Habits

Contents

Task 1: Predicion modelling

Part 1: Inspecting, cleaning and imputing the data set

In this section, we will train and rund three different prediction models based on a data set containing financial records of a list of companies. In total, there are 24 variables in the data set and xxx observations. Prior to training our prediction models, we load and examine the data set.

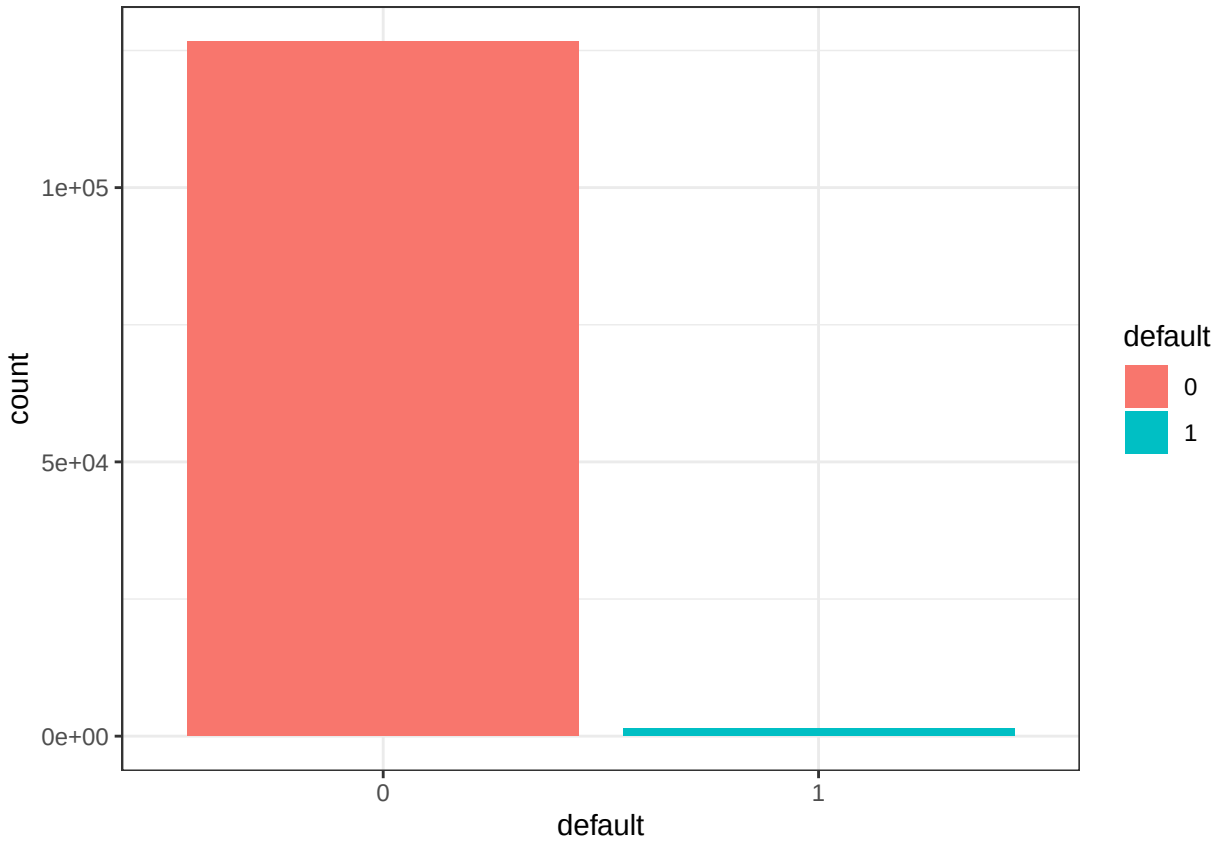
The first thing we notice is that some variables are classified as numeric when they should be factor variables, and vice versa. We reclassify these variables.

Table 1:

Statistic	Min	Max
default	0	1
profit_margin	−10,000,000,000,000,000,000.000	25,553.330
gross_operating_inc_perc	0.000	1.000
operating_margin	−10,000,000,000,000,000,000.000	586.175
EBITDA_margin	−10,000,000,000,000,000,000.000	586.171
interest_coverage_ratio	−10,000,000,000,000,000,000.000	10,000,000,000,000,000,000.000
cost_of_debt	−8,131.722	165,944.600
interest_bearing_debt	−10,000,000,000,000,000,000.000	7,523.000
revenue_stability	−10,000,000,000,000,000,000.000	4,473.242
equity_ratio	−10,000,000,000,000,000,000.000	168.999
equity_ratio_stability	−10,000,000,000,000,000,000.000	168.000
liquidity_ratio_1	−2,473.828	10,000,000,000,000,000,000.000
liquidity_ratio_2	−2,473.790	10,000,000,000,000,000,000.000
liquidity_ratio_3	−2,473.809	10,000,000,000,000,000,000.000
equity	−771,200	182,466,000
total_assets	−9,685	544,267,000
revenue	−2,883,588	588,422,000
age_of_company	2	22
unpaid_debt_collection	−5.000	10,000,000,000,000,000,000.000
paid_debt_collection	−5.500	10,000,000,000,000,000,000.000
adverse_audit_opinion	0	6
industry	0	11
amount_unpaid_debt	−121,226	10,000,000,000,000,000,000
payment_reminders	0	3

Table xx displays summary statistics of our data. As shown in the table, the value xx is present in many of the observations. Most likely, these values represent measurement errors or missing variables, and we therefore choose to replace these values with NAs. After completing this step, there are still some quite extreme outliers remaining in the data. Our assessment is that these values do not represent real observations, as they deviate from typical values for the financial key figures represented in the variables in question. Applying a threshold of 2,5 percent at each end of the variables' distribution, we replace values exceeding this threshold with NAs.

We also notice that the variable default, which will be the dependent variable in our prediction models, is imbalanced. That is, observations of defaulting companies are underrepresented, which will make it difficult to predict defaults. The underrepresentation of defaults is illustrated in figure **??**. we will handle this imbalance when we implement the prediction models in section xx.


