NEW YORK CABS DATA: 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.

Deliverable is concerned with Multivariant Data Analysis and model building for response variables: Y- Total Amount (Numeric Target) and binary factor Y.bin- 'TipIsGiven' (Binary Target) for trips in the green Taxi Trip Records.

The aim of the practice is to determine a decision rule for providing tips in green cab trips and to divide the sample units into groups: green (automatic yes) and red (automatic no).

A random sample containing 5000 registers of green taxi records has to be generated by each group and kept for all the practice.

Conclusions to *exploratory multidimensional statistics* are useful to enlighten some aspects of possible behavior and "predictors" for the total amount ($), or the binary discrete 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. Predictiive 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 (Preu/Price)
- Factor (Final Decision, TipisGiven)

R Markdown script should be included. A report (pdf file) has to describe decisions, procedures, criteria, etc.

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 for your binary target (column) and 3 factors:

- Eigenvalues and dominant axes analysis. How many axes we have to consider
- Are there any row categories that can be combined/avoided to explain the binary target.

MCA analysis for your data should contain:

- Eigenvalues and dominant axes analysis. How many axes we have to consider for next Hierarchical Classification stage?
- Individuals point of view: Are they any individuals "too contributive"? Are there any groups?
- Interpreting map of categories: average profile versus extreme profiles (rare categories)
- Interpreting the axes association to factor map.
- 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)

- Description of clusters
- Parangons and class-specific individuals.
- Comparison of clusters obtained after K-Means (based on PCA) and Hierarchical Clustering (based on PCA) focusing on Acceptance/Rejection binary target.

Parts IX to XI for Statistical Modelling are detailed:

- **Total Amoubnt (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 binary response. Variable **'TipIsGiven' (Binary Target)** will be the response variable for Binary Regression Models included in Statistical Modeling Part II**. TipisGiven has to be defined as positive when Tip >0.**

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

SHL Taxi Trip Records in New York

This data dictionary describes SHL trip data (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) - A sample of 5000 trips has been randomly selected from green taxi trip records.

VendorID	A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc
lpep_pickup_datetime	The date and time when the meter was engaged.
lpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value. Trip_distance
Pickup_longitude	Longitude where the meter was engaged.
Pickup_latitude	Latitude where the meter was engaged.
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server: Y= store and forward trip N= not a store and forward trip
Dropoff_longitude	Longitude where the meter was timed off.
Dropoff_ latitude	Latitude where the meter was timed off.
Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use. Improvement_surcharge
Tip_amount	Tip amount - This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.
Trip_type	A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver. 1= Street-hail 2= Dispatch