Report on HGC prediction

Péter ENDES-NAGY

Introduction

This is a report on *predicting high growth companies (HGC)*, prepared for a(n imaginary) Client active in the banking sector that aims to introduce a new product targeting potential HGCs.

The Client is interested in estimating the share of HGCs and their classification as they expect future revenue generated by HGCs, but the product includes consulting services coming with a high cost on the bank's side.

4 models (2 logit, a logit with LASSO and a random-forest model) were built for predicting HGC status, based on a cross-section of Bisnode data for the year 2012.

Data

Data is from the `bisnode-firms` dataset, collected, maintained, and cleaned by Bisnode, a major European business information company. It covers the entire population of companies between 2005 and 2016 for a few industries in a given European country.

The report is focusing on a cross-section of companies in 2012, other parts of the dataset is used for calculating high-growth for subsequent years as target variable and some historical growth variables as predictors.
The dataset is available (here).

Target variable and sample design

Details on the data preparation and sample design is available in the Technical Documentation (here).

Our *target variable* is the binary HGC status, if a company experiences high-growth in the 2012-2014 period. High-growth is defined as average yearly growth in sales above 20% in a 2 years period.

In the general HGC literature, HGC is usually defined as average yearly 20% growth in terms of sales or employment, in a 3 years period, or at least 20% yearly growth in 3 consecutive years. My definition is less strict as the Client would like to target companies with good potential and not only the top of the segment that experience growth in a 3 year period. I chose 2 years instead of 1, because it decreases the chance of identifying an ad-hoc, one time growth as high-growth.

Furthermore, the HGC literature recommends excluding companies with a very low base, as it is too easy to grow from there. Usually the micro enterprise segment is dropped (under 10 employees) as a good practice.

I had to keep in mind that I need a reasonable sample size, so I was less strict with the filtering on the bottom end. The good practice wouldn't have worked as the number of employees variable is very unreliable and dropping observations under 10 employees would have resulted in a sample of less than 300 observations.

For choosing the bottom threshold, 10-50-100.000 EUR cuts were inspected in terms of sales, they would have resulted in roughly 15-9-6 thousand observations. Even the 100.000 EUR threshold would be relatively low in light of the good practice (10k EUR sales per employees, not enough even to cover salaries), so the decision was based on potential sample size: 50.000 EUR sales. It is also in line with the Client's expectation that doesn't want companies that are too small (10.000), the 50.000 EUR minimum sales value is acceptable

for the business case. Furthermore the Client targets the SME segment, so observations above 10m EUR sales were dropped as well.

To sum it up, the cross-sample of companies for the year of 2012 was filtered for companies between 50k and 10m EUR sales, resulting in sample size of 9116 observations.

Feature engineering

For details on the feature engineering steps, please consult the Technical Documentation.

The potential *predictor variables* available in the dataset can be categorized into 4 main groups: size, management, financials and other characteristics (see detailed table in the appendix).

The financials were recalculated into ratios; balance sheet elements divided by total assets, variables from the profit and loss account divided by sales, which transformation deals with extreme values already. The remaining extreme values were winsorized, in most cases the upper and bottom 1% were replaced and flagged, or treated in other meaningful way, for instance the value of a given ratio is usually between -1 and 1.

Furthermore, quadratic and cubic forms were calculated to capture non-linearity based on `loess` plots, missing values were replaced with means for some variables, for others these observations were dropped. Interaction terms were also created for the LASSO model based on field knowledge.

As a result, our sample shrunk further to 8203 observations, the prevelance of HGC's in the sample is 24.82%.

Model building and prediction

4 Models were built: 2 logit models, a logit model with LASSO and a Random Forest model.
The variable types included in the models:
- X1: sales (log), financial ratios, firm, historical growth + quadratic forms + flags - X2: sales (log), financial ratios + their quadratic forms + flags , firm, human capital, historical growth + quadratic forms + flags, data quality - LASSO: X2 and interactions - RF: untransformed financials, historical growth (not winsorised), hr, firm, human capital and data quality

For predictive purpose, I took random 25% of the data for holdout, the remaining 75% is my working set. The work set was divided into test and training sets as I'm doing 5-fold cross validation. Cross-validated average RMSE and AUC was calculated both on the working, both of the holdout sets. The following table presents the predictive performance of each model, as well as the number of variables and coefficients:

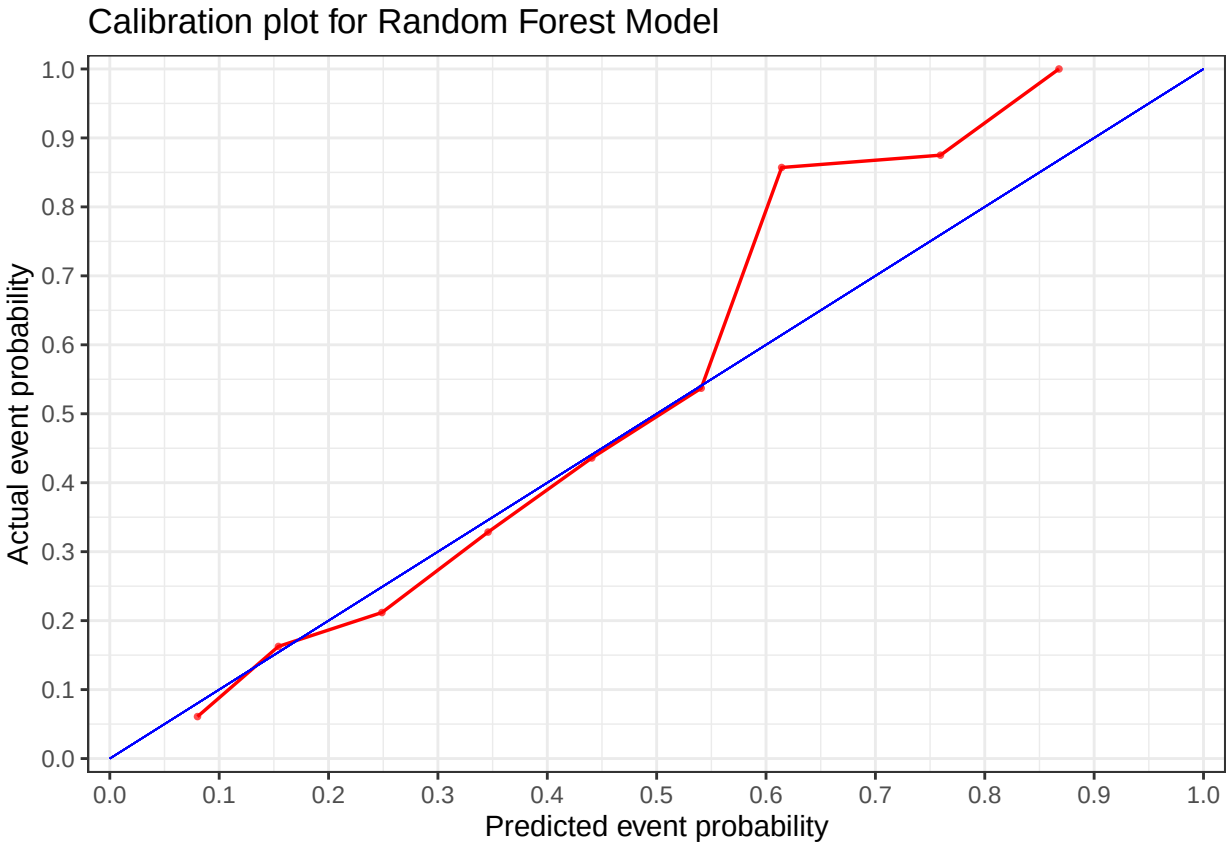
Table 1: Model performace on the holdout (HO) and working set (cross-validated: CV)

	N vars	N coeffs	CV RMSE	CV AUC	HO RMSE	HO AUC
X1	22	35	0.4209	0.6501	0.4169	0.6571
X2	29	75	0.4201	0.6514	0.4155	0.6563
LASSO	29	25	0.4201	0.5117	0.4182	0.6562
RF	29	43	0.4185	0.6540	0.4144	0.6704

Random Forest slightly dominates the others both in terms of RMSE (lowest) and AUC (highest), both in the working and hold out samples. Surprisingly, each model performed better on the hold out set, with lower RMSE and higher AUC, even though the sample size is smaller in the hold out set, so the models did not overfit.

As the models' performances are this close to each other, we could decide on X1 as it has the least number of variables and comes with interpretable coefficients, but Random Forest showed off its predictive power on the hold-out sample: the largest relative increase/decrease of AUC/RMSE compared to the working sample.

The following plot shows the calibration curve for the chosen Ranfom Forest model. Predicted probabilities were grouped into 10 bins, and the average predicted probability was plotted for each bin with the proportion of y = 1 cases for those observations. The calibration curve shows that the model is relatively well calibrated (X2 is the best calibrated, see Annex), but becomes biased for high propabilities.



