**MFIN7034 Problem Set 3 – Risk Analysis**

Version: 2025/02/25

Due Date: 2025/03/04 23:55:00 UTC+8

This problem set aims to provide some experience applying machine learning methods in risk analysis. The dataset “credit\_risk.csv” is available to you on Moodle. Your main task is to establish machine learning models that predict the default label using available information (covariates).

A table of variable explanations is provided here:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Note** | **Explanations** |
| age | Age of borrower | Age in number of years |
| edu | Education level | 0: below high school, 1: high school, 2: college, 3: master, 4: above master |
| gender | Gender | 0: female, 1: male |
| housing | Housing ownership | 0: not own, 1: own |
| income | Income | Monthly income-level |
| job\_occupation | Job type | 0: unemployed/temporarily employed, 1: employed, 2: manager/senior worker |
| past\_bad\_credit | Historical default label | 0: non-default, 1:default |
| married | Marital status | 0: unmarried, 1: married |
| default\_label | Default indicator | 0: non-default, 1:default |

**Submission format: .ipynb notebook with runnable code and all the steps shown, and a PDF report.** The final report should contain results generated by your program. Simple, presentable, coherent English, clean graphs. Proper visualization and clear interpretations & discussions, such as explaining why a factor can predict default or what your logic is in pursuing higher AUC, will also be graded.

**1. Machine Learning Trials (60 Marks)**

The first part of this problem set contains three practical tasks for machine learning algorithm applications:

**1.1 Logistic Model (25 Marks)**

Run logistic regression: regress default label on available variables. Besides the original variables, also try to add more interaction term variables and/or non-linear transformation variables (polynomials, log transformations, dummy variables, etc.) to the model. Summarize your result. Obtain prediction values in the regression above. Compute and plot the ROC curve. Compute AUC value. Explain your main results. Also compare the AUC performance from different model specifications. Briefly discuss the outcomes.

**1.2 SVM/Random Forest (15 Marks)**

You might wonder whether non-linearity in model specifications can help. Try SVM or Random Forests method. You can select either one. Then, report the key parameters of your model, the AUC value, and the ROC plot as your main result.

**1.3 LightGBM (20 Marks)**

LightGBM has been one of the most popular gradient boosting algorithms since it was developed. This algorithm is very popular on Kaggle and also productive in the real-world production scenarios. Try LightGBM method. Describe the procedure in detail, such as data preprocessing, model specification, feature selection and hyper-parameter tuning. Report the AUC value and plot the ROC curve. Compare this model’s performance with outcomes in the previous two questions.

**2. Deeper Explorations (40 Marks)**

Think deeper, ask further, and explore more:

**2.1 Data Preprocessing (15 Marks)**

Introduce the detailed target for the step-by-step data preprocessing procedures towards Logistic model and LightGBM model respectively. Note that the prodecures should match with your code in Question 1.1 and 1.3. An example answer would be in the following format:

For Logitsit model:

…: …;

Standardization: In order to …

…: …

For LightGBM model: …

**2.2 Feature Importance Analysis (15 Marks)**

For each model you use in Question 1.1, 1.2 and 1.3, list one **model-dependent** method to provide feature importance measurements for the feature inputs. Also use the nominated method to output the feature importance ranking for the top 5 features. You will produce a table like (as an example):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1st | 2nd | 3rd | 4th | 5th |
| Logistic | age | edu | … | … | … |
| SVM/Random Forest (the one you used) | … | … | … | … | … |
| LightGBM | age | edu | … | … | … |

**2.3 Go Deeper towards Feature Importance Analysis! (10 Marks)**

Do you think there could be any method that can apply to all above four models (i.e., Logistic regression, SVM, random forest, LightGBM)? Please discuss your idea and thoughts. The mark of this question will be given very generously, so if your answer is yes, just give a try and show what you can get!