Credit-Card Default

Classification and Misclassification

Problem Statement

Given a dataset with demographic and borrowing history data for Taiwanese credit-card customer accounts classified as defaulting or not defaulting in October-2005, can I build a supervised model that performs better than identifying only members of the negative non-default class (baseline model) while minimizing the misclassification of either group? In this context, if I can predict accounts as belonging to the defaulting group, I want to minimize the number of predicted defaulters who did not actually default that October (lost revenues) while minimizing the number of predicted non-defaulters who did end up defaulting (lost profits).

Welcome to Taiwan, R.o.C. (a.k.a. Formosa)

Where is Taiwan?

What is it known for?

Population and income?

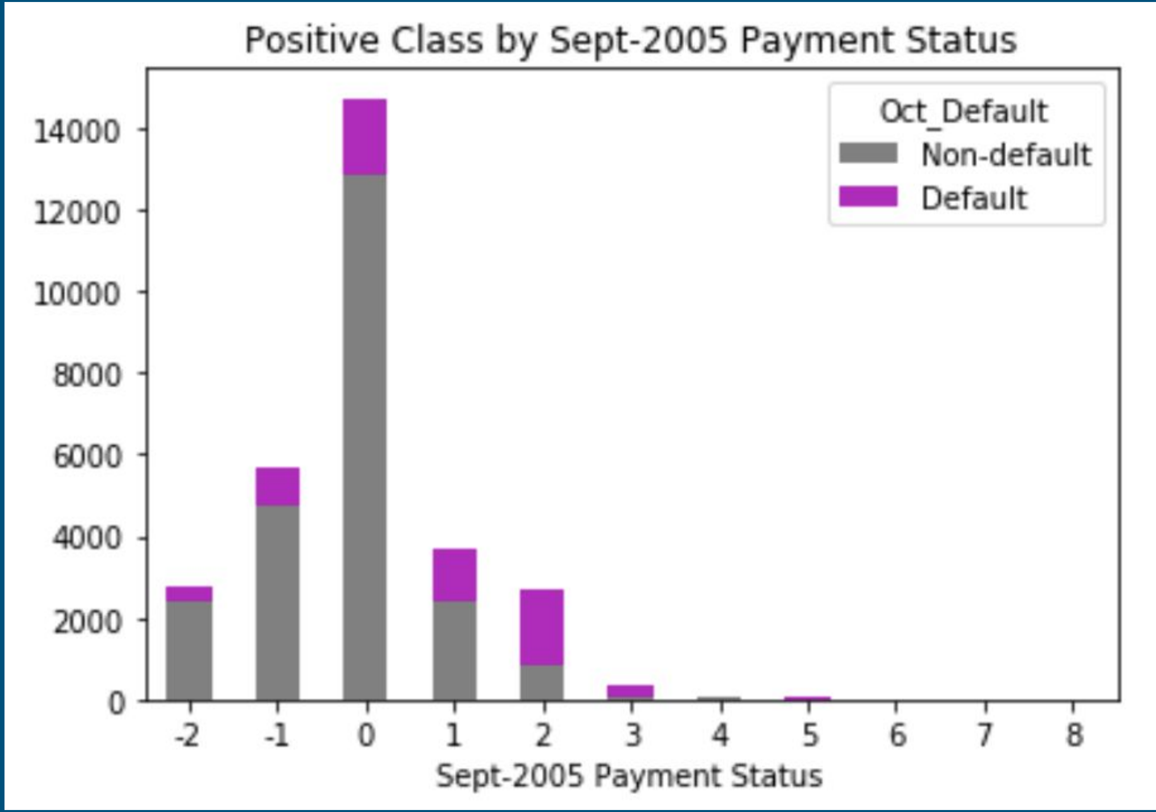
The Data

- 30,000 customer accounts
- One dependent variable for <u>defaulted</u> or <u>not-defaulted</u> in October of 2005
- Demographic information on account-holders
    - Age and Gender
    - Highest level of education
    - Marital status

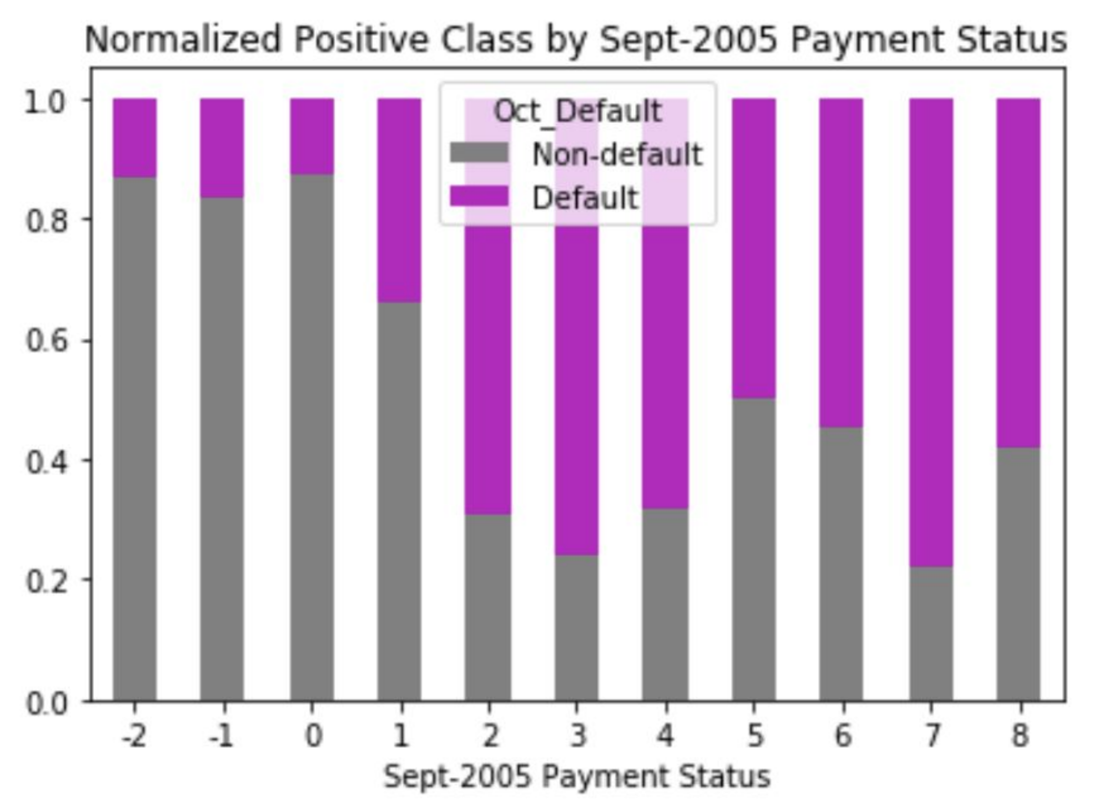
The Data (continued)

