Default of Credit Card Clients: An ISM 4545 Business Intelligence Report

Published April 16, 2021 – 12 min read

By Analysts Rachel Arcangeli, Volga Acebes del Castillo

Credit card default payments can be both revenue and loss for card issuers, as interest accumulates but also risks debt never being paid. This report informs organisations of customer behaviour & possible strategies to increase revenue and prevent losses.

Overview

Opportunities & Challenges

- The credit limit of people likely to default must be on the lower side to balance losses in loans vs interest payments.
- The credit limit of people likely to pay back must be high, to balance risk of losses vs revenue from transactions.
- Marketing efforts should be focused towards clients with a graduate degree education, as they bring the highest revenue with an acceptable amount of risk.

What You Need to Know

- Previous behaviour is predictive of future behaviour, including default payments.
- People with high credit limits are likely not to default.
- People with a graduate degree aged 43 to 53 have the highest credit limit & least likelihood of default, whilst people with “Other” as education aged between 21 and 31 are most likely to default.

Business Understanding

Business Context

We are increasingly seeing how data analytics is changing many aspects of our lives, and many times without realising it. From loyalty programs to SEO, analytics is everywhere and is here to stay.

Data analytics provides businesses the power of information to be transformed into data and then knowledge, increasing efficiency and producing better results. If applied to the banking world, we can determine which characteristics are likely held by someone who will default on their credit card payments, and decision makers can make informed choices about these users, such as their maximum credit limit or customer segments marketing should target.

To default is failing to complete a payment on a debt by the due date. Credit card companies can take different actions; increasing interest rates, decreasing the maximum credit, or even legal action to enforce the payment.

User behaviour can be collected & used to create models that then predict said behaviour. Credit card issuers will find these useful because predictive analytics will allow them to decide if a person should even be approved for a credit card in the first place using an automatised process, allowing for a better allocation of resources and a better experience for applicants, as they would know sooner if they have been approved or not.

Business Questions

1. What are the main indicators a customer will default next month?
2. Which user creates most losses to the card issuer?
3. Which market segment should be targeted by marketing efforts?

Data Understanding

We selected out data from UCI Machine Learning Repository, by I-Cheng Yeh from Chung Hua University in Taiwan. It was originally used to compare the predictive accuracy of default payments among six data mining methods, and it was found that an artificial neural network worked best. We wanted to investigate this, thus we used different methods to analyse the data, including AUC & hold out method to evaluate the models.

Dataset Description

The dataset originally consisted of 30,000 records and 25 attributes (more were derived for analysis purposes):

1. ID
2. Credit line in New Taiwan dollar: includes individual consumer credit & their family's supplementary credit.
3. Sex: 1 = male, 2 = female
4. Education: 1 = others, 2 = high school, 3 = university, 4 = graduate

   *Others may include no education, or over graduate level. Not explained clearly in original data.
5. Marital status: 1= married, 2 = single, 3 = other
6. Age in years

7. – 12. History of past payments: Pay_1 = repayment status in September 2005, Pay_2 = August 2005, …, Pay_6 = April 2005

*Measurement scale: -1 = paid on time, 1 = payment delayed 1 month, 2 = delayed 2 months, …

*Summarised on new columns: "Late_Pmt_Sum" and "Late_Pmt_Avg"

13. – 18. Amount of bill statement in NT$: Bill_Amt1 = amount of bill statement in September 2005, Bill_Amt2 = Aug 2005, …, Bill_Amt6 = April 2005

