APPLIED STATISTICAL MODELLING – CS7DS3

Main Assignment

Ketan Patil – 22303876

INTRODUCTION:

The Chowdary dataset, which was originally examined by Chowdary et al. (2006) and then by de Souto et al. (2008), is made up of gene expression data from tissue samples taken from these two forms of cancer. The dataset is high-dimensional, with 104 observations and 182 numeric columns, posing unique obstacles in data processing. Our major goal is to identify the genes that have the greatest effect on cancer type identification and, to the highest level feasible, to assess the characteristics of the relation (increasing/decreasing) among these genes and cancer type classification.

Identifying important genes that drive cancer growth can give vital insights into the underlying processes and aid in the development of targeted therapeutics. In this assignment, we will examine the Chowdary dataset to determine the most significant genes in cancer type identification, with a focus on lymph node-negative breast tumours (type B) and Dukes' B colon tumours (type C).

This was achieved using various statistical tests, feature selection techniques such as lasso and elastic net. After using these techniques, we use logistic regression to identify the significant genes and their influence on the type of Cancer B or C. This is accompanied by visualisations for better analysis of the dataset. All of this was done in using both python and R.

METHODOLOGY

Below is the figure illustrating the flow of the task.

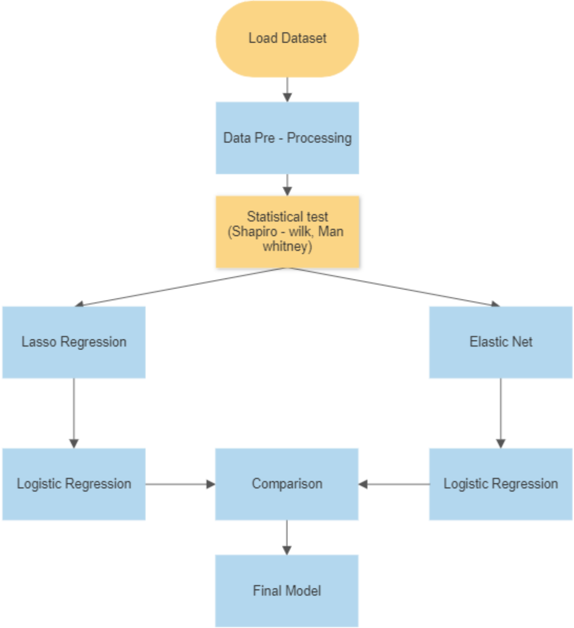


Fig 1. Flowchart of the task

Data Pre-Processing:

As the data is quite high dimensional with 104 rows and more than 180 columns. So, we check for missing values. But there were no missing values so there was no need to impute the values for any of the columns.

However, there was a column with categorical values which had values like 'GSM85961' which seemed like the serial number as it the values followed had an increase on 1 like 'GSM85962'. So, we dropped the column.

Statistical Tests:

To perform Statistics test such as t-test we need to check whether the data is normally distributed which was done as follows:

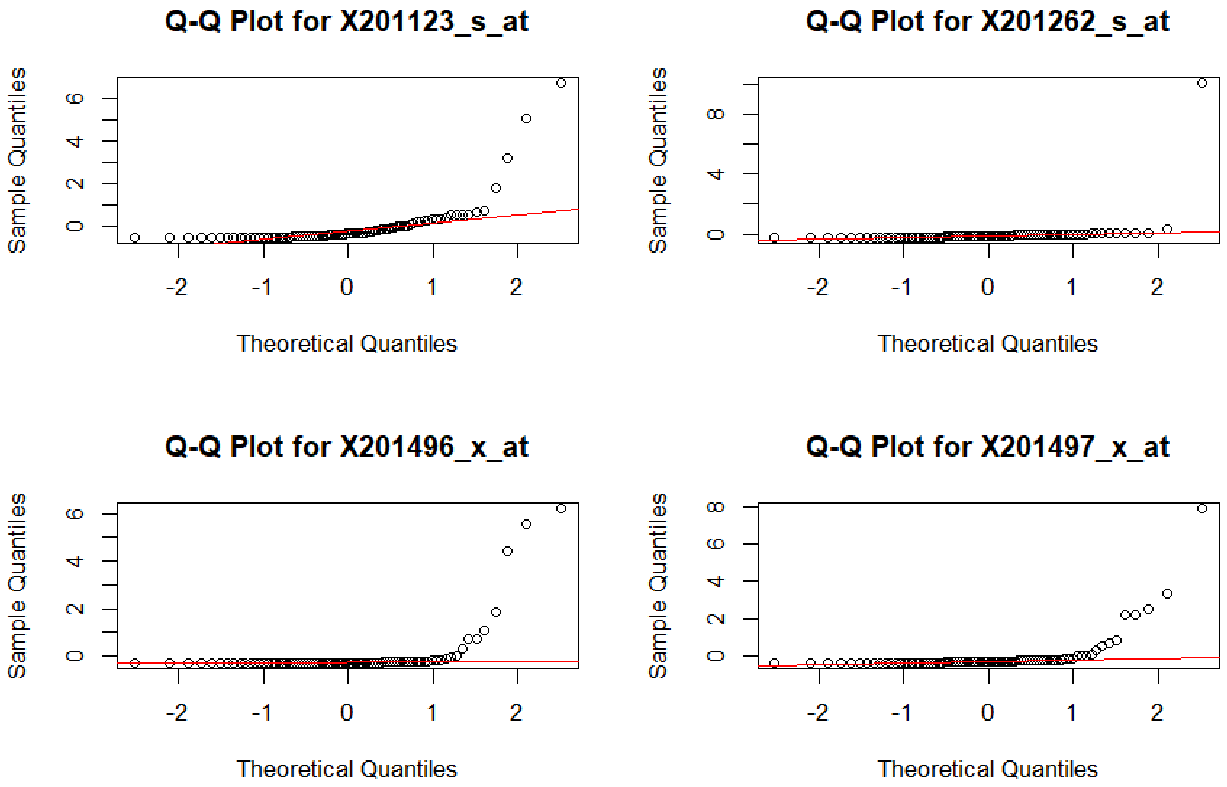
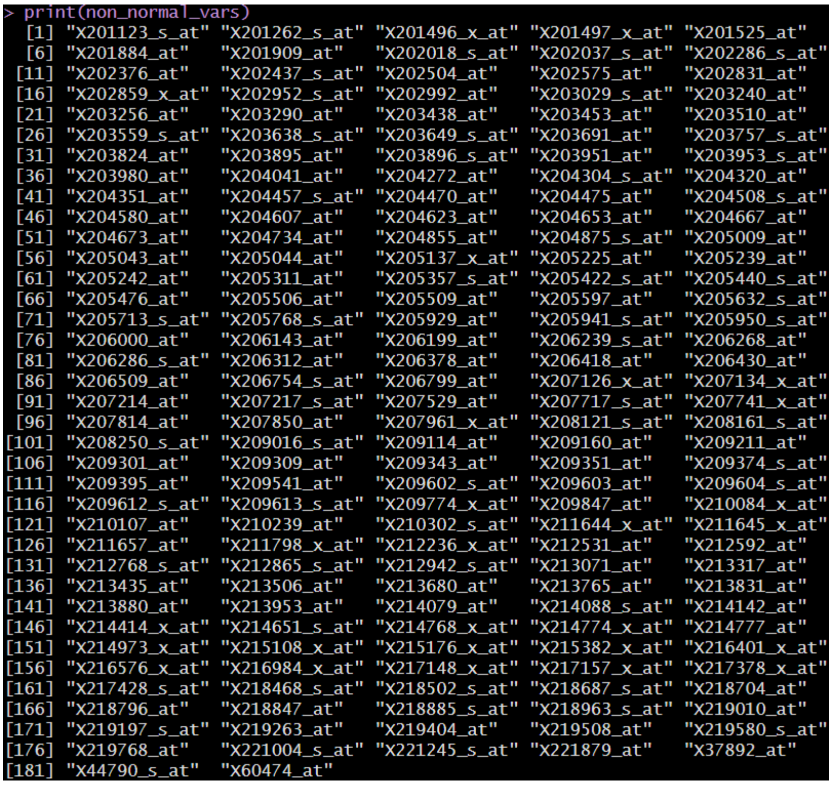


Fig 2. Q-Q plots for checking normal distribution.

