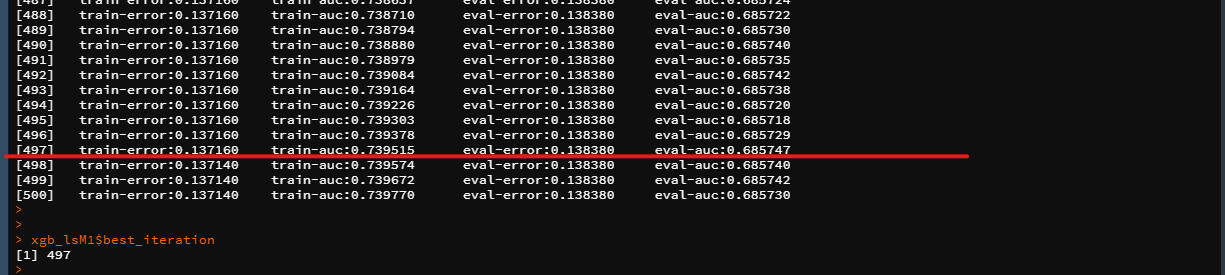
Assignment 2

# Models for investment decisions in LendingClub loans

1. (a) Develop boosted tree models (using either gbm or xgBoost) to predict loan\_status. Experiment with different parameters using a grid of parameter values. Use cross-validation. Explain the rationale for your experimentation. How does performance vary with parameters, and which parameter setting you use for the 'best' model? Model performance should be evaluated through use of same set of criteria as for the earlier models - confusion matrix based, ROC analyses and AUC, cost-based performance. Provide a table with comparative evaluation of all the best models from each method; show their ROC curves in a combined plot. Also provide profit-curves and 'best' profit' and associated cut-off. At this cut-off, what are the accuracy values for the different models?

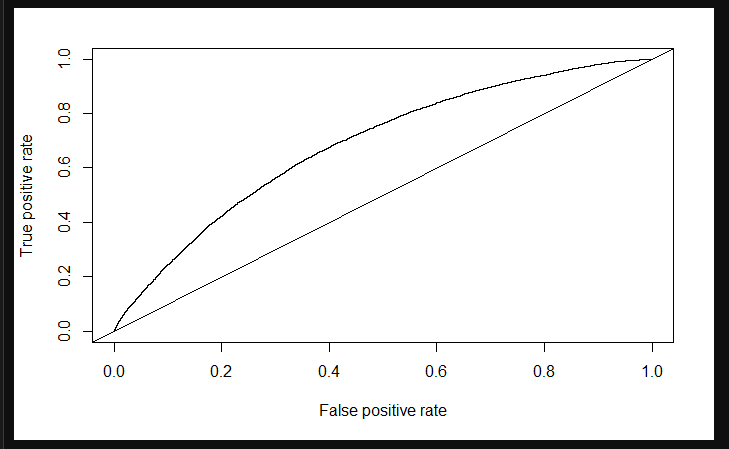
For a boosted tree model, we used the dataset obtained from the Assignment 1 Part A, wherein we did a univariate analysis and found out variables that impact the loan status. Now based on these variables we developed a boosted tree model. We used XGBoost library in R.

We created training data set and a test data set with a split of 0.5. Then we setup the prediction dataset vs actual dataset to evaluate the outcomes. In the same process we used various variables like max-depth, eta, objective, eval\_metric, nrounds, early\_stopping\_rounds. Once configured we trained the model on the training dataset and then find the “best” model on the basis of result parameters such as train-error, train-auc, eval\_error and eval\_auc.   


And we run prediction using the best iteration. We have also checked for the cross-validation as seen at the bottom of the below screenshot.

