Credit Risk and Default Probability

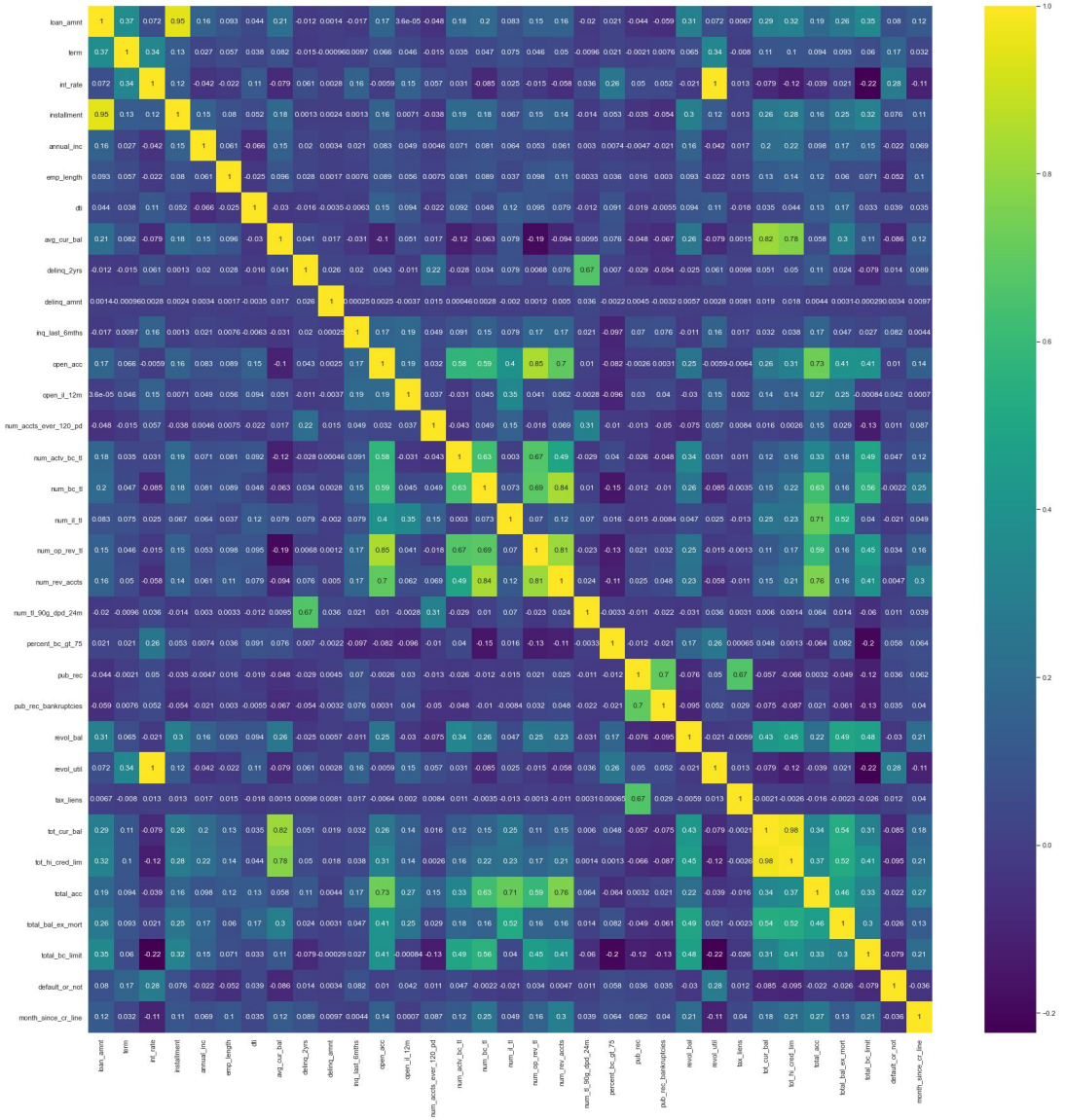
Big Data for Finance I

Analysis

- Variable of interest: Default or non-default
- How can a lender predict the borrowers default probability and adjust their decisions to the risk at hand?
  - Interest rate?
  - Loan amount?
  - Term structure?

