CONSUMER CREDIT RISK

WILL THEY BE PAYING ON TIME?

BY: JEN, MANU, POOJITA,RISHA, VERLISA

MAY 05, 2018

UCB BOOTCAMP, DATA ANALYTICS

FINANCE 101 ON CREDIT CARD DEFAULT

- Credit card default happens when you’ve become severely delinquent on your credit card payment; if you miss the minimum credit card payment six months in a row.
- In most cases, delinquency can be remedied by simply paying the overdue amount, plus any fees or charges resulting from the delinquency.
- In contrast, default status usually triggers the remainder of your loan balance to be due in full, ending installment payments set forth in the original loan agreement.
- Delinquency adversely affects the borrower’s credit score, but default reflects extremely negatively on it and on the borrower’s consumer credit report, which makes it difficult to borrow money in the future.
- Credit card default borrower may have trouble obtaining a mortgage, purchasing homeowners insurance, getting approval to rent an apartment.

CREDIT CARD DELINQUENCY STATISTICS

Transunion's Q4 2017 Industry Insights Reports

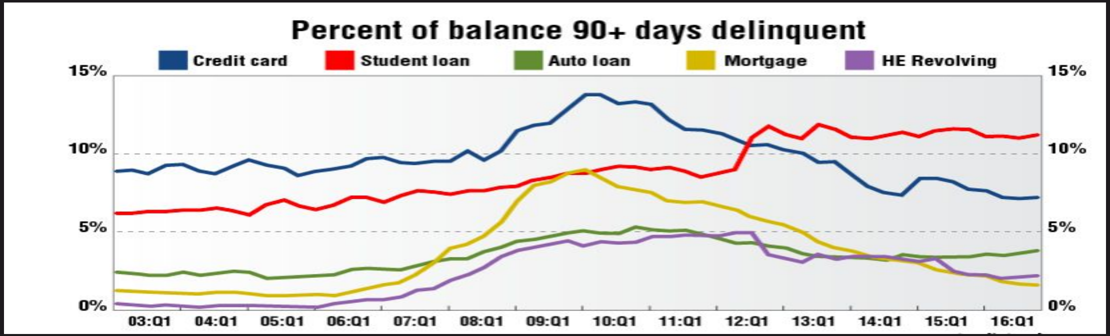
Q4 2017 Credit Card Performance by Age Group

Generation	90 + DPD	Annual Pct. Change	Average Loan Balances Per Consumer	Annual Pct. Change
Gen Z (1995 - present)	2.69%	+4.3%	$1,189	+26.5%
Millennials (1980 - 1994)	2.77%	+0.7%	$4,243	+11.2%
Gen X (1965 - 1979)	2.35%	+2.6%	$7,212	+4.6%
Baby Boomers (1946 - 1964)	1.21%	+5.2%	$6,501	+0.8%
Silent (Until 1945)	0.78%	+6.8%	$4,025	+0.2%

Q4 2017 Credit Card Trends

Credit Card Lending Metrics	Q4 2017	Q4 2016	Q4 2015	Q4 2014
Number of Credit Cards	419M	404M	381M	364M
Borrower-level Delinquency Rate (90 + DPD)	1.87%	1.79%	1.59%	1.48%
Average Debt Per Borrower ($)	5,644	5,486	5,337	5,329
Prior Quarter Originations	16.3M	17.5M	15.4M	14.4M
Average New Account Credit Lines ($)	5,194	5,373	5,068	+0.2%

CREDIT CARD DELINQUENCY STATISTICS



Based on Transunion's insights, Credit Card delinquency has been a cause for concern in the United States which is reflective in the QoQ trendline

HIGH RATES:

States with High Delinquency Rates	Rate
Mississippi	3.14%
Louisiana	2.46%
Arkansas	2.41%
Georgia	2.37%
West Virginia	2.28%

LOW RATES:

States with Low Delinquency Rates	Rate
Wisconsin	1.11%
Washington	1.12%
Utah	1.14%
Minnesota	1.15%
Montana	1.19%

DATASET SUMMARY

	TRAIN DATASET		REAL DATASET
Timeframe:	04/2005 - 09/2005	*Timeframe:*	2018
Transactions:	30K	*Transactions:*	30
Attributes:	24	*Attributes:*	24
Location:	Taiwan	*Location:*	USA
Source:	Center of ML & Intelligent Systems, UCI	*Source:*	Team Personal Transactions