- Credit Limit
- Borrowing activity for each of the six months prior to October-2005:
    - Payment-history status (number of months current or behind)
    - Billing amount
    - Amount of payments received

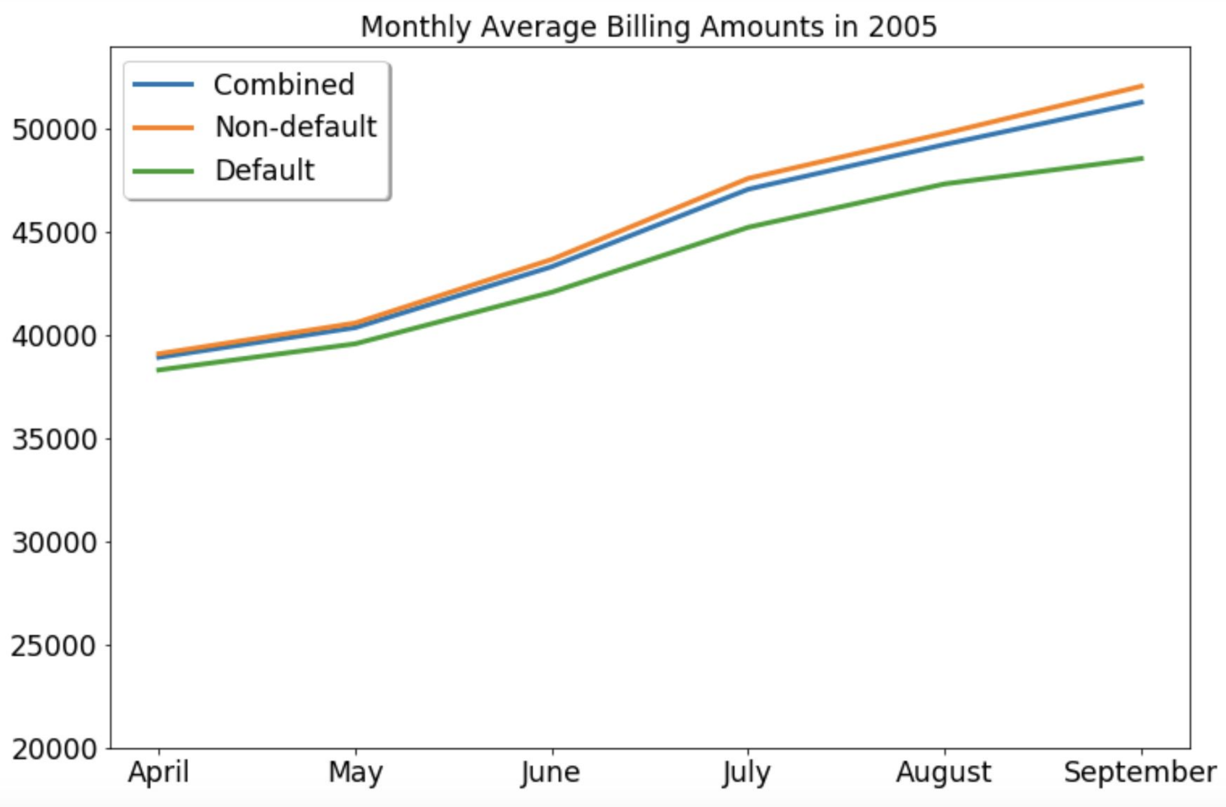
Sept-2005 Payment Status



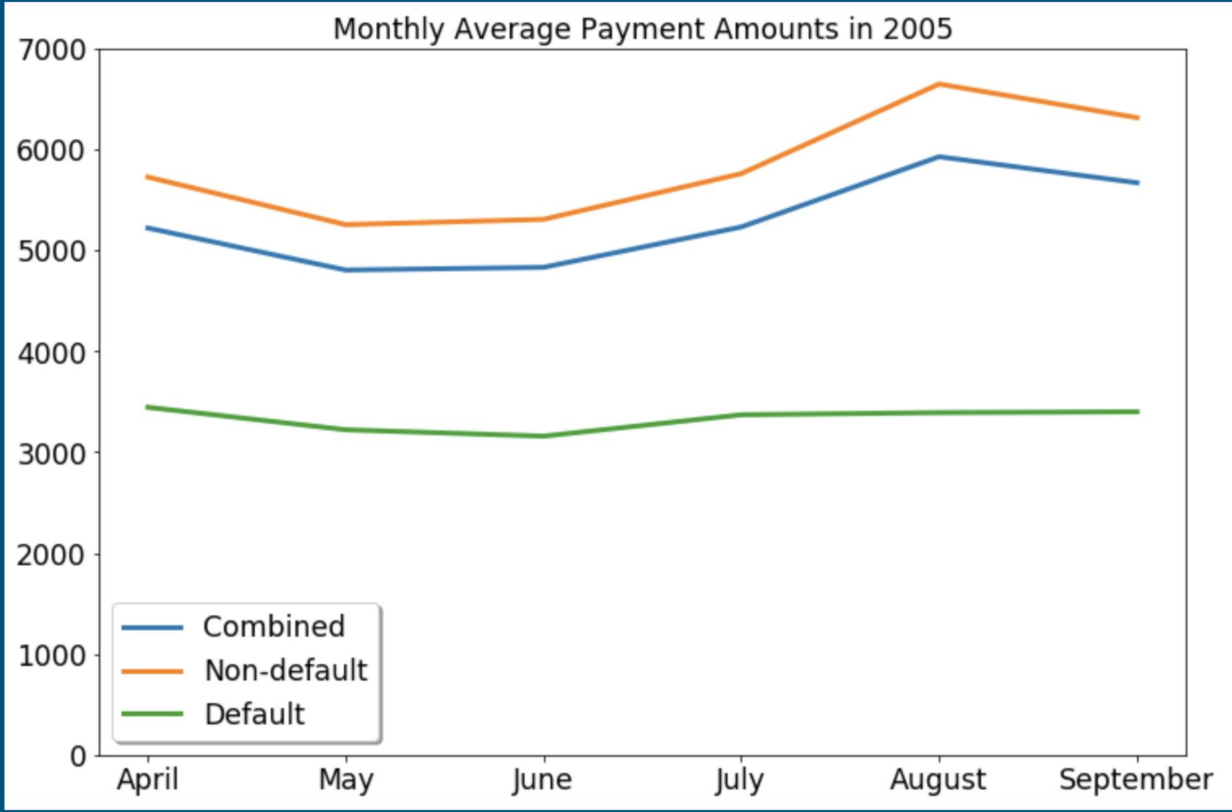
Sept-2005 Payment Status by Percentage



Credit-card Bill Amounts



Credit-card Payment Amounts



The Pitfalls of Default-Classification

