Regularized Logistic Regression

Shahzeb Naveed, Dafe Eboh

MSCI 718 – Final Project

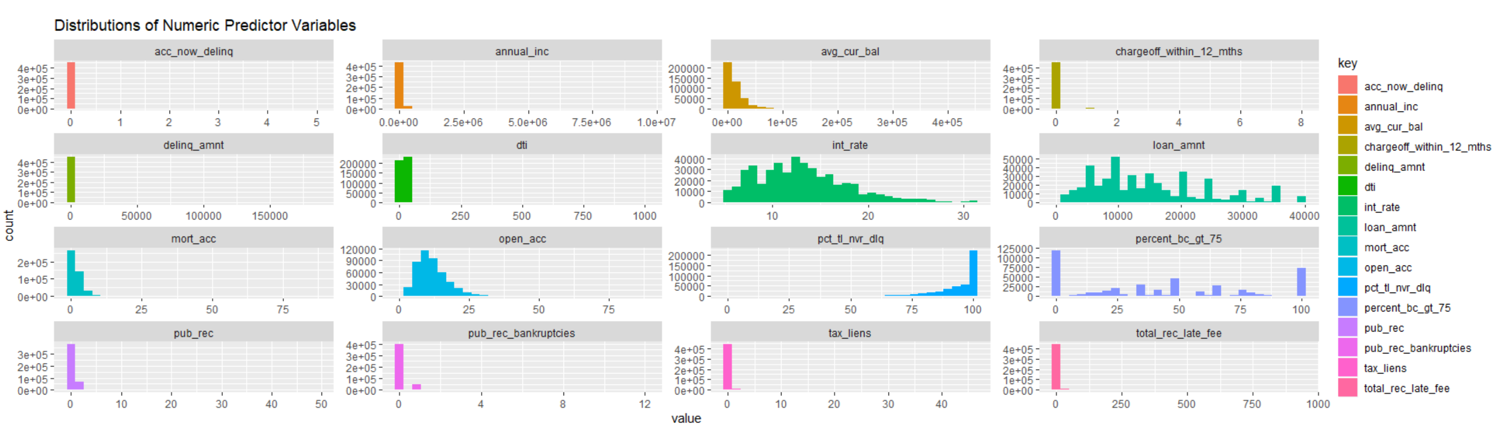
Introduction

Post financial crisis of 2008, a great emphasis has been laid on risk management within financial institutions to enhance transparency, consumer protection and better business decisions. To aid banks in identifying the creditworthiness of loan applicants, we apply regression modelling to predict whether a loan borrower will default or not. The Lending Club Loan Data was employed that initially contained 2260668 rows and 145 features. Using the given data dictionary, we hypothesized 19 variables as potential significant predictors which are tabulated below:

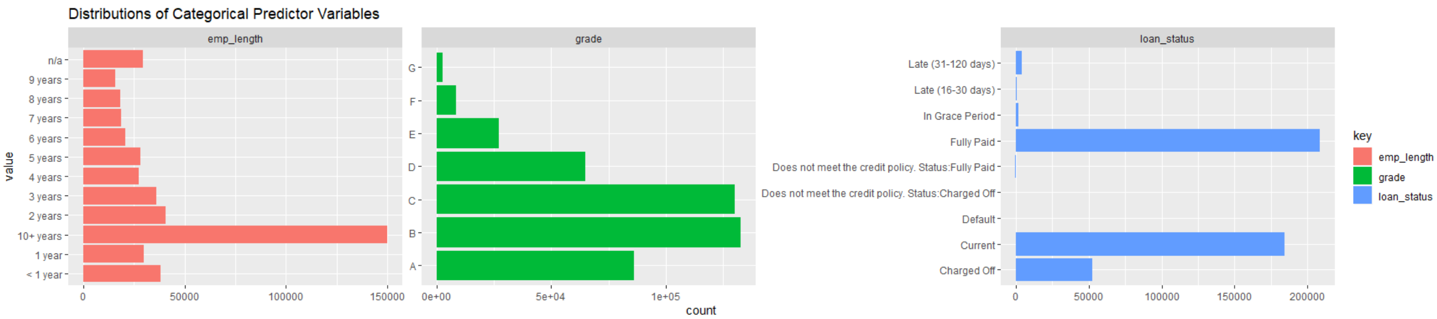
Variable	Type	Description
acc_now_delinq	Integer	The number of accounts on which the borrower is now delinquent.
annual_inc	Integer	The self-reported annual income provided by the borrower during registration.
avg_cur_bal	Integer	Average current balance of all accounts
chargeoff_within_12_mths	Integer	Number of charge-offs within 12 months
delinq_amnt	Integer	The past-due amount owed for the accounts on which the borrower is now delinquent.
dti	Integer	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations,
emp_length	Factor	Employment length in years. Possible values are between 0 and 10 where 0 means ten or more years.
grade	Factor	LC assigned loan grade
int_rate	Integer	Interest Rate on the loan
loan_amnt	Integer	Listed amount of the loan applied for by the borrower. If at some point the loaned in this value.
loan_status	Factor	Current status of the loan
mort_acc	Integer	Number of mortgage accounts.
open_acc	Integer	The number of open credit lines in the borrower's credit file.
pct_tl_nvr_dlq	Integer	Percent of trades never delinquent
percent_bc_gt_75	Integer	Percentage of all bankcard accounts > 75% of limit.
pub_rec	Integer	Number of derogatory public records
pub_rec_bankruptcies	Integer	Number of public record bankruptcies
tax_liens	Integer	Number of tax liens

Exploratory Data Analysis

After ensuring the data is in a Tidy form, we explore the distributions of our variables of interest.

