Project Report

FE 582 – Foundations of Data Science

Title: Prediction of Credit Card Default

Team Members

Simran Jariwala, Anirudh R Sundararaghavan, Ramesh Krishnan Balasubramani

Abstract

This report analyzes the relationship between demographic factors and trends in repayment of credit card with risk of default of credit card. The objective of this paper was to understand which variables are strongest predictors of credit card default and train a predictive model to identify risk of credit card default.

Our analysis showed that education, age, default trend and latest default status are the strongest predictors of default under various models. We compared the accuracy rate and recall rate of multiple machine learning algorithms to determine which model performed the best. Compared to all other training models, a Quadratic Discriminant Analysis model gave the highest recall rate while maintaining a good accuracy rate.

1. Introduction:

Default risk is the probability that a company or individual will be unable to make the required payments on their debt obligation. Default risks represent a significant problem to banks and the economy as a whole. Every instance of default results in a financial loss for the bank impacting the profitability, solvency and share price of the bank.

Banks are exposed to default risk across their business divisions from home loans, mortgages to credit card lending. Predicting accurately which customers are most probable to default represents significant business opportunity for all banks.

Credit card lending is one of the major consumer lending products in the United States, representing roughly 30% of total consumer lending (USD 3.6 trillion in 2016). Furthermore, studies have shown that 15 percent of American families are living beyond their means and are spending more than they receive. As credit cards have a high annual percentage rate, this means many Americans are seeing their debts compound and grow at a staggering rate. on a monthly basis increasing risk of default daily.

Therefore, the ability to predict which customers are more likely to default on their credit cards or identifying factors that are strongest predictors of credit card default can significantly help banks protect against default risk.

2. Research Questions and Objectives

- How does the probability of default payment vary by categories of different demographic variables?
- Which variables are the strongest predictors of default payment?
- Train a predictive model to identify future default

3. Data Description[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.

We have a total of 30,000 rows (i.e. customer details and we have 24 features about each customer.

We have the following features with respect to each customer and the scale of each feature is also provided below:

3.1 Demographic Factors

- Limit Balance - Amount of given credit in NT dollars
- Gender (1=male, 2=female)
- Education: 1=graduate school, 2=university, 3=high school, 4=others
- Marriage: Marital status (1=married, 2=single, 3=others)
- Age: Age in years

[1] *Source:*
Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. The original dataset can be found here at the UCI Machine Learning Repository.

3.2 Repayment Details

- Repayment Status

  PAY_0: Repayment status in September, 2005
  PAY_2: Repayment status in August, 2005
  PAY_3: Repayment status in July, 2005
  PAY_4: Repayment status in June, 2005
  PAY_5: Repayment status in May, 2005
  PAY_6: Repayment status in April, 2005

  -2 = No consumption;
  -1 = Paid in full;
  0 = The use of revolving credit;
  1=payment delay for one month,
  2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above

- Amount of Bill Statement

  BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
  BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
  BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
  BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
  BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
  BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

  The range of Amount of Bill Statements is between (-)170,000 to 1,664,089. The negative balance could be a result of overpayment by customers in the previous payment. Refer point a) below.

- Amount of Previous Payment

  PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
  PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
  PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
  PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
  PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
  PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

  The range of Amount of previous payments vary between 0 to 0 1,684,259.

- Next Month Default Status - 1=yes, 0=no

3.3 Key Notes about the dataset

a) Repayment Status Feature

The Pay_X columns provide details of repayment status of customers as at the end of each month. It should be noted that the while repayments may be made in bulk, the number of months of delay in month cannot make jumps more than 1 since a customer can delay payment only once per month.

b) Excess over limit balance

For the purpose of this dataset we have assumed that situations where the bill amount exceeds the limit is not an error and that the bank does allow the customers to exceed balance in exceptional circumstances.

Consequently, the bill amount may also be negative in the subsequent months

c) Monthly payment excess

It is also to be noted that the monthly payment may exceed the bill amount as this could be additional payment for future months.

d) Key Variables

Education – Level of education may play a key role in risk of default as a higher education is likely to ensure constant income and consequently, less risk of default.

Marriage – Married Status may not be a very important variable as marriage could have a dual effect – a) marriage could indicate a state of stability in income and career and hence less risk of default or b) marriage may lead to higher expenditure due to requirement to finance a bigger family.

