MDA - Final Exercises

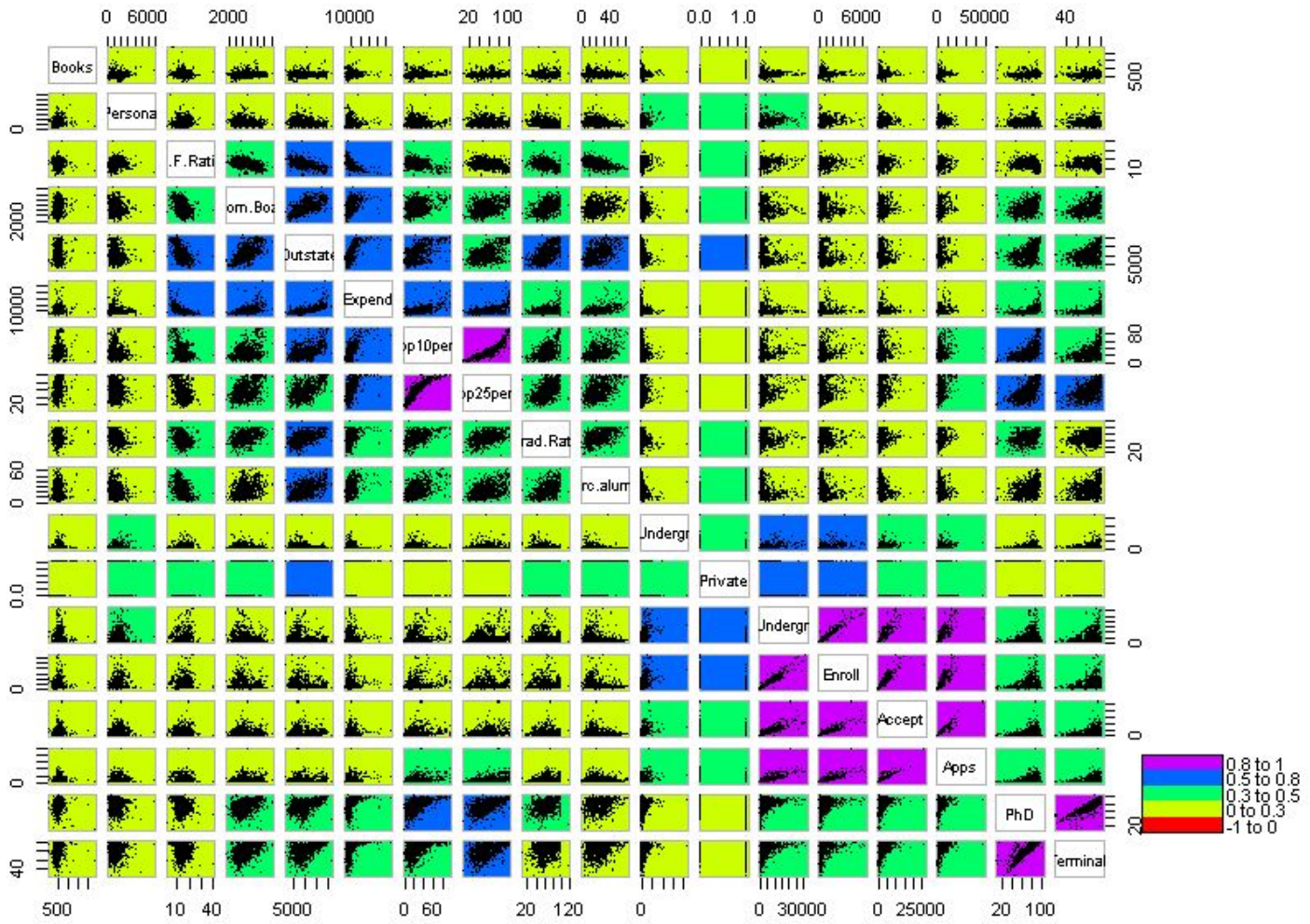
Lorenz Linhardt (a1247418)
Raphael Mitsch (a1006529)
Magdalena Schwarzl (1209910)

Regression

Regression - Data

College (ISLR)

Dependent:	Apps
Independent (17):	Private (bool) Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
Rows:	777
Condition:	pristine


