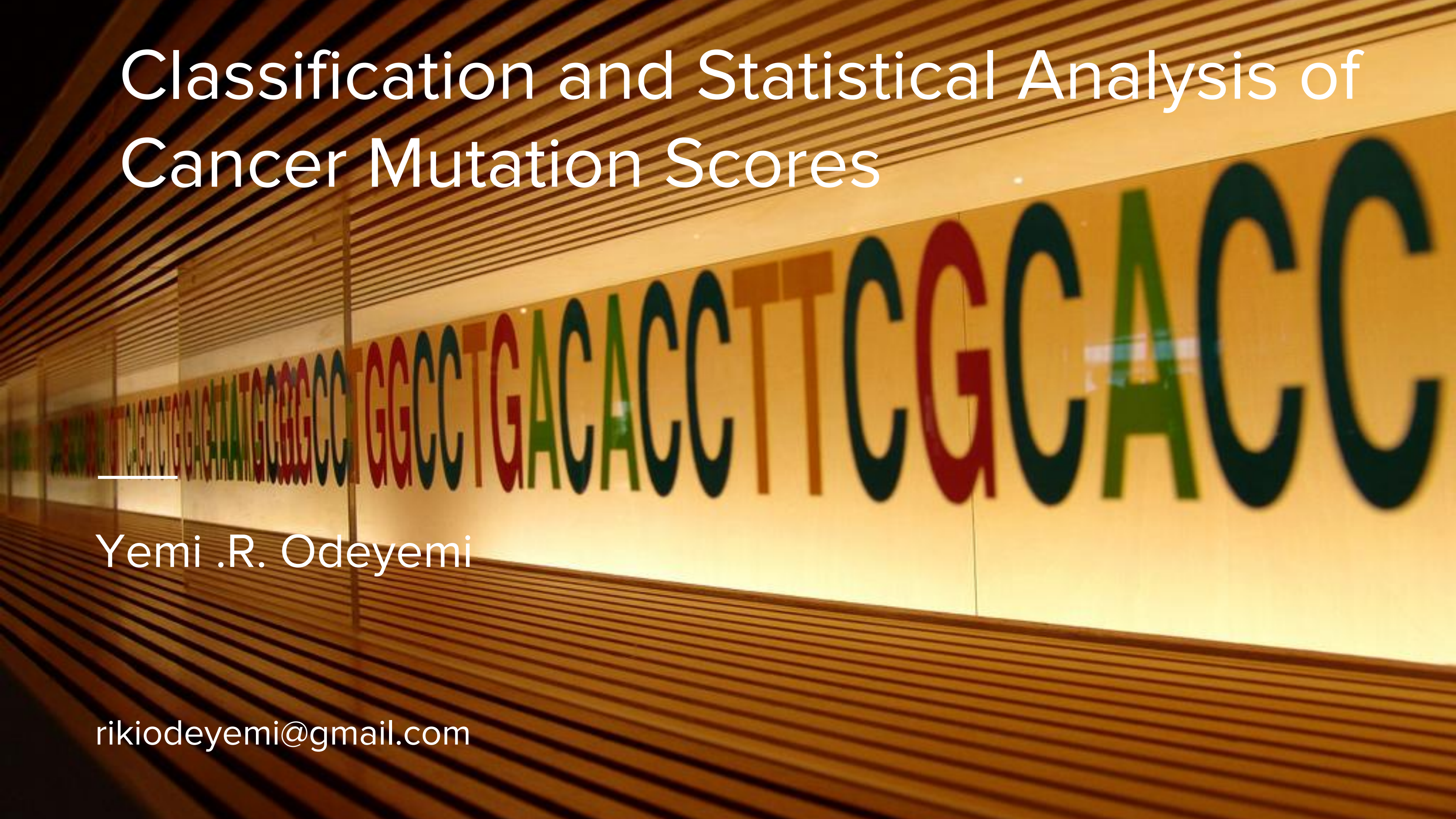


Classification and Statistical Analysis of Cancer Mutation Scores

Yemi .R. Odeyemi

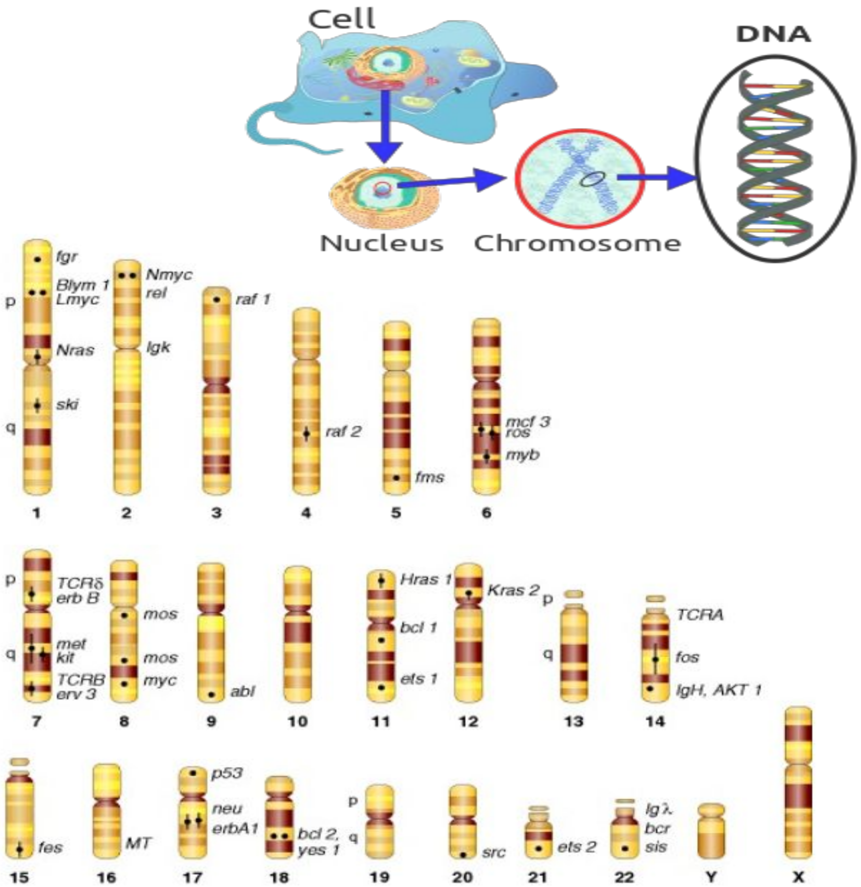
rikiodeyemi@gmail.com

Objective

- ➔ To build a predictive model to classify driver-passenger mutation
- ➔ To determine the optimal class boundary for the mutation label

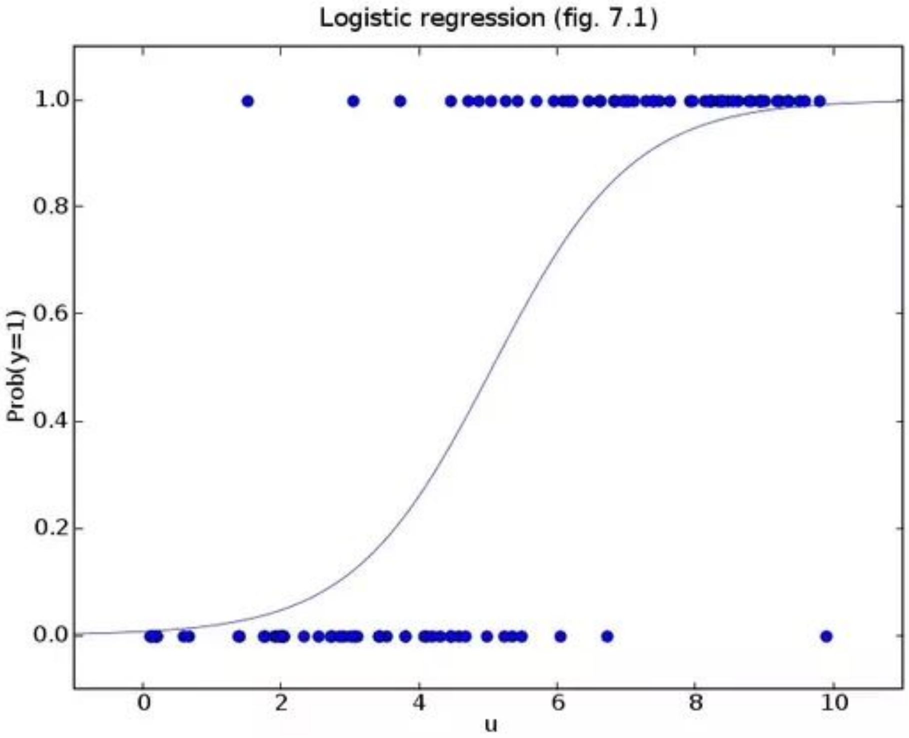
Background

- ➔ Cancer is a genetic disease
- ➔ Characteristics of cancer
  - ◆ uncontrolled cell division leading to an overgrown group of cells called a ***tumor***
  - ◆ the spread of tumor cells throughout the body to form new tumors, a process called **metastasis**
- ➔ Mutation of proto-oncogenes to oncogenes
- ➔ Mutation is a random process
- ➔ **Passenger** mutations are mutations that have no impact on a cell's phenotype (Neutral)
- ➔ Mutations that drive cancer progression are known as **Driver** mutations



