**Data Mining Project**

**Title: Pre-owned Car Evaluation**

**Team 5**

Logo

Description automatically generated

**Team Members:**

Sudeep Kakarla

Taraka Sai Srinath Reddy Vanukuri

Harish Reddy Lallapati

Sai Sravani Nalamalpu

Mohith Chandra Kanti

**INTRODUCTION:**

Cars are a necessity. Without one, most people couldn’t hold a job, buy groceries, or shuttle kids to school. But cars are also a major expense, devouring a large chunk of their owners’ incomes. Before jumping into the car market, you need to honestly decide what you can afford, a calculation that includes more than sticker prices. Most people finance their purchases, taking on years of monthly loan payments that could cause more problems than owning a car solves. Creating a realistic (not optimistic) budget is the key to making it work. Car costs shouldn’t exceed 20% of your income. Since insurance, maintenance, tolls, parking expenses, and other costs are part of that 20%, you should limit your car-loan payment to less than 10% of take-home pay. Remember that car values rapidly depreciate and the vehicle you bought new last year would be worth far less today if you had to sell it. So, it’s very important to understand the prices of cars. We have adopted the data set from Kaggle called Car Evaluation. Here we will predict the prices of the used cars using the same dataset.

This dataset contains a set of variables, each representing a custom parameter in a used car. For example, a variable could be Automatic transmission or a Fuel type.

**Problem Specification:**

Automobile companies and resellers want to understand the factors that are affecting the pricing of used cars following an algorithmic approach to set up their reselling units. In order to predict the price of used cars significant variables must be identified that truly impact the price factor. Hence the name Car Evaluation is chosen as the title for this project. Here Evaluation means looking into the key parameters that can get us the estimated price as a conclusion.

**Questions Answered:**

We aim to perform a few data mining techniques by following certain procedures with an algorithmic approach to predict the price of used cars. With the use of this dataset, we were able to put some of the principles we had learned in class and a few fresh ideas to use, such as data cleaning, hyperparameters tuning, etc. So, now we can look at this problem as predicting the price of the used car with available independent variables as far as we could. We have to make sure that the independent variables we are considering should not impact the quality of prediction of the price of used cars. With all the data we have, the only standard we know is that the price prediction should not be impacted too far when we use certain regression models.

**Data Characteristics:**

Table

Description automatically generated

A screen snapshot of the dataset with several attributes is shown in the above figure. As per the data we considered “Price” as the target variable and the rest of the attributes are independent variables. Again, all 16 independent variables are categorized as Numerical and categorical variables.

**Numerical variables:** [‘Levy’, ‘Production Year’, ‘Engine volume’, ‘Mileage’, ‘Cylinders’, ‘Doors’, ‘Air bags’]

**Categorical variables:**[‘manufacturer’, ‘model’, ‘Category’, ‘Leather interior’, ‘Fuel type’, ‘Gearbox type’, ‘Wheel’, ‘color’]

**Visualizing the Dataset:**

The below figure is a screen capture from the Azure ML studio which shows the characteristics of the data. It shows that there are 13418 entries or rows of data available against 17 columns or features.

Table

Description automatically generated with medium confidence

**Heatmap to show the correlation between variables:**

In order to visualize multiple correlations and understand the correlation between the Target variable “Price” and numerical variables heatmap is plotted.

A screenshot of a computer

Description automatically generated with low confidence

Correlation of **target variable “Price”** with **Numerical variables:**

* Looking at the above correlation plot in the heatmap we can interpret that there is no strong correlation between the Target variable “Price” and other Numerical variables
* We could also notice that Price is negatively correlated to numerical attributes such as Mileage, Doors, and Airbags.

**Train and Test Split:**

As we are using the Regression model, training the data gives more probability for the model to perform better. The data set is passed through a data transformation module called Split Data for training and testing. A 70-30 split approach is used here. It means 70% of the data we considered will be used for training and the rest 30% of data will be used for testing the model.

To avoid imbalanced or partial train and test split a stratified split is enforced for this dataset. For the model to perform well when tested with a case that is different from the training data, we wanted to make sure that it has been trained against all the possible scenarios.

**Data Mining Model Construction:**

**Data Cleaning:**

In the dataset that we have considered, a few missing values in some attributes were identified. While the models that we are going to build would not perform well with missing values there is a great need of cleaning the data. We have replaced a few instances with the mean value and excluded a negligible amount of data.

**Model Building:**

To see which model or algorithm is best suited for this type of problem the following Regression algorithms are used on the dataset by comparing their metrics in Azure ML.

1. **Decision Forest Regression**

* It is a type of supervised machine learning algorithm that is used to regress the data using true or false to certain questions.

**2.** **Boosted Decision Tree Regression**

* It is a type of additive model that combines decisions from a set of base models to make predictions.

**3. Linear Regression**

* It is a type of supervised machine learning method which is used to find a linear equation that best describes the correlation of explanatory variables with the dependent variable.

**Data Mining Models justification:**

In our dataset, we know our dependent variable. Hence it comes under the supervised learning section. Supervised learning is mainly classified into classifiers and regression types.

In the regression type, we have chosen the Decision Forest Regression, Boosted Decision Tree Regression, and Linear Regression models preferably over other models for the following reasons.

* For the data set we considered these models are giving the best accuracy as expected when compared with other models.
* As we are dealing with business problems making informed decisions is very necessary. The above models will assist us to analyze the results effectively. We can also use these models to find new patterns in business relations.

For Example, these data mining models can be used to examine car purchasing or engagement patterns by customers to predict when their services or products might reach higher demands

**Metrics for Evaluation:**

Since this is a regression problem, we have the following metrics to compare its performance with the model above:

* Mean Absolute Error
* Root Mean Squared Error
* Relative Absolute Error
* Relative Squared Error
* Coefficient of Determination (this is very important)

**Visualization of Metrics:**

**Decision Forest Regression:**

Diagram

Description automatically generated with medium confidence

**Boosted Decision Tree Regression** **Linear Regression**

Table

Description automatically generatedTable

Description automatically generated

Chart

Description automatically generated Chart

Description automatically generatedChart, waterfall chart

Description automatically generated

**Decision Forest Regression Boosted Decision Tree Regression** **Linear Regression**

**Interpretation of Metrics:**

* Mean Absolute Error has been higher for Linear Regression and the other two models performed a lot better in terms of the same.
* The coefficient of determination which is also called the R^2 value is also better for the Boosted Decision Tree Regression and Decision Forest Regression. There is not much difference in the values for the latter two models.
* This shows that our model is working fine and is not showing any bigger issues.

**Screen Capture of DM model showing all the modules:**

Diagram

Description automatically generated

**Hyperparameter Tuning:**

Hyperparameter tuning means finding the optimal set of hyperparameter values ​​for a learning algorithm while applying the optimized algorithm to an arbitrary dataset

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

* As per the results above we can clearly say that after applying Hyper Tuning on the best model, which is Boosted Decision Tree Regression, the coefficient of Determination has increased from 0.68 to 0.74 which makes the model perform better prediction of the price used cars.

**The business value of this Analytics:**

With all the Analytics and techniques used above, we have the following business value that would interest or help a business stakeholder:

* For most of the automotive companies and car resellers this model helps a lot in automating the process of prediction of price for used cars.
* As this model gives the accuracy of giving data as input which is very helpful to maintain transparency between business and customers that makes deals profitable.
* Consumption of time can be saved due to this above model which helps customers to do Remote evaluations at the same place.