MECON6102 Problem Set 2

Xing Mingjie

April 25, 2024

1 Data

1.1 Description

Table 1 shows the summary statistics of the data. The data set contains 13982 observations and 9 variables. The dependent variable is the default label, which is a binary variable indicating whether the individual defaults.

Figure 1 shows the heat map of the correlation matrix of the features and target. Most of the features are arguably uncorrelated. There is a high correlation between housing and age at 0.55. The correlation between income and education level is 0.51, which captures the wage premium of education.

1.2 Data Preprocessing

Figure 2 shows the distribution and the skewness of feature `income`. The distribution is right-skewed. The report uses the log transformation to reduce the skewness of the feature for better performance in models.

Figure 3 shows the distribution of the feature `income` after the log transformation. The distribution is more symmetric after the transformation and helps boost model performance in our exercise.

Table 1: Data Description

	count	mean	std	min	max
default_label	13982.00	0.02	0.15	0.00	1.00
age	13982.00	41.66	14.56	17.00	66.00
gender	13982.00	0.46	0.50	0.00	1.00
edu	13982.00	1.69	1.10	0.00	4.00
housing	13982.00	0.63	0.48	0.00	1.00
income	13982.00	7426.48	6226.68	650.42	37515.37
job_occupation	13982.00	0.34	0.56	0.00	2.00
past_bad_credit	13982.00	0.96	0.19	0.00	1.00
married	13982.00	0.53	0.50	0.00	1.00

