EED-497 : End Sem PPT

Fraud Detection in Financial Systems using Statistical Learning

Machine Learning Group-1
By: Yatharth Jain and Ojas Srivastava | Guided By: Prof. Madan Gopal

Abstract

- ➔ This project will discover methods to detect ***default in Financial Systems.***
- ➔ Our aim is to create a robust system that would ***detect patterns of default*** and predict any potential default which could be encountered by the system.
- ➔ We will use ***UCI Machine Learning Credit Card Default Data***.(1) This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

The Problem - Background of Data

- In Taiwan, in February 2006, ***debt from credit cards*** and cash cards reached ***$268 billion USD.***
- More than ***half a million people*** were not able to repay their loans. They became ***"credit card slaves".***
- At the end of 2005, card debts crisis was exploded in Taiwan, ***700 thousand people*** have become card-slaves, and the ***average money owed was one million NTD***. (2)
- Some debtors and their families committed ***suicide because of the debt***, some became homeless due to repossession of their homes, and others could not afford to pay their children's tuition.
- The ***suicide rate increased 22.9%*** compared to the rate in 2005, and the main reason is unemployment and credit card debt. (3)

Problem Formulation & Dataset Description

- ➔ The default detection problem in credit cards can be formalized based on the data provided for paid amount, billed amount, and other demographic data like gender, age, education background, etc.
- ➔ The task is to predict the ***customer who could become a defaulter*** in coming month.
- ➔ Any bank would like to detect as many defaulters as possible while not wrongly labelling any of its loyal & honest customers.

The Dataset has 25 variables, a brief snapshot is given below:

- ID [Qualitative, Customer ID]
- LIMIT_BAL [Qualitative, Limit on Credit Card]
- SEX [Qualitative]
- EDUCATION [Qualitative]
- MARRIAGE [Qualitative]
- AGE [Qualitative]
- PAY_*n* [Quantitative, Payment Status delay]
- BILL_AMT_*n* [Quantitative, Billed Amount]
- PAY_AMT_*n* [Quantitative, Paid Amount]
- Default in next month [Decision Variable]

