Question 1 (10 Points)

List as many use cases for the dataset as possible.

1. Cleansing the data
2. Running the linear regression model since independent variables follow linear trend considering variable “Price“ as dependent variable.
3. Apply cluster analysis and observe the pattern and characteristcs od each cluster and behaviour of the customer how does the customzing the car to buy. Segment the characteristcs of the vehicles associated by risk or by price
4. Apply decision tree model like CART or CHAID identify the statistical model for optimum accuracy and getting predictive value of a catergorical variable and cut the points for continous variables results to maximum accuracy. However, regression analysis gives dependent variable in numerical variables and hierarchical discriminant analysis gives dependent variable in class variable

Question 2 (10 Points)

1. Pick one of the use cases you listed in question 1 and describe how building a statistical model based on the dataset could best be used to improve Auto1’s business.

Firstly primary steps need to be consider before analysis

1. Loading and handling the data
2. Reading data, dimensions, columns i.e variables involed
3. Idtentfying the data types for analysis
4. Duplicates and removing
5. Missing values imputaion of missing values and replacing if required
6. Descriptive statistics
7. Maps & Figures : Boxplot – outliers
8. Scatter plot
9. Histogram
10. Corrleation
11. Spliting the dataset into train dataset and test dataset

Consider the model which is linear regression for identifying the impact of each independent variable impact on the depedent variable significance.

Consider the classification problem, allowing for the supervised learning technique. Spliting the dataset into train dataset and test dataset 80% and 20% respectively. Set-up the test harness to use cross validation. Build five different models to predictive the price and behaviour of the customer at the time of buying customized car. Select the best model. Create the validation dataset, used the 10-fold cross validation for accuracy, obtain the best model from the mix of linear (Linear Discrimanant Analysis, Logistic regression) and non-linear (NB, SVM, KNN and CART) algorthims. We observe the largest estimated accuracy score. Compare the algorthims. Observe the Whisker and boxplot at the top range and store at 100% accuracy

CART model algorithm is the most accurate after testing. Condiser the final check of the accuracy of the best model to obtain the behaviour of the customer for the customized car buying. Also summarize the results as a score, confusion matrix, and classification report.

Question 3 (20 Points)

Implement the model you described in question 2 in R or Python. The code has to retrieve the data, train and test a statistical model, and report relevant performance criteria.

When submitting the challenge, send us the link for a Git repository containing the code for analysis and the potential pre-processing steps you needed to apply to the dataset. You can use your own account at github.com or create a new one specifically for this challenge if you feel more comfortable.

Ideally, we should be able to replicate your analysis from your submitted source-code, so please explicit the versions of the tools and packages you are using (R, Python, etc).

Question 4 (60 Points)

A. Explain each and every of your design choices (e.g., preprocessing, model selection, hyper parameters, evaluation criteria). Compare and contrast your choices with alternative methodologies.

B. Describe how you would improve the model in Question 3 if you had more time.

Worked on Python IDE : Anoconda - Jupyter Notebook

Step by Step procedure

Install required and significant libraries and useful packages for Machine Learning alogorthms in Python

Load the dataset

Load the dataset naming “Auto1“ and understanding the variables, datatypes, statistics of the dataset. There are 25 variables observed which parameter leads to impact on the price of the car.

Handling the Missing Values :

Cleaning procedure is performed on the dataset, observed few records with “?“ values converted them into interger. Replaced ? with mean value stragegy. Also observed few string datatype varaible