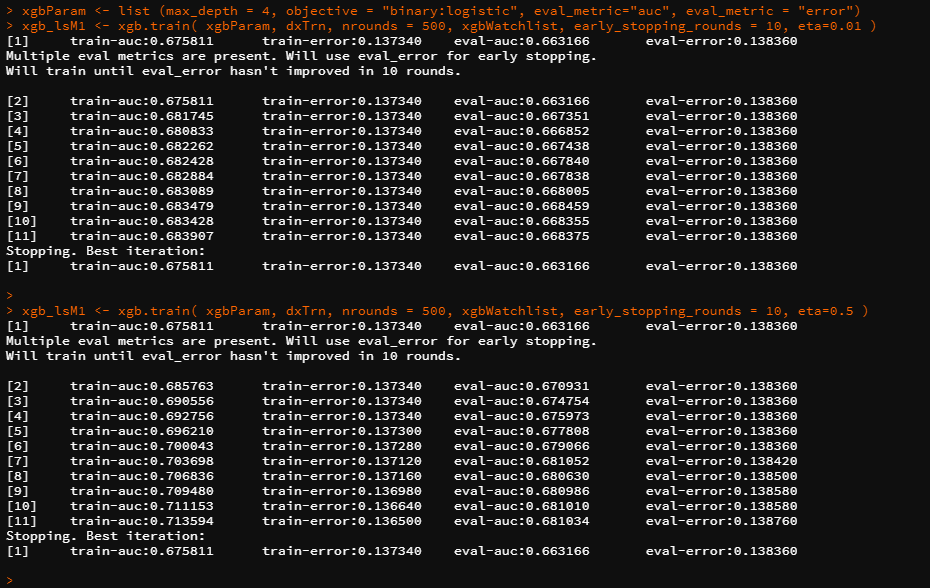


The AUC and ROC curve obtained for the model is

