CAPSTONE-PH125.9x-Indian Patient Liver Diseases

Laxmansingh Rajput

6/3/2019

OBJECTIVES

Creating a recommendation system using the Indian Liver Patient dataset. Train a machine learning algorithm using the inputs in one subset to predict liver patients in the validation set.

This project is being completed in R version 3.6.0 (*Planting of a Tree*)

The submission for the Indian Liver Patient project will be three files:

- A report in the form of an Rmd file
- A report in the form of a PDF document knit from your Rmd file
- An Rmd/R script file that generates the predicted movie ratings and calculates RMSE

The overall accuracy is used as predictions are compared to the value in the validation.

EXECUTIVE SUMMARY

As this is a classification problem only certain models are being considered from the caret package.

1. Dataset
    a. Indian Liver Patient dataset from the Kaggle curated datasets (more information on next page)
2. Goals of the project
    a. Predict the patients with liver disease for the validation dataset
    b. Compute the accuracy based on the confusion matrix
3. Key Steps
    a. Explore the dataset
    b. Identify and impute missing values
        i. Use 'Mice' (Multivariate Imputation by Chained Equations) package to identify and impute missing values
        ii. Use VIM and marginplot to display the missing value (Okay - just exploring the various packages/libraries)
    c. Split the dataset into train and test
        i. Convert the Predictor information from 1 and 2 to "YES" and "NO"
        ii. Check for imbalance in the dataset (using receiver operating characteristic (ROC curves) and area under the curve (AUC))
    d. Check for correlation between the features
    e. Principle Component Analysis
    f. Identify the top 3 features using Random Forest
    g. Train the model using select set of models
    h. Make predictions the dataset based on the test set
    i. Identify the top model and corresponding accuracy

Dataset Information

Indian Liver Patient Records - Patient records collected from North East of Andhra Pradesh, India

Context

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.

Content

This data set contains 416 liver patient (1) records and 167 non liver patient (2) records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.

Any patient whose age exceeded 89 is listed as being of age "90".

Columns:

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Protiens
- Albumin
- Albumin and Globulin Ratio

Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

Acknowledgements: This dataset was downloaded from the UCI ML Repository: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Inspiration Use these patient records to determine which patients have liver disease and which ones do not.

METHODS & ANALYSIS

Explore the Indian Liver Patient dataset prior to splitting (train and test)

1. First 10 rows of the dataset

	1	2	3	4	5	6	7	8	9	10
Age	65	62	62	58	72	46	26	29	17	55
Gender	Female	Male	Male	Male	Male	Male	Female	Female	Male	Male
Total_Bilirubin	0.7	10.9	7.3	1.0	3.9	1.8	0.9	0.9	0.9	0.7
Direct_Bilirubin	0.1	5.5	4.1	0.4	2.0	0.7	0.2	0.3	0.3	0.2
Alkaline_Phosphotase	187	699	490	182	195	208	154	202	202	290
Alamine_Aminotransferase	16	64	60	14	27	19	16	14	22	53
Aspartate_Aminotransferase	18	100	68	20	59	14	12	11	19	58
Total_Protiens	6.8	7.5	7.0	6.8	7.3	7.6	7.0	6.7	7.4	6.8
Albumin	3.3	3.2	3.3	3.4	2.4	4.4	3.5	3.6	4.1	3.4
Albumin_and_Globulin_Ratio	0.90	0.74	0.89	1.00	0.40	1.30	1.00	1.10	1.20	1.00
Dataset	1	1	1	1	1	1	1	1	2	1

Note:
First 10 rows displayed

2. The list of the variables.

Variable Names	Mean
Age	44.746140651801
Gender	NA
Total_Bilirubin	3.29879931389365
Direct_Bilirubin	1.4861063464837
Alkaline_Phosphotase	290.576329331046
Alamine_Aminotransferase	80.7135506003431
Aspartate_Aminotransferase	109.910806174957
Total_Protiens	6.48319039451115
Albumin	3.14185248713551
Albumin_and_Globulin_Ratio	0.94706390328152
Dataset	NA

Note:
Mean computed with na.rm = TRUE.
[1] Albumin_and_Globulin_Ratio has 4 missing values.