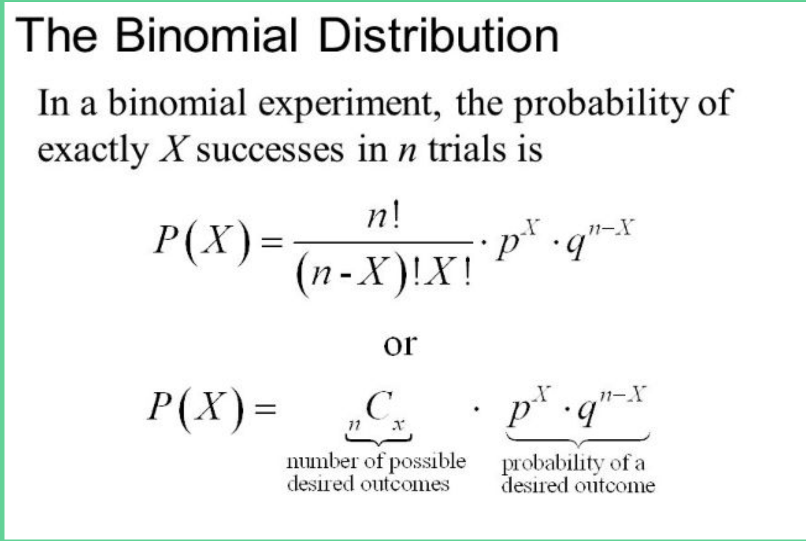
****Human chromosomes showing bands from Giemsa staining and the positions (shown by black dots) of known proto-oncogenes; mutations in proto-oncogenes lead to cancer.

Algorithm: Logit Model



Concept:

- Probabilistic in nature



The Binomial Distribution

In a binomial experiment, the probability of exactly X successes in n trials is

$$P(X) = \frac{n!}{(n-X)!X!} \cdot p^{X} \cdot q^{n-X}$$

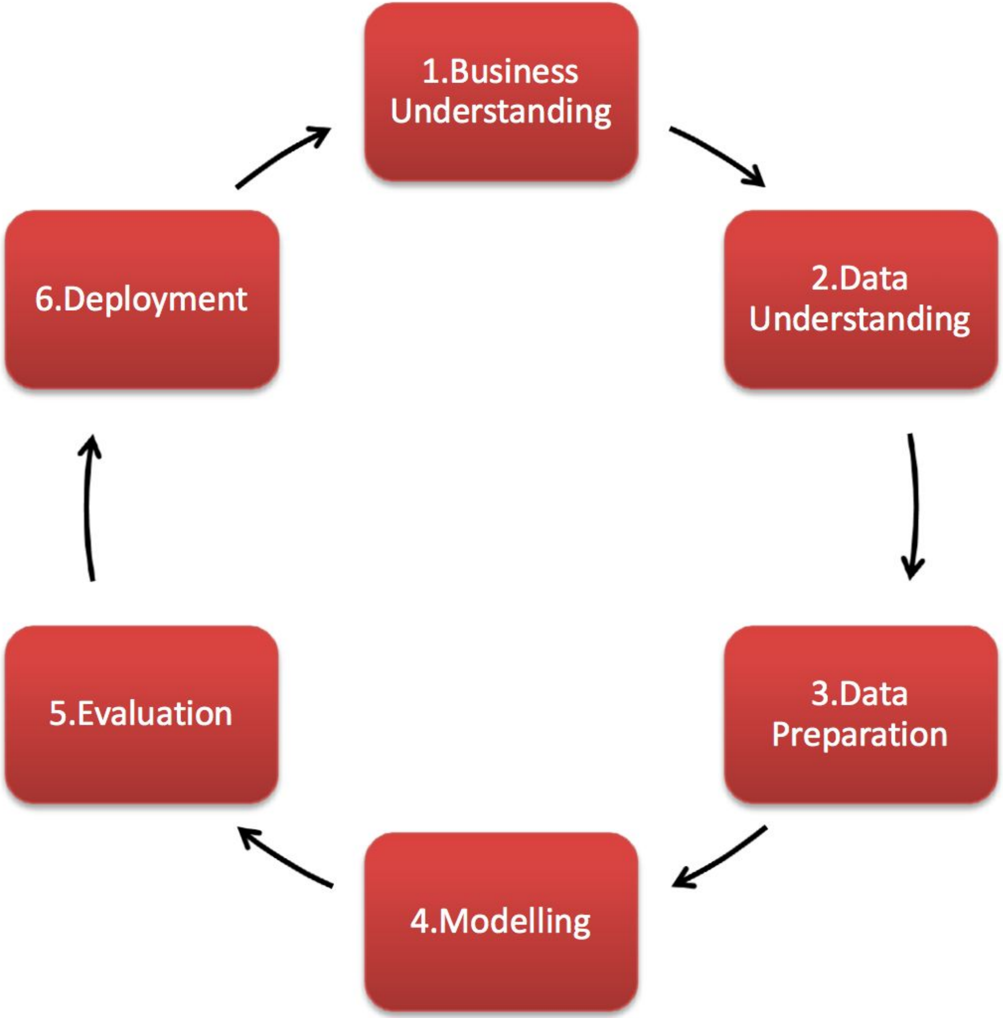
or

$$P(X) = \underbrace{{}_{n}C_{x}}_{\text{number of possible desired outcomes}} \cdot \underbrace{p^{X} \cdot q^{n-X}}_{\text{probability of a desired outcome}}$$