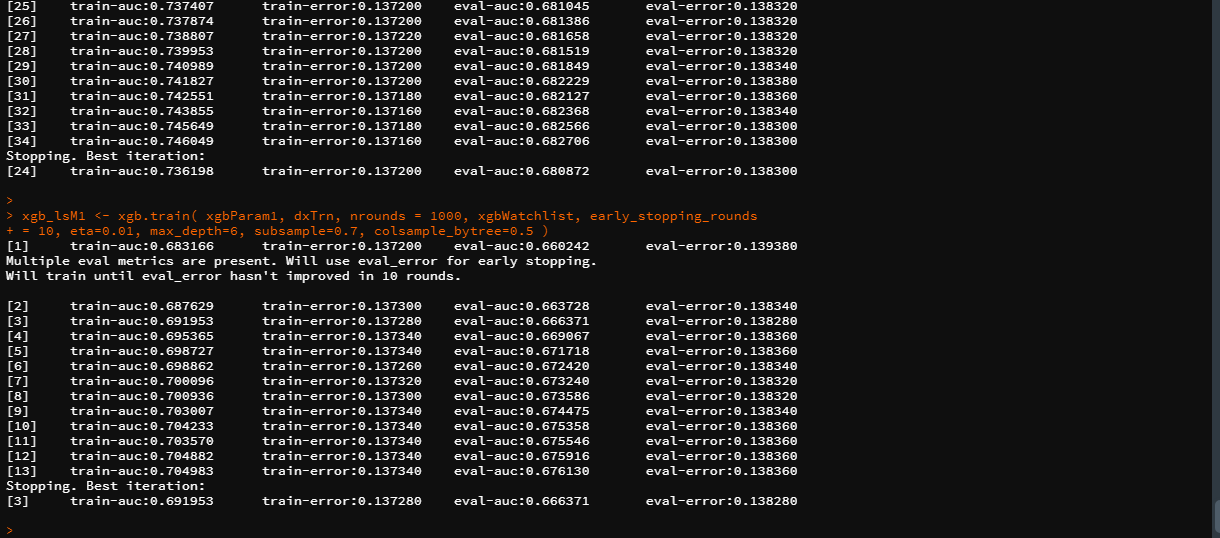
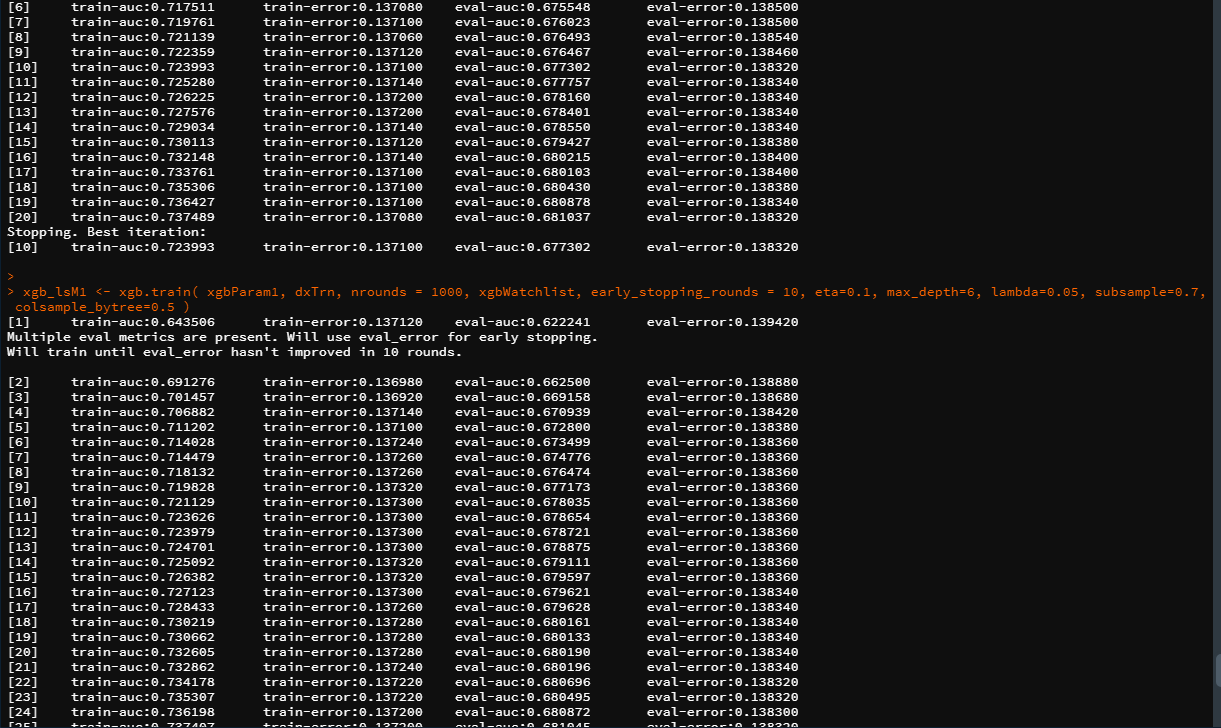
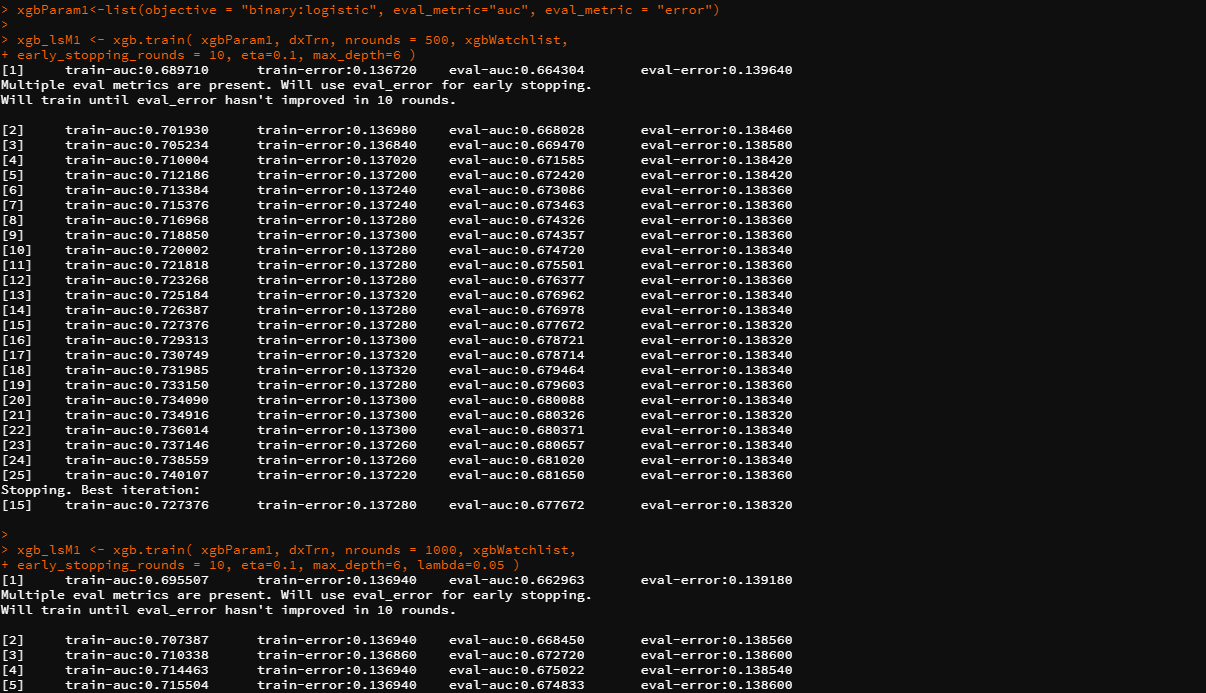


