MIE1626 Data Science Methods and Quantitative Analysis
Project 1: Home Credit Default Risk [90 Marks]

Deadline: February 11 at 21:00

Academic Integrity

This project is individual: It is to be completed on your own. Do not share your code with others, or post any parts of your work online. You can only submit code that you have written yourself. If you use any online resource for developing parts of your code, acknowledge the source in a comment in your code. Students suspected of plagiarism on a project will be referred to the university for formal discipline according to the the Code of Behaviour on Academic Matters.

Introduction

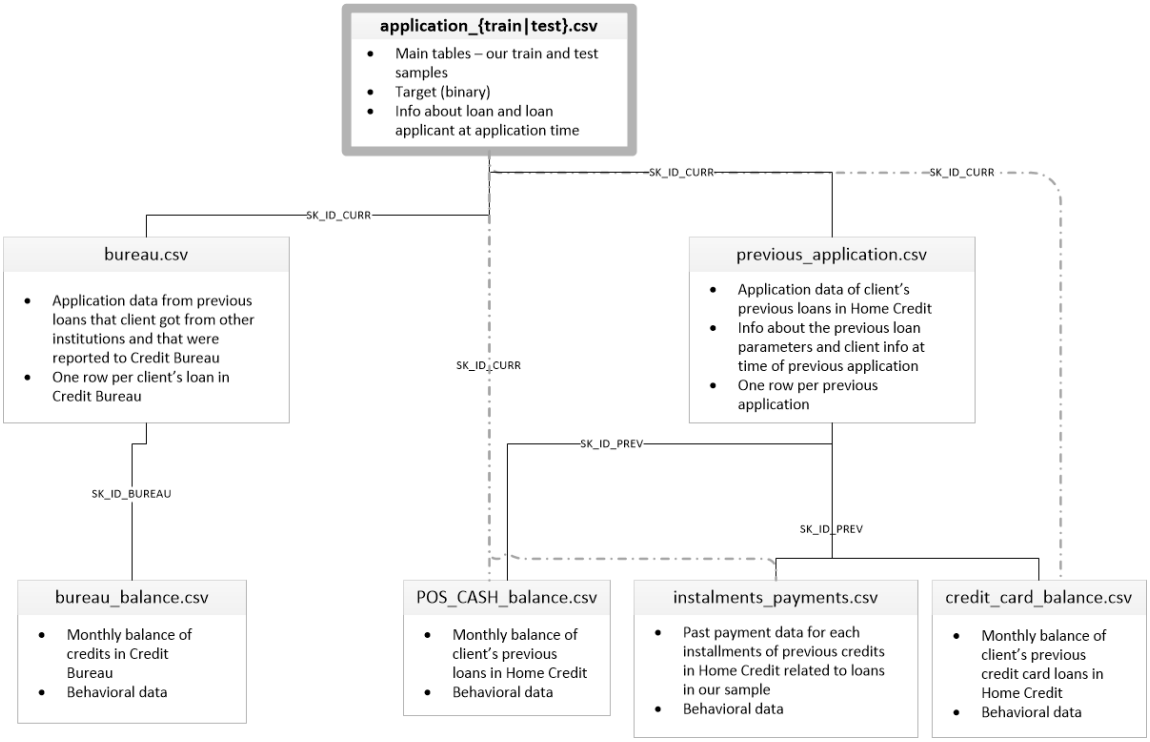
This project is about developing credit default risk models using logistic regression and an alternative method. It covers materials in Chapter 2 to 5 of the textbook "An Introduction to Statistical Learning" as well as additional data science and machine learning skills. The main purpose of a credit default risk model is to predict the repayment abilities of credit applicants including those with insufficient or non-existent credit histories. With such a predictive model in place, clients who are capable of repayment would not be rejected and loans will be given to appropriate candidates based on their predicted repayment ability.

The data come from a past Kaggle competition entitled "Home Credit Default Risk" which is accessible at **https://www.kaggle.com/c/home-credit-default-risk/data** (you need to agree to the competition rules). To complete this project, you should submit a Jupyter Python notebook on Quercus with code, saved output, text in markdown cells, and descriptive comments for different parts of the project. Use your UTorID as the filename.

Include your name, UTorID, and student number at the beginning of your Jupyter file. You should provide complete and correct code, text, code output to complete each part and get its mark. The code outputs (including answers to questions) should also be readable in your Jupyter notebook without the marker running the code. Enumerate different parts of your Jupyter notebook with numbers matching the sections and subsections of this handout.

1 Getting started [20 Marks]

The dataset is comprised of several files, which follow the schematic logic shown in the graph below. The csv file `HomeCredit_columns_description.csv` contains descriptions for all the variables from each data file.



Load the data

Load the following six csv files as pandas dataframes using a string 'path' for the location of files on your system (to be then updated by the marker for evaluation):
`application_train.csv`,
`bureau.csv`,
`bureau_balance.csv`,
`POS_CASH_balance.csv`,
`credit_card_balance.csv`,
`installments_payments.csv`

Note that the main dataset is `application_train` and the unique identifier for a sample is `SK_ID_CURR`. The other datasets contain additional information which you should extract and combine with the main dataset following the detailed instructions provided next.