The Statistical Assumption:

- Response variable has to be binary in nature
  - 0, 1 : Passenger vs Driver mutation
- No high Intercorrelations among the predictors
- Linear relationship between the logit of the outcome and each predictor
- $$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \epsilon$$

CRISP-DM



1.Business Understanding

2.Data Understanding

3.Data Preparation

4.Modelling

5.Evaluation

6.Deployment

Data Source: CbioPortal

- ➔ **Cancer genomic database for interactive exploration of multidimensional cancer genomics data sets.**
- ➔ **Primary source of mutation**
  - ◆ **Mutation organized by Cancer type and genes**
- ➔ **Data types:**
  - ◆ **DNA Copy number data**
  - ◆ **mRNA**
  - ◆ **MicroRNA**

cBioPortal
FOR CANCER GENOMICS

About the data by the **numbers**

➔ Extracted mutation from **17** subtypes where cancer is known to be present

- Glioblastoma
- Ovarian & Peritoneal carcinoma
- Prostate adenocarcinoma and sarcoma
- Apoptosis regulation signaling pathway

➔ Mutation **50000+**

➔ For each mutation the following was extracted

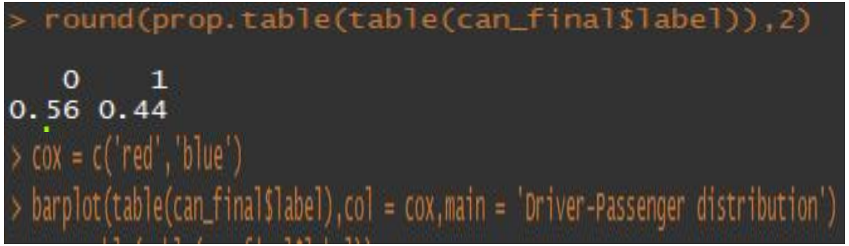
- Chromosome
- Start and ending position
- Reference and alternate nucleotide

➔ Mutation data inputted into the dbWGFP and **48** scores were generated

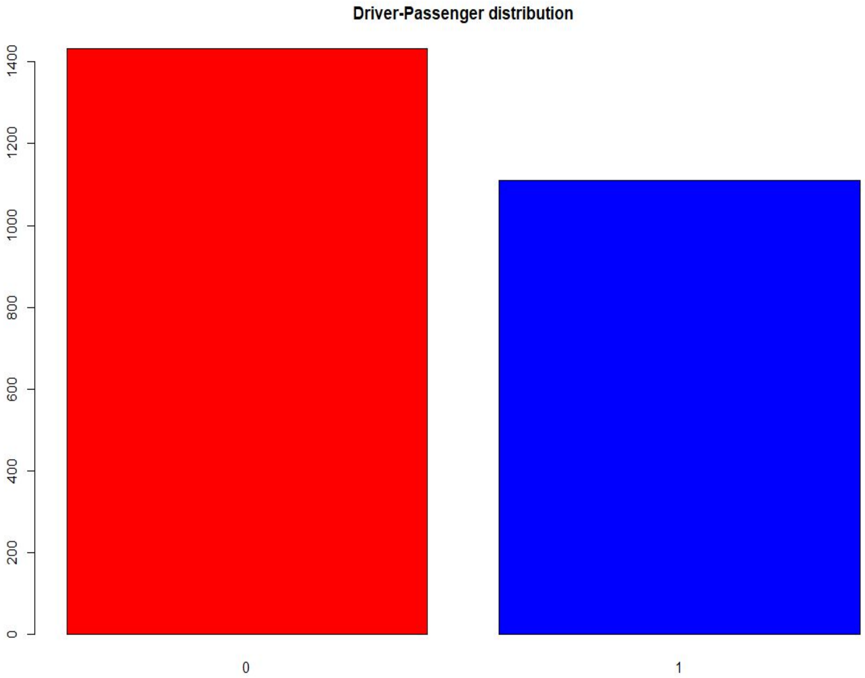
- MutationTaster_score
- Grantham
- FATHMM_score

Exploratory Data Analysis

- Summary statistics
  - Descriptives/Measure of central tendency
  - Gaussian distribution
- Data structure
- Predictors-response features identification
- Response feature class identification and class distribution



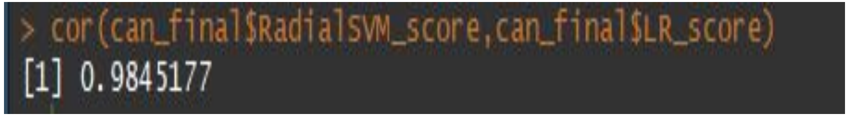
```
> round(prop.table(table(can_final$label)),2)

   0    1
0.56 0.44
> cox = c('red','blue')
> barplot(table(can_final$label),col = cox,main = 'Driver-Passenger distribution')
```



