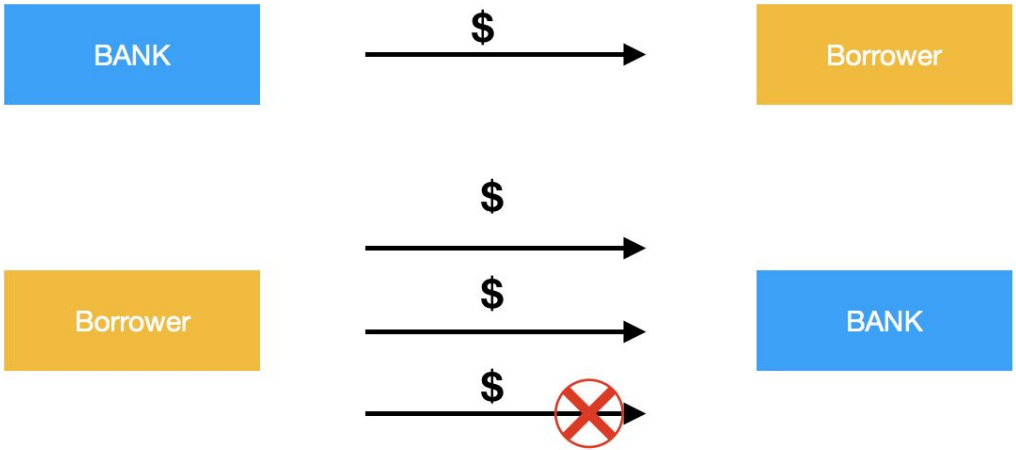
Credit Risk Prediction

Belle Pandya . Kate Weber

Introduction

Banks determine whether to lend money to a customer based on many factors, including the customer's age, home ownership status, annual income, and credit rating. These factors are intended to be used to predict whether the customer will pay back the loan or defect, in which case, the bank would not lend money to them.



Dataset Description (Before and After)

Size (n)	~30,000 rows
Number of Predictors	8
Responding Variable	1, Loan_status (Binary)
Continuous Variables	5, (age, emp_length, loan_amount, int_rate,annual_income)
Categorical Variables	2, grade (A to G), home_ownership(own, mortgage, rent, others)

Sample distribution	Train = 23274; Test = 5818
Number of Predictors	After spatial sign transformation and scaling = 12
Responding Variable	1, Loan_status (Binary)
Continuous Variables	11, (age, emp_length, loan_amount, int_rate,annual_income, grade (1-5), homeownership (mortgage,own,rent)
Categorical Variables	Converted into dummy variables

Pre- Processing steps

Outliers

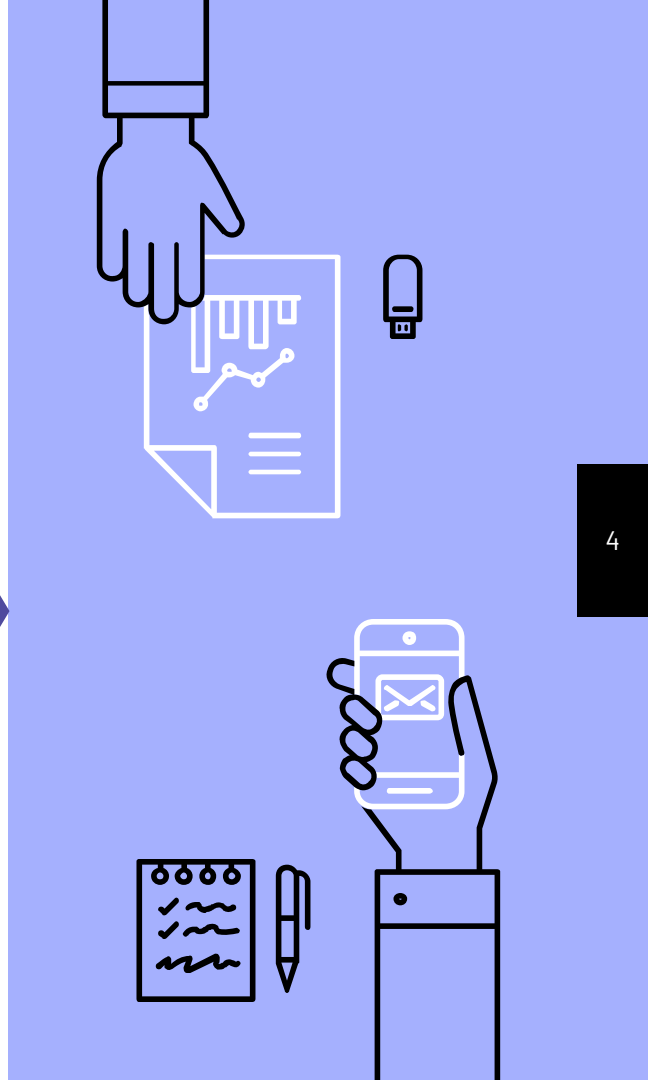
- Exploratory Data Analysis
- Spatial Sign Transformation

