Prediction of Credit Card Default

By : Anirudh Sundararaghavan, Ramesh Balasubramani, Simran Jariwala

INTRODUCTION

- Default risk is the chance 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.
- 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.

Objectives:

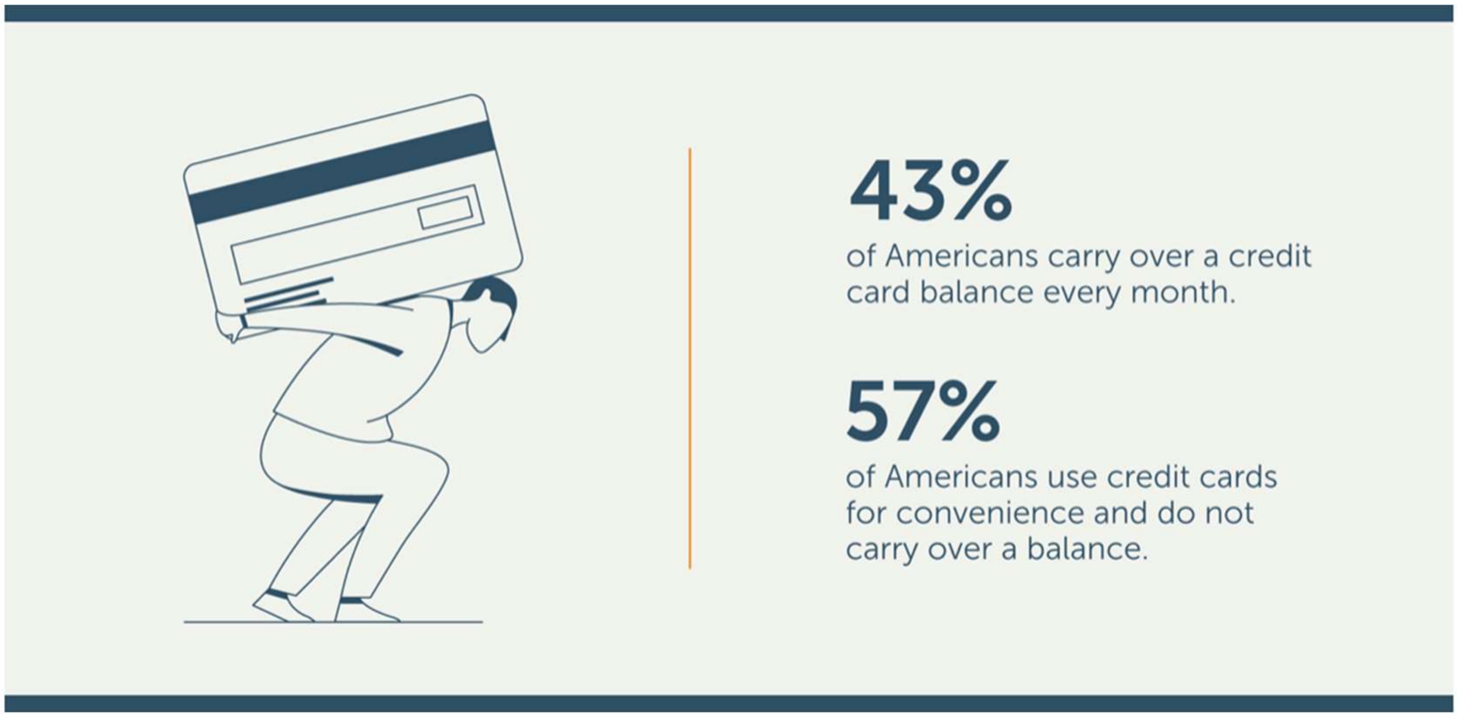
✓ Relationship between probability of default payment and different demographic variables?

✓ Identify which variables are the strongest predictors of default payment?

✓ Train a predictive model to identify future default

STATISTICS

- 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.



DATASET

- This dataset contains information on credit card 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).

Limit Balance	• Credit limit of customers
Sex	• Female or Male
Marital Status	• Single, married or others
Age	• Age of customers
Repayment Status	• Defaults by customers in previous months
Education	• Graduate, University, High school or others
Bill Statement	• Amount due in previous months
Previous Payment	• Amount paid in previous months

- *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.