Exploring XGBoost models (different parameters)

1st -



2nd –



Analyzing above images, and plotting auc values for all the models we get that our best model has eval-auc: 0.68 when the parameters for the model are set to - **nrounds=1000, early\_stopping\_rounds=10, eta=0.1, max\_depth =6, lambda= 0.05, subsample= 0.7, colsample\_bytree= 0.5**

1. (a) Develop linear (glm) models to predict loan\_status. Experiment with different parameter values, and identify which gives ‘best’ performance. Use cross-validation. Describe how you determine ‘best’ performance. How do you handle variable selection? Experiment with Ridge and Lasso, and show how you vary these parameters, and what performance is observed.

(b) For the linear model, what is the loss function, and link function you use? (Write the expression for these, and briefly describe).

(c) Compare performance of models with that of random forests (from last assignment) and gradient boosted tree models.

(d) Examine which variables are found to be important by the best models from the different methods, and comment on similarities, difference. What do you conclude?

(e) In developing models above, do you find larger training samples to give better models? Do you find balancing the training data examples across classes to give better models?

For this, in the training dataset columns like loan\_status, actualTerm, annRet, actualReturn and total\_paymnt were omitted and then various regression models using ‘glmnet’ library with cross-validation were observed –

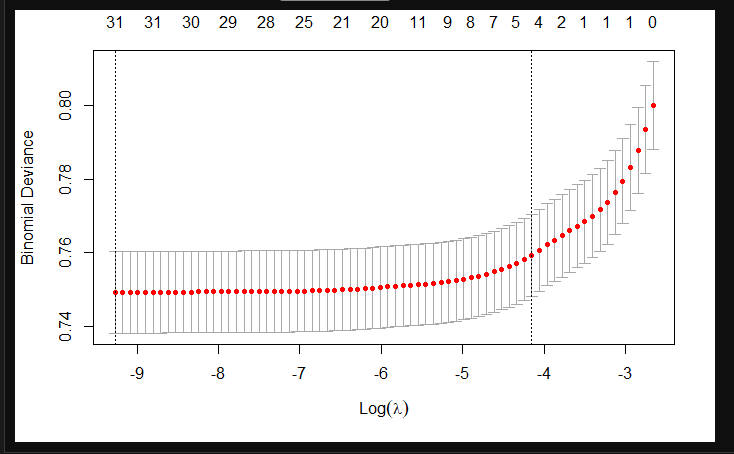
Lasso Regression (Least Absolute Shrinkage and Selection Operator)