Skewness

- Centered
- Scaled
- BoxCox

Missing Values

- kNN Imputation



Data Splitting/Resampling

Data Splitting Method

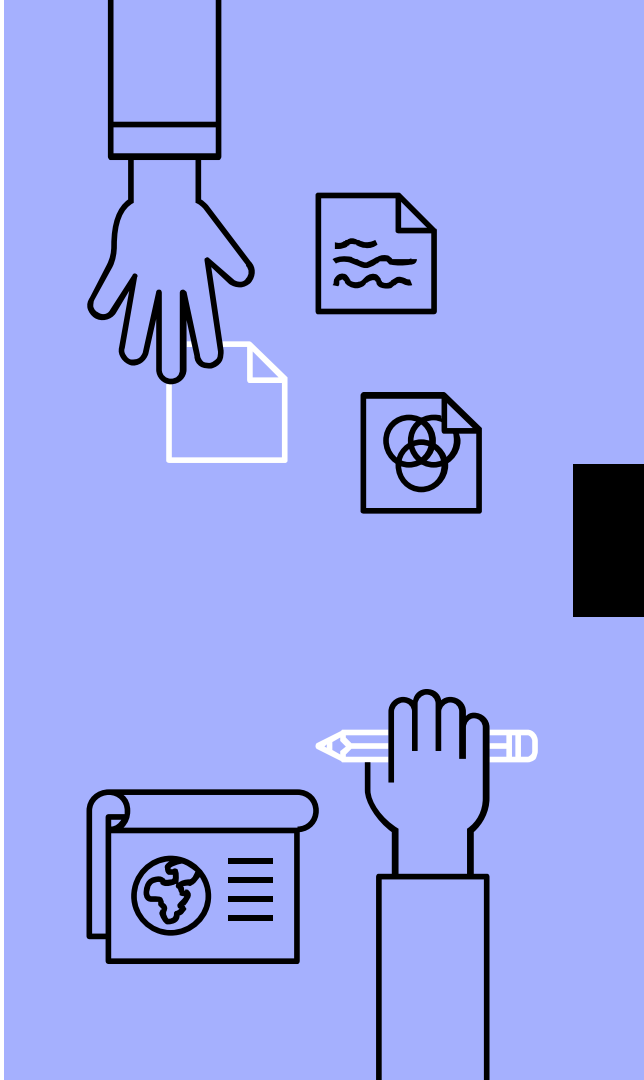
- Stratified Random Sampling

Train: 23,274 observations

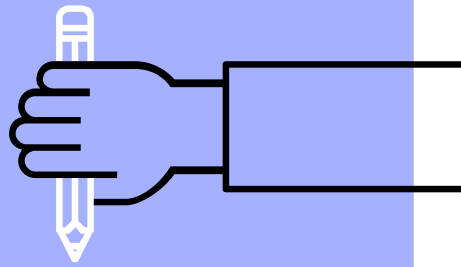
Test: 5818 observations

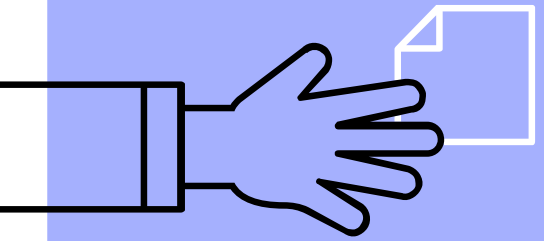
Resampling Method

- Leave Group Out Cross Validation (LGOCV)
- Up-sampling



Linear Classification Models





Logistic Regression

Accuracy: 58.6%

Kappa: 0.096

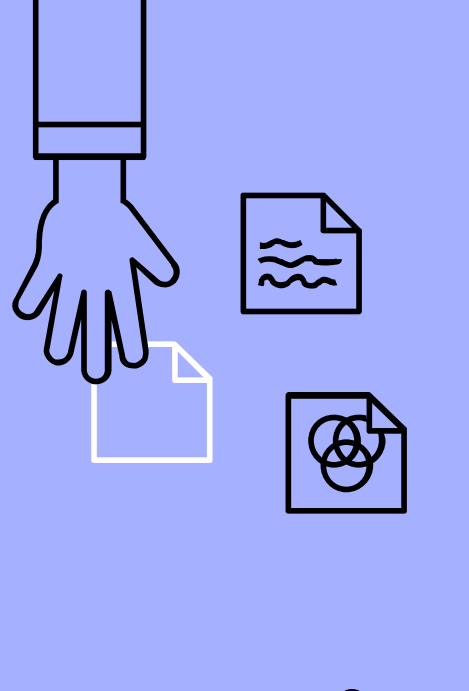
(no tuning parameters)

