UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

USED CAR PRICES CASE STUDY – DELIVERABLE DESCRIPTION

Projects form an important part of the education of software engineers. They form an active method of teaching, as defined by Piaget, leading to a "*training in self-discipline and voluntary effort*", which is important to software engineering professionals. Two purposes served by these projects are: education in professional practice, and outcomes-based assessment.

The course project is concerned with Multivariant Data Analysis and model building for response variables for scraped data of used cars, which have been separated into files corresponding to each car manufacturer (only Mercedes, BMW, Volkswagen and Audi cars are to be considered): Y- Price (Numeric Target) and binary factor Y.bin-'Audi' (Binary Target) are the targets. A binary target indicating whether an Audi car is referred in the register will be produced to fully cover the contents of the course.

Aim is to predict how much you should sell your old car. It involves a numeric outcome. As a secondary objective, given car characteristics you will have to predict if it refers to an Audi car.

A random sample containing 5000 registers combining Audi, VW, Merc and BMW registers has to be retained by each group. Data from:
https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

Conclusions to *exploratory multidimensional statistics* are useful to enlighten some aspects of possible behavior and "predictors" for price (£) or Audi binary target. Their relationships with the other variables has to be established previously to model building.

Models to be considered are: general linear model for General Score (extension of classical regression to allow the presence of factors as explicative variables) included in Statistical Modeling Topics I and II and binary regression for Y.bin (Statistical Modeling III).

Common Sections for the Documentation of Final Deliverable

The following pages are to be included in the deliverable:

I. Cover page (contents & layout)
   a. Name of Document
   b. Author's name(s)
   c. Date
II. Validation of the Data Set: description of the process (Univariate Descriptive Analysis should be included for each variable).
III. Data Imputation for selected numeric variables
IV. Data Imputation for selected categorical variables
V. Feature Selection for **Numeric Target and Binary Target**: description of the process and conclusions.
VI. Principal Component Analysis.
VII. Multiple Correspondence Analysis

VIII. Clustering: population segmentation.
IX. Description of Model Building process for prediction of numeric response (**Numeric Target**).
   a. Statistical summary of considered variables.
   b. Best Model Selection. Relationship to Feature Selection results.
   c. Model building: goodness of fit, interpretation of estimators in the final model
   d. Model Validation: outliers and influent data.
X. Feature Selection for **Binary Target**: description of the process and conclusions
XI. Description of Model Building process for prediction of binary response
   a. Statistical summary of considered variables.
   b. Best Model Selection. Relationship to Feature Selection results.
   c. Model building: goodness of fit, interpretation of estimators in the final model
   d. Model Validation: outliers and influent data.
   e. Sample Split into Work and Test samples. Predictive Capacity for each sample.
XII. Conclusions

Description of the Process

Parts II to V

Univariate Descriptive Analysis (to be included for each variable**):**

- Original numeric variables corresponding to qualitative concepts have to be converted to factors.
- Original numeric variables corresponding to real quantitative concepts are kept as numeric but additional factors should also be created as a discretization of each numeric variable.
- Exploratory Data Analysis for each variables (numeric summary and graphic support).

Data Quality Report:

Per variable, count:

- Number of missing values
- Number of errors (including inconsistencies)
- Number of outliers
- Rank variables according the sum of missing values (and errors).

Per individuals, count:

- number of missing values
- number of errors,
- number of outliers
- Identify individuals considered as multivariant outliers.

Create variable adding the total number missing values, outliers and errors.

Describe these variables, to which other variables exist higher associations.

- Compute the correlation with all other variables. Rank these variables according the correlation
- Compute for every group of individuals (group of age, size of town, singles, married, …) the mean of missing/outliers/errors values. Rank the groups according the computed mean.

Imputation:

- Numeric Variables
- Factors

Profiling:

- Numeric Target (Price)
- Factor (Audi)

R Markdown script should be included. A report (pdf file) has to describe decisions, procedures, criteria, etc no longer than 40 pages. Additional results can be included in appendixes.