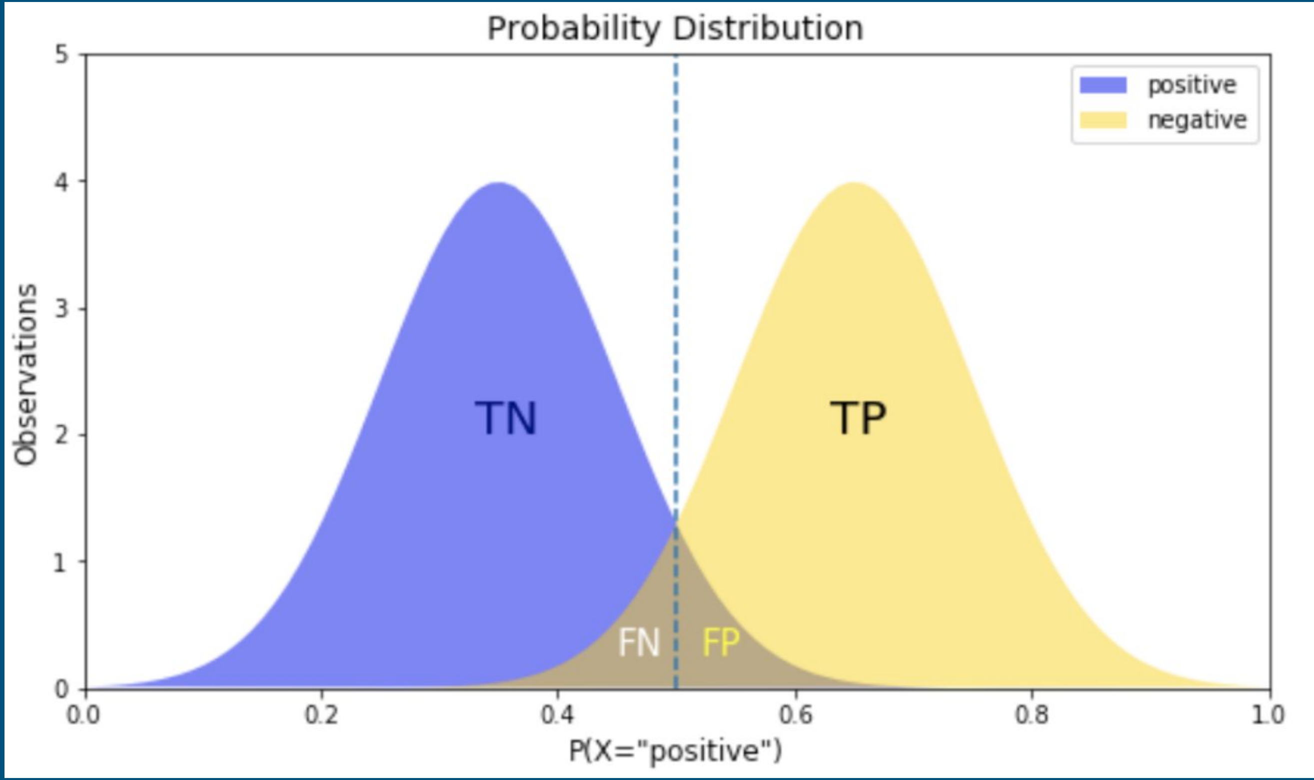
Financial Loss:

“Default classified as Non-default” 👎 👎

Lost Business:

“Non-default classified as Default” 👎

Tale of Two (Balanced?) Classes



Miss Rates and Fall-Outs

$$MissRate = \frac{FN}{(FN + TP)}$$

$$FallOut = \frac{FP}{(FP + TN)}$$

Analytical Techniques

Regression Modeling:

Linear Regression

Logistic Regression

Time Series

Classification Trees

etc.

Machine-learning Modeling:

Artificial Neural Networks

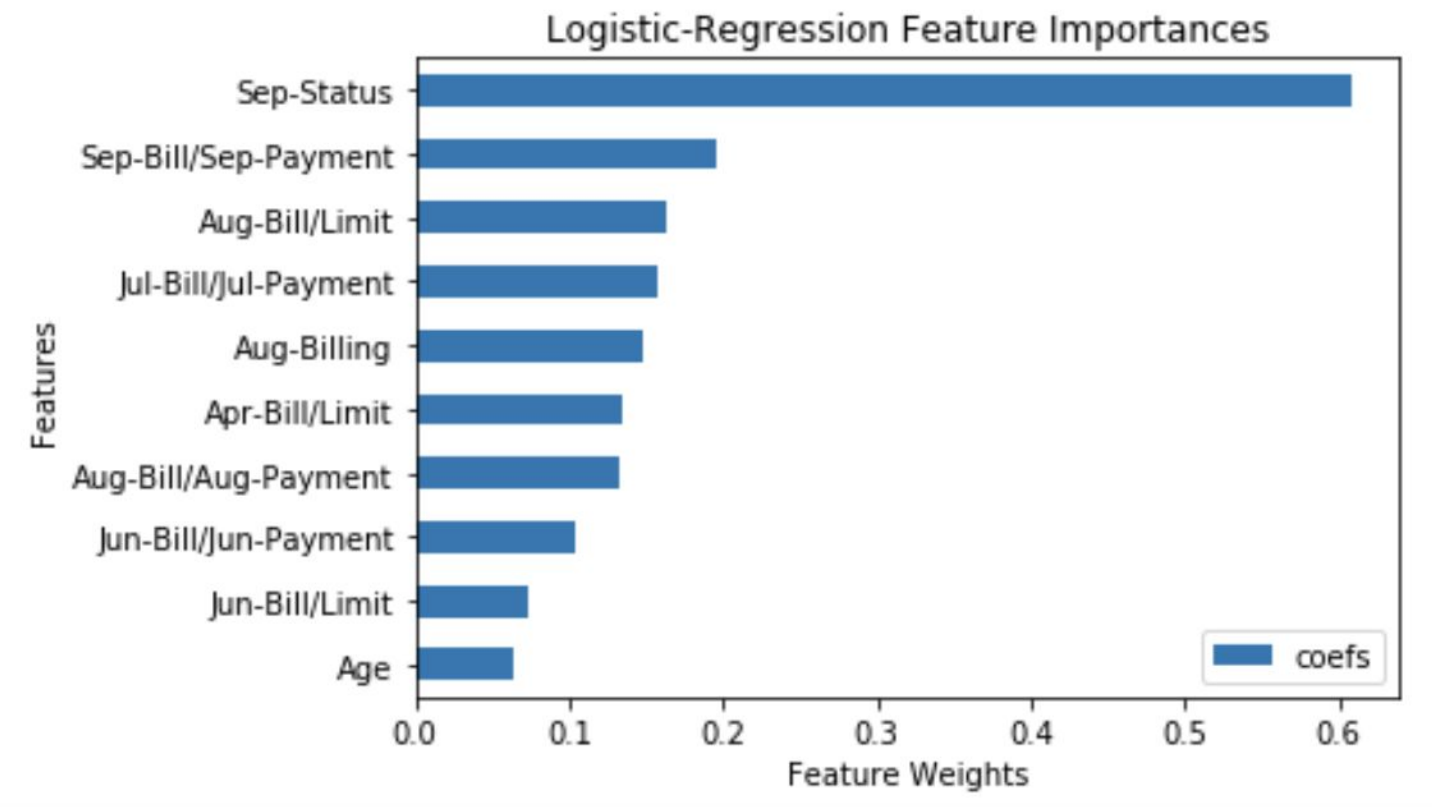
Support Vector Machines

Naive Bayes

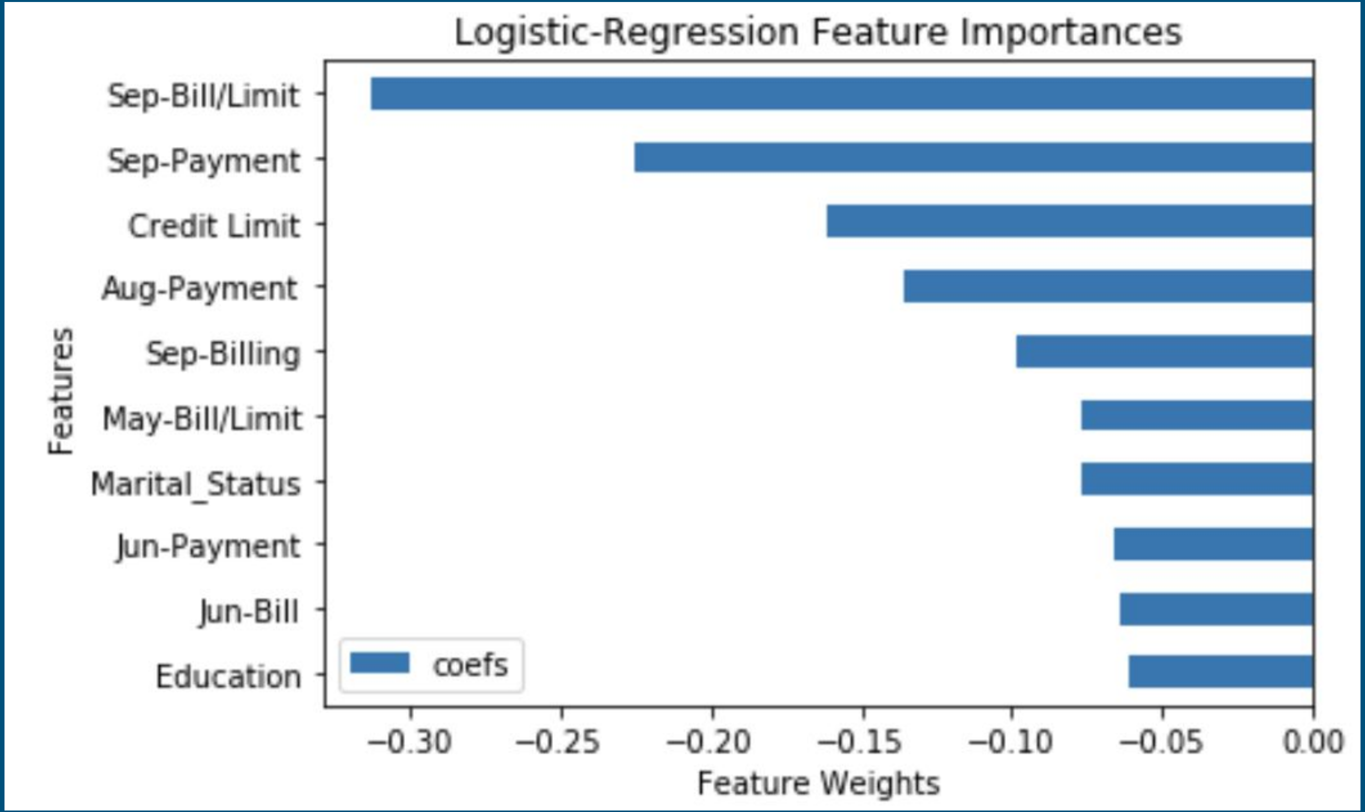
k-nearest Neighbor

etc.