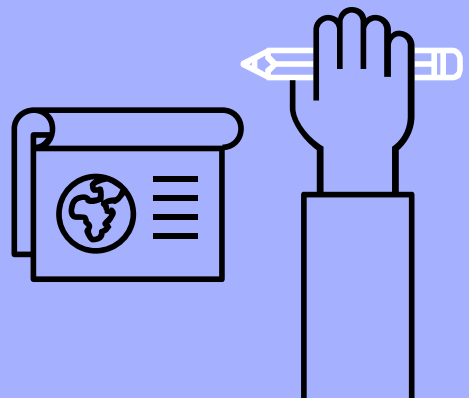
Linear Discriminant Analysis

Accuracy: 58.5%

Kappa: 0.092

(tuning parameter dimen held at constant of 1)



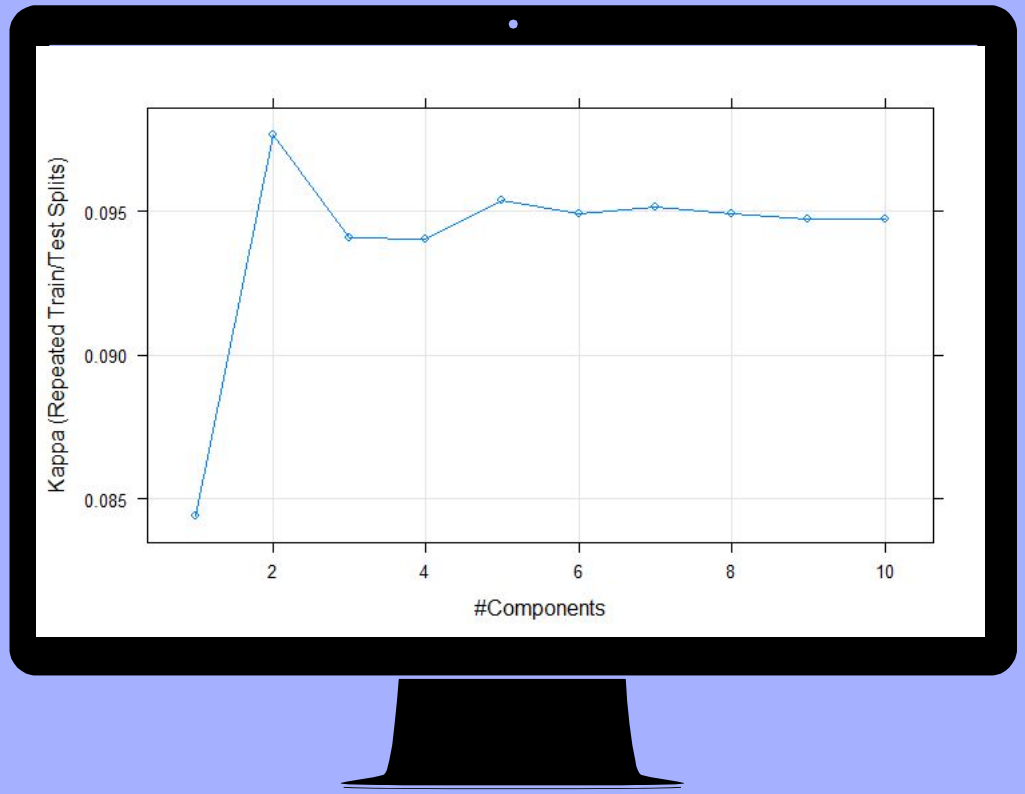


Partial Least Squares Discriminant Analysis

Components = 2

Accuracy = 59.3%

Kappa = 0.098



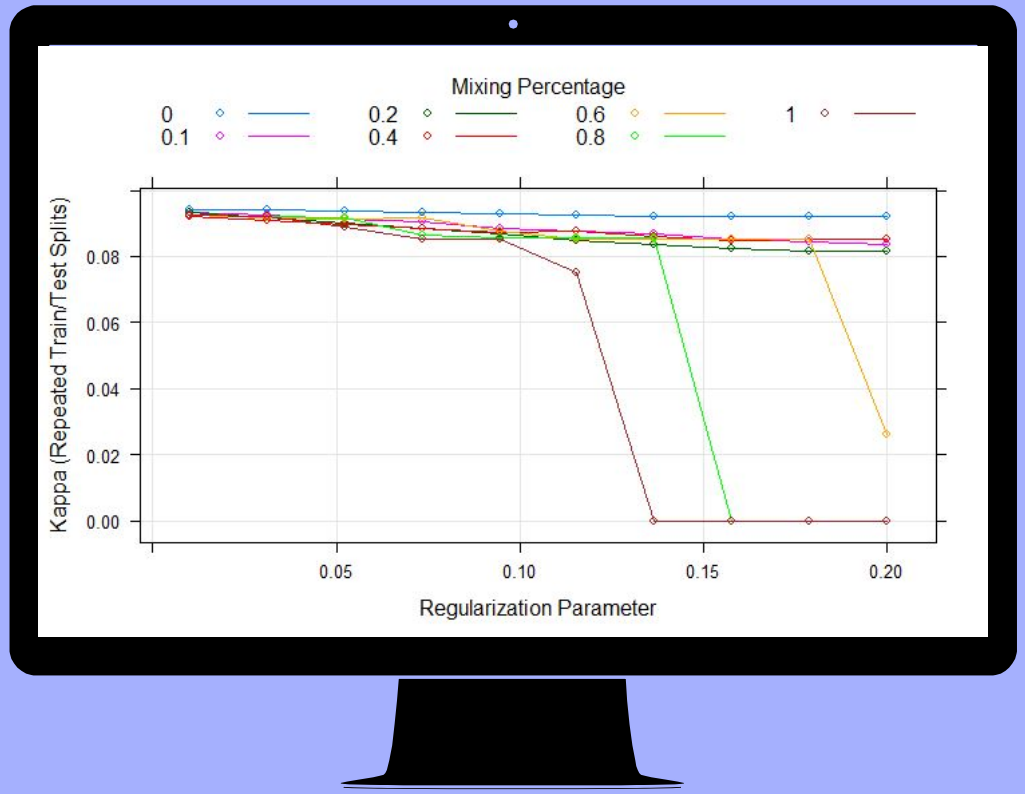
Penalized Models

Alpha = 0 (mixing percentage)

Lambda = 0.0311 (regularization)

