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

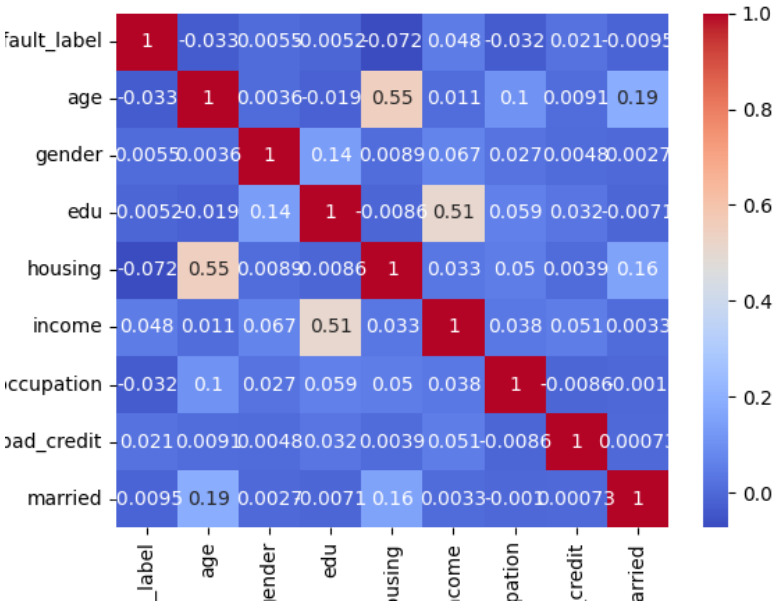

Figure 1: Heat map of the correlation matrix of the variables

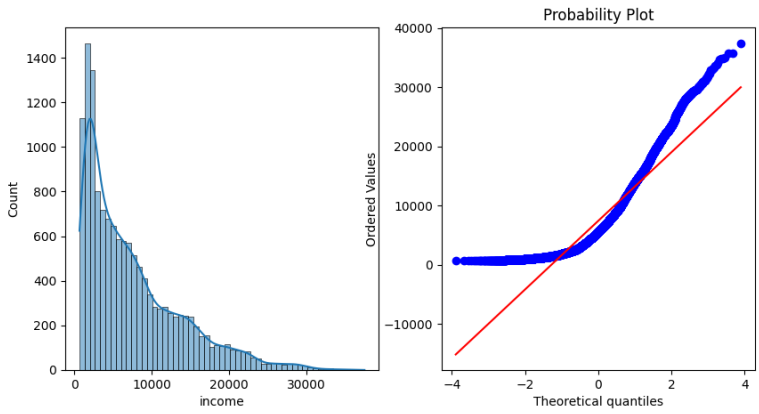

Figure 2: Income Distribution

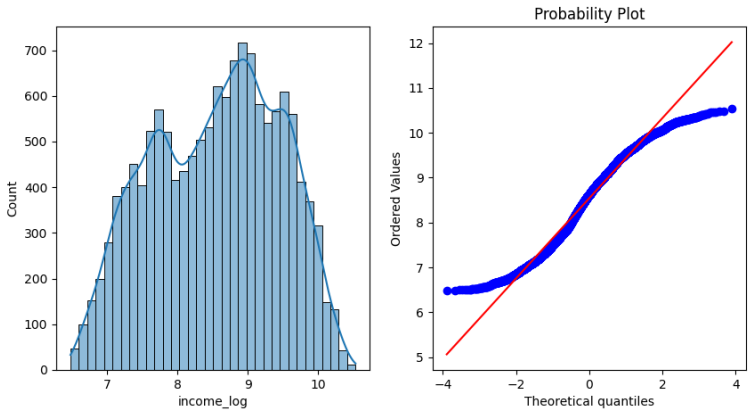

Figure 3: Income Distribution After Log Transformation

Table 2: AUC Comparison

Model	auc
Simple Logistic	0.56
Logistic with log income	0.59
Full Logistic	0.69
SVM RBF	0.67
SVM Sigmoid	0.45
SVM Poly	0.64
Random Forest	1.00
XGBoosting	0.98
Neural Network	0.71

2 Model Comparison

2.1 Simple Logistic Model

2.2 Within-Sample Performance

Table 2

2.3 Out-of-Sample Performance

Table 3

Table 3: AUC Comparison (Out of Sample)

Model	auc
Logistic	0.69
SVM RBF	0.52
Random Forest	0.58
XGBoosting	0.57
Neural Network	0.61

Bibliography