Repayment Status – Increasing cumulation of number of counts of delay would be a very key indicator of default in upcoming months

Amount of previous payment – Where there is a shortfall in the previous payment or where this are consecutive lower payments, these may be signs of upcoming default

<<This space has been left blank intentionally>>

4. Data and Methods

4.1 Data Transformation

While the dataset provides details with respect to the bill statement and amount payment of the clients. This data directly does not provide any information on the default risk of the customers. Given the same, we analyzed the way we could use these variables to provide information on risk of default.

Payment Proportion

We matched the bill amount and repayment amount of each month and using this we created a new variable – *'Payment Proportion'* for each month. We can now observe the amount each customer pays as a proportion to his bill amount. Where there is very low proportion this may indicate that there is a higher risk of default.

Payment and Default Trend

Since we had information on payment proportion and default of recent months, we used this information to create two new variables for each customer – *'Payment Trend'* and *'Default Trend'* based on the repayment behavior in those months.

We used the below methodology to classify customers into different categories based on repayment behavior in the past.

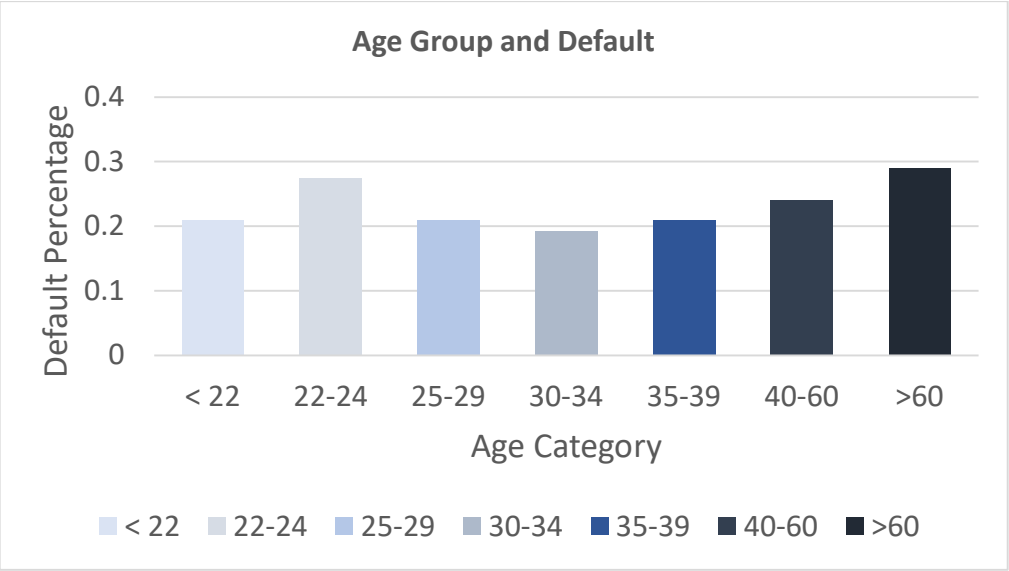
Step 1 – Compute month-wise increase or decrease in default/ repayment proportion
Step 2 – Summarize the number of increases/ decreases
Step 3 – Use the below table to categorize customers based on number of increases and decreases.

S No	Decision Point	Category
1	Where there is only decrease	Negative
2	Where majority of the months there is a decrease in payment	Majorly Negative
3	Where there are equal amounts of increase and decrease	Mixed
4	Where majority of the months there is an in increase in payment	Majorly Positive
5	Where there is only decrease	Positive
6	No Increase or Decrease	No Trend