Variables Ordered and Colored by Correlation
Books
Persona
.F.Rati
om.Boa
Outstate
Expend
p10per
p25per
rad.Rat
rc.alum
Undergr
Private
Undergr
Enroll
Accept
Apps
PhD
Terminal
0.8 to 1
0.5 to 0.8
0.3 to 0.5
0 to 0.3
-1 to 0

Target Variable: Apps

Min. 81

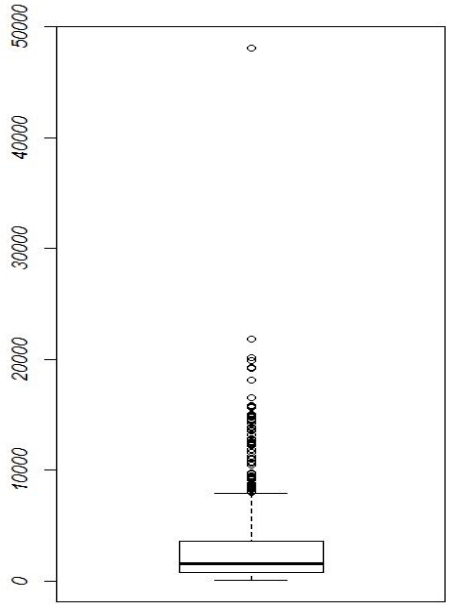
1st Qu. 776

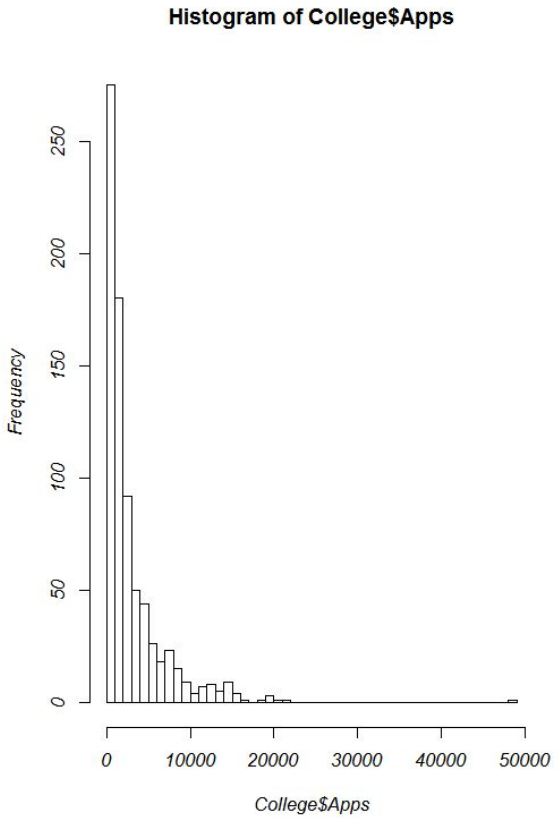