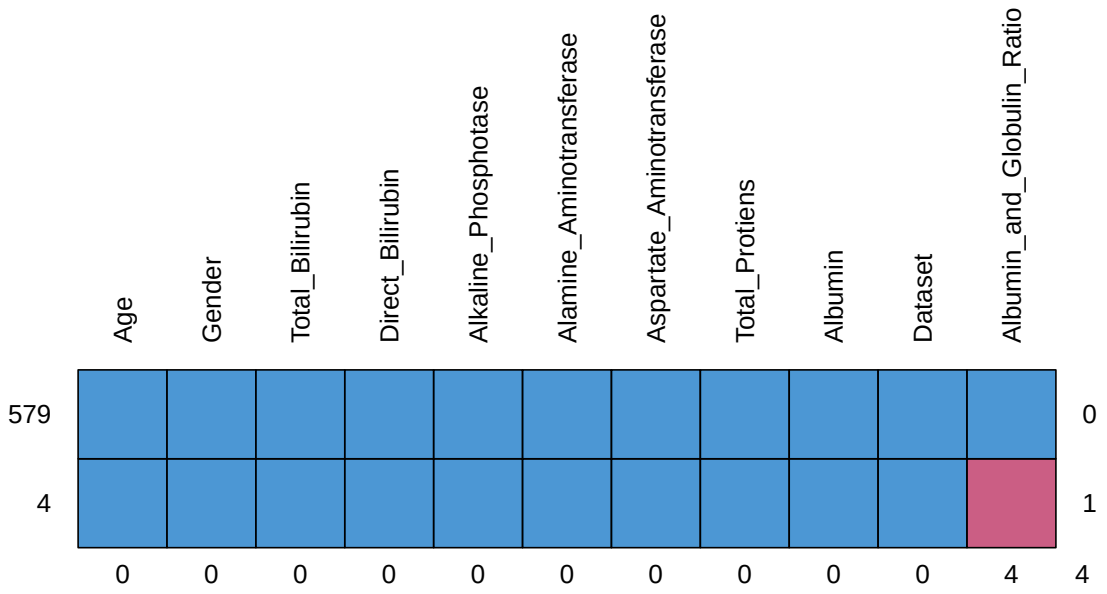
3. Describe the data

variable	missing	complete	n	n_unique	top_counts	ordered
Dataset	0	583	583	2	1: 416, 2: 167, NA: 0	FALSE
Gender	0	583	583	2	Mal: 441, Fem: 142, NA: 0	FALSE

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
Age	0	583	583	44.75	16.19	4	33	45	58	90
Alamine_Aminotransferase	0	583	583	80.71	182.62	10	23	35	60.5	2000
Albumin	0	583	583	3.14	0.8	0.9	2.6	3.1	3.8	5.5
Albumin_and_Globulin_Ratio	4	579	583	0.95	0.32	0.3	0.7	0.93	1.1	2.8
Alkaline_Phosphotase	0	583	583	290.58	242.94	63	175.5	208	298	2110
Aspartate_Aminotransferase	0	583	583	109.91	288.92	10	25	42	87	4929
Direct_Bilirubin	0	583	583	1.49	2.81	0.1	0.2	0.3	1.3	19.7
Total_Bilirubin	0	583	583	3.3	6.21	0.4	0.8	1	2.6	75
Total_Protiens	0	583	583	6.48	1.09	2.7	5.8	6.6	7.2	9.6