Imbalance in Dataset →

Non-Defaulters: 78%
Defaulters: 22%

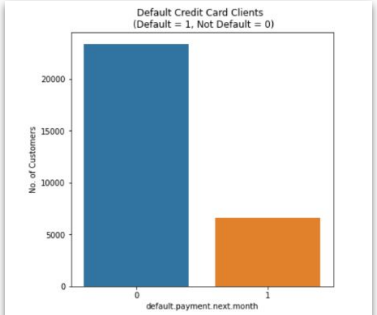


Figure 3.5: Imbalance in Dataset

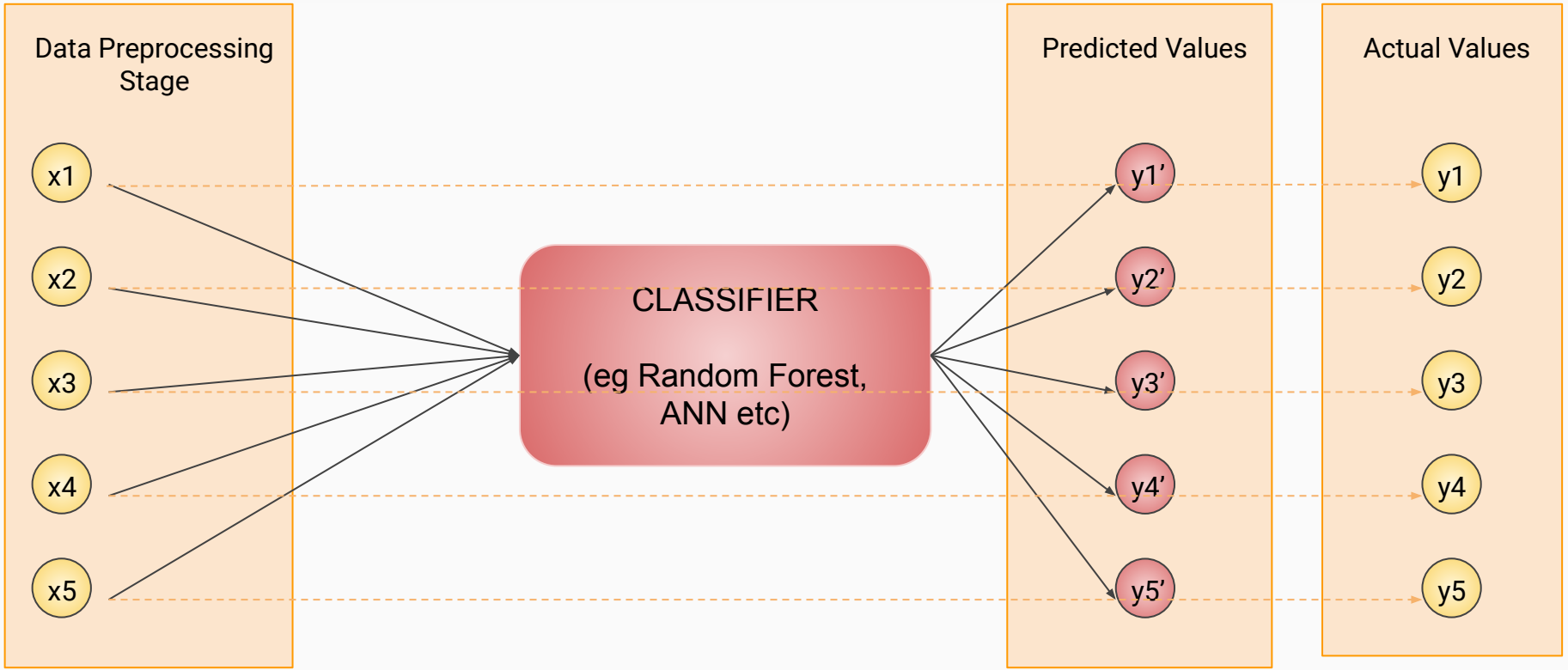
Data Pre-Processing & Train-Test Split

- **Binary Encoding** : All the categorical variables were **binary encoded by us**. Our categorical variables which were in *integers* and were transformed into a binary number. After that binary value is split into different columns, **then we made a different column for each digit of this binary equivalent.**

- **Data normalisation** is the step which is used to **remove the dominance of a particular feature** by bringing all the features on same scale. **Normalized data to zero mean and unit variance.**
  - The formula used is: $\mathbf{x_j^{(i)\prime} = (x_j^{(i)} - \mu)/\sigma}$

- Throughout the project we have employed a **75-25 train-test split**. Separate data frames for the following were made:
  - X_train: Data frame of all features in the training set. (csv file: See Here)
  - Y_train: Dependant variable data frame of the training set. (csv file: See Here)
  - X_ test: Data frame of all features in the testing set. (csv file: See Here)
  - Y_test: Dependant variable data frame of the testing set. (csv file: See Here)

Working of a Binary Classifier (Positive Class=1)



There is bound to be differences in the predicted output and ground truth, hence it will also give *True Positives, False Positives etc*

Common Metrics for Evaluating Performance

BALANCED DATASET

- **Misclassification Error:** Total number of incorrect classifications divided by the total number of classifications.

*The decisions made on the basis of classifications that are based on misclassification error rate result in poor performance when **training data is unbalanced.** For the datasets that are unbalanced as ours an alternative strategy is to consider the ratio of one cost to another.*

IMBALANCED DATASET

- **True Positive Rate or Recall:** The tp rate tells us how sensitive our decision method is in the detection of the abnormal event. A classification method with high sensitivity will rarely miss the abnormal event.
- **Precision:** Ratio of correctly classified positives to total predicted positives.
- **F_1 Score:** Harmonic mean of precision and recall. Higher the F_1 Score better the performance of model.
- **Precision-Recall Curve:** Plots precision on y-axis and recall on x-axis. The **optimal threshold point** on the curve is the one having highest F_1 Score. The **break-even point** is the point on the curve that has same value for both precision and recall.