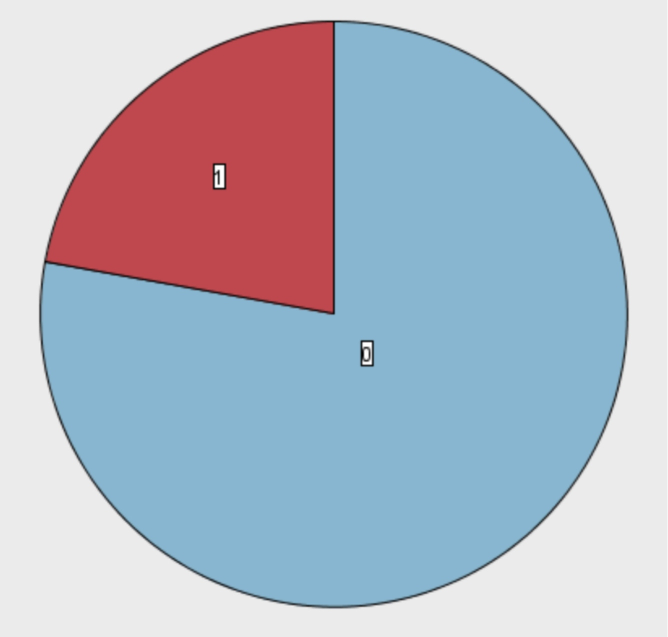
*Summarised on new columns: "Bill_Amt_Sum" and "Bill_Amt_Avg"

19. – 24. Amount of previous payment in NT$: Pay_Amt1 = amount paid in September 2005, Pay_Amt2 = Aug 2005, …, Pay_Amt6 = April 2005

*Summarised on new columns: "Pay_Amt_Sum" and "Pay_Amt_Avg"

25. Default payment: 1 = yes, 0 = no

Data Summary



Most customers (77.88%) did not default on payments, whilst 22.12% did.

Figure 1

There are more female customers than male, but a nearly equal amount of default payments for both sexes.

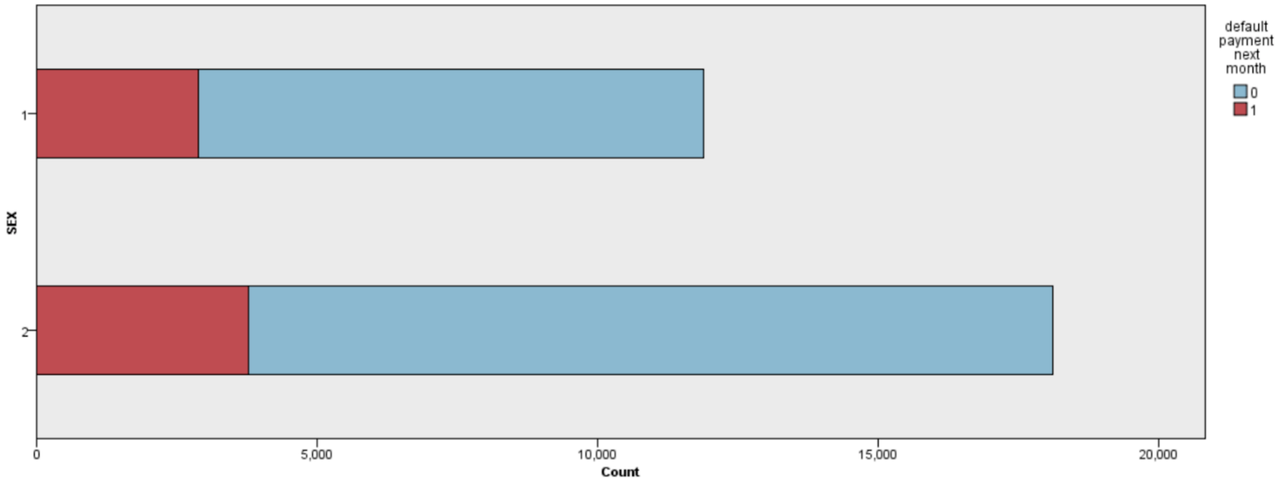
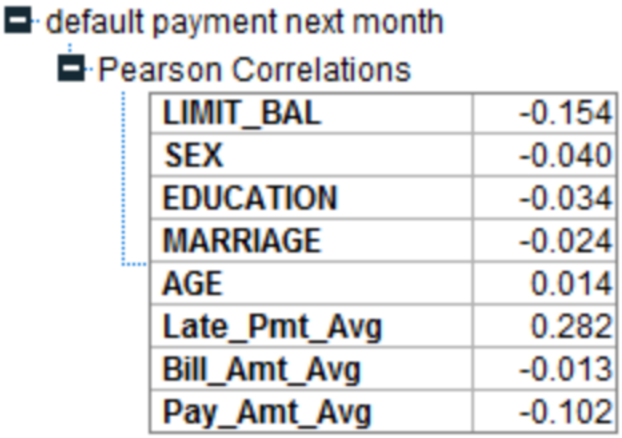