1.1 Data transformation and cleaning [15 Marks]

To conduct data cleaning, the first step is create and select relevant features. This includes the following sub-steps:
1. Remove/disregard features that have a constant value.
2. Remove/disregard features that are irrelevant to credit risk or can't be explained. This includes time variables that are only related to the application and other variables that are not related to the response.
3. Handle missing values.
4. Calculate the proportion of each outcome in dichotomous value to transform categorical variables to continuous variables (explained below).
5. Create new variables.

For each dataset, perform the following steps before combining the datasets:

• `bureau_balance.csv`

For each `SK_ID_BUREAU`, calculate the proportion of `0` in the variable `STATUS`. It is an indication of 'good' status of credit bureau loan. Drop `MONTHS_BALANCE` as it's irrelevant to the clients' behaviour.

Then, merge `bureau_balance` to `bureau` using the key `SK_ID_BUREAU`.

• `bureau.csv`

1. For each `SK_ID_CURR`, calculate the proportion of `Closed` and `Active` of `CREDIT_ACTIVE`.
2. Remove `CREDIT_CURRENCY` as it's constant.
3. Remove `DAYS_CREDIT`, `CREDIT_DAY_OVERDUE`, `DAYS_CREDIT_ENDDATE`, `DAYS_ENDDATE_FACT` and `DAYS_CREDIT_UPDATE` as these are irrelevant variables.
4. For each `SK_ID_CURR`, calculate the proportion of `Consumer credit` and `Credit card` of `CREDIT_TYPE`.

Note that there might be some duplicated `SK_ID_CURR` in some of these datasets and several values may exist for a relevant feature (that we want to use) per `SK_ID_CURR`. Under such circumstances, calculate the average for the numerical variable per `SK_ID_CURR` and use the calculated average to present the corresponding feature value for that `SK_ID_CURR`. For example, the variable that you created based on the proportion of `0` in the variable `STATUS` is to be averaged for each unique `SK_ID_CURR` before being merged with `application_train`.

Then, merged all transformed variables from `bureau` which also include the information from `bureau_balance` to `application_train` using the key `SK_ID_CURR`.

In the next steps, prepare some relevant features (further explained below) from `installment_payment.csv`, `credit_card_balance.csv` and `POS_CASH_balance.csv` to be merged to the final dataset:

• `installment_payment.csv`

Combine `AMT_INSTALMENT` and `AMT_PAYMENT` into one variable indicating whether the client has paid out the installment on time. One new dummy variable should be created based on the follow-

ing criterion: if `AMT_PAYMENT` $\geq$ `AMT_INSTALMENT`, then the dummy equals 1, otherwise it equals 0. Then, calculate the maximum of that dummy variable for each `SK_ID_CURR`.

• `credit_card_balance.csv`

For the variables `AMT_DRAWINGS_ATM_CURRENT`, `AMT_DRAWINGS_CURRENT`, `AMT_DRAWINGS_OTHER_CURRENT`, `AMT_DRAWINGS_POS_CURRENT`, `AMT_RECIVABLE`, and `AMT_TOTAL_RECEIVABLE` follow this procedure: replace all `NA` with `0` and then compute average for each `SK_ID_CURR`.

• `POS_CASH_balance.csv`

Pick variable `SK_DPD`, replace `NA` with `0` if any, and make it into a dummy variable based on the following criterion: If the original value is 0 then keep it as 0, else replace original value with 1. Then, take the maximum of the newly created dummy variable for each `SK_ID_CURR`.

Replace `NA` with `0` for `CNT_INSTALMENT_FUTURE`, and take its average for each `SK_ID_CURR`.

Then, merge the new variables picked or transformed from `POS_CASH_balance`, `credit_card_balance` and `installment_payment` to `application_train` using the key `SK_ID_CURR`.
Report the shape of the combined dataset that you have obtained.
Provide a data dictionary (i.e. an organized list of variables you have added to the main dataset and their descriptions) before performing any statistical modelling. This can be provided as text in your Jupyter notebook.

1.2 Visualizing the response variable [5 Marks]

Determine the response variable in your notebook and explain in plain English what it represents. Provide a Seaborn plot for the response variable and one arbitrary pair of the features that you have created in the previous parts. Use appropriate figure size, title/caption, legend, and axis titles.

2 Model Construction using Logistic Regression [10 Marks]

Split the combined dataset into a training (70%) and testing set (30%). Make sure to separate out the column corresponding to the target.

Logistic Regression: Suppose there are n explanatory variables: $x_1, x_2, x_3, \cdots, x_n$, and one binary response variable Y. Assuming it represents the client who has payment difficulties ($Y = 1$ means that the client has payment difficulties, and $Y = 0$ refers that the client is of all other cases). The model assumes that there exists a linear relationship between the logit (log-odds) of the event that $Y = 1$ and the explanatory variables $x_1, x_2, x_3, \cdots, x_n$. Note that log(odds)= $\log \frac{p}{1-p}$, where p is the probability of the event $Y = 1$. The mathematical form of logistic regression can be explained as follows:

$$log(odds) = log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where $log(odds)$ is the log-odds, $log(\cdot)$ is the natural logarithm, β_i for $i \in \{1, 2, \cdots, n\}$ are the parameters of the logistic model.