Data Processing

- The dataset in question is very large and we have cleaned it ahead of hour analysis:
  - Dropping features with more than 50% of the data missing, and removed data little related to the default results
  - Addressed highly correlated predictors
- Transform categorical variables into dummies
  - Address state, Home Ownership, Purpose, Verification Status, Sub Grade, Employment Title, Loan Status
- Split our dataset into training- and test subset
  - Split 0.8/0.2 by random allocation



Logistic Regression

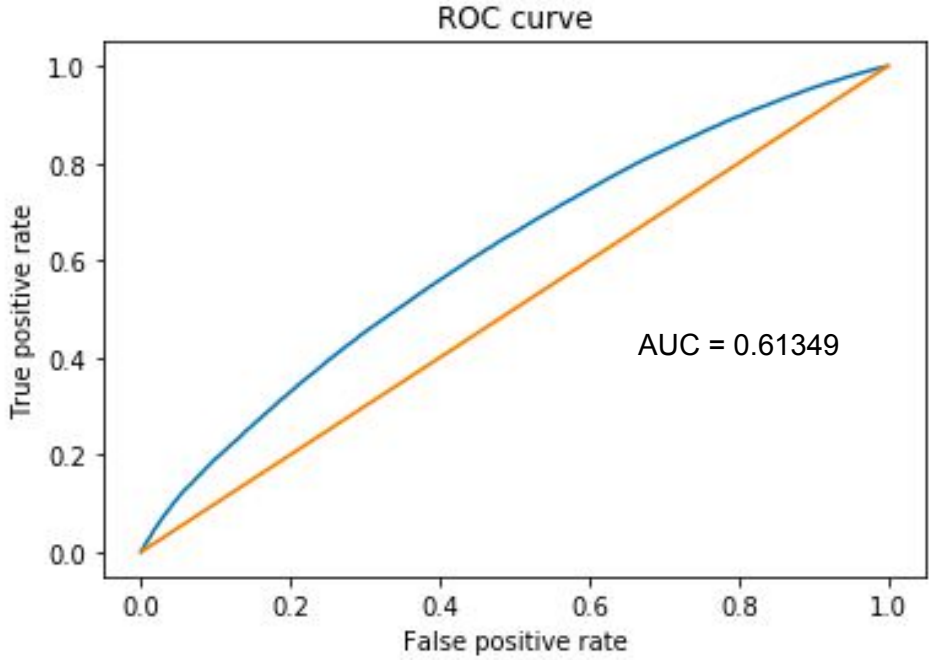
- We model the probability of default rather than the actual response directly.
- We run our logistic regression using all 139 variables.
  - This includes the dummy variables. As expected we can see small coefficients for many of these, especially the dummy variable purpose.

	explanatory_var	Coefficient	p_value
0	Intercept	-1.074687e-05	-
1	loan_amnt	4.581936e-05	0
2	term	-1.688865e-04	0
3	int_rate	4.527696e-05	0
4	installment	-5.724970e-04	0
5	annual_inc	-2.059007e-06	0
6	emp_length	-7.914974e-05	0
7	dti	-1.139441e-04	0
8	avg_cur_bal	-2.299647e-05	0
9	delinq_2yrs	-1.386987e-07	9.36935e-83
90	purpose:house	8.226447e-09	0.000344909
91	purpose:major_purchase	-2.648119e-07	0.0392279
92	purpose:medical	-4.774022e-08	0.406977
93	purpose:moving	-4.232730e-08	0.000161956
94	purpose:other	-8.204824e-07	0.0259119
95	purpose:renewable_energy	4.416485e-09	0.0499377
96	purpose:small_business	3.094579e-07	6.21646e-79
97	purpose:vacation	-1.220634e-07	1.03855e-05
98	verification_status:Not Verified	-7.909142e-06	0
99	verification_status:Source Verified	-4.641696e-06	0.202702
100	sub_grade:A1	-2.536475e-06	0