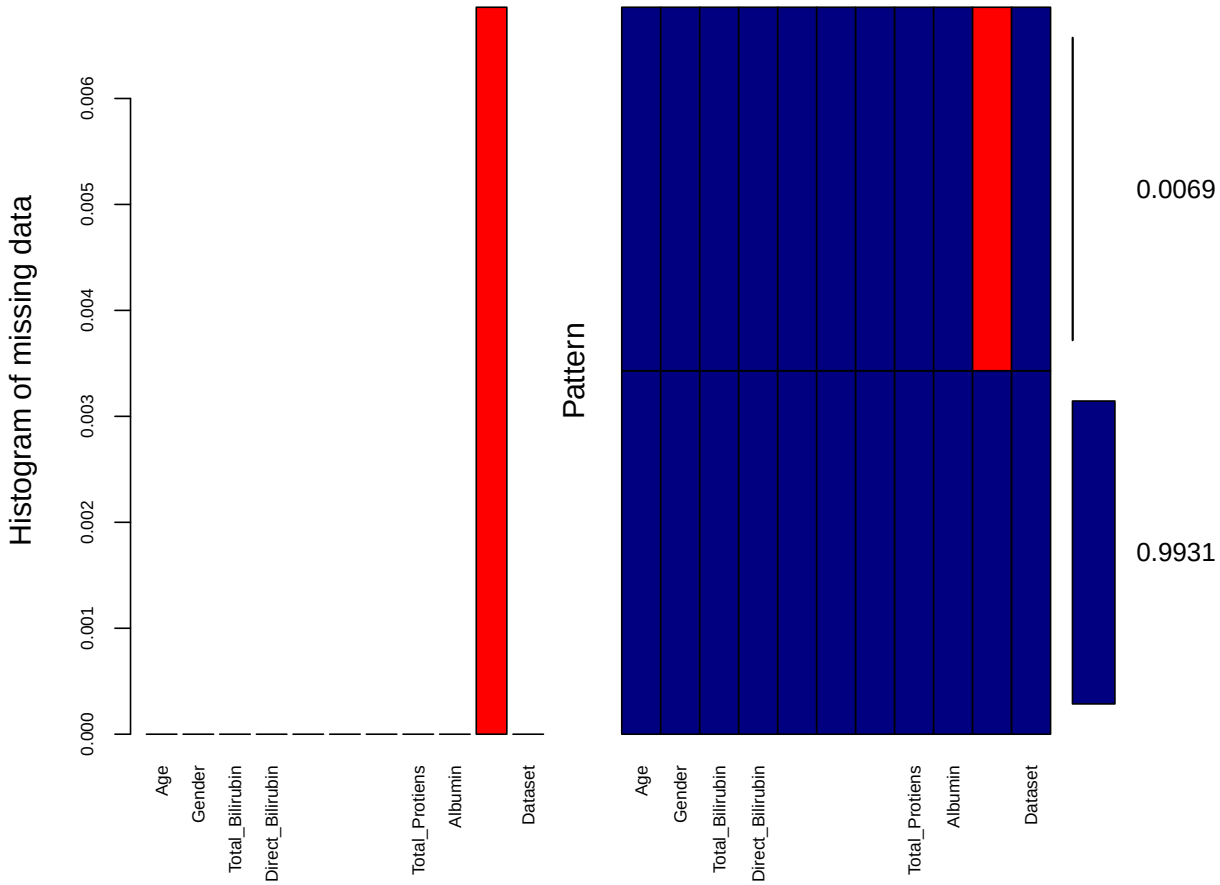
Missing values and Imputation

Find the missing value(s) and use various plots to explore the dataset

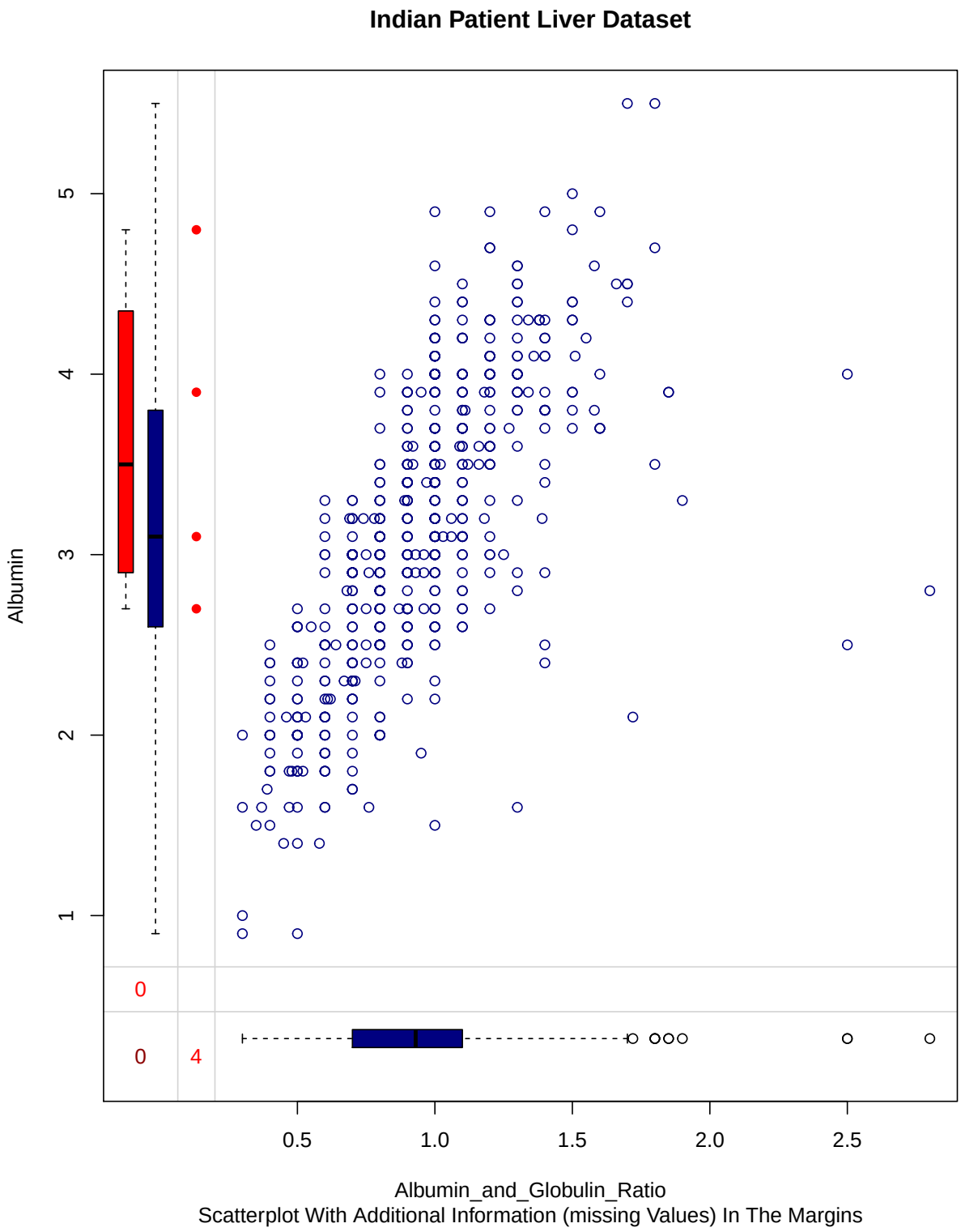


As can be seen in the chart above the feature “Albumin_and_Globulin_Ratio” is missing 4 values.