Accuracy = 57.9%

Kappa = 0.094

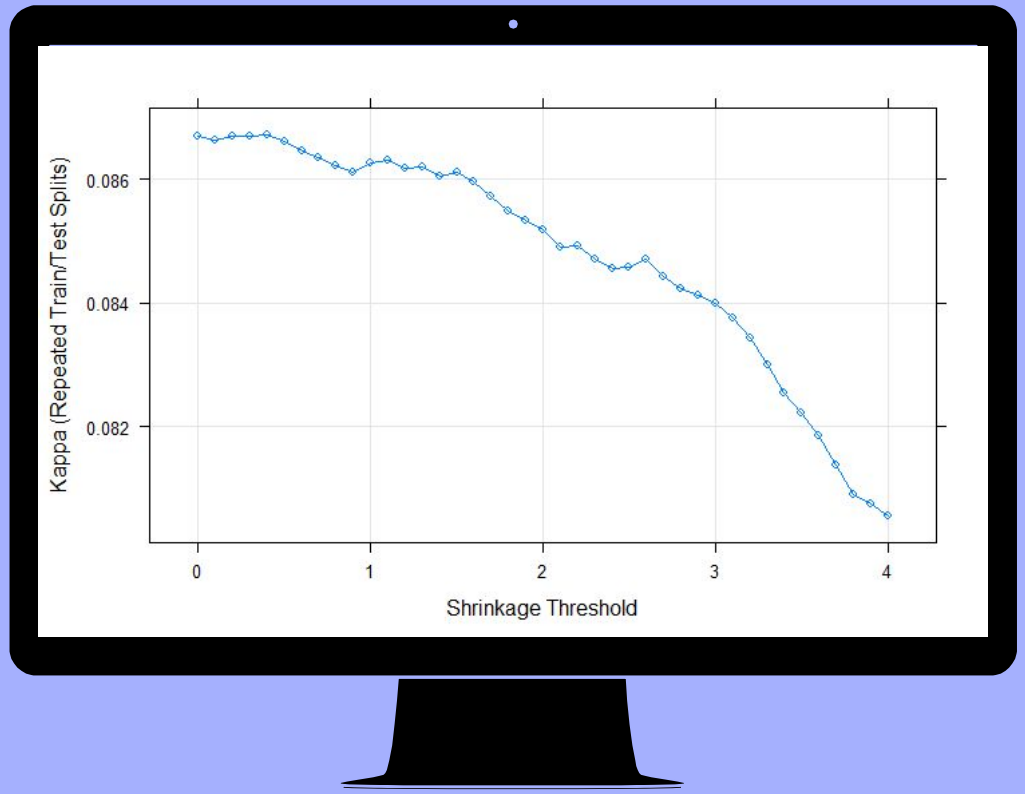


Nearest Shrunken Centroids

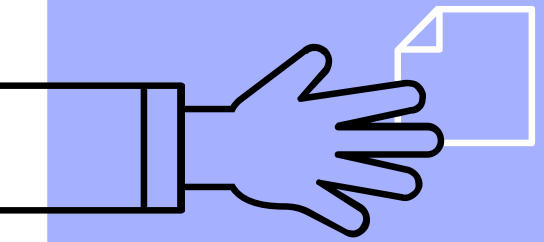
Threshold = 0.4

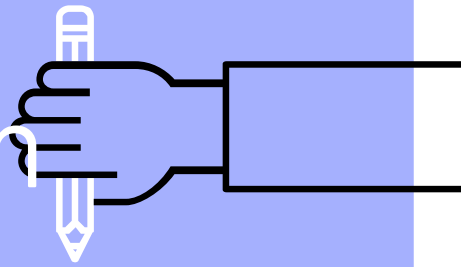
Accuracy = 54.3%

Kappa = 0.087