<<This space has been left blank intentionally>>

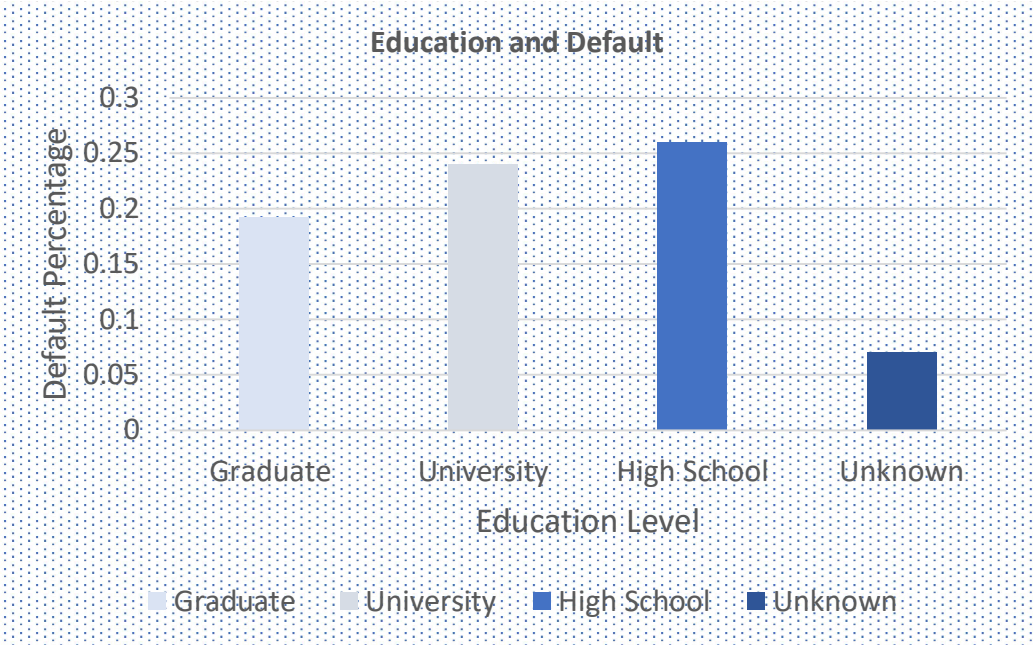
4.2 Exploratory Data Analysis

Relationship between Age Group and Default



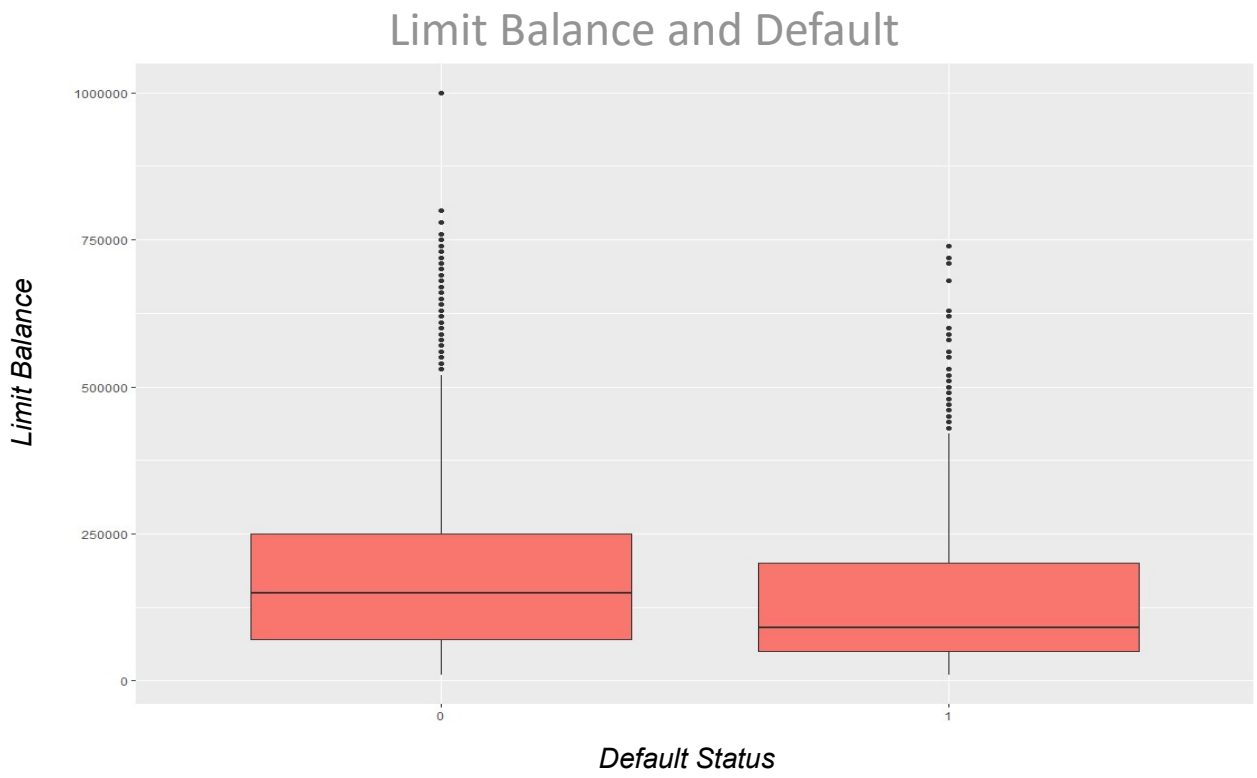
- Consolidating age groups into generations, there are some unique patterns pertaining to each group. The Default payments for age groups greater than 60 and 22-24 seem to be highest.
- The defaults for less than 22 is average since they focus more on building a good credit history whereas for holders that belong to age groups from 30-34 have a more mature credit history, more stable and a good salary progression.