PROJECT WORKFLOW

DISCOVERY

Standardized data

Descriptive Analytics

EXPLORATION

Compiled personal dataset

Feature Engineering

Dimensionality Reduction

MODELING

Trained and fit dataset into numerous ML models

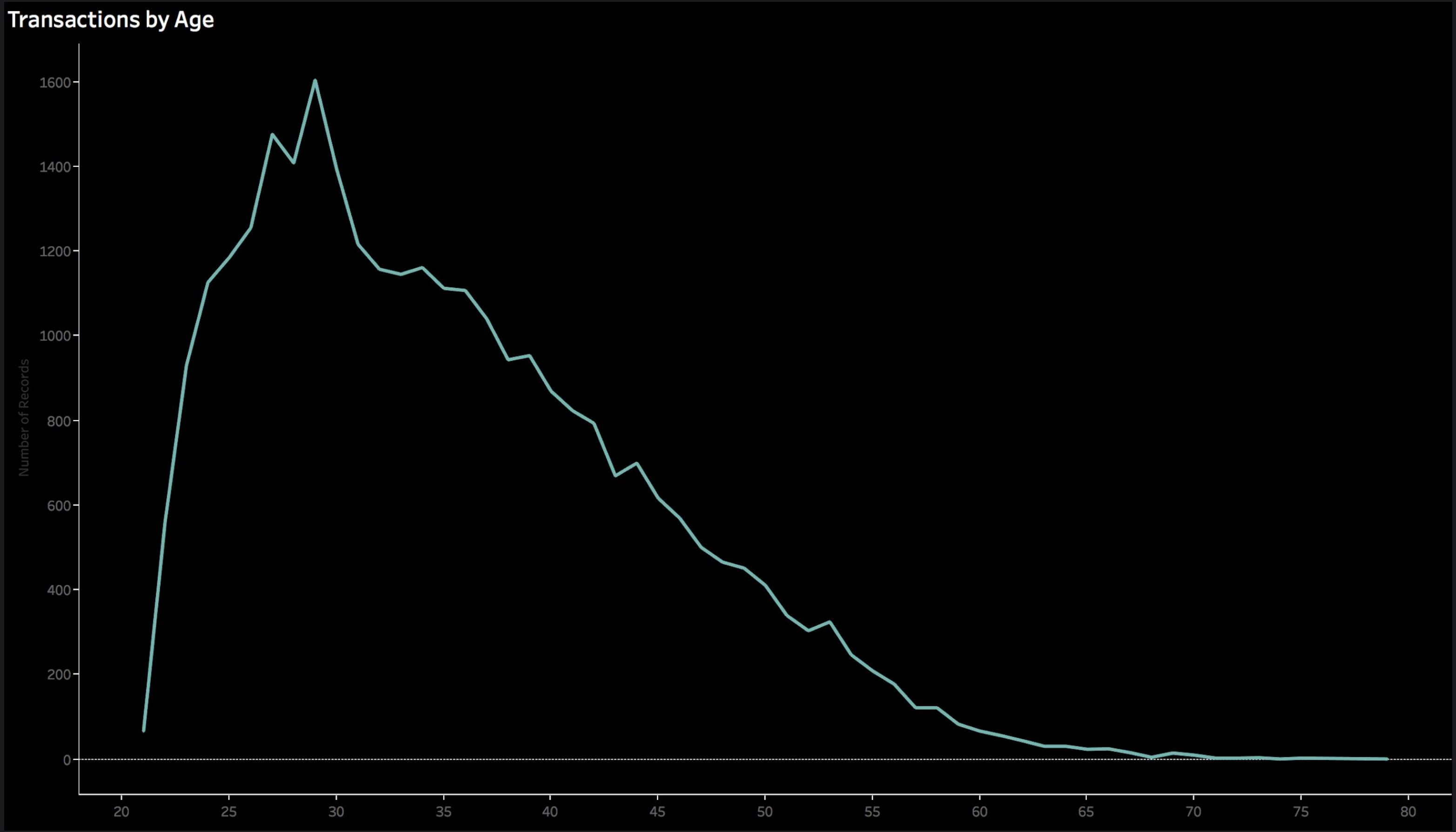