Logistic Regression

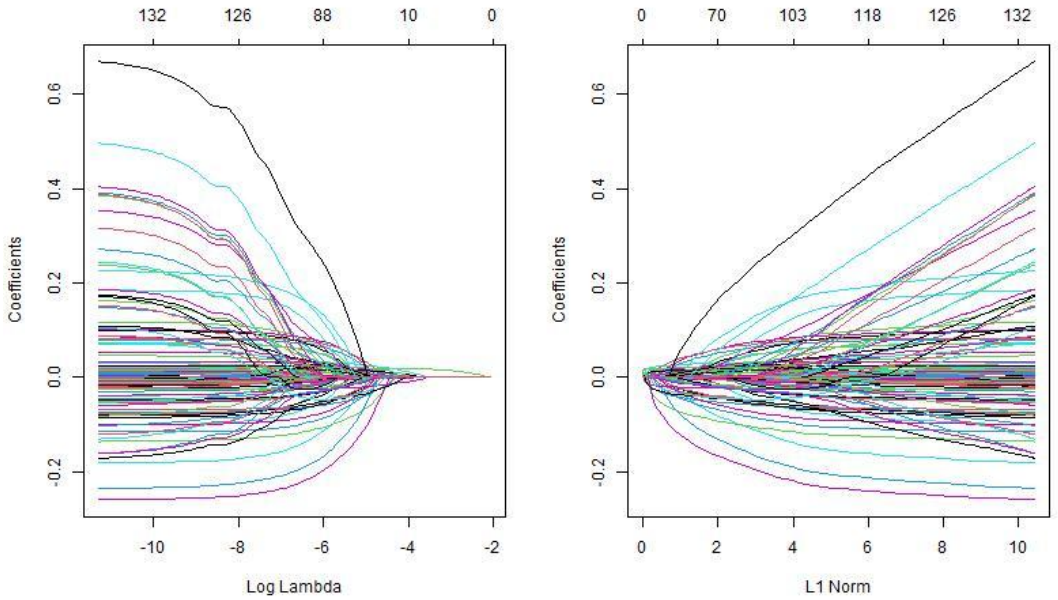
- We model the probability of default rather than the actual response directly.
- We run our logistic regression using all 139 variables.
  - This includes the dummy variables. As expected we can see small coefficients for many of these, especially the dummy variable purpose.
- We can see that a logistic regression has a degree of predictability from the ROC-curve. We will now try to improving predictability by looking at more advanced models.

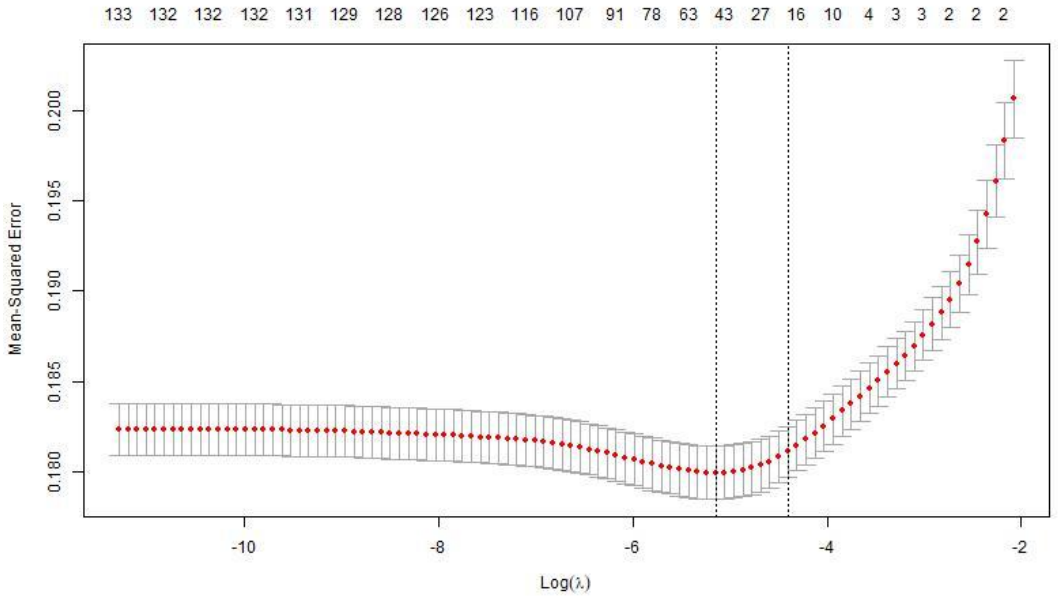
predict	0	1
actual		
0	0.697777	0.007175
1	0.287306	0.007741