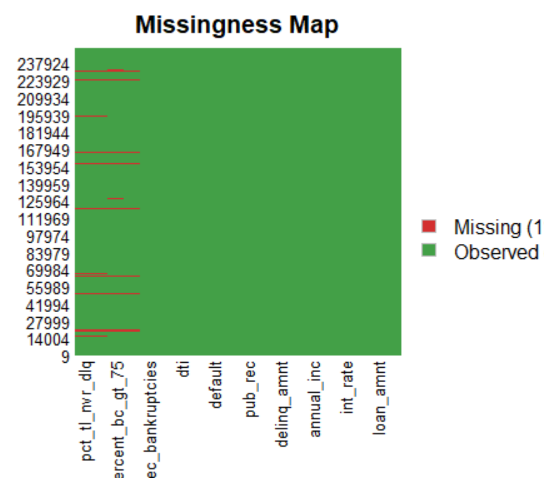


For outcome variable `default`, we used `loan_status` to combine categories of timely and late payments as "Not Default" and default and charged-off as "Default".



We now plot a missingness map to explore the proportion of NAs concluding that only 1% of the values are missing.

We now box-plot our numeric variables to visualize outliers. Practically speaking, the extreme points do not indicate any data entry errors as they might be just some very rich people, people who got loans on very high interest rates or people with unusual financial circumstances for example. For the purpose of generalizability, we decided not to remove these outliers and will incorporate these into our prediction models. (See Appendix)



Incomplete Information: A very important requirement of logistic regression is Incomplete Separation that can lead to unusually high standard errors. A 3-way crosstabulated table was drawn to make sure we have some data in every possible combination. (See Appendix)

For Complete Separation, we plotted every predictor against variable to visualize it. Note that complete separation may arise even when predictors do not exhibit it individually but that is beyond the scope of our project. (See Appendix)

Model Building

Since `default` is a binary variable, a logistic regression model was used. But before diving into modelling, we first take a look at some of the underlying assumptions below.

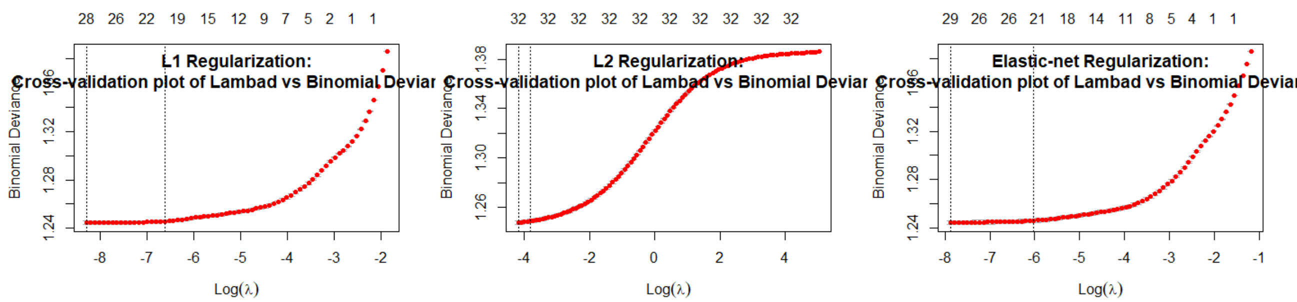
Assumptions of Logistic Regression

1. **Large sample size:** `250104` rows. Enough said.
2. **No or Less Multicollinearity:** We'll come to that in a while.
3. **Linearity of predictors and logit of outcome variable:** We'll come to that in a while as well.
4. **Complete Information:** As mentioned earlier, we have `0` NAs and have data for every possible combination. (See Appendix)
5. **Incomplete Separation:** We can clearly see that there is no complete separation in the scatterplots. (See Appendix)

Class Imbalance: Since our data contains `201859` default cases and `48245` not default cases, we up-sampled our minority class before training our model.

Feature Selection

We trained 4 logistic regression models: Simple, L1-Regularized, L2-Regularized and Elastic-Net with *alpha* arbitrarily chosen mid-way between L1 and L2 as `0.5`. With regularization, the optimal penalty measure *lambda* is selected such that it minimizes the cross-validated out-of-sample accuracy error. In short, L1 forces 4 of the co-efficients to exactly zero thus aiding us in variable selection and model simplification. L2 forces some co-efficients close to zero while Elastic-net forces 3 of the variables exactly zero while forcing some of the rest close to zero. In the graph below, Binomial Deviance is plotted against Log(lamda), where the left dashed line indicates the value of lambda that minimizes out-of-sample accuracy error. (See Appendix)