Firstly Q-Q plots were used to check the whether the data is normally distributed or not as seen none of the four features are in normal distribution. And it is not feasible to plot the Q-Q plot for all the 182 features we do the Shapiro wilk test.

Shapiro-wilk test: It is statistical test used to check the normality of the data. It calculates the correlation between the data and the corresponding values estimate from a normal distribution with equivalent mean and variance as the sample data. It produces p-value which helps to determine the normality of the data. If the p-value is less than 0.05 then it is not a normal distribution and if it is greater then the data is in normal distribution.

Below are the list non normal features:

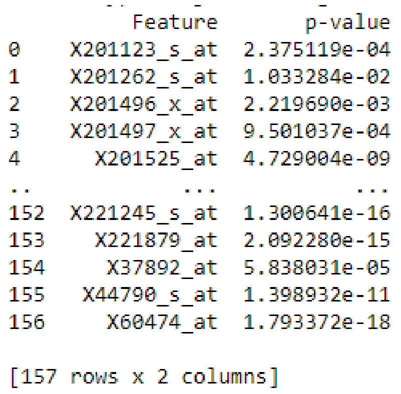


```
> print(non_normal_vars)
  [1] "X201123_s_at" "X201262_s_at" "X201496_x_at" "X201497_x_at" "X201525_at"
  [6] "X201884_at"   "X201909_at"   "X202018_s_at" "X202037_s_at" "X202286_s_at"
 [11] "X202376_at"   "X202437_s_at" "X202504_at"   "X202575_at"   "X202831_at"
 [16] "X202859_x_at" "X202952_s_at" "X202992_at"   "X203029_s_at" "X203240_at"
 [21] "X203256_at"   "X203290_at"   "X203438_at"   "X203453_at"   "X203510_at"
 [26] "X203559_s_at" "X203638_s_at" "X203649_s_at" "X203691_at"   "X203757_s_at"
 [31] "X203824_at"   "X203895_at"   "X203896_s_at" "X203951_at"   "X203953_s_at"
 [36] "X203980_at"   "X204041_at"   "X204272_at"   "X204304_s_at" "X204320_at"
 [41] "X204351_at"   "X204457_s_at" "X204470_at"   "X204475_at"   "X204508_s_at"
 [46] "X204580_at"   "X204607_at"   "X204623_at"   "X204653_at"   "X204667_at"
 [51] "X204673_at"   "X204734_at"   "X204855_at"   "X204875_s_at" "X205009_at"
 [56] "X205043_at"   "X205044_at"   "X205137_x_at" "X205225_at"   "X205239_at"
 [61] "X205242_at"   "X205311_at"   "X205357_s_at" "X205422_s_at" "X205440_s_at"
 [66] "X205476_at"   "X205506_at"   "X205509_at"   "X205597_at"   "X205632_s_at"
 [71] "X205713_s_at" "X205768_s_at" "X205929_at"   "X205941_s_at" "X205950_s_at"
 [76] "X206000_at"   "X206143_at"   "X206199_at"   "X206239_s_at" "X206268_at"
 [81] "X206286_s_at" "X206312_at"   "X206378_at"   "X206418_at"   "X206430_at"
 [86] "X206509_at"   "X206754_s_at" "X206799_at"   "X207126_x_at" "X207134_x_at"
 [91] "X207214_at"   "X207217_s_at" "X207529_at"   "X207717_s_at" "X207741_x_at"
 [96] "X207814_at"   "X207850_at"   "X207961_x_at" "X208121_s_at" "X208161_s_at"
[101] "X208250_s_at" "X209016_s_at" "X209114_at"   "X209160_at"   "X209211_at"
[106] "X209301_at"   "X209309_at"   "X209343_at"   "X209351_at"   "X209374_s_at"
[111] "X209395_at"   "X209541_at"   "X209602_s_at" "X209603_at"   "X209604_s_at"
[116] "X209612_s_at" "X209613_s_at" "X209774_x_at" "X209847_at"   "X210084_x_at"
[121] "X210107_at"   "X210239_at"   "X210302_s_at" "X211644_x_at" "X211645_x_at"
[126] "X211657_at"   "X211798_x_at" "X212236_x_at" "X212531_at"   "X212592_at"
[131] "X212768_s_at" "X212865_s_at" "X212942_s_at" "X213071_at"   "X213317_at"
[136] "X213435_at"   "X213506_at"   "X213680_at"   "X213765_at"   "X213831_at"
[141] "X213880_at"   "X213953_at"   "X214079_at"   "X214088_s_at" "X214142_at"
[146] "X214414_x_at" "X214651_s_at" "X214768_x_at" "X214774_x_at" "X214777_at"
[151] "X214973_x_at" "X215108_x_at" "X215176_x_at" "X215382_x_at" "X216401_x_at"
[156] "X216576_x_at" "X216984_x_at" "X217148_x_at" "X217157_x_at" "X217378_x_at"
[161] "X217428_s_at" "X218468_s_at" "X218502_s_at" "X218687_s_at" "X218704_at"
[166] "X218796_at"   "X218847_at"   "X218885_s_at" "X218963_s_at" "X219010_at"
[171] "X219197_s_at" "X219263_at"   "X219404_at"   "X219508_at"   "X219580_s_at"
[176] "X219768_at"   "X221004_s_at" "X221245_s_at" "X221879_at"   "X37892_at"
[181] "X44790_s_at"  "X60474_at"
```