Lasso

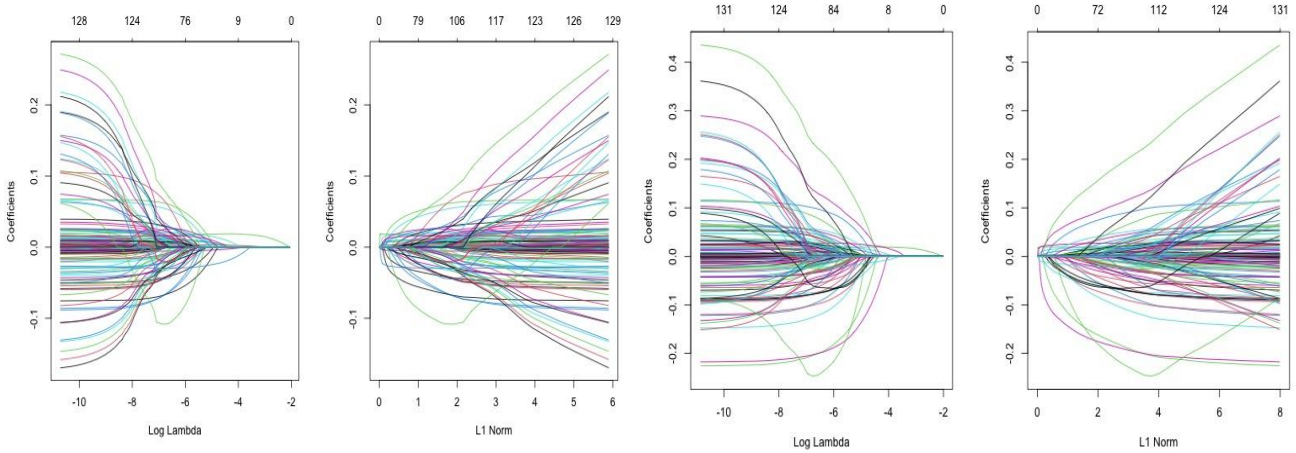
- Best subset selection computationally infeasible, and we therefore want to use a shrinkage method – the Lasso.
- In the total dataset, the Lasso decreases total number of predictors from 139 to 110.
- We get the most regularized lambda by doing cross validation on training sample.
  - The lambda selected is the most regularized one within 1 standard error from the lowest MSE lambda.
- When we run the Lasso on a subset of the original dataset we see that the shrinkage is more protruding.