A perhaps more helpful visual representation can be obtained using the VIM package as follows:



Another helpful visual approach (it is possible to explore all the features against one another)



Implementing the imputation in 2-steps, using mice() to build the model and complete() to generate the completed data.

```
[1] "Mean Pre Imputation of Albumin_and_Globulin_Ratio 0.94706390328152"
```

```
[1] "Mean Post Imputation of Albumin_and_Globulin_Ratio 0.947358490566038"
```

Missing Index	Pre Imputation	Post Imputation
210	NA	1.16
242	NA	0.90
254	NA	0.90
313	NA	1.00

Check the training set for class imbalance

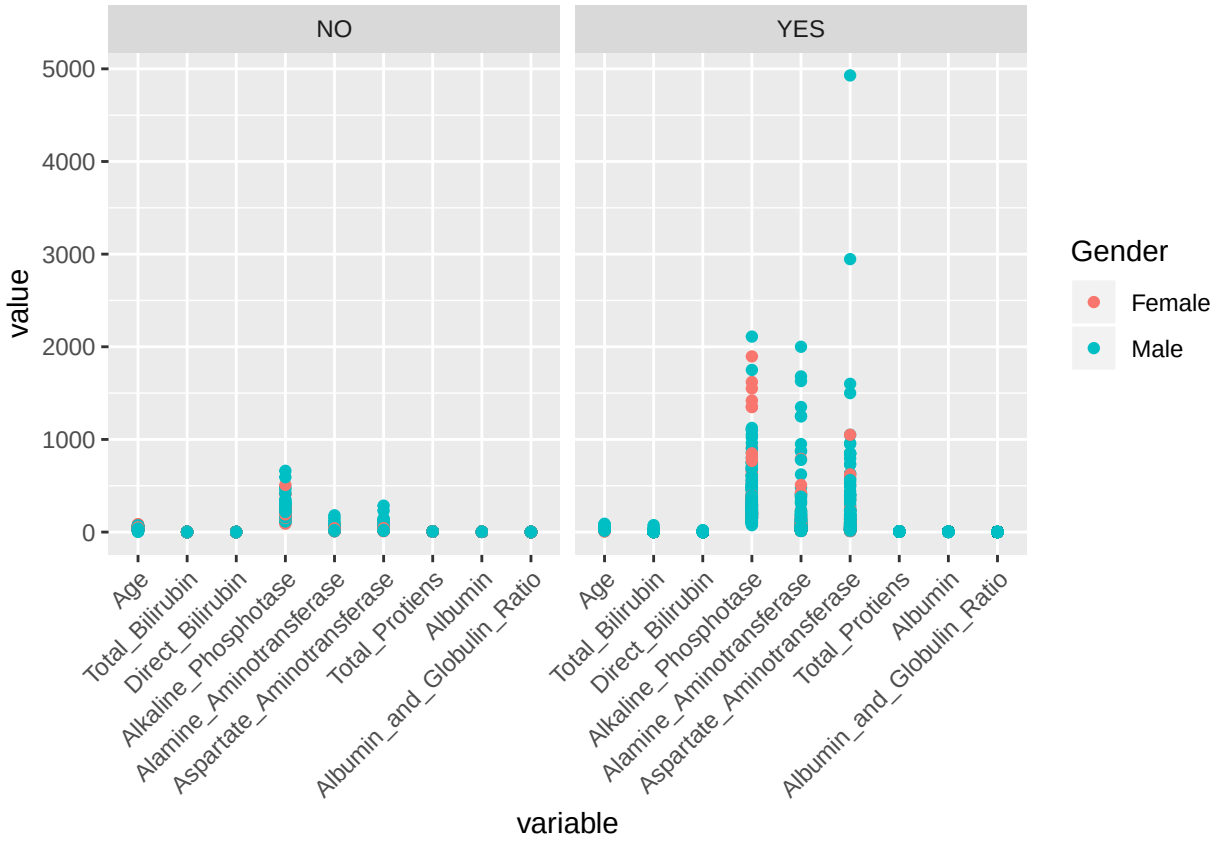
Cross Tabulation counts/Marginal Table

Value	Count	Percentage
NO	150	0.2862595
YES	374	0.7137405

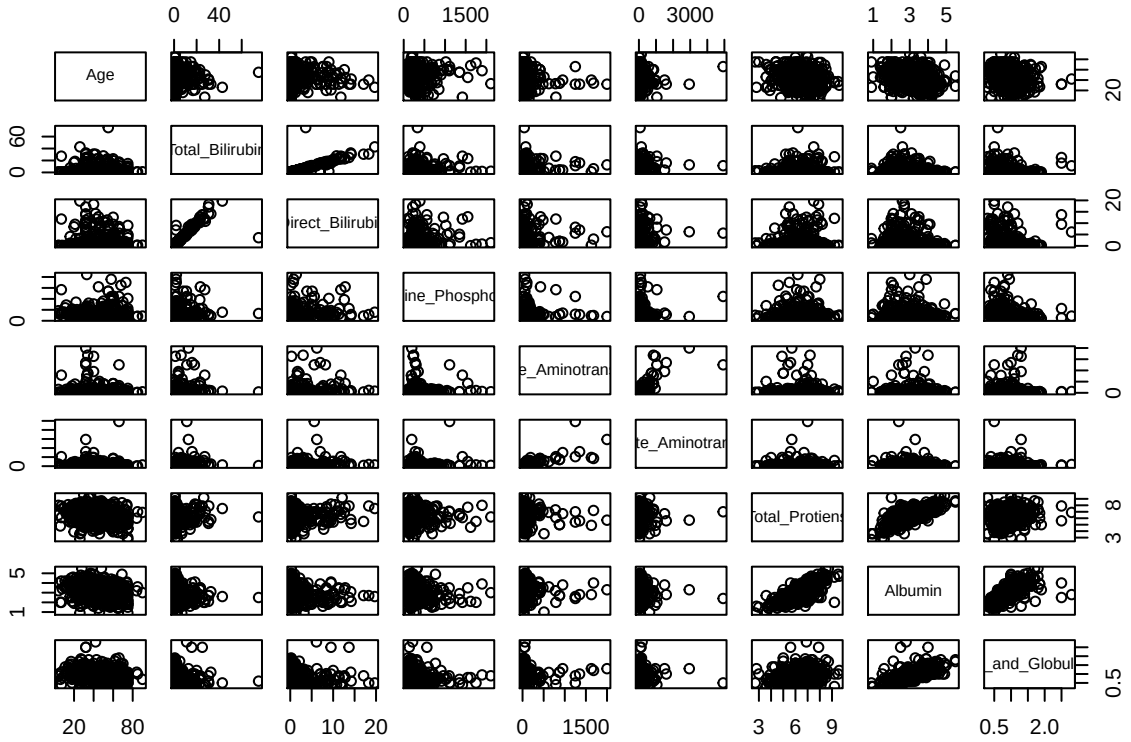
More information - Accuracy measurements/Area under the Curve

Description	Value
precision	0.7021277
recall	0.7857143
F	0.3707865
Area under the curve (AUC)	0.5952381