Important lambda values - 

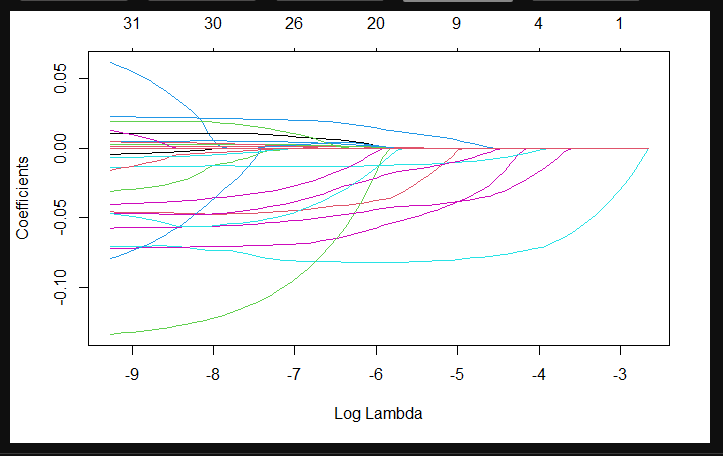
Table

Description automatically generated with low confidence

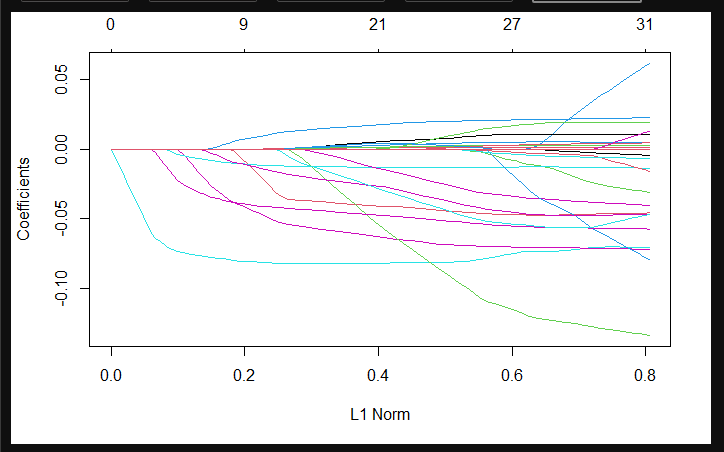
In the first image, we can see the best performing minimum lambda value and the best performing lambda value within 1 standard deviation. And the second image highlights the working of lasso regression and drops the attributes that meagerly contribute to the performance of the model. In our model the lasso regression retains only 13 attributes after training the model.

Plotting cross validation glm (generalized linear model) model -

From the above figure we can visualize how the binomial deviance varies with the change in log of lambda value.

Plotting the glm fit against the lambda values -

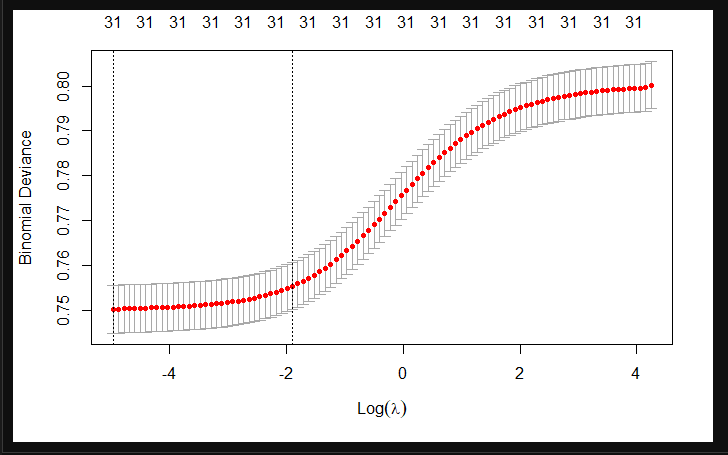
Plotting glm fit –

**

From the above image, we can infer the point where the lasso regression model drops the attribute since its contribution to the model drops low. To identify this point, we look at the kink in the curve and for that value of the kink, the variable is dropped corresponding to that lambda value. As the lambda value increases, the punishment factor increases and there will be a point where that attribute does not contribute enough to performance of the model, that is when Lasso regression will drop that attribute.