Classification

The Client intends to introduce the new product for the HGC segment. If a firm turns out to be a HGC, the bank expects 5.000 EUR extra revenue. If the firm isn't a HGC, no extra revenue but the cost of consultation is lost, 1.000 EUR.
Therefore the false negative case means the loss of extra revenue, but consultation cost saved, so 4.000 EUR. A false positive case is the loss of 1.000 EUR consultation cost.

For each model, expected loss and best threshold was calculated for each fold, then averaged and saved into a list. For the best threshold, the Youden index was maximized. The calculations were carried out on the hold out set as well (using the best threshold calculated on the Folds).

The following tables show the Model performance on the working and hold out sets:

The optimal threshold is very similar across models, except for LASSO. The expected loss (both on the working, both on the hold out sets) is the lowest for the Random Forest model. The best threshold for the RF model means, that our model classifies firms as HGC if the model predicts an at least 20% probability. If external validity is high, then the bank can expect to make a loss of 638-647 EUR per classification. The difference isn't too much compared to other models that performed similarly (expect for LASSO). Like previously, picking RF isn't the very strong preference.

Table 2: Model performance on the working set (cross-validated)

	N vars	N coeffs	RMSE	AUC	Thresholds	Expected loss
X1	22	35	0.4209	0.6501	0.1875	0.6475
X2	29	75	0.4201	0.6514	0.1721	0.6483
LASSO	29	25	0.4201	0.5117	0.2465	0.7482
RF	29	43	0.4185	0.6540	0.1999	0.6384

Table 3: Model performance on the hold out set

	N vars	N coeffs	RMSE	AUC	Expected loss
X1	22	35	0.4169	0.6571	0.6532
X2	29	75	0.4155	0.6563	0.6488
LASSO	29	25	0.4182	0.6562	0.6576
RF	29	43	0.4144	0.6704	0.6468

Discussion of results and recommendations

As it was discussed earlier, the models do not perform very differently. In terms of prediction, their performances are very close, so a simple logit might be a good choice, but Random Forest increased it's performance both in terms of AUC and RMSE on the hold-out sample, suggesting a better predictive power.

By looking at our business case and quantifying the expected losses related to misclassification, the models again performed very similarly (except for LASSO). The advantage of Random Forest is less than 10 EUR per classification.

The confusion table shows a similar idea, displaying that with optimal threshold calculated on the working sample, which percentage of the companies are predicted into given groups and what the actual status they had in the hold out sample. As the Client loses more on missing a HGC (4.000 EUR loss) than wrongly offering the new product to a not HGC (1.000 EUR loss), the share of wrongly categorized not HGC's is high. With a different loss function, it would look different.

\begin{table}[!h]