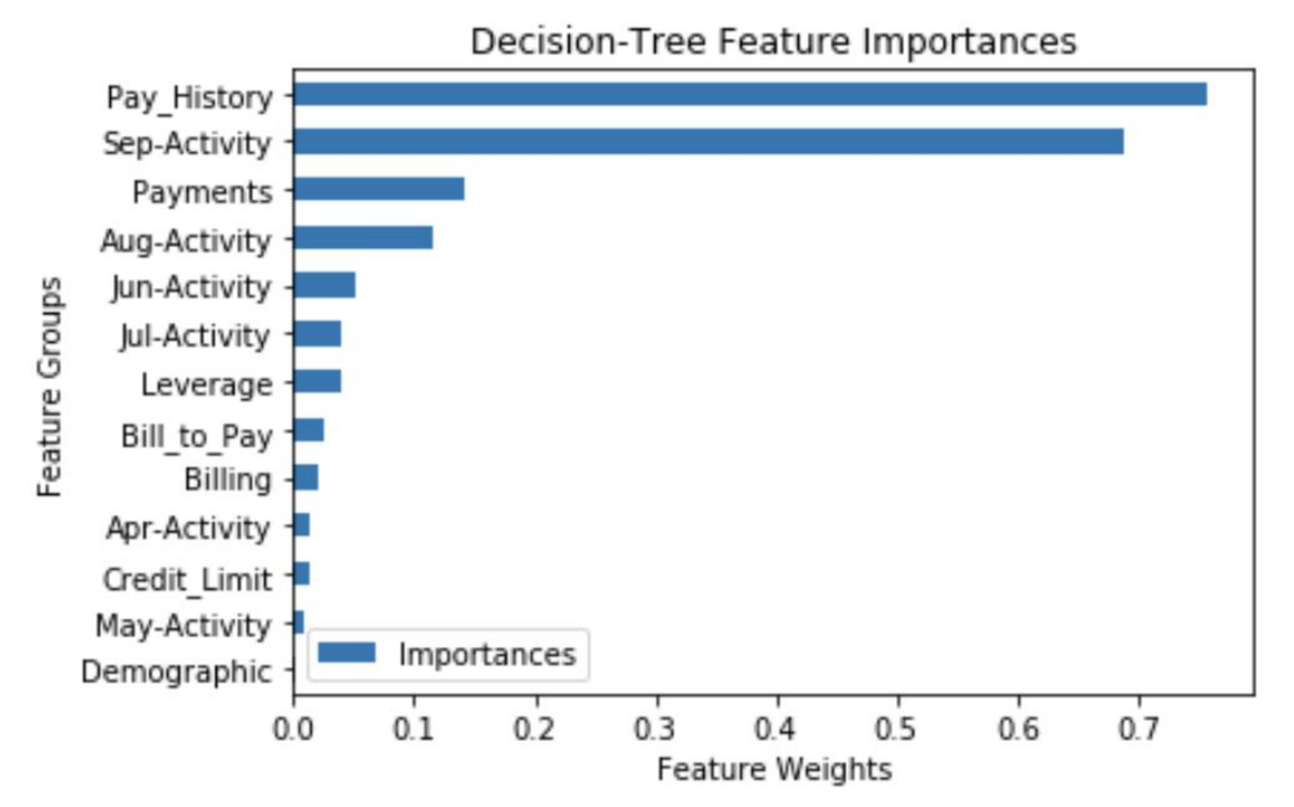
Logistic-regression Feature Coefficients



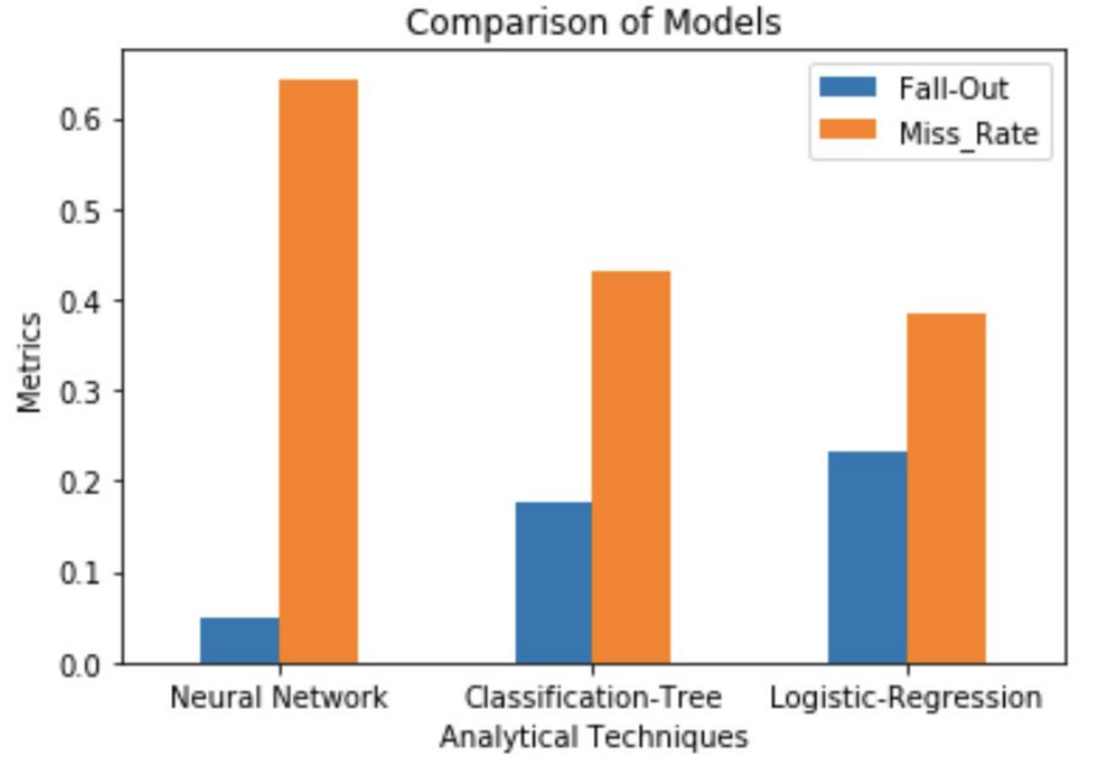
Logistic-regression Feature Coefficients (cont'd)