Performance Metrics Used (Imbalanced Training Data)

We will use Precision & Recall Metrics. Formula for both have been provided below:

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

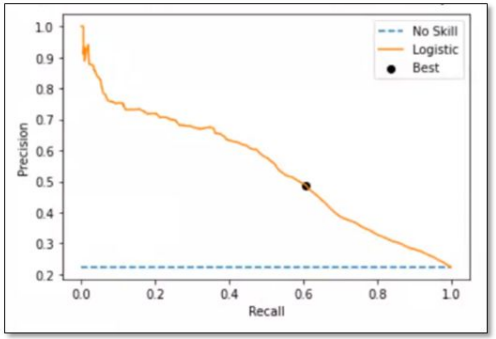
A few points to keep in mind:

- According to the formula of the metric it is evident to see that we need to have higher recall for class - 0. Which means lesser number of honest customers being wrongly labelled
- At the same time we need high recall for class -1 which will imply higher detection of default and lower number of wrongly labelled dishonest customers
- It is also desired to have high precision values for both the classes.

Basic Models (Imbalanced Dataset) - Best Model

Algorithm	*Class 0* *Precision & Recall*	*Class 1* *Precision & Recall*	*Accuracy Rate*	*Misclassification Rate*	*F_1-score* *Class -0* *Class-1*
Random Forest	**0.84** **0.94**	**0.65** **0.39**	**81.54%**	**18.46%**	**0.89** **0.48**

Precision-Recall Curve of the Best Model (in yellow)



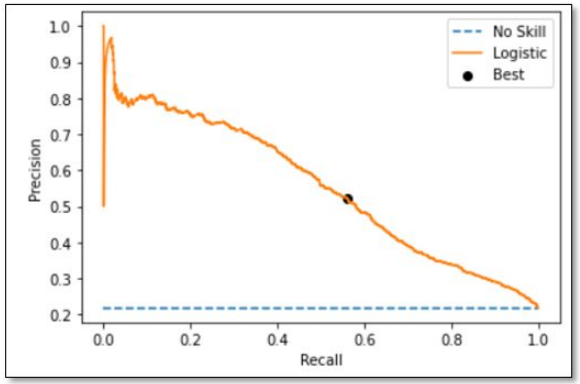
Confusion Matrix

	Actual	
Predicted	0	1
0	5470	354
1	1030	646

Cost Sensitive Learning (Best Point) - Best Model

Algorithm	Class 0 Precision & Recall	Class 1 Precision & Recall	Accuracy Rate	Misclassification Rate	Threshold
ANN-2	0.87 0.86	0.52 0.56	79.32%	20.68%	0.27

Precision-Recall Curve of the Best Model (in yellow)



Confusion Matrix

	Actual	
Predicted	0	1
0	1083	353
1	317	99

Data Balancing Techniques & Performance Metrics used

We used 3 different techniques to balance the data in our training sets:

- **SMOTE (Synthetic Minority Oversampling TEchnique):** It creates synthetic (not duplicate) samples of the minority class. Hence making the minority class equal to the majority class.

- **Over-Sampling:** It creates duplicate samples of the minority class. Hence making the minority class equal to the majority class.

- **Under-Sampling:** Under sampling consists of reducing the data by eliminating examples belonging to the majority class with the objective of equalizing the number of examples of each class.

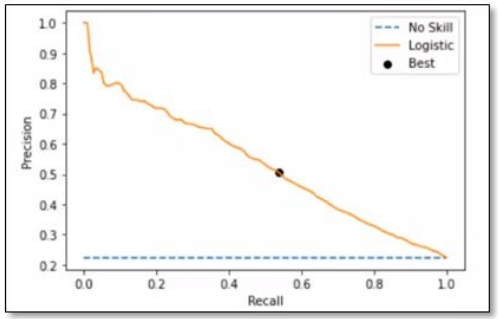
The above techniques of data manipulation are always applied only on the training set while keeping the test set untouched.

