**VEHICLE-RESALE VALUE PREDICTION**

**BACKGROUND**

**1.1 INTRODUCTION**

Artificial intelligence is a technique for teaching a computer, a robot that is controlled by a computer, or a piece of software to think critically, much like an intelligent person would. It is possible to create intelligent software and systems by first studying how the human brain functions, as well as how people learn, make decisions, and work when attempting to solve a problem.

An application of artificial intelligence that gives computer systems the ability to learn from data and get better with experience without being explicitly programmed is known as machine learning. It is a subfield of computer science.

Basically, the main goal of machine learning is to let computers learn on their own, without assistance from humans. Starting with data observations is a good place to start. The data can be some examples, instruction or some direct experiences too. Then on the basis of this input, machine makes better decision by looking for some patterns in data.

**1.2 OBJECTIVES OF RESEARCH**

* The main objective of vehicle-resale prediction model is to determine the resale value of a second-hand vehicle by considering vehicle attributes.
* This Model is very much useful for vehicle makers in second-hand market by predicting the resale value of the vehicle.

**1.3 PROBLEM STATEMET**

Car Value Prediction is essential for many individuals and companies like Banks, Insurance, Car Retailers, Vehicle Manufacturers. Vehicle makers face several challenges in the second-hand market. The depth crisis in the European Union, the general problem of overcapacity, increasing competition from Asian manufacturers, and the trend toward more eco-friendly cars are only a few factors that add to the difficulty of selling used vehicles in the second-hand market and decrease sales margins. Therefore, Vehicle makers require sophisticated decision support systems to sustain the profitability of the used vehicle business. A core component of such systems is a prediction model that estimates resale prices on the basis of vehicle attributes .Although a statistical modelling of resale prices has been considered in previous work (e.g., Purohit, 1992), only very few studies have explicitly attempted to predict resale prices with maximal accuracy to support decision making. As a consequence, we don’t know to which degree are resale prices predictable, what is the relative accuracy of different prediction methods and are some methods particularly effective.

Thousands of used automobiles are sold each day all around the world. The ability to predict the price of a used car serves as a crucial benchmark for both private buyers and sellers, as well as business experts like auto dealers, lenders, and insurance providers.

Banks need to be aware of the precise value of used cars because they frequently hold liens or are transferring loans from one borrower to another. Since they will be determining premiums when determining their risk profile, insurance companies must be able to evaluate the value of the used cars.

On doing the literature survey of various methods for vehicle-resale value prediction, we come to the conclusion that to predict the resale value there are multiple approaches like:

* Multi-Linear Regression
* Decision Tree Regression
* Random Forest Regression

Multi-Linear Regression:

Multiple Linear Regression is a simple and common way to analyse linear regression. The model is often used for predictive analysis since it defines the relationship between two or more variables. Multiple Linear Regression attempts to model the Relationship between two or more features and a response by fitting a linear equation to observed data.

Decision Tree Regression:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Random Forest Regression:

Random forest Regression is an ensemble machine learning algorithm that is used for regression problems. This algorithm is mainly used to forecast values based on some attributes. Random Forest algorithm is easy to use and accurate Machine Learning Algorithm.

**2. DATA COLLECTION**

Dataset: This project's dataset was obtained from Kaggle.com. Each data point in the original dataset is a personal advertisement posted on a German used automobile sales website. As a result, the dataset is extremely random and has a high percentage of missing data, the majority of which cannot be filled in using interpolation or other common fill techniques. Approximately 232 000 data points out of 377 000 data points, or 62% of the original data, were left after all incorrect data points and non-fillable missing values were eliminated. Since data precision has a significant impact on model quality, data fidelity was favoured over the total number of data points still available. At first, all strings were capitalized, and unnecessary columns were deleted. There were some variable names and elements that be translated from German to English to provide ease of use.Over 370000 used vehicles scraped with scrapy from Ebay-Kleinanzeigen .