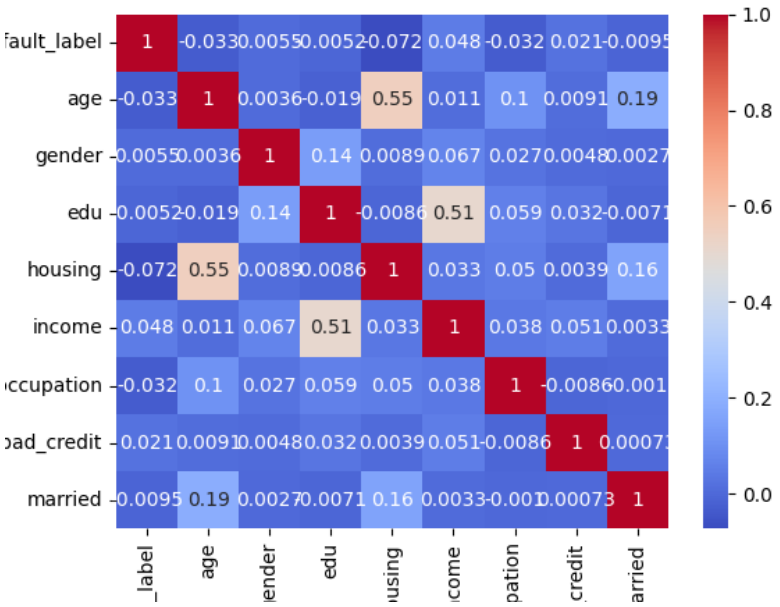


Figure 1: Heat map of the correlation matrix of the variables

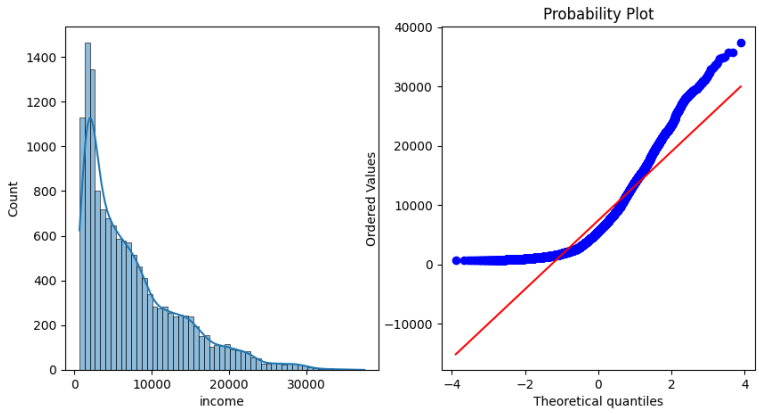


Figure 2: Income Distribution

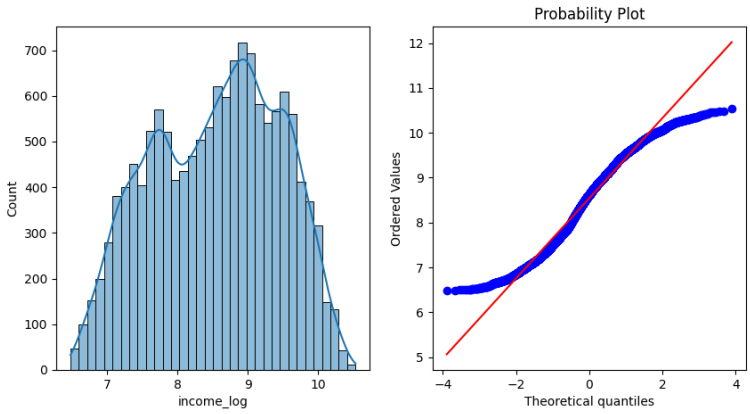
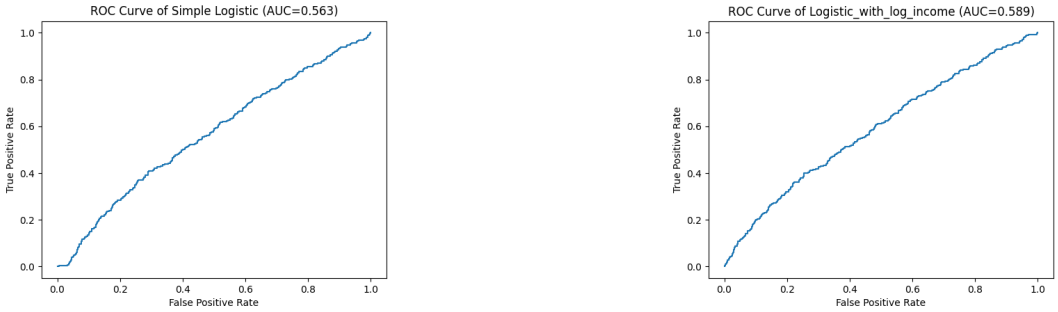


Figure 3: Income Distribution After Log Transformation



(a) Panel A

(b) Panel B

Figure 4: ROC Curve of Simple Logistic Model

2 Model Comparison

2.1 Simple Logistic Model

Panel A of figure 4 shows the ROC curve of the simple logistic regression model with feature `past_bad_credit` and `income` and the Panel B is the simple model with log-transformed `income`. The AUC is 0.56 for simple model, and 0.59 for simple logistic with log-transformed income. The predictive power is close to a random draw, but the log transformation add to the performance of model. In the rest of the exercise, we use the log-transformed income instead of the original income.

2.2 Full Logistic Model

In a full logistic model with log-transformed income in Figure 5, we have an AUC of 0.69, which is a significant improvement over the simple model. The full model includes all the features in the data set.