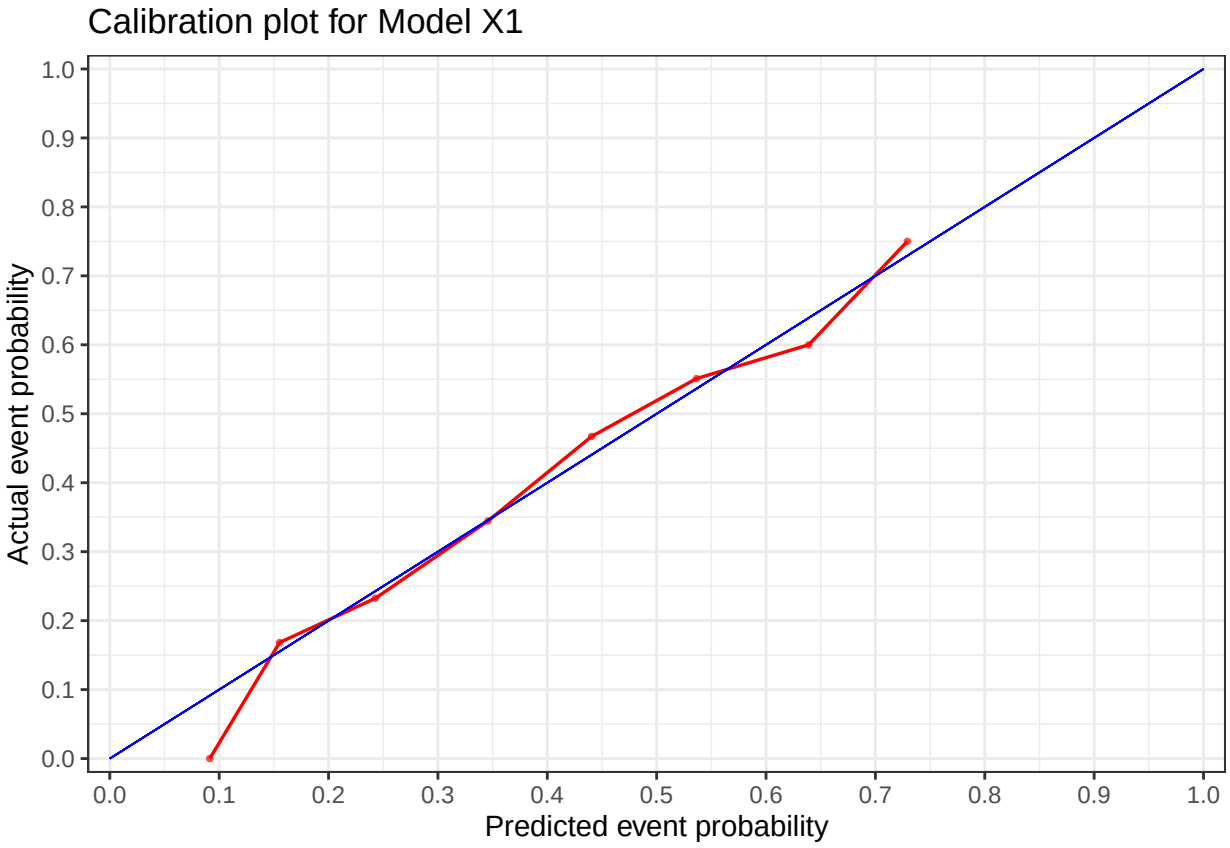
\caption{Confusion table of Random Forest Model in %}

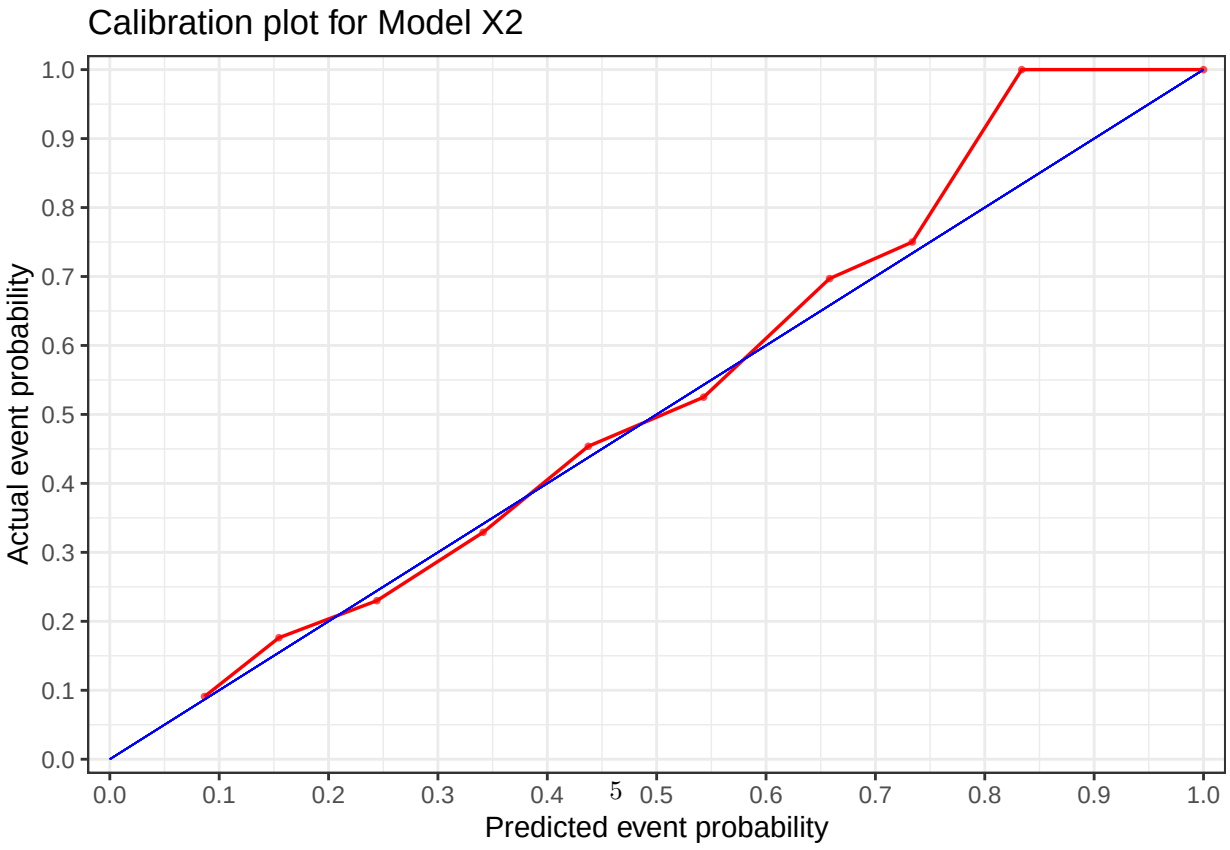
	Actual not HGC	Actual HGC
Predicted not HGC	30.4	5.2
Predicted HGC	44.9	19.5