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 clients.
- Hence, we created a few transformed variables which categorize customers based on the trend in repayment status and proportion of bill repaid.
- Using R, we split the customers into the following categories based on trend in repayment and proportion paid.

Trends in Repayment Status and Proportion

Majorly Positive

Majorly Negative

No Trend

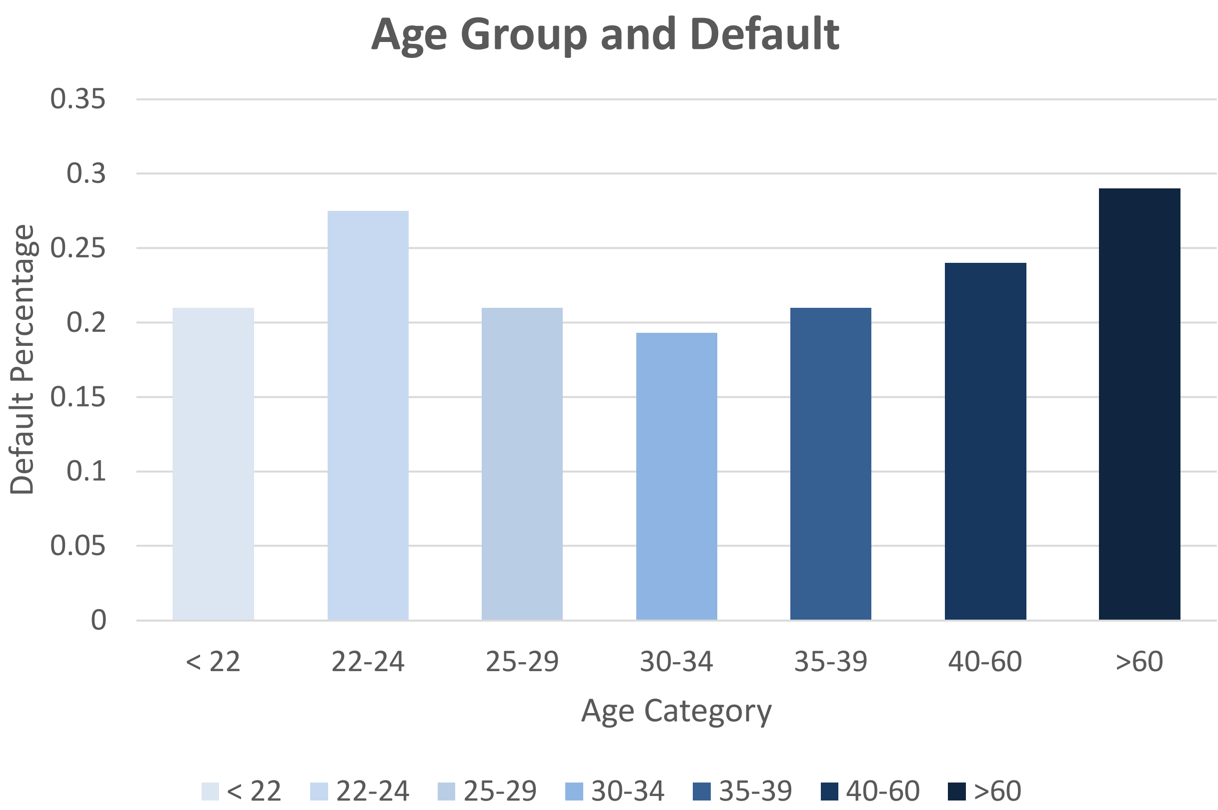
Mixed

Positive

Negative

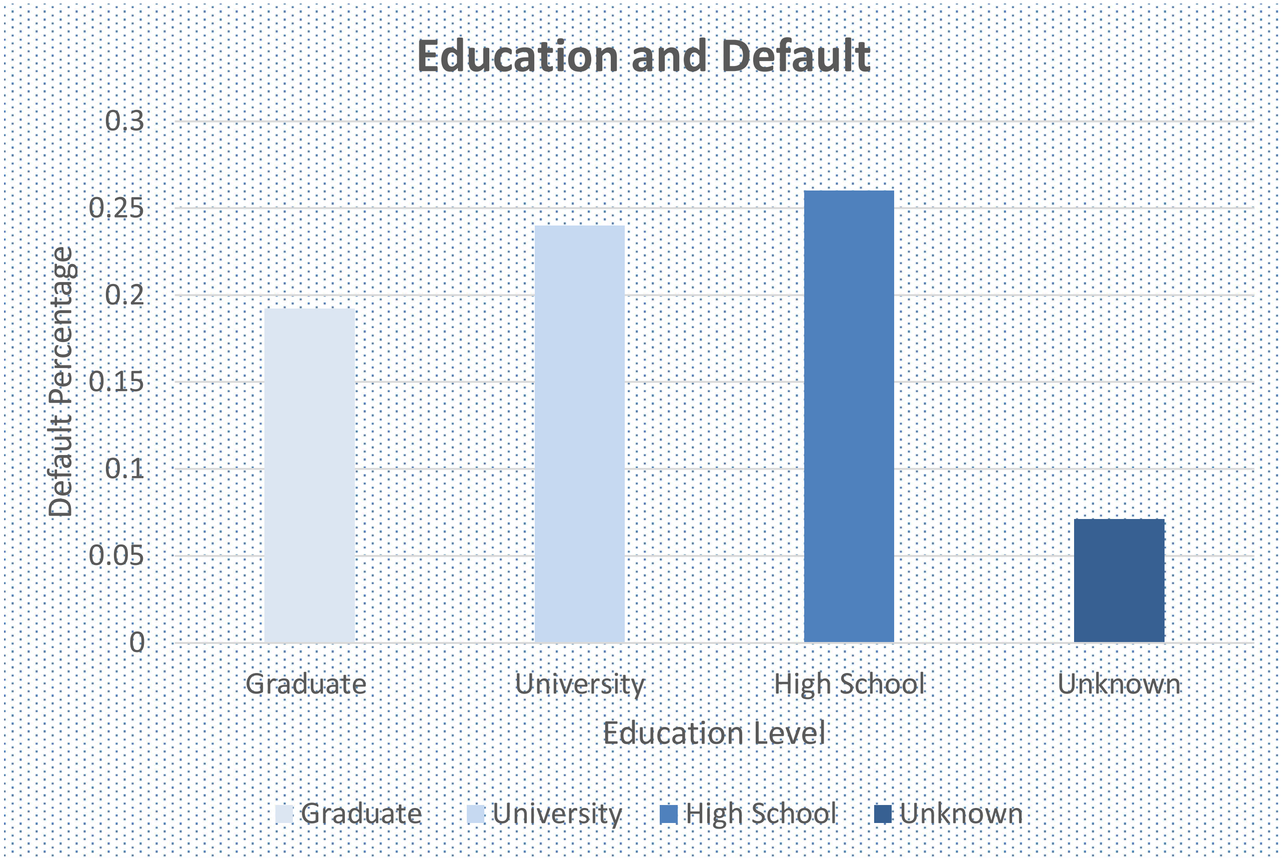
EXPLORATORY DATA ANALYSIS

Age Group