As you can see from the image that all the features are not normally distributed which makes us unable to do t-test as it assumes that the data is in normal distribution, using it will not lead to an appropriate result.

But now that we know that no feature is normally distributed, we can do Man Whitney U test to determine if there is a huge difference in the distribution of the two independent samples. It also gives an p-value with the help of which we can eliminate/drop the features, e.g., if the p-value is greater than 0.05 we drop that feature and if it is less, we keep that feature.

In the image to the right are the features that remain after calculating the p-value using the Man Whitney U test. There were 182 features at the start now there are 157 features left which we will use further in the selection techniques as the number of features is still high.

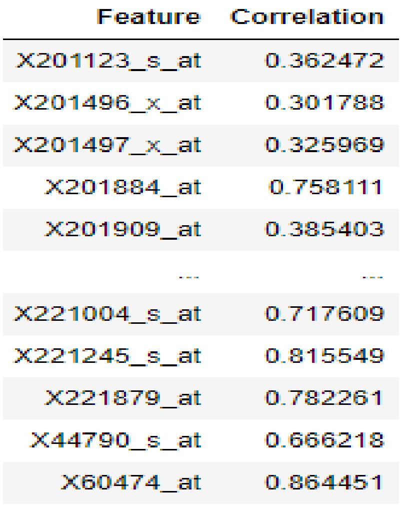


```
          Feature       p-value
0    X201123_s_at  2.375119e-04
1    X201262_s_at  1.033284e-02
2    X201496_x_at  2.219690e-03
3    X201497_x_at  9.501037e-04
4      X201525_at  4.729004e-09
..            ...           ...
152  X221245_s_at  1.300641e-16
153    X221879_at  2.092280e-15
154     X37892_at  5.838031e-05
155   X44790_s_at  1.398932e-11
156     X60474_at  1.793372e-18

[157 rows x 2 columns]
```

To check the correlation in the data we use spearman rank test as we can't do Pearson correlation coefficient as it need the data to be linear and normally distributed. However, we already know that our data is not normally distributed which is why we go ahead with spearman rank correlation test as it a nonparametric test which does need the data to be normally distributed. It measures the strength and relation between the two variables based on their ranks.

The image to right shows the correlation as it not possible to plot all the values due to the number of features being high it will not be readable this is why it was made into a data frame.



Feature	Correlation
X201123_s_at	0.362472
X201496_x_at	0.301788
X201497_x_at	0.325969
X201884_at	0.758111
X201909_at	0.385403
...	...
X221004_s_at	0.717609
X221245_s_at	0.815549
X221879_at	0.782261
X44790_s_at	0.666218
X60474_at	0.864451

Lasso Regression:

Lasso Regression in a regularisation technique that is used for selecting the features and apply regularisation to increase accuracy and interpretability. It adds an L1 penalty that shrinks the unimportant features to zero resulting in dropping of those features from the model.

$$\arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

It is used with cross validation to find optimal value of the lambda that decreases the prediction error by running multiple times till it finds an optimal value.