Data Preparation/Preprocessing/Munging/Cleaning 1

- ➔ Removal of redundant predictors
- ➔ Handle missing values
- ➔ Regular expression

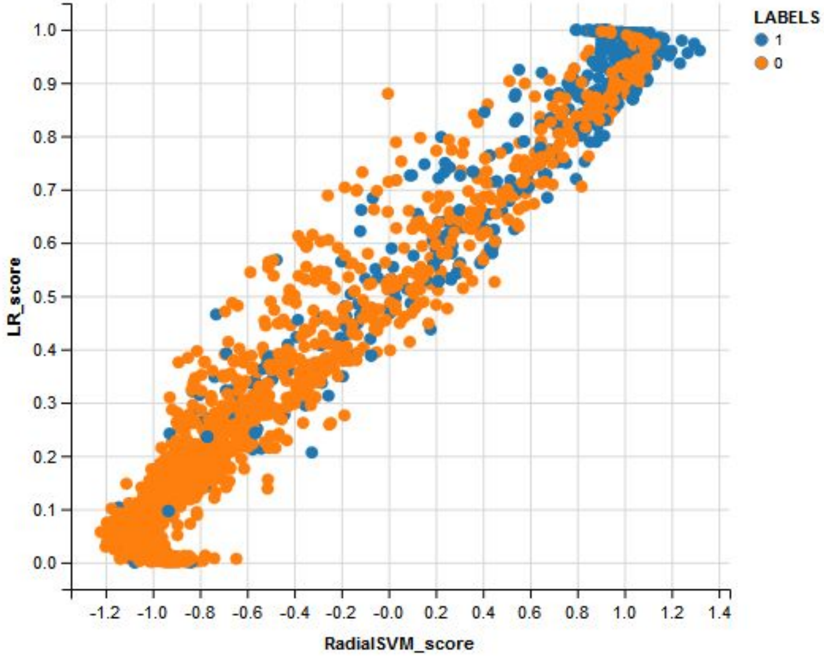
Data Preparation/Preprocessing/Munging/Cleaning 2



```
> cor(can_final$RadialSVM_score,can_final$LR_score)
[1] 0.9845177
```

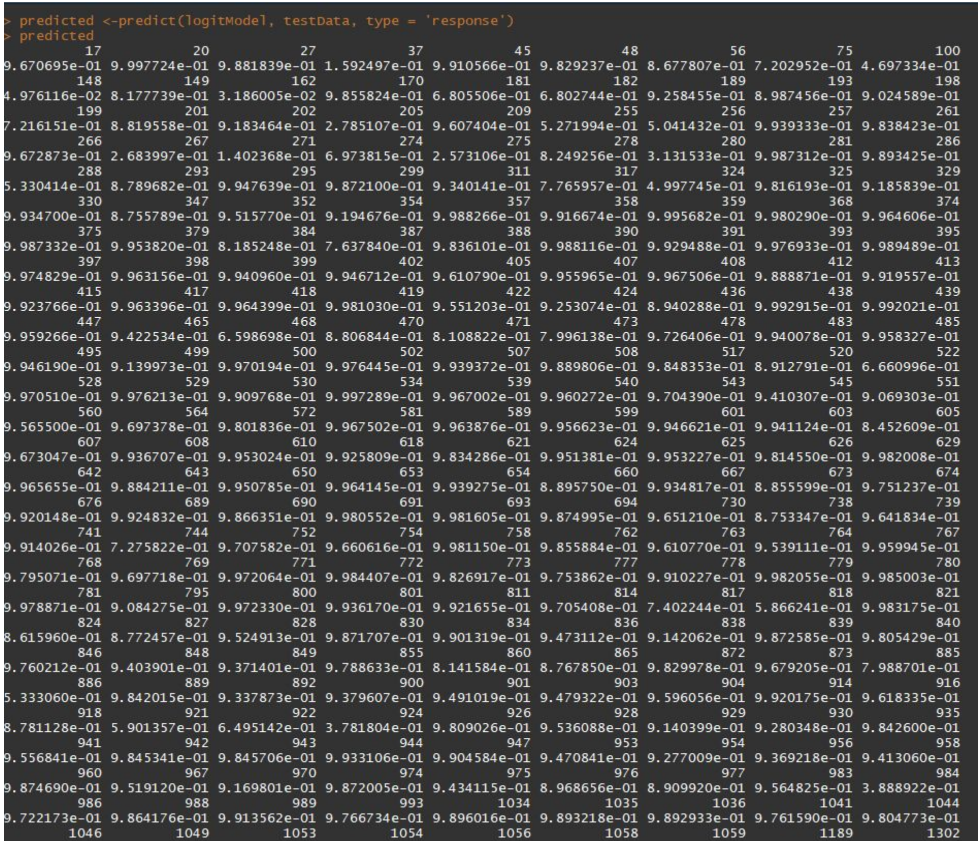
- ➔ Conversion of dichotomous response feature to binary format
- ➔ Multicollinearity check using a matrix
  - ◆ Correlation Coefficient Matrix
  - ◆ Pearson correlation coefficient
    - ● 0.98