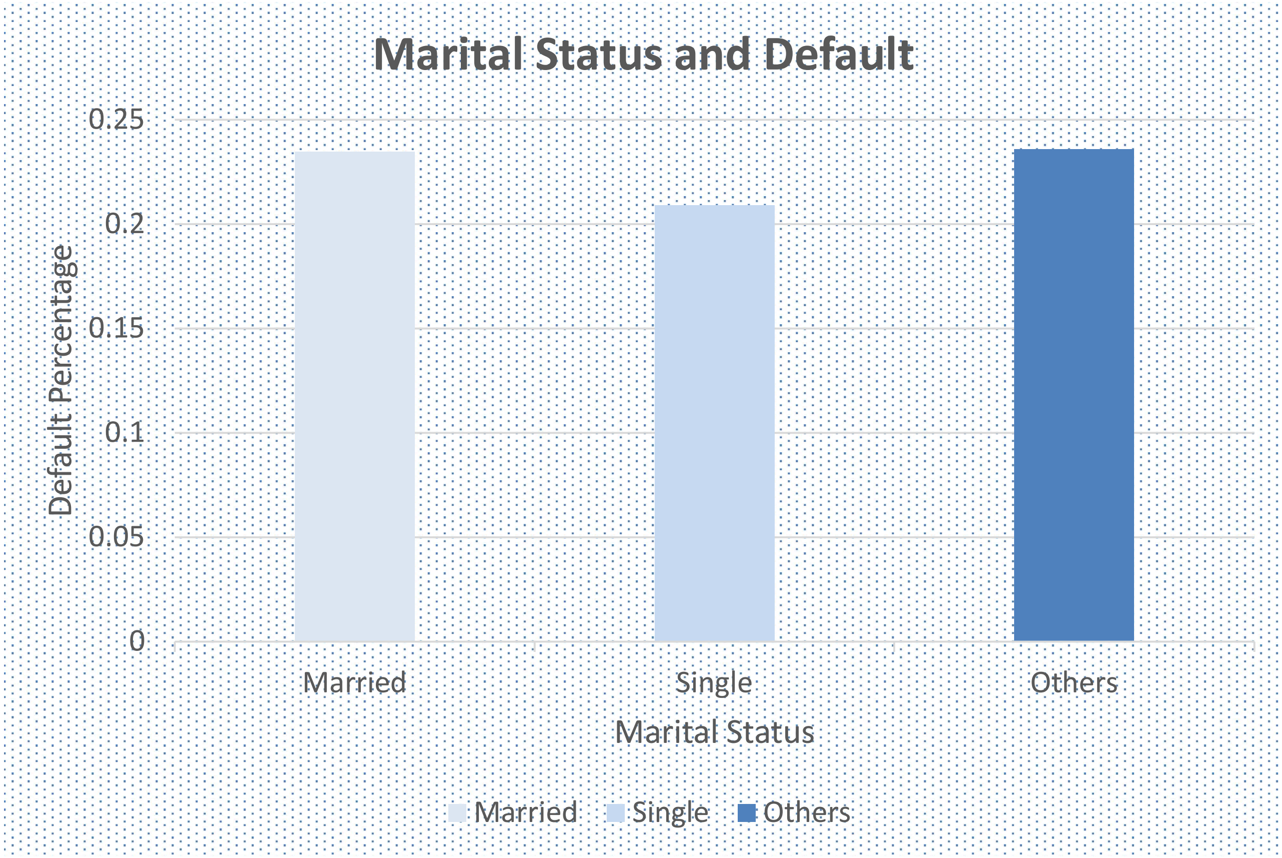
EXPLORATORY DATA ANALYSIS

Education Level



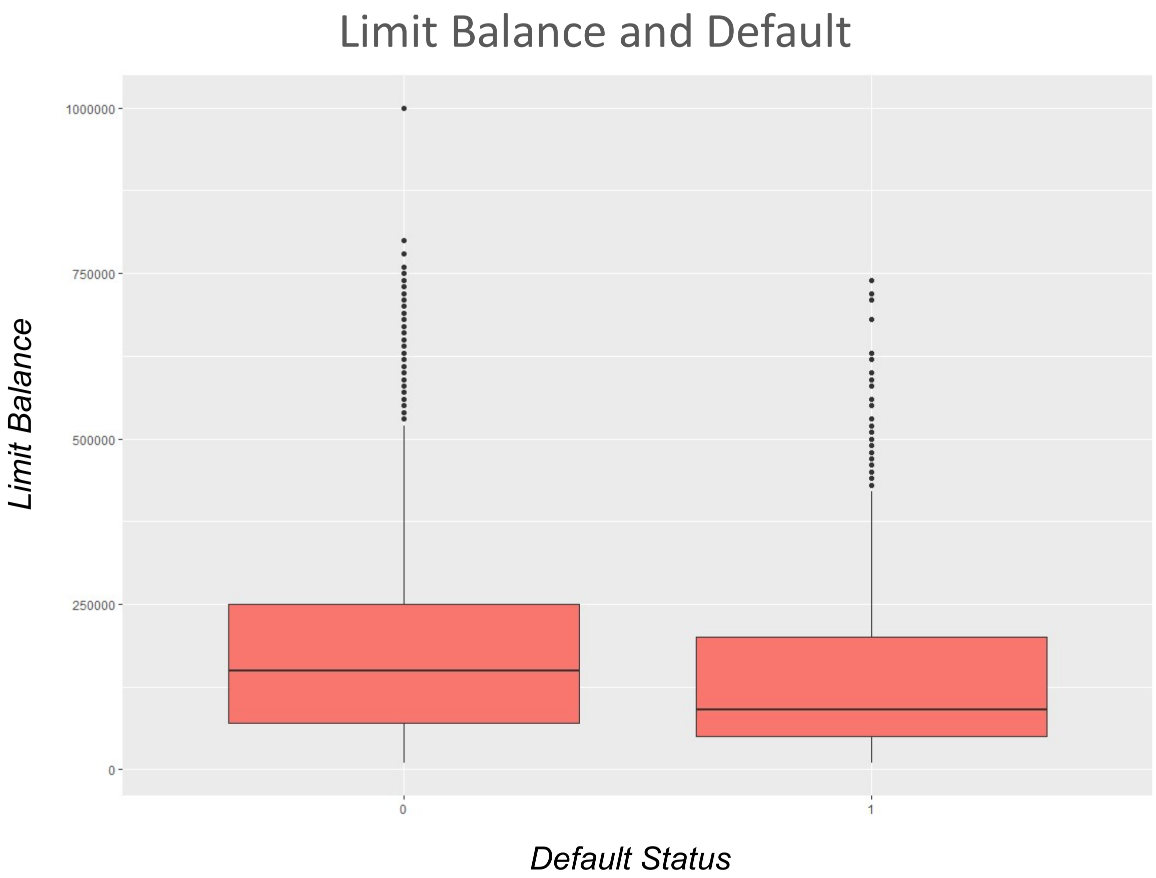
EXPLORATORY DATA ANALYSIS

Marital Status



EXPLORATORY DATA ANALYSIS

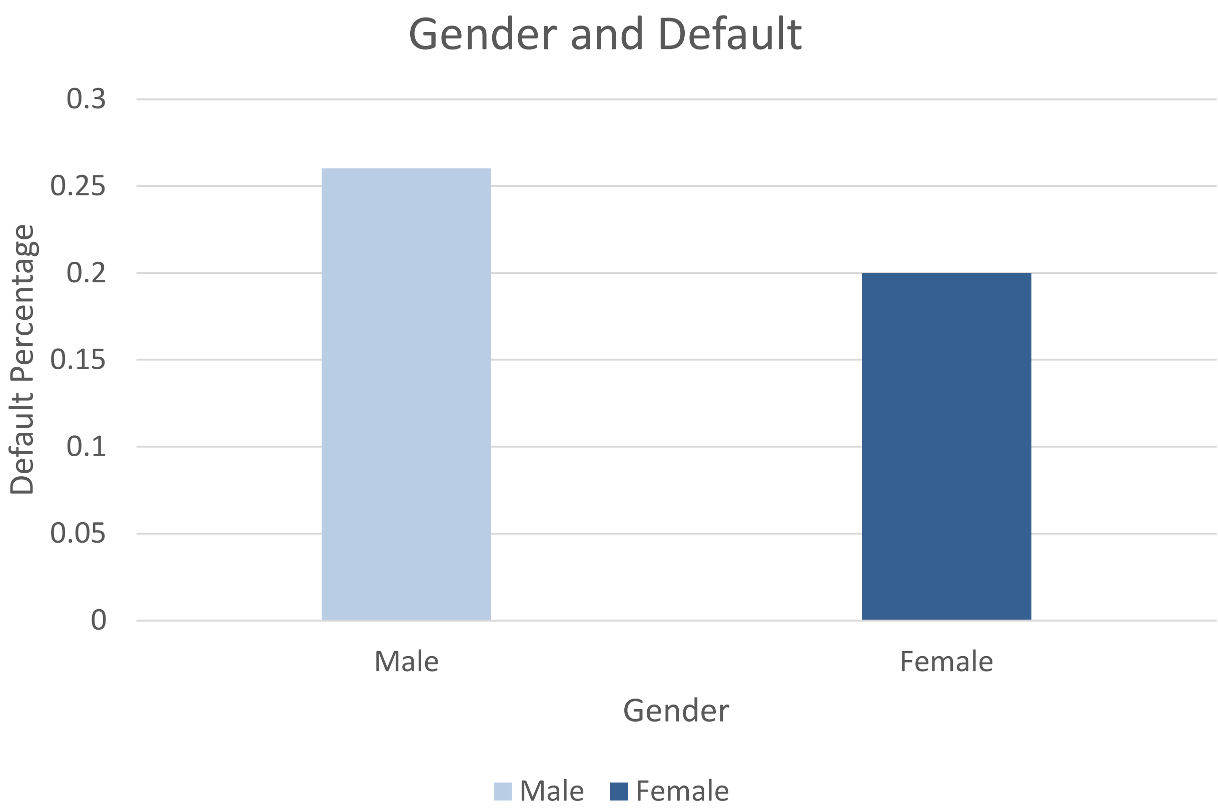
Limit Balance



It is interesting to note that default is inversely proportional to limit balance. This highlights how banks may have higher scrutiny while issuing larger limit balances as compared to smaller limit balances. This could pose a risk of several lower limit balances accumulating to a large loss on the bank