Classification-Trees Feature Importances



Metrics Summary



Conclusions

- Of the three modeling techniques, a neural-network provides the best Fall-Out rate (customer-retention metric), while logistic-regression provides the best Miss Rate (loss-mitigation metric)
- Alternatively, the Classification-Tree model provides a good balance of both lower Fall-Out and Miss Rates
- (Limiting feature selection to billing, payments and history-status for the prior two months, is worth further analysis)

Appendix

1. Original publication of the sources dataset; actual Taiwan default rate
2. Level-of-education and credit-limit distributions; by bin-percentages
3. Advantages of neural-network modeling
4. List of available Keras-on-TensorFlow loss-functions
5. Comparison of misclassification types by loss functions tested

Well-known Dataset

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Expert Systems with Applications 36 (2009) 2473–2480

Expert Systems with Applications

www.elsevier.com/locate/eswa

The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients

I-Cheng Yeh [a,*], Che-hui Lien [b]

[a] *Department of Information Management, Chung-Hua University, Hsin Chu 30067, Taiwan, ROC*

[b] *Department of Management, Thompson Rivers University, Kamloops, BC, Canada*

Actual Default Rates

A Two- Stage Cardholder Behavioural Scoring Model Using Artificial Neural Networks and Data Envelopment Analysis

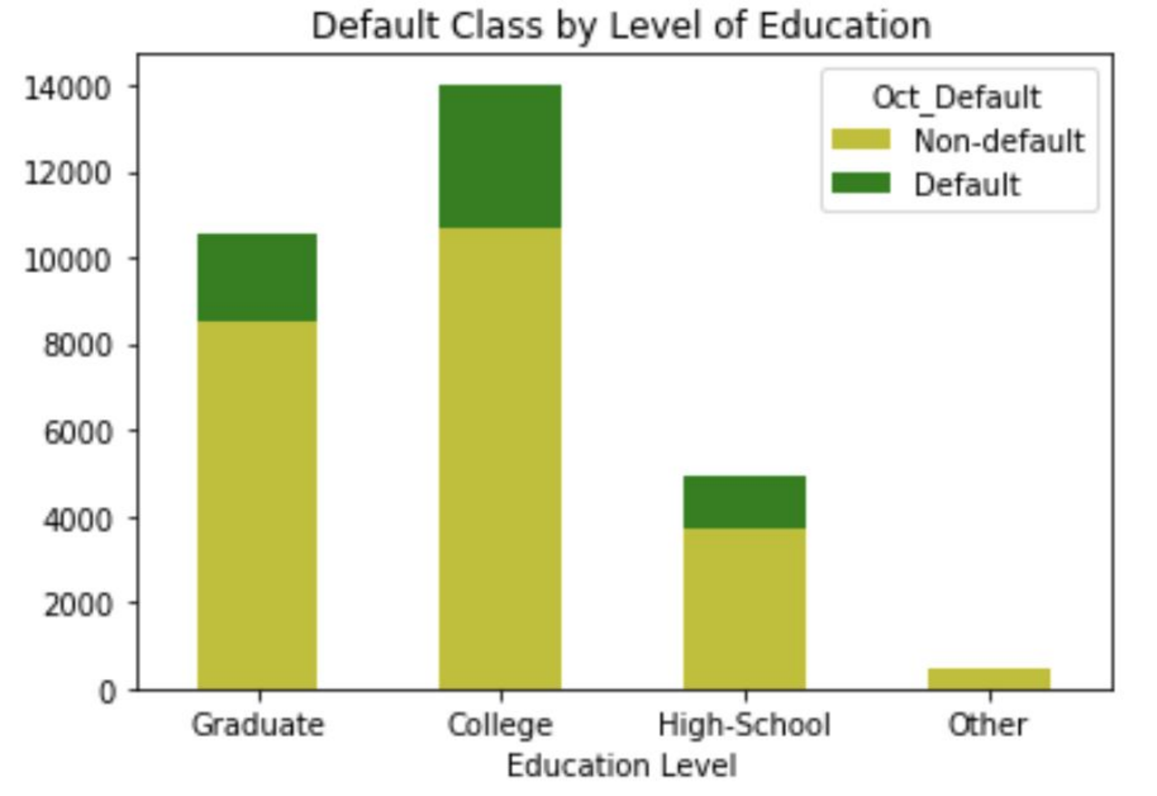
I-Fei, Chen
Tamkang University, Department of Management Science and Decision Making,
151 Ying-chuan Road, Danshui Dist., New Taipei City 25137 Taiwan, enfa@mail.tku.edu.tw
doi:10.4156/ijact.vol3.issue2.11