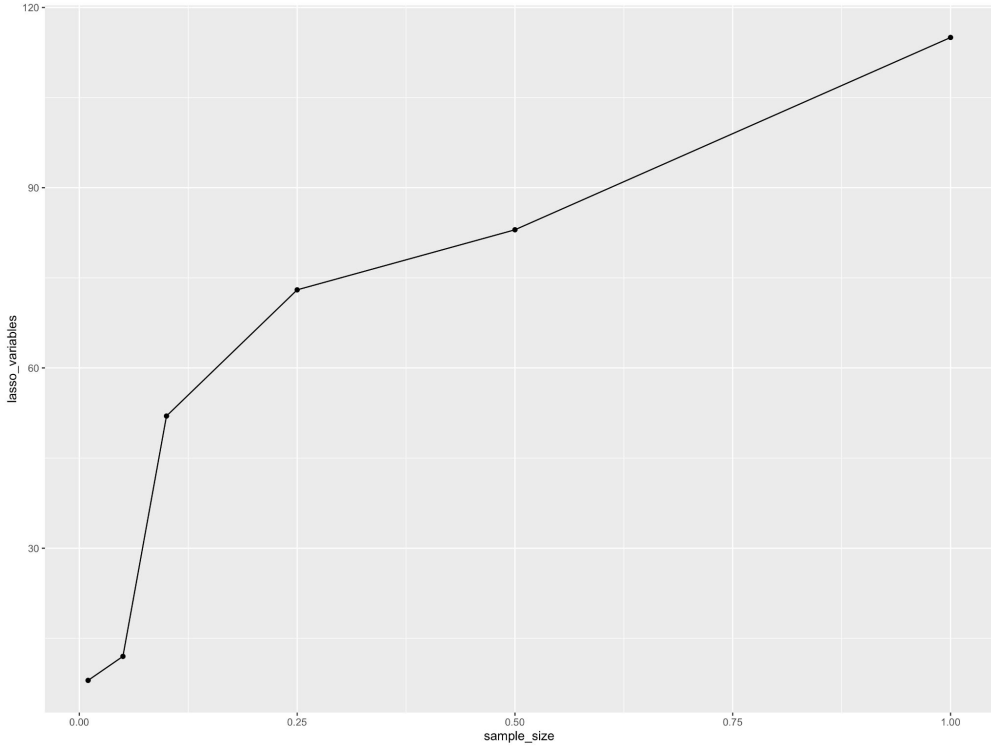




Lasso 2

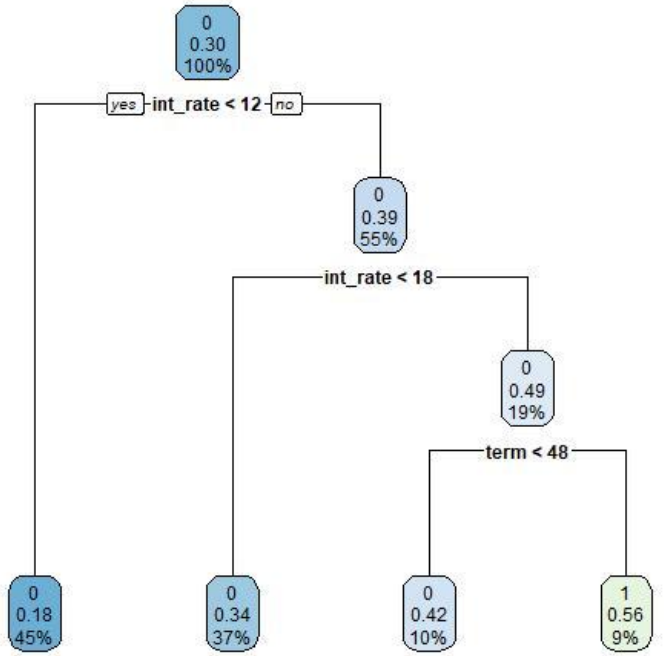
- In the subset that we drew for computational reasons, we see that the lasso has increased shrinking effect on parameter number.
- The number of predictors selected by the lasso increases in general with the sample size. When the sample size is small one is more likely to overfit the training data, and as we can see the lasso eliminates more predictors the smaller our subset is.

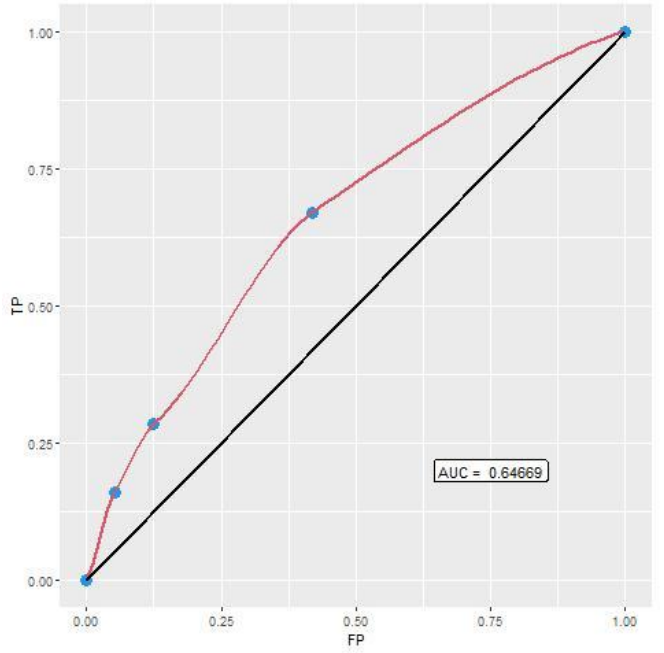




Classification Tree

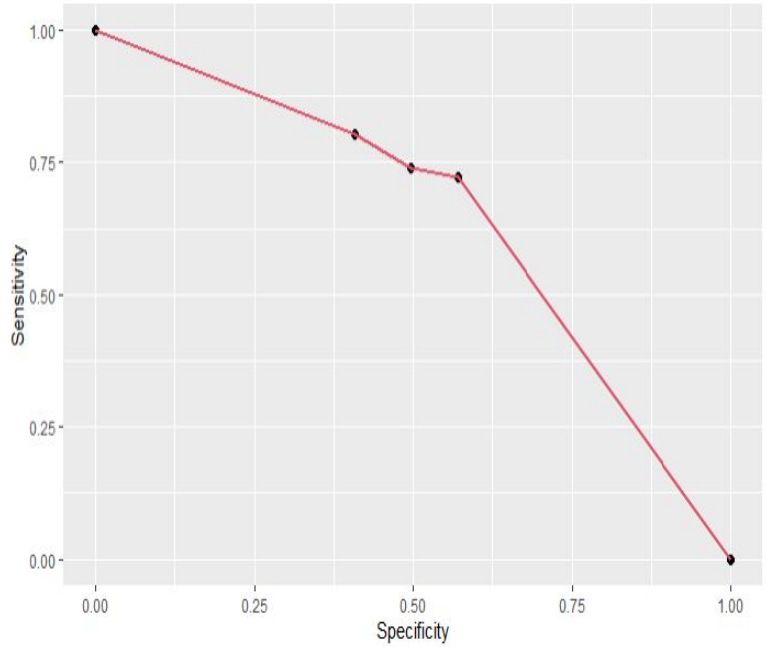
- From the logistic regression, we saw that the interest rate on the loan had the highest coefficient among the predictors. In the single classification tree we see that the classifier reoccurs. Also, the second classifier is term which suggest that number of payments above 48 indicates default, while below indicates non-default.
    - Interestingly, these are both factors decided by the lender.
- When running the tree model on the dataset adjusted for lasso-regressors we get identical results.
    - Some predictors seem to be highly important and therefore dominate the split in the nodes in both cases.