Figure 2



LIMIT_BAL	-0.154
SEX	-0.040
EDUCATION	-0.034
MARRIAGE	-0.024
AGE	0.014
Late_Pmt_Avg	0.282
Bill_Amt_Avg	-0.013
Pay_Amt_Avg	-0.102

Figure 3

When examining Pearson Correlation between default payment and the other attributes, the average of late payments (Late_Pmt_Avg) clearly is the biggest tell sign for predicting defaults, followed by balance limit (LIMIT_BAL).

Data Preparation

1. Removed the first row of miscellaneous titles from Excel.
2. Transformed from .xls to .csv file.
3. Education transformed using Excel’s “Find and Select” to fit ordinal type better: Graduate 1 to 4, University 2 to 3, High School 3 to 2, Other 4 to 1 (includes 0, 5, and 6 values, their meaning is unknown)
4. Pay_0 changed to Pay_1 to fit the format of the other labels.
5. Changed default payment role from input to target attribute.
6. Filter out ID, as it is irrelevant to analysis.
7. Changed Sex and Default payment types from continuous to flag.
8. Changed Marriage type to nominal.
9. Changed Education & Pay 1-6 to ordinal type.
10. Performed data audit to asses data quality.
11. Partition node used for the hold out method model evaluation. 70% used for training & 30% used for testing.

Business Question 1

What are the main indicators a customer will default next month?

For the first model, a CHAID decision tree was created, as it is a fast multi-way tree algorithm that explores data quickly and efficiently. To avoid overfitting, tree depth was modified to 3 levels.

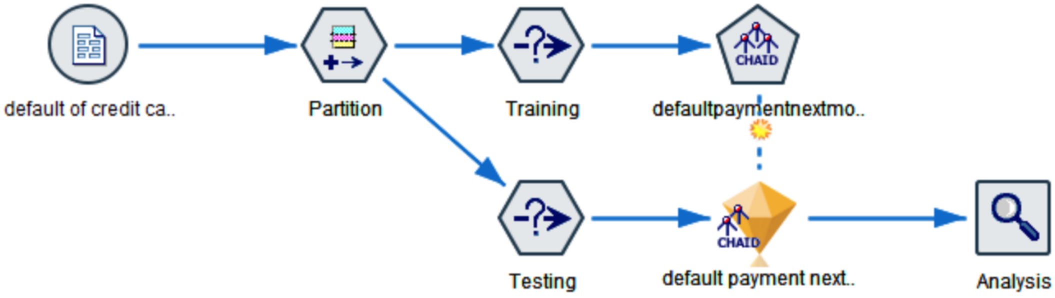


Figure 4

The largest default node had 1,501 users:

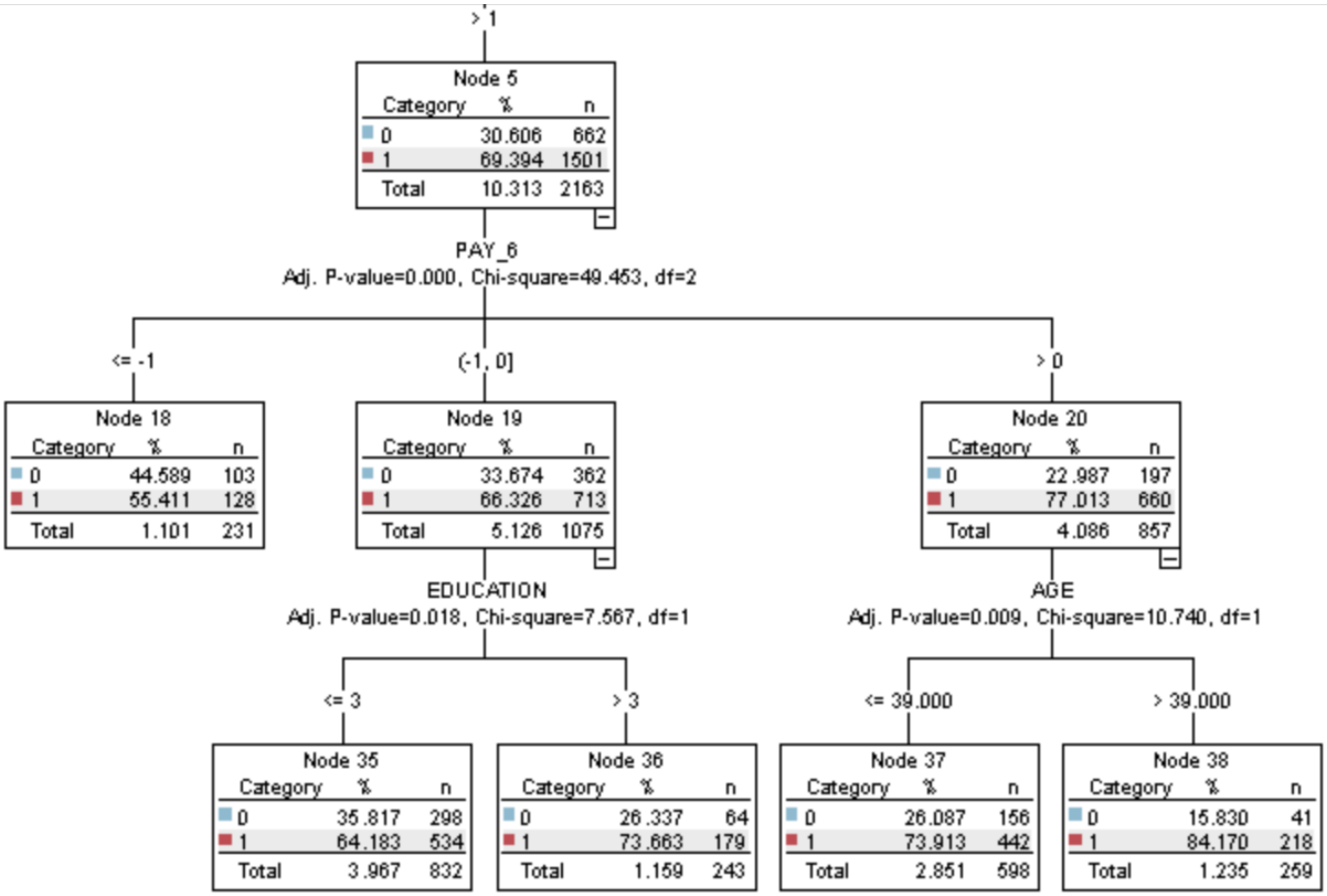


Figure 5

The main indicator for future defaults (1 or more months late) is Pay_1, followed by Pay_6, and age. These users paid late by at least one month in September 2005, they also paid late by at least one month in April 2005, and were aged 39 or younger. Model evaluation showed it has good accuracy, with 81.08% of predictions correct & AUC of 0.748.

For the second model, we used an artificial neural network, following the original data's recommendations:

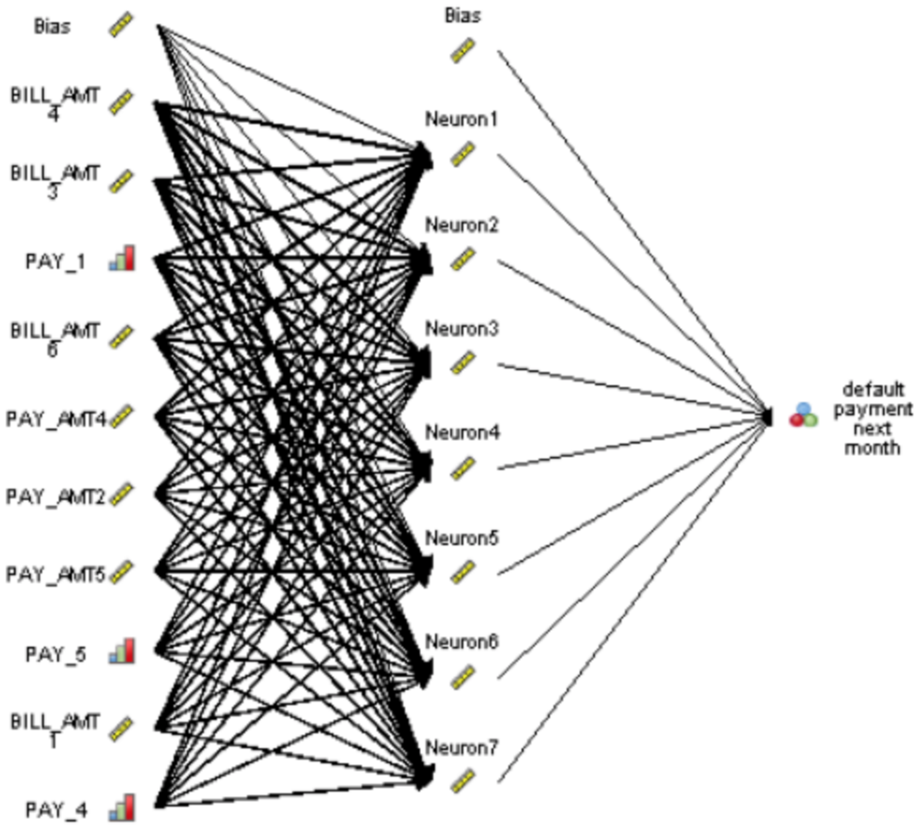


Figure 6

Predictor importance showed Bill_Amt4 as most relevant, followed by Bill_Amt3, and Pay_1. Pay_4 was the least relevant.

Model evaluation showed very similar, albeit better, results to the decision tree, with 80.96% of predictions correct and AUC of 0.764.

The business benefits from these observations by applying them to decisions involving maximum credit balance & interest rates applied, as credit cards make most revenue from interest payments, but default losses have to be cut too.

Business Question 2

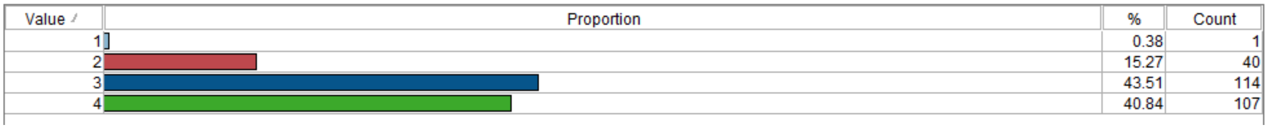
Which user creates most losses to the card issuer?

Analysing “Late_Pmt_ Avg”, users usually pay on time, but 5.89 standard deviation means that there is a high variability from the mean. “Unpaid total” was then created to find which users were causing the largest loss of revenue; found by subtracting total paid amount from total bill amount.

Unpaid total & Late payment have a 0.308 correlation, meaning the later the payment is, the more is owed, and the less likely they are to pay it back. It is possible that it generates high interest payments, but it is also likely that the issuer will have to resort to legal action, leading to more losses over time.

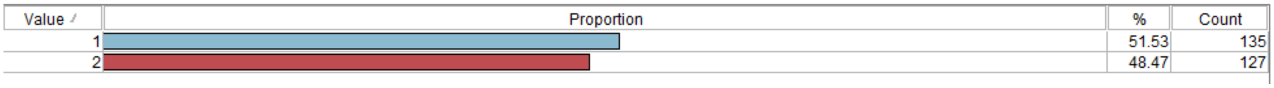
To find this data, users with at least 6 months of unpaid balance & over 722,238 NT$ were considered, as most people are below these amounts when following a binning procedure to find groups of users.

Applying these filters, we find that unpaid total amount had the highest correlation with age and sex, and that late payments had the highest correlation with sex and education, thus these attributes are evaluated further:



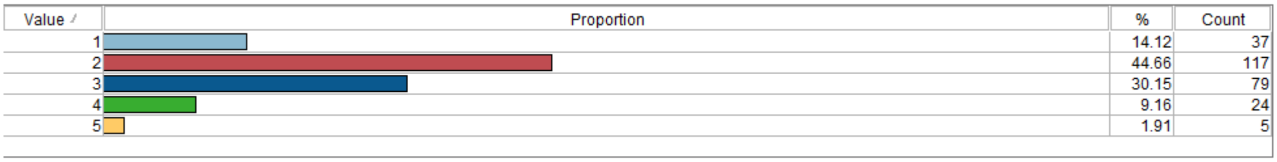
Value	Proportion	%	Count
1	🟦	0.38	1
2	🟥	15.27	40
3	🟦	43.51	114
4	🟩	40.84	107

