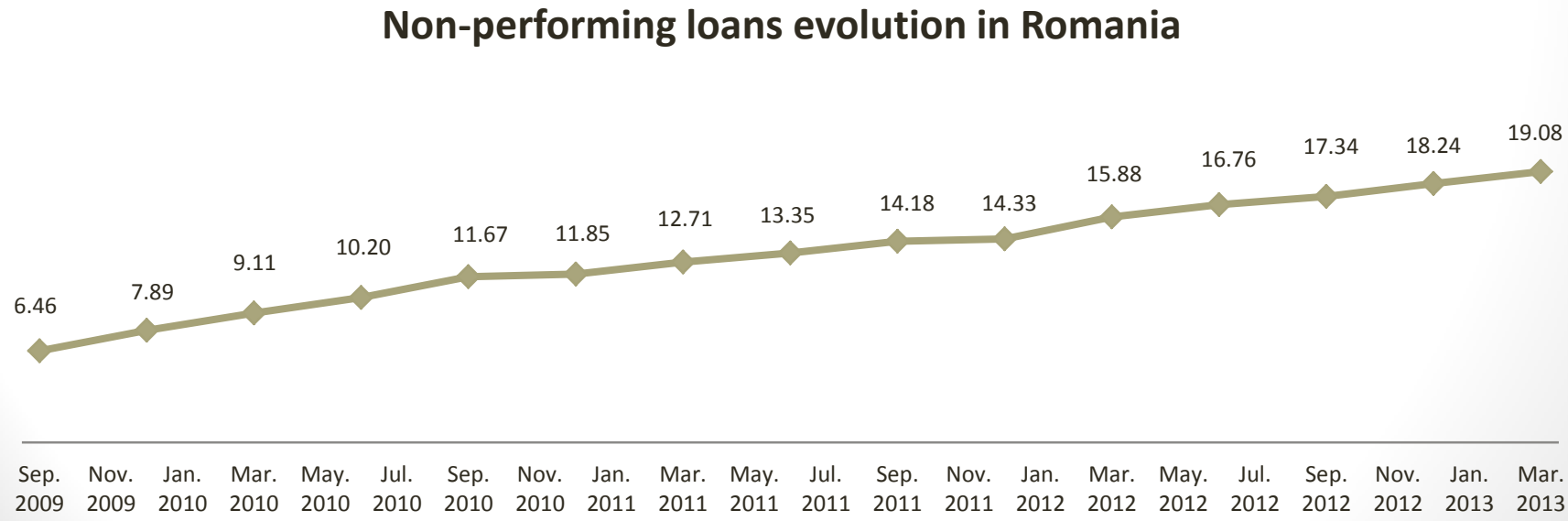
CREDIT SCORING MODELS for default probabilities in a corporate bank portfolio

Supervisor: Phd. MOISĂ ALTĂR

MSc. Student: GAFAR ŞENIS

Motivation

- Recent high and constantly growing default frequency underlined the importance for a bank to develop early warning systems that can help **prevent or avert** corporate default and that facilitate the **selection of firms** to collaborate with or invest in
- Under Basel II regulations, banks are allowed to use their own estimated risk parameters for calculating regulatory capital for credit risk, known as Internal Ratings-Based (IRB) Approach.



Source: National Bank of Romania interactive database

Paper Objectives

- The paper aims to assess and compare different credit scoring techniques in order to predict the default probability of corporate clients, using qualitative, quantitative and macroeconomic variables for a commercial bank's corporate credit client's portfolio.
- I use different credit scoring techniques to identify which one fits better the Banks portfolio in order to **reduce capital requirements using internal rating based approach**.
- Comparing different logistic regression models and classification trees, shows that for the database used, the best model is the logistic regression model with all available variables included.
- A Bayesian logistic regression model using an informative prior from the logistic regression model is performed, in order to improve the logistic regression model.

Table of content

- Literature review
- Methodology used
- Data analysis
- Empirical results
- Conclusions

Literature review

- In **1994, Altman *et al*.** provided one of the first assessments of neural networks in credit scoring, by comparing the neural networks to linear discriminant analysis (LDA), when LDA performed better,
- In **1996, Desai et al.** obtained for a credit union data set, a neural network performed better than LDA but did not perform significantly better than the logistic regression.
- In **1997, Hand and Henley** give a larger overview of different models used for credit scoring: they compare discriminant analysis, regression analysis, logistic regression, probit analysis, mathematical programming, recursive partitioning (decision trees), expert systems, neural networks, nonparametric smoothing methods and time varying models.
- They state that **"there is no overall best model",** because **the best model depends on the data structure.**

Methodology used

- Logistic regression model
- Bayesian logistic regression model
  - With non-informative prior
  - With informative prior
- Classification and Regression Trees
  - Without prior
  - With prior
- Conditional inference Tree

- Validation
  - ✓ Analysis of Variance test (ANOVA)
  - ✓ Chi-square statistic test
- Discriminatory Performance
  - ✓ ROC Curve
  - ✓ Area under the curve (AUROC)
  - ✓ Accuracy ratio (AR)
  - ✓ KS-statistic

Logistic regression

- Estimates the probability of a certain event occurring
- Is given by

$$\pi\,(\mathbf{x}) = \frac{exp\left(\beta_0 + \sum_{j=1}^{k} \beta_0\, x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{k} \beta_0\, x_{ij}\right)}$$

- Applies maximum likelihood estimation after transforming dependent variables into logit variable
- Likelihood function is:

$$L\,(y \backslash \beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

- Conditional probability of the default is given by

$$\mathbf{P}\left(\mathbf{y_i} = 1 \,\middle|\, \mathbf{x_i^k}\right) = \pi\,(\mathbf{x_i^k})$$

- Estimates changes in the log odds of the dependent variable, not changes in the dependent itself.
- The odds that the firm defaults are: $\mathbf{odds_i} = \frac{\pi\,(\mathbf{x_i^k})}{1 - \pi\,(\mathbf{x_i^k})}$

Bayesian Logistic regression

- Formulated by specifying prior distributions on the regression coefficients:

$$\beta_j \sim p\left(\beta_j \middle| \theta_j\right), j = 0, \ldots M$$

- posterior expectations of a function f(θ_j) is

$$\boldsymbol{E}\left[\boldsymbol{f}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\int \boldsymbol{f}(\boldsymbol{\theta})\, \boldsymbol{p}(\boldsymbol{\theta})\, \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \boldsymbol{p}(\boldsymbol{\theta})\, \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}\right.$$

- The posterior distribution is proportional to the product of the prior distribution and likelihood:

$$\boldsymbol{\pi}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \;) \; \boldsymbol{\alpha}\, \boldsymbol{L}\, (\boldsymbol{y}|\, \boldsymbol{\beta}) \,, \boldsymbol{\pi}\left(\boldsymbol{\beta} \;\right)$$

$$\boldsymbol{\pi}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \;) \boldsymbol{\alpha} \prod_{i=1}^{n} \boldsymbol{\pi}_i^{y_i} (\mathbf{1} - \boldsymbol{\pi}_i)^{1-y_i} \prod_{j=1}^{k} \frac{\mathbf{1}}{\sqrt{2\boldsymbol{\pi}\boldsymbol{\sigma}_j^2}} \boldsymbol{exp}\left(-\frac{\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{0j}\right)^2}{2\boldsymbol{\sigma}_j^2}\right)$$

Classification and regression trees

- Same as a linear model, a CART model can be used for classification or regression.
- Is created by splitting a population into smaller and smaller segments, using splitting rules based on the value of the predictor variables
- Individuals in different segments display different behavior, and decisions about how to treat someone are based on the properties of the final segment into which they fall after all splitting rules have been applied.
- The development sample is segmented into two parts, based on the properties of the predictor variables and the relationship they display with the dependent variable. the processes is repeated for each of the two “child” segments that resulted
- The process stops when there are too few individuals in a parent node for it to make sense to split it further, or when no further differentiation between groups can be found;

Conditional inference Trees

- overcomes the instability and biases found in traditional recursive partitioning like CART
- offer a concept of statistical significance based on bonferroni metric unlike traditional tree methods
- A regression model describing the conditional distribution of a response variable Y given the status of m covariates by means of tree-structured recursive partitioning.
- The response Y from some sample space Y may be multivariate as well. The m-dimensional covariate vector **X = (X1, . . . , Xm)** is taken from a sample space **X = X1 × · · · × Xm**.
- The conditional distribution D(Y|X) of the response Y given the covariates X depends on a function f of the covariates.

$$\mathbf{D(Y|X) = D(Y|X1, \ldots, Xm) = D(Y|f(X1, \ldots, Xm))}$$

- where we restrict ourselves to partition based regression relationships, i.e., r disjoint cells *B1, . . . , Br* partitioning the covariate space X = $\bigcup_{k=1}^{r} Bk$

Data analysis

- The data set contains information about 6263 private firms' bank loans, granted in 2010 and 2011
- For Exposure, a material threshold of 20,000 RON (EUR 4500) is applied, in order to exclude very small loans.
- Clients are randomly separated into 3 data sets: training data set, validation data set, and testing data set
- The overall default percentage of the database is 14%
- All entries with missing or extreme values were removed from database
- Companies within financial, real estate or public sector and some categories of organizations (e.g. charitable or religious nature) were eliminated, because of different structures and bankruptcy environment.
- **Collinearity** of independent variables is verified using the correlation matrix. No worrying correlations occur.

Database used

Data	Description	Type
Client ID	Client Identification number	unique number
Default	Default event	Binary (Default event ocurred = 1, not ocurred = 0)
REL	Relationship with bank in years	continuous
AGE	Experience on market	continuous
EXP	Client's exposure at Bank	continuous
ACTIV	Activity	categorical (agr/constr/ind/services)
CIF	Turnover trend	categorical (1/2/3/4/5)
CAL	Shareholder quality	categorical (1/2/3/4/5)
MAN	Management quality	categorical (1/2/3/4/5)
STA	Strategy	categorical (1/2/3/4/5)
PIA	Market conditions	categorical (1/2/3/4/5)
RAP	Reporting quality	categorical (1/2/3/4/5)
COL	Collateral quality	categorical (0/1/2/3/4/5)
CurrLiq	Current Liquidity (Current Assets/ Current Liabilities)	continuous
DebtCov	Debt coverage (Total liabilities/Debt service > 1 Y)	continuous
IntCov	Interest coverage (EBIT/interest expenses)	continuous
TARot	Total assets rotation (Turnover/Total assets)	continuous
TanARot	Tangible assets rotation (Turnover/Tangible assets)	continuous
ROCE	Profitability (EBIT/Total liabilities)	continuous
GDPgr	GDP growth for Romania (-1.1 for 2010, 2.2 for 2011)	continuous

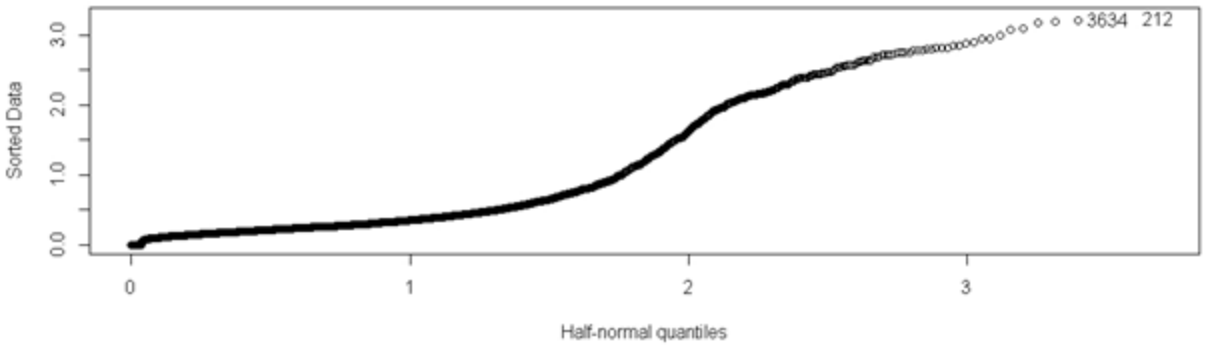
Logistic regression models

All variables	Model developed using all 19 variables
Stepwise selection	Model using variables chosen by stepwise selection method
Significant variables by p value	Model using variables chosen by p-value significance level of 0.05
Removed outliers and influential observations	Model developed after outliers and influential observations were removed using half-normal plot of residuals, , leverages and Cooks statistics

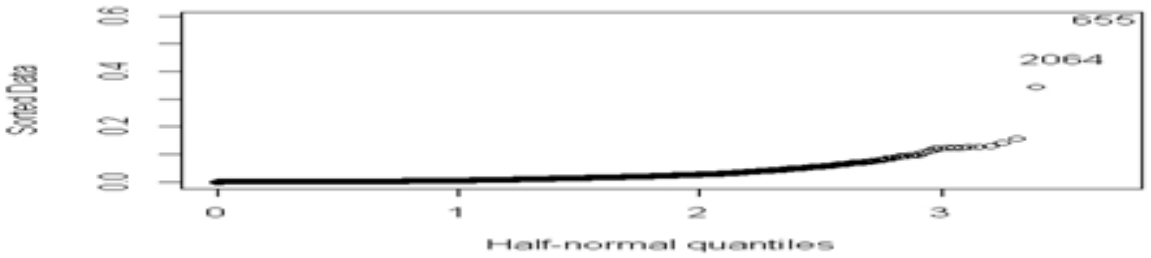
Outliers and influential observations

- To test if the model has outliers or influential observations, a half normal plot of residuals, leverages and Cooks statistics are considered.

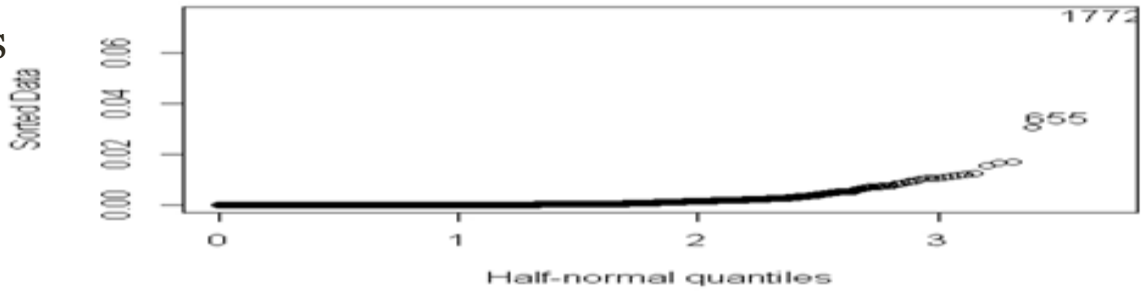
Half Normal Plot of residuals shows no outliers



Half normal plot of leverages shows that observations numbered 2064 and 655 may affect the fit of the model.



Half normal plot of Cook's distance shows that observations with number 655 and 1772 are influential



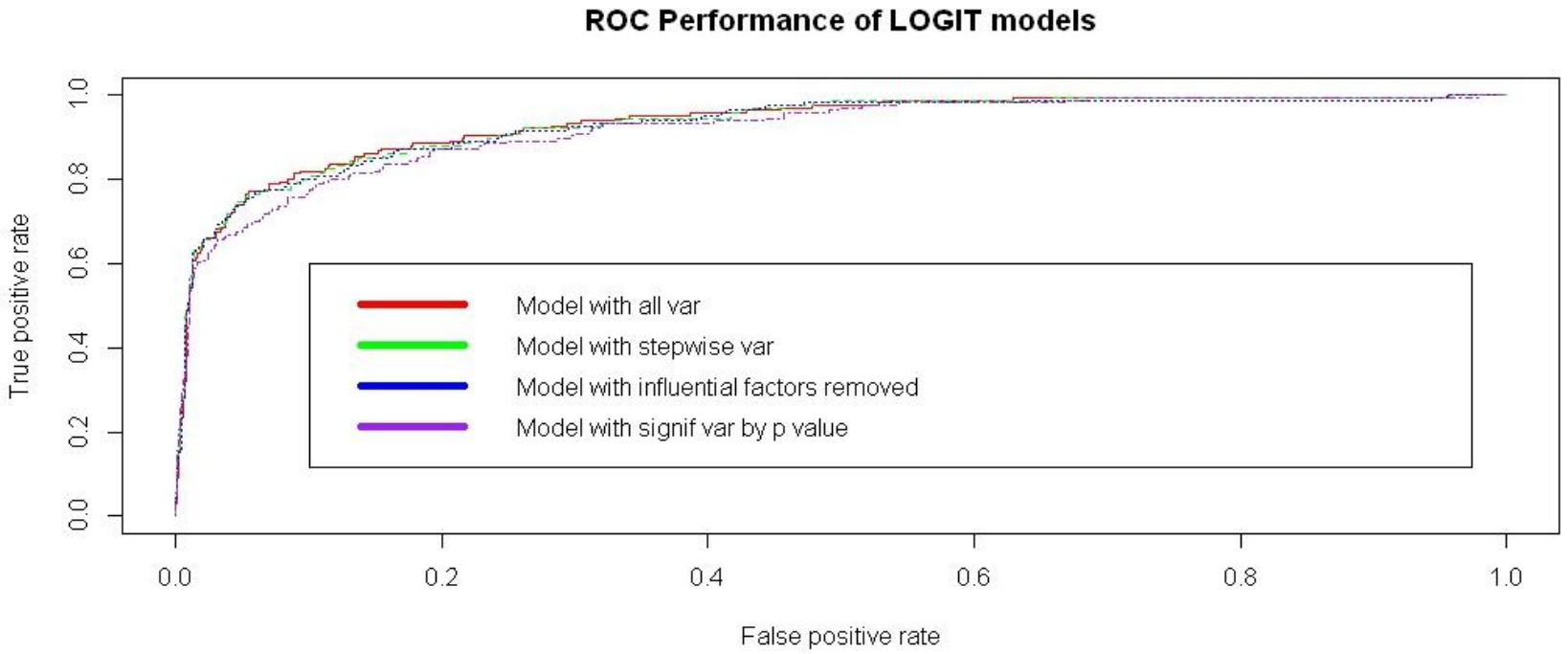
Empirical results

Logistic regression models

Summary of logistic regression models using different estimation techniques

Model	No. of variables used	No. of Fisher Scoring iterations	Null deviance	Degrees of freedom	Residual deviance	Degrees of freedom	AIC
Model 1 - All variables	19	16	5054.7	6261	2463.4	6215	2557.4
Model 2 - Model with stepwise var	10	16	3529.1	4381	1713	4350	1777
Model 3 - Model with signif var by p-value	9	7	5054.7	6261	2717.6	6234	2773.6
Model 4 - Model with influential factors removed	10	16	3520.9	4378	1704.8	4347	1768.8

Discriminatory Performance of Logistic regression models



Model	AUROC	AR	KS Statistic
Model 1 with all variables	**0.933370**	**86.67%**	**0.7256132**
Model 2 with stepwise selection variables	0.931373	86.27%	0.7129786
Model 3 with influential factors removed	0.927833	85.57%	0.7113966
Model 4 with significant var by p value	0.918019	83.60%	0.6864206

Bayesian Logistic Regression Model

- To obtain the posterior estimates, a Markov Chain with 510,000 samples is generated for both models. To allow enough time for the Markov Chain to converge to the stationary distribution, the first 500,000 samples were excluded. Thus, we have left a Markov Chain with 10,000 samples and a burn-in period of 500,000.
- **Bayesian Logistic Regression Model** with informative priors is using parameters from the reduced logistic regression performed earlier
- **Bayesian Logistic Regression Model** with non-informative prior will use a uniform prior.

Bayesian Logistic Regression Model

With informative prior

- Only variables for ACTIVITY=services, Market conditions = stable and emerging market (PIA 2 and PIA 5), collateral quality = Guarantees received from international companies, financial /insurance companies, credit guarantee funds with good reputation **are insignificant.**
- Geweke diagnostics statistics show that variables
  - COL1 (good collateral quality)
  - COL4 (Non-cash guarantees that can damage the market value in time)
  - Current Liquidity **did not converged.**

With non-informative prior

- There is a larger number of non-significant variables
- The Strategy factor describing defined objectives, the stable market indicator and emerging markets indicator **are now insignificant**, but stable market, an insignificant factor for Bayesian regression with informative prior, **is now significant.**
- Geweke diagnostics for Bayesian Logistic regression model with non-informative prior is showing that all variables converged.

Logistic regression model vs. Bayesian models

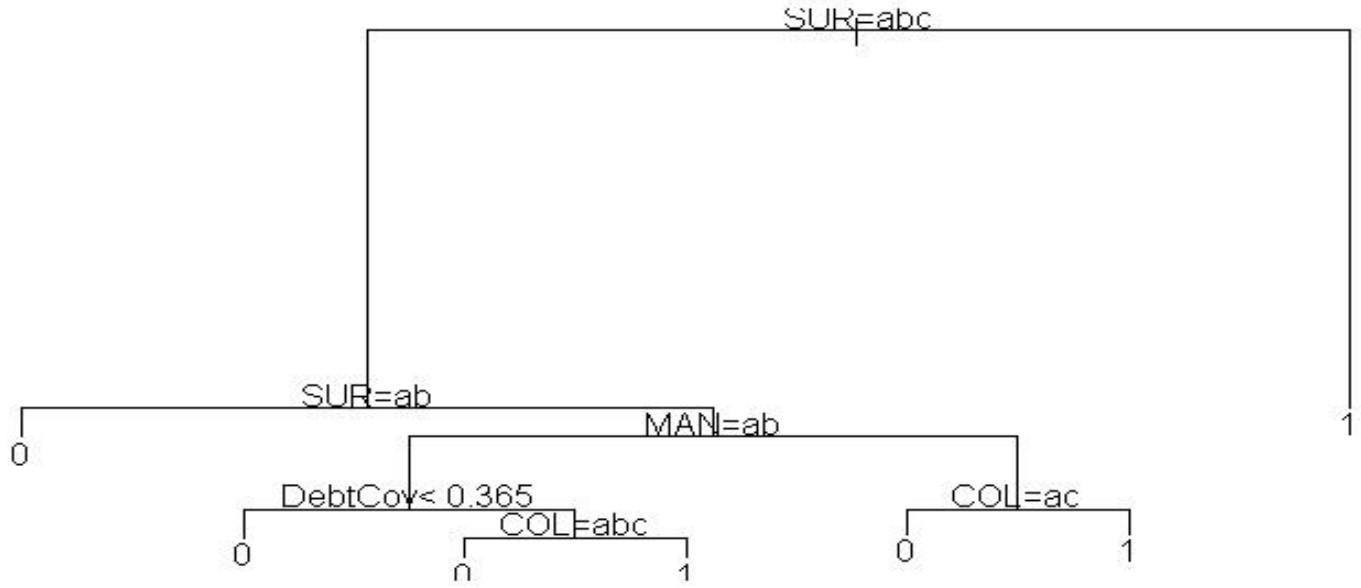
There are no large differences between coefficients estimated by the three models.
Given the fact that we suppose that the bank is implementing for the first time an internal rating system for capital requirements determination, will choose to use an easier method, with less variables, which will be updated later using Bayesian regression models.

Coefficients	Logistic regression	Bayesian with informative prior	Bayesian with non-informative prior
(Intercept)	-6.98801	-6.92414	-7.90960
ACTIVITYconstr	0.97437	0.97462	0.98029
ACTIVITYind	0.37226	0.39007	0.37676
ACTIVITYserv	-0.08430	-0.14381	-0.16041
CIF2	0.77114	0.70642	0.71375
CIF3	0.65239	0.58174	0.58027
CIF4	0.95616	0.98763	1.00776
CIF5	1.68458	1.74127	1.74453
MAN2	1.19441	1.25405	1.24132
MAN3	2.48681	2.65541	2.64991
MAN4	3.00168	2.90678	2.83776
MAN5	2.23170	1.82269	1.70704
STA2	0.18042	0.22817	0.28778
STA3	0.83533	0.78914	0.86895
STA4	1.59791	1.88733	2.07490
STA5	2.19834	2.73757	2.94594
PIA2	0.11876	-0.05275	-0.01260
PIA3	0.65873	0.45740	0.50940
PIA4	1.29943	1.14845	1.17869
PIA5	0.23092	-0.06836	-0.07016
COL1	-0.57600	-0.42826	-0.41274
COL2	1.09376	1.13197	1.40301
COL3	1.99379	2.11218	2.38540
COL4	2.36813	2.41198	2.71316
COL5	3.15494	3.28078	3.62850
CurrLiq	-0.03377	-0.04036	-0.04553
DebtCov	1.80633	1.88540	1.88904
GDPgr	-0.11951	-0.11471	0.12072

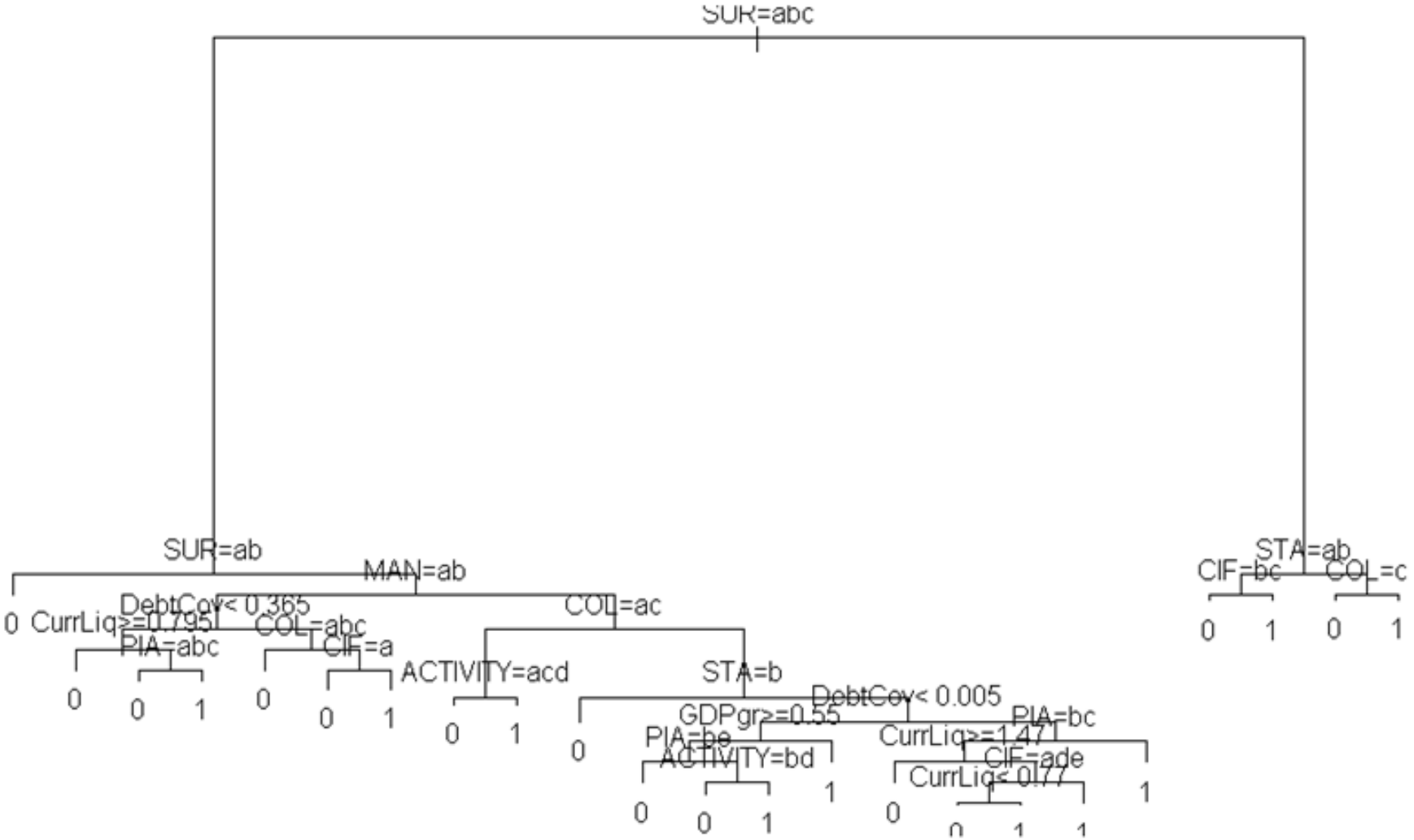
Classification and Regression Trees

- For our dataset, a two different classification and regression trees were performed: one using a prior – the logistic regression estimated earlier, and one without prior.

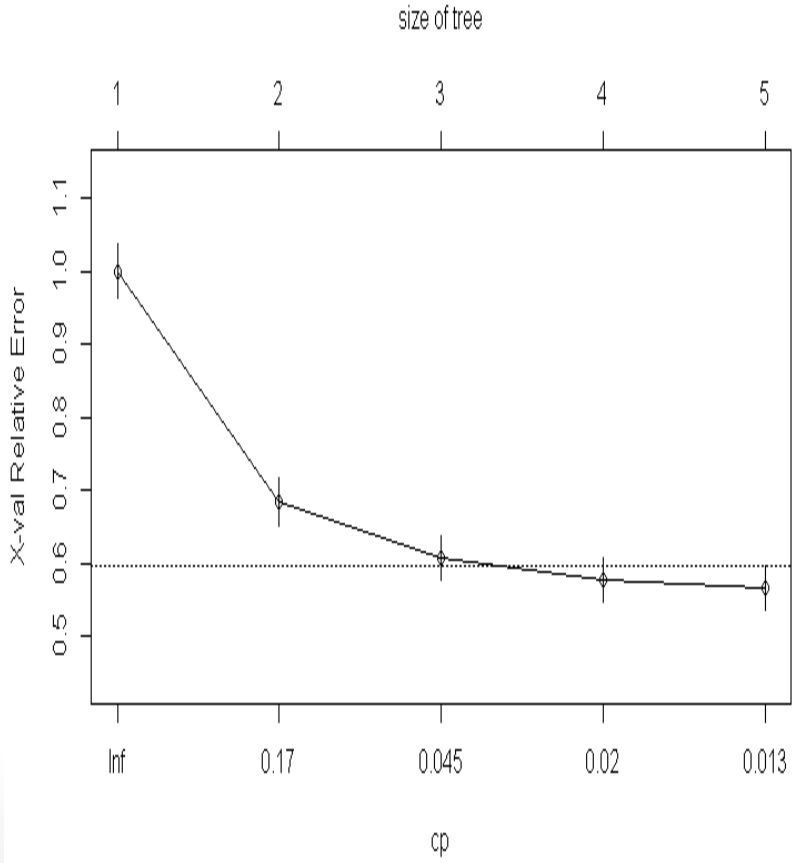
Classification Tree without prior



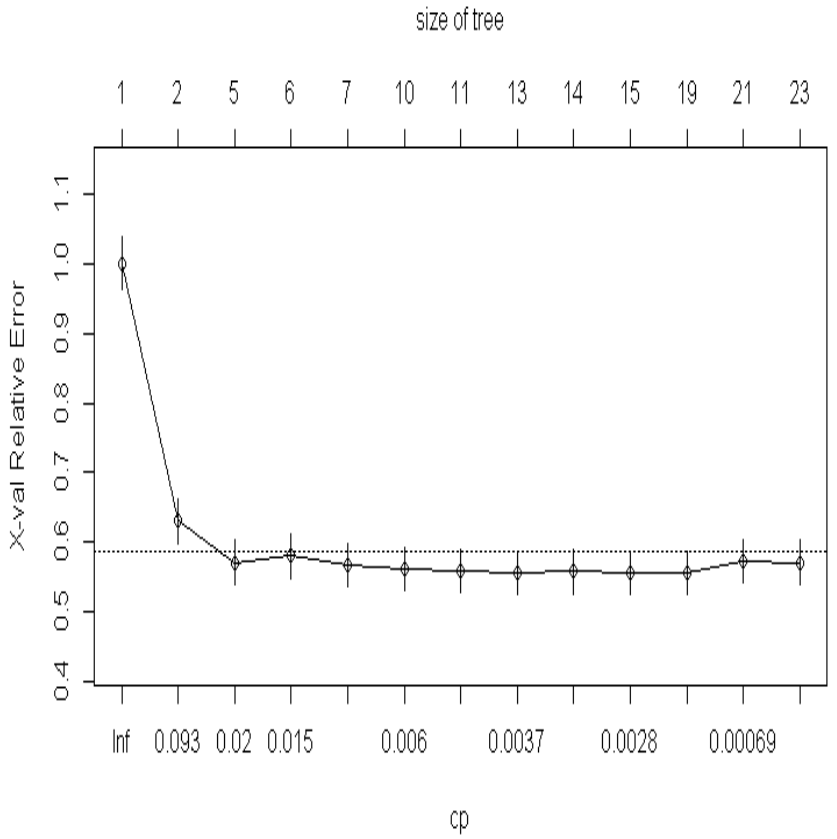
Classification Tree with prior



Comparing and Complexity out of Sample Error for Classification Trees

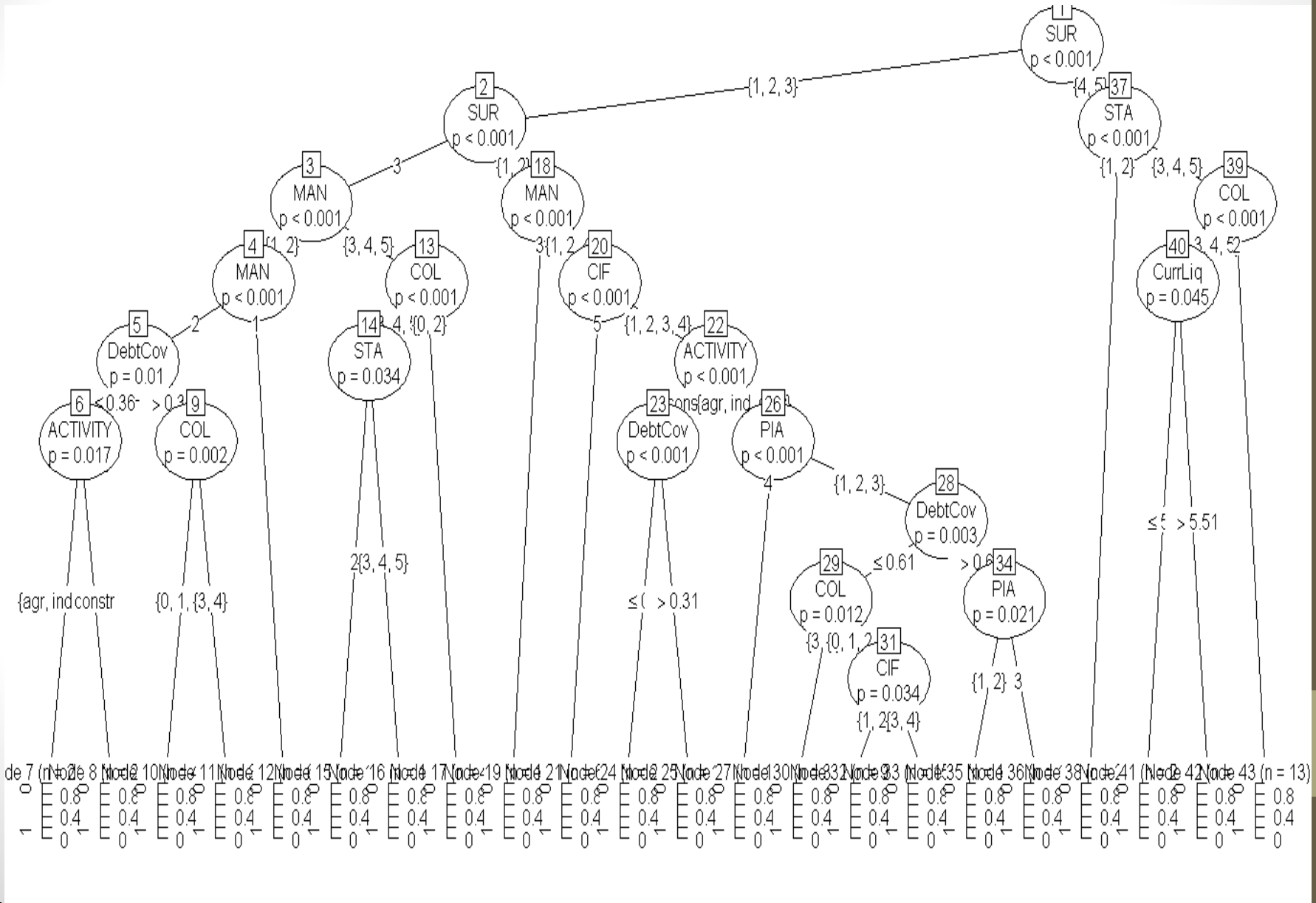


Complexity and out of Sample Error plot for tree without prior

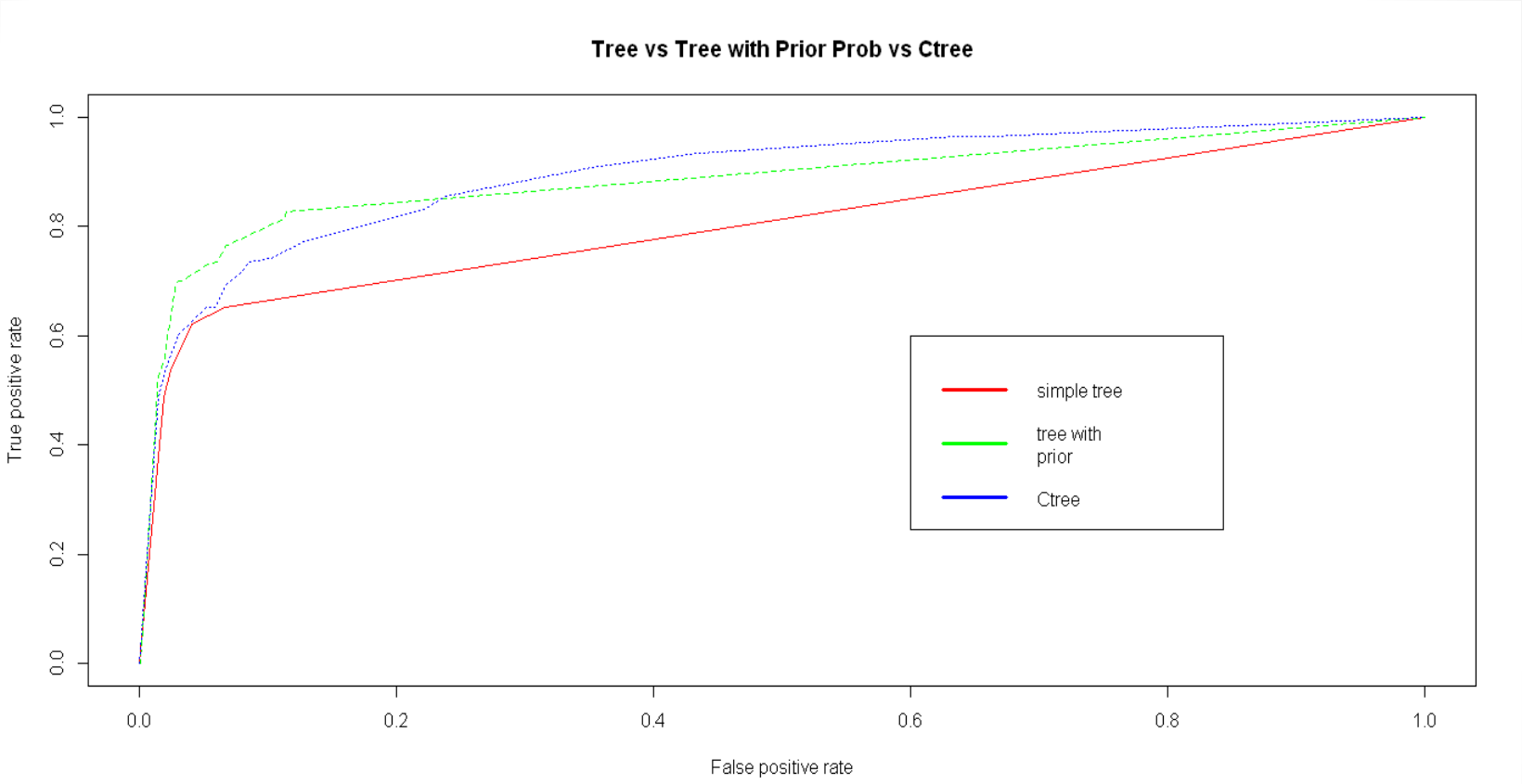


Complexity and out of Sample Error plot for tree with prior

Conditional inference Trees



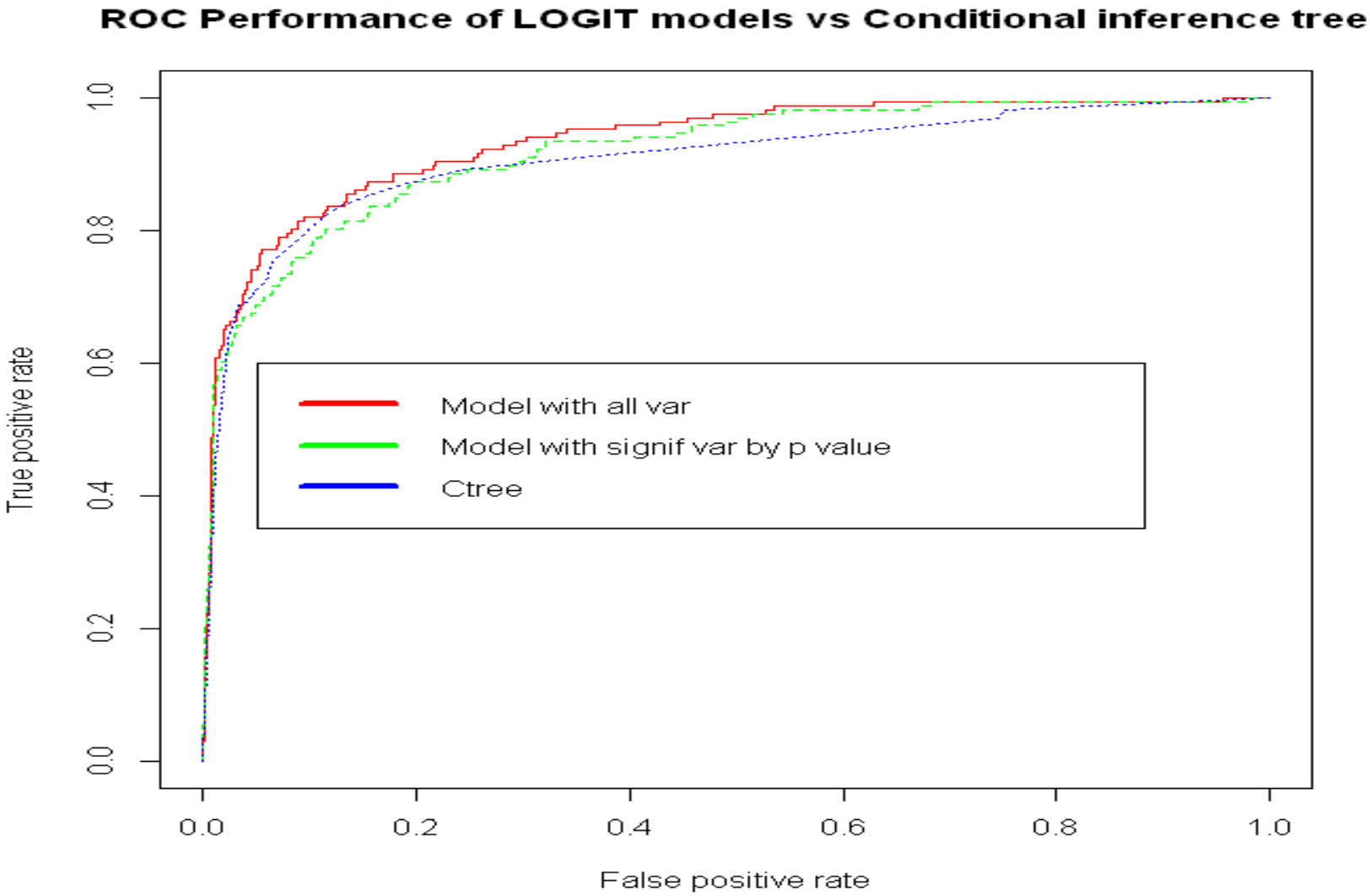
Compare ROC Performance of Trees



Model	AUROC	AR	KS Statistic
Ctree	**0.8942187**	**79%**	**0.6495403**
Tree with prior	0.8855145	**77%**	0.7114353
Simple tree	0.8039375	**61%**	0.584487

Conclusions

Comparing logistic regression model and conditional inference tree, we can see that logistic model is performing better than the conditional tree



Conclusions

- On the specific database analysed, we see that logistic regression method is performing the best, as shown by the ROC curve, AUROC, Accuracy Ratio and KS Statistic.
- The values of the discriminatory performances for logistic regression method are very close, so if we consider the effort needed by a bank to collect accurate data in order to use it within the credit scoring system, the bank may consider the reduced logistic model.
- The aim of the Bayesian regression model is to help the bank improve logistic regression method by updating the model when new information is available, when logistic regression becomes prior expert information

Conclusions

- The use of prior estimators showed its usefulness also in applying classification trees. We saw how only the use of the prior develops the classification tree with more conditions and also better discriminatory performance.
- The new approaches on credit scoring showed that classification trees are able to assess a significant and easy to use scoring method, which is also performing good on the database analysed.
- Classification and conditional inference trees are also a method to help the bank understand the the data structure and the links between the information provided by the client

Thank you for your attention!

References

- Altman, E., Marco, G., and Varetto, F. (1994). Corporate distress diagnostics: Comparison using linear discriminant analysis and neural networks (the Italian Experience). *Journal of Banking and Finance* 18: 505-529.
- Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory.* John Wiley & Sons, Chichester.
- Biçer, I., Sevis, D., and Bilgiç, T. (2010). Bayesian credit scoring model with integration of expert knowledge and customer data. *Twenty-fourth Mini EURO Conference on Continuous Optimization and Information-Based Technologies in the Financial Sector*, Vilnius Gediminas Technical University Publishing House, Technika, pp. 324–329.
- Altman, E., 1968, Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy
- Altman, E., Herbert A. Rijken, 2005, The effects of rating through the cycle on rating stability, rating timeliness and default prediction performance
- Altman, E., Gabriele Sabato, 2005, Modeling Credit Risk for SMEs: Evidence from the US Market
- James A. Ohlson, 1980, Financial Ratios and the Probabilistic prediction of Bankruptcy
- Sjur Westgaard, Nico van der Wijst, 2000, Default probabilities in a corporate bank portflio – a Logit model approach
- Claudia Czado, Carolin Pfluger, 2007, Modeling dependencies between rating categories and their effects on prediction in a credit risk portfolio
- Alexander J. McNeil, Jonathan P. Wendin, 2006, Bayesian inference for generalized linear mixed models of portfolio credit risk
- Elisabeth Van Laere, Bart Baesens, 2010, The development of a simple and intuitive rating system under Solvency II

References

- Sharma, Guide to Credit Scoring in R, http://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf
- Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*, Series A 160: 523-541.
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika* 57: 97–109.
- Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1: 145-168.
- Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* 2nd Edition. John Wiley & Sons, Inc., New York.
- Mira, A. and Tenconi, P. (2004). Bayesian estimate of credit risk via MCMC with delayed rejection. In: *Seminar on Stochastic Analysis, Random Fields and Applications IV.* Centro Stefano Franscini, Ascona, pp. 277-291. Birkhauser Verlag, Basel.
- Mok, J-M. (2009). *Reject Inference in Credit Scoring.* Available from http://www.few.vu.nl/en/Images/werkstuk-mok_tcm39-91398.pdf.
- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* 2nd Edition. Springer-Verlag, New York.
- Robert, C.P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R.* Springer-Verlag, New York.
- Thomas, L.C. (2009). *Consumer Credit Models*: *Pricing, Profit and Portfolios.