Median 1558

Mean 3002

3rd Qu. 3624

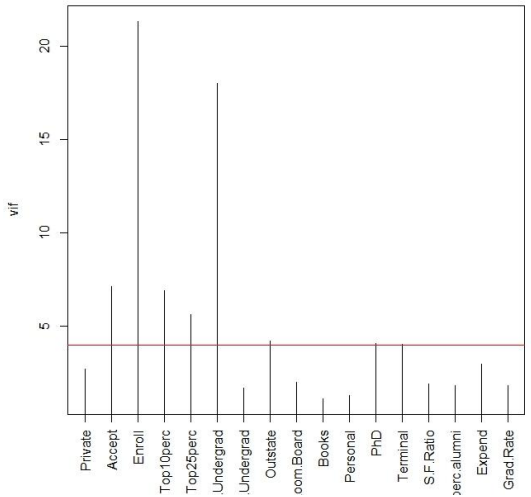
Max. 48090



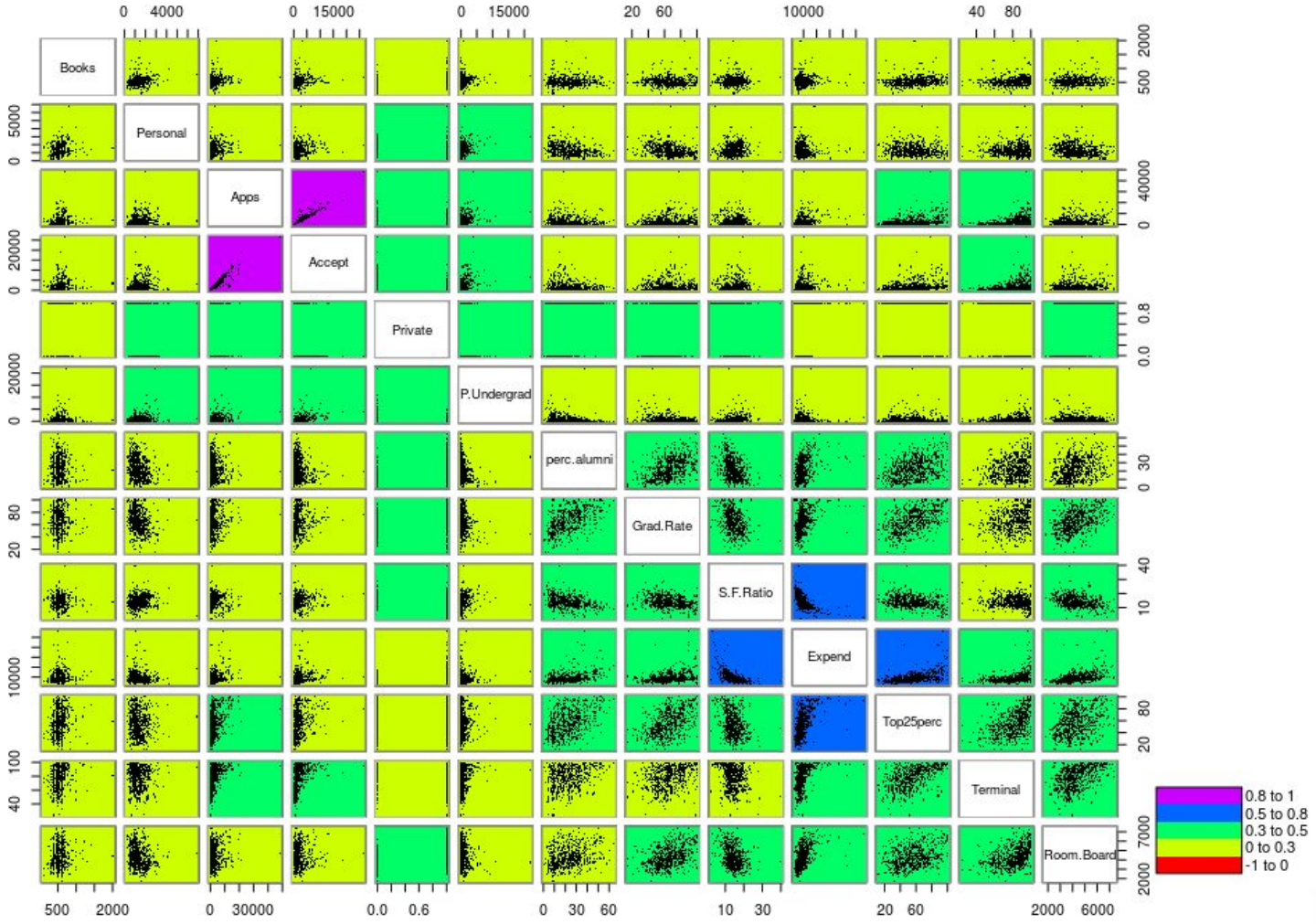


Regression - preprocessing

- Remove outliers
- Collinearity
  - remove columns with vif > 4
- Split into training (75%) and test (25%) set
- Split training set into 10 folds for cross validation