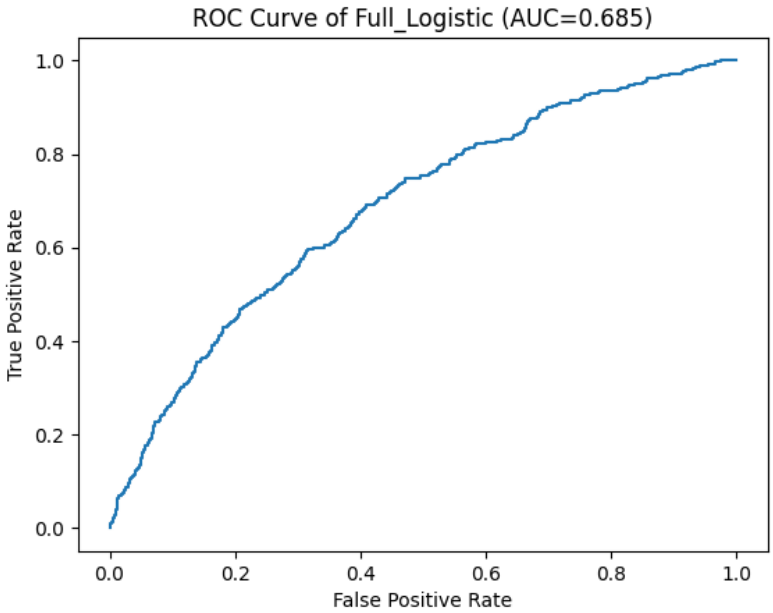


Figure 5: ROC Curve of Full Logistic Regression

2.3 non-linear models

This part reports the non-linear models. The SVM model with RBF as kernel has an AUC of 0.67, with sigmoid 0.45, with polynomial 0.64, as reported in Figure 6. The RBF kernel has the best performance among the three kernels. Sigmoid performs even worse than random classification. In the rest of the exercise, we use the RBF kernel for SVM.

In the literature, Bazzana et al. (2023) and Wu (2022) proposes Neural Network, XGBoosting and Random Forest as the best predictive algorithms for default risk, with AUC around 0.97.

2.4 Within-Sample Performance

Table 2 reports the with-in sample of each model as we discussed above. The two best performers are Random Forest and XGBoosting, with AUC of 1 and 0.98, showing an overfitting

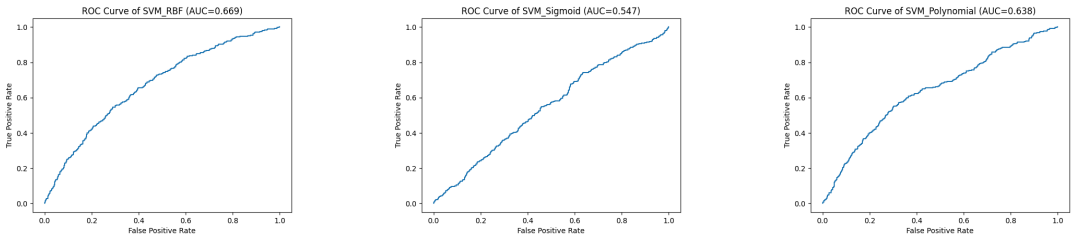


Figure 6: SVM with RBF, sigmoid, polynomial kernel

Table 2: AUC Comparison

Model	auc
Simple Logistic	0.563
Simple Logistic with log income	0.589
Full Logistic	0.685
SVM RBF	0.669
SVM Sigmoid	0.547
SVM Poly	0.638
Random Forest	1.000
XGBoosting	0.980
Neural Network	0.720

Table 3: AUC Comparison (Out of Sample)

Model	auc	auc_tuned
Logistic	0.686	0.687
SVM RBF	0.482	0.555
Random Forest	0.567	0.665
XGBoosting	0.567	0.650
Neural Network	0.601	0.669

problem. The rest part of the report focuses on hypertuning models to improve the out-of-sample performance.

2.5 Out-of-Sample Performance

Table 3 reports the baseline (default parameter suggested by the package) out-of-sample performance of each model in the `auc` column. The full logistic model has the best performance with an AUC of 0.686. The SVM model with RBF kernel has a lowest AUC of 0.518.

2.5.1 Hypertuning

The baseline, hypertuned parameters, and best parameters are reported in the jupyter notebook. We leverage `GridSearchCV` to hypertune the models to maximize out-of-sample roc-auc performance. We provide a dictionary of hyperparameters to each model and do cross validation of 5-fold. As we can see from the `auc_tuned` column of table 3, after hypertuning there is generally a 10% increase in the performance of each model. The best performer is still the Logistic model, though with only marginal improvement.

Bibliography

Bazzana, F., Bee, M., and Hussin Adam Khatir, A. A. (2023). Machine learning techniques for default prediction: an application to small italian companies. *Risk Management*, 26(1):1.

Wu, W. (2022). Machine learning approaches to predict loan default. *Intelligent Information Management*, 14:157–164.