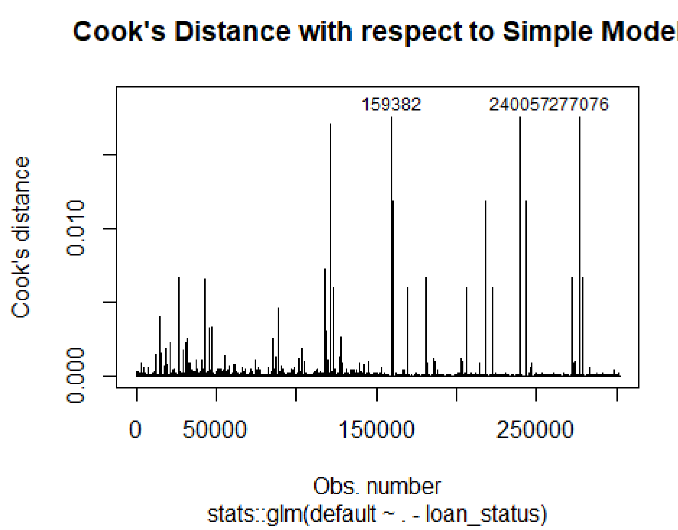
Model Evaluation: Goodness-of-Fit

To evaluate model fit, the Deviance Statistics and H&L R^2 values (that can also be used as effect size) for all of our models have been compared below. We see that there is not much difference in the model fit with Model 1 doing slightly better. (See Appendix)

Method	Chi.Statistic	df	p.value	R.Square
Simple Regression	108090.9	32	0	0.1032000
L1	108053.5	28	0	0.1031495
L2	105928.8	32	0	0.1011213
Elastic-net	108046.6	28	0	0.1031430

Diagnostics

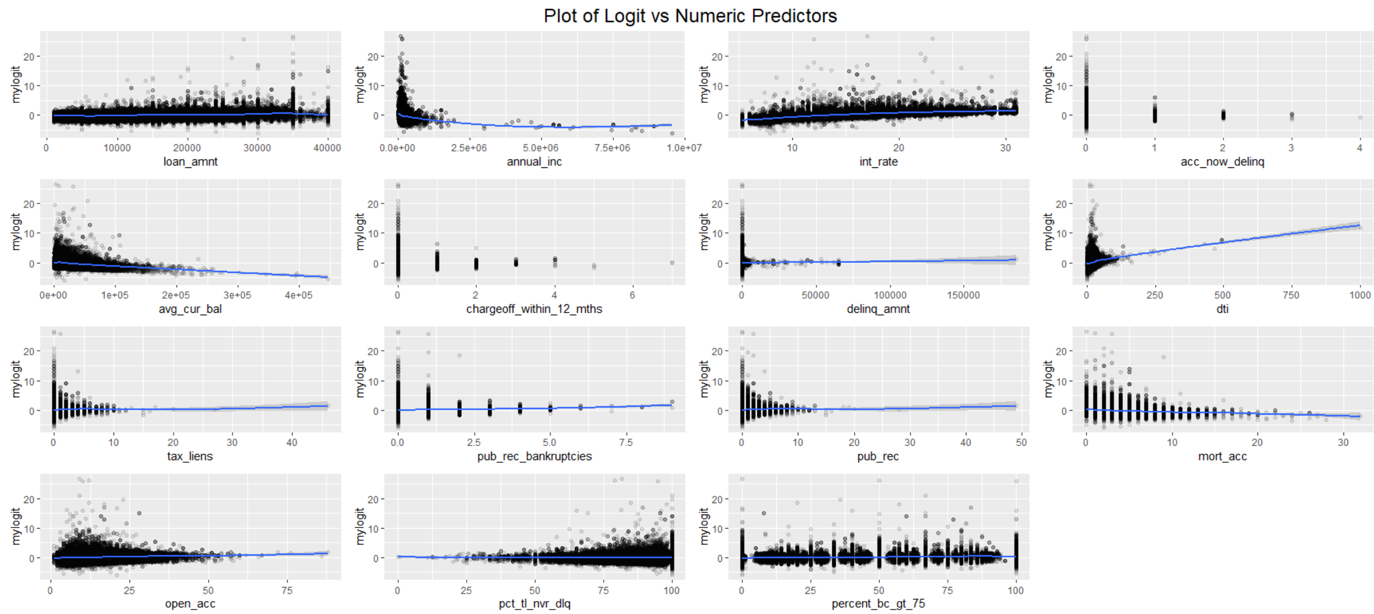
To make sure none of our observations have an unfair influence on our model, we plot a Cook's distance plot using predicted probabilities from Model 1 and observe that even the most influential observation has a cook's distance of `0.0175372 so we are good to go.`



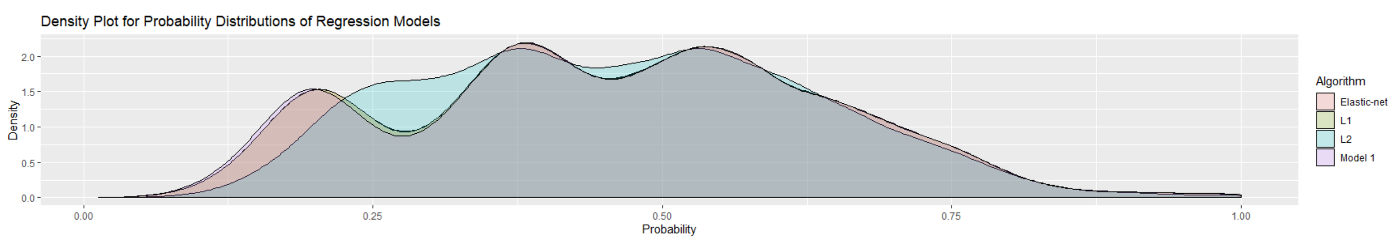
Remaining Assumptions

To test for multicollinearity among our predictors, we calculate GVIF and observe that none of the variables have GVIF's greater than 10 which, if adjusted for the degrees of freedom, are even less. (See Appendix)

For ensuring logit-linearity, we employ Box-Tidwell test on Model 1 and find that all the log-interaction terms are signficant indicating non-linearity. But this is highly probable because of the massive sample size that always results in small standard errors, resulting in extremely significant z-statistics. A better and an easier way it to simply plot the logit against the predictor variables. As can be seen below, the relationship can safely be approximated as linear. Note that we don't need a non-linearity test for categorical variables since they are coded as dummy variables with values 0 and 1 so the relationship becomes "linear" by definition since we have only two points to connect.



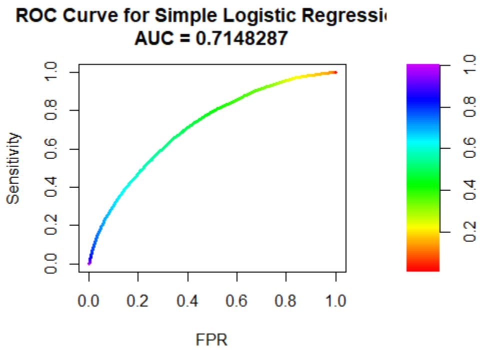
To visualize our models, probability distributions have been compared below.

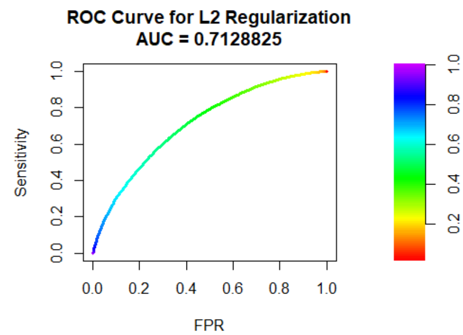


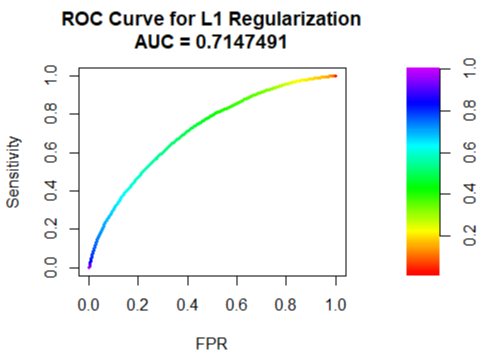
Testing Prediction Accuracy

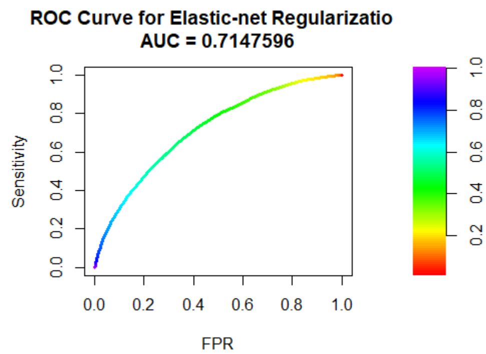
We now evaluate our models based on out-of-sample accuracy and observe that for our data, all the 4 models designed perform, in a similar fashion. Their ROC curves alongwith AUCs as well as their various evaluation metrics are calculated below. In terms of all the metrics, all the models perform similarly on our data.

Method	Accuracy	Precision	Recall	F1.Score
Simple Regression	0.2854371	0.2109362	0.9803	0.347171
L1	0.2792111	0.2097339	0.9823	0.345663
L2	0.2353092	0.2014629	0.9938	0.335015
Elastic-net	0.2788273	0.2096316	0.9822	0.345518









Gap Analysis and Future Work

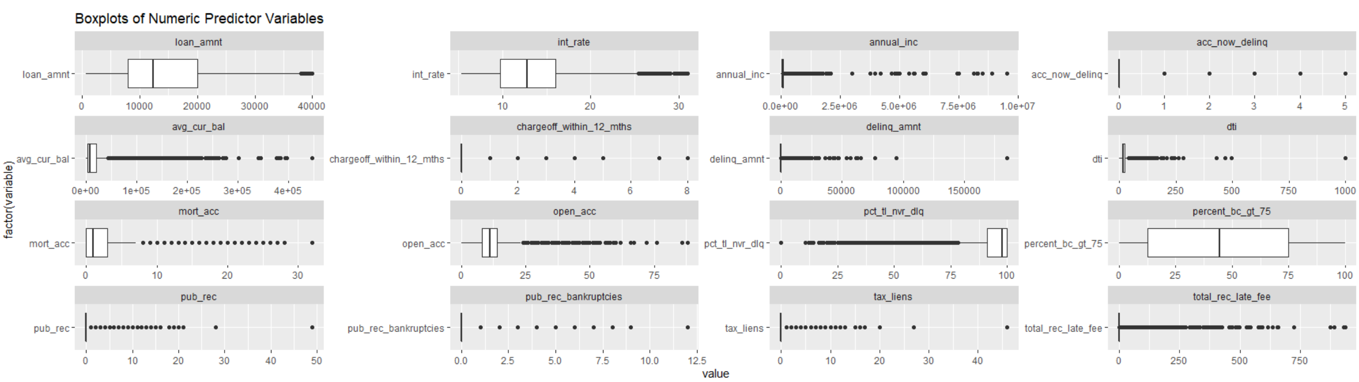
For future work, outlier observations may be considered to be removed to evaluate their performance on out-of-sample observations. Furthermore, *alpha* may be calculcated using cross-validation from `caret` package for a more accurate Elastic-Net models. Moreover, new variables might be included in the model after careful study from the data dictionary.

Conclusion

To conclude, regardless of the levels of regularization, logistic regression gives us a similar performance. With a slight difference, Model 1 has better accuracy, precision and F1-scores of 0.28, 0.31 and 0.347 respectively and L2 Model being better than others in terms of Re-call. With L1 and Elastic Regularization, we were able to safely remove 4 and 3 predictors respectively resulting in a simpler model without any significant decrease in prediction accuracy. The final magnitudes of co-efficients are shown in the Appendix.

Appendix

1. Boxplot of Numerical Predictors



2. Incomplete Information

```
##           < 1 year 1 year 10+ years 2 years 3 years 4 years 5 years 6 years 7 years
8 years 9 years
##
## FALSE A      2473   1873     10982    2739    2508    1871    1928    1458    1375
1455    1180
##       B      3940   3252     16761    4418    4025    2928    3186    2330    2206
2307    1930
##       C      3597   2928     15155    4078    3545    2648    2709    2040    1961
2018    1773
##       D      1631   1396      7016    1886    1675    1222    1318    1032    1001
955     752
##       E       702    554      2953     787     646     479     515     403     385
395     360
##       F       212    182       946     231     203     147     170     124     135
152     109
##       G        56     58       232      55      64      43      42      32      26
37      32
## TRUE  A       685    428      2853     725     502     409     511     309     395
354     225
##       B      2550   2178      9820    2686    2531    1772    1920    1485    1287
1455    1278
##       C      4614   3523     15956    4737    4105    2911    3293    2249    2395
2240    2005
##       D      3094   2524     12022    3437    3146    2153    2261    1591    1641
1610    1288
```

```
##          E       1716    1418       7237     2075     1922     1300     1210     1092      856
1051      881
##          F        620     466       2973      862      722      526      490      488      378
450      341
##          G        219     187       1000      273      193      182      186      162      113
118       68
```

3. Complete Separation

4. Model 1 Summary:

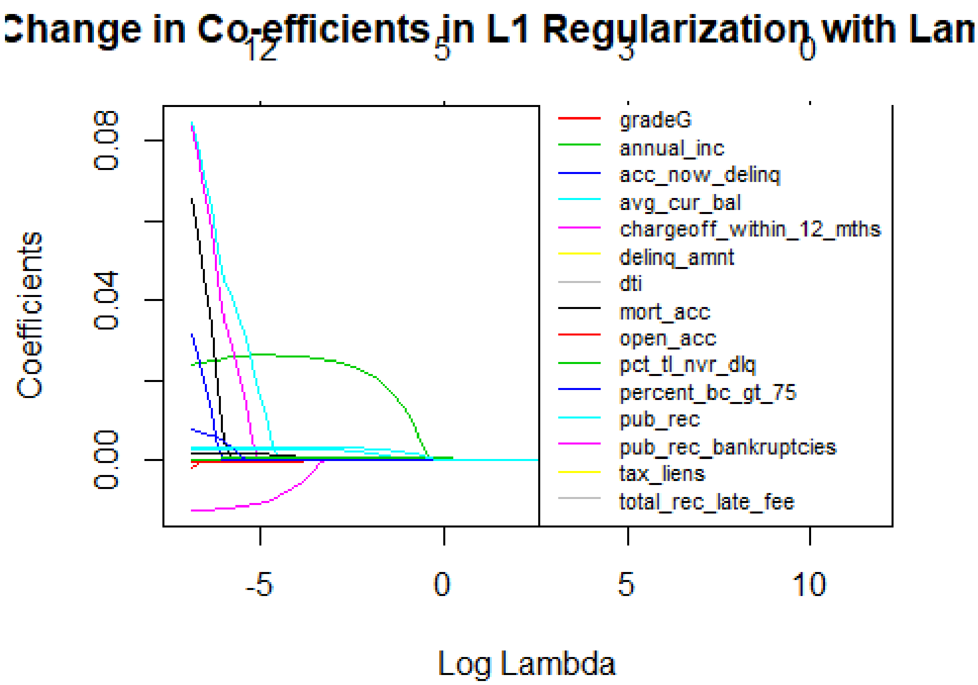
```
##
## Call:
## stats::glm(formula = default ~ . - loan_status, family = binomial(link = "logit"),
##     data = train.data.1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.3059  -1.0564  -0.0357   1.0490   2.6931
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.777e+00  5.177e-02 -34.321  < 2e-16 ***
## loan_amnt                  1.449e-05  5.133e-07  28.234  < 2e-16 ***
## emp_length1 year          -2.985e-02  1.984e-02  -1.505 0.132361
## emp_length10+ years       -7.062e-02  1.495e-02  -4.725 2.30e-06 ***
## emp_length2 years         -3.743e-02  1.829e-02  -2.047 0.040664 *
## emp_length3 years         -3.376e-02  1.881e-02  -1.795 0.072728 .
## emp_length4 years         -7.959e-02  2.059e-02  -3.866 0.000111 ***
## emp_length5 years         -5.323e-02  2.021e-02  -2.633 0.008460 **
## emp_length6 years         -8.523e-02  2.209e-02  -3.858 0.000114 ***
## emp_length7 years         -5.267e-02  2.238e-02  -2.354 0.018571 *
## emp_length8 years         -6.600e-02  2.217e-02  -2.976 0.002918 **
## emp_length9 years         -6.140e-02  2.347e-02  -2.616 0.008894 **
## int_rate                   2.473e-02  2.689e-03   9.195  < 2e-16 ***
## gradeB                     7.234e-01  1.800e-02  40.188  < 2e-16 ***
## gradeC                     1.188e+00  2.381e-02  49.869  < 2e-16 ***
## gradeD                     1.477e+00  3.275e-02  45.106  < 2e-16 ***
## gradeE                     1.719e+00  4.208e-02  40.847  < 2e-16 ***
## gradeF                     1.829e+00  5.473e-02  33.411  < 2e-16 ***
## gradeG                     1.956e+00  7.206e-02  27.151  < 2e-16 ***
```

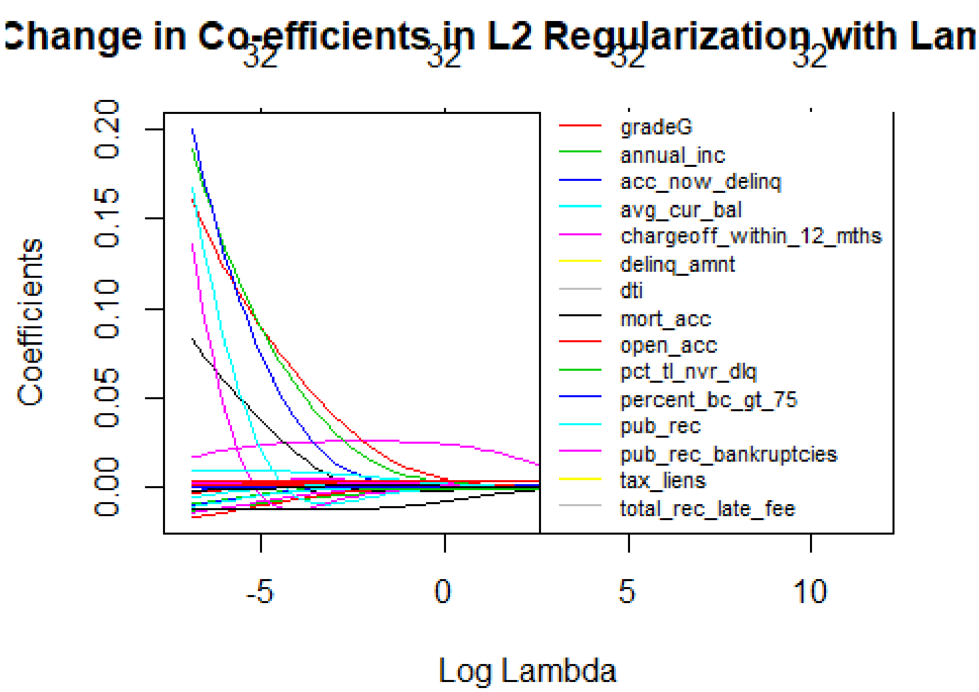
```
## annual_inc                -4.828e-07  8.722e-08  -5.535 3.11e-08 ***
## acc_now_delinq            -1.471e-02  5.129e-02  -0.287 0.774318
## avg_cur_bal               -9.197e-06  3.413e-07 -26.947  < 2e-16 ***
## chargeoff_within_12_mths -1.249e-02  3.541e-02  -0.353 0.724252
## delinq_amnt                7.444e-06  4.619e-06   1.612 0.107013
## dti                        1.289e-02  5.003e-04  25.764  < 2e-16 ***
## mort_acc                  -5.586e-02  2.436e-03 -22.928  < 2e-16 ***
## open_acc                   8.019e-03  7.879e-04  10.178  < 2e-16 ***
## pct_tl_nvr_dlq            -4.030e-04  4.583e-04  -0.879 0.379289
## percent_bc_gt_75           1.044e-03  1.150e-04   9.080  < 2e-16 ***
## pub_rec                    4.139e-02  1.766e-02   2.344 0.019060 *
## pub_rec_bankruptcies       9.999e-03  2.023e-02   0.494 0.621103
## tax_liens                 -8.906e-03  2.064e-02  -0.431 0.666128
## total_rec_late_fee         2.778e-02  5.011e-04  55.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 418461  on 301855  degrees of freedom
## Residual deviance: 375776  on 301823  degrees of freedom
## AIC: 375842
##
## Number of Fisher Scoring iterations: 5
```

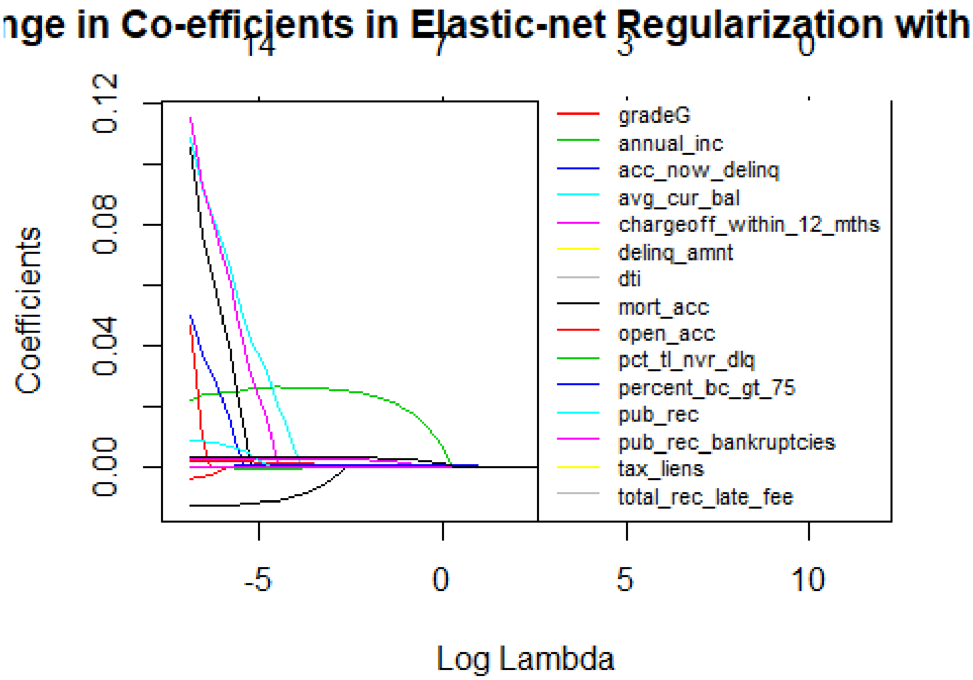
5. CheckInfiniteEstimates() for Model 1:

```
## Warning: package 'detectseparation' was built under R version 3.6.3

## Warning: package 'brglm2' was built under R version 3.6.3

## Registered S3 method overwritten by 'brglm2':
##   method                   from
##   print.detect_separation detectseparation

##
## Attaching package: 'brglm2'

## The following objects are masked from 'package:detectseparation':
##
##     check_infinite_estimates, checkInfiniteEstimates,
##     detect_separation, detect_separation_control, detectSeparation,
##     detectSeparationControl
```

6. Effect of Regularization parameter Lambda on Model Co-efficients


Change in Co-efficients in L1 Regularization with Lan
12
5
3
0
0.00
0.04
0.08
Coefficients
gradeG
annual_inc
acc_now_delinq
avg_cur_bal
chargeoff_within_12_mths
delinq_amnt
dti
mort_acc
open_acc
pct_tl_nvr_dlq
percent_bc_gt_75
pub_rec
pub_rec_bankruptcies
tax_liens
total_rec_late_fee
-5
0
5
10
Log Lambda

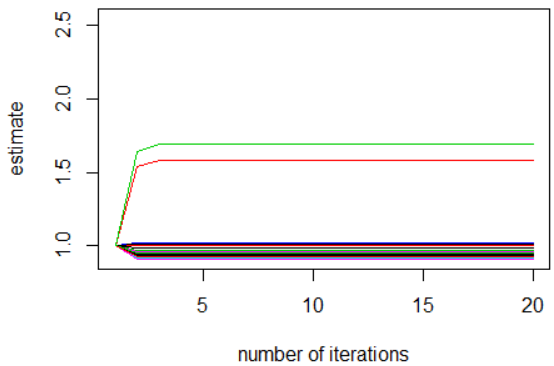

Change in Co-efficients in L2 Regularization with Lan
32
32
32
32
0.00
0.05
0.10
0.15
0.20
Coefficients
gradeG
annual_inc
acc_now_delinq
avg_cur_bal
chargeoff_within_12_mths
delinq_amnt
dti
mort_acc
open_acc
pct_tl_nvr_dlq
percent_bc_gt_75
pub_rec
pub_rec_bankruptcies
tax_liens
total_rec_late_fee
-5
0
5
10
Log Lambda



7. Box-Tidwell Logit Linearity Test

```
##
## Call:
## glm(formula = default ~ loan_amnt + int_rate + annual_inc + log.loan_amnt +
##     log.int_rate + log.annual_inc, family = binomial(link = "logit"),
##     data = train.data.1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.64200  -1.09941   0.00212   1.05025   2.19876
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.012e+00  6.230e-02  -80.44   <2e-16 ***
## loan_amnt       3.804e-04  1.488e-05   25.56   <2e-16 ***
## int_rate        8.525e-01  1.425e-02   59.83   <2e-16 ***
## annual_inc     -1.549e-05  6.100e-07  -25.39   <2e-16 ***
## log.loan_amnt  -3.387e-05  1.392e-06  -24.34   <2e-16 ***
## log.int_rate   -1.956e-01  3.823e-03  -51.16   <2e-16 ***
## log.annual_inc  9.897e-07  4.198e-08   23.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 418461  on 301855  degrees of freedom
## Residual deviance: 383688  on 301849  degrees of freedom
## AIC: 383702
##
## Number of Fisher Scoring iterations: 4
```

8. Plot of co-efficients over iterations for Model 1:



9. Co-efficients from all Models:

10.
```
## (Intercept) loan_amnt emp_length1 year
## -1.776615e+00 1.449353e-05 -2.985149e-02
## emp_length10+ years emp_length2 years emp_length3 years
## -7.061771e-02 -3.742846e-02 -3.376260e-02
## emp_length4 years emp_length5 years emp_length6 years
## -7.959469e-02 -5.322884e-02 -8.522664e-02
## emp_length7 years emp_length8 years emp_length9 years
## -5.267467e-02 -6.599771e-02 -6.139978e-02
## int_rate gradeB gradeC
## 2.472607e-02 7.234120e-01 1.187598e+00
## gradeD gradeE gradeF
## 1.477224e+00 1.718794e+00 1.828576e+00
## gradeG annual_inc acc_now_delinq
## 1.956472e+00 -4.828151e-07 -1.470723e-02
## avg_cur_bal chargeoff_within_12_mths delinq_amnt
## -9.197487e-06 -1.249191e-02 7.444340e-06
## dti mort_acc open_acc
## 1.289021e-02 -5.586117e-02 8.019244e-03
## pct_tl_nvr_dlq percent_bc_gt_75 pub_rec
## -4.029750e-04 1.044348e-03 4.138993e-02
## pub_rec_bankruptcies tax_liens total_rec_late_fee
## 9.999224e-03 -8.906389e-03 2.777615e-02
```

11.
```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                  s0
## (Intercept)           -1.847658e+00
## loan_amnt              1.422975e-05
## emp_length1 year       .
## emp_length10+ years   -3.867739e-02
## emp_length2 years     -3.095326e-03
## emp_length3 years      .
## emp_length4 years     -4.444441e-02
## emp_length5 years     -1.870353e-02
## emp_length6 years     -5.007175e-02
## emp_length7 years     -1.705085e-02
## emp_length8 years     -2.939853e-02
## emp_length9 years     -2.508878e-02
```

```
## int_rate                   3.312855e-02
## gradeB                     6.657357e-01
## gradeC                     1.102932e+00
## gradeD                     1.360958e+00
## gradeE                     1.572992e+00
## gradeF                     1.647162e+00
## gradeG                     1.742970e+00
## annual_inc                -4.447212e-07
## acc_now_delinq             .
## avg_cur_bal               -9.199038e-06
## chargeoff_within_12_mths -1.080676e-03
## delinq_amnt                5.914870e-06
## dti                        1.290664e-02
## mort_acc                  -5.532619e-02
## open_acc                   7.751861e-03
## pct_tl_nvr_dlq            -2.974232e-04
## percent_bc_gt_75           1.036150e-03
## pub_rec                    3.398496e-02
## pub_rec_bankruptcies       1.496708e-02
## tax_liens                  .
## total_rec_late_fee         2.753079e-02
```

12.
```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                      s0
## (Intercept)               -1.813113e+00
## loan_amnt                  1.387320e-05
## emp_length1 year          -5.844324e-03
## emp_length10+ years       -5.095583e-02
## emp_length2 years         -1.354122e-02
## emp_length3 years         -1.110203e-02
## emp_length4 years         -5.325444e-02
## emp_length5 years         -3.274214e-02
## emp_length6 years         -6.294400e-02
## emp_length7 years         -3.159602e-02
## emp_length8 years         -3.786288e-02
## emp_length9 years         -3.722938e-02
## int_rate                   7.474430e-02
## gradeB                     2.676262e-01
## gradeC                     5.528181e-01
## gradeD                     6.454266e-01
## gradeE                     7.067517e-01
## gradeF                     6.221125e-01
## gradeG                     5.981727e-01
## annual_inc                -4.746307e-07
## acc_now_delinq             1.254937e-02
## avg_cur_bal               -8.793879e-06
## chargeoff_within_12_mths -3.591454e-03
## delinq_amnt                6.803631e-06
## dti                        1.283330e-02
## mort_acc                  -5.421686e-02
## open_acc                   7.793877e-03
## pct_tl_nvr_dlq            -1.257132e-03
## percent_bc_gt_75           1.365738e-03
## pub_rec                    3.780301e-02
## pub_rec_bankruptcies       2.576824e-02
```

```
## tax_liens                  1.337643e-03
## total_rec_late_fee         2.297527e-02
```

13.
```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)               -1.844875e+00
## loan_amnt                  1.428800e-05
## emp_length1 year          .
## emp_length10+ years       -4.225138e-02
## emp_length2 years         -7.115876e-03
## emp_length3 years         -3.532623e-03
## emp_length4 years         -4.871429e-02
## emp_length5 years         -2.305999e-02
## emp_length6 years         -5.462609e-02
## emp_length7 years         -2.159102e-02
## emp_length8 years         -3.384735e-02
## emp_length9 years         -2.973042e-02
## int_rate                   3.398941e-02
## gradeB                     6.608839e-01
## gradeC                     1.094686e+00
## gradeD                     1.349235e+00
## gradeE                     1.558191e+00
## gradeF                     1.629525e+00
## gradeG                     1.724159e+00
## annual_inc                -4.544987e-07
## acc_now_delinq             .
## avg_cur_bal               -9.199234e-06
## chargeoff_within_12_mths -3.640448e-03
## delinq_amnt                6.235629e-06
## dti                        1.291600e-02
## mort_acc                  -5.541942e-02
## open_acc                   7.818056e-03
## pct_tl_nvr_dlq            -3.529974e-04
## percent_bc_gt_75           1.047713e-03
## pub_rec                    3.437590e-02
## pub_rec_bankruptcies       1.558951e-02
## tax_liens                  .
## total_rec_late_fee         2.751639e-02
```