**3. DETAIL DESIGN OF FEATURES**

Those fields are included in autos.csv:

* Date Crawled : when this ad was first crawled, all field-values are taken from this date
* name : name of the vehicle
* seller : private or dealer
* offer Type
* price : the price on the ad to sell the vehicle
* vehicle Type
* year Of Registration : at which year the vehicle was first registered
* gearbox
* power PS : power of the vehicle in PS
* model
* kilo meter : how many kilo meters the car has driven
* month Of Registration : at which month the vehicle was first registered
* fuel Type
* brand
* not Repaired Damage : if the vehicle has a damage which is not repaired yet
* Date Created : the date for which the ad at e-bay was created

**4. METHODOLOGY**

**4.1 Data Cleaning**

To eliminate inaccurate data, a multi-stage filtering process was implemented. Since the data set was created by amateurs, errors such as 20000 horsepower engines or car values of 0 euros were discovered. Some automobiles were grossly underpriced, while others had inflated prices. These outliers were quickly identified thanks to exploratory data analysis using graphs and some web research on average costs. Special attention had to be given to automobiles priced under 1000 euros because it is typical for dealers to rent out cars in this price range rather than sell them and because they are using this platform illegally to post such advertising. Due to the impossibility of having registration years lower than 1800 and higher than 2016, these values had to be deleted using a two-stage filter.

Since registration dates cannot be later than the data crawl date, registration year values as low as 1800 and as high as 2016 have to be excluded using a two-stage filter.

A function to fill in the missing model information was coded after the initial cleaning stage.

The developed algorithm searches for terms in the ad title to construct a brands-models dictionary and locates all entered model values for the appropriate brand. While informative, this technique also revealed a flaw in the dataset: Many of their models were referred to as "andere," which means "others," and non-German vehicles were typically categorized under the designation "sonstige autos," which means "other cars." Foreign brands and their models were added to the brands-models dictionary in order to lower the deleted number of rows.

The trim information, engine horsepower, and other crucial strings might be recovered.

However, this would involve data mining and go beyond the parameters and purpose of this study.NLP and algorithms were therefore not used.

There will still be some missing values after the code that was built to retrieve information from the name column to fill in the blanks in the brand and model columns has been executed. These pertain to the gasoline type, vehicle type, and transmission. For these columns, there are no intermediate values, therefore interpolation would not be possible. Filling either forward or backward would result in inaccurate data. For the benefit of the model, leaving those rows out was the best course of action.

Since some vehicles and models have different body types and engines, some missing values in the type of vehicle and type of fuel columns cannot be filled. It would also be imprecise to say whether a body type missing value for a particular brand or model is sedan or hatchback because there is no way to determine the actual value. A Mercedes Benz E Klasse, for instance, has both a sedan and a station wagon body type, as well as both a diesel and a gasoline engine; therefore, if any of these numbers were absent, the entire row was rejected because it would only be possible to determine those values by guessing.

Another EDA was performed after coding algorithms to fill in missing values, and any remaining outliers were deleted.

The data frame is ready for examination after the big dataset has been cleaned and any unnecessary values have been removed. The covariance matrix's size will be constrained because the dataset's variables are largely binary, like the "salvaged" value, or mostly strings, like "brand" and "model," rather than continuous numeric variables.

**4.2 EXPLORATORY DATA ANALYSIS**

**4.2.1 Figures and Tables**

Chart

Description automatically generated

Fig 1 : Heat-Map

It was time to assess the key elements influencing the value of the used automobile after the data had been cleaned. Below is a heat map that was created to show the correlation coefficients.

Transmission type, fuel type, and other columns that are not continuous numeric variables are automatically blocked by an internal algorithm built into Pandas. During the machine learning phase, these will be handled as a component of the classification issue.

More graphical analysis were conducted to examine the factors influencing the pricing in greater detail.

In the age-price graph shown below, the regression is represented by the red line. To overlay, the regression line was moved upward. Age and price have a definite negative correlation, which makes the pattern obvious. On the top of the following page, you can see the heat map created to show the covariance.

Please be aware that the algorithm ignores factors such as fuel type, transmission type, and others and only processes continuous numeric data. The classification issue will be dealt with in the machine learning stage using binary or discrete variables.

Table

Description automatically generated

Fig 2 : Correlation Table

A picture containing text

Description automatically generated

Fig 3.1. Erroneos data for YOR vs Price value

Chart

Description automatically generated

Fig 3.2. Erroneos data for PowerPS vs Price value

Chart, histogram

Description automatically generated

Fig 4 : Joint Plot of name length and price

**4.3 STATISTICAL TECHNIQUES AND DATA VISUALIZATION**

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral to yourself and stakeholders than measures of association or significance.

The following list of visualization techniques are used in the Predicting vehicle-resale value are:

**Bar charts:** Rectangles (bars) are used to represent different entities, where the height or the width encodes quantitative values. For example, we have an example of stacked bar charts that is being utilized for ranking. This visualization technique is well suited to represent relative differences. There are different types of bar charts such as horizontal bar charts, stacked bar charts, and range bar charts that were classified equally in this category.