Summarized data into key visualizations

VISUALIZATION

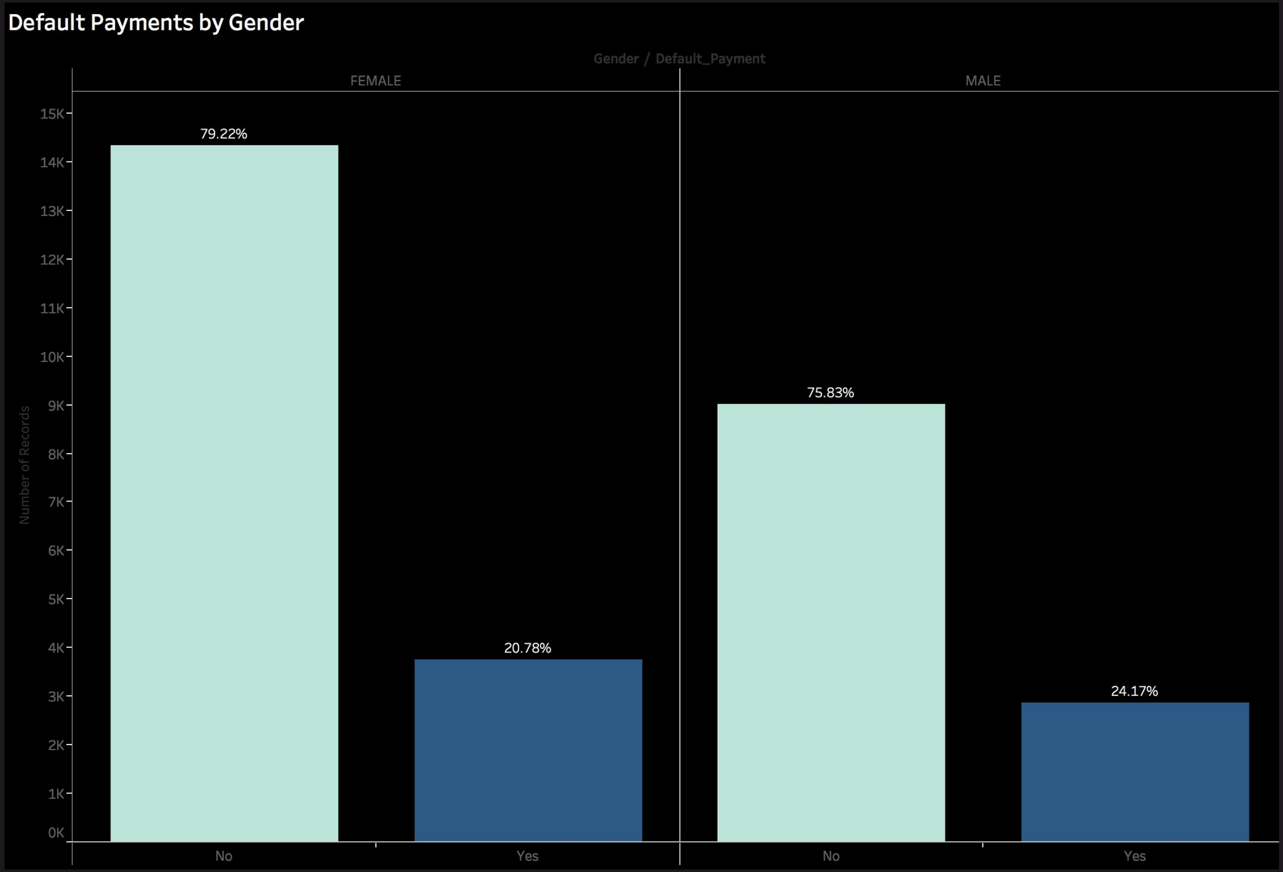
CONCLUSIONS

ATTRIBUTE RELATIONSHIPS TO DEFAULTER STATUS

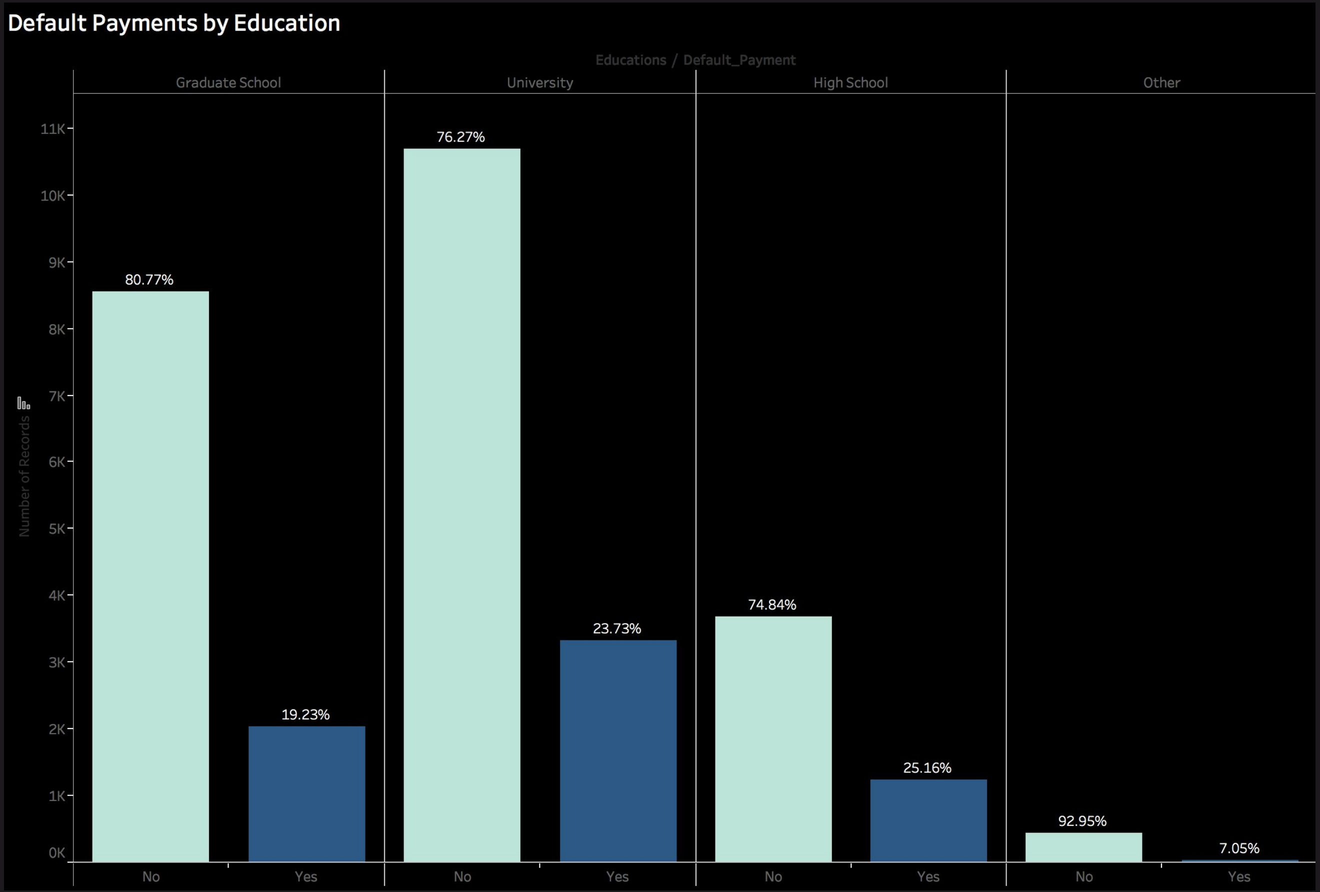
CONSUMERS BETWEEN 25-35 HAVE MOST TRANSACTIONS



