Problem Statement

In EDA case study, you studied the behavior of consumers who had defaulted and who had not with respect to certain variables.

Now suppose you are working in the risk management team at a german bank and you have to assist the credit manager to decide whether to approve the loan for a prospective customer or not. So you have to make a logistic regression model to predict the chances of the customer defaulting on the loan or not. The amount the customer will default is not to be predicted. You have to just predict whether the customer will default or not.

Downloads:

You can download the data set and data dictionary from the link below.

German Credit Data
*file_download*Download

Data Dictionary
*file_download*Download

How to start the assignment?

As you have learnt in the logistic regression lectures, there is a 5-step process, you ideally need to follow while solving any regression problem. The chart below will guide you through the process of solving the assignment.

Approach To So

1. Business Understanding

2. Data Preparation

3. Model Development

4. Model Evaluation and Testing

5. Model Acceptance or Rejection

Figure 1: Logistic Regression Flow Chart

Note:

Please make sure the below points are to be followed strictly for evaluation purpose:

- Store the dataset into "*german_credit*" object.
- Divide the dataset into **70:30** ratio and set seed to 100 for the reference. It should be renamed as "train" and "test" respectively.

Packages Required:

You have to install the below packages below starting this assignment.

- install.packages("car") for VIF
- install.packages("Hmisc")
- install.packages("ROCR")
- install.packages("Caret")

Checkpoints

Checkpoint 1: Data Understanding and Data Exploration

Since you already know the business context of the problem, this is the first stage of solving the problem. In this checkpoint, you are supposed to get well versed with the dataset. For example, you should know the aim of solving this assignment and should also be attentive to explore your data completely.

Once the data is loaded into the R working environment, do basic exploration of data set which includes:

- Summary and structure of the dataset
- Univariate plots of any 5 variables. Present your insight from this univariate analysis in a word document

NOTE:- Keep writing every step of commented R codes properly in the .R file (provided in the next segment) which you have to submit at the end of this assignment for the evaluation purpose.

Checkpoint 2: Data Cleaning and Transformation

- In this step, you first need to identify the missing values and outliers for each variable and impute them accordingly.
- Generate dummy variables for factor variables. (Certain variables which should be ideally of factor type might be of character or integer type. you need to first convert such variables into factor variables using “as.factor()”)

NOTE:- Keep writing every step of commented R codes properly in the .R file (provided in the next segment) which you have to submit at the end of this assignment for the evaluation purpose.

Checkpoint 3: Splitting the Dataset into train and test

As you know to learn your model you should use only a portion of the data and keep the rest aside for testing the model. Split the given data set such into the train (70%) and test data set (30%). (Hint: Use seed of 100 while splitting.)

Checkpoint 4:Modeling

In this step, you will be actually building the logistic regression model on the data set.

- Make your initial model including all the variables and then select variables using step-wise function in R
- The model from the step-wise procedure should be checked using for multicollinearity using VIF in R (use a VIF threshold of 3).
- Report the AIC value and Null deviance and Residual Deviance of the final model

(Although in the lectures we discussed the importance of subject matter knowledge for selection of variables, for this exercise selection of variables has to be done only on the basis of step-wise selection and VIF procedure).

Checkpoint 5: Model Evaluation

Once the model is built evaluate the model using C-statistic and KS-statistic for both train and test data. Based on the values of C-statistic and KS-statistic, determine whether your model has good accuracy or not.

Checkpoint 6: Threshold value

After model evaluation, determine the threshold value of probability using ROC curve. Once the optimal value of threshold value is determined, generate misclassification table for both train and test data and report the following:

- Sensitivity
- Specificity
- Overall Accuracy

NOTE:- Keep writing every step of commented R codes properly in the .R file (provided in the next segment) which you have to submit at the end of this assignment for the evaluation purpose.

Submission

This is a graded assignment. The deadline for submission of this case study is **October 09, 2016, 11:59 PM**. For submissions obtained within 1 week of the deadline, there will be a 30% penalty. Submissions beyond 1 week of the deadline will not be accepted.

Important Note: All your code has to be submitted in one main .R file. Please make sure to rename your R script with "**Roll_Number_main.R**".

Also, you have to prepare a document summarizing the results of your model. This should briefly describe the important results and recommendations. Name the document "Roll_Number_main.docx".

Document template (Mandatory submission): Download the sample .docx file below. Also, download the commented R file which you have to use to write your codes and submit for evaluation.

Sample

*file_download*Download

Commented .R file

*file_download*Download

In total, you have to upload the two files as **one zip file named "Roll_no.zip".**The zip file should contain the **main R file, one .docx file.**

Submission

Default Prediction Model

*file_upload*Upload File

Marks Distribution:

Correctness of code: 50%

Comments to explain the code: 25%

Document: 25%