Discipline Project 2 – **Executive Summary (480388575)**

Guiding aim

Whilst kidney transplants already suffer from limited donor resources, acute renal rejection further impairs patient recovery and frustrates progress. Thus, this project aims to create an accessible classifier from gene expression data to predict acute renal rejection. More specifically, three sub-goals emerge: accuracy, transparency and accessibility.

The classifier has to be foremost accurate and robust to a diverse patient demographic. However, performance must balance with transparency, allowing users to easily interpret how various genes have quantitatively contributed as features to the model. Finally, it should be deployed in a user-friendly/customisable way to match differing needs.

Approach & Rationale

The key technique in the approach is LASSO regression. This essentially involves a penalised logistic regression with non-trivial (5468) gene expressions as starting parameters. Particularly, L1 regularisation penalises the absolute values of the parameter coefficients with respect to the Average Mean Squared Error (AMSE), inducing coefficient shrinking to zero and thus reducing the number of features based on a tuning parameter, λ. In implementation, this involves firstly finding the λ that minimises AMSE through Cross-Validation. Then, the final LASSO classifier is rebuild using that λ on full training data to optimise parameter shrinkage in maximising accuracy.

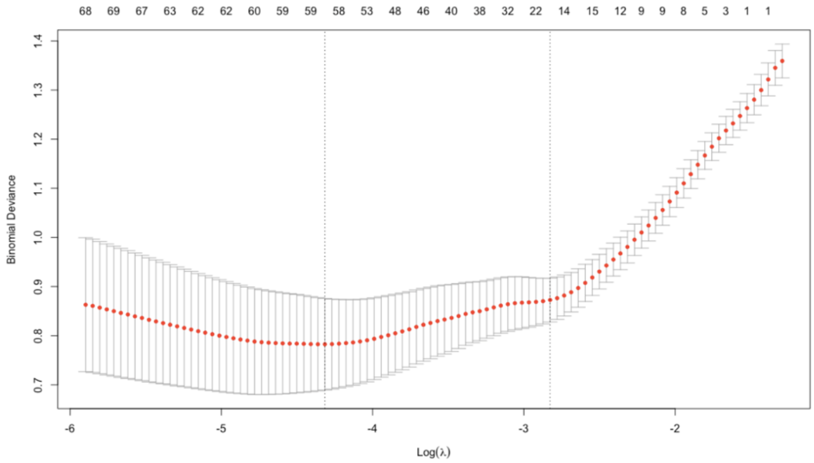


Figure 1: Plot of CV AMSE for optimising λ, with dotted lines indicating the chosen λs

The project classifier incorporated both GSE15296 and GSE14346 microarray data. As seen in Figure 1, the optimal λ corresponds to 58 genes being selected. Glmnet also provided a "next-best" model with substantially less features (14 genes) with comparable accuracy.

The reduction of features in this approach reduces the number of genes for better user interpretation compared to ridge regression. Domain knowledge that certain genes have minimal/no effect on renal rejection shows this is appropriate and makes performance time reasonable for deployment.

In this way, we can retain the transparency of various genes' quantitative contribution to the prediction outcome inherent in logistic regression (as compared to Random Forest and PCA), whilst addressing potential multi-collinearity that can invalidate interpretation of feature parameters, thus fulfilling our aim.

Issues & solutions

The first potential issue that arises is the robustness of the classifier, particularly as it still relies on some logistic regression distribution assumptions and could potentially overfit. To address this, ensemble bagging incorporates random sampling whilst not sacrificing model transparency. Average feature parameters can be used instead in Shiny App visualisations.

However, the main issue is a consideration of ethics. As this prediction can majorly influence patient kidney allocation, the model cannot be biased on unethical grounds like disproportionate positive predictions for certain races. Whilst data only consists of gene expressions, there is established research in racial differentiation in MicroRNA & gene expressions so this is a possibility.