Classification Tree

- From the logistic regression, we saw that the interest rate on the loan had the highest coefficient among the predictors. In the single classification tree we see that the classifier reoccurs. Also, the second classifier is term which suggest that number of payments above 48 indicates default, while below indicates non-default.
  - Interestingly, these are both factors decided by the lender.
- When running the tree model on the dataset adjusted for lasso-regressors we get identical results.
  - Some predictors seem to be highly important and therefore dominate the split in the nodes in both cases.



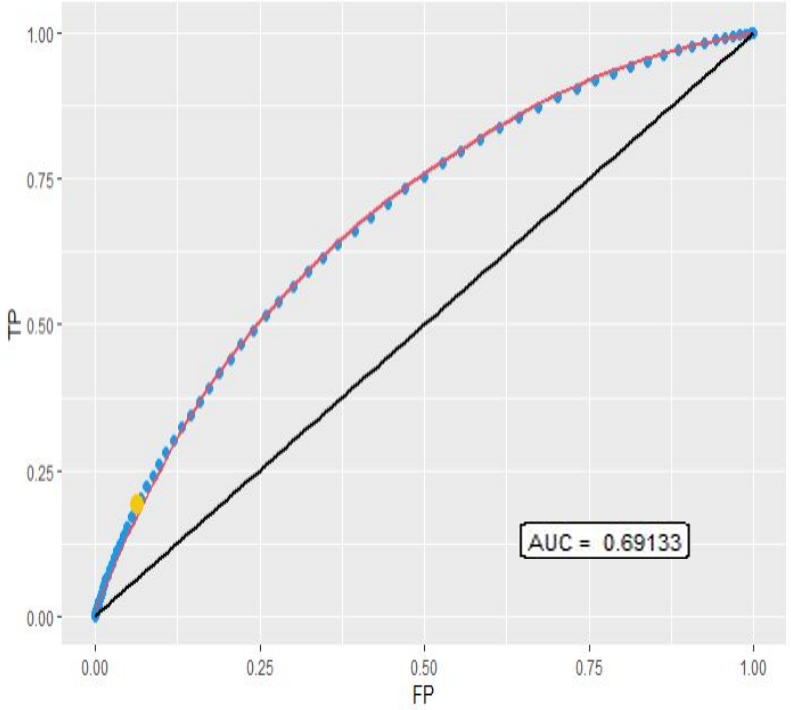

```
           0          1
0 0.66272696 0.03614046
1 0.25295613 0.04817644
```

```
> accuracy_is
[1] 0.7175794
> accuracy_oos
[1] 0.7109034
```

Random Forest

- When we expand on the framework used in the classification tree model and use random forest, we can see some change. In the table we have presented the variable importance.
  - Interest rate is still ranked on top, but the term is not considered as important any longer. Instead we see utilization of credit limit, account balance and loan amount.
- The AUC has increased from 0.65 to 0.69
- When we use lasso-regressors, the AUC slightly increased to 0.696
- 12 variables per iteration and 100 trees.
  - Optimization of hyperparameters was restricted due to computational limitations.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
int_rate	18.38317	10.99998	22.29592	172.21324
revol_util	16.43631	12.15982	20.99068	182.54982
tot_cur_bal	13.13991	-8.71955	13.01552	109.78619
loan_amnt	11.64533	-1.95627	12.4977	105.53672
avg_cur_bal	11.19703	-4.09172	11.74284	115.51975
tot_hi_cred_lim	11.73458	-6.87522	11.5185	113.70733



	0	1
0	0.65361957	0.04514271
1	0.24361318	0.05762454