Nonlinear Classification Models



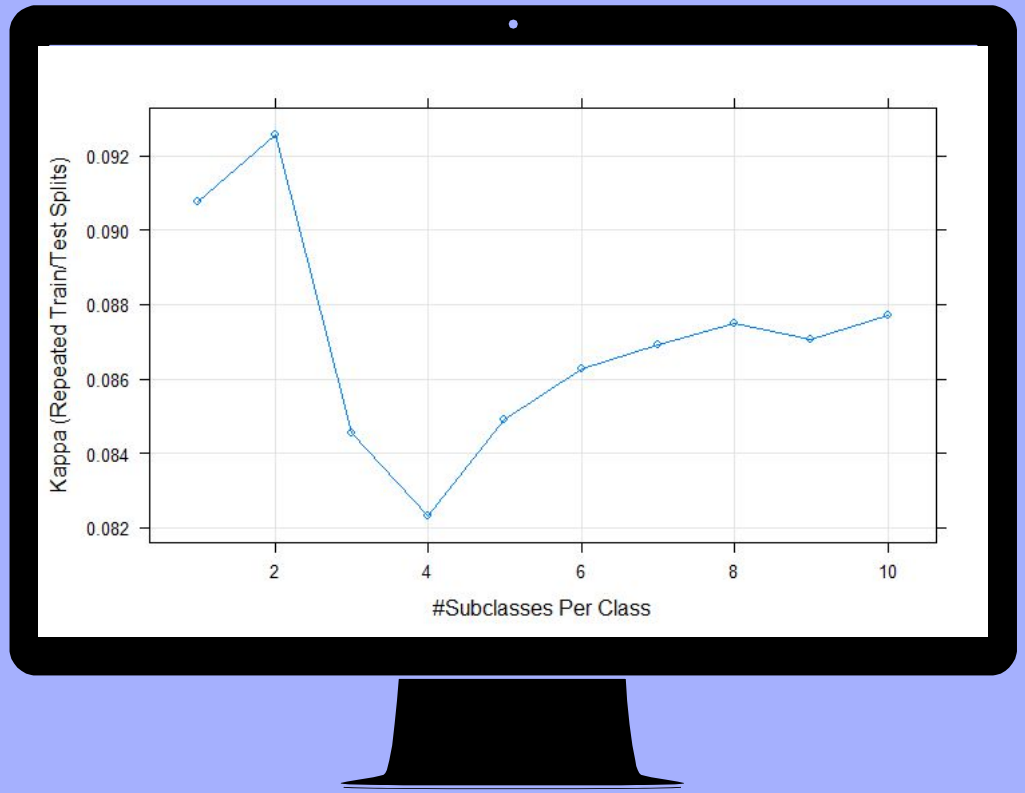


Mixture Discriminant Analysis

Subclasses = 2

Accuracy = 58.8%

Kappa = 0.093



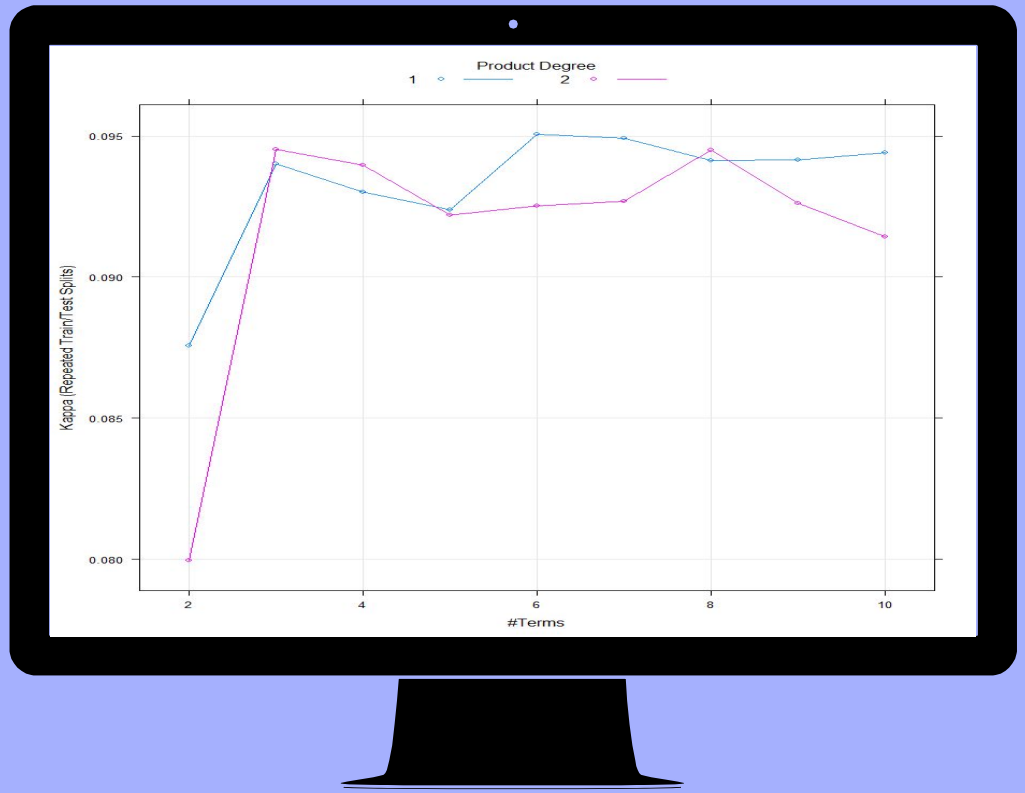
Flexible Discriminant Analysis