- ➔ Test for gaussian distribution
  - ◆ Wilk -shapiro test



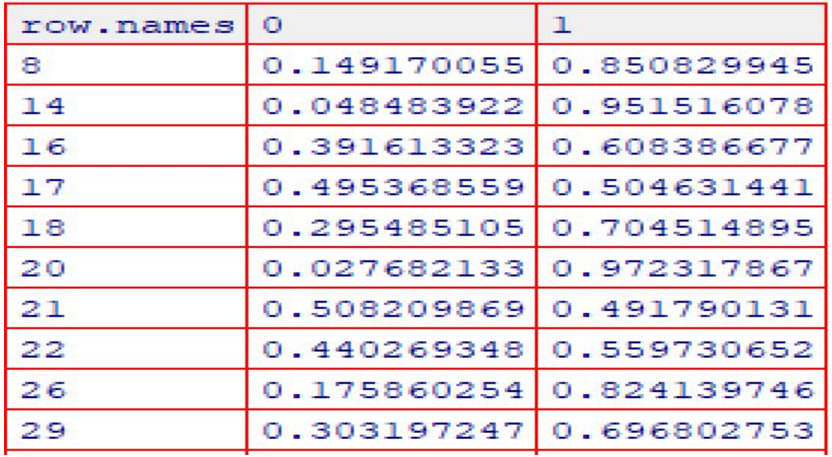
Modeling

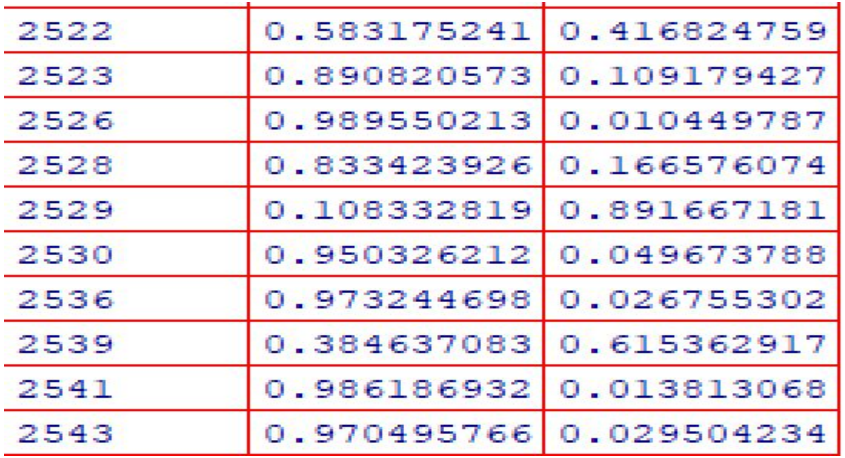
- ➔ Synthetic Minority Over-Sampling Technique (SMOTE)***
- ➔ Data Partitioning
  - ◆ Training set 80% : Test set 20%
  - ◆ Cross validation
    - ● 10-fold
- ➔ Train the model
  - ◆ Logit regression
- ➔ Test the model