This is done using 'glmnet' package in R with cv. The alpha is set to 1 as 1 signifies L1 penalty which means it is lasso regression.

After finding the optimal lambda, that value is used to fit into a final model to extract the features from the data.

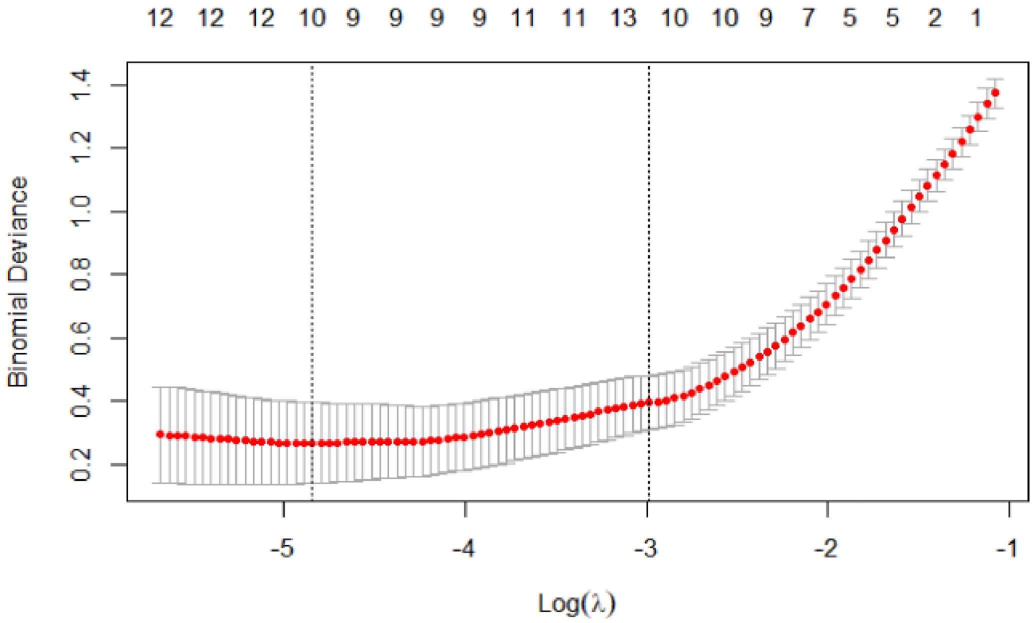


Fig 3. Binomial deviance plot for Lasso Regression

The vertical dotted line on left represents the lambda that has minimum mean deviance and the one on the right represents largest lambda value within one standard error of minimum deviance.

Elastic Net:

Elastic Net is regularisation technique that combines lasso regression and ridge regression i.e., the combination of L1 and L2 penalty.

This technique is useful when the data is high dimensional which is the case here and data has multicollinearity.

This is done using 'glmnet' package in R with cv. The alpha is set to 0.5 which is between 0 (Ridge regression) and 1(lasso regression) by taking a value between 0 and 1 for alpha it means the model is elastic net.

Same as lasso regression we find optimal lambda and then use it fit a final model to get features.