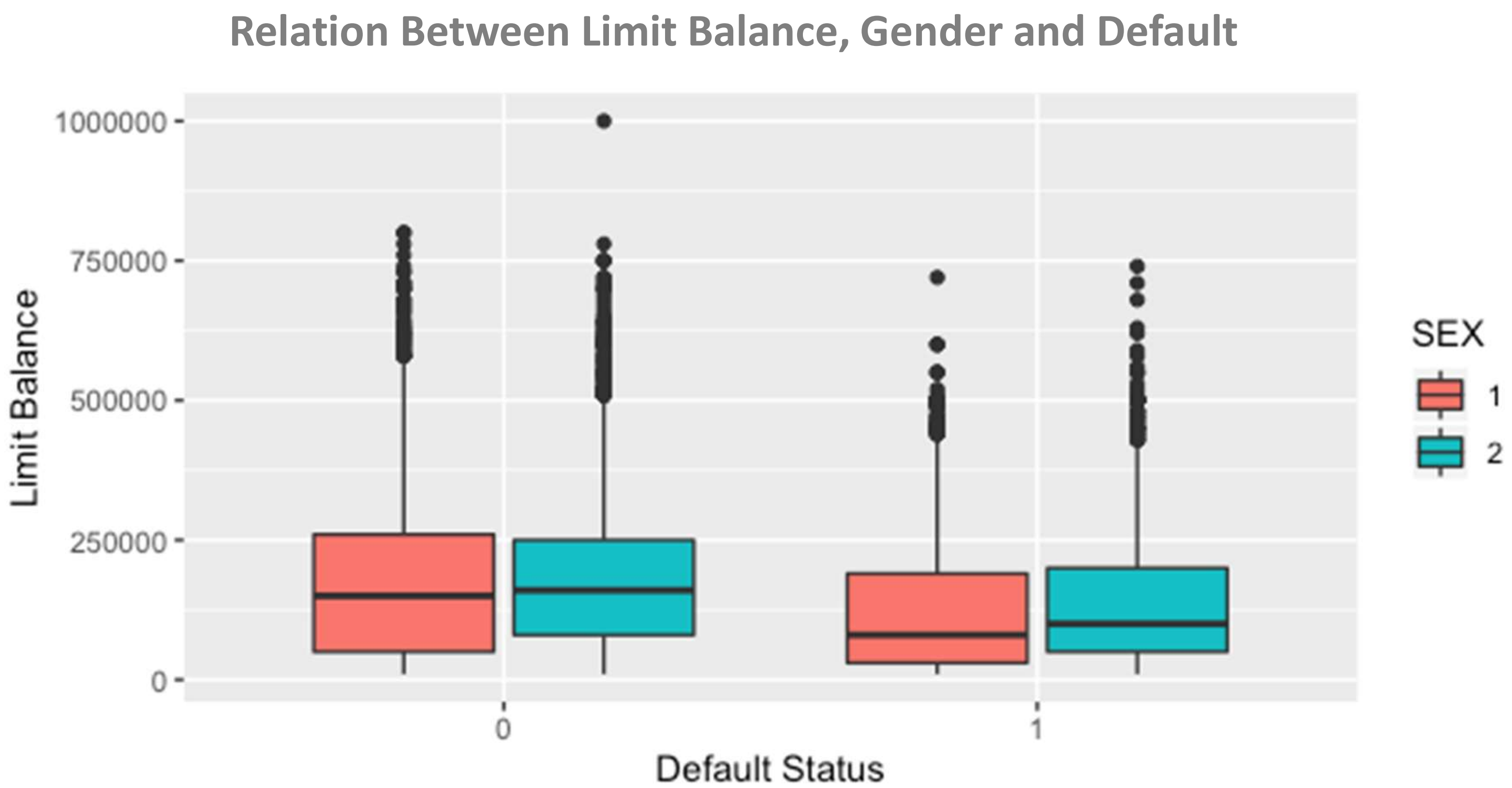
EXPLORATORY DATA ANALYSIS

Gender



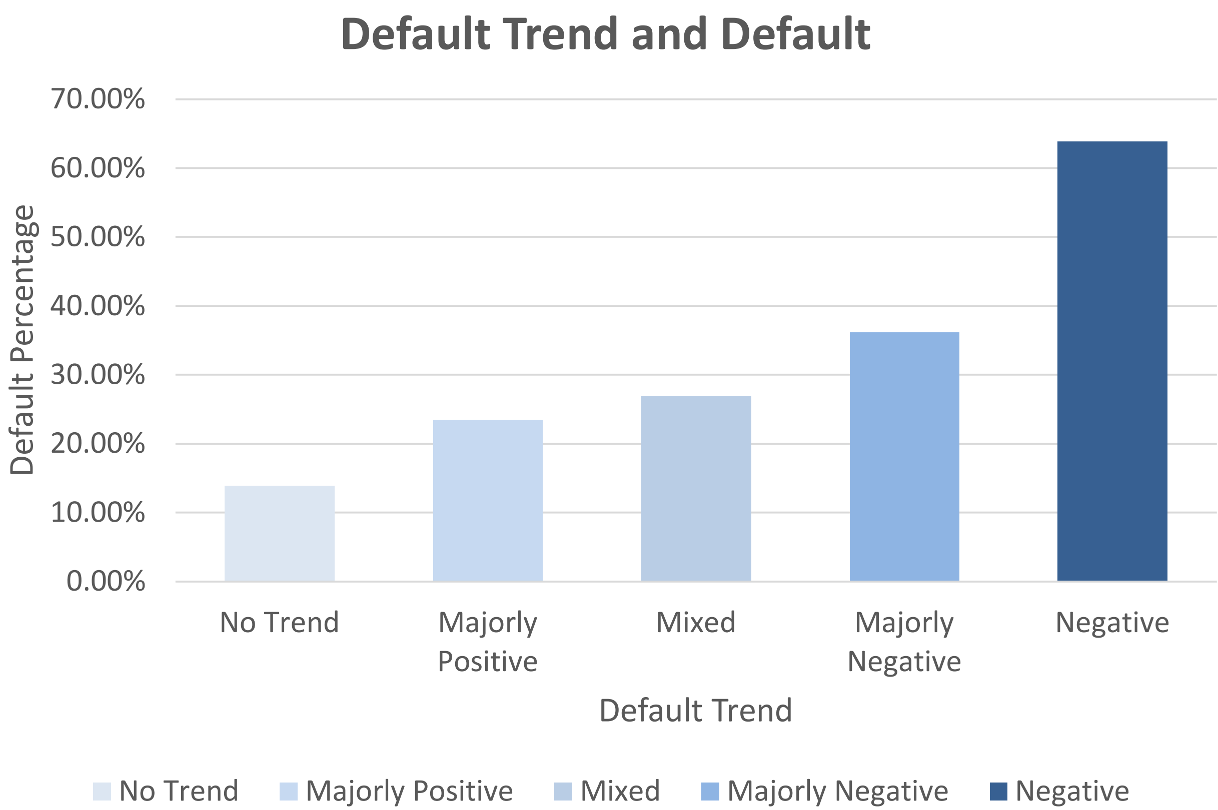
EXPLORATORY DATA ANALYSIS

Limit Balance and Gender



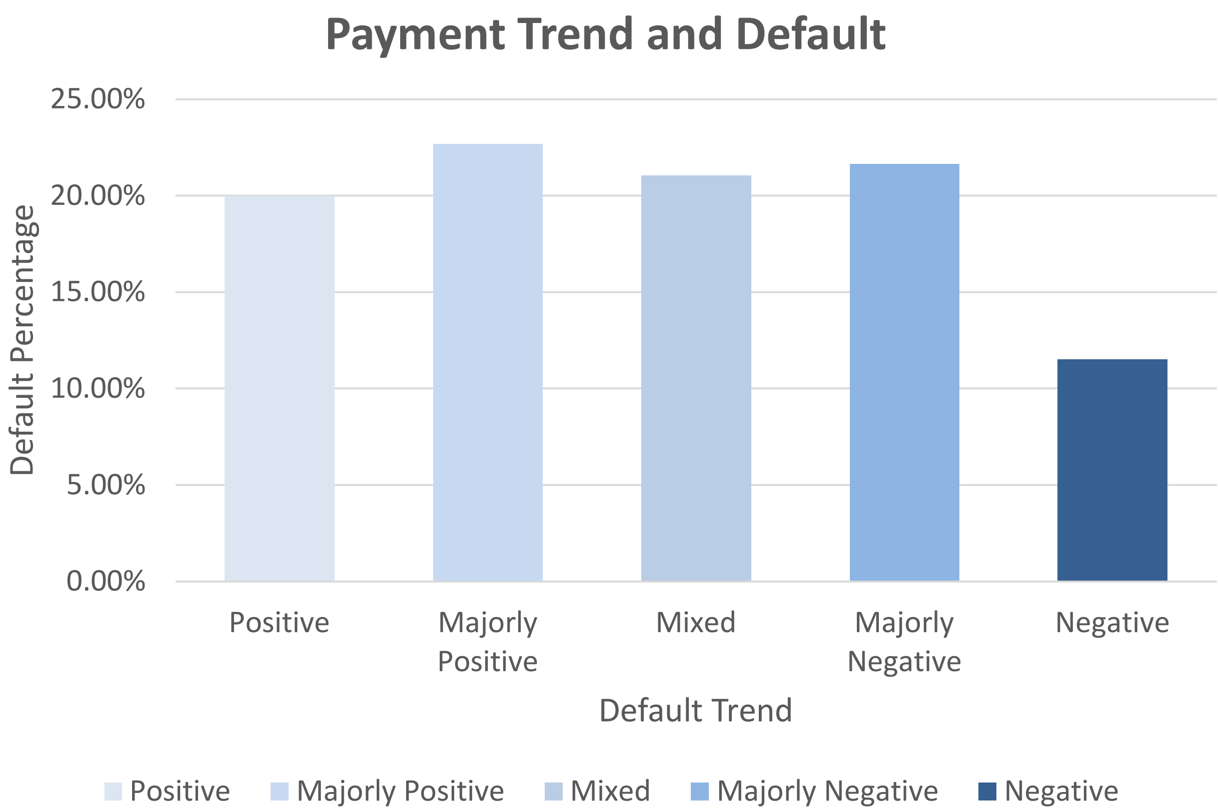
EXPLORATORY DATA ANALYSIS

Default Trend



EXPLORATORY DATA ANALYSIS

Payment Trend



OUR APPROACH

Logistic Regression

- We selected different combinations of variables and fit logistic regression models to identify which variables were statistically significant

Stepwise Model

- Next, we ran a stepwise logistic regression model to identify the predictors that provide an optimum prediction model

Classification Tree

- We also ran a classification regression tree algorithm to identify which predictor variables were crucial in decision making

Other Methods

- Once, we completed the above, we had a good understanding of the strong predictor variables, we now started fitting different classification models to identify which gave the best predictive accuracy

RESULTS

Logistic Regression Results

- As per the regression model, the following factors were statistically significant
- Marriage, Education, Default trend, Recent Payment Proportion and recent default status

Classification Tree Results

- We got a very short tree classification tree which considered only 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

Other Training Models

- While the accuracy rate was high across all models (mostly due to the imbalanced nature of the data), only the QDA gave a very high recall rate but with a lower accuracy rate.
- This shows that a balance would need to be obtained between accuracy and recall would depend on each bank based on the risk appetite.

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%

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 upto present would help in identifying trends in lifecycle of credit cards that may indicate risk of default

The learning from these analysis can also be extended to predicting defaults in home loans, mortgages etc.

THANK YOU !