After completing the preliminary steps above, the variables listed below are to be selected and used for step-wise logistic regression. This is according to the advice you have received from a domain expert who has looked into WOE (Weight of Evidence) values and analyzed the colleanirity between features (using variable clustering).

Explanatory Variable		
WOE_ENTRANCES_MEDI	WOE_AMT_GOODS_PRICE	WOE_CODE_GENDER
WOE_NONLIVINGAREA_MODE	WOE_AMT_CREDIT	WOE_DAYS_REGISTRATION
WOE_AMT_CREDIT_MAX_OVERDUE	WOE_REG_REGION_NOT_WORK_REGION	WOE_NAME_FAMILY_STATUS
WOE_AMT_REQ_CREDIT_BUREAU_WEEK	WOE_REG_CITY_NOT_WORK_CITY	WOE_NAME_HOUSING_TYPE
WOE_AMT_DRAWINGS_CURRENT	WOE_FLOORSMAX_AVG	WOE_OWN_CAR_AGE
WOE_AMT_TOTAL_RECEIVABLE	WOE_ELEVATORS_AVG	WOE_AMT_INCOME_TOTAL
WOE_LIVINGAPARTMENTS_MEDI	WOE_CNT_INSTALMENT_FUTURE	WOE_FLAG_WORK_PHONE
WOE_COMMONAREA_MODE	WOE_AMT_PAYMENT_GREATER_EQUAL_INSTALMENT	WOE_FLAG_OWN_REALTY
WOE_DAYS_EMPLOYED	WOE_EXT_SOURCE_3	WOE_NAME_CONTRACT_TYPE
WOE_FLAG_EMP_PHONE	WOE_CREDIT_TYPE_CREDIT_CARD	WOE_CNT_FAM_MEMBERS
WOE_DEF_30_CNT_SOCIAL_CIRCLE	WOE_FLOORSMIN_AVG	WOE_NAME_TYPE_SUITE
WOE_OBS_30_CNT_SOCIAL_CIRCLE	WOE_FLOORSMIN_MEDI	WOE_WEEKDAY_APPR_PROCESS_START
WOE_EXT_SOURCE_2	WOE_EXT_SOURCE_1	WOE_DAYS_LAST_PHONE_CHANGE
WOE_REGION_RATING_CLIENT	WOE_OCCUPATION_TYPE	

2.1 Constructing the model [5 Marks]

Fit a logistic regression model using the selected explanatory variables and the training data.

2.2 Interpreting the model [5 Marks]

Interpret the estimated coefficients of the logistic regression model and explain their impact on the response variable in plain English.

3 Model Evaluation [25 Marks]

3.1 Confusion matrix [5 Marks]

Calculate the confusion matrix using the testing set.

3.2 Explaining the confusion matrix [5 Marks]

Explain a weakness or a strength of the model based on the confusion matrix.

3.3 Graphic illustration of the performance [5 Marks]

Plot the ROC curve (receiver operating characteristic curve) showing the performance of the model. Calculate the AUC (area under the curve) respectively.

3.4 Compute the performance measures [5 Marks]

Compute the following measures to assess the performance of the logistic regression model: Precision, recall, F_1-score, accuracy, and total misclassification rate; using the testing set.

3.5 Suitable performance measures for reporting [5 Marks]

If you were to only choose two of the five performance measures for reporting, which ones do you choose? Justify your choice.

4 Alternative classification model [20 Marks]

Select an alternative classification model that has at least one tunable hyper-parameter.

4.1 Train the alternative classification model [5 Marks]

Train your alternative model on the same training set.

4.2 Tune the alternative classification model [5 Marks]

Use a k-fold cross validation to tune the hyper-parameters of the alternative model based on accuracy.

4.3 Confusion matrix [5 Marks]

Calculate the confusion matrix using the testing set.

4.4 Graphic illustration of the performance [5 Marks]

Plot the ROC curve (receiver operating characteristic curve) showing the performance of the model. Calculate the AUC (area under the curve) respectively.

5 Compare the two models [15 Marks]

Use the test data to compare the logistic regression model and the tuned alternative model.

5.1 Compare predictive performance [5 Marks]

Choose a suitable plot and illustrate all the five performance measures (precision, recall, F_1-score, accuracy, and total misclassification rate) for the two models. Use appropriate figure size, title/-caption, legend, and axis titles.

5.2 Recommend model to the client (scenario 1) [5 Marks]

Among the two models you have developed, what is you suggested model to Home Credit Group (client of the project), if they are more concerned about rejecting applications of individuals who may otherwise end up having difficulty in repayment? Provide a numeric measure for each of the two models to justify your response.

5.3 Recommend model to the client (scenario 2) [5 Marks]

Among the two models you have developed, what is you suggested model to Home Credit Group (client of the project), if they are more concerned about not rejecting applications of individuals who may be capable of repayment according to the data? Provide a numeric measure for each of the two models to justify your response.