1e+05
5e+04
0e+00
count
0
1
default
default
0
1

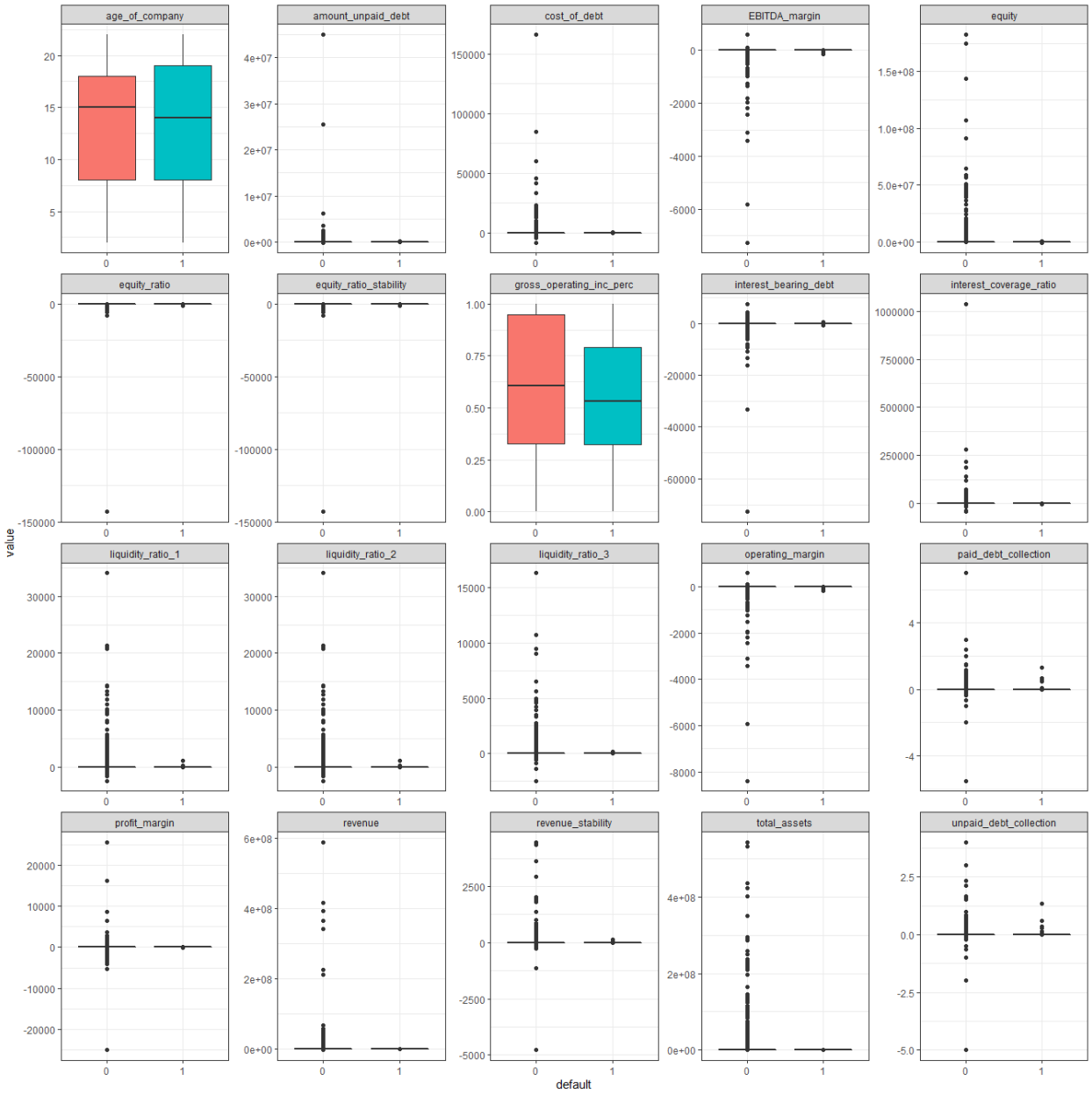
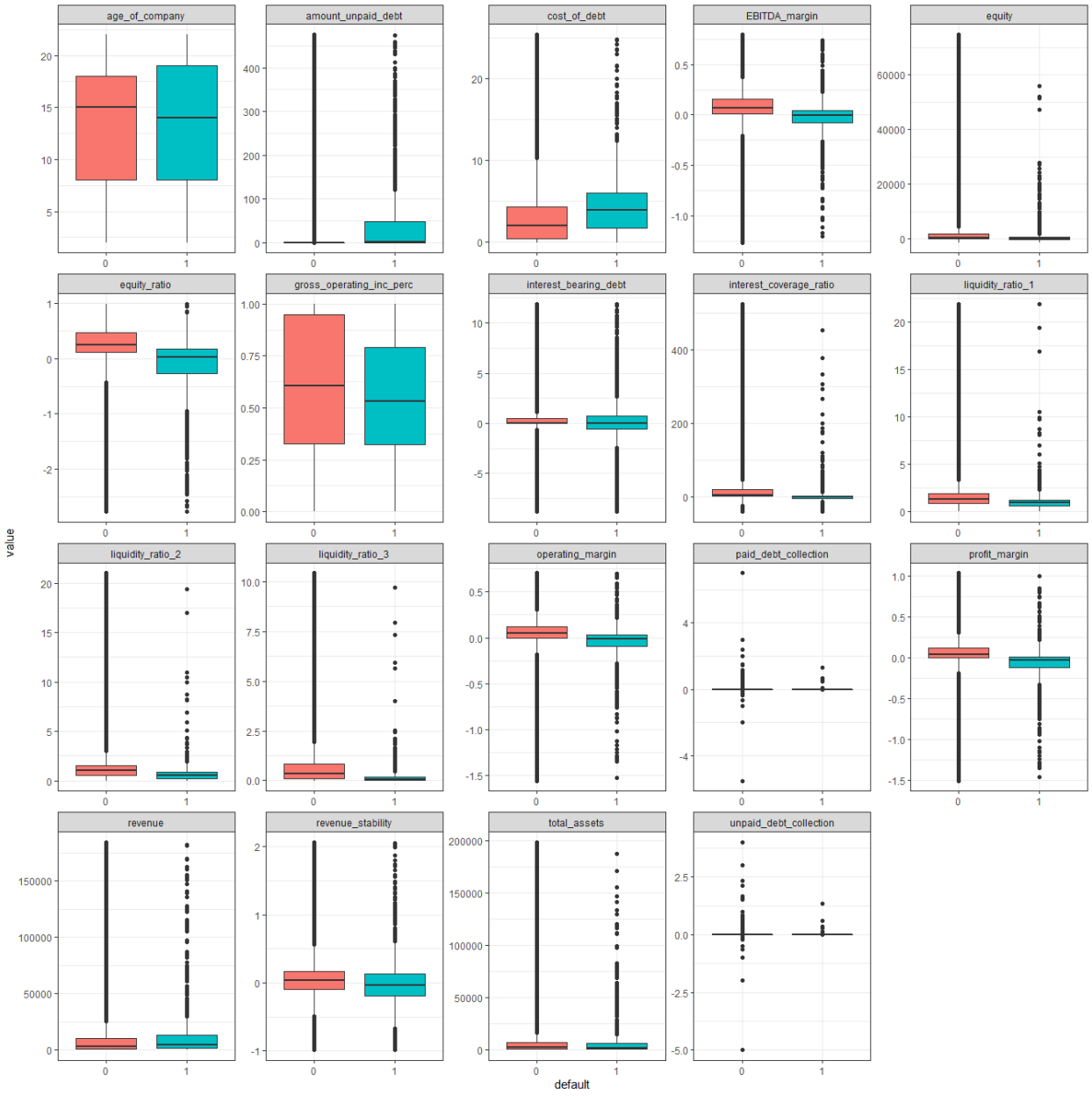
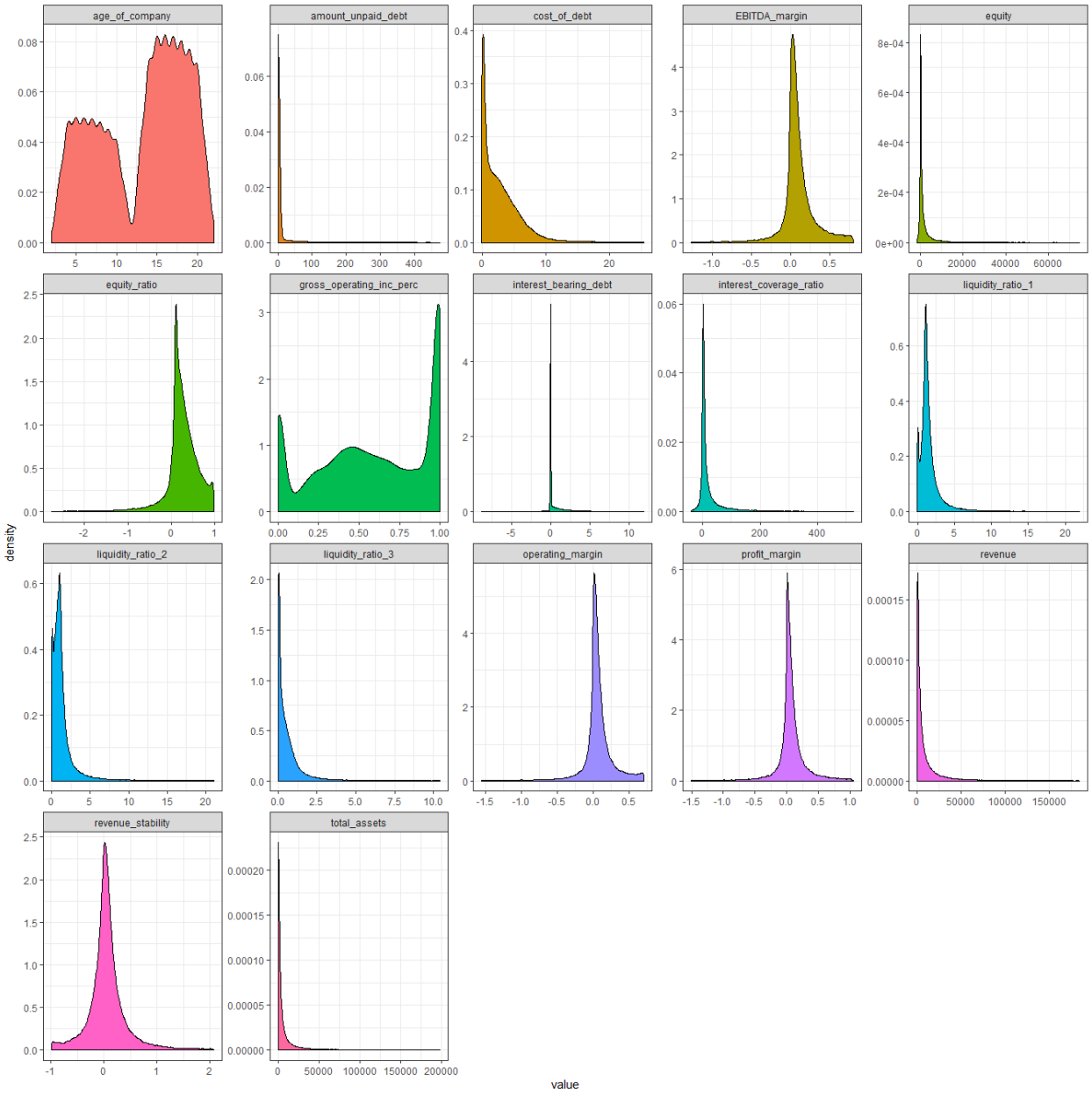


Figure **??** and **??** contain box- and density plots of the numeric variables in our data set after cleaning:


age_of_company
amount_unpaid_debt
cost_of_debt
EBITDA_margin
equity
equity_ratio
gross_operating_inc_perc
interest_bearing_debt
interest_coverage_ratio
liquidity_ratio_1
liquidity_ratio_2
liquidity_ratio_3
operating_margin
paid_debt_collection
profit_margin
revenue
revenue_stability
total_assets
unpaid_debt_collection
value
default



After cleaning the data set, we are left with xx missing values. As these account for a substantial fraction of the total number of observations in the data set, we choose to impute these using the MICE package. An illustration and explanation of the imputation added to the appendix.