Chart, bar chart

Description automatically generated

Fig.5 Bar chart for ‘Fuel type’ vs ‘gear box’

**Scatter plots.** These are graphs where each sample is represented by a point or symbol. Each point or symbol position is defined according to two dimensions, or two generated features of these samples. Those graphs are useful to illustrate trends and correlations.

**Heat Maps.** This graphical representation represents values by colors. In fraud detection this technique is usually used to visually query for patterns or outliers in a large amount of data.

**5. DATA MODELLING USING SUPERVISED ML TECHNIQUES - IMPLEMENTATION**

Random forest is a type of ensemble supervised machine learning algorithm. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

To implement a Random Forest Regression Model using Python. we need to import a fews libraries:

**from sklearn.ensemble import RandomForestRegressor**

An object is created for RandomForestRegressor

**forest = RandomForestRegressor()**

we may also consider some parameters like criterion, random\_state.

Then we have to fit the model by using training data by using fit method.

**forest.fit(X\_train, y\_train)**

At last, we have to predict the values by using predict method and predicted values are stored in a variable.

**y\_predict=forest.predict(X\_test)**

If we want to check the accuracy of a model, r2\_score is used which is to be imported from sklearn.metrics

The parameters for r2\_score is y\_test which contains independent

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

**r2\_score(y\_test,y\_predict)**

**6. PRELIMINARY RESULTS**

In the below figure, we see that the model has predicted the value of the used car to be $1366 for the below values of input:

Power PS = 110

Kilo = 150000

YOR = 2003

MOR = 3

Gearbox = 1

Namelen = 18

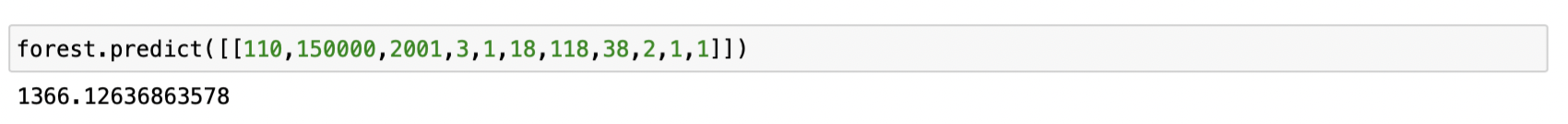
Model = 118

Brand = 38

Not repaired = 2

Fueltype = 1

Vehicletype = 1



The model has given out an accuracy of 82%.

Background pattern

Description automatically generated

**REFERENCES**

<https://www.kaggle.com/ddmngml/trying-to-predict-used-car-value/data>

<https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

<https://www.datascience.com/blog/classification-random-forests-in-python>

**CONCLUSION**

A significant amount of data from websites that sell cars was cleaned, filtered, and evaluated. Power, year of registration, and kilometers, as well as the importance of which have been identified thus far in our effort. These preliminary results will be used in the prediction model together with the classification data. To achieve better results when needed, the cleaning phase may need to be revisited.

Each machine learning problem is closely related to the dataset and the problem-solving phase starts with a clean dataset and good understanding of the weight of those parameters in building up the model. The dataset was not prepared by professionals and a rigorous clean-up process had to be carried out to minimize the loss of data and to zero out faulty data feed to the prediction models. The observations being quite arbitrary, coupled with great percentage of missing data posed great difficulty. Missing values, most of which could be normally completed with interpolation or other conventional fill methods had to be picked out deleted, since all were non-fillable missing values. This resulted in loss of approximately %38 percent of the original data. However, despite the spread and diversity of the data, adequate data fidelity was achieved and the extensive data wrangling efforts paid off with satisfactory model outputs.

Any machine learning model would find it challenging to deal with second hand cars with values ranging from 200 euros to 1.25 million euros that were produced between 1950 and 2016. The model had to handle data with a mean price of 6894€ and a standard deviation of prices of 10558€. There are other brands with comparable characteristics but a broad variety of costs. The first plan was to use a Scikit-learn-based linear regression model, however the results were unsatisfactory because the data included both continuous variables and classifiers, and linear models cannot accept classifiers as input and must instead disregard them. Regression tree models, once more from Scikit learn, were the next strategy used to better model this heterogeneous dataset.

In this project, three different machine learning techniques are used to forecast the price of used vehicles. The accuracy obtained for Multi-Linear Regression is 0.5, for Decision Tree Regression is 0.7 and for Random Forest Regression is 0.8. So, we prefer Random Forest Regression as it has more accuracy. Finally, we are building a machine learning model by using Random Forest Regression to predict the resale value of used vehicles by considering some attributes of a vehicle.