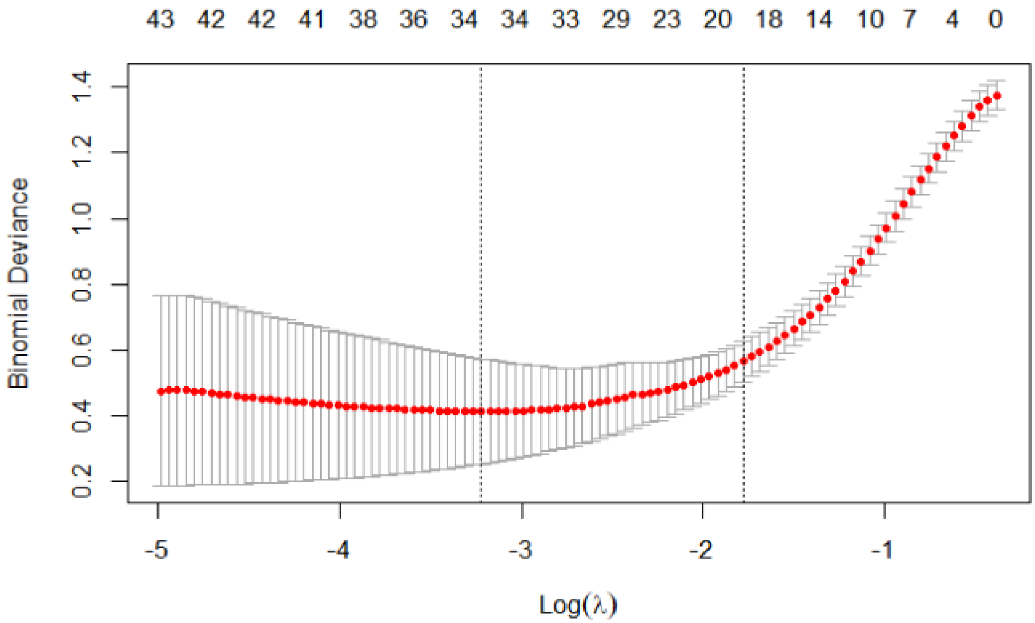
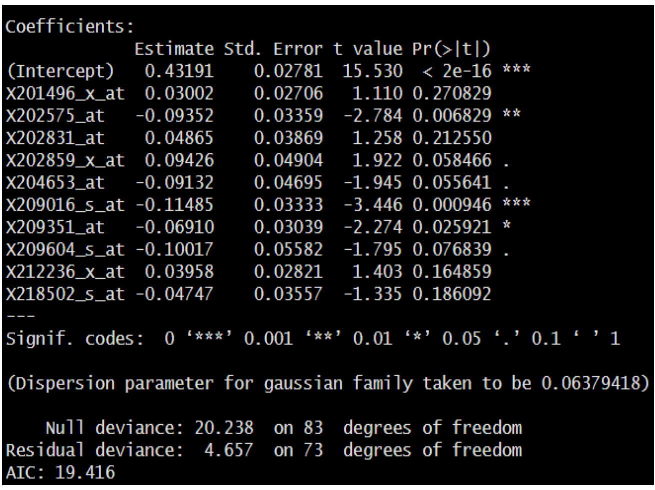
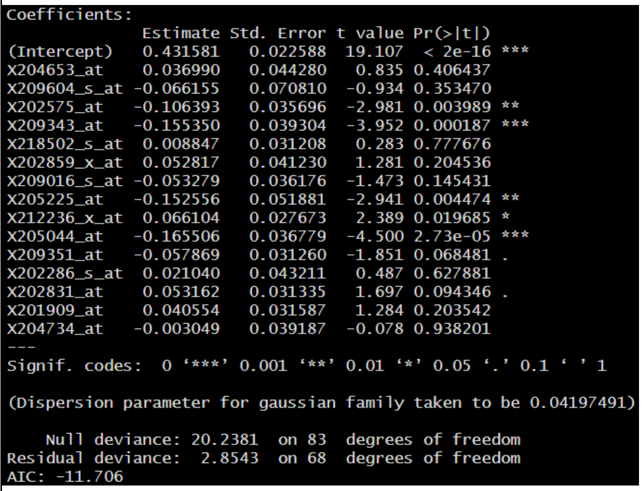


Fig 4. Binomial deviance plot for Elastic Net

Features selected by Lasso Regression	Features selected by Elastic Net
"X201496_x_at" "X202575_at" "X202831_at" "X202859_x_at" "X204653_at" "X209016_s_at" "X209351_at" "X209604_s_at" "X212236_x_at" "X218502_s_at"	"X204653_at" "X209604_s_at" "X202575_at" "X209343_at" "X218502_s_at" "X202859_x_at" "X209016_s_at" "X205225_at" "X212236_x_at" "X205044_at" "X209351_at" "X202286_s_at" "X202831_at" "X201909_at" "X204734_at"
Total Number of features selected = 10	Total Number of features selected = 15

Using the features from above two different logistic regression model were fitted to find out which one is better. Below is the summary for both the models.

