PD Model

Xiaogang (Mark) Li

August 24, 2021

Packages Required in Python

- numpy (for data processing)
- pandas (for data frame)
- matplotlib.pyplot (for plots)
- seaborn (for plots)
- sklearn (for modeling)
- scipy (for modeling output)
- pickle (for model usage)

General Comments on the Data

- No data issues were identifies.
- There is no missing values or outliers, as such, no missing value and outlier imputation needed.
- Prepayment delays, bill statements, and previous payments are monthly data from April 2005 to September 2005, which are not good features to set up the model
- Personal information, including gender, age, education level, and marital status are good indicators so should be included in the model.
- Personal information may be sensitive and may breach the policy. If so, probably could remove them from the model.

Define New Variables

- There are 6 new variables were defined, covering average prepayment delays in 6 months and 3 months, average bill statement to credit limit (similar to LTV) in 6 months and 3 months, and average prepayments to credit limit in 6 months and 3 months
- In particular, for 6 months:
  1. **Average Payment Delays** = Average of payment delays from April to September, 2005
  2. Average Bill Statement = Average amount of bill statement from April to September, 2005
  3. Average Previous Payment = Amount of previous payment from April to September, 2005
  4. **Bill to Credit** = Average Bill Statement / Amount of the given credit
  5. **PrePay to Credit** = Average Previous Payment / Amount of the given credit

Define New Variables (Cont')

- For 3 months:
  1. **Average Payment Delays 3mon** = Average of payment delays from July to September, 2005
  2. Average Bill Statement 3mon = Average amount of bill statement from July to September, 2005
  3. Average Previous Payment 3mom = Amount of previous payment from July to September, 2005
  4. **Bill to Credit 3mon** = Average Bill Statement 3mon / Amount of the given credit
  5. **PrePay to Credit 3mon** = Average Previous Payment 3mon / Amount of the given credit

Summary Statistics

	count	mean	std	min	25%	50%	75%	max
X2	30000	1.6037	0.4891	1	1	2	2	2
X3	30000	1.8531	0.7903	0	1	2	2	6
X4	30000	1.5519	0.5220	0	1	2	2	3
X5	30000	35.4855	9.2179	21	28	34	41	79
X6	30000	-0.0167	1.1238	-2	-1	0	0	8
Avg_pay_delay	30000	-0.1824	0.9822	-2	-0.8333	0	0	6
Bill_to_Credit	30000	0.3730	0.3519	-0.2326	0.0300	0.2848	0.6879	5.3643
PrePay_to_Credit	30000	0.0389	0.0526	0	0.0113	0.0261	0.0439	2.4277
Avg_pay_delay_3m	30000	-0.1056	1.0370	-2	-1	0	0	7
Bill_to_Credit_3m	30000	0.4090	0.3907	-0.4652	0.0272	0.3087	0.7819	6.0757
PrePay_to_Credit_3m	30000	0.0426	0.0701	0	0.0107	0.0269	0.0448	4.2723

Baseline Model - Logistic Regression

- The baseline model selected is Logistic Regression.
- The probability of default can be expressed as

$$PD = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

where $n = 11$ and x_i represents for independent variables.

- Reasons on Why Logistic Regression was selected
  - Logistic Regression is a classic statistical model which is widely used in industry to dealing with classification data.
  - Logistic Regression learns a linear relationship and then introduce a non-linearity in terms of a Sigmoid function. The linear relationship helps to explain the importance of features.
  - Logistic Regression is easier to implement, interpret, and very efficient to train
  - Logistic Regression is less inclined to overfitting comparing to other machine learning algorithms.

Model Training Output

Feature name	Coefficients	p_values
Intercept	-0.93055	
X2	-0.13433	0.00000
X4	-0.10202	0.00006
X5	0.004494	0.00021
Bill_to_Credit	0.743688	0.00008
PrePay_to_Credit	-1.55733	0.00014
Avg_pay_delay_3m	0.766735	0.00000
Bill_to_Credit_3m	-0.91978	0.00000

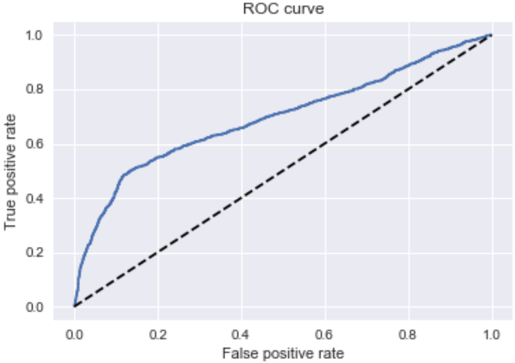
Feature Importance

Features	Importances
Avg_pay_delay_3m	0.76673534
Bill_to_Credit	0.743687759
X5	0.004493617
X4	-0.102017352
X2	-0.134333188
Bill_to_Credit_3m	-0.919783464
PrePay_to_Credit	-1.55732726

- Average of payment delays from July to September 2005 and Average Bill Statement / Amount of the given credit (Bill statement to Credit Limit ratio) are the two most important features. The higher of these two variables, the higher probability that a customer will default
- In contrast, PrePay to Credit ratio (Average Previous Payment / Amount of the given credit) and 3 month Bill to Credit ratio (Average Bill Statement 3mon Amount of the given credit) are the two most important features that have negative effect on PD.

Out-of-sample Validaiton

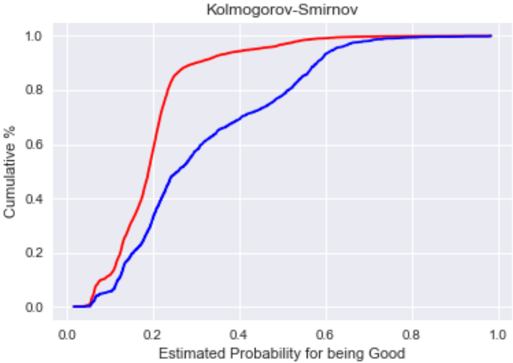
- Accuracy ratio is 0.783 which indicates that the model has a excellent predication power
- ROC Curve



- AUC is 0.695 which also indicates a excellent predication power

Gini and Kolmogorov-Smirnov

- Gini is 0.3903 which is smaller than 0.4
- KS Curve



- KS is 0.366 which also indicates a strong predication power

Cross Validation

- Minimin, Average, and Maximum Accuracy Score using 10-folder cross validation are 0.79875, 0.80512, and 0.81875, respectively
- Minimin, Average, and Maximum AUC under ROC using 10-folder cross validation are 0.66299, 0.68251, and 0.72191, respectively
- Cross Validation confirms that the model has a strong prediction power

Conclusion

- The Logistic Regression model has a strong predication power in predicting customer's probability of default given using the given data
- Average of payment delays from July to September 2005 and Average Bill Statement / Amount of the given credit (Bill statement to Credit Limit ratio) are the two most important features. The higher of these two variables, the higher probability that a customer will default
- Business should monitor closely on customer's Bill statement to payment delays and Credit Limit ratio. If these two variables are relatively large then further steps should be applied to avoid default.