Parts VI to VIII

PCA analysis for your data should contain:

- Eigenvalues and dominant axes analysis. How many axes we have to interpret according to Kayser and Elbow's rule?
- Individuals point of view: Are they any individuals "too contributive"? To better understand the axes meaning use the extreme individuals. Detection of multivariant outliers and influent data.
- Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables
- Perform a PCA taking into account also supplementary variables the supplementary variables can be quantitative and/or categorical

K-Means Classification

- Description of clusters

Hierarchical Clustering

- Description of clusters

CA analysis for your data should contain your factor version of the numeric target (f.total_amount) and 2 factors:

I. Eigenvalues and dominant axes analysis. How many axes we have to consider
II. Are there any row categories that can be combined/avoided to explain transformed price target into f.price .

MCA analysis for your data should contain:

I. Eigenvalues and dominant axes analysis. How many axes we have to consider for next Hierarchical Classification stage?
II. Individuals point of view: Are they any individuals "too contributive"?Are there any groups?
III. Interpreting map of categories: average profile versus extreme profiles (rare categories)
IV. Interpreting the axes association to factor map.
V. Perform a MCA taking into account also supplementary variables (use all numeric variables) quantitative and/or categorical. How supplementary variables enhance the axis interpretation?

Hierarchical Clustering (from MCA)

I. Description of clusters
II. Parangons and class-specific individuals.
III. Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on f.price target.
IV. Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on Audi binary target.

Parts IX to XI for Statistical Modelling are detailed:

- **Price (Numeric Target)** – **y** is the numeric response variable. **This variable will be the target for linear model building (Statistical Modeling I and II).**

- **Outcome – y.bin:** A new variable that is going to be the binary target. Variable **'Audi' (Binary Target)** will be the response variable for Binary Regression Models included in Statistical Modeling Part II**I.**

Explicative Variables for modeling purposes are any of the initial variables (except target) and new variables developed during the analysis process.

Multivariant Analysis conducted in previous deliverables has to be used to select the initial model. Students have some degrees in freedom in model building, but the following conditions are requested:

- For **General Regression Models** (Statistical Modeling Part I and II) **Numeric Target as the response variable**.
    - At least two numerical variables have to be considered as explicative variables for initial steps in model building, called covariates. Non-linear models have to be checked for consistency.
    - Select the most significant factors found in Multivariant Data Analysis as initial model factors. Put some reasonable limits to initial model complexity.
    - **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
    - Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable).
- For **Binary/Logistic Regression Models** (Statistical Modeling Part III) target analysis.
    - Split the sample in work and test samples (consisting on a 70-30 split). Working data frame has to be used for model building purposes.
    - At least two numerical variables have to be considered as explicative variables for initial steps in model building.
    - Select the most significant factors according to feature selection as initial model factors. Put some reasonable limits to initial model complexity.
    - **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
    - Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable).
    - You have to predict ***Y.bin (Binary Target)*** in the Working Data Frame vs the rest according to the best validated model that you can find and make a confusion matrix.
    - Make a confusion matrix in the Testing Data Frame for ***Y.bin (Binary Target)*** according to the best validated model found.

Confusion Matrix: When referring to the performance of a classification model, we are interested in the model's ability to correctly predict or separate the classes. When looking at the errors made by a classification model, the confusion matrix gives the full picture. Consider e.g. a three class

problem with the classes A, and B. The confusion matrix shows how the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data. The columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.

Data Description

100,000 UK Used Car Data set

This data dictionary describes data (https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes) - A sample of 5000 trips has been randomly selected from Mercedes, BMW, Volkwagen and Audi manufacturers. So, firstly you have to combine used car from the 4 manufacturers into 1 dataframe.

The cars with engine size 0 are in fact electric cars, nevertheless Mercedes C class, and other given cars are not electric cars,so data imputation is requered.

manufacturer	Factor: Audi, BMW, Mercedes or Volkswagen
model	Car model
year	registration year
price	price in £
transmission	type of gearbox
mileage	distance used
fuelType	engine fuel
tax	road tax
mpg	Consumption in miles per gallon
engineSize	size in litres