Proposal

Machine Learning Nanodgree capstone

Minxia Ji

Background:

When a customer accept a credit card, he or she needs to agree to certain terms, such as make minimum payment by the due date listed on theircredit card statement. If the customer missed the minimum credit card payment six months in a row, his or her credit card will be in default. The credit card issuer will likely close the customer's account and report the default to the credit bureaus.

But to the credit card issuer(usually a bank), the lost might be irretrievable.Eventually, usually after a period of 90 days of nonpayment, the loan/payment is written off. Banks are required by law to maintain an account for loan loss reserves to cover these losses.

Banks could reduce credit risk by conducting a credit risk analysis on credit card applicants/holders. Banks can substantially reduce their credit risk by lending to their customers, since they have much more information about them than about others, which helps to reduce adverse selection. Checking and savings accounts can reveal how well the customer handles money, their minimum income and monthly expenses, and the amount of their reserves to hold them over financially stressful times. Banks will also verify incomes and employment history, and get credit reports and credit scores from credit reporting agencies.

Reference:
Bank Risks
http://thismatter.com/money/banking/bank-risks.htm
Banks that make the most money, and the least, on credit card loans
http://www.creditcards.com/credit-card-news/bank-yields-loans-1276.php

Dataset and inputs

Data:default of credit card clients Data Set
Data Source:UCI https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Overview of the dataset:

3000 instances and 23 features are included in the dataset. Most of features are numerical features, i.e. bill statement, and we need to check the skewness and may need to do some transformation. Some of features, which represent customer information, i.e.gender, are categorical. We need to transfer variables under these features to dummies variables.

Features in the dataset:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = 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.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

2 Classes:

default payment next month:

1: stands for the customer who will conduct default payment next month.Has **6636** instances.

0: stands for customer who will not conduct default payment next month. Has **23364** instances.

Spliting data issues are discussed in the solution statement part.

A problem statement:

Inputs:

3000 instances with 23 features and one label are included in the dataset. Most of features are numerical features, i.e. bill statement. Some of features, which represent customer information, i.e.gender, are categorical.

Output:

a trained model which performed well on predicting whether a customer would default or not given his/her information.

Learning task:

Binary classification: detecting the clients who are likely to default or not(default:1;not default:0)

A solution statement:

Data:

-Deal with class imbalance: try both random undersampling the majority class and oversampling minority class

-Splitting data: training set: 75% testing set 25%. In order to maintain class balance in training and testing sets, use train_test_split function in sklearn and specify a param ‘stratify’.

-Check the skewness of features and deal with it.(i.e.log transformation.)

Models:

-Classification models: employ RandomForestClassifier, GradientBoostingClassifier and LogisticRegressionClassifier.

-Train and find a best classifier and optimize it using grid search method

-Conduct predicting default based on new dataset using the classifier

A benchmark model

Detect credit card fraud by decision trees and svm

http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp442-447.pdf

A set of evaluation metrics

-**$F_{0.5}$score** =$(1+0.5^2)\cdot$ (precision $\cdot$ recall/$(0.5^2\cdot$ precision)+recall)

Why $F_{0.5}$score: our aim is to find out who will conduct default payment in the future.We hope that we could improve the precision(true positive/classified positive). Thus, we can introduce F-beta score, choose beta = 0.5 so that more emphasis is placed on precision.

-**False negative rate =** false negative/(true positive + false nagetive)

In our problem, if the false negative rate is low, then our model is good. Otherwise, if we have high false negative rate, which means among the clients who are going to default, we classified a lot of them as 'not going to default'. That is extremely bad and we don't want to see that happens.

An outline of the project design

-Data exploration/visualization and preprocessing

Techniques: log transformation, undersampling and oversampling

-Build classifiers:

GrandientBoostingClassifier

RandomForestClassifier

LogisticRegressionClassifier

and choose a best classifier with certain criteria.Training and testing on 2 datasets(dataset with oversampling on minority class and dataset with undersampling on majority class) to see which model and dataset combination performs best based on the evaluation metrics.

-Parameter tunning:

Do some search and reading to understand which parameters should be tunned for the current dataset.

Use grid search to find the best parameters for the classifier.

-Performance evaluation:False nagative rate and $F_{0.5}$score on testing dataset

-Conclusion and ways to improve the performance.