

Assignment

Thank you for taking the time to complete MUFG Credit Risk Analyst data assignment.

Please prepare a short presentation for the task outlined below using the provided data. You will have 10 minutes to talk through your points with the team. What you present should encourage discussion of the results, as well as the team’s challenges, as in a normal team meeting.

Modeling probability of default (PD) is one of the major tasks that MUFG wholesale stress testing team performs as part of MUFG’s overall capital adequacy and planning program. The wholesale stress testing team is working on building a PD model in order to understand risk factors that drive PDs, and assess the risk profile for each portfolio. As a part of the team, you are asked to build a PD model by exploring and analyzing the below data:

1. **Financial Data:** This data includes obligor specific financial information. LOAN\_ID distinguishes individual obligors and columns labeled X1 to X15 are features available for modeling. These features include both numeric and categorical values.

****

1. **Default Data:** This data includes information on defaulted loans. LOAN\_ID distinguishes individual obligors.



**OBJECTIVE**

*Construct a machine learning model that achieves the best F-1 Score in predicting whether an obligor will default.* Join the financial and default data by ‘LOAN\_ID’ to construct a variable ‘Default\_Ind’ which will serve as your target variable. When ‘Default\_Ind’ is 1, a customer is in default, and 0 otherwise. Treat the first 7,000 observations as your training data and preserve the remaining 3,000 observations for testing model performance. Your model’s F-1 Score will be judged ONLY on its performance on the testing data.

Here are some questions to consider in your analysis:

* What, if any, data treatment needs to be performed before model training? (i.e. feature engineering, imputation, standardization, etc.)
* Why did you select your specific machine learning algorithm and model architecture? Did you employ any hyperparameter tuning techniques? (For example, you used logistic regression because it best fit the target variable and you added lasso regularization to compensate for overfitting. A grid search was used to determine the optimal strength of regularization.)
* Did you define a validation hold-out data set, implement cross-validation, or use other techniques to assist your model decision-making process? How and why did you implement your chosen strategy? (For example, you utilized 5-fold cross-validation to better estimate out-of-sample performance for various hyperparameters for your algorithm.)
* Which feature(s) influence the model output the most and which have little influence? What techniques did you utilize to reach these conclusions? (i.e. You plotted partial dependence plots for the various features and concluded … or you extracted feature importance calculations from a tree-based model to determine …)

As a part of your take home exercise please follow the below steps:

1. Prepare 10 minute presentation for a cross-functional panel in PowerPoint
2. There is no slide limit or requirement
3. Please include code for your analysis and any supporting materials in PDF format.
4. We will assess your results based on technical soundness and depth, business insight, structure and organization.