MALE DEFAULTER RATE IS *SLIGHTLY* HIGHER THAN FEMALES

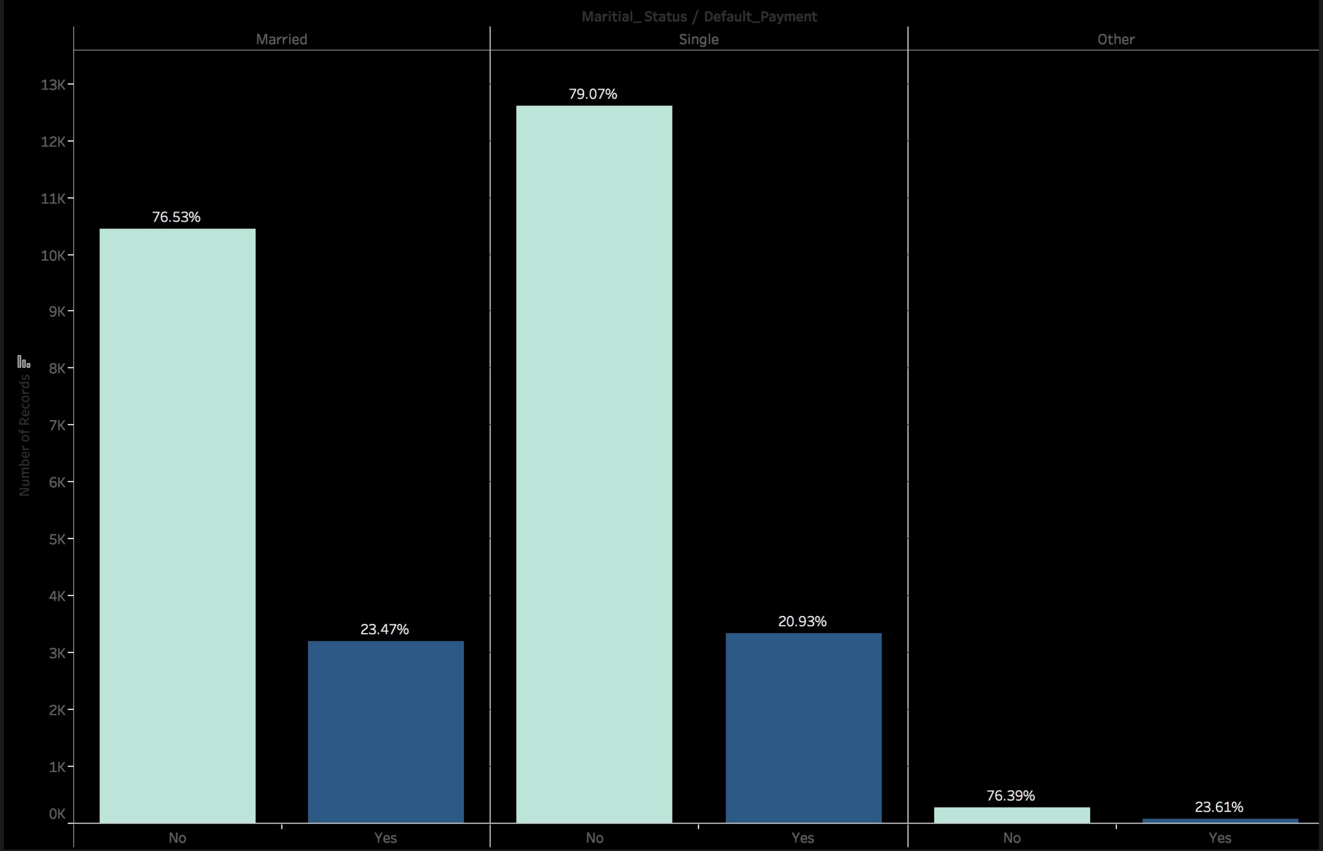


HIGHER EDUCATION IS RELATED TO LOW DEFAULTER RATE

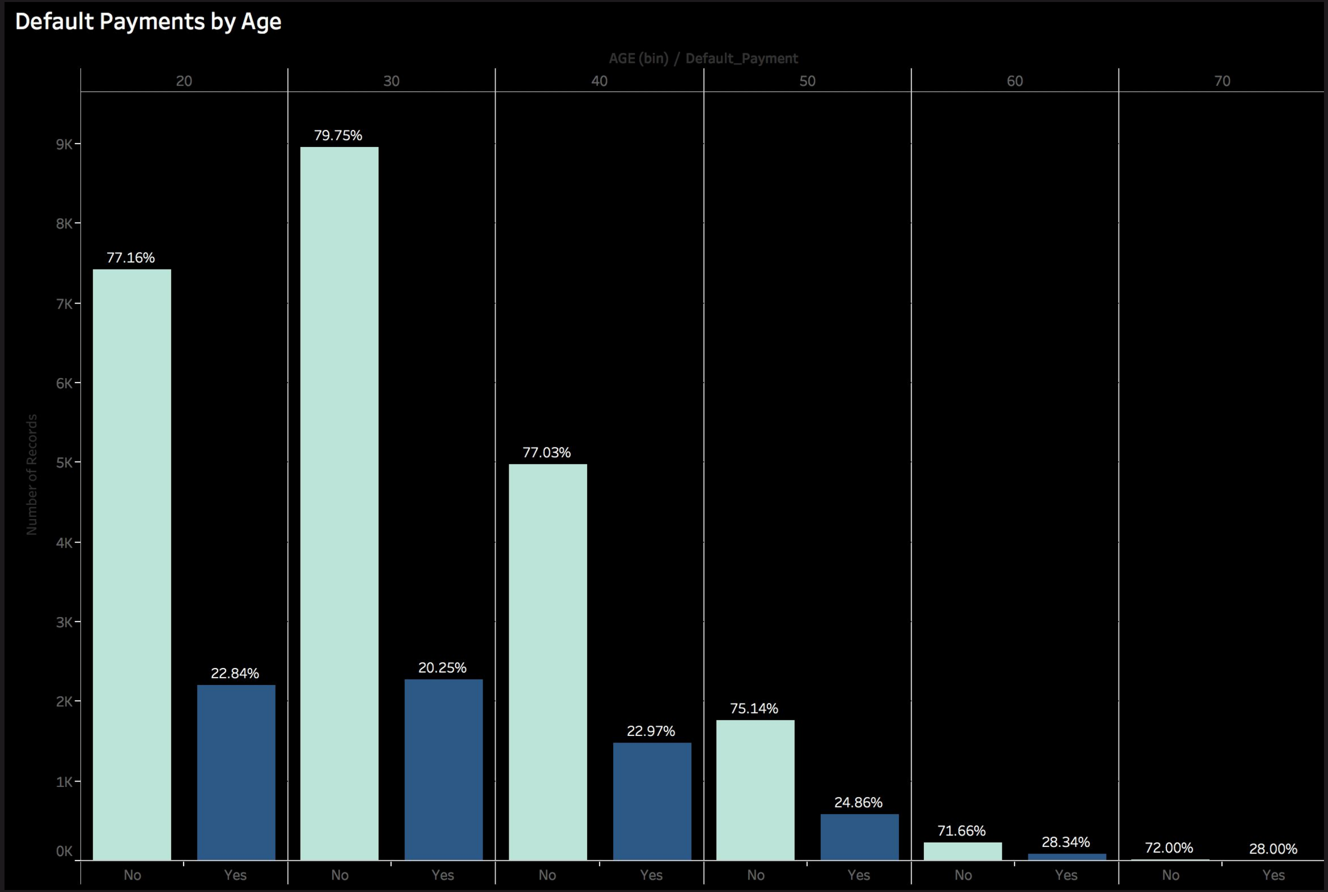


MARRIED DEFAULTER STATUS TENDS TO BE SLIGHTLY HIGHER

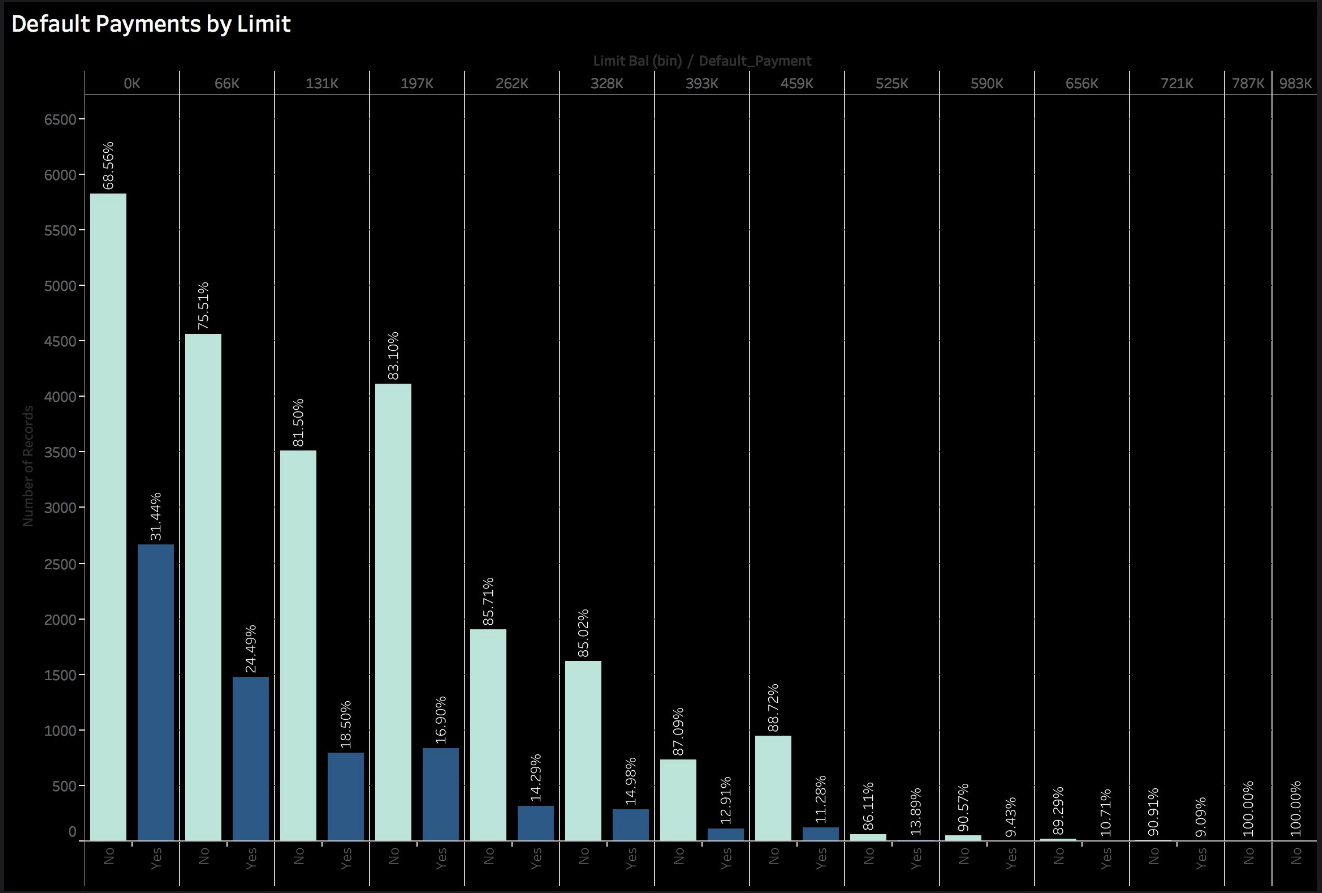
Default Payments by Marital Status