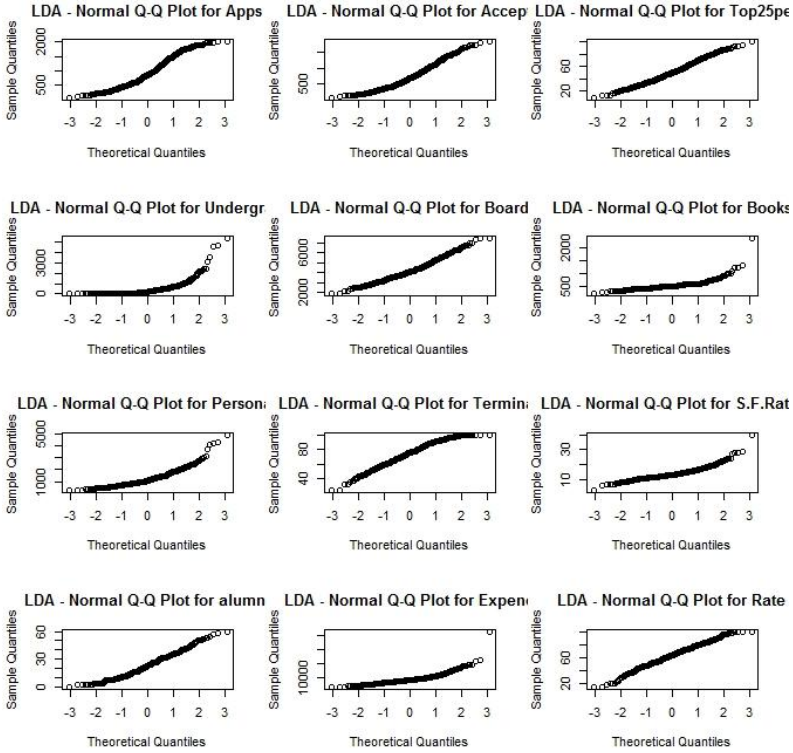


remove: Enroll, F.Undergrad, Top10perc, Outstate, PhD


Variables Ordered and Colored by Correlation
Books
Personal
Apps
Accept
Private
P.Undergrad
perc.alumni
Grad.Rate
S.F.Ratio
Expend
Top25perc
Terminal
Room.Board
0.8 to 1
0.5 to 0.8
0.3 to 0.5
0 to 0.3
-1 to 0

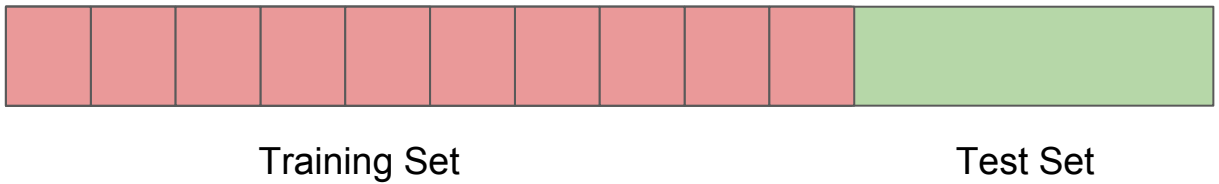
Regression - Model Assumptions

- Independence of data → collinearity
- Normality of errors (QQ-plots, Shapiro-Wilkinson)
- Heteroskedasticity (Breusch-Pagan)
- Differences in covariance matrices (in dependence of private)



Evaluation Strategy

- Training set (75%): Model selection, parameter tuning
- Compare **RMSE**
- Test set (25%): Evaluate best model, trained on training set
- Training without outliers (except rob. regr.), test with outliers



Regression - Models

method	rmse (full dataset)	method	rmse
k-nn	1110.486	lts -fit	1554.008 (1458.397)
cubic spline	1176.283	robust (Huber)	1530.724 (1208.946)
linear regression	1528.623	robust (bisqu)	1559.48 (1336.591)
loess	1153.059	log transform	4798.845
regression tree	3938.292	box cox	4710.03