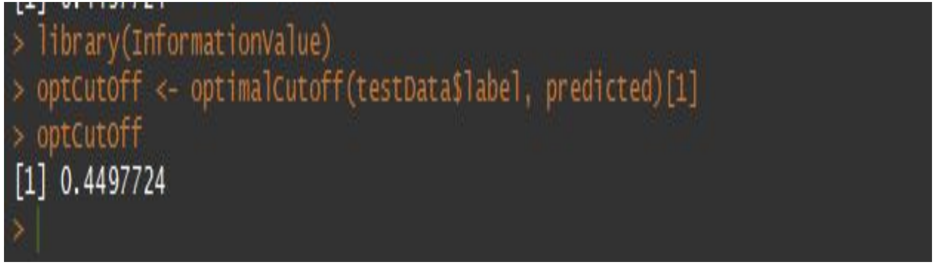
```
> predicted <-predict(logitModel, testData, type = 'response')
> predicted
          17           20           27           37           45           48           56           75          100 
9.670695e-01 9.997724e-01 9.881839e-01 1.592497e-01 9.910566e-01 9.829237e-01 8.677807e-01 7.202952e-01 4.697334e-01 
         148          149          162          170          181          182          189          193          198 
4.976116e-02 8.177739e-01 3.186005e-02 9.855824e-01 6.805506e-01 6.802744e-01 9.258455e-01 8.987456e-01 9.024589e-01 
         199          201          202          205          209          255          256          257          261 
7.216151e-01 8.819558e-01 9.183464e-01 2.785107e-01 9.607404e-01 5.271994e-01 5.041432e-01 9.939333e-01 9.838423e-01 
         266          267          271          274          275          278          280          281          286 
9.672873e-01 2.683997e-01 1.402368e-01 6.973815e-01 2.573106e-01 8.249256e-01 3.131533e-01 9.987312e-01 9.893425e-01 
         288          293          295          299          311          317          324          325          329 
5.330414e-01 8.789682e-01 9.947639e-01 9.872100e-01 9.340141e-01 7.765957e-01 4.997745e-01 9.816193e-01 9.185839e-01 
         330          347          352          354          357          358          359          368          374 
9.934700e-01 8.755789e-01 9.515770e-01 9.194676e-01 9.988266e-01 9.916674e-01 9.995682e-01 9.980290e-01 9.964606e-01 
         375          379          384          387          388          390          391          393          395 
9.987332e-01 9.953820e-01 8.185248e-01 7.637840e-01 9.836101e-01 9.988116e-01 9.929488e-01 9.976933e-01 9.989489e-01 
         397          398          399          402          405          407          408          412          413 
9.974829e-01 9.963156e-01 9.940960e-01 9.946712e-01 9.610790e-01 9.955965e-01 9.967506e-01 9.888871e-01 9.919557e-01 
         415          417          418          419          422          424          436          438          439 
9.923766e-01 9.963396e-01 9.964399e-01 9.981030e-01 9.551203e-01 9.253074e-01 8.940288e-01 9.992915e-01 9.992021e-01 
         447          465          468          470          471          473          478          483          485 
9.959266e-01 9.422534e-01 6.598698e-01 8.806844e-01 8.108822e-01 7.996138e-01 9.726406e-01 9.940078e-01 9.958327e-01 
         495          499          500          502          507          508          517          520          522 
9.946190e-01 9.139973e-01 9.970194e-01 9.976445e-01 9.939372e-01 9.889806e-01 9.848353e-01 8.912791e-01 6.660996e-01 
         528          529          530          534          539          540          543          545          551 
9.970510e-01 9.976213e-01 9.909768e-01 9.997289e-01 9.967002e-01 9.960272e-01 9.704390e-01 9.410307e-01 9.069303e-01 
         560          564          572          581          589          599          601          603          605 
9.565500e-01 9.697378e-01 9.801836e-01 9.967502e-01 9.963876e-01 9.956623e-01 9.946621e-01 9.941124e-01 8.452609e-01 
         607          608          610          618          621          624          625          626          629 
9.673047e-01 9.936707e-01 9.953024e-01 9.925809e-01 9.834286e-01 9.951381e-01 9.953227e-01 9.814550e-01 9.982008e-01 
         642          643          650          653          654          660          667          673          674 
9.965655e-01 9.884211e-01 9.950785e-01 9.964145e-01 9.939275e-01 8.895750e-01 9.934817e-01 8.855599e-01 9.751237e-01 
         676          689          690          691          693          694          730          738          739 
9.920148e-01 9.924832e-01 9.866351e-01 9.980552e-01 9.981605e-01 9.874995e-01 9.651210e-01 8.753347e-01 9.641834e-01 
         741          744          752          754          758          762          763          764          767 
9.914026e-01 7.275822e-01 9.707582e-01 9.660616e-01 9.981150e-01 9.855884e-01 9.610770e-01 9.539111e-01 9.959945e-01 
         768          769          771          772          773          777          778          779          780 
9.795071e-01 9.697718e-01 9.972064e-01 9.984407e-01 9.826917e-01 9.753862e-01 9.910227e-01 9.982055e-01 9.985003e-01 
         781          795          800          801          811          814          817          818          821 
9.978871e-01 9.084275e-01 9.972330e-01 9.936170e-01 9.921655e-01 9.705408e-01 7.402244e-01 5.866241e-01 9.983175e-01 
         824          827          828          830          834          836          838          839          840 
8.615960e-01 8.772457e-01 9.524913e-01 9.871707e-01 9.901319e-01 9.473112e-01 9.142062e-01 9.872585e-01 9.805429e-01 
         846          848          849          855          860          865          872          873          885 
9.760212e-01 9.403901e-01 9.371401e-01 9.788633e-01 8.141584e-01 8.767850e-01 9.829978e-01 9.679205e-01 7.988701e-01 
         886          889          892          900          901          903          904          914          916 
5.333060e-01 9.842015e-01 9.337873e-01 9.379607e-01 9.491019e-01 9.479322e-01 9.596056e-01 9.920175e-01 9.618335e-01 
         918          921          922          924          926          928          929          930          935 
8.781128e-01 5.901357e-01 6.495142e-01 3.781804e-01 9.809026e-01 9.536088e-01 9.140399e-01 9.280348e-01 9.842600e-01 
         941          942          943          944          947          953          954          956          958 
9.556841e-01 9.845341e-01 9.845706e-01 9.933106e-01 9.904584e-01 9.470841e-01 9.277009e-01 9.369218e-01 9.413060e-01 
         960          967          970          974          975          976          977          983          984 
9.874690e-01 9.519120e-01 9.169801e-01 9.872005e-01 9.434115e-01 8.968656e-01 8.909920e-01 9.564825e-01 3.888922e-01 
         986          988          989          993         1034         1035         1036         1041         1044 
9.722173e-01 9.864176e-01 9.913562e-01 9.766734e-01 9.896016e-01 9.893218e-01 9.892933e-01 9.761590e-01 9.804773e-01 
        1046         1049         1053         1054         1056         1058         1059         1189         1302 
```

