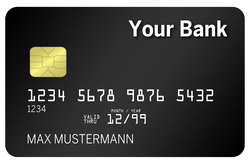


PREDICTING CREDIT CARD CLIENT CREDIBILITY

Michael McNamara

Oral: https://imperial.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=725ce8ae-8ee6-47fa-97ae-ad49006ce1e0

Imperial College London

Intro: What are we doing here?

In this project we will find a model that will allow us to predict whether someone will be able to pay back a credit card loan or if they will end up defaulting on it. This is incredibly useful especially in the credit industry as this allows a lender to score possible clients to work out if it is worth the risk of giving the loan.

What is this Dataset?

This dataset(1) documents the default payments of 30,000 customers in Taiwan along with information about their education, marital status and age.

LIMIT_BALANCE	SEX	EDUCATION	MARRIAGE	AGE	default payment next month
200000	2	3	2	34	0
260000	2	1	2	51	0
630000	2	2	2	41	0
70000	1	2	2	30	1
250000	1	1	2	29	0
50000	2	3	3	23	0
20000	1	1	2	24	1

1. **LIMITBALANCE**: How much the person borrowed
2. **SEX**: 1=male; 2=female
3. **EDUCATION**: 1=graduate school; 2=university; 3=high school; 4,5,6=others below high school
4. **MARRIAGE**: 1=married; 2=single; 3=others
5. **AGE**: Years
6. **default payment next month**: defaulted=1, did not=0

Adaptations of the Dataset

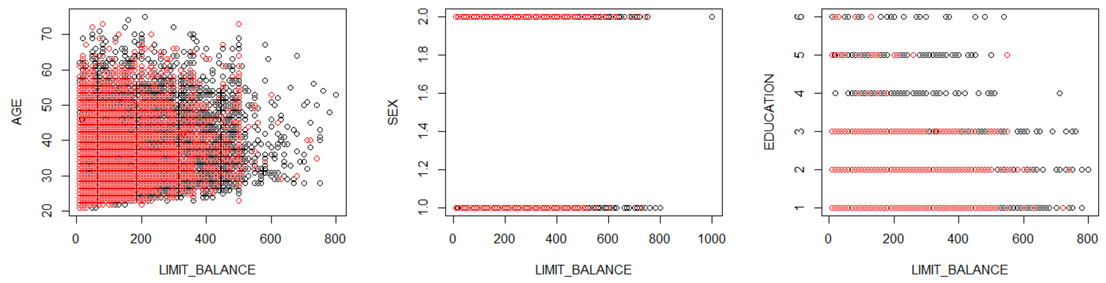
Feature removal: there were originally 24 different features in the data, but the majority were removed for simplification.

Undefined datapoint removal: some clients had an education value of 0 which is undefined, so I removed these clients from the set.

Anomaly removal: After plotting each feature, I removed datapoints that were too far away to be feasibly incorporated.

Choosing which Features to Compare

Here are some of the comparisons I made between different features of the dataset. A red point indicates a default on a payment and black one does not and each **LIMIT BALANCE** is in thousands:



What I was looking for when comparing these attributes was a nice, clear split between 'defaulted'(red) and not(black).

When I compared the amount borrowed (LIMIT BALANCE) with age, there was clearly a pattern of where red points lied, however they overlapped with black points so creating a good boundary would be difficult.

I then compared the amount borrowed with the gender of the customer. Interestingly, there appeared to be no significant impact of gender on whether a person would default or not so this also would not be a sensible choice of features.

Finally, I compared the amount borrowed against the level of education of the customer and noticed how disjoint the red and black points appeared to be. Therefore, I selected these 2 features to begin developing models that could be used for classification.

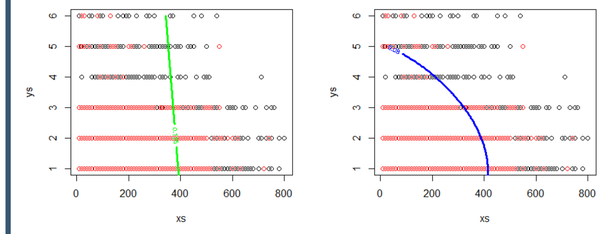
Method for Choosing the Best Model

The 2 methods of classification I will be comparing are called **LDA**(linear discriminant analysis) and **QDA**(quadratic discriminant analysis). **LDA** allows us to 'maximize the separation between multiple classes'(2). **QDA** is very similar except the 'covariance of each of the classes [do not need to be] identical'(3)

1. Split the dataset into a larger training set and smaller testing set using the sklearn module within python
2. Import the data subsets into R
3. Plot **Education** level against **Limit Balance**
4. Plot both an **LDA** boundary and **QDA** boundary into the data
5. Calculate the accuracy of how well each separated the data into classes
6. Select the more accurate classifier.

The Results

LDA (green); **QDA** (blue)



The following errors are calculated based on the percentage of coloured points on the incorrect side of the corresponding classifying boundary.

Model	Training Error	Testing Error
LDA	7.55%	7.73%
QDA	14.74%	14.96%

Conclusions

Observing the resulting errors of the **LDA** and the **QDA**, we can conclude that an **LDA** would be the preferable model to use in the case of this dataset. Therefore, we can predict whether someone would be able to pay back a loan just based on their education and the amount the client borrowed with an accuracy of 92.45To improve next time, I would try more classifying methods such as k-nearest-neighbour to look for any further decreases in error rate.

Real-World Applications

The main real-world application of this credibility model is credit scoring. Lenders need a way to fairly score potential borrowers to decide whether they are eligible for the amount of credit they are applying for. If they are below the minimum requirement, they will not receive the credit. This simple **LDA** model does not take into account discriminatory factors such as race, disability or religion which is extremely important both ethically and legally (in most countries).

References

[1] I-Cheng Yeh. 'default of credit card clients Data Set'. UCL machine learning repository, 2016-01-26.

[2] Sebastian Raschka. 'Linear Discriminant Analysis – Bit by Bit'. sebastianraschka.com, Aug 3, 2014.

[3] Bradley Boehmke. 'Linear Quadratic Discriminant Analysis · UC Business Analytics R Programming Guide'. 2020-03-29.