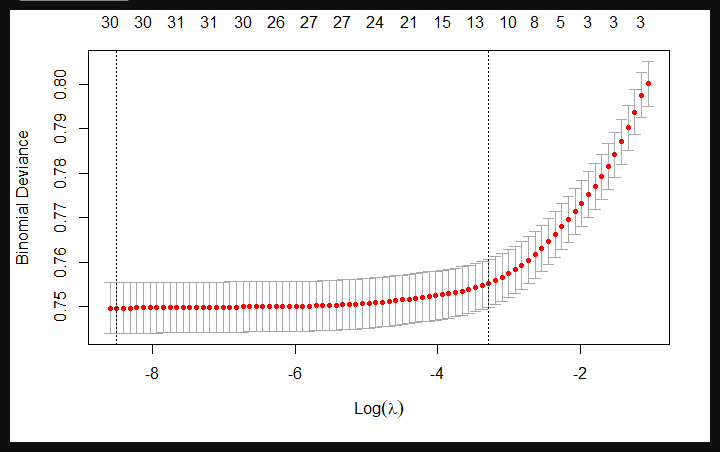
Ridge Regression

Ridge regression minimizes the weightage of the variable as the slope of the variable increases because of the punishment factor implied by the lambda value.

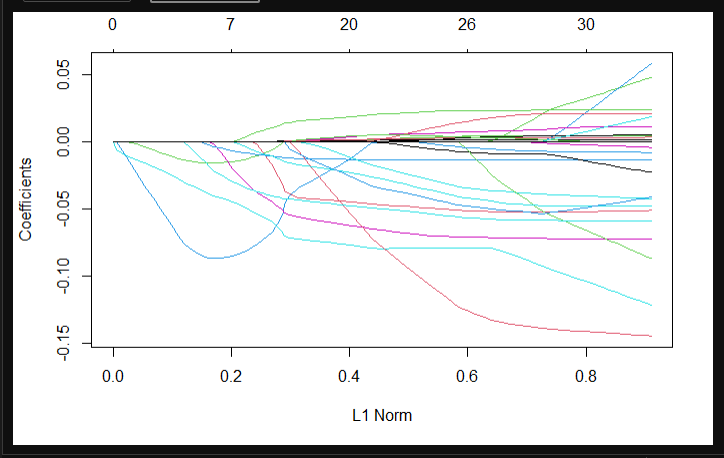


From the above figure we can visualize how the binomial deviance varies with the change in log of lambda value.

Elastic-net Regression

Plotting cross-validation glm model with alpha 0.2 -

From the above figure we can visualize how the binomial deviance varies with the change in log of lambda value.

Plotting glm model with alpha 0.5 (no cross-validation) –**

Elastic net regression - combines the strength of lasso regression and ridge regression. We have two lambda values, one from lasso regression and one from ridge regression. Elastic net regression groups and shrinks the parameters associated with the correlated variables and leaves them in the equation or removes them all at once.

The loss function used is –

ℓ(𝛽∣𝑦,𝑥) =∑𝑖 = 1𝑛𝑦𝑖 log𝑔−1(𝑥𝑇𝑖𝛽) + (1−𝑦𝑖) log(1−𝑔−1(𝑥𝑇𝑖𝛽))