Class Probability Boundary & Optimal Cutoff





row.names	0	1
8	0.149170055	0.850829945
14	0.048483922	0.951516078
16	0.391613323	0.608386677
17	0.495368559	0.504631441
18	0.295485105	0.704514895
20	0.027682133	0.972317867
21	0.508209869	0.491790131
22	0.440269348	0.559730652
26	0.175860254	0.824139746
29	0.303197247	0.696802753
2522	0.583175241	0.416824759
2523	0.890820573	0.109179427
2526	0.989550213	0.010449787
2528	0.833423926	0.166576074
2529	0.108332819	0.891667181
2530	0.950326212	0.049673788
2536	0.973244698	0.026755302
2539	0.384637083	0.615362917
2541	0.986186932	0.013813068
2543	0.970495766	0.029504234



```
> library(InformationValue)
> optCutoff <- optimalCutoff(testData$label, predicted)[1]
> optCutoff
[1] 0.4497724
>
```

Model Selection : Akaike Information Criteria

Statistical tool that compares the quality of a set of models to each other

Ranks each model from best to worst

$$AIC = 2K - 2\ \log(\mathscr{L}(\hat{\theta}|y)),$$

Where

- K is the number of model parameters (the number of variables in the model plus the intercept)
- Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output.

Feature Importance of the Mutation Scores

- ➔ Random Forest Method
- ➔ Stepwise Regression
- ➔ Information value and Weight of evidence

Feature Importance_RF

Diagnostics: Area Under the Curve

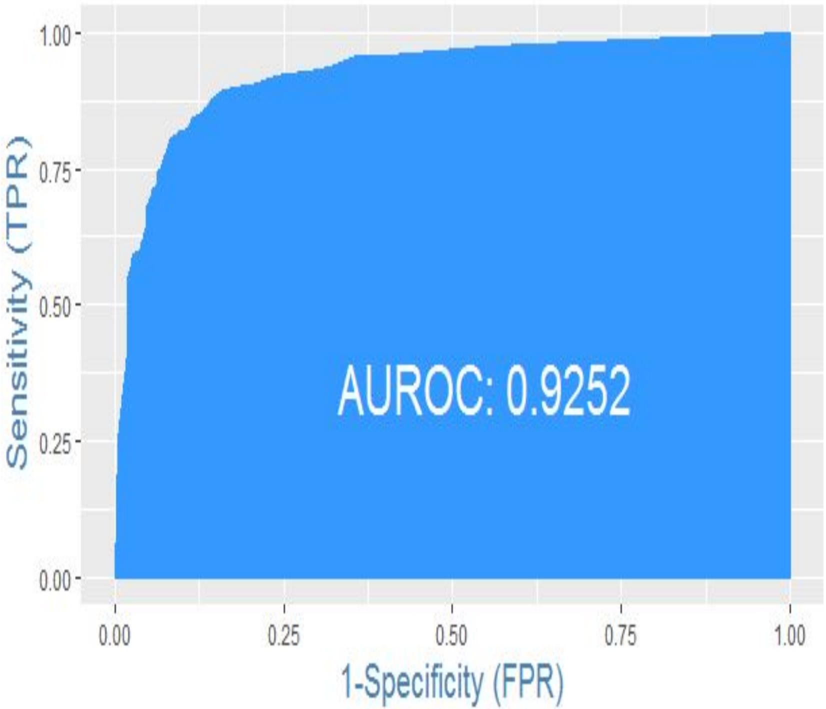
- ◆

➔ Sensitivity/Recall/True Positive Rate

- ◆ Proportion of positive data points that are correctly considered as positive, with respect to all positive data points
- ◆ Higher TPR ➔ less misclassification of positive data points

➔ Fall Out/False Positive Rate

- ◆ Proportion of negative data points that were mistakenly classified as positive with respect to all negative data points
- ◆ Higher FPR ➔ more misclassification of negative data points



Conclusion

- ➔ Theoretical class boundary is not always the same as the optimal class boundary.
- ➔ Comparative studies of the logit model with random forest model yielded almost similar AUC
- ➔ Removal of the top 3 mutation scores had a significant impact on the models’ AUC
- ➔ Hypothesis testing yielded no significant difference between the top 3 mutation scores in the variable importance
- ➔ Exploration of SVD,PCA and t-SNE for dimension reduction would enable holistic view of the features

Miscellany: R packages

{DMwR} - Functions and data for the book "Data Mining with R" and SMOTE algorithm

{caret} - modeling wrapper, functions, commands

{pROC} - Area Under the Curve (AUC) functions

{Dplyr} - Data manipulation