Taiwan Credit Defaults

Group 14: Atindra Bandi, Henry Chang, Avani Sharma, Abraham Khan

Introduction

The Taiwanese economy experienced tremendous growth during the 1990's, almost doubling in value along with the other countries known as the "Asian Tigers". The country's financial sector was heavily involved in the growth of real estate during this period. However, in the early 2000's, this growth slowed and banks in Taiwan turned towards consumer lending to continue the expansion. As a result, credit requirements were loosened and consumers were encouraged to spend by borrowing capital.

We will be analyzing data on Taiwanese credit card holders from mid-2005, as the flood of debt was reaching its peak. The following algorithms will be tested for accuracy in predicting if an individual will miss their next payment (predictor variable).

- Logistic Regression
- Random forest
- Naïve Bayes

The banking data is skewed with only 22.1% defaulters which could potentially lead to biased predictions. Thus, we compared the models in terms of their specificity and sensitivity - the proportion of correct positive predictions (Non-defaulters) and proportion of correct negative predictions (Defaulters). Each method was compared on the basis of **R**eceiver **O**perating **C**haracteristic curve.

Exploratory Data Analysis

The data contains five categorical variables: sex, education level, marital status for the prior six months, and whether or not the person missed their next month payment. There are also four numerical variables: the person's credit limit, age, bill and payment amounts for the last six months.

Demographical measures:

University	Graduate	High School	Others
47%	35%	16%	2%

	Married	Single	Others
	53%	45%	2%
Defaulters	23%	21%	24%
Non - Defaulters	77%	79%	76%

	Male	Female
	40%	60%
Defaulters	24%	21%
Non - Defaulters	76%	79%

Feature Engineering

We created variables spend per month, weighted spending; pay spend ratio and payment month status as the demographics did not show a strong indication of defaulting.

1. Spend = Bill amount this month – Bill amount last month +Payment this month
2. Pay Spend Ratio = ∑ Payment / ∑ Spending
3. Mean Spend ratio = Weighted Spending /Limit Balance

4. Weighted Payment Delay = 0.4* PAY_0+0.25* PAY_2+0.15* PAY_3+0.1* PAY_4+0.05* PAY_AMT5+0.05* PAY_6
5. Grouped ages
6. Interactions terms SEX*MARRIAGE, MARRIAGE*AGE, EDUCATION*AGE

Analysis

We performed a lasso analysis on logistic regression with all variables for both variable selection and regularization. From the cross validation we got the most optimum lambda along with variables that significantly impacted our dependent variable. Following are the most significant variables from the analysis:

1. The dummy variables derived from education, marriage & payment status come significant and hence were incorporated.
2. The most recent pay status, payment and bill amount also came out as significant variables.
3. From the feature engineering the two most significant variables were weighted spend and mean spend

We moved to techniques like Naives Bayes and Random Forest to improve our results further keeping the variables selected from Lasso. The optimum number of trees from the random forest plot is 700. We selected random forest as it had a higher AUC in ROC curve which depicts that it could reach a higher sensitivity with low false positive rate. Although we could reach a very high accuracy of 94% but there was a trade-off between both sensitivity (96%) and specificity (37%). For our problem statement we selected a lower accuracy of 80% with a sensitivity of 88% and specificity of 45%. Since the bank has to take remedial measures to lower the credit debt it has to identify the probable defaulters correctly (specificity) and at the same time keep their market hold by not penalizing the non-defaulters (sensitivity). Our sensitivity, specificity and accuracy are selected in the same spirit.

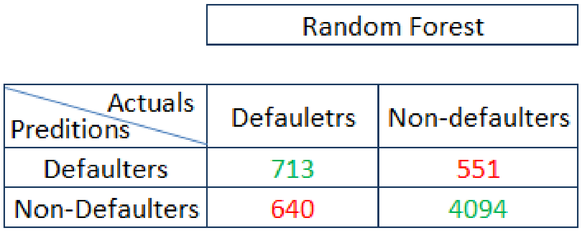
Conclusion

1. Random forest has the best balance between TPR and FPR. The model is accurate as well as can predict the number of defaulters the best among all algorithms.
2. Recent payments, billed amounts and payment status are the most important variables for prediction
3. Demographic variables are not important predictors for defaulting as could be actually seen from our data exploration exercise.
4. Feature engineering led us to many interesting variables of which few actually turned out to be significant in our final model equation.
5. We could predict our test set with 80% accuracy and defaulter rate prediction of 53% and non-defaulter rate prediction accuracy of

Dataset:

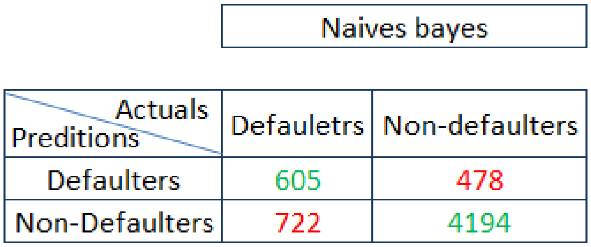
http://archive.ics.uci.edu/ml/machine-learning-databases/00350/

Cross Validation Results:



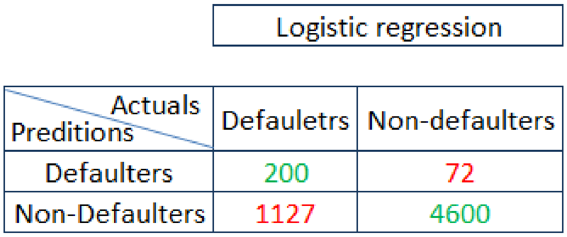
Random Forest

Actuals Preditions	Defauletrs	Non-defaulters
Defaulters	713	551
Non-Defaulters	640	4094



Naives bayes

Actuals Preditions	Defauletrs	Non-defaulters
Defaulters	605	478
Non-Defaulters	722	4194



Logistic regression

Actuals Preditions	Defauletrs	Non-defaulters
Defaulters	200	72
Non-Defaulters	1127	4600

Data Set Information:

The data contains 30000 observations with the predictor variables as well as the response variable. Our test set contained 6000 random observations with the response variable removed.

Variable descriptions:
This employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
LIMIT_BAL: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
SEX: Gender (1 = male; 2 = female).
Education: (1 = graduate school; 2 = university; 3 = high school; 4 = others).
Marital status: (1 = married; 2 = single; 3 = others).
Age: In years.
PAY_1 – PAY_6: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: PAY_1 = the repayment status in September, 2005; PAY_2 = the repayment status in August, 2005; . . .; PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
BILL_AMT1 – BILL_AMT6 -: Amount of bill statement (NT dollar). BILL_AMT1 = amount of bill statement in September, 2005; BILL_AMT2 = amount of bill statement in August, 2005; . . .; BILL_AMT6 = amount of bill statement in April, 2005.
PAY_AMT1 – PAY_AMT6: Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September, 2005; PAY_AMT2 = amount paid in August, 2005; . . .; PAY_AMT6 = amount paid in April, 2005.