Degree = 1

# of Terms = 6

Accuracy : 0.5840083

Kappa : 0.09504791

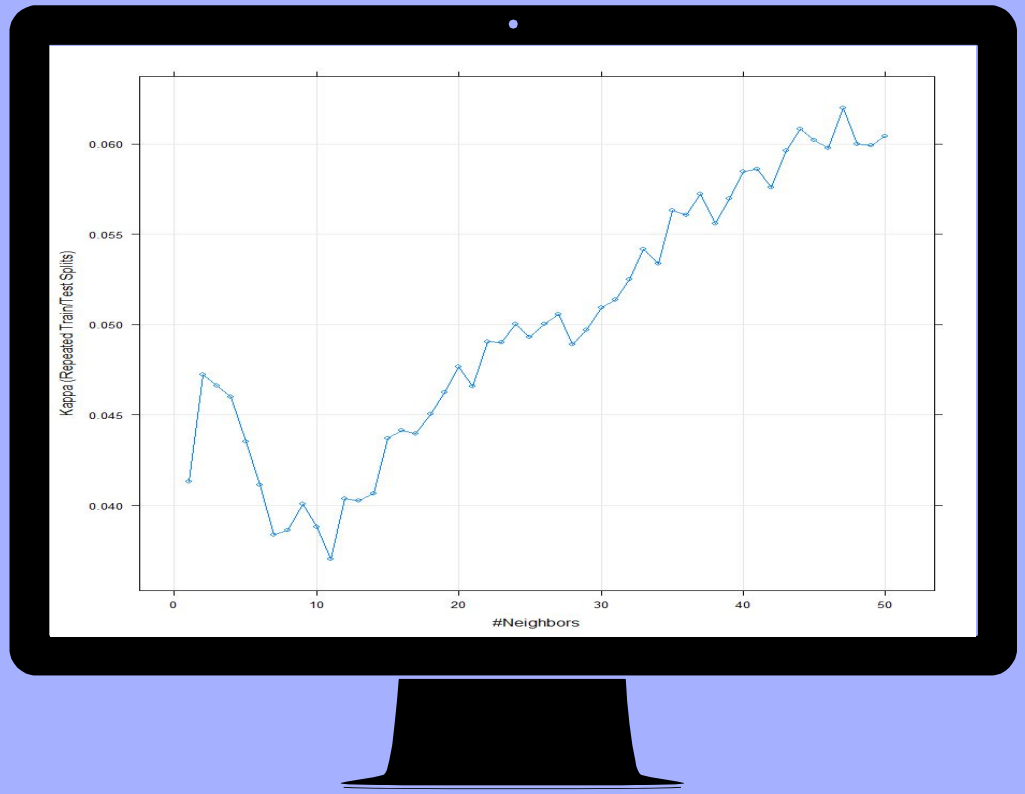


K-Nearest Neighbors

Accuracy: 0.5563286

Kappa: 0.06197283

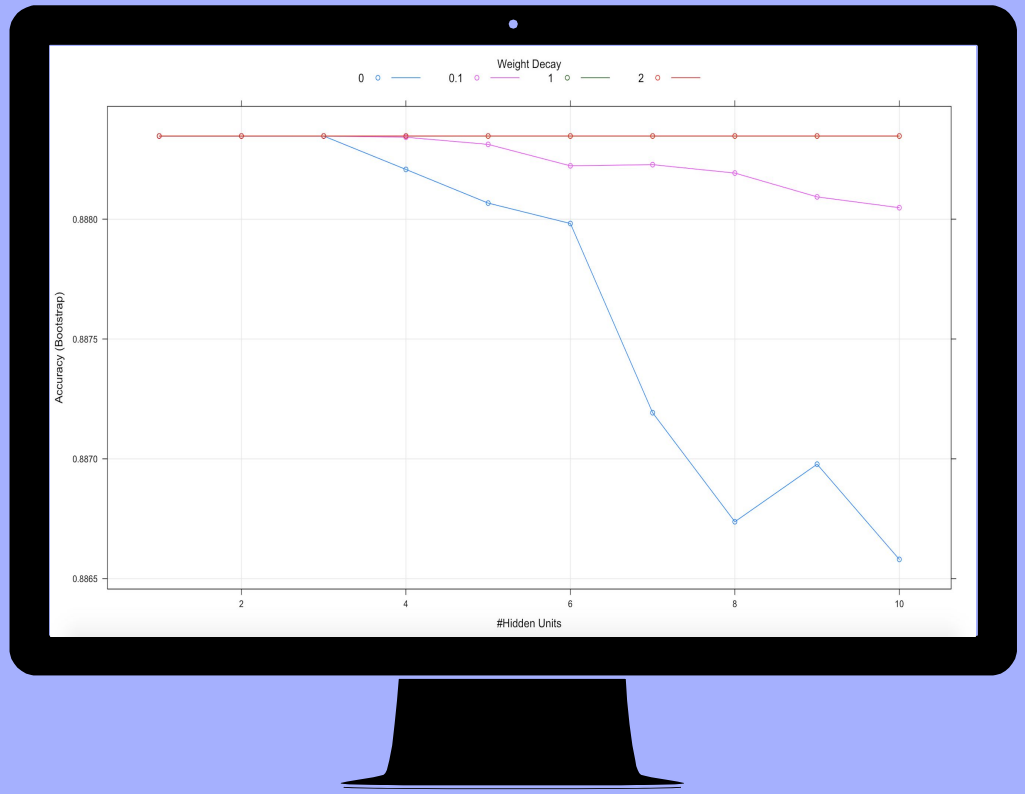
k = 47.