Relationship between Education and Default



- Lack of education does play a significant role to predict credit card debt. People with education level up to high school have the highest defaults.
- There is a discrepancy as financial literacy increases education level which gives access to jobs quickly thus financial stability and ample savings.
- This is clearly visible from the graph where people with higher education level have lowers defaults thus, there lies an inverse relationship between the two.

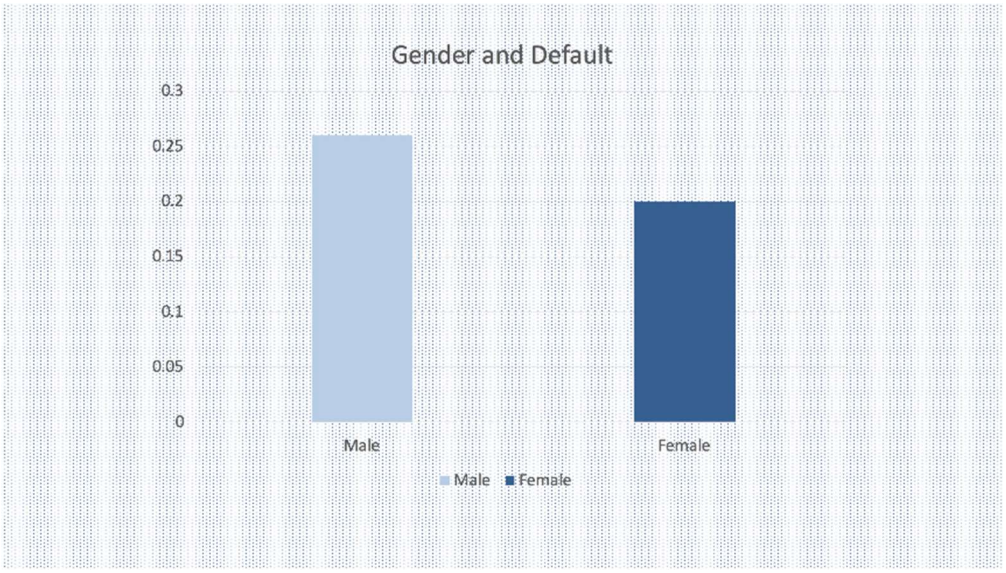
Relationship between Limit Balance and Default



- Limit Balance is the most significant predictor to predict credit card defaults.
- It is interesting that customers with lower limits tend to default the most. Whereas customers with higher limit balance generally default less because of high potential.
- Banks usually tend to disregard these customers because of lower balances and focus more on the higher side but these small defaults tend to pile up, affect overall credit management for lenders and cause major bankruptcies.

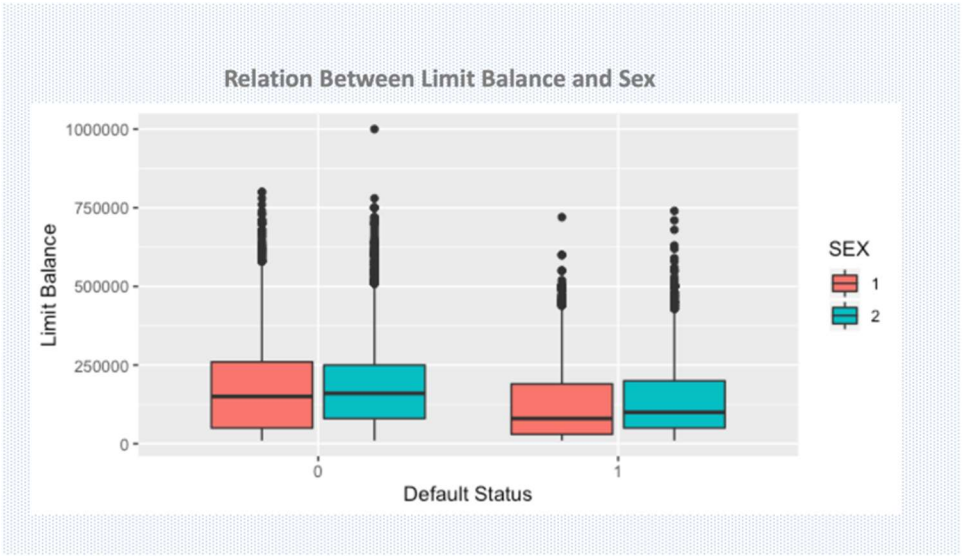
<<This space has been left blank intentionally>>