HIGHER AGE BINS ARE RELATED TO HIGHER DEFAULTER RATE



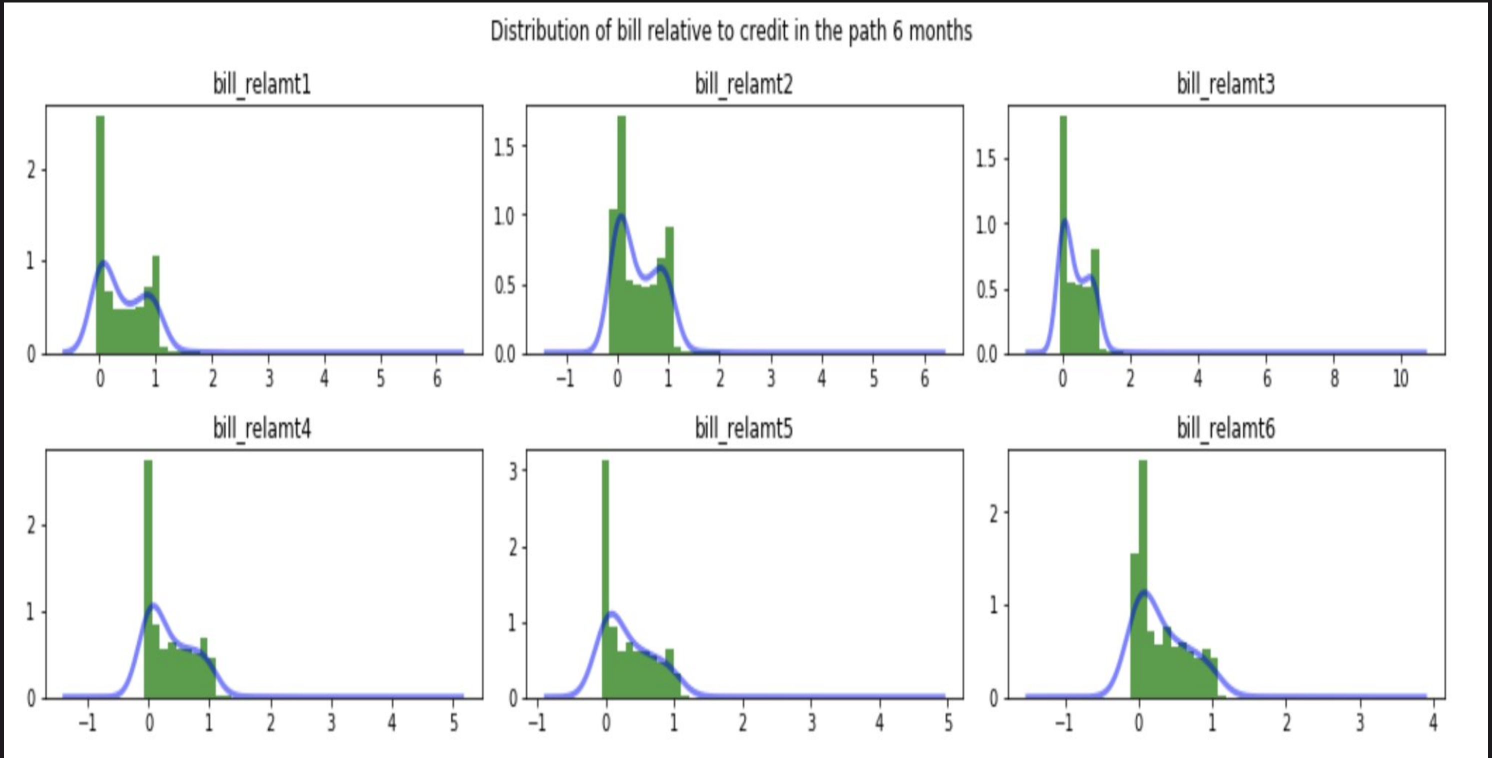
LOWER CREDIT CARD LIMITS RELATE TO HIGH DEFAULTER RATE



MACHINE LEARNING MODELS FOR PREDICTABILITY & CLASSIFICATION

FEATURE ENGINEERING

Process of using domain knowledge of the data to create features that make machine learning algorithms work.



