FA 590 Statistical Learning in Finance

2024 Fall

Assignment 1 Loan Default and Return Prediction using Linear Models

In this assignment, you will explore a dataset containing over one million personal loans originated between 2008 and 2017 on LendingClub. The dataset includes detailed loan characteristics (e.g., interest rate, loan amount), borrower characteristics (e.g., FICO score, income), and loan outcomes (e.g., early default and return). Your task is to analyze the dataset, build linear models, and make predictions of early defaults and returns on the test set.

Files Provided: You can access all the data files via the following link: https://stevens0-my.sharepoint.com/:f:/g/personal/zyang99_stevens_edu/Ek4DFNqCmh5EoCwCV_SUAD4Bvuwu-mWQGiPP_qi8miWLfw?e=nbvNas

1. lc_loan.csv: The training data, containing loans originated between 2008 and 2016.

2. lc_loan_test.csv: The test data, containing loans originated in 2017. Loan outcomes have been removed.

3. submission_test.csv: A submission template for your test set predictions. It includes the following columns:

- id: Unique loan identifier (to match loans in lc_loan_test).
- early_default: Placeholder for your predictions (whether the loan defaulted within one year). The values presented in this column are randomly generated.
- return: Placeholder for your predictions for loan returns. The column is filled with 0.

4. Feature Dictionary.xlsx: Describes all features in the dataset.

Submission Guidelines:

- This assignment is **NOT** a group assignment.
- Submit a detailed report in .pdf, .ipynb, or .Rmd format that explains your methodology, analysis, and model evaluation. The report should disclose whether and how you have used generative AI tools, like ChatGPT, for this assignment. You are welcome to share good use cases of generative AI.
- Submit your final predictions in submission_test.csv.

- Deadlines: Please submit all files on Canvas by **3 PM, October 8, 2024**.

Computing Resources:

If your personal laptop does not have sufficient computing resources to handle the data size for this assignment, you have access to the computing resources at the Hanlon Financial Systems Center, available at two locations:

- **Hanlon Lab 1** (4th floor, Babbio): 30 workstations equipped with Intel i9-11900 CPUs, 64GB RAM, RTX 3070 GPUs (8GB GPU memory), 1TB SSD for Windows 11, and 2TB HDD for Ubuntu 20.04.
- **Hanlon Lab 2** (1st floor, Babbio): 33 workstations equipped with Intel i9-14900 CPUs, 32GB RAM, RTX 4000 Ada GPUs (20GB GPU memory), 1TB SSD for Windows 11, and 1TB SSH for Ubuntu 22.04.

The center is open Monday through Saturday. You can view detailed availability on their calendar [Link to Calendar]. Reservations are not required—simply log in with your Stevens account to use any available workstation.

Requirements:

1. Exploratory Data Analysis

Perform exploratory data analysis to understand the patterns and relationships within the dataset. Your analysis should include:

- Report the summary statistics, such as average, minimum, maximum, for key variables.
- Histograms, scatter plots, and box plots for key variables.
- Correlation plots to assess relationships between features.

2. Data Preprocessing

2.1 Sampling

You are required to further sample the training data into training and validation sets.

2.2 Handling Missing values

Missing data is present in the dataset. You may choose to either impute missing values or filter out incomplete observations. Justify your choice of approach (e.g., forward filling, backward filling, or mean).

2.3 Feature Engineering

2.3.1 Generate at least one interaction term and one nonlinear (e.g., quadratic) term. Provide justification for the chosen interaction/nonlinear terms and examine whether they are statistically significant in your model.

2.3.2 Scale the numerical features using either standardization or normalization techniques. Explain how it improves the dataset for machine learning models.

2.3.3 Apply one-hot encoding and label encoding to transform qualitative features into numerical features.

3. Model

3.1 Linear Regression Model

Fit a simple linear regression model to predict the loan return. Report the in-sample and out-of-sample R-squared.

3.2 Regularized Regression Model: Lasso Regression, Ridge Regression, and Elastic Nets

Fit Lasso regression, Ridge regression, and Elastic Nets to predict the loan return. Select the hyperparameters using the results in the validation set. Report the in-sample and out-of-sample R-squared.

3.3 Logistic Regression

Fit a logistic regression model to predict loans' early default. The logistic regression model will return a probability of early default, and it requires a threshold to classify the loans into early default or not (the default value for the threshold is 0.5). Select the threshold using the results in the validation set. Report the in-sample and out-of-sample model performance, including accuracy, F1 score, and AUC score.

4. Predicting on the Test Set

Use your best regression and classification models (including the hyperparameters) to predict the outcomes for the test set (lc_loan_test.csv). Replace the values in submission_test.csv for the columns:

- early_default: Your logistic regression predictions. Note that the submitted predictions should be 0's and 1's with 1 being early defaulted.
- return: Your regression predictions for loan returns.
- Make sure the predictions match the 'id' column to ensure correct evaluation.