**Task 1**

***Predictive model of Y1***

(1)First, we can see below the distribution of Y1 is close to the gaussian distribution.



Chart, scatter chart

Description automatically generated

(2)Hence, since 50 predictors seem to be too many, I used Lasso Regression to apply penalization for a better overall MSE. Also, by allowing coefficient to shrink to zero, Lasso Regression performed feature selection, in case of multicollinearity existing in the data, and to achieve a simpler model.

We set the lambda value that minimizes the MSE as the most optimal lambda value. As we can see in the plot of MSE against log(lambda) below, the optimal lambda value is 0.0001739218.

Text

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

(3) Then I used the optimal lambda value to fit into the final model. From the result below, we can see 34 out of 50 predictors are left after feature selection by lasso regression. 16 predictors were filtered out since they are not influential enough to Y1.

Text

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generatedA picture containing text, watch, gauge

Description automatically generated(4) I used the fitted lasso regression model to predict Y1, and calculate the R-squared value using the equation: , where SSE (Sum Squared Regression Error) =  , SST (Sum Squared Total Error) = . And RMSE (Root Mean Square Error) = .

Here,  is coefficient of determination and can measure how well the fitted model fits the observations. As shown below the  is approximately 0.65, which means the fitted model can correctly fit 65% of the data in the training dataset. RMSE is another measurement of the model usefulness. Here, RMSE is low at 0.02255695. Both and RMSE indicate acceptable performance.

Graphical user interface, text, application

Description automatically generated with medium confidence

***Predictive model of Y2***

(1) I plotted Y2 as shown below and found it is also close to gaussian distribution.

Icon

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

(2) Hence, I went through the same workflow as for Y1 as shown below. We can see the predictors after filtering is only 5. The and RMSE are also not ideal at about 0.0028 and 0.0155, respectively. is too low, showing a poor prediction performance.



A picture containing application

Description automatically generated

Chart

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

(3) I then reconsidered a better model to fit Y2. I tried using Lasso Regression model to fit the |Xi...50| and Y2 as well as (Xi...50)2  and Y2, as shown below. We can see both the two model can achieve a good performance with and RMSE at about 0.93, 0.0041 for |Xi...50| model, and 0.96, 0.0020 for (Xi...50)2  model. In comparison, (Xi...50)2  model has a better performance with higher usefulness.

**# Try |Xi...50|**



Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

**# Try (Xi...50)2**



Graphical user interface

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

***Predictive model of Y3***

Y3 is a binary outcome, so I first tried using Lasso logistic regression model to fit it. The workflow is similar to the previous ones for Y1 and Y2, except for some arguments that were changed as shown below in R codes. By changing the ‘type.measure’ into ‘class’, the cross-validation provided the lambda value that corresponded to the minimum misclassification error. We can see below the usefulness evaluated by Lasso is 98.46%, which is high enough.

A picture containing graphical user interface

Description automatically generated

Chart

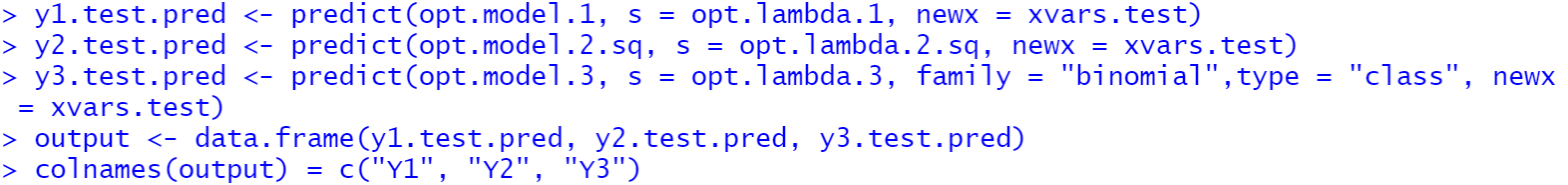
Description automatically generated

A picture containing graphical user interface

Description automatically generated

***Y1, Y2, Y3 prediction for data in test.csv***

I used R codes shown below to predict Y1, Y2, Y3 for the data in test.csv, and write them into a new csv file.



Text

Description automatically generated

