```
## Loading required package: ggplot2

## Loading required package: lattice
```

Logistic Regression and LASSO

Since this is a binary classification problem, it's possible to use logistic regression as one of the models. First we will fit the full model as a baseline.

```
##
## Call:
## glm(formula = Default ~ ., family = "binomial", data = train)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.585e+00  8.293e-02  19.110  < 2e-16 ***
## Age                         -3.971e-02  7.104e-04 -55.901  < 2e-16 ***
## Income                      -8.354e-06  2.572e-07 -32.476  < 2e-16 ***
## LoanAmount                   3.987e-06  1.453e-07  27.447  < 2e-16 ***
## CreditScore                 -6.751e-04  6.423e-05 -10.511  < 2e-16 ***
## MonthsEmployed              -9.693e-03  2.982e-04 -32.508  < 2e-16 ***
## NumCreditLines               8.652e-02  9.124e-03   9.483  < 2e-16 ***
## InterestRate                 7.098e-02  1.582e-03  44.877  < 2e-16 ***
## LoanTerm                    -1.240e-03  6.012e-04  -2.063   0.0392 *
## DTIRatio                     2.646e-01  4.420e-02   5.986 2.16e-09 ***
## EducationBachelor's         -1.117e-01  2.826e-02  -3.953 7.71e-05 ***
## EducationMaster's           -2.224e-01  2.882e-02  -7.717 1.19e-14 ***
## EducationPhD                -2.792e-01  2.889e-02  -9.665  < 2e-16 ***
## EmploymentTypePart-time      2.787e-01  2.943e-02   9.470  < 2e-16 ***
## EmploymentTypeSelf-employed  2.476e-01  2.963e-02   8.357  < 2e-16 ***
## EmploymentTypeUnemployed     4.712e-01  2.915e-02  16.163  < 2e-16 ***
## MaritalStatusMarried        -1.676e-01  2.522e-02  -6.644 3.05e-11 ***
## MaritalStatusDivorced        5.228e-02  2.464e-02   2.122   0.0339 *
## HasMortgageYes              -1.782e-01  2.040e-02  -8.733  < 2e-16 ***
## HasDependentsYes            -2.286e-01  2.042e-02 -11.197  < 2e-16 ***
## LoanPurposeBusiness          4.186e-02  3.192e-02   1.311   0.1897
## LoanPurposeEducation        -4.417e-02  3.215e-02  -1.374   0.1694
## LoanPurposeHome             -1.893e-01  3.266e-02  -5.796 6.79e-09 ***
## LoanPurposeOther            -1.038e-02  3.222e-02  -0.322   0.7475
## HasCoSignerYes              -2.786e-01  2.043e-02 -13.635  < 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: 65774  on 47445  degrees of freedom
## Residual deviance: 56105  on 47421  degrees of freedom
## AIC: 56155
##
## Number of Fisher Scoring iterations: 3
```

Since our sample size is so large, almost every coefficient is significant with the exception of some of the Loan Purpose coefficients. However statistical significance does not always imply practical significance and

the large sample size makes the test always pick up on tiny miniscule differences. To actually help us find the important predictors, we will use a LASSO logistic regression to do so.

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

The way LASSO works is that when we optimize the likelihood to estimate the coefficients, we add a penalty term called the l_1 norm where it will penalize us for fitting useless variables and thus force some coefficients to become zero to weed out these useless variables. This can help us with model selection and find some important predictors.

The penalty term has a hyperparameter called λ which will control how harsh the penalty will be for fitting more terms. The larger the lambda, the less terms we will fit. We want to find a lambda that balances parsimony and model performance so we turned to cross validation and tried many values of lambda and we ended up using $\lambda = 0.03$.

```
## 25 x 1 sparse Matrix of class "dgCMatrix"
##                                         s0
## (Intercept)                  8.589759e-01
## Age                         -2.815760e-02
## Income                      -4.589212e-06
## LoanAmount                   1.904345e-06
## CreditScore                  .
## MonthsEmployed              -5.142304e-03
## NumCreditLines               .
## InterestRate                 4.572493e-02
## LoanTerm                     .
## DTIRatio                     .
## EducationBachelor's          .
## EducationMaster's            .
## EducationPhD                 .
## EmploymentTypePart-time      .
## EmploymentTypeSelf-employed  .
## EmploymentTypeUnemployed     .
## MaritalStatusMarried         .
## MaritalStatusDivorced        .
## HasMortgageYes               .
## HasDependentsYes             .
## LoanPurposeBusiness          .
## LoanPurposeEducation         .
## LoanPurposeHome              .
## LoanPurposeOther             .
## HasCoSignerYes               .
```

From the results of our LASSO, we ended up selecting 5 variables which are Age, Income, Loan Amount, Months Employed, and Interest Rate and reduced our model from 16 variables to 5 variables.

Now keep in mind these variables are all quantitative and have a large range of values so it's not shocking that the effects look small.

An interpretation of the coefficients in simpler terms is that older people are less likely to default due the negative coefficient. Similarly those who work longer and have higher income are less likely to default. However those who borrow more money and have higher interest rates are more likely to default due to the positive coefficient.

	Accuracy	AUC	TPR	TNR
Full	0.678	0.737	0.684	0.672
LASSO	0.664	0.723	0.679	0.650

We can see across all these metrics, the LASSO almost performs just as well as the full model with only a 1-2% difference especially with an 19 predictor difference. Interpreting these metrics for the LASSO model, the accuracy indicates that we correctly predicted 66.4% of the observations in the test/unseen dataset. While not great, it's not exactly bad either. The AUC is .723 which means this model is better than random guessing quite a bit; .723 is also somewhere between .5 and 1 indicating an OK but not amazing performance. The LASSO model also classfies 67.9% of the defaulted loans correctly with the TPR and 65% of the non-defaulted loans correctly with the TNR and doesn't have a specific class that it's really good/bad at classifying. However, that doesn't mean that it could be better at classifying both 0 and 1 as 67.9% and 65% aren't exactly great numbers and still quite far from 1.