Relationship between Gender and Default



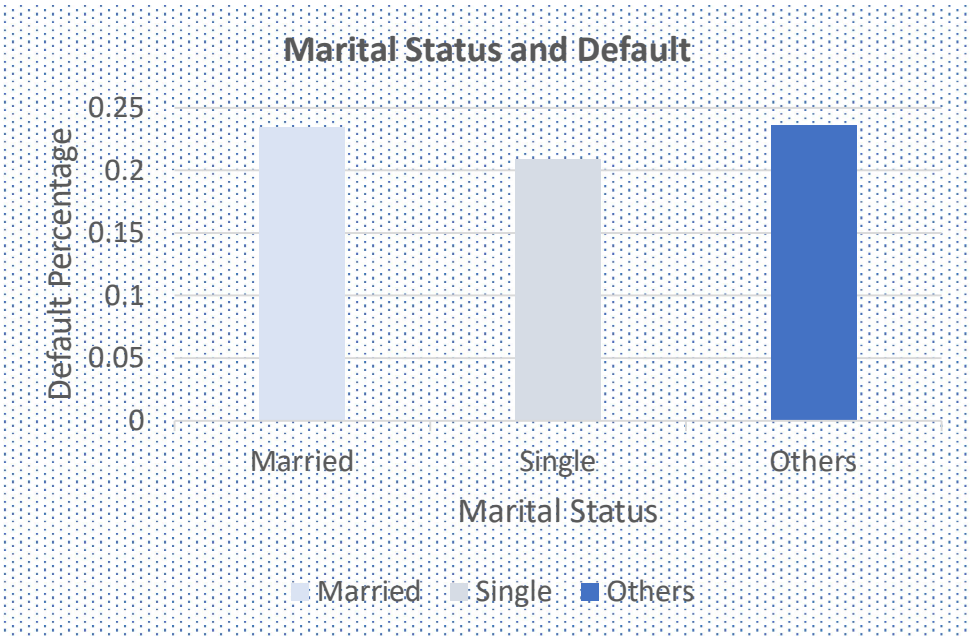
The above graph shows that male tend to default more than female.

Relationship between Limit Balance, Sex and Default



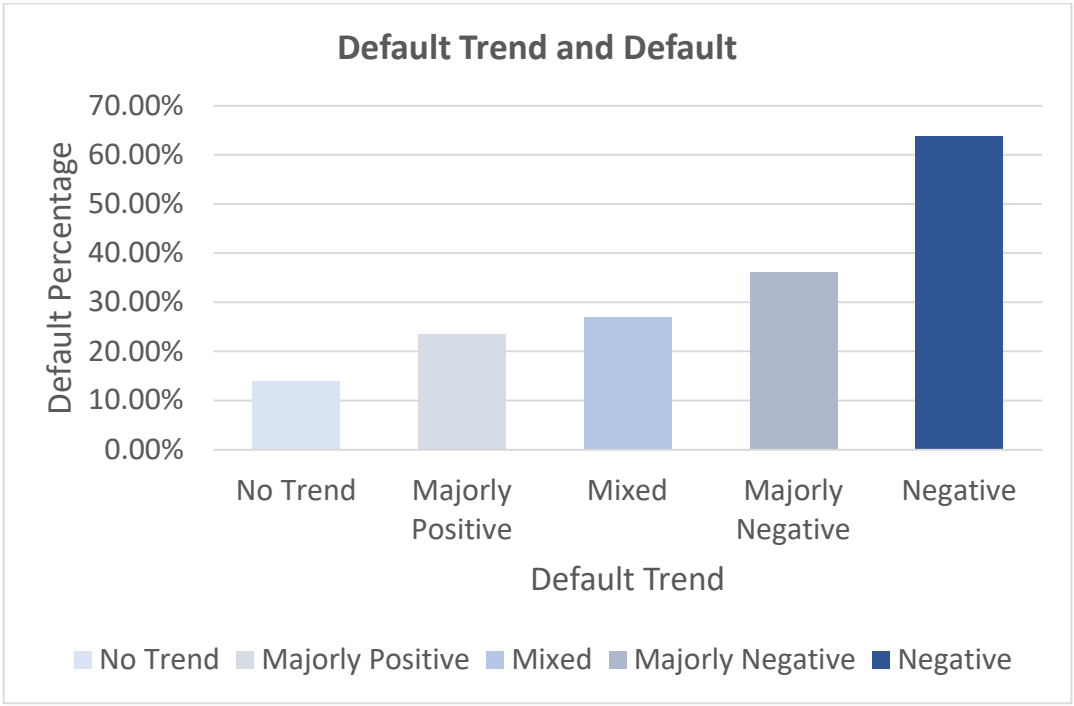
We can observe that among the people who default, limit balance of female tends to be slightly more than male.