Neural Network

Accuracy: 0.560086

Kappa: 0.091

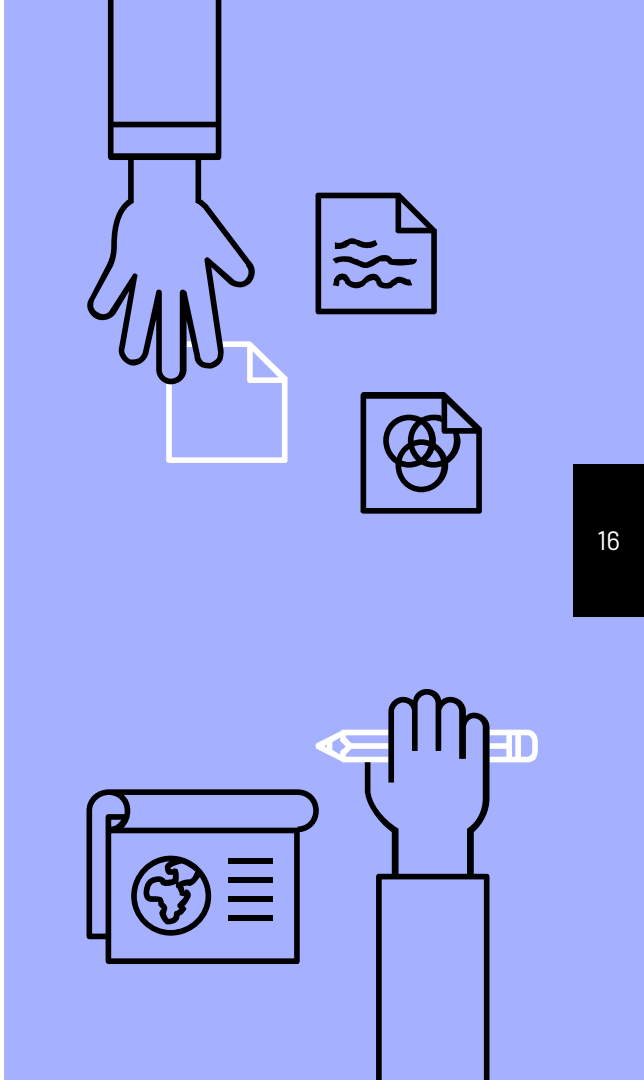


Naive Bayes

Accuracy : 0.6188

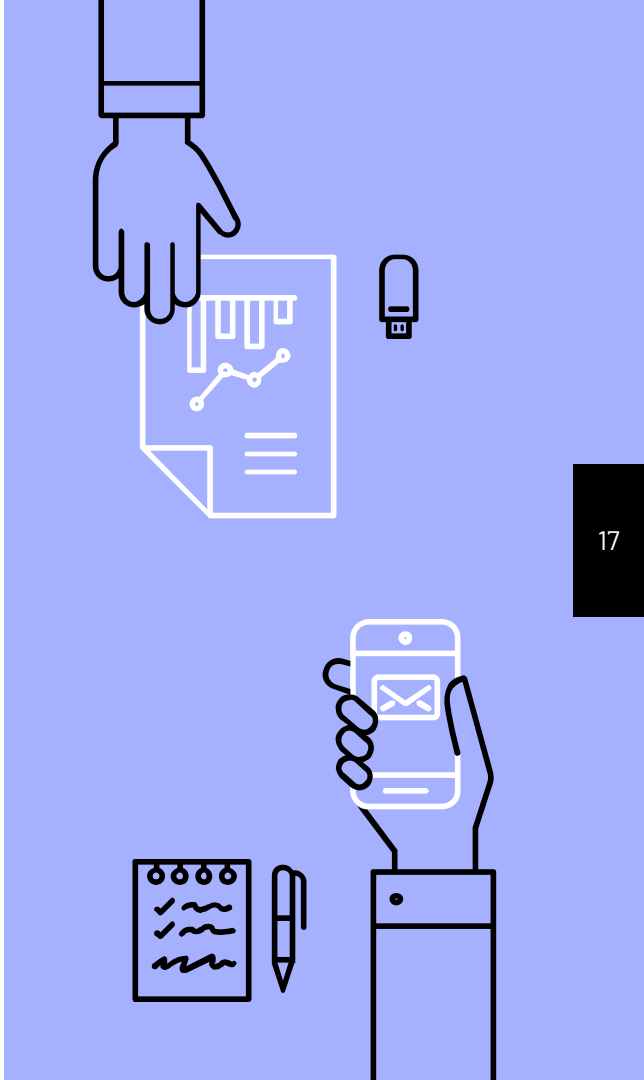
Kappa : 0.084

(no tuning parameters)



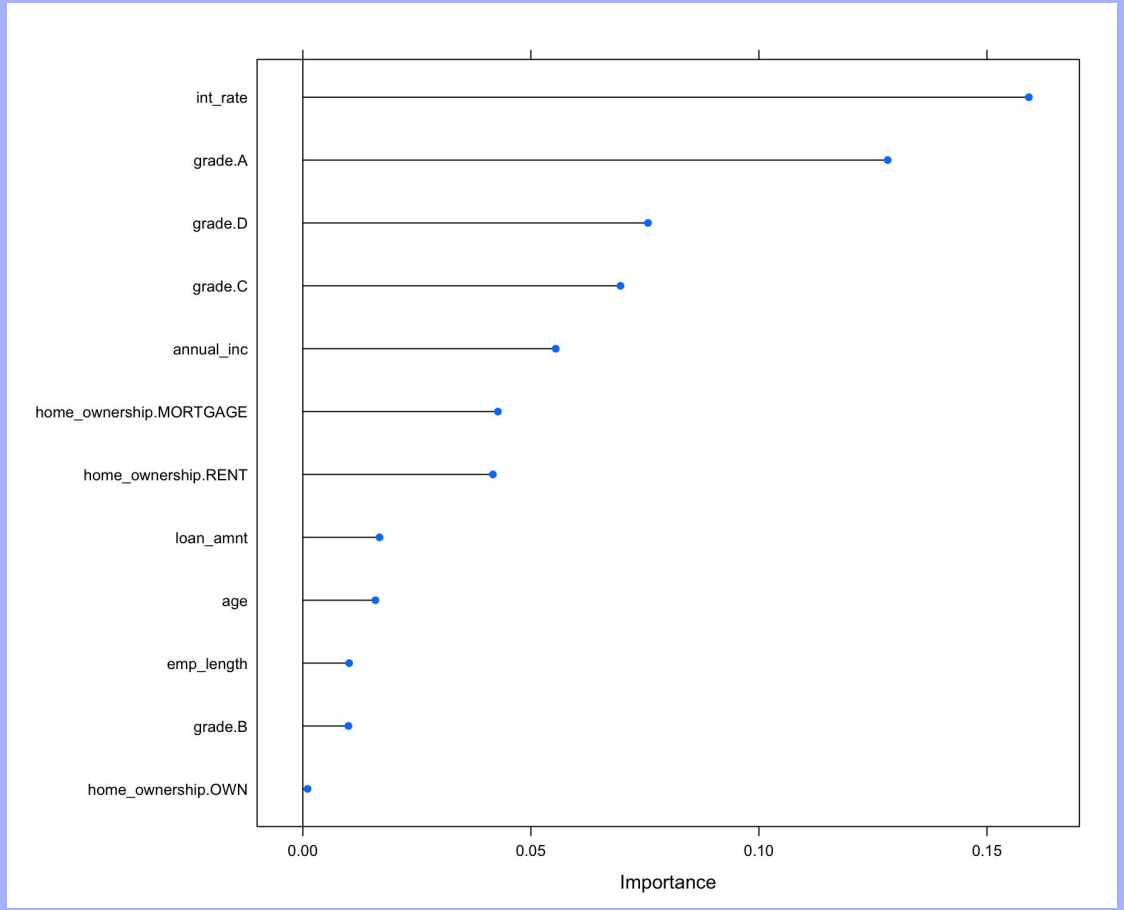
Top two models

	Accuracy	Kappa	Tuning Parameter
Partial least square discriminant analysis	59.04%	0.096	- # of comp= 2
Logistic Regression	57.7%	0.1008	none



Important predictors

(PLS-DA)



Thanks!