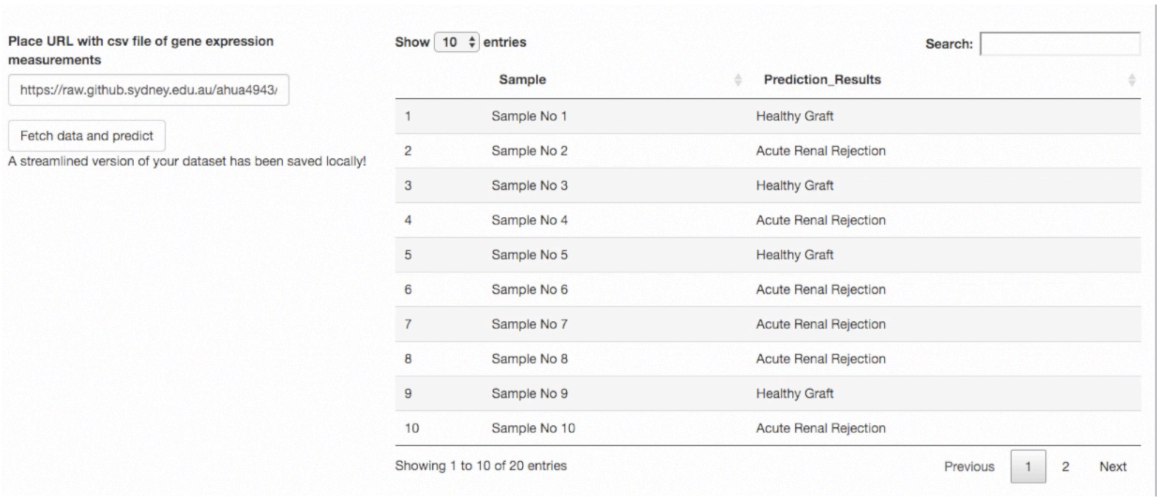
Whilst the classifier deployed has not accounted for potential violations of ethics, the current approach can mitigate this by calculating the predictive/statistical parity of the most influential genes in the model (shown in Shiny part (2)). Those failing a pre-determined threshold can be further analysed, and the gene can be removed from the training data and the LASSO classifier recalibrated if needed.

Shiny Deployment
(https://github.com/ahuang5045/DATA3888PROJ2)

So how does differing components of the Shiny App address our issues of performance, transparency and accessibility in predicting rejections?

1) Prediction with user-inputted data

Firstly, the app allows users to apply the LASSO classifier (optimal as default) to predict acute rejection based on inserting an URL to sampled csv data. This deployment is robust to large patient samples and differing genes, provided those in the LASSO classifier are present.


Place URL with csv file of gene expression measurements

https://raw.github.sydney.edu.au/ahua4943/

A streamlined version of your dataset has been saved locally!

Show 10 entries

Search:

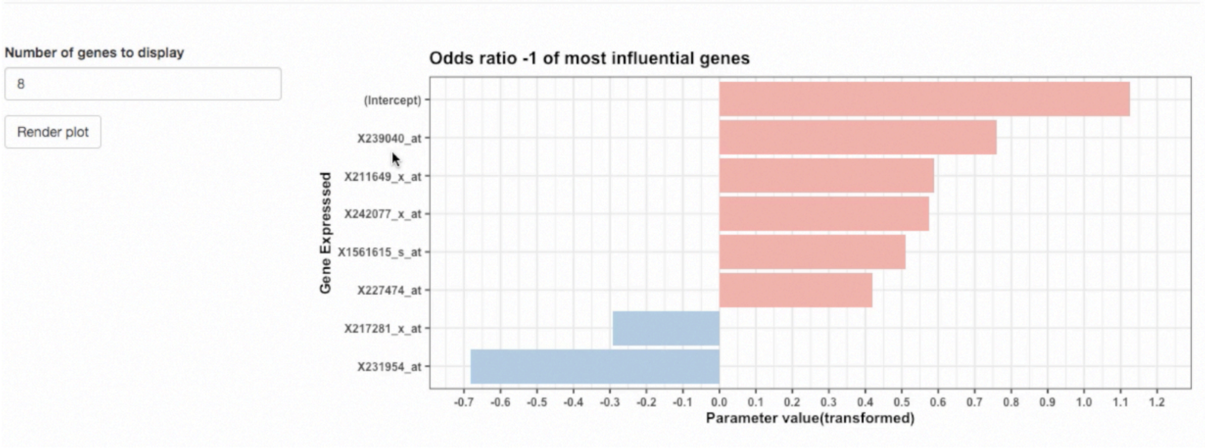
	Sample	Prediction_Results
1	Sample No 1	Healthy Graft
2	Sample No 2	Acute Renal Rejection
3	Sample No 3	Healthy Graft
4	Sample No 4	Acute Renal Rejection
5	Sample No 5	Healthy Graft
6	Sample No 6	Acute Renal Rejection
7	Sample No 7	Acute Renal Rejection
8	Sample No 8	Acute Renal Rejection
9	Sample No 9	Healthy Graft
10	Sample No 10	Acute Renal Rejection