Visualize the data



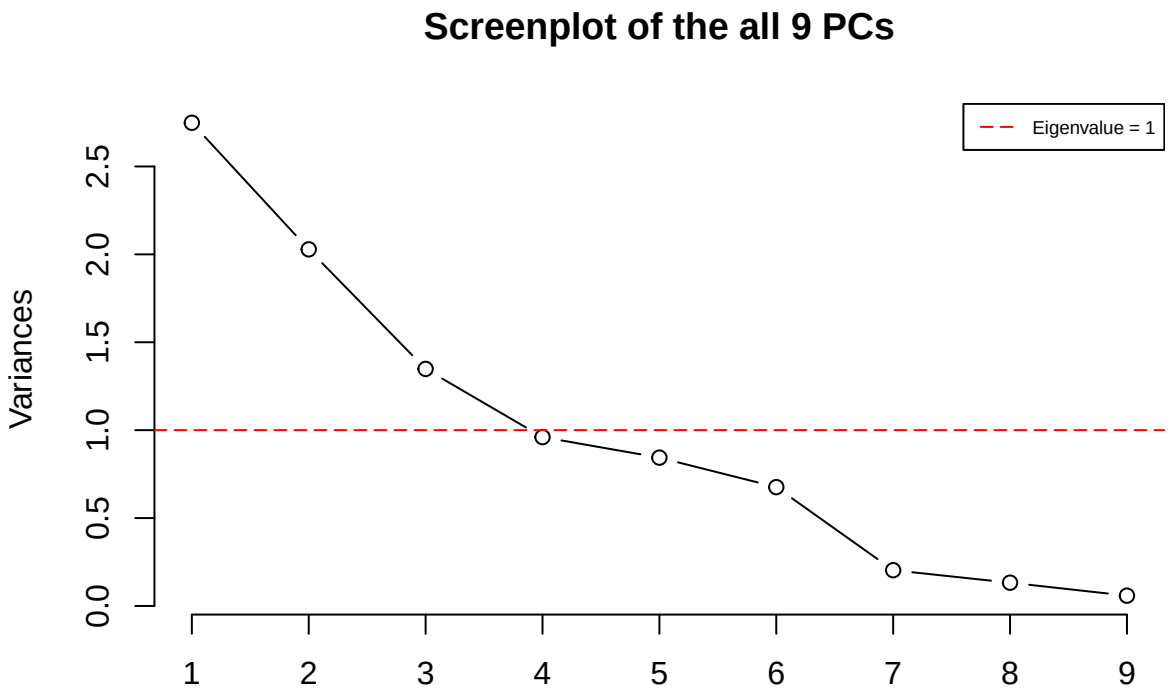
Correlation

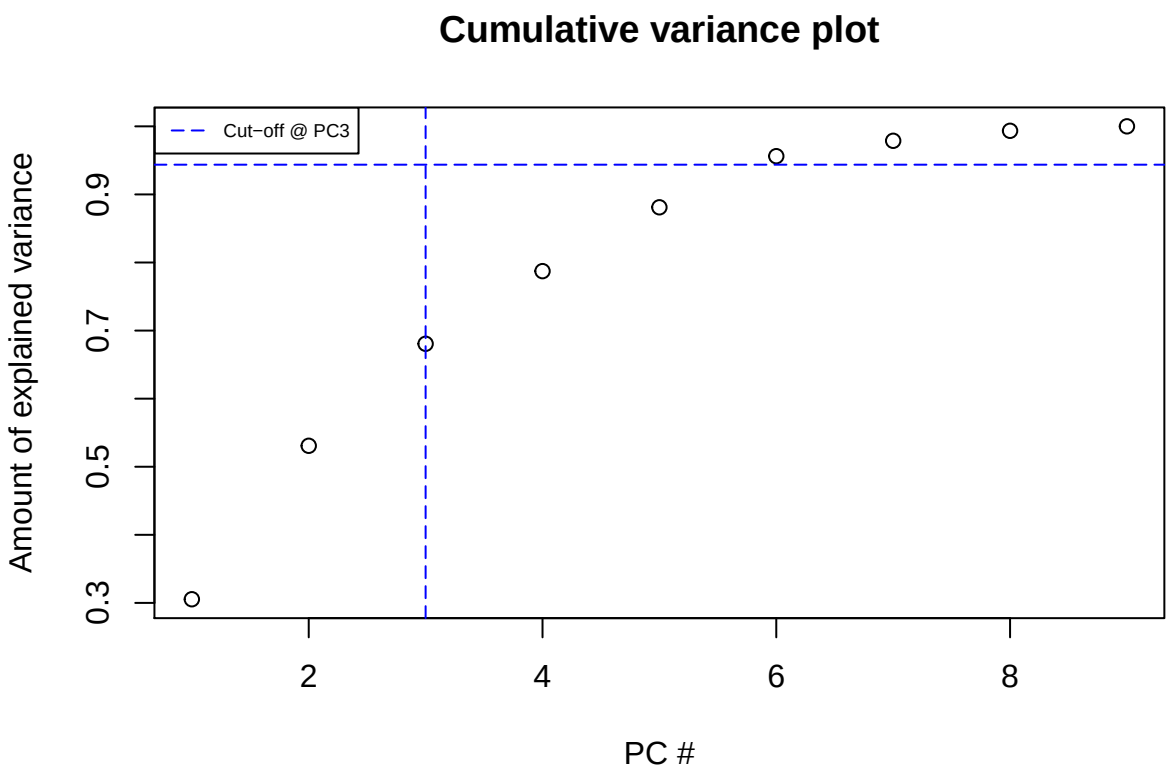


	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio
Age	1.0000	0.0075	0.0019	0.0781	-0.0854	-0.0142	-0.1867	-0.2586	-0.2029
Total_Bilirubin	0.0075	1.0000	0.8679	0.2110	0.2198	0.2462	0.0015	-0.2257	-0.2084
Direct_Bilirubin	0.0019	0.8679	1.0000	0.2437	0.2437	0.2705	0.0122	-0.2320	-0.2010
Alkaline_Phosphotase	0.0781	0.2110	0.2437	1.0000	0.1326	0.1839	-0.0173	-0.1571	-0.2289
Alamine_Aminotransferase	-0.0854	0.2198	0.2437	0.1326	1.0000	0.7913	-0.0408	-0.0288	-0.0016
Aspartate_Aminotransferase	-0.0142	0.2462	0.2705	0.1839	0.7913	1.0000	-0.0183	-0.0814	-0.0688
Total_Protiens	-0.1867	0.0015	0.0122	-0.0173	-0.0408	-0.0183	1.0000	0.7807	0.2172
Albumin	-0.2586	-0.2257	-0.2320	-0.1571	-0.0288	-0.0814	0.7807	1.0000	0.6714
Albumin_and_Globulin_Ratio	-0.2029	-0.2084	-0.2010	-0.2289	-0.0016	-0.0688	0.2172	0.6714	1.0000

Principal Component Analysis (PCA)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.657869	1.424267	1.161196	0.9800466	0.9185626	0.8221695	0.4509479	0.3638875	0.2420298
Proportion of Variance	0.305390	0.225390	0.149820	0.1067200	0.0937500	0.0751100	0.0225900	0.0147100	0.0065100
Cumulative Proportion	0.305390	0.530780	0.680600	0.7873300	0.8810800	0.9561800	0.9787800	0.9934900	1.0000000




Cumulative variance plot
Cut−off @ PC3
Amount of explained variance
0.3
0.5
0.7
0.9
2
4
6
8
PC #

Random Forest to identify the variable that are most important

Top 3 variables
Alkaline_Phosphotase
Aspartate_Aminotransferase
Age

RESULTS

Train the following models:

- adaboost
- avNNet
- LogitBoost
- lda
- loclda
- naive_bayes
- wsrf
- gamLoess
- kknn
- knn
- monmlp
- mlp
- mlpML
- nnet
- svmLinear3
- svmLinear
- svmRadial
- svmRadialCost
- svmRadialSigma
- rf
- Rborist
- nodeHarvest