*The link function –*

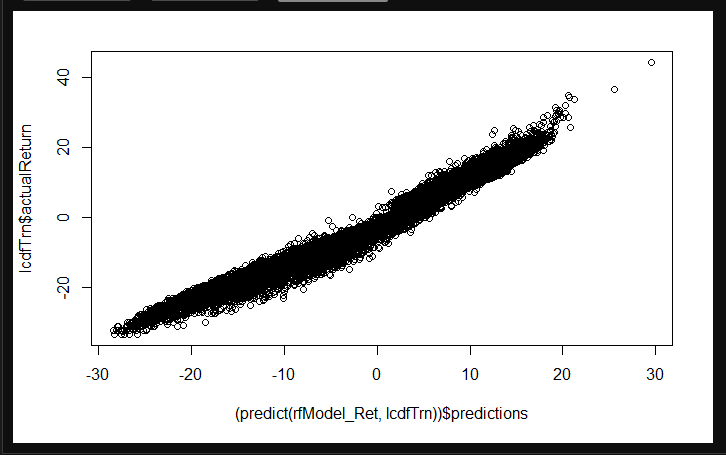
𝑔=logit

1. Develop models to identify loans which provide the best returns. Explain how you define returns? Does it include Lending Club’s service costs? Develop glm, rf, gbm/xgb models for this. Show how you systematically experiment with different 2 parameters to find the best models. Compare model performance – explain what performance criteria do you use, and why.

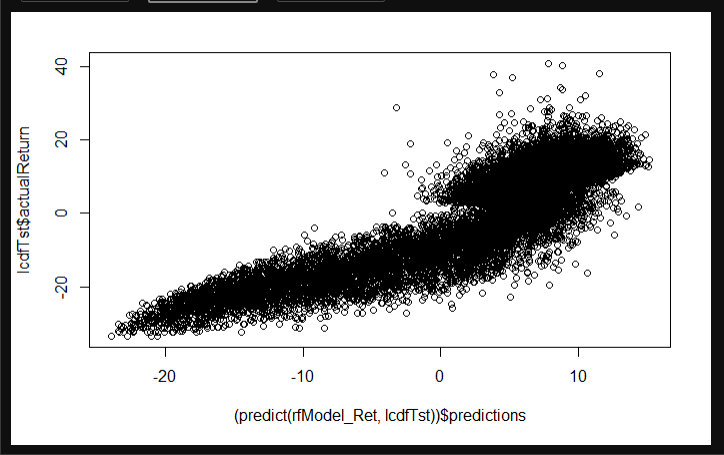
We have trained two models – Random Forest model and XGBoost for predicting the returns. Returns is the total amount returned to the investor at the end of the closing period of the entire loan cycle and this amount does not include the Lending Club’s service cost.

Random Forest Model

Root mean square error –**

Plotting random forest model predictions on training data against actual return –**

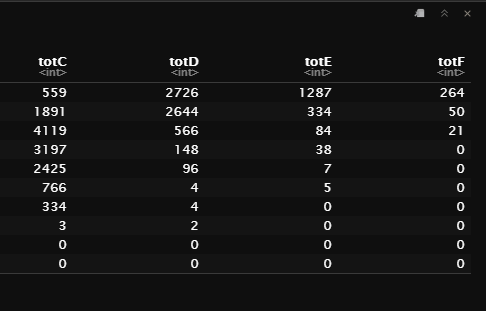
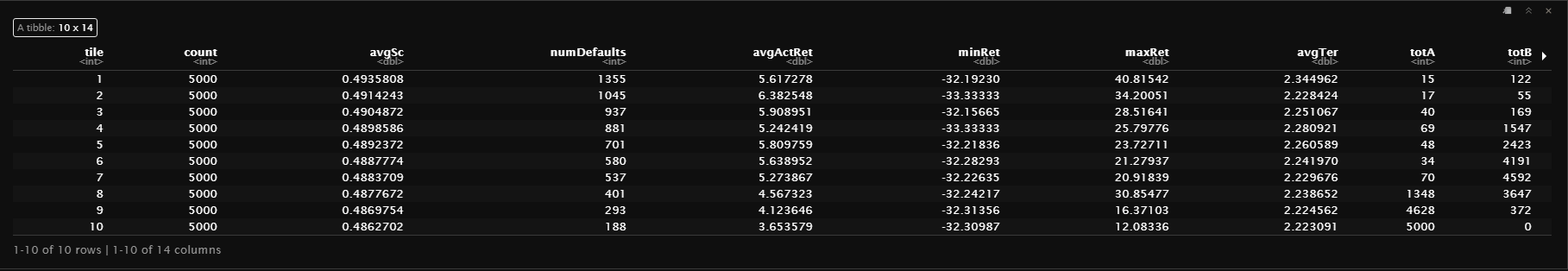
We can see in the above image that for the most part, our predicted value is almost close to the actual value.