Regression - best model: k-nn

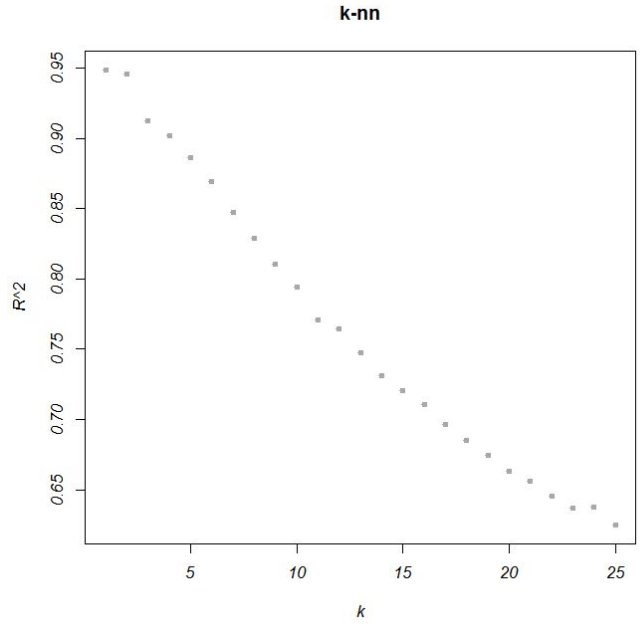
k = 1

library: FNN

Preprocessing: boolean to numeric {0,1}

+ Reduzierte Attribute

RMSE on training set: 1110.486
RMSE on test set: 746.334

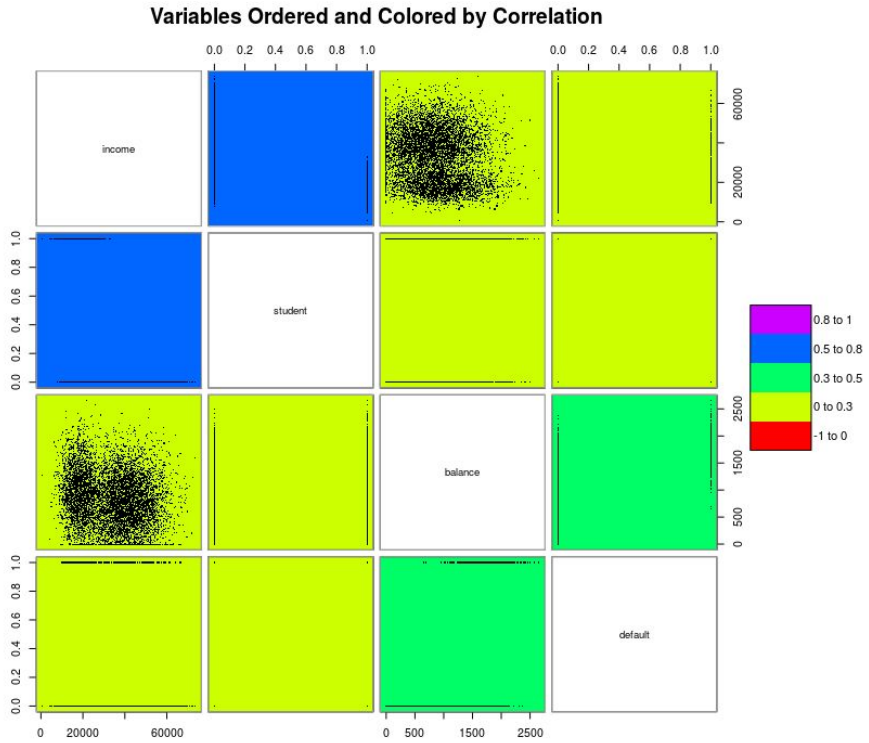


Classification

Classification - Data

default (ISLR)

Dependent:	default (bool)
Independent (3):	student (bool) balance income
Rows:	10.000
Class proportions:	9.667:333 ~ 29:1
Condition:	pristine

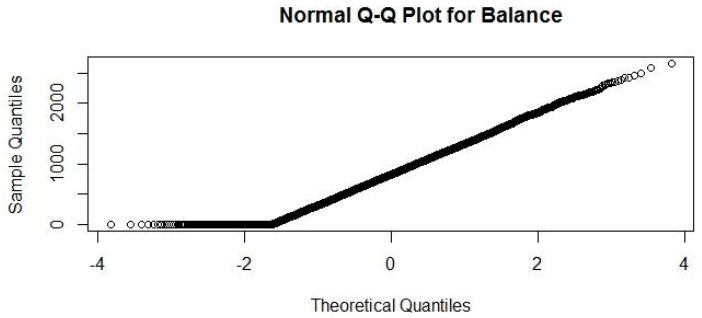


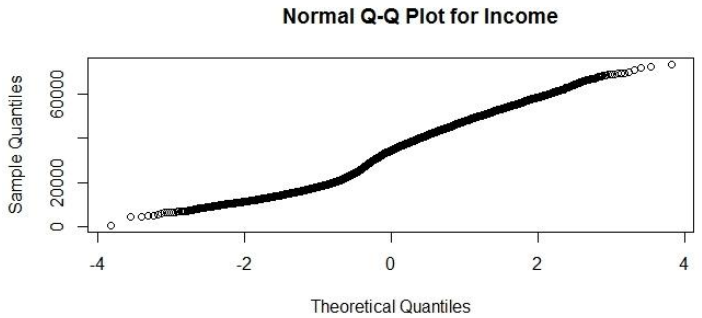
Classification - Preprocessing

- Split into training (75%) and test (25%) set
- Split training set into 10 folds for cross validation
- Stratified folds - balanced classes