Abstract

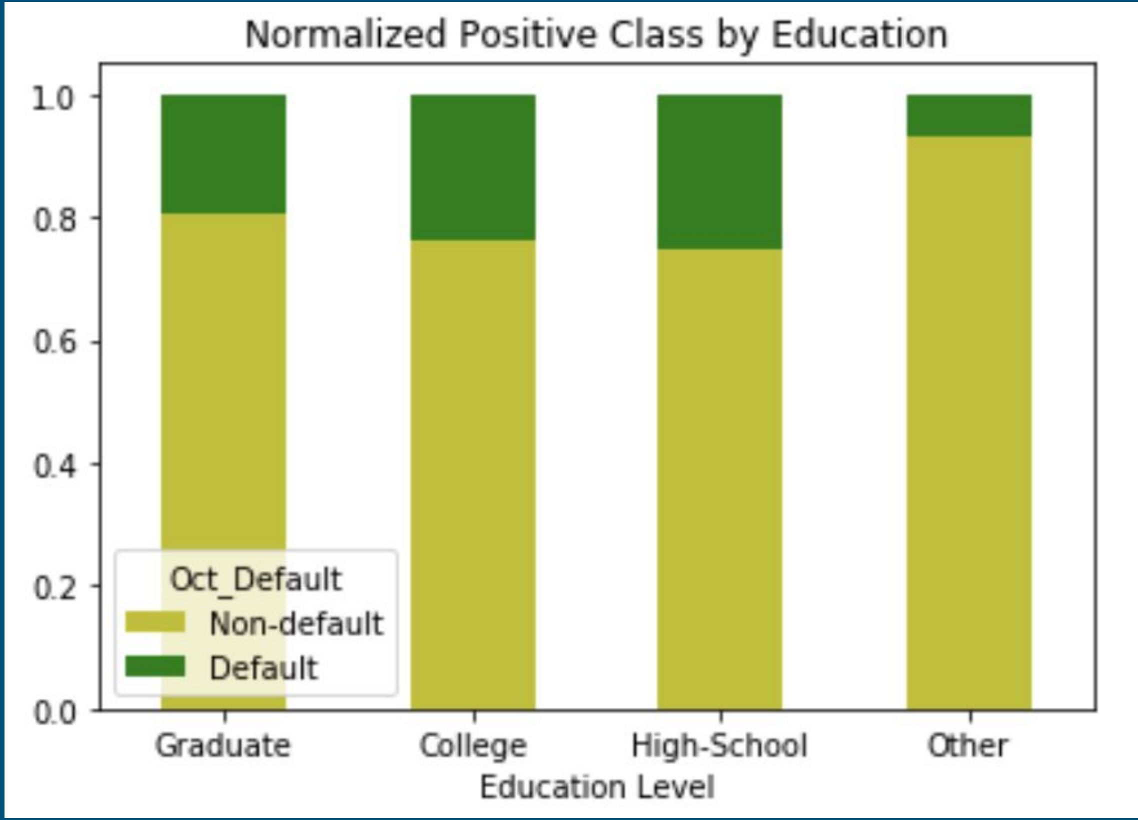
Since the databases that banks use for analysis of cardholders' repayment behaviours are usually large and complicated, and the extant classification techniques hardly offer 100% correct

used in this study. There are totally 700 cardholders in the dataset with 500 good credit customers (class 0), 175 revolvers (class 1) and the remaining 25 are bad credit customers (class 3). The relative ratios of bad customers to total customer is 3.57% very close to the national standard in Taiwan and hence should be a representative dataset for testing the practicability of the proposed scheme. Each cardholder in the dataset contains 34 independent variables containing demographic characteristics