Logistic Model using Features from Lasso	Logistic Model using Features from Elastic Net
	

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.43191    0.02781  15.530  < 2e-16 ***
X201496_x_at  0.03002    0.02706   1.110 0.270829    
X202575_at   -0.09352    0.03359  -2.784 0.006829 ** 
X202831_at    0.04865    0.03869   1.258 0.212550    
X202859_x_at  0.09426    0.04904   1.922 0.058466 .  
X204653_at   -0.09132    0.04695  -1.945 0.055641 .  
X209016_s_at -0.11485    0.03333  -3.446 0.000946 ***
X209351_at   -0.06910    0.03039  -2.274 0.025921 *  
X209604_s_at -0.10017    0.05582  -1.795 0.076839 .  
X212236_x_at  0.03958    0.02821   1.403 0.164859    
X218502_s_at -0.04747    0.03557  -1.335 0.186092    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06379418)

    Null deviance: 20.238  on 83  degrees of freedom
Residual deviance:  4.657  on 73  degrees of freedom
AIC: 19.416
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.431581   0.022588  19.107  < 2e-16 ***
X204653_at    0.036990   0.044280   0.835 0.406437    
X209604_s_at -0.066155   0.070810  -0.934 0.353470    
X202575_at   -0.106393   0.035696  -2.981 0.003989 ** 
X209343_at   -0.155350   0.039304  -3.952 0.000187 ***
X218502_s_at  0.008847   0.031208   0.283 0.777676    
X202859_x_at  0.052817   0.041230   1.281 0.204536    
X209016_s_at -0.053279   0.036176  -1.473 0.145431    
X205225_at   -0.152556   0.051881  -2.941 0.004474 ** 
X212236_x_at  0.066104   0.027673   2.389 0.019685 *  
X205044_at   -0.165506   0.036779  -4.500 2.73e-05 ***
X209351_at   -0.057869   0.031260  -1.851 0.068481 .  
X202286_s_at  0.021040   0.043211   0.487 0.627881    
X202831_at    0.053162   0.031335   1.697 0.094346 .  
X201909_at    0.040554   0.031587   1.284 0.203542    
X204734_at   -0.003049   0.039187  -0.078 0.938201    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.04197491)