##Prediction modelling

Preparations

XXX Blalala the data is split into a training set and a test set.

Prior to training our prediction models, we run the data set through a function which detects whether any of the variables correlate. It turns out that the variables *total assets*, *revenue*, *industry* and *paid debt collection* correlate with one or more of the other variables. To avoid multicollinearity, these variables are removed from the data set.

Logistic Regression Model

Our first prediction model is a logistic regression model. - Using cross validation: ten fold, - Oversampling method (smote) implemented here - Variable selection: choosing varibles, and why. Not relative values etc etc

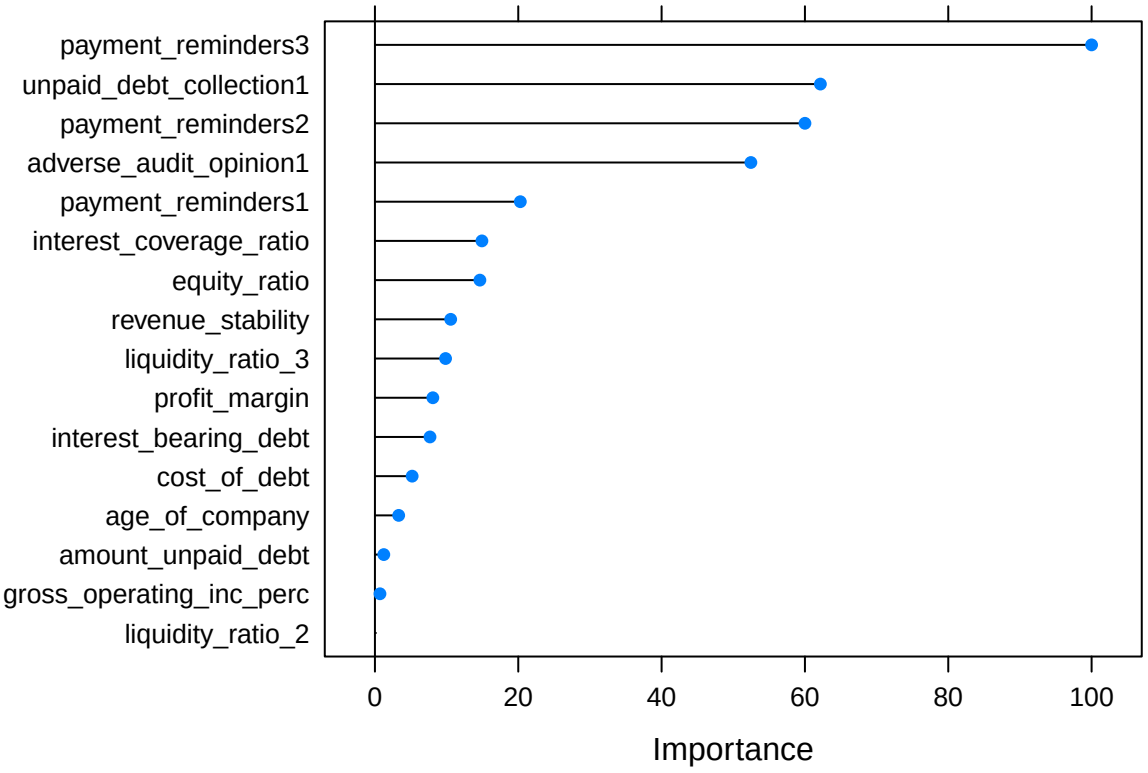
Summary statistics from the logistic regression model are presented in table xxxx. - comments: which variables are significant, and do they make economically sense

Figure xxx gives and overview of the variable importance.
- Comments:

A confusion matrix from the glm model is presented in table xx. - Comments:

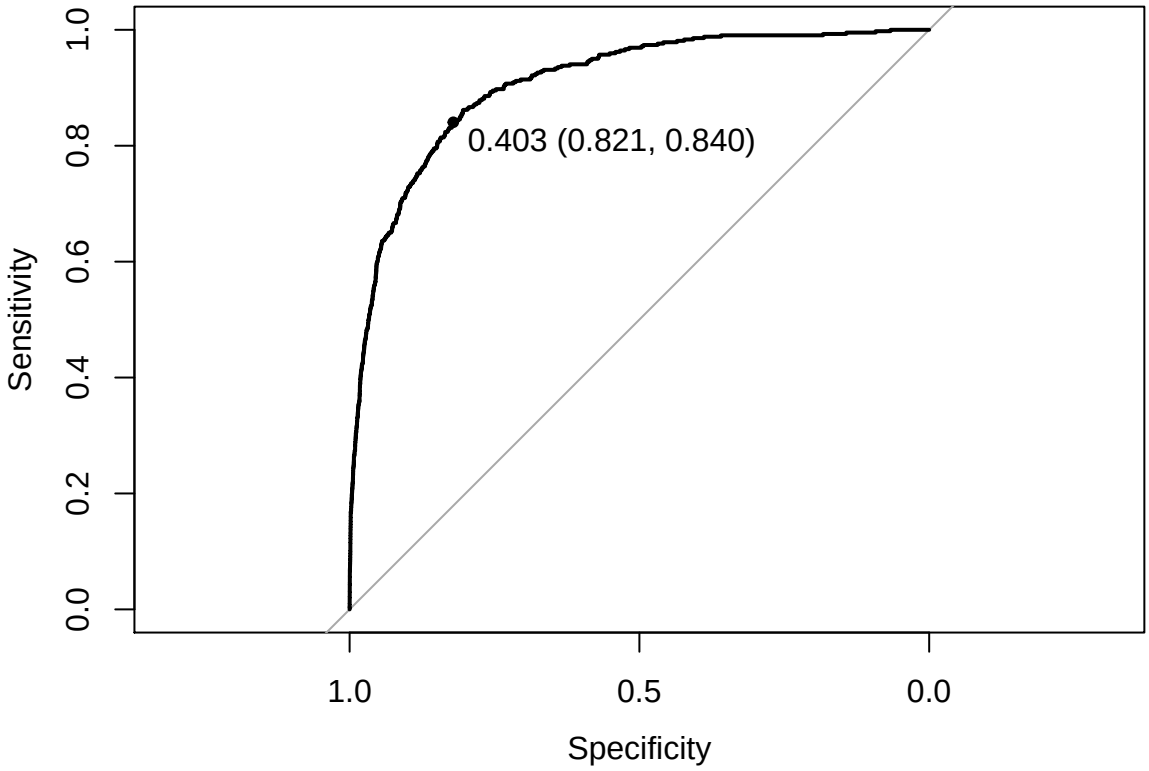
Figure xx shows the AUC for the glm model. - Comments:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.4972	0.1450	-17.22	0.0000
profit_margin	-0.7085	0.1700	-4.17	0.0000
gross_operating_inc_perc	-0.2626	0.1183	-2.22	0.0264
interest_coverage_ratio	-0.0071	0.0012	-5.97	0.0000
cost_of_debt	0.0375	0.0110	3.41	0.0007
interest_bearing_debt	-0.0602	0.0148	-4.06	0.0000
revenue_stability	-0.4814	0.0998	-4.82	0.0000
equity_ratio	-0.5196	0.0881	-5.90	0.0000
liquidity_ratio_2	-0.0874	0.0430	-2.03	0.0419
liquidity_ratio_3	-0.4105	0.0886	-4.63	0.0000
age_of_company	0.0194	0.0067	2.91	0.0036
unpaid_debt_collection1	1.4323	0.0777	18.44	0.0000
adverse_audit_opinion1	1.2293	0.0774	15.88	0.0000
amount_unpaid_debt	0.0012	0.0005	2.36	0.0182
payment_reminders1	0.6615	0.0895	7.39	0.0000
payment_reminders2	1.7179	0.0961	17.87	0.0000
payment_reminders3	3.1267	0.1100	28.42	0.0000



	0	1
0	32849	93
1	5151	327

	Values
Sensitivity	0.86
Specificity	0.78
Pos Pred Value	1.00
Neg Pred Value	0.06
Precision	1.00
Recall	0.86
F1	0.93
Prevalence	0.99
Detection Rate	0.85
Detection Prevalence	0.86
Balanced Accuracy	0.82

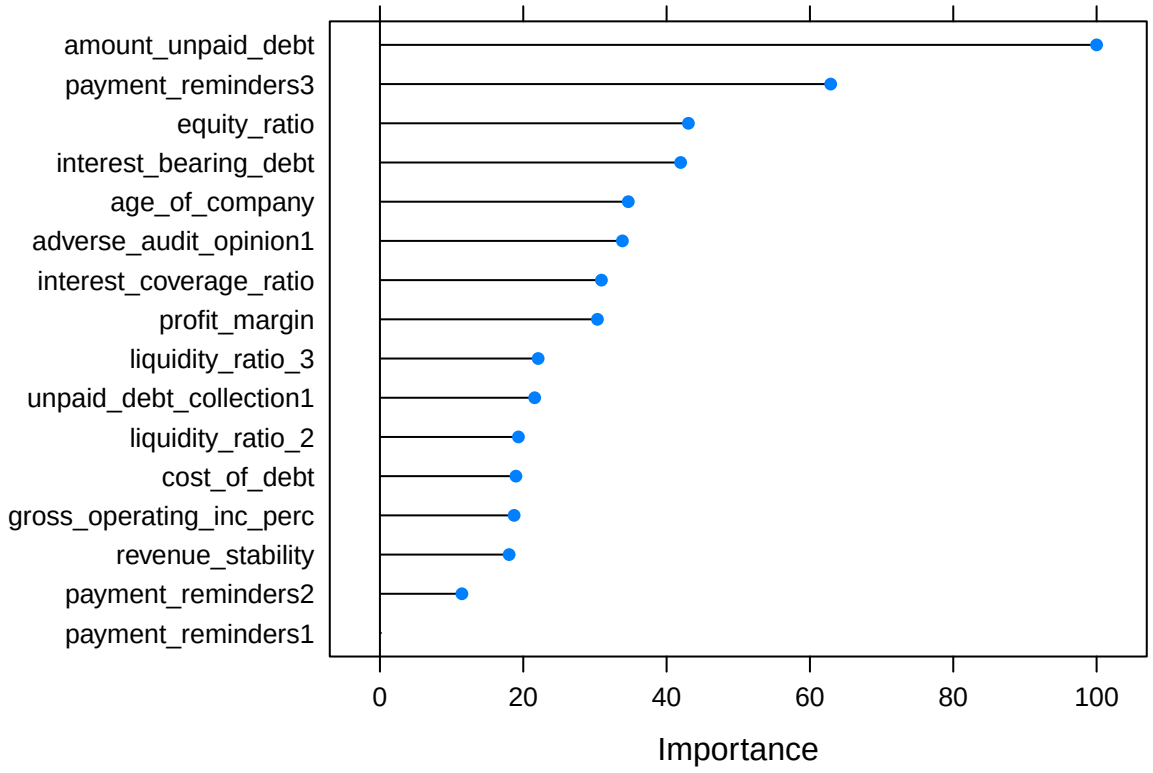


threshold accuracy 0.4028848 0.8213170

Random forest

Our second predicion model is a random forest model. Comments: - We have implemented cross validation - Fine tuning? - Oversampling technique is also implemented here.