Figure 7 Education



Value	Proportion	%	Count
1	🟦	51.53	135
2	🟥	48.47	127

Figure 9 Sex



Value	Proportion	%	Count
1		14.12	37
2		44.66	117
3		30.15	79
4		9.16	24
5		1.91	5

Figure 8 Age

In conclusion, credit card issuers must be careful approving cards for customers that are males aged between 32 and 42, who have education of High School or Other, as they may decrease the bank’s revenue by not paying large bills.

Business Question 3

Which market segment should be targeted by marketing efforts?

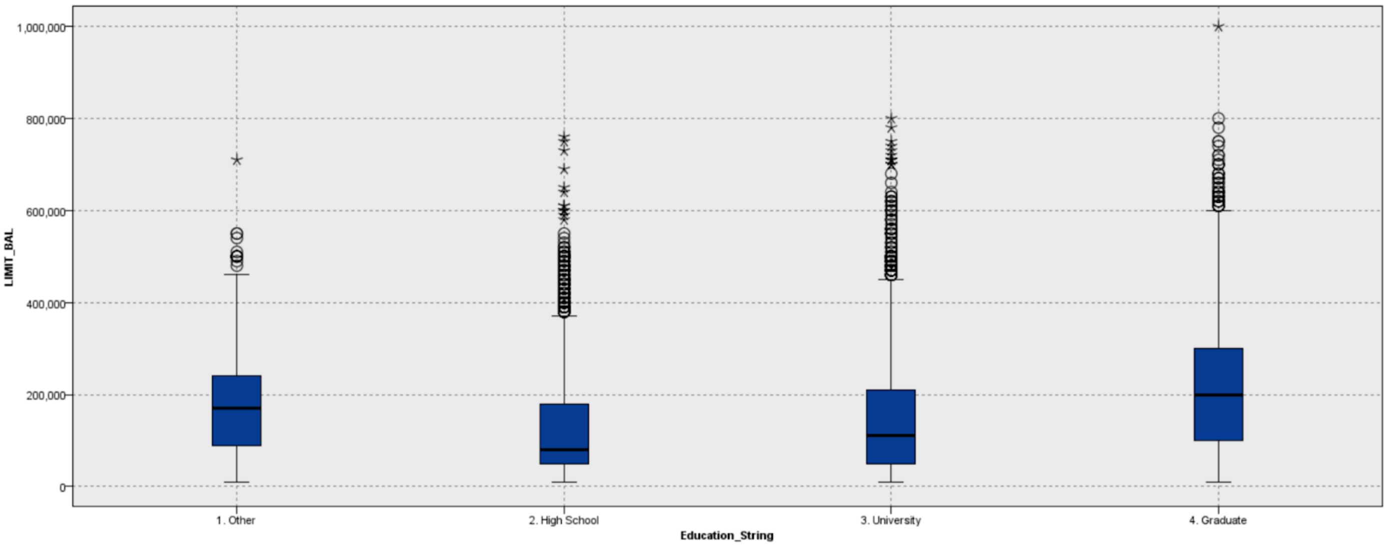
Conversely, it is also interesting to find those users that will ensure the bank higher revenues, as this segment should be targeted by advertising & marketing to attract and keep these customers. These departments would find these users using demographic data available in the census to for example, place local banners, send specific physical mail, or create local promotions. The attributes used will therefore be age, marriage, and education, and they will be correlated to high credit and few late payments.

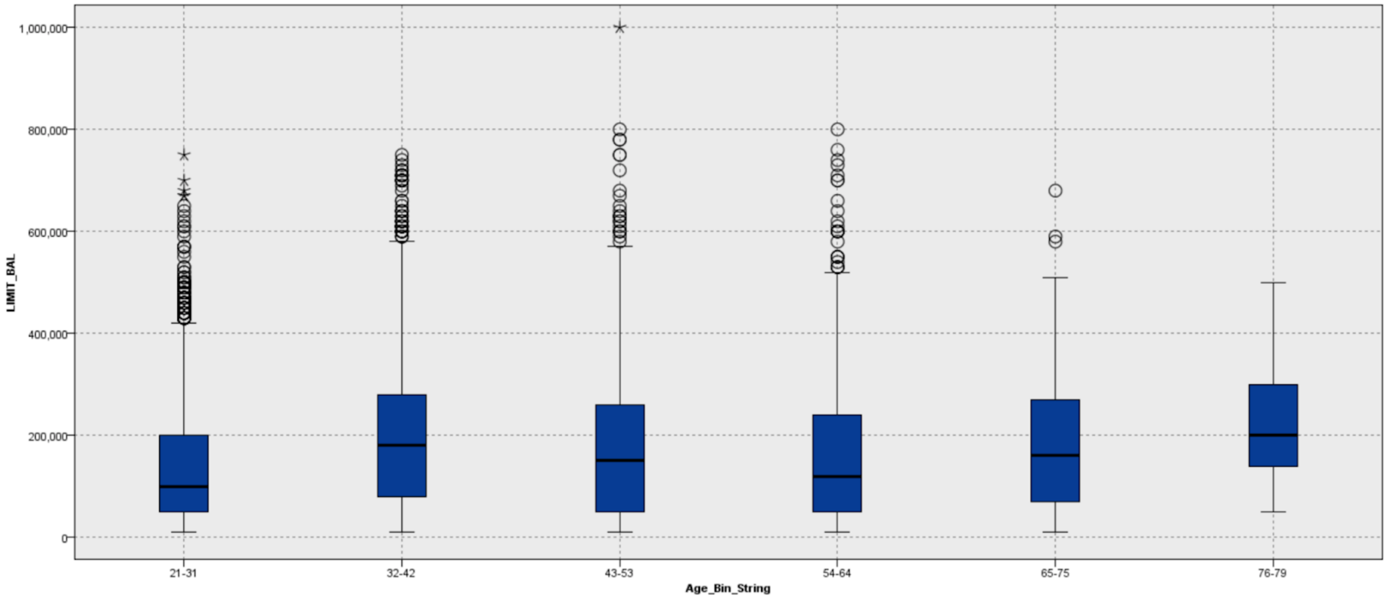
Age groups were created by binning as follows: 1 = 21-31, 2 = 32-42, 3 = 43-53, 4 = 54-64, 5 = 65-75, and 6 = 76-79.

Correlations:

Attribute	Education	Marriage	Age
High credit	0.231	-0.108	0.119
Few late pmts	-0.103	0.035	-0.047

Education with high credit & Age with high credit show the highest correlation, thus groups with these characteristics should be targeted, and they were found next.





In conclusion, graduates aged 43 to 53 should be given higher credit limits, as they will not default, thus bring in revenues without a higher risk.

Executive Summary

The aim of this Business Intelligence Report was to examine the credit card issuer's past data on default payments in order to predict future default payments, and thus make recommendations to increase revenue through credit payments, as well as prevent approving customers that will likely default & not pay for long periods of time.

We used multiple methods to determine which attributes were significant indicators of future default payments, which characteristics of customers are significant & undesirable for credit card approval, and which characteristics of customers were significant and desirable for high credit limits. These methods include a CHAID decision tree, an artificial neural network, box plots, and correlation values.

Through our analysis, we found that:

- History of past payments is the main indicator for future defaults
- Pay_1 & Pay_6 were the most significant months of payment history that determine whether a customer makes a default payment next month.
- Age & Education have the highest impact on credit limit, with those aged 43-53 with a graduate degree having the highest credit limit.
- Males aged 32-42 with an Education of "Other" are most likely to impact revenue negatively.

Solutions to increase revenue & decrease costs:

- Invest in personal finance tools for customers that are negatively impact revenue with very late & high unpaid bills. They could be offered after a small amount of unpaid bills accumulate.
- Directly call customers with multiple late payments to avoid default payments.
- Maintain low credit limits for young and less educated customers to ensure their financial success by avoiding large unplayable debts.
- Target graduates aged 43-53 to find customers that can have a high credit limit with few debts, as they will make many purchases that translate to usage fees for the firm.