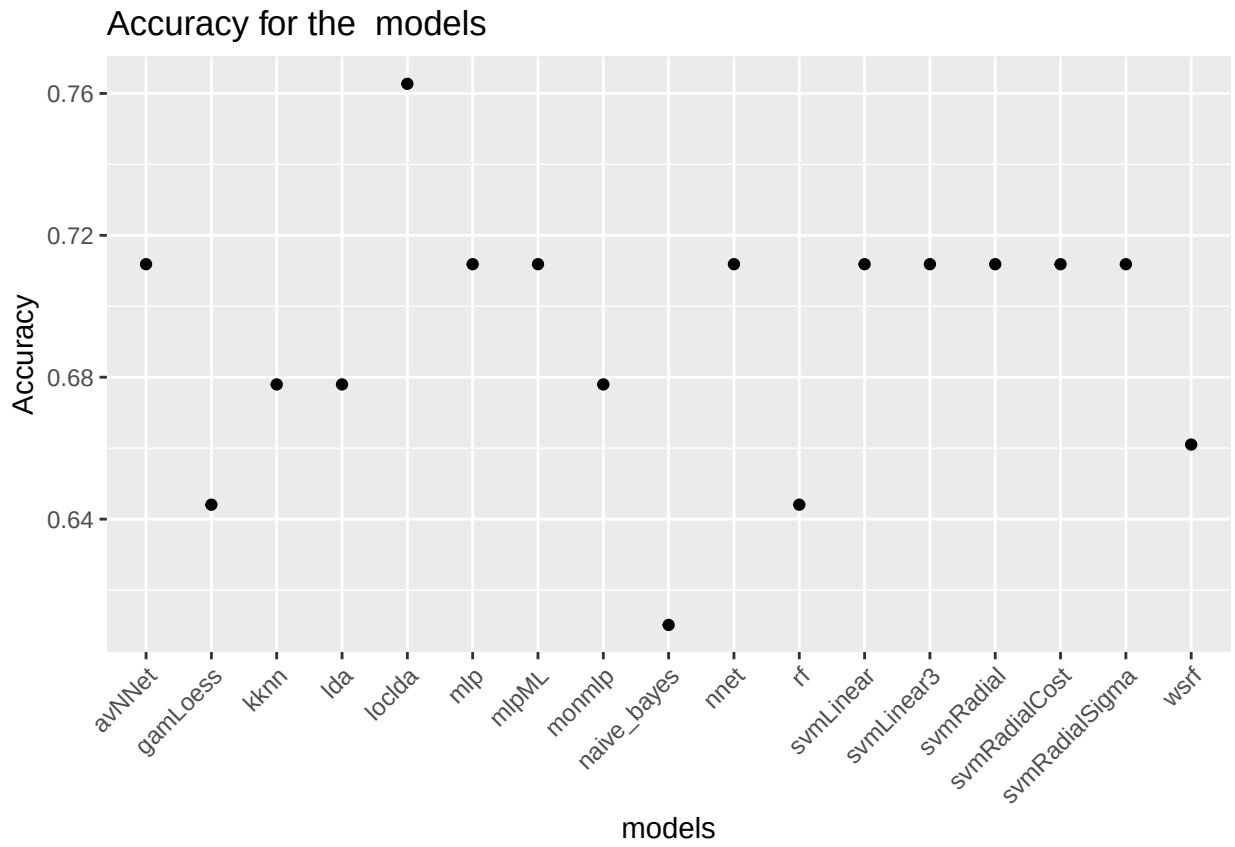
Matrix of predictions

Prediction for Various models & overall accuracy

avNNet	lda	loclda	naive_bayes	wsrf	gamLoess	kknn	monmlp	mlp	mlpML	nnet	svmLinear3	svmLinear	svmRadial	svmRadialCost	svmRadialSigma	rf
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	NO	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	NO	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	YES	NO	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	NO	NO	NO	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	NO	YES	NO	YES	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	NO	NO	NO	NO	NO	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	NO	NO	NO	NO	NO	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	NO	NO	NO	NO	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
YES	YES	NO	NO	YES	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	NO	YES	NO	NO	NO	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	NO	YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	NO	NO	NO	NO	NO	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO
YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	NO	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

models	accuracy
loclda	0.76271186440678
avNNet	0.711864406779661
mlp	0.711864406779661
mlpML	0.711864406779661
nnet	0.711864406779661
svmLinear3	0.711864406779661
svmLinear	0.711864406779661
svmRadial	0.711864406779661
svmRadialCost	0.711864406779661
svmRadialSigma	0.711864406779661
lda	0.677966101694915
kknn	0.677966101694915
monmlp	0.677966101694915
wsrf	0.661016949152542
gamLoess	0.644067796610169
rf	0.644067796610169
naive_bayes	0.610169491525424

models	accuracy
avNNet	0.711864406779661
lda	0.677966101694915
loclda	0.76271186440678
naive_bayes	0.610169491525424
wsrf	0.661016949152542
gamLoess	0.644067796610169
kknn	0.677966101694915
monmlp	0.677966101694915
mlp	0.711864406779661
mlpML	0.711864406779661
nnet	0.711864406779661
svmLinear3	0.711864406779661
svmLinear	0.711864406779661
svmRadial	0.711864406779661
svmRadialCost	0.711864406779661
svmRadialSigma	0.711864406779661
rf	0.644067796610169