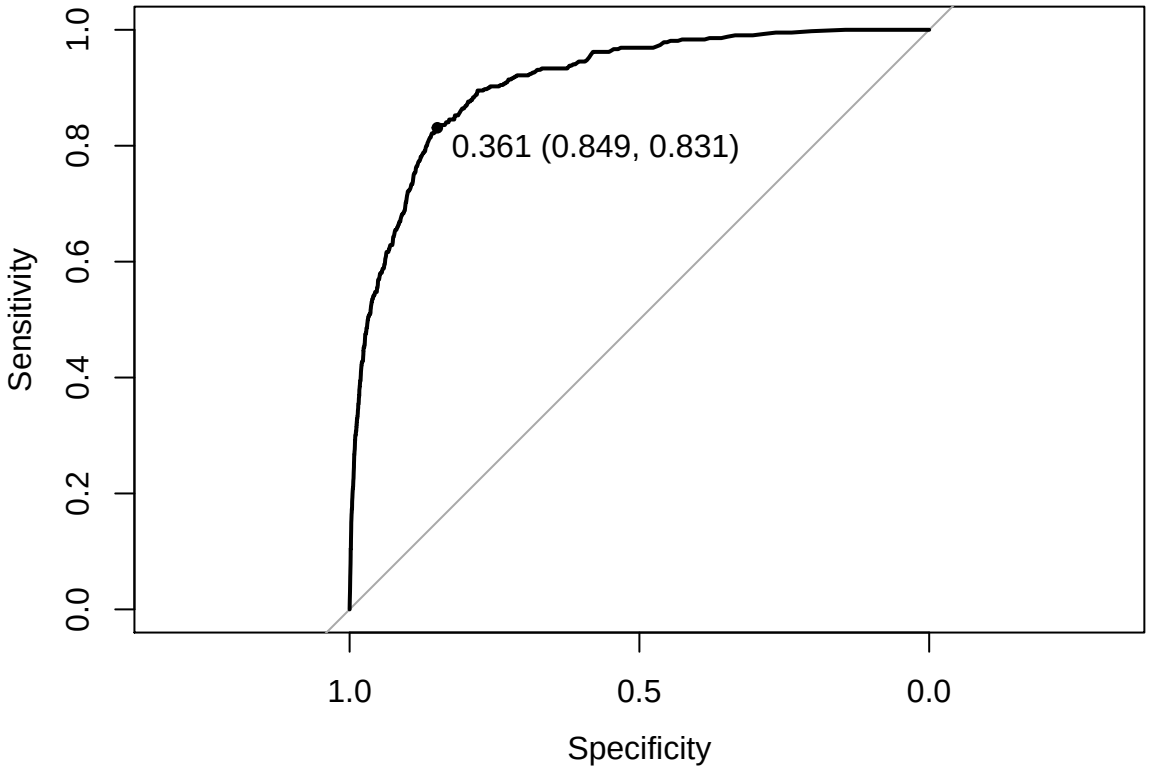
The model's confusion matrics is presented in figure xx, and the ROC curve is presented in figure xx. - Discuss performance



	0	1
0	34738	139
1	3262	281

	Values
Sensitivity	0.91
Specificity	0.67
Pos Pred Value	1.00
Neg Pred Value	0.08
Precision	1.00
Recall	0.91
F1	0.95
Prevalence	0.99
Detection Rate	0.90
Detection Prevalence	0.91
Balanced Accuracy	0.79

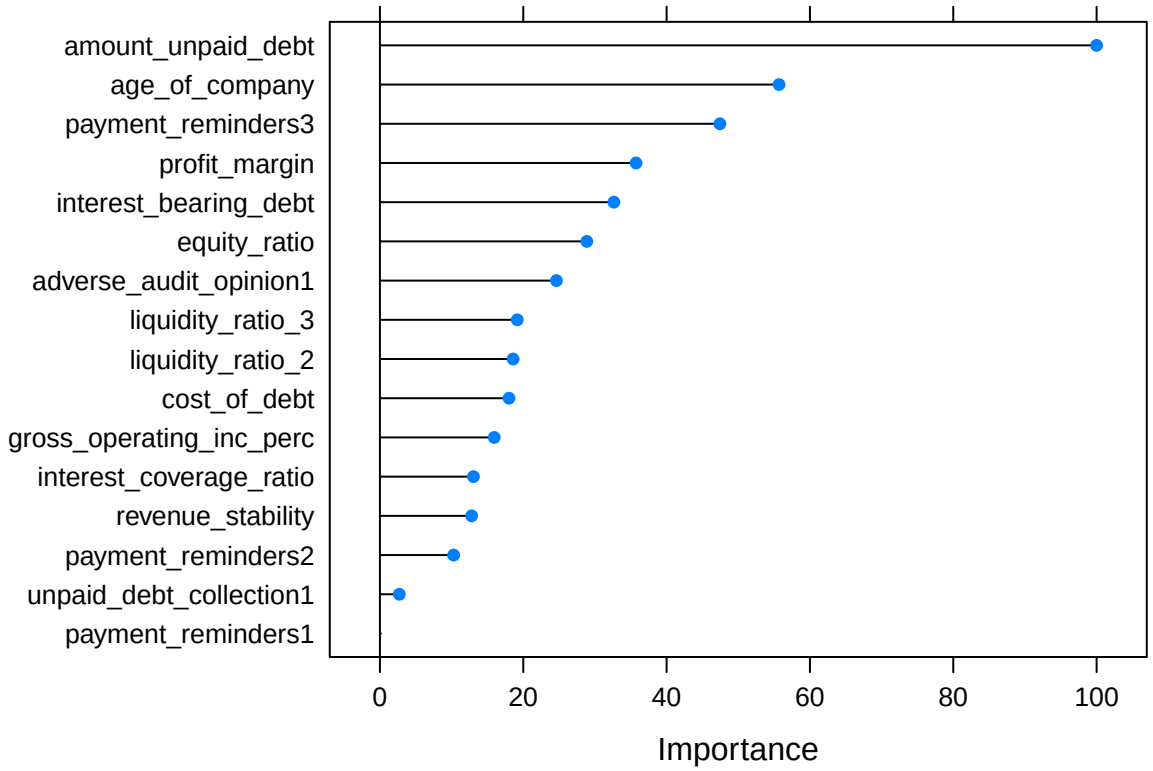
Comments: About the confustion matrix How well the model performs relative to glm Accuracy, specificity and sensitivity etc.