Relationship between Marital Status and Default



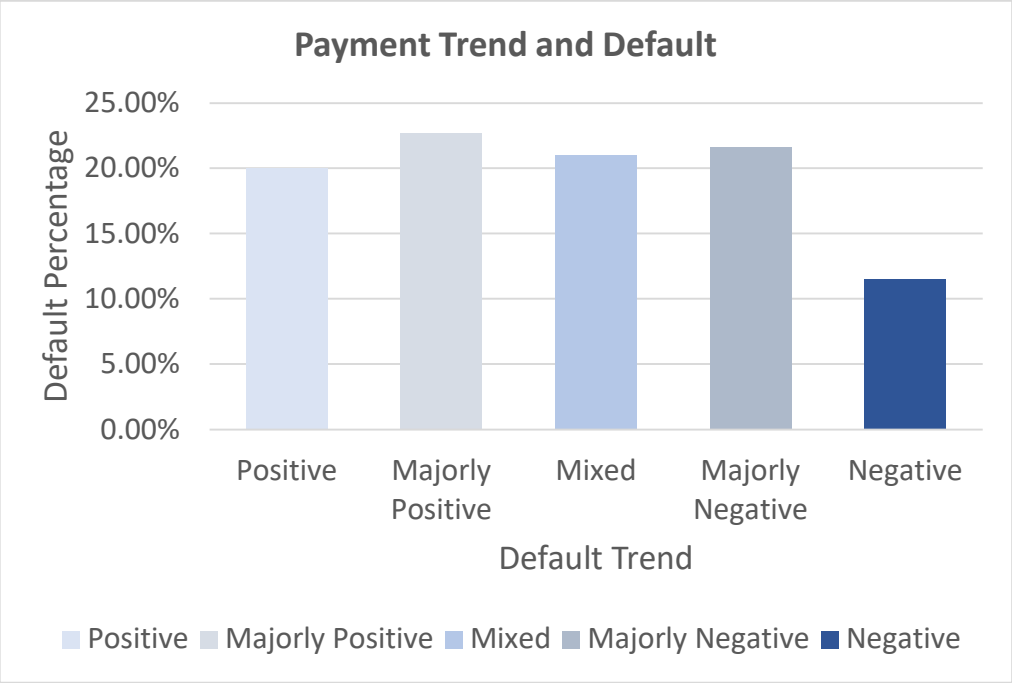
We can observe that while people who are single tend to have less default rate than people with a different marital status, the difference is minimal, and the default rate seems to be consistent around 20%.

Relationship between Default Trend and Default



As mentioned above, we created this variable called default trend based on the past months data of the customers. As we can observe from the above graph the percentage of default steadily increases from no trend to negative. This shows that default trend has a clear impact on future default

Relationship between Payment Trend and Default



Unlike default trend, payment trend does not show any clear indication of default. We can observe that majorly negative and majorly positive payment trends seem to exhibit high percentage of default whereas negative trend contributes to a low percentage of default which is contrary to the belief that negative payment trend contributes to high percentage of default.

<<This space has been left blank intentionally>>

4.3 Our Analysis

Upon completion of the exploratory data analysis, we understood the nature of relationship of each variable with the risk of default. We decided to use various statistical methods to test the statistical significance of each variable.

Logistic Regression

We selected different combinations of variables and fit logistic regression models to identify which variables were statistically significant. We categorized variables into demographics, payment related, proportion related etc. and ran regression models with each group of variables to identify which variables were the best predictors.

Stepwise Logistic Regression Model

Next, we ran a stepwise logistic regression model to identify the predictors that provide an optimum prediction model. A stepwise Logistic Regression Model uses AIC (Akaike information criterion) to determine the optimum model.

AIC provides a score based on the value of residuals and includes a cost complexity factor which penalizes the model for every additional predictor variable. Thus, unless a predictor variable reduces residuals more than the penalty incurred on addition of the variable it shall not be included in the model.

The stepwise model follows the following methodology:

- The algorithm starts with a null model and computes the AIC
- The algorithm then computes AIC posting add each of the variables and picks the variable that has the lowest AIC
- Similarly, at each step the algorithm computes the AIC for each combination of addition and removal of variables.
- At the point, where there is no further reduction in AIC, the algorithm stops as the optimum model with lowest AIC has been reached

Classification Tree

We also ran a classification regression tree algorithm to identify which predictor variables were crucial in decision making. A classification tree attempts to split the predictor space into various regions based on which decision points in each predictor which have most impact on the outcome. This is done using recursive binary splitting.

The end output is a decision tree which splits the data based on various decision points in the important predictor variables. Thus we can observe which are the predictor variables which are useful in classifying a customer as high risk of default or not.

Other Methods

The methods discussed above are not only used to build predictive models but also help us understand the importance of each variable. Given this advantage of interpretability we use the above methods to build models and understand relationships.

Now we shall move on to more complex methods which have slightly lower interpretability but have good predictive power

We shall also use the following methods to build models to predict risk of default:

Random Forest

Random forests is an advanced form of classification trees where the algorithm builds separate prediction models from many data sets and averages the resulting predictions. Further, random forests also help in decorrelation of the models by randomly selecting only a few of the predictor variables to build the tree at each node.

Linear Discriminant Analysis and Quadratic Discriminant Analysis:

LDA and QDA use Bayes Theorem and conditional probability to predict the class of the outcome based on the assumption that the distribution of predictors X are approximately normal. The difference between LDA and QDA is that LDA assumes equal variances between classes while QDA does not.

5. Results

Logistic Regression

As per the regression model, the following factors were statistically significant

- Limit Balance
- Education
- Default trend
- Marriage
- Recent Payment Proportion and
- Recent default status

Classification Tree

The classification tree algorithm provided very short tree classification tree which considered only the following variables:

- Latest default status and
- Limit balance.

On further analysis, we understood that no other variable gave a more significant information gain while improving accuracy power of the model

Other Training Models

In order to compare performance of the various models we used the following performance parameters:

- Accuracy Rate
  Accuracy rate is the percentage of total correct predictions against the total testing data

- Recall Rate
  Recall rates is the percentage of total correct default predictions against the actual defaults in the testing data

From a bank's perspective the recall rate is very significant as it computes the number of correct predictions of default. However, it should be noted that a high recall rate at the cost of a lower

accuracy rate may not be beneficial. This would mean that while the bank is correctly predicting defaulters and not issuing credit cards, the bank is also turning away credit worthy customers who may not default but are categorized as defaulters due to lower accuracy rate.

Every bank would need to be have identify a balance between accuracy and recall based on the risk appetite of the respective bank.

Summary of Model Performances

Method Used	Accuracy Rate	Recall Rate
Logistic Regression (Stepwise)	82.03%	25.75%
Classification Tree	83.07%	32.54%
Random Forest	83.17%	37.05%
LDA	82.28%	30.81%
QDA	69.93%	67.54%

Of all the models, QDA has a significantly higher recall rate while maintaining a decent accuracy rate of 69.9%.

6. Way Forward:

This analysis gives us an indication of which factors are strong predictors of default risk and how to identify them. However, this is the starting point we would need to build more robust predictive models. We believe that the following should be the focus points

- Importance should also be given to more data collection. 30K customers is a very small number when it comes users of credit card.
- Data Collection would also need to be more detailed as features such as nature of occupation and credit scores of customers would be important factors in predicting default
- Further, detailed data with respect to repayment by customers from issuance of credit card up to present would help in identifying trends in lifecycle of credit cards that may indicate risk of default
- The learning from these analyses can also be extended to predicting defaults in home loans, mortgages etc.