regressor = RandomForestRegressor(n\_estimators=1000,max\_depth=10,random\_state=34)**

**regressor.fit(X\_train, np.ravel(Y\_train,order='C'))**

**y\_pred = regressor.predict(X\_test)**

**print(r2\_score(Y\_test,y\_pred))**

**filename = 'resale\_model.sav'**

**pickle.dump(regressor, open(filename, 'wb'))**