threshold accuracy 0.3610000 0.8484643

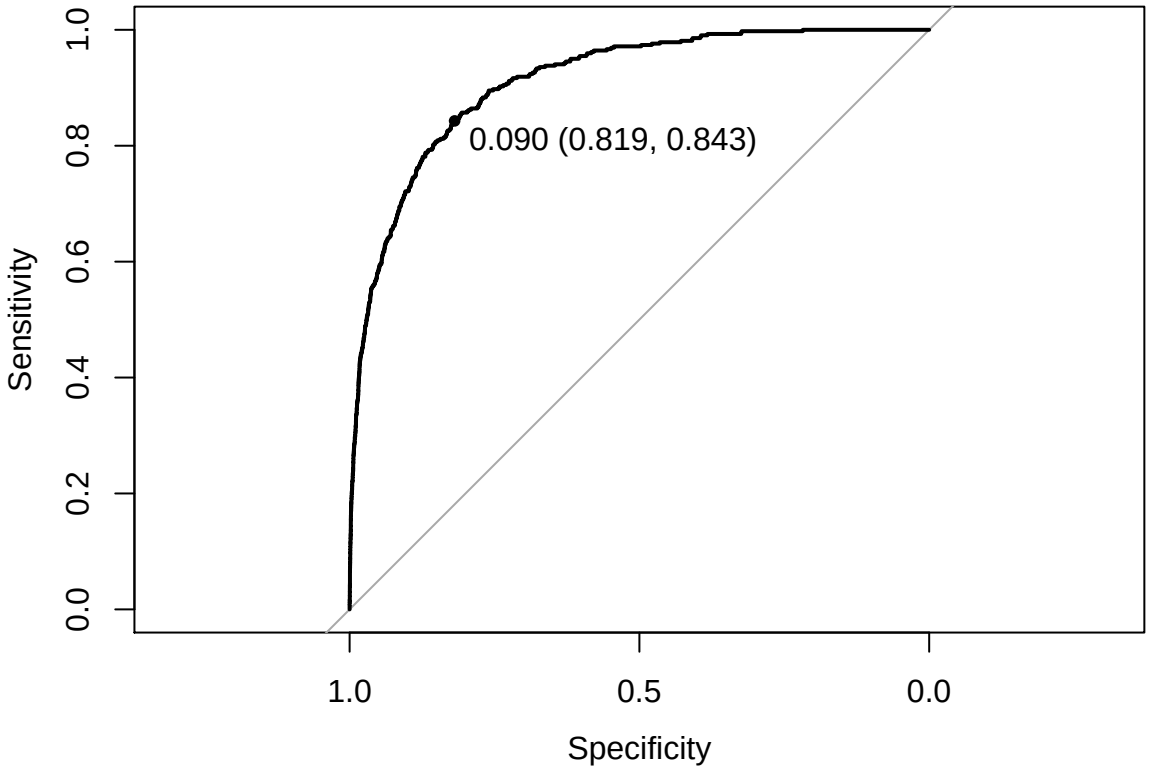
•

Show ROC here # -

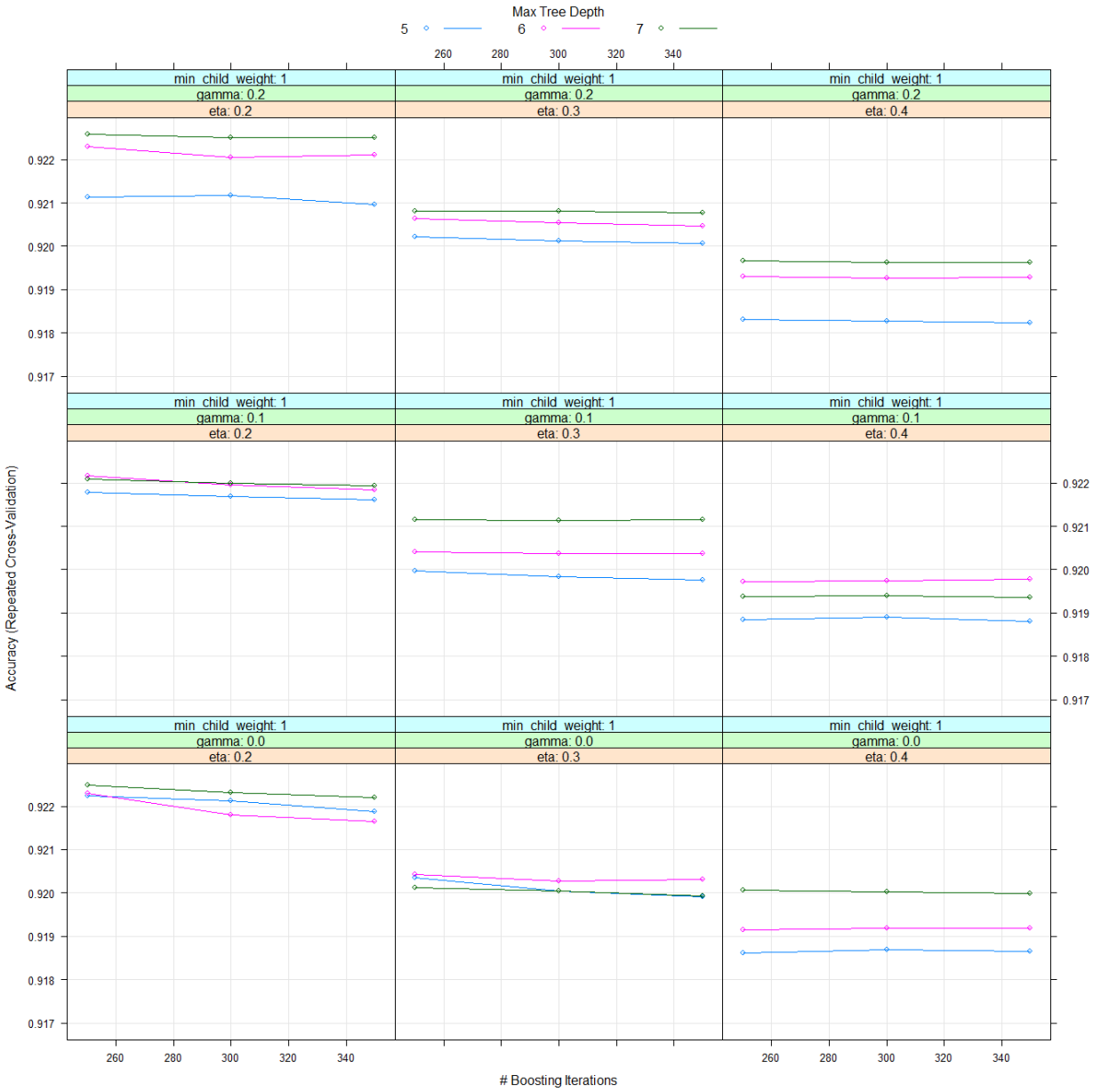


	0	1
0	35184	144
1	2816	276

	Values
Sensitivity	0.93
Specificity	0.66
Pos Pred Value	1.00
Neg Pred Value	0.09
Precision	1.00
Recall	0.93
F1	0.96
Prevalence	0.99
Detection Rate	0.92
Detection Prevalence	0.92
Balanced Accuracy	0.79



threshold accuracy 0.08979452 0.81907860



Accuracy and Kappa These are the default metrics used to evaluate algorithms on binary and multi-class classification datasets in caret.

Accuracy is the percentage of correctly classifies instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes (e.g. you need to go deeper with a confusion matrix). Learn more about Accuracy here.

Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes (e.g. 70-30 split for classes 0 and 1 and you can achieve 70% accuracy by predicting all instances are for class 0). Learn more about Kappa here.

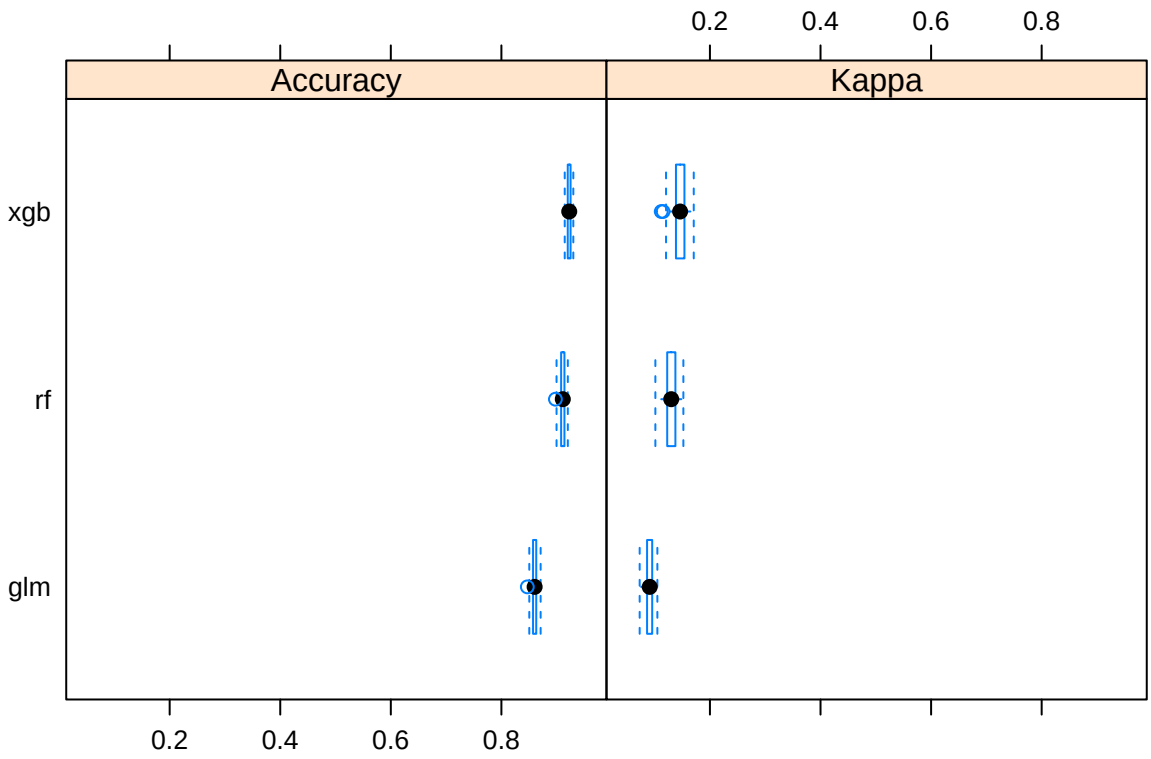


Table 2: Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
glm	0.847	0.857	0.860	0.860	0.863	0.871	0
rf	0.898	0.908	0.911	0.911	0.914	0.920	0
xgb	0.915	0.920	0.923	0.923	0.925	0.930	0

Table 3: Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
glm	0.073	0.086	0.091	0.091	0.096	0.105	0
rf	0.101	0.123	0.130	0.130	0.137	0.152	0
xgb	0.112	0.139	0.146	0.146	0.154	0.171	0

#Appendix