Plotting random forest model predictions on test data against actual return – **

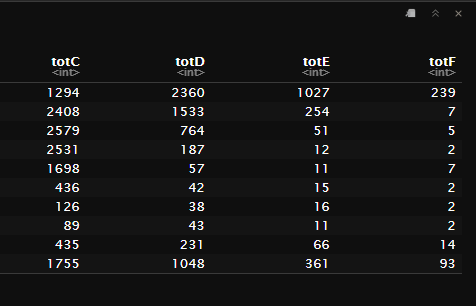
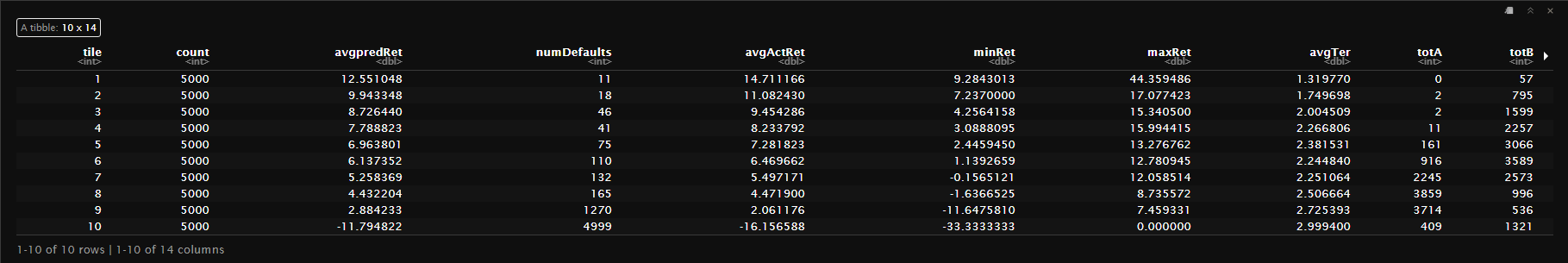
From the above image we can see that our model is working quite well on the testing data as well.

XGBoost Model compared to Random Forest Model using decile

XGBoost –



Random Forest –

**

From the above set of images, we can see that XGBoost is working better than Random Forest as it reflects the reality much better. We can see that there are only 11 defaults out of 5000 loans where majority of the 5000 loans fall under the risky category. Had this been the truth then all the investors could blindly invest their money in these types of loans and in turn enjoy high return (of around 14%) with next to no risk of the person defaulting.

1. Considering results from Questions 1 and 2 above – that is, considering the best model for predicting loan-status and that for predicting loan returns -- how would you select loans for investment? There can be multiple approaches for combining information from the two models - describe your approach, and show performance. How does performance here compare with use of single models?
2. As seen in data summaries and your work in the first assignment, higher grade loans are less likely to default, but also carry lower interest rates; many lower grad loans are fully paid, and these can yield higher returns. One approach may be to focus on lower grade loans (C and below), and try to identify those which are likely to be paid off. Develop models from the data on lower grade loans, and check if this can provide an effective investment approach – for this, you can use one of the methods (glm, rf, or gbm/xgb) which you find to give superior performance from earlier questions.

Can this provide a useful approach for investment? Compare performance with that in Question 4.

From our observations we found GLM to be the best performing model for predicting the loan\_status. Then we sorted the predicted output in the descending order of decile. If we look at the avgActRet ( i.e. the average actual return ) and the avgpredRet ( i.e. the average predicted return ), we find that our a slightly underperforming from predicting the actual value. Apart from that, if we look at the right side in the image below, we find that the loans with least defaults and good interest rates are majorly the ‘grade C’, ‘grade D’ and ‘grade E loans’.

Graphical user interface

Description automatically generated with medium confidence

Using this model for predicting which loans to invest in can prove to be extremely beneficial for both the investors ( since they would enjoy good returns with less risk of the other person defaulting ) and Lending Club ( since the borrower would not default and return the borrowed money from time to time, Lending Club would receive their transaction fee and also attract more investors ).