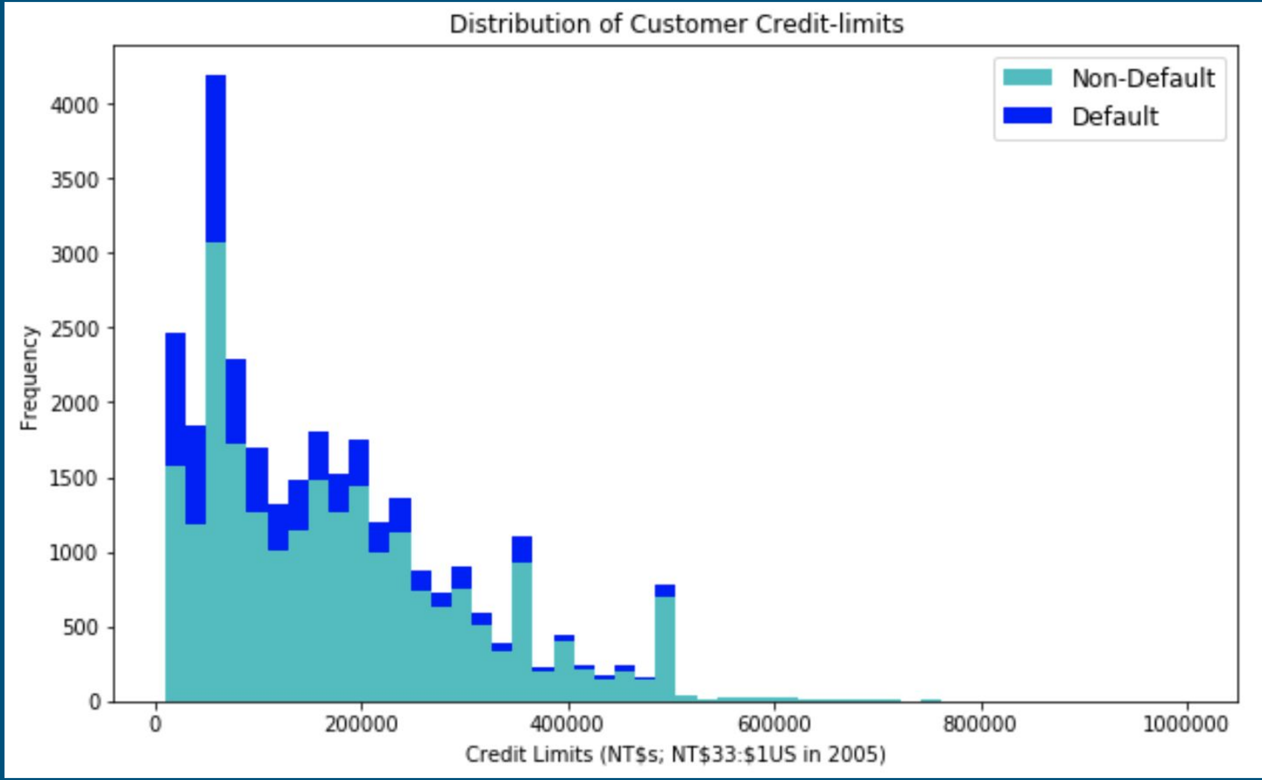
Demographic Category



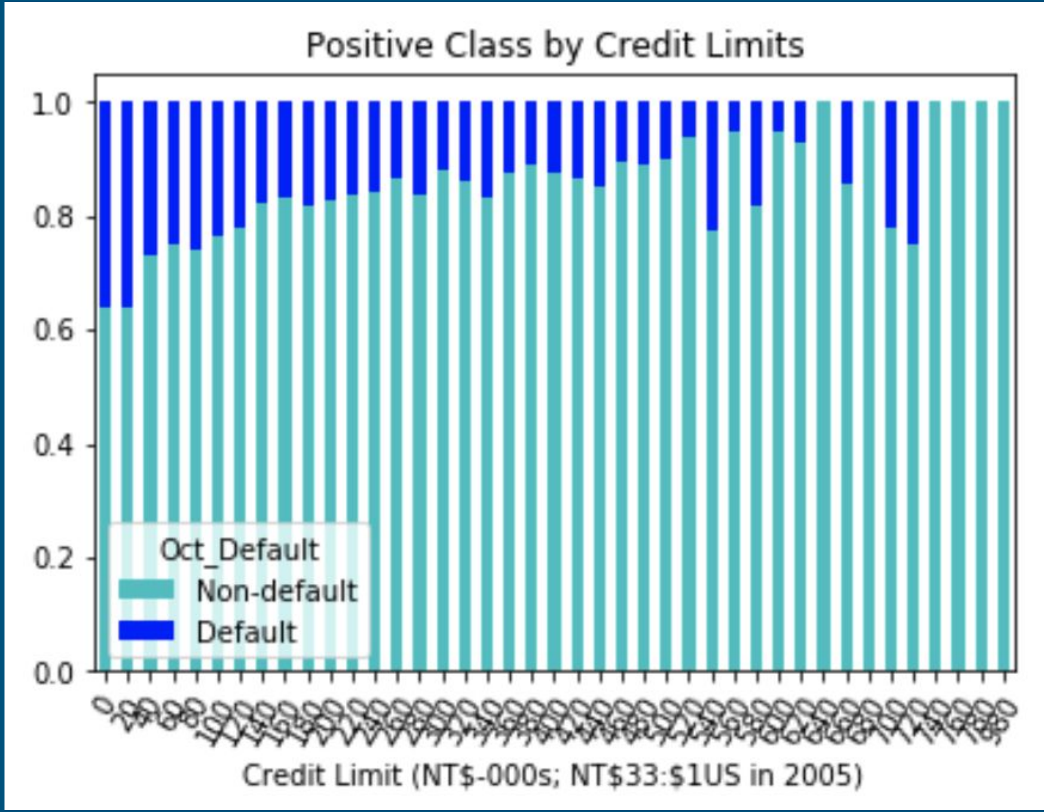
Demographic Category by Percentage



Credit-Limit Distribution



Credit-Limit Distribution by Percentage



Why Neural Networks?

1. Neural networks can model complex functions in nonlinear ways
2. They are useful when the relationship between inputs and output is unknown
3. They have been shown to provide superior results

Neural-Network Objective Functions

Mean-squared Error

Mean-absolute Error

Hinge

Squared Hinge

Binary Crossentropy

Categorical Crossentropy

“Keras”

Mean-squared_log Error

Categorical Hinge

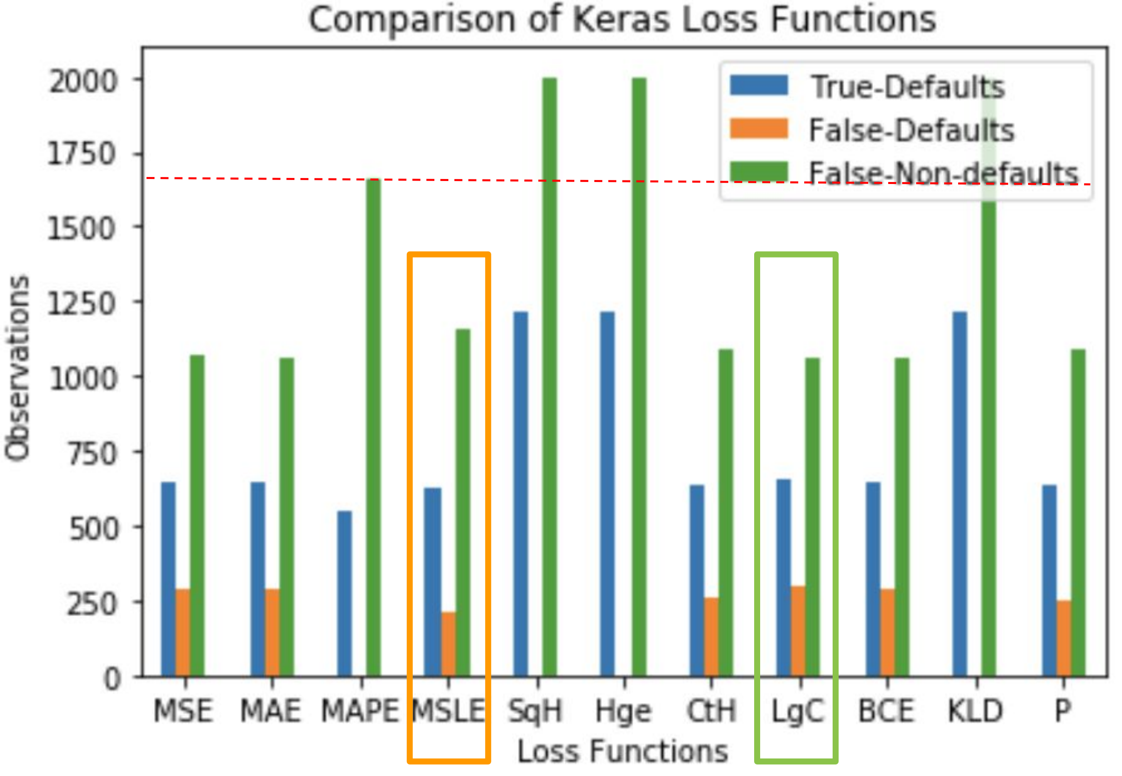
Log-Cosh

Poisson

Mean-absolute-% Error

Kullback-Leibler Divergence

How Keras' Available Loss Functions Differ



Thank you