Now, since our models were trained on balanced datasets, the performance metrics which we used for comparison were simply the **Misclassification Rate** or **Accuracy Rate.**

Data Manipulation - Best Model

Algorithm	*Class 0* *Precision & Recall*	*Class 1* *Precision & Recall*	*Accuracy Rate*	*Misclassification Rate*	*F_1-score* *Class -0* *Class-1*
Random Forest (OverSampled Data)	**0.85** **0.91**	**0.59** **0.45**	**80.8%**	**19.2%**	**0.88** **0.51**

Precision-Recall Curve of the Best Model (in yellow)



Confusion Matrix

	Actual	
Predicted	0	1
0	4998	826
1	800	876

Use of Break-Even Point for Thresholding

We are getting two types of models:

1. There is a visible trade-off between identifying higher number of defaulters at the cost of genuine customers being labelled wrongly.
2. Identification of good customers at a higher level but with a lower level of detecting defaulters.

This is a business decision depending on a particular bank as to what kind of model it would like to opt for. If a bank is well established then it can opt for first model in which there is a need of setting up of a department which would monitor the results of the model while if a bank is still finding its foot in the industry then it can opt for second model where less number of its honest customers are troubled on the cost predicting less number of defaulters

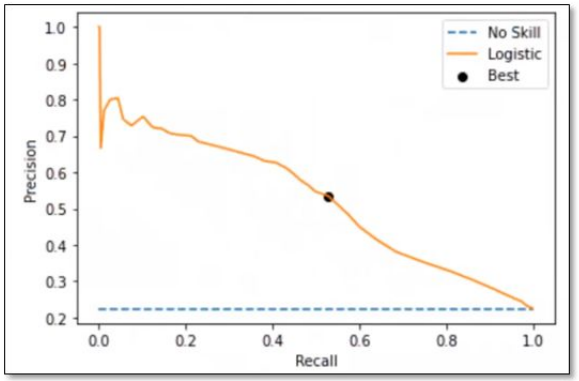
Since we were not able to make a decision on our own, we decided to opt for break-even point for deciding the threshold for our models.

➢ **Break-Even Point:** The **break-even point** is the point on the precision-recall curve that has same value for both precision and recall. This point is used for thresholding when we are not sure about business requirements from our model.

Cost Sensitive Learning (Break Even Point) - Best Model

Algorithm	Class 0 Precision & Recall	Class 1 Precision & Recall	Accuracy Rate	Misclassification Rate	Threshold
KNN	0.87 0.87	0.52 0.52	79.59%	20.41%	0.30

Precision-Recall Curve of the Best Model (in yellow)



Confusion Matrix

	Actual	
Predicted	0	1
0	5134	690
1	841	835

References

(1) : UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

(2) Chih Hsiung Chang , Heidi H. Chang , Jui-Chu Tien , "A Study on the Coping Strategy of Financial Supervisory Organization under Information Asymmetry: Case Study of Taiwan's Credit Card Market," Universal Journal of Management, Vol. 5, No. 9, pp. 429 - 436, 2017. DOI: 10.13189/ujm.2017.050903.

(3)sevenpillarsinstitute.org, [Online]. Available: https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/.

(4) S. S. H. Padmanabhuni, A. S. Kandukuri, D. Prusti and S. K. Rath, "Detecting Default Payment Fraud in Credit Cards," 2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, India, 2019, pp. 15-153, doi: 10.1109/ICISGT44072.2019.00018

(5) A. Lawi and F. Aziz, "Classification of Credit Card Default Clients Using LS-SVM Ensemble," 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-4, doi: 10.1109/IAC.2018.8780427.

(6) A. Lawi and F. Aziz, "Classification of Credit Card Default Clients Using LS-SVM Ensemble," 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-4, doi: 10.1109/IAC.2018.8780427.

(7) Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.