Showing 1 to 10 of 20 entries

This provides accessible, accurate results in fulfilling the overall aim/issue of providing predictions.

2) Interactive graph

In maximising model transparency, this graph visualises the a user-defined number of genes by their quantitative influence in the classifier. The user can clearly interpret how specific genes expressions relatively increase/decrease rejection probability. For example, a one-fold increase of *X239040_at* expression increases the odds of a rejection outcome by 75% (parameter value 0.75), whilst the same for *X231954_at* reduces odds by 68%.

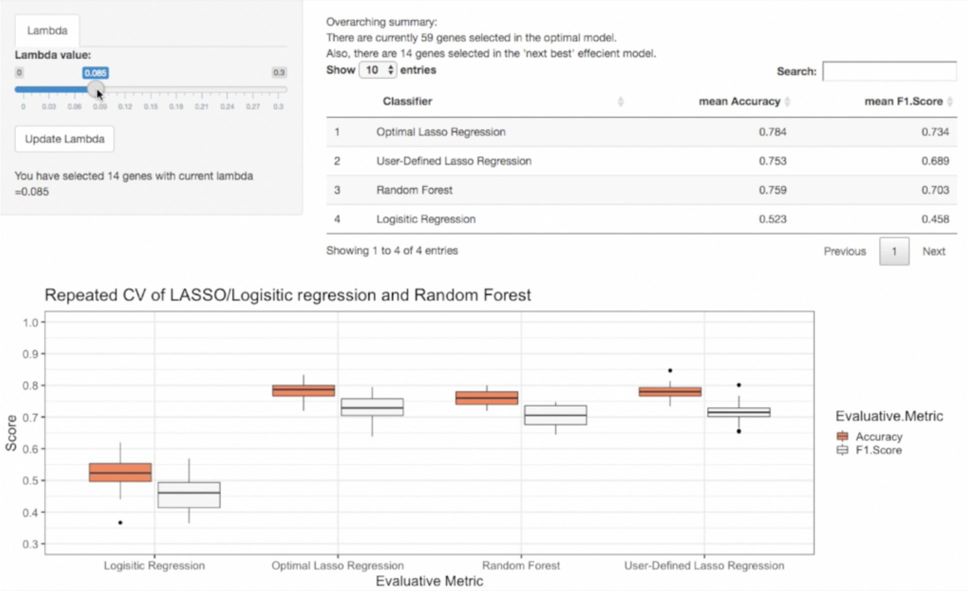


Number of genes to display: 8

This "biomarks" important genes for further antibody research and provides the basis for ethical accountability.

3) CV-evaluation with user-defined LASSO

This section allows for further feature reduction by altering λ (the hyper-parameter), illustrating its effect on the accuracy and F1 score in repeated CV. Furthermore, it shows the comparable accuracy of LASSO regression to random forest, and its improvement on logistic regression in validating the chosen approach and prediction accuracy at 78%.



Lambda value: 0.085

You have selected 14 genes with current lambda =0.085

Overarching summary:
There are currently 59 genes selected in the optimal model.
Also, there are 14 genes selected in the 'next best' effecient model.

Show 10 entries

	Classifier	mean Accuracy	mean F1.Score
1	Optimal Lasso Regression	0.784	0.734
2	User-Defined Lasso Regression	0.753	0.689
3	Random Forest	0.759	0.703
4	Logisitic Regression	0.523	0.458

Showing 1 to 4 of 4 entries

This allows the user to lower parameters for a faster prediction in part (1) after they have assured comparable accuracy, granting accessibility and transparency whilst ensuring performance.