Classification - Model Assumptions

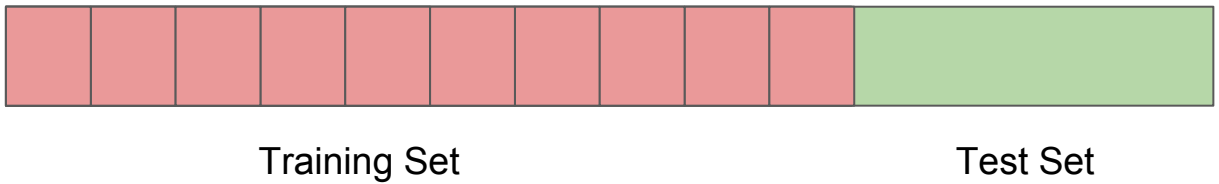
- Independence of data → collinearity
- Normality of data (QQ-plots, Shapiro-Wilkinson)
- Heteroskedasticity (Breusch-Pagan)
- Differences in covariance matrices
- Indication of influence by student status





Evaluation Strategy

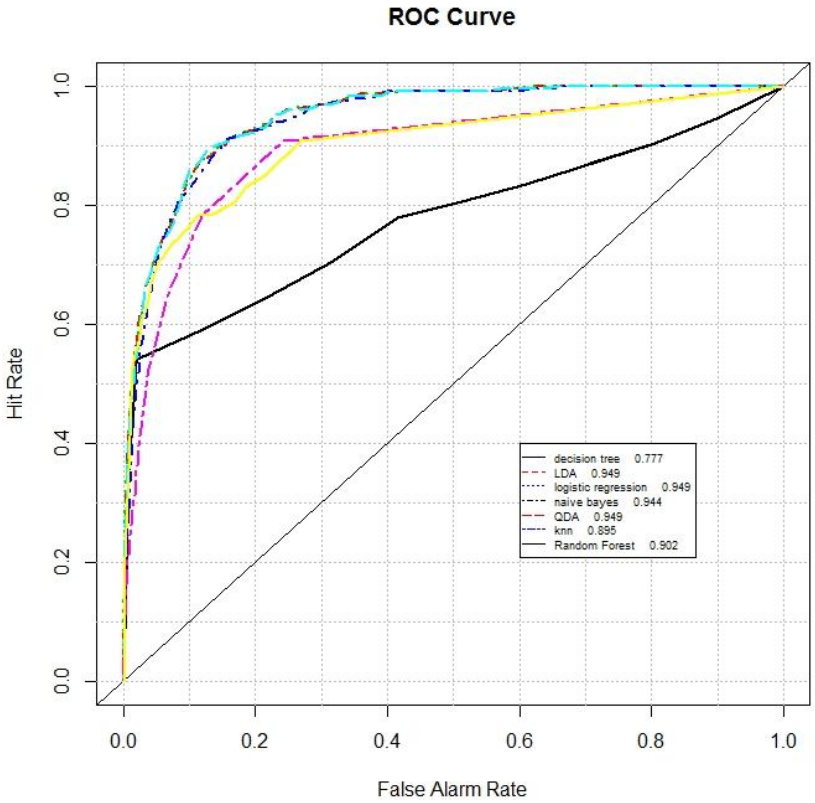
- Training set (75%): Model selection, parameter tuning
- Compare **AUC, accuracy, balanced accuracy**
- Test set (25%): Evaluate best model, trained on training set

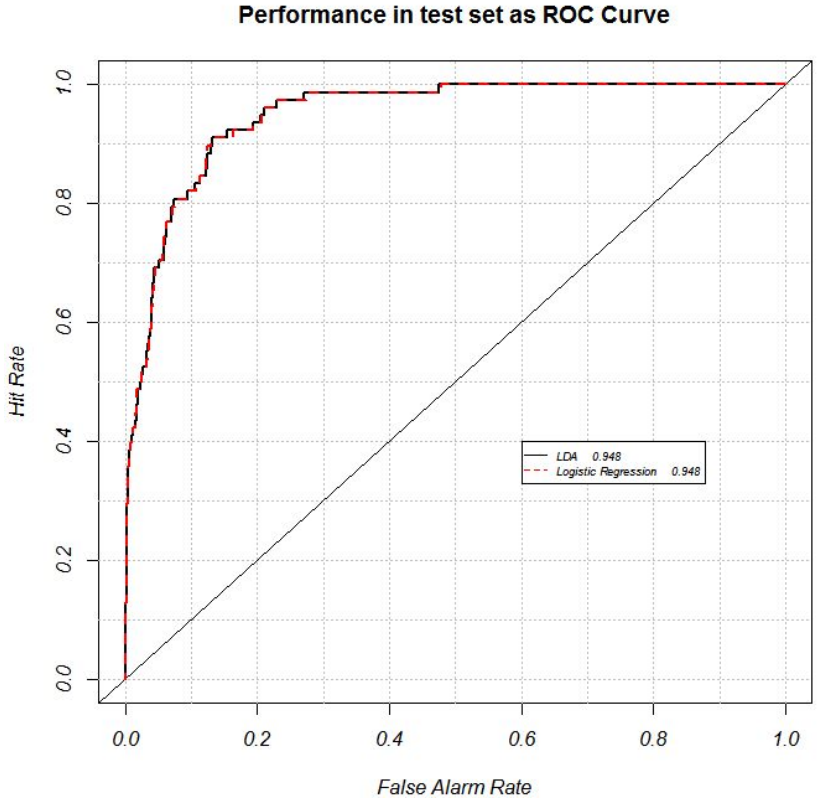


Classification - Models

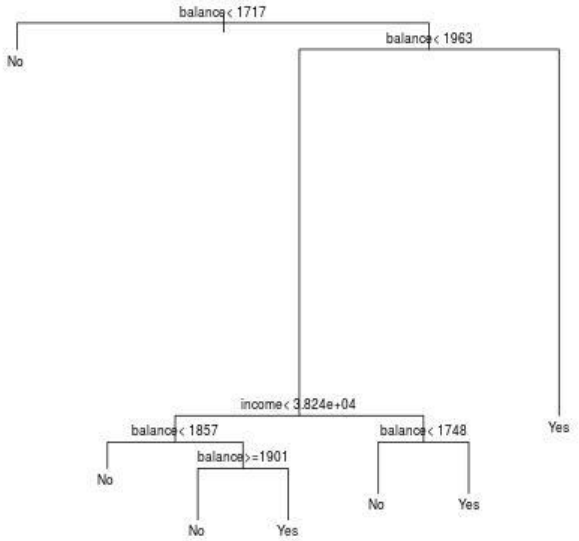
method	AUC	Accuracy	Balanced
k-nn	0.895	0.9512	0.6342
Naive Bayes	0.944	0.9704	0.6404
Decision Tree	0.777	0.9709	0.6634
LDA	0.949	0.9744	0.6208
Logistic Regression	0.949	0.9744	0.6456
QDA	0.949	0.9727	0.6453
Random Forest	0.902	0.9705	0.6253

Classification - Evaluation

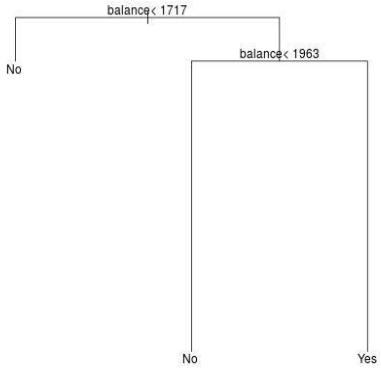




Classification - best model: Decision Tree



original



post- pruned

Classification - best model: Decision Tree

Confusion Matrix and Statistics

Prediction	Reference 1	2
1	7202	172
2	43	83

On training set (CV)

Confusion Matrix and Statistics

	No	Yes
No	2415	59
Yes	7	19

On test set (trained on training set)