CONCLUSION

The top model with its corresponding accuracy is:

```
[1] "Model   :  loclda"

[1] "Accuracy:  0.76271186440678"
```

- Please note that just based on the models created above that this is not the end all, as we can use weights to see if this can be improved further in the future.

Session Information for the execution

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 14393)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
 [1] splines   parallel  grid      stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] gam_1.16            Rborist_0.1-17      doParallel_1.0.14
 [4] iterators_1.0.10    foreach_1.4.4       xtable_1.8-4
 [7] skimr_1.0.6         randomForest_4.6-14 factoextra_1.0.5
[10] reshape2_1.4.3      ROSE_0.0-3          rpart_4.1-15
[13] VIM_4.8.0           data.table_1.12.2   colorspace_1.4-1
[16] mice_3.5.0          pROC_1.15.0         DMwR_0.4.1
[19] caret_6.0-84        lattice_0.20-38     kableExtra_1.1.0
[22] knitr_1.23          forcats_0.4.0       stringr_1.4.0
[25] dplyr_0.8.1         purrr_0.3.2         readr_1.3.1
[28] tidyr_0.8.3         tibble_2.1.1        ggplot2_3.1.1
[31] tidyverse_1.2.1

loaded via a namespace (and not attached):
  [1] readxl_1.3.1       backports_1.1.4    igraph_1.2.4.1
  [4] plyr_1.8.4         lazyeval_0.2.2     sp_1.3-1
  [7] optimx_2018-7.10   digest_0.6.19      htmltools_0.3.6
 [10] gdata_2.18.0       magrittr_1.5       ROCR_1.0-7
 [13] openxlsx_4.1.0.1   recipes_0.1.5      modelr_0.1.4
 [16] gower_0.2.1        xts_0.11-2         rvest_0.3.4
 [19] ggrepel_0.8.1      haven_2.1.0        pan_1.6
 [22] xfun_0.7           crayon_1.3.4       jsonlite_1.6
 [25] lme4_1.1-21        survival_2.44-1.1  zoo_1.8-6
 [28] glue_1.3.1         gtable_0.3.0       ipred_0.9-9
 [31] webshot_0.5.1      wsrf_1.7.17        questionr_0.7.0
 [34] kernlab_0.9-27     car_3.0-3          quantmod_0.4-14
 [37] DEoptimR_1.0-8     jomo_2.6-8         abind_1.4-5
 [40] scales_1.0.0       miniUI_0.1.1.1     Rcpp_1.0.1
 [43] viridisLite_0.3.0  laeken_0.5.0       foreign_0.8-71
 [46] stats4_3.6.0       lava_1.6.5         prodlim_2018.04.18
 [49] vcd_1.4-4          httr_1.4.0         gplots_3.0.1.1
 [52] pkgconfig_2.0.2    nnet_7.3-12        labeling_0.3
 [55] tidyselect_0.2.5   rlang_0.3.4        later_0.8.0
 [58] munsell_0.5.0      cellranger_1.1.0   tools_3.6.0
 [61] cli_1.1.0          generics_0.0.2     ranger_0.11.2
```

```
 [64] broom_0.5.2        evaluate_0.13      yaml_2.2.0
 [67] kknn_1.3.1         monmlp_1.1.5       RSNNS_0.4-11
 [70] ModelMetrics_1.2.2 zip_2.0.2          robustbase_0.93-5
 [73] caTools_1.17.1.2   mitml_0.3-7        nlme_3.1-139
 [76] mime_0.6           xml2_1.2.0         compiler_3.6.0
 [79] rstudioapi_0.10    curl_3.3           e1071_1.7-1
 [82] klaR_0.6-14        stringi_1.4.3      highr_0.8
 [85] naivebayes_0.9.6   Matrix_1.2-17      nloptr_1.2.1
 [88] pillar_1.4.0       LiblineaR_2.10-8   combinat_0.0-8
 [91] lmtest_0.9-37      bitops_1.0-6       httpuv_1.5.1
 [94] R6_2.4.0           promises_1.0.1     KernSmooth_2.23-15
 [97] rio_0.5.16         codetools_0.2-16   boot_1.3-22
[100] MASS_7.3-51.4      gtools_3.8.1       assertthat_0.2.1
[103] withr_2.1.2        hms_0.4.2          timeDate_3043.102
[106] class_7.3-15       minqa_1.2.4        rmarkdown_1.12
[109] carData_3.0-2      TTR_0.23-4         numDeriv_2016.8-1
[112] shiny_1.3.2        lubridate_1.7.4
```