    Null deviance: 20.2381  on 83  degrees of freedom
Residual deviance:  2.8543  on 68  degrees of freedom
AIC: -11.706
```

Before comparing the summaries, we first interpret the values.

Estimate – The positive coefficient in the model shows that an increase in the relevant independent variable leads to a greater log-odds of the dependent variable being equal to 1 and the negative coefficient suggests that an increase in the relevant independent variable leads to a reduced log-odds of the dependent variable being equal to 1,

Std. Error – It represents the error in the uncertainty of the estimated coefficients.

t value - The t-statistics for each coefficient, obtained by dividing the coefficient estimate by the standard error.

Pr(>|t|) – This helps in determining whether the coefficient is significant or not. Value less than 0.05 is means the feature is more significant.

Null deviance: It represents the goodness-of-fit of a model by taking no predictors and uses it as a baseline.

Residual deviance: It represents the goodness-of-fit of a model by taking the included predictors.

AIC: It is a criterion for model selection that uses both fit and complexity. Lower the AIC value the better.

Comparison of both the models.

Metric	Logistic with Lasso	Logistic with Elastic Net
More Significant features (p-value <=0.05)	X202575_at, X202859_x_at, X204653_at, X209016_s_at, X209351_at, X209604_s_at	X202575_at, X209343_at, X205225_at, X212236_x_at, X205044_at, X209351_at, X202831_at
Accuracy	0.95	0.95
Precision	0.8888889	0.8888889
Recall	1	1
F1 - score	0.9411765	0.9411765
AUC - ROC	0.9895833	0.9791667
AIC	19.46	-11.706

Using the table above we can see that almost all the metric for both the models are same except for features and AIC.

But as our data is multicollinear and based on the lower AIC value for Elastic Net we will select Logistic Model with Elastic net.

Below is the plot for estimated coefficients of logistic model with Elastic net we and we already know how positive and negative coefficient work. So, this is our central figure with our main findings.

The coefficients were extracted from the summary of the logistic model with elastic net. And then plotted where red represents Positive coefficients and green represents Negative coefficients.

From the graph we can infer that the significant features from the logistic model are influencing the more for the prediction.

For 0 i.e., B Cancer Type - X202575_at, X209343_at, X205225_at, X205044_at, X209351_at,

For 1 i.e., C Cancer Type - X202831_at, X212236_x_at,

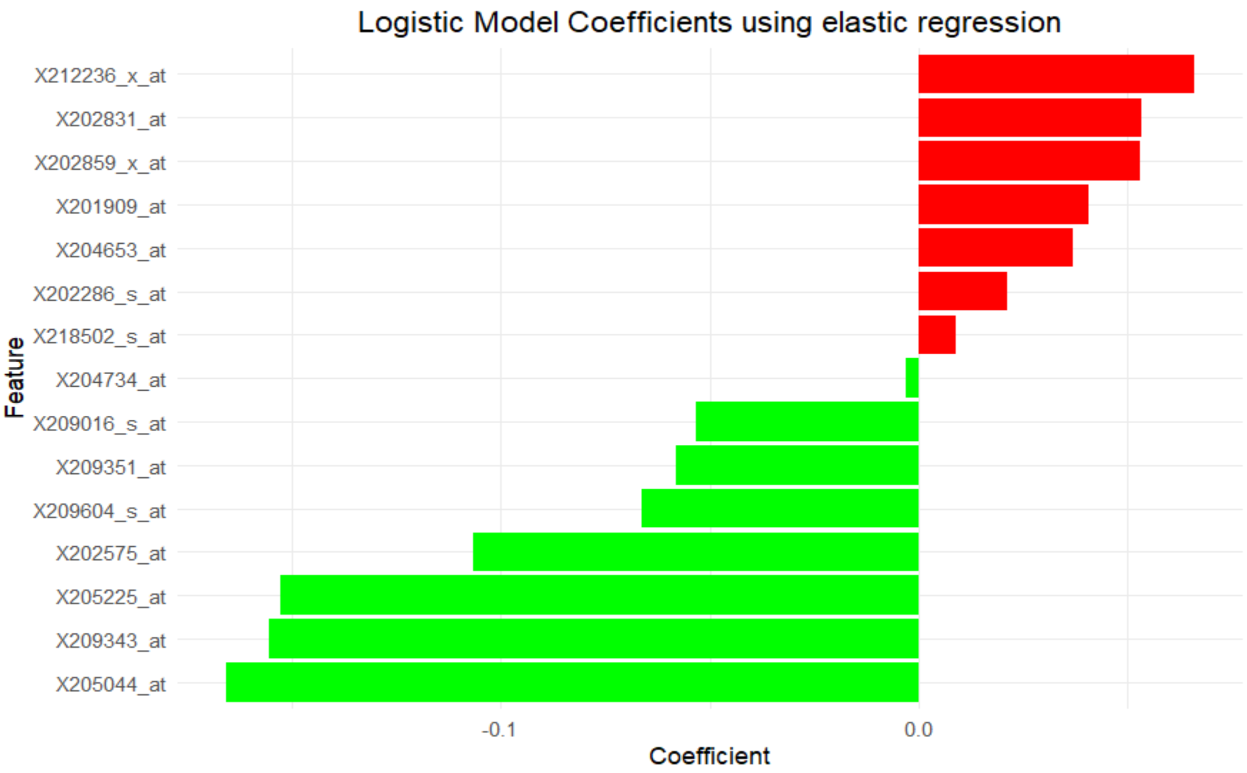


Fig 5. Central Figure of Findings

Conclusion:

The most influential genes are listed below which is also seen from the central figure for both of the cancer types.

B Cancer Type - X202575_at, X209343_at, X205225_at, X205044_at, X209351_at,

C Cancer Type - X202831_at, X212236_x_at

These are the more significant features but there are few genes that contribute comparatively less than the significant features such as X209016_s_at, X209604_s_at, X204734_at for Type B and "X204653_at", "X218502_s_at", "X202859_x_at" ,"X202286_s_at" ,"X201909_at" .