LOGISTIC REGRESSION

Logistic regression

Command: Statistics
- Regression
  - Logistic regression

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

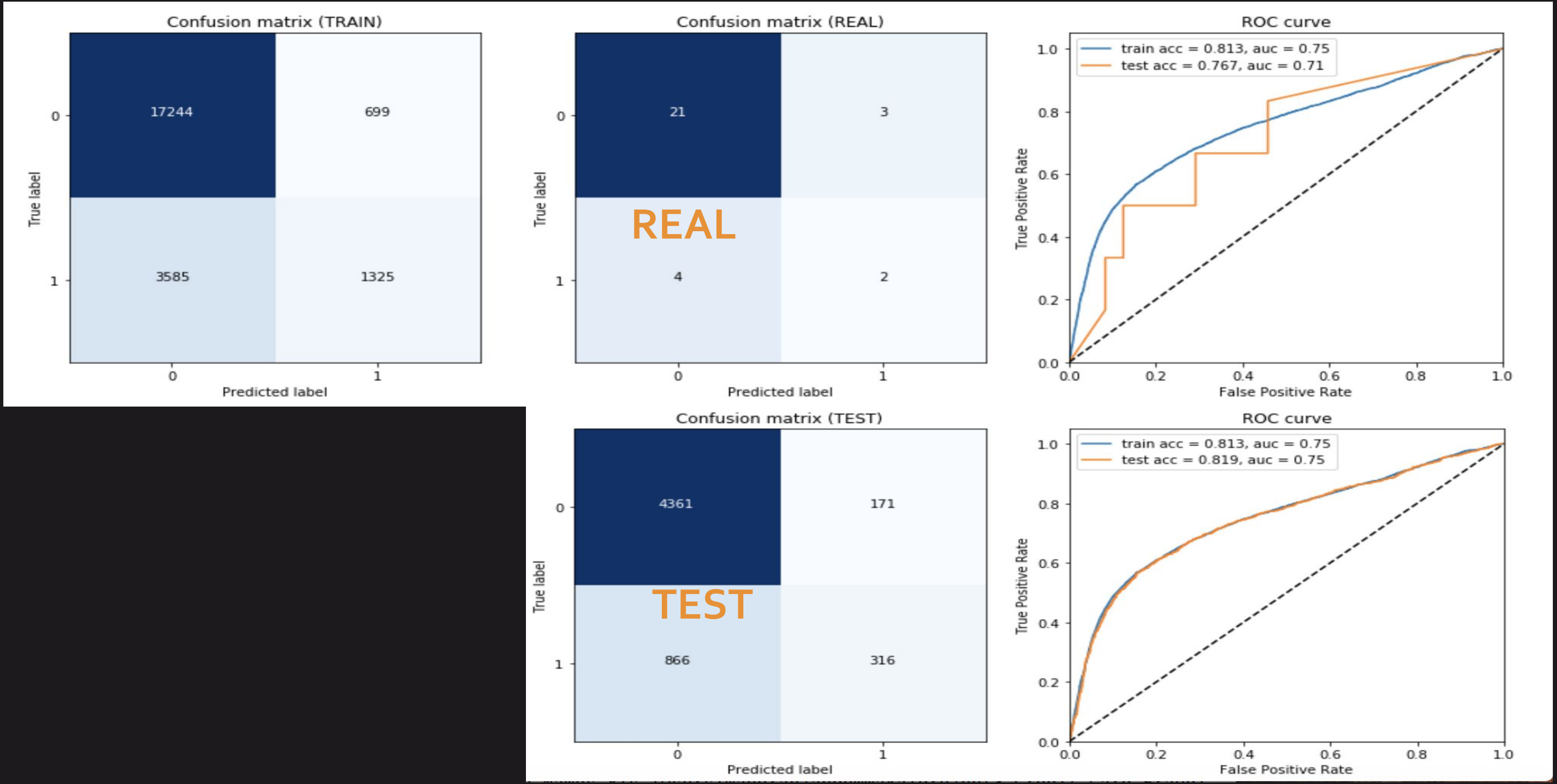
where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\,of\,presence\,of\,characteristic}{probability\,of\,absence\,of\,characteristic}$$

and

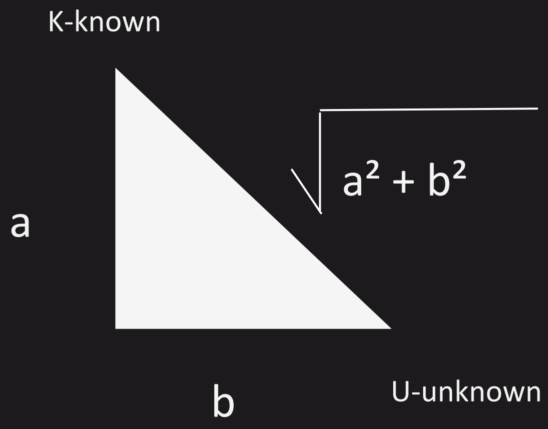
$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

LOGISTIC REGRESSION



K-NEAREST NEIGHBORS (KNN)

- Euclidean distance between the new point and its nearest neighbors
- Pythagorean Theorem: $a^2 + b^2 = c^2$
- Real data

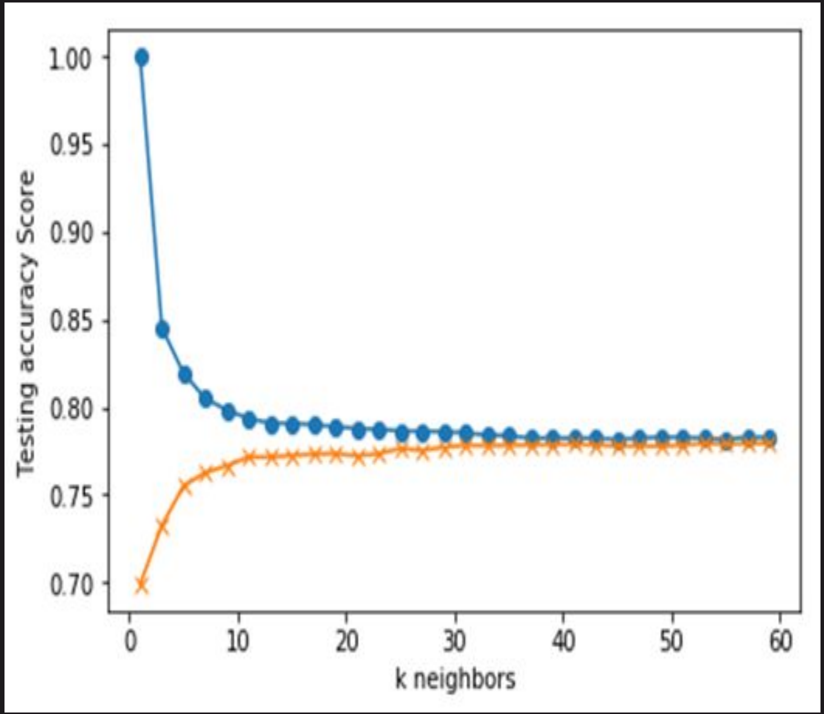


Other distance functions:

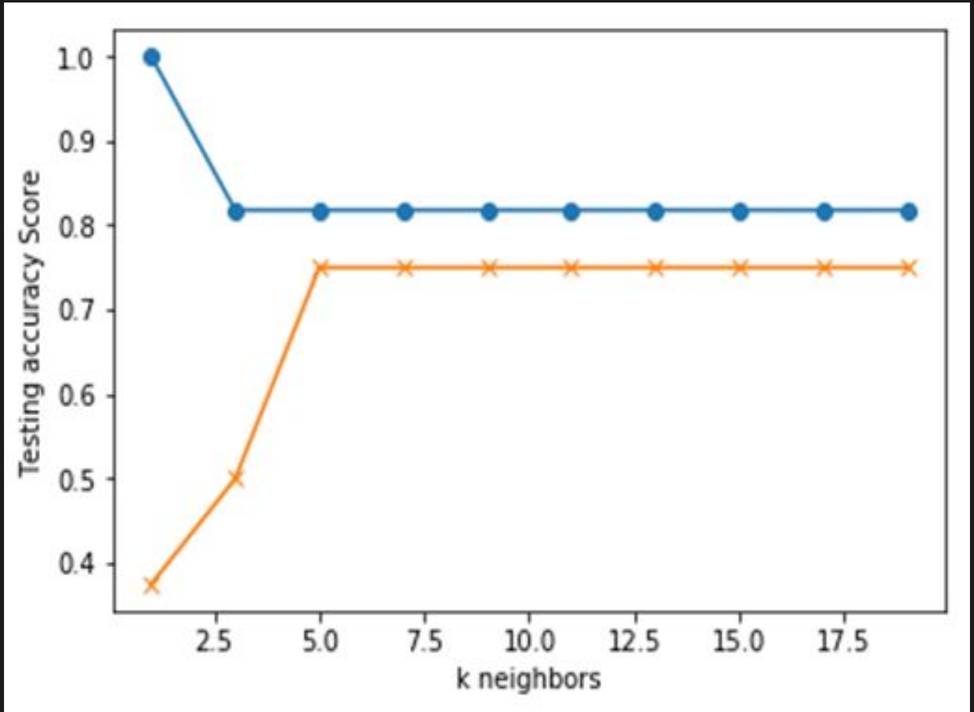
- Manhattan
- Minkowski
- Hamming - categorical data

TRAIN vs. REAL

TRAIN DATA



REAL DATA



SUMMARY: KNN, STRUCTURED, LAZY

ADVANTAGES

- Simple to use
- Follows familiar steps
    - Split data into test/train
    - Predict using trained model
- Use with multiple features
- High degree of accuracy
- Both Classification and Regression

DISADVANTAGES

- Time consuming
    - Run for each k
    - More features = more time
- Scaling affects results - PCA
- Data must be clustered – can't be too random
- Assumes straight line between points – may not always be true

DECISION TREES

WHAT?

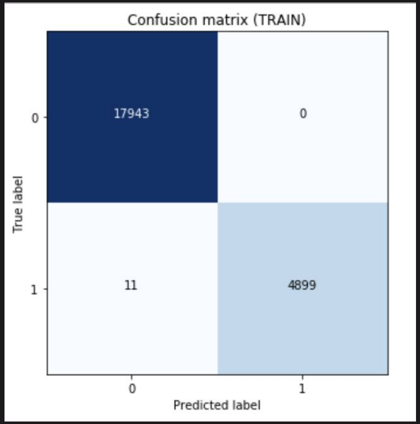
It looks at the variables in a data set, determines which are the most important, and then comes up with a tree of decisions that best partitions the data.

RELATED TERMS:

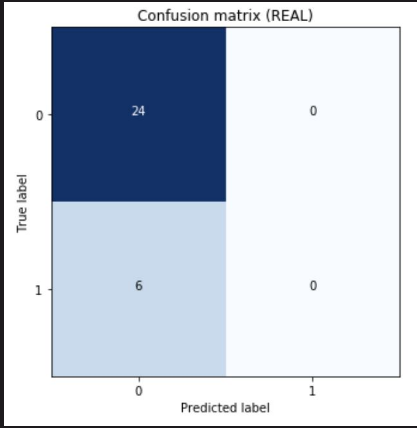
Impurity: level of uncertainty
Information Gain: how much uncertainty is reduced

GOAL:

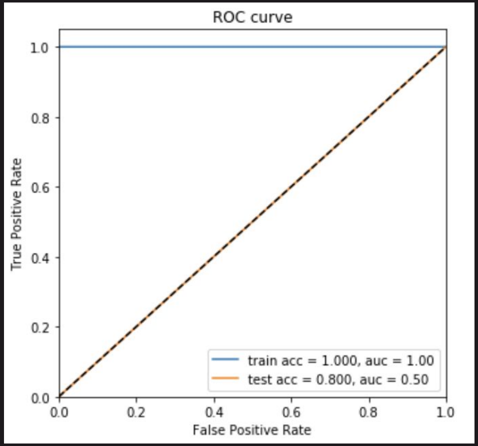
Unmix the data to produce the purest possible distribution of the labels.



>90%



80%



DECISION TREES

ADVANTAGES:

- Implicility perform feature selection
- Can easily handle qualitative (categorical) features
- Requires little data preparation
- Nonlinear relationships between parameters do not affect tree performance

DISADVANTAGES:

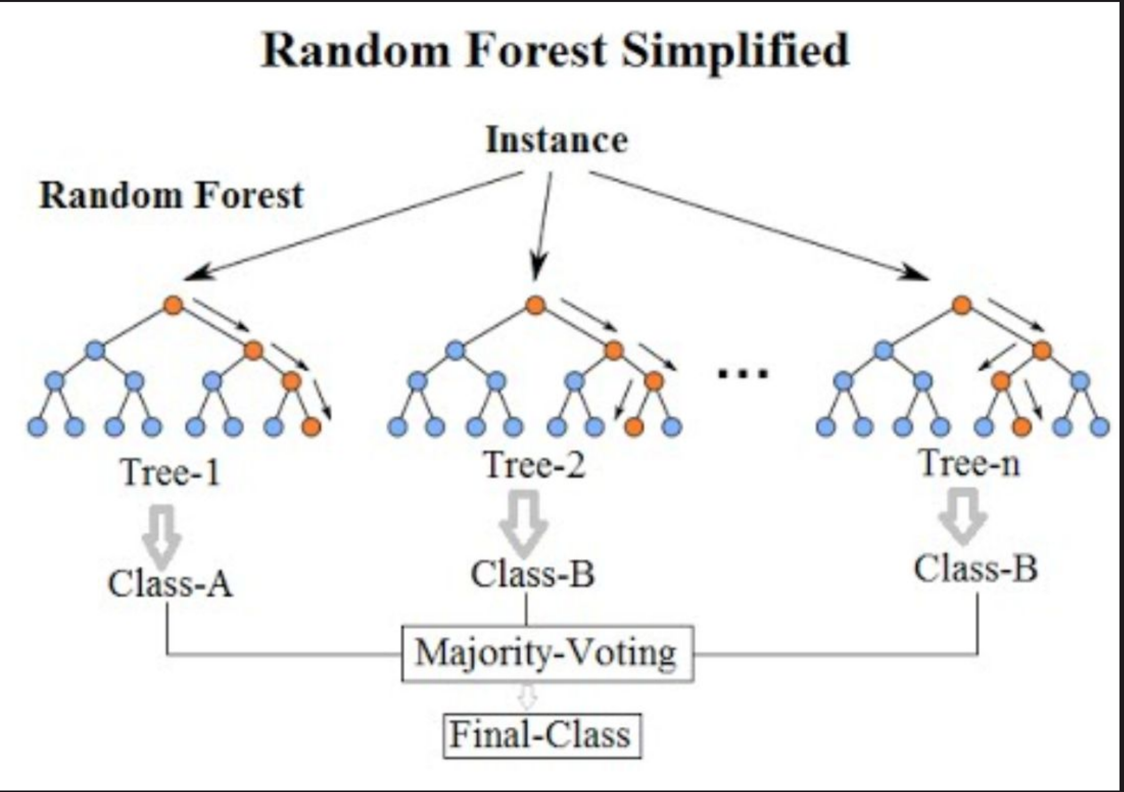
- Prone to overfitting
- Unstable, small change in data can lead to a large change in structure
- Tree structure prone to sampling

RANDOM FOREST

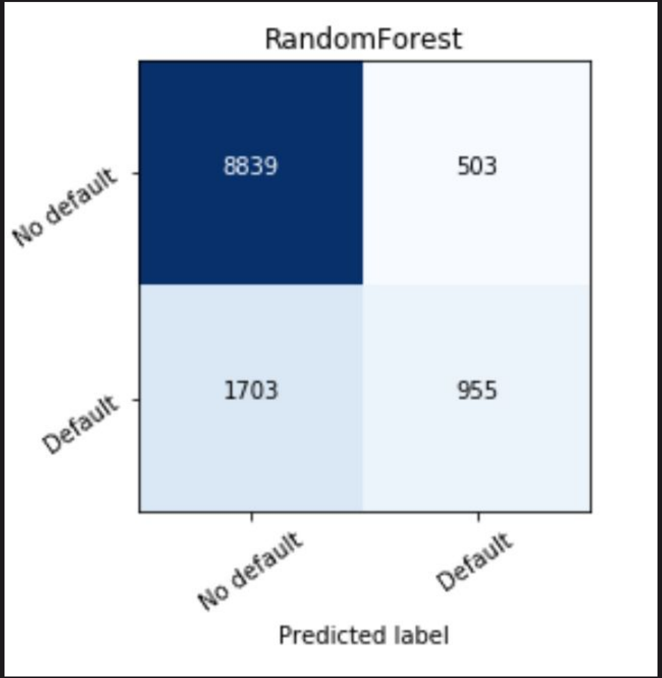
Random Forests train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data. There are typically two parameters in RF - number of trees and no. of features to be selected at each node.

- RF is good for parallel or distributed computing.
- Almost always have lower classification error and better f-scores than decision trees.
- Almost always perform as well as or better than SVMs, but are far easier for humans to understand.
- Deal really well with uneven data sets that have missing variables.
- Gives you a really good idea of which features in your data set are the most important
- Generally train faster than SVMs.
- Not as easy to visually interpret

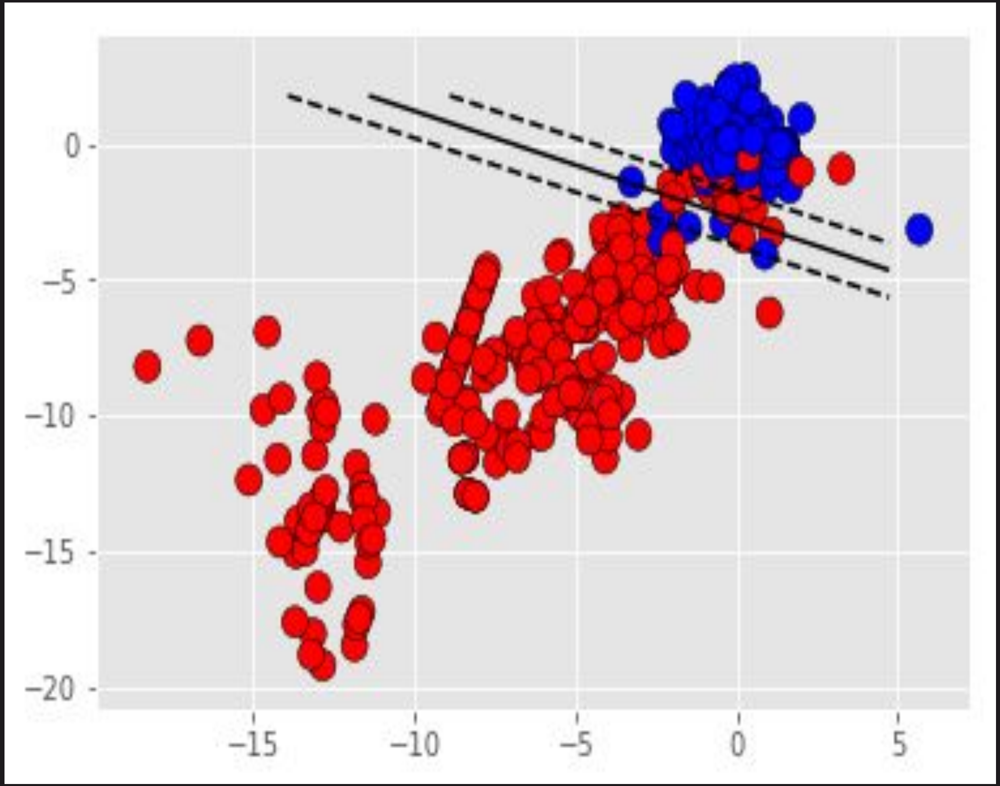
RANDOM FOREST



CONFUSION MATRIX:



SUPPORT VECTOR MACHINE (SVM)



Summary

- ➔ Linear Classifier (SVM)
- ➔ Features Used:
  - Total Pay Amount
  - Total Bill Amount
  - Education
  - Age
  - Credit Card Limit
- ➔ **Accuracy: 77.9%**
- ➔ **Accuracy(Real Data): 73%**

SUPPORT VECTOR MACHINE - FINDINGS AND ANALYSIS

- ➔ **Balanced data for accuracy in prediction model**
- ➔ **Scaled data set for performance efficiency**
- ➔ **Leverage appropriate feature set**
- ➔ **Recommended : Yes**

Advantages

- ❏ Enables Kernel engineering based on data and applications
- ❏ Accurate classifier
- ❏ Less overfitting

Disadvantages

- ❏ Limited on multi-class classification
- ❏ Computationally Expensive

MODEL PERFORMANCE SUMMARY

ALGORITHM	ACCURACY (TRAIN)	ACCURACY (REAL)
LOGISTIC REGRESSION	77.9%	73%
K-NEAREST NEIGHBORS (KNN)	77.9%	75%
DECISION TREE	>90%	80%
RANDOM FOREST	80.5%	80%
SUPPORT VECTOR MACHINE (SVM)	77.9%	73%

QUESTIONS?

https://github.com/jmtchen/Project3_Credit_Card_Fraud

APPENDIX

SYNOPSIS:

This project unfolds the following phases.

- Getting the Data
- Data Preparation
- Descriptive analytics
- Feature Engineering
- Dimensionality Reduction
- Modeling
- Explainability

MODELING:

We are comparing the predictive power of below algorithms.

- Logistic Regression (scikit-learn)
- Support Vector Machine (scikit-learn)
- KNN (scikit-learn)
- Decision Trees (scikit-learn)
- Random Forest (scikit-learn)

SOURCE:

The dataset is availble at the Center for Machine Learning and Intelligent Systems, Bren School of Information and Computer Science, University of California, Irvine: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

SOURCE:

The dataset is availble at the Center for Machine Learning and Intelligent Systems, Bren School of Information and Computer Science, University of California, Irvine: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients