Risk of business failure

Manuel Luci, Salvatore Fergola, Carlo Volpe

1 Dataset preparation (aida_preparation.r)

In this report a failed company is chosen to be the one whose legal status was "Bankruptcy" or "Dissolved (bankruptcy)", all the others are chosen to be active.
Useless features were deleted (e.g."Company name").
A new dataset was outlined creating new rows with respect to the features "-1 & -2" referred to the "Last accounting closing date".
A failed company at a given year ("Last accounting closing date") was considered as an active company for the two previous years.
The new feature "Age" was created and the Ateco code was replaced by the proper character (see this link ATECO CODE).

For question A, B, C, only values that made sense were selected (e.g. Age > 0), histograms of failed and active companies and "Cullen&Fray" graphs were chosen as a visual suggestion for theoretical distributions. Consequently, ks.tests were performed for continuous data to check if they were distributed as the theoretical distribution suggested (chi.tests for discrete ones). The same procedure was adopted to check if distributions of failed and active were drawn from the same distribution ; if the test failed a t.test was performed to check if mean were equal in the large sample hypothesis.
The significance level α was chosen to be 0.05, multiple corrections were accounted for changes in α.

2 Distributions of size/age/liquidity between failed and active at a specific year (questionA.r)

For this task the distributions of size (log ('Total assetsth EURLast avail. yr')), age and liquidity ratio between failed and active companies were compared at specific year 2011 (risk of default for Italy) fixing :

— Specific company form : everything but S.R.L's ;
— Ateco 2007 code : C.

1. Each test between active and failed companies was rejected (i.e. we have to reject H0 :"drawn by the same theoretical distribution").
2. A chi.test for the Age of active companies led not to reject the H0 : discrete power-law with a $p_{value} = 0.071$.
3. A ks.test for the size of failed companies with the specific "C" ATECO led not to reject the H0 : logistic distribution with a $p_{value} = 0.21$.

3 Distributions of size/age/liquidity for failed companies over different years (questionB.r)

For this task the distributions of size (log ('Total assetsth EURLast avail. yr')), age and liquidity ratio between failed companies were compared over 2006-2010 years (Great Recession) fixing :

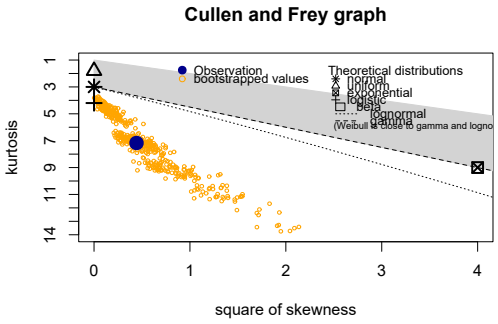

FIGURE 1 – Cullen and Frey Graph for Size (ATECO)

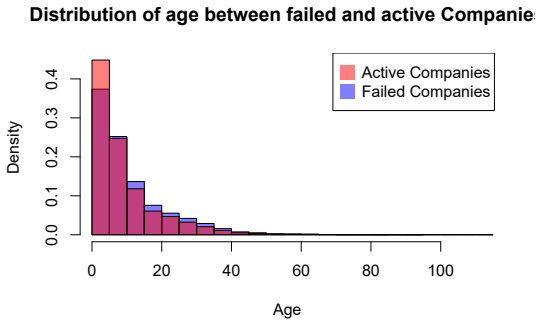

FIGURE 2 – Histogram Active and Failed graph distribution (General)

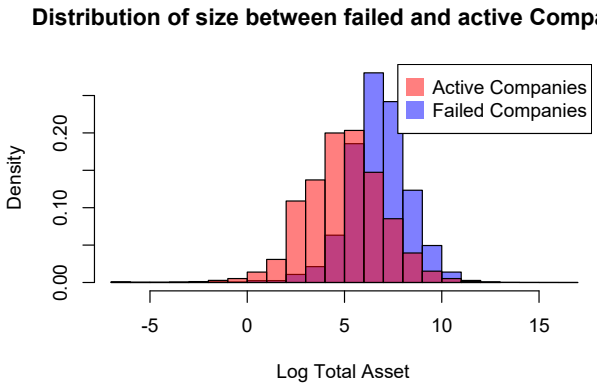

FIGURE 3 – Histogram Size General

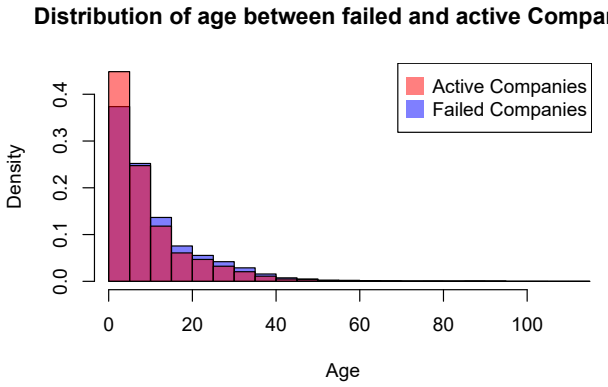

FIGURE 4 – Histogram Age General

— Specific company form : S.R.L ;
— Regions : Lombardia.

TABLE 1 – General

	Type of test	Year(s)	p-value(s)
Size	ks.test(plogis)	2006	0.12
	ks.test	2009 – 2010	0.21
	t.test	2006 – 2010	0.60
Age	chi.test(power-law)	2006	0.06
	chi.test(power-law)	2008	0.06
	chi.test	2007 – 2010	0.48
	chi.test	2009 – 2010	0.93
	t.test	2006 – 2007	0.67
Liquidity	ks.test	2009 – 2010	0.56
	t.test	2006 – 2008	0.50
	t.test	2006 – 2009	0.53
	t.test	2006 – 2010	0.31
	t.test	2007 – 2008	0.29

TABLE 2 – Specific company

	Type of test	Year(s)	p-value(s)
Size	ks.test(plogis)	2006	0.54
	ks.test(plogis)	2008	0.07
	ks.test	2009 – 2010	0.19
	t.test	2006 – 2009	0.34
	t.test	2006 – 2010	0.54
	t.test	2007 – 2010	0.33
Age	chi.test(power-law)	2006	0.08
	chi.test(power-law)	2008	0.09
	chi.test(power-law)	2010	0.08
	t.test	2006 – 2007	0.31
Liquidity	chi.test	2006 – 2007	0.66
	chi.test	2009 – 2010	0.31
	t.test	2006 – 2008	0.86
	t.test	2006 – 2009	0.31
	t.test	2008 – 2009	0.35
	t.test	2008 – 2010	0.34

TABLE 3 – Region

	Type of test	Year(s)	p-value(s)
Size	ks.test(plogis)	2006	0.39
	ks.test(plogis)	2007	0.17
	ks.test(plogis)	2008	0.42
	ks.test(plogis)	2009	0.40
	ks.test(plogis)	2010	0.66
	ks.test	$2007 - 2010$	0.48
	ks.test	$2009 - 2010$	0.93
	t.test	$2007 - 2009$	0.69
Age	chi.test	$2006 - 2007$	0.34
	chi.test	$2006 - 2009$	0.24
	chi.test	$2007 - 2008$	0.32
	chi.test	$2009 - 2010$	0.45
Liquidity	ks.test	$2006 - 2007$	0.31
	ks.test	$2006 - 2008$	0.69
	ks.test	$2009 - 2010$	0.43

In general the distributions over different years of failed companies were not comparable with 2008 ones (with respect to Size Age and Liquidity), an exception was the focus on Lombardia for 2007-2008 (Age) and 2006-2008 (Liquidity).
It is not possible to reject that 2009-2010 General distributions were drawn from the same one with respect the three features.
It could not be rejected that the Size distribution of 2006 was drawn always by a logistic one and the Age distribution of the same year from power-law one (exept for Region).
It is interesting to notice that it can not be rejected that the Size distributions for Lombardy were drawn from logistic ones.
There are many positive tests for the Liquidity ratio, this is something expectable due to the meaning of the ratio itself.

4 Probability of failures conditional to size/age/other of firms at a specific year(questionC.r)

For this task probabilities of failures conditional to size ($\log('\text{Total assetsth EURLast avail. yr}')$), age and liquidity ratio were compared at specific year 2016.
In the first place three "General" probabilities were computed.
Firms were split into four groups :"nord";"centro";"sud";"isole", then chi-test was performed inside groups, consequently between groups.
In the end probabilities of failure conditional to Ateco C-F-G and General probability were compared and the same for Company form S.R.L.

According to the chi-test is not possible to reject H0 hypothesis : probabilities of failure conditional to Liquidity and Age for Nord-Centro-Sud were drawn from the same distribution.
The hypothesis that Val d'Aosta probability of failure is drawn from distribution of another Nord region has to be rejected, same for Lombardy but size (drawn by the same distribution of Veneto). The hypothesis that Umbria probability of failure is drawn from distribution of another Centro region has to be rejected but Age (drawn by the same distribution of Marche).

Table 4

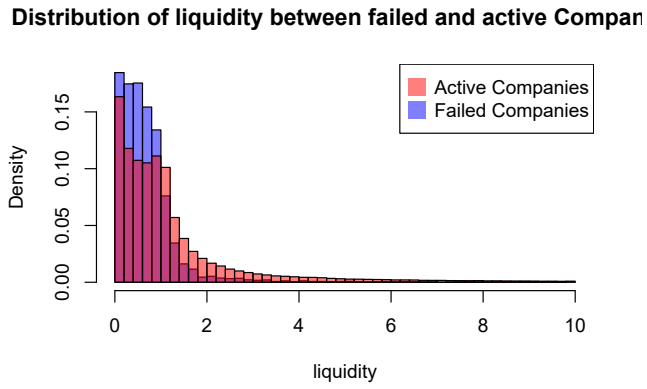

FIGURE 5 – Histogram Liquidity General

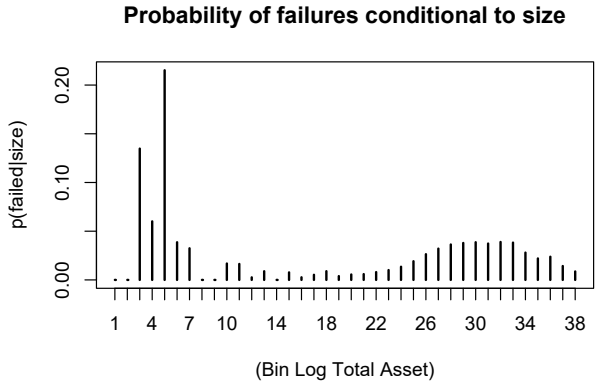

FIGURE 6 – Prob of failures conditional to size (General)

	Ateco	p-value(s)	Company Form	p-values
Size	C-F	0.39	S.R.L.-General	0.24
	C-G	0.42		
	F-G	0.42		
Liquidity	C-F	0.38	S.R.L.-General	0.25
	C-G	0.21		
	F-G	0.33		
Age	C-F	0.16	S.R.L.-General	0.42
	C-G	0.25		
	C-General	0.27		
	F-G	0.32		

The hypothesis that the most numerous Ateco (C-G-F) probabilities of failure were drawn by the same distribution has not to be rejected.

Table 5

	Regions	p-value(s)
Size	Piemonte-Emilia Romagna	0.38
	Lombardia-Veneto	0.40
	Trentino-Alto Adige -Liguria	0.39
	Trentino-Friuli-Venezia Giulia	0.37
	Liguria-Friuli-Venezia Giulia	0.44
	Veneto-Emilia-Romagna	0.39
	Lazio-Marche	0.25
	Lazio-Toscana	0.46
	Marche-Toscana	0.23
	Abruzzo-Calabria	0.38
	Basilicata-Molise	0.32
	Campania-Puglia	0.38
	Sicilia-Sardegna	0.17
	Nord-Centro	0.33
Liquidity	Piemonte-Veneto	0.44
	Piemonte- Emilia- Romagna	0.48
	Trentino-Liguria	0.38
	Trentino- Friuli-Venezia Giulia	0.43
	Liguria-Friuli	0.43
	Veneto - Emilia-Romagna	0.41
	Lazio-Toscana	0.18
	Marche-Toscana	0.25
	Abruzzo-Calabria	0.40
	Abruzzo-Puglia	0.32
	Basilicata-Molise	0.40
	Calabria-Puglia	0.26
	Campania-Puglia	0.24
	Sicilia-Sardegna	0.18
	Nord-Centro	0.33
	Nord-Sud	0.45
	Centro-Sud	0.26
Age	Piemonte-Veneto	0.40
	Piemonte-Emilia	0.42
	Trentino-Friuli-Venezia Giulia	0.43
	Liguria-Friuli-Venezia Giulia	0.39
	Veneto- Emilia-Romagna	0.45
	Lazio-Toscana	0.39
	Marche-Toscana	0.16
	Marche-Umbria	0.22
	Abruzzo-Calabria	0.48
	Basilicata-Molise	0.42
	Campania-Puglia	0.48
	Sicilia-Sardegna	0.40
	Nord-Centro	0.16
	Nord-Sud	0.18
	Centro-Sud	0.42

5 Failure prediction

5.1 Test and train preparation (test_train_preparation.R)

First of all, the missing values were deleted, then the number of active companies were undersampled to avoid an heavily misclassification [1]. The firms with "Last accounting closing date" < 2018 were included in the train set, the other ones in the test, this led to a train-test $\sim 88 - 12\%$.

5.2 Logistic regression(questionD1.r)

A Vif test was performed as a multicollinearity test in order to choose uncorrelated numerical variables for the logistic regression model. The omitted variable were :'Cash Flowth EURLast avail. yr', 'Total assetsth EURLast avail. yr', 'Return on asset (ROA)%Last avail. yr', 'Interest/Turnover (%)%Last avail. yr'.
A stepAIC "backward" algorithm reduced the number of variables.
In the end "Last accounting closing date" was considered misleading so it was removed.
The final variables used for the model were the following :

- 'Banks/turnover%Last avail. yr'
- 'Cost of debit (%)%Last avail. yr'
- 'Current liabilities/Tot ass.%Last avail. yr'
- 'Current ratioLast avail. yr'
- 'Debt/EBITDA ratio%Last avail. yr'
- 'Debt/equity ratio%Last avail. yr'
- 'EBITDA/Vendite%Last avail. yr'
- 'EBITDAth EURLast avail. yr'
- 'LeverageLast avail. yr'
- 'Liquidity ratioLast avail. yr'
- 'Net financial positionth EURLast avail. yr'
- 'Return on equity (ROE)%Last avail. yr'
- 'Return on investment (ROI) (%)%Last avail. yr'
- 'Solvency ratio (%)%Last avail. yr'
- 'Total assets turnover (times)Last avail. yr'
- 'Age'
- 'ATECO 2007code'
- 'Legal form'
- 'Registered office address - Region'.

Then the logistic regression was trained via repeated cross-validation (repeats=6 & number=12) trying to optimize the metric "AUC".

Table 6 – Logistic Regression Deviance Residual

Min	1Q	Median	3Q	Max
-3.04	-0.97	0.55	0.88	3.20

"AUC" was computed over the validation folds performing a shapiro test (p-value=0.50) and a z.test with confidence level=0.95, the results were : mean = 0.774 (CI 0.771-0.777). The final "AUC" computed over the test-set is 0.787 (CI 0.748-0.826 [2]). From the feature plot the interception is more or less at 0.5.
The confusion matrix with cutoff=0.5 is shown below :

		Reference	
		Active	Failed
Prediction	Active	2150	38
	Failed	723	92

The following plots display some quality measures for the logistic regression over the test-set.
Even if the majority class of the dataset was undersampled there is still a misclassification of the negative class (precision ~ 11 %, $F_1 \sim 20\%$).The Binary ECE is 0.31 and the Brier Score is 0.16.
Binary ECE is a usefull measure for the interpretation of the calibration plot, Brier Score is a measure of misclassification error.

1. They were almost 60 times more than failed ones
2. DeLong

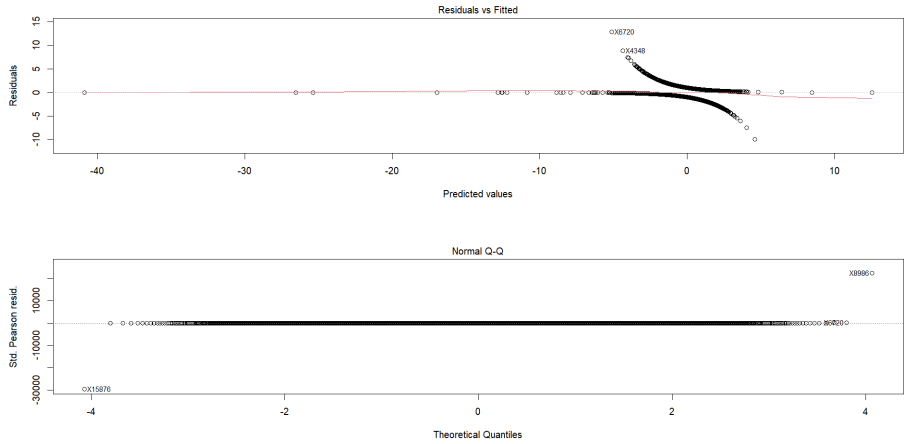

FIGURE 7 – Diagnostic Plot logistic regression

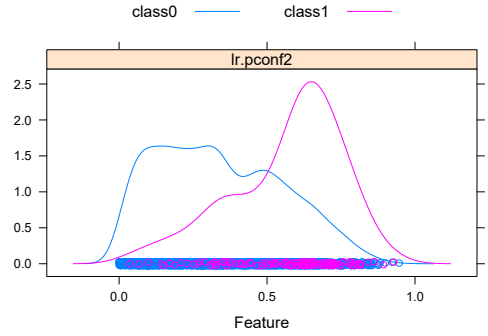

FIGURE 8 – Feature Plot logistic regression

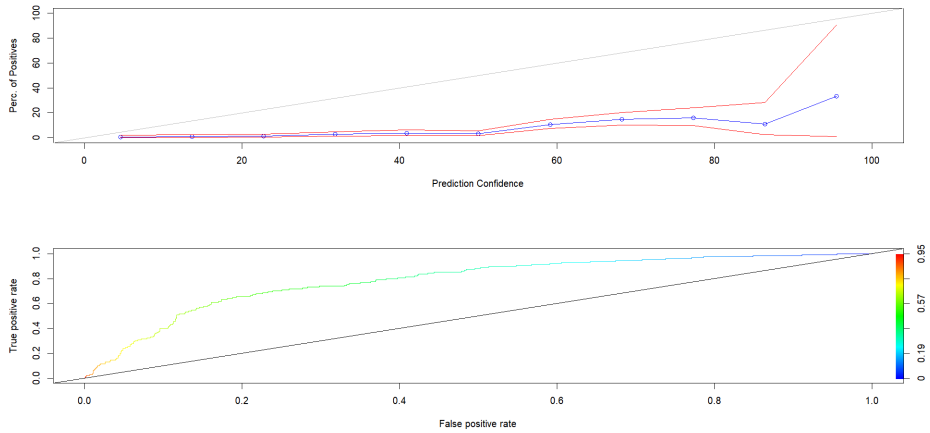

FIGURE 9 – Calibration and ROC plot Logistic Regression

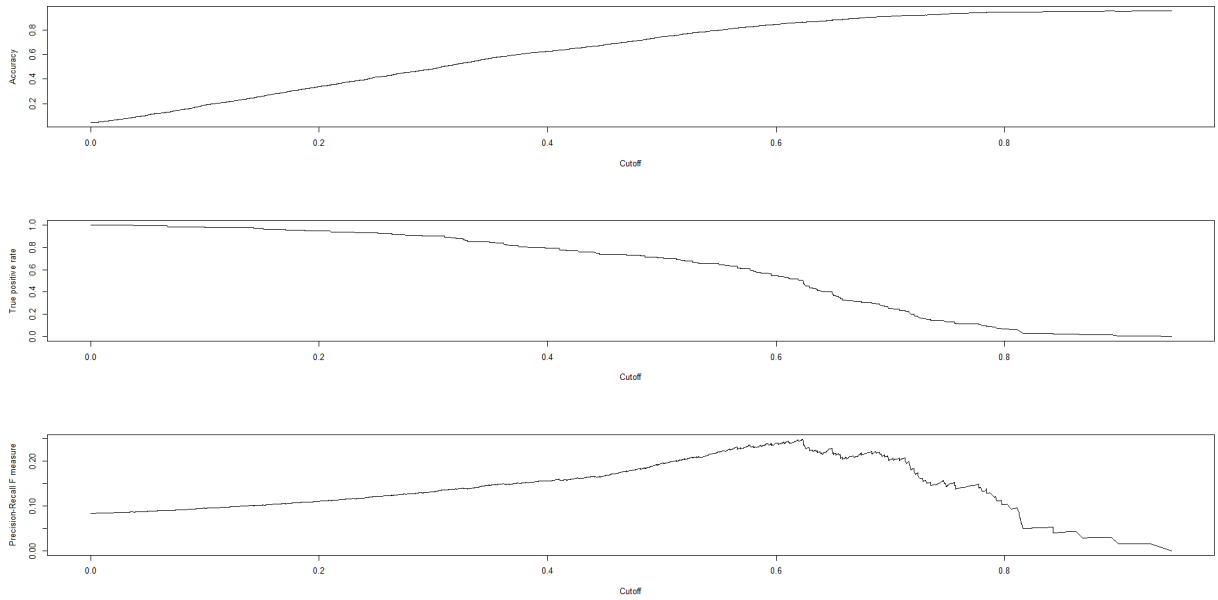

FIGURE 10 – Quality Plot Logistic Regression

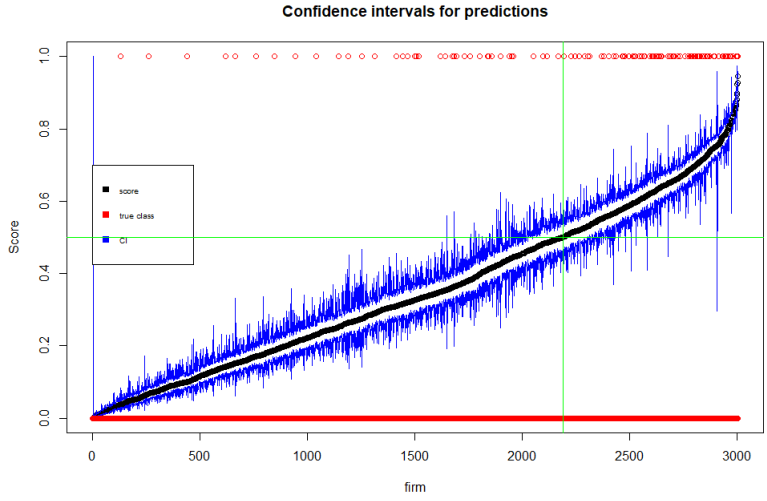

FIGURE 11 – CI Logistic Regression

The "Confidence Interval Plot" gives a further proof of misclassification of this model.
The mean value of the Confidence Intervals' width should be the lower bound of the ratings' binwidth, in this case is 0.0814 (CI= 0.0797 - 0.0833) [3].

From the scoring model a rating model was built : firms were classified in five classes, with ratings from "A" to "E", the probability cutoffs were set as $q_{0.2} = 0.33, q_{0.4} = 0.53, q_{0.6} = 0.68, q_{0.8} = 0.78$ over the scores of the train-set.
The procedure accounts for the Feature Plot : the interception between the two classes' densities at certain cutoff (future percentile p_c) explains in some sense the score at which our classifier splits the two classes. The middle class "C" of the rating model could be defined as a neighbourhood of the chosen p_c ($p_c - 0.1 \leq$ "C" $\leq p_c + 0.1$) [4].
Percentiles referring to other classes were chosen as midpoints between 0-(p_c-0.1) and (p_c+0.1)-1 in order to have a balanced rating model.
The barplot shows that this rating model is not the best one, in fact for the "D"-"E" rating classes there are more Active than failed company (the model classifies the active in "A"-"B" better).

5.3 Random forest(questionD2.r)

The train and test set used for the random forest are the same indicated in *subsection 5.1.*
The training was performed via repeated cross-validation (repeats=6 & number=12) trying to optimize the metric "AUC". The number of trees were chosen to be 15
"AUC" was computed over the validation folds performing a shapiro test (p-value=0.81) and a z.test with confidence level=0.95, the results were : mean = 0.846 (CI 0.843-0.848).
The confusion matrix with cutoff=0.5 in shown below :

		Reference	
		Active	Failed
Prediction	Active	2732	86
	Failed	263	44

From the feature plot the interception is more or less at 0.3 ; this choice could be usefull from an investor's point of view : there are less FN [5].
The confusion matrix with cutoff=0.3 is shown below :

		Reference	
		Active	Failed
Prediction	Active	2065	38
	Failed	930	92

The following plots display some quality measures over the testset for the random forest.
The BinaryECE is 0.22 and the Brier Score is 0.096.
The final "AUC" computed over the test-set is 0.776 (CI 0.741-0.812).
Even if the majority class of the dataset was undersampled there is still a misclassification of the negative class (precision $\sim$ 14 %, $F_1 \sim 20\%$)

The rating model procedure is the same as before, this time the cutoff from FeaturePlot Random-Forest is $p_c = 0.3$. The probability cutoffs were set as $q_{0.1} = 0.13, q_{0.2} = 0.33, q_{0.4} = 0.67, q_{0.7} = 0.87$ over the scores of the train-set. The barplot shows that this rating model is slightly better with respect to the one obtained from Logistic Regression (there are almost no failed companies with "A" rating and more failed companies for the others).

3. In the rating model this condition will be satisfied
4. This because a range of 0.2 was chosen for the middle "C" class
5. There is less probability to invest in a riskier firm

The rcv Random Forest and rcv Logistic Regression were compared using a t.test (shapiro and var.test confirmed they were drawn from normal distribution with same variances [6]).

6 Implementation with library ROSE

6.1 Dataset prepration (questionE_data_preparation.r)

The previous models showed an evident problem of misclassification, one of the causes of this problem was a unbalanced dataset. A better way to balance dataset than the one used previously could be ROSE algorithm : it produces a synthetic, possibly balanced, sample of data simulated according to a smoothed-bootstrap approach, as described in Menardi and Torelli (2013).
The splitting between train and test was done as before.

6.2 Logistic Regression (questionE.r)

This time the Vif procedure returned 0 variables deleted.

The final variables used for the model after step_AIC("backward") were the following :

- 'Banks/turnover%Last avail. yr'
- 'Cash Flowth EURLast avail. yr'
- 'Cost of debit (%)%Last avail. yr'
- 'Current liabilities/Tot ass.%Last avail. yr'
- 'Current ratioLast avail. yr'
- 'Debt/equity ratio%Last avail. yr'
- 'EBITDA/Vendite%Last avail. yr'
- 'Interest/Turnover (%)%Last avail. yr'
- 'LeverageLast avail. yr'
- 'Net financial positionth EURLast avail. yr'
- 'Return on asset (ROA)%Last avail. yr'
- 'Return on equity (ROE)%Last avail. yr'
- 'Return on investment (ROI) (%)%Last avail. yr'
- 'Solvency ratio (%)%Last avail. yr'
- 'Total assets turnover (times)Last avail. yr'
- 'Age'
- 'ATECO 2007code'
- 'Legal form'
- 'Registered office address - Region'.

The logistic regression was trained via repeated cross-validation (repeats=6 & number=12) trying to optimize the metric "AUC" (as before).

Table 7 – Log Regr Dev Residual with ROSE

Min	1Q	Median	3Q	Max
-3.03	-1.00	0.55	0.92	2.73

"AUC" was computed over the validation folds performing a shapiro test (p-value=0.05) and a z.test with confidence level=0.95, the results were : mean = 0.760 (CI 0.758-0.763). The final "AUC" computed over the test-set is 0.744 (CI 0.712-0,768) [7].
BinaryECE=0.275, slightly better than before.
From the feature plot the interception is more or less at 0.5.
The confusion matrix with cutoff=0.5 is shown below :

		Reference	
		Active	Failed
Prediction	Active	1945	131
	Failed	829	289

6. p-values for shapiro were alredy reported, p-value for var.test was 0.85
7. DeLong

The following plots display some quality measures for the logistic regression over the test-set.
The actual "AUC" is lower than the one computed for logistic regression without "ROSE", nevertheless there is a significant improvment of precision $\sim$ 26 %, $F_1 \sim 38\%$.

6.3 Random Forest (questionE.r, questionE$_test.r$

The train and test set were the same used in *subsection 6.2.*
The training was performed via repeated cross-validation (repeats=6 & number=12) trying to optimize the metric "AUC", the number of trees were chosen to be 15.
"AUC" was computed over the validation folds performing a shapiro test (failed) and a t.test with confidence level=0.95, the results were : mean = 0.9992 (CI 0.9991-0.9993).
The BinaryECE is 0.145. The final "AUC" computed over the test-set is 0.998 (CI 0.997-1) [8].
The confusion matrix with cutoff=0.5 in shown below :

		Reference	
		Active	Failed
Prediction	Active	2766	2
	Failed	8	418

The tails intercept at 0.5, the two densities are separated from each other. The following plots display some quality measures over the testset for the random forest with ROSE implementation.
The BinaryECE is 0.15 and Brier Score is 0.03.
These are the best results of the report (precision $\sim$ 98 %, $F_1 \sim 99\%$).
These results seem to be an overfitting symptom, in fact performing a prediction over the test set of section D (without ROSE) verifies this hypothesis.
The BinaryECE is 0.72. The Brier Score is 0.58.
The precision is 5% and F_1= 8%. The confusion matrix is :

		Reference	
		Active	Failed
Prediction	Active	320	1
	Failed	2675	129

8. Delong

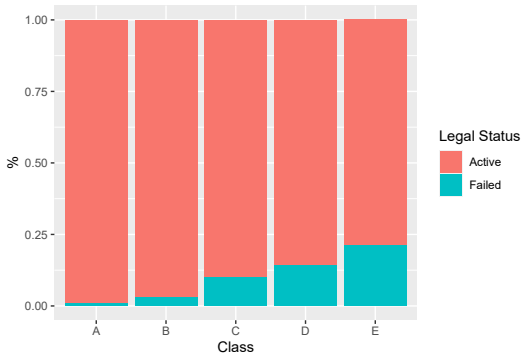

FIGURE 12 – Barplot Rating Model Logistic Regression

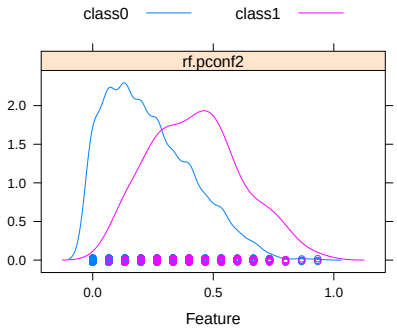

FIGURE 13 – Feature Plot Random-Forest

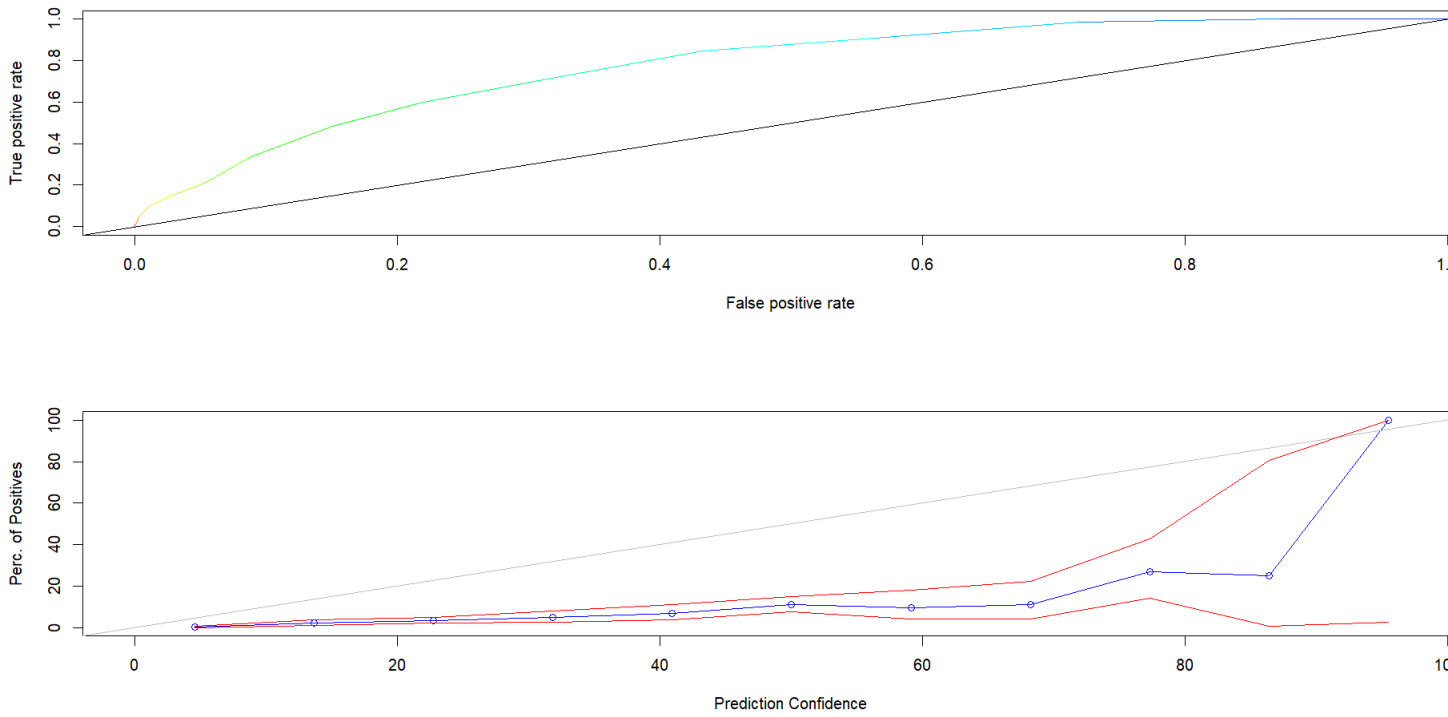

FIGURE 14 – Calibration and ROC plot Random-Forest

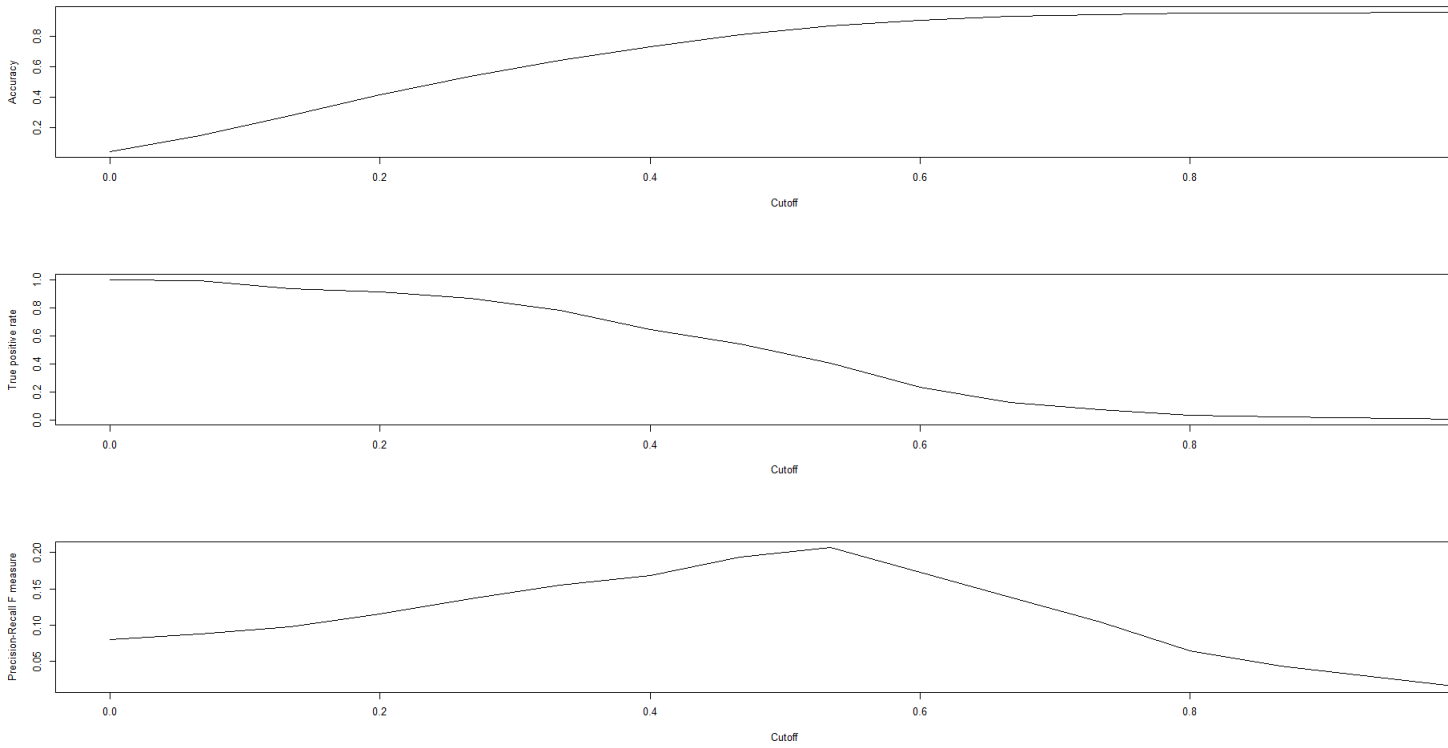

FIGURE 15 – Quality Plot Random-Forest

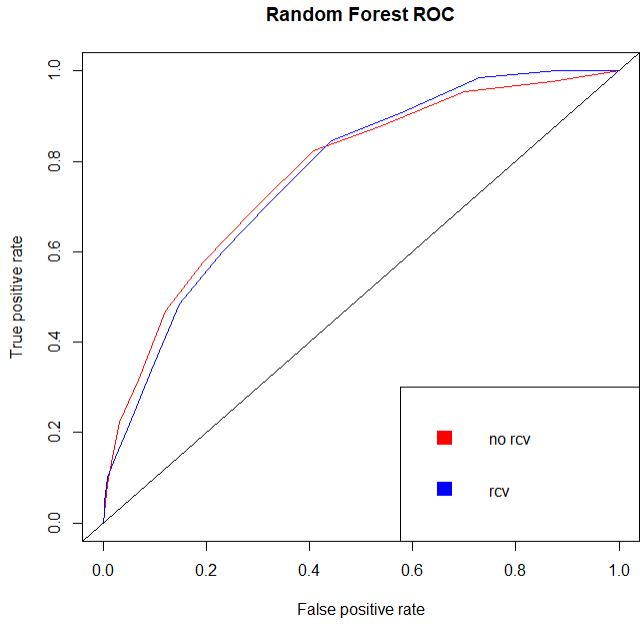

FIGURE 16 – ROC curve rcv Random Forest- Roc curve no rcv Random Forest

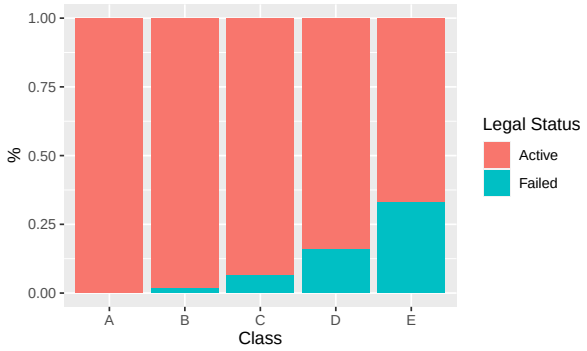

FIGURE 17 – Barplot Rating Model Random-Forest

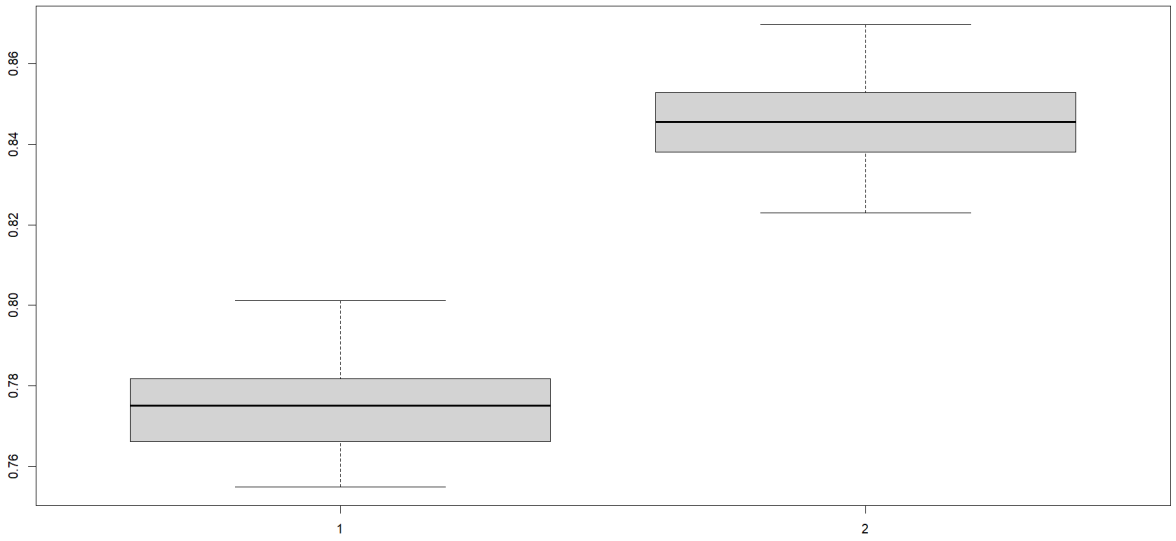

FIGURE 18 – Boxplot over validation folds for rcv Logistic (1) and rcv RF (2)

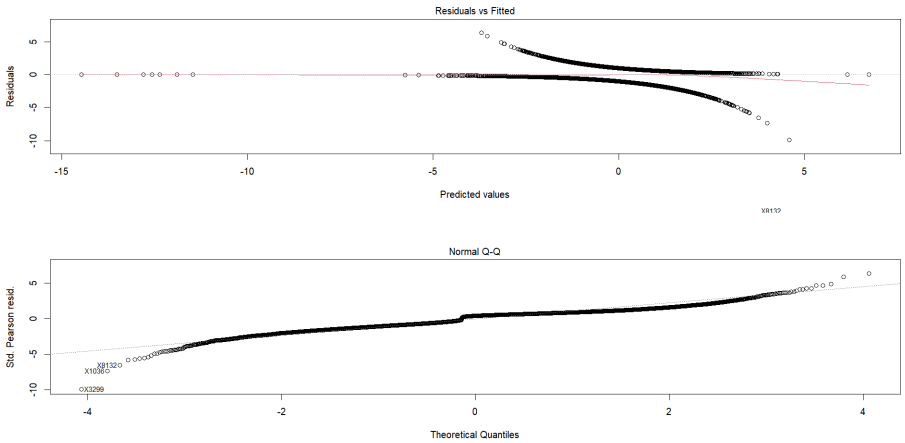

FIGURE 19 – Diagnostic Plot logistic regression

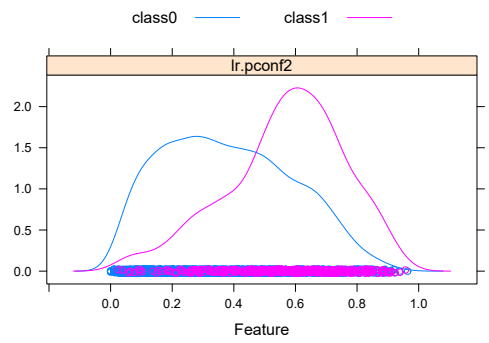

FIGURE 20 – Feature Plot logistic regression

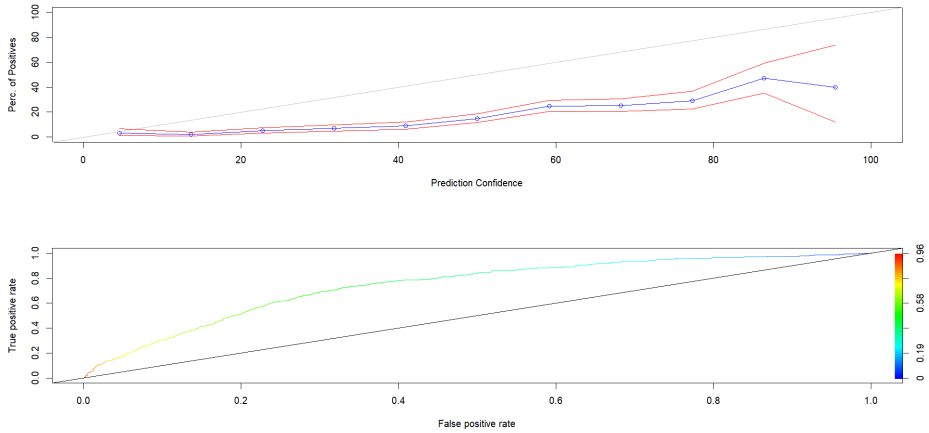

FIGURE 21 – Calibration and ROC plot Logis Regr with ROSE

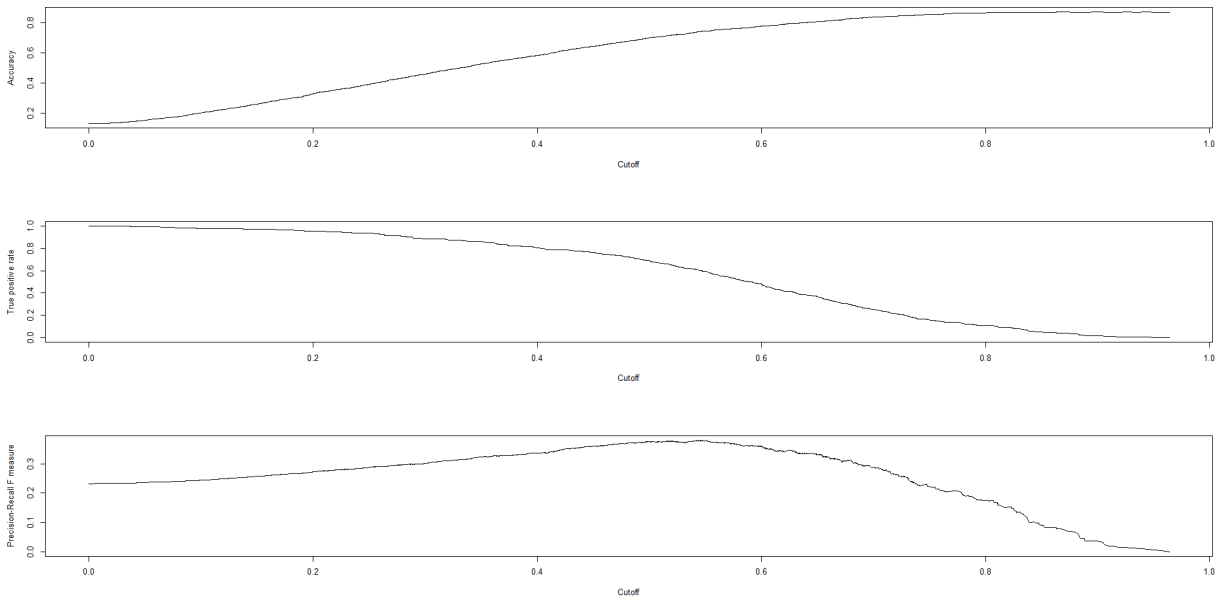

FIGURE 22 – Quality Plot Logis Regr with ROSE

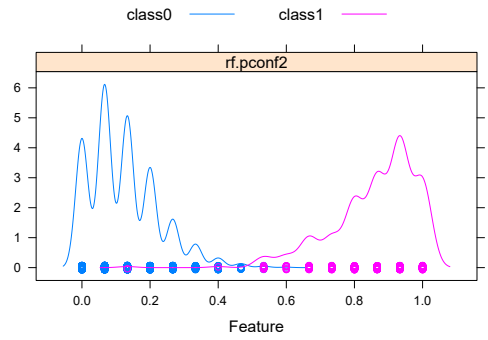

FIGURE 23 – Feature Plot Random-Forest ROSE

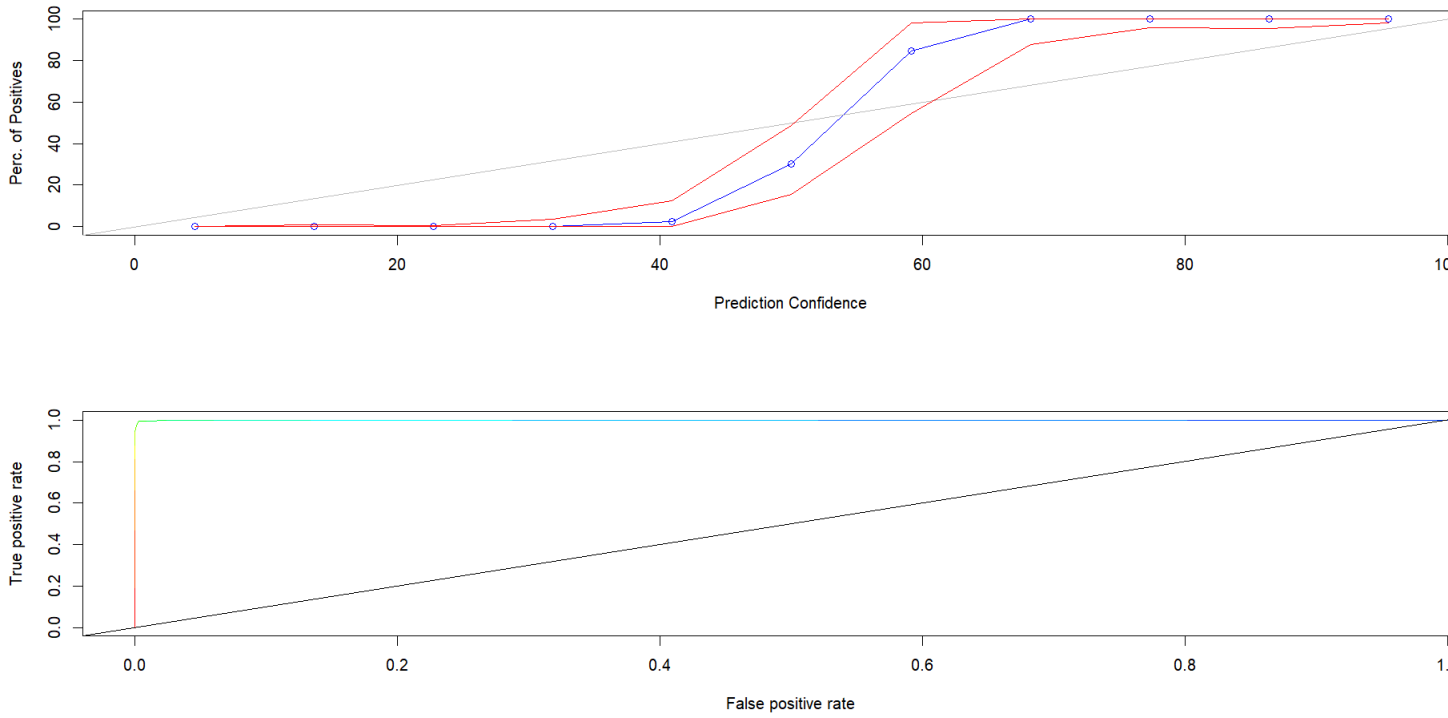

FIGURE 24 – Calibration and ROC plot Random-Forest ROSE

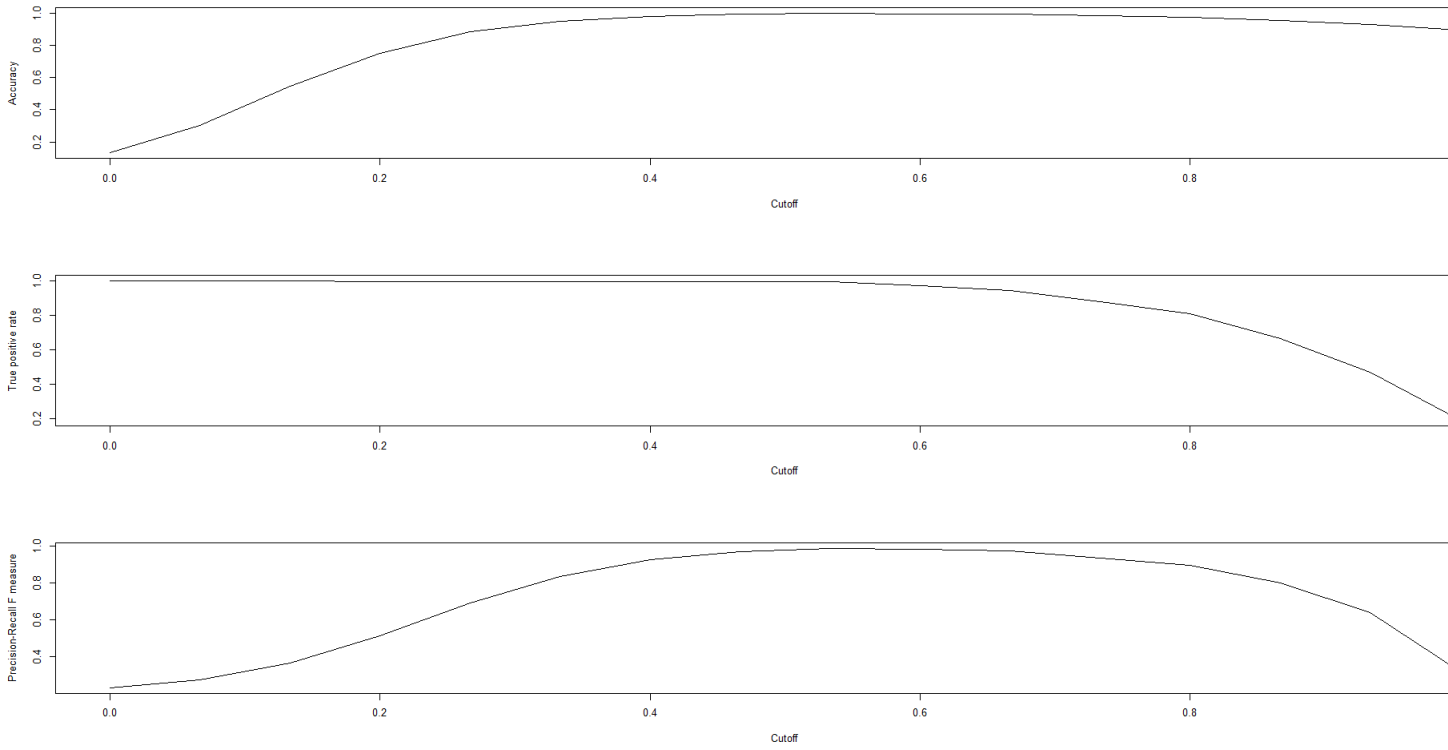

FIGURE 25 – Quality Plot Random-Forest ROSE

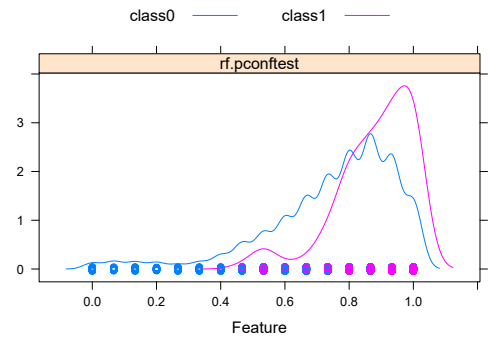

FIGURE 26 – Feature Plot RF ROSE over old test set