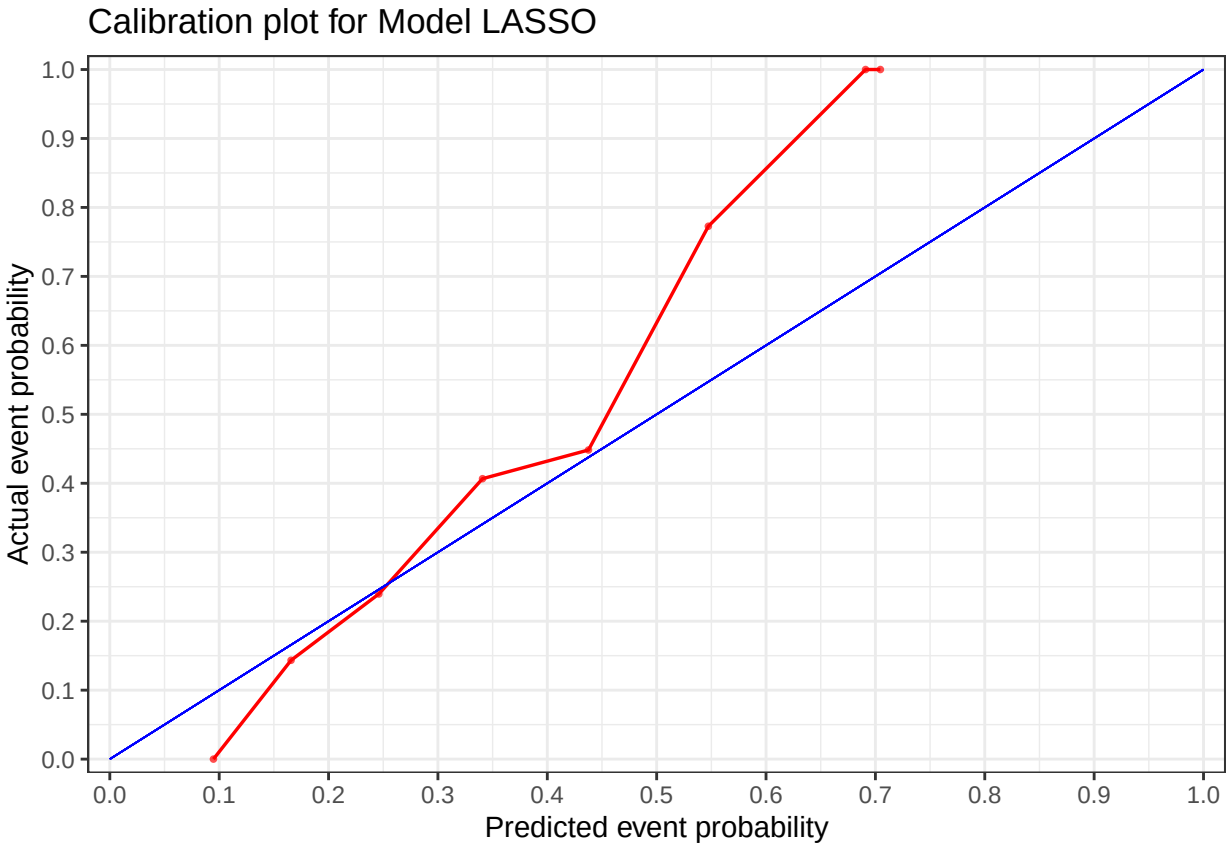
\end{table}

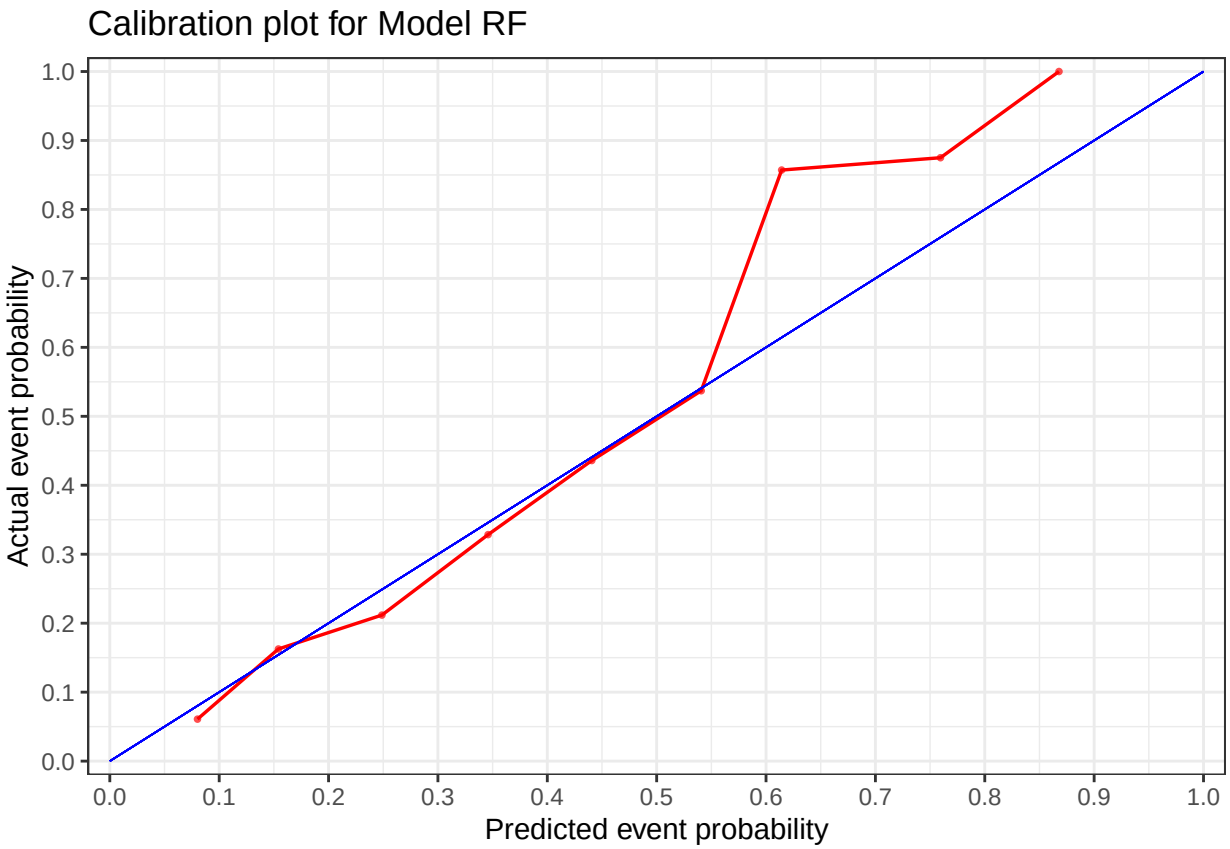
Annex

Calibration curves






Calibration plot for Model LASSO
Actual event probability
Predicted event probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0



Confusion tables

\begin{table}[!h]

\caption{Confusion table of X1 Model in %}

	Actual not HGC	Actual HGC
Predicted not HGC	31.8	5.7
Predicted HGC	43.5	19.0

\end{table}

\begin{table}[!h]

\caption{Confusion table of X2 Model in %}

	Actual not HGC	Actual HGC
Predicted not HGC	26.4	4.0
Predicted HGC	48.9	20.7

\end{table}

\begin{table}[!h]

\caption{Confusion table of LASSO Model in %}

	Actual not HGC	Actual HGC
Predicted not HGC	47.1	10.2
Predicted HGC	28.2	14.5

\end{table}

\begin{table}[!h]

\caption{Confusion table of Random Forest Model in %}

	Actual not HGC	Actual HGC
Predicted not HGC	30.4	5.2
Predicted HGC	44.9	19.5

\end{table}