MAFS6010Z Project1: Warm-up of Statistical Machine Learning

Li Chenghai 20828446 Liu Shifei 20734530 Nie Jialei 20747874 Wu Jiajun 20666111

1 Introduction

We used Linear Discriminant Analysis method to analyze default probability of unbanked population. We used application_train and application_test datasets. And we selected 120 features from the application_train dataset.

2 Methodology

Linear Discriminant Analysis

binary response: $Y = \begin{cases} 0, & \text{not default} \\ 1, & \text{default} \end{cases}$

Probability of not default:

$$p_0(x) = \frac{\pi_0 \exp[(-1/2)(x-\mu_0)^T \sum -1 (x-\mu_0)]}{\sum_{i=0}^{1} \pi_i \exp[(-1/2)(x-\mu_i)^T \sum -1 (x-\mu_i)]}$$

Probability of default:

$$p_1(x) = \frac{\pi_1 \exp[(-1/2)(x-\mu_1)^T \sum -1 (x-\mu_1)]}{\sum_{i=0}^{1} \pi_i \exp[(-1/2)(x-\mu_i)^T \sum -1 (x-\mu_i)]}$$

3 Prediction - confusion matrix

sensitivity=TP/P=13593/(11232+13593)=0.55

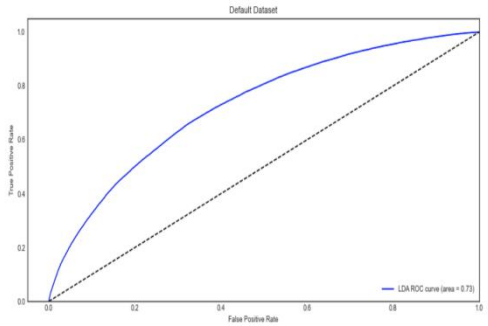
specificity=TN/N=216154/(216154+66532)=0.76

accuracy rate= （216154+13593）/307511=0.75

		True default status	
		no	yes
Predicted status	no	216154	11232
	yes	66532	13593

4 Prediction - ROC curve

area under ROC = 0.73



5 Conclusion

We set decision_prob equal to 0.1 and by comparing the predicted results with the true results, we got the confusion matrix. Among all customers who actually default, we predicted 55% of them to be default. The area under ROC curve is 0.73, which means the model performs well with different thresholds.

6 Contribution

Li Chenghai	code
Liu Shifei	code
Nie Jialei	poster
Wu Jiajun	poster

7 Kaggle team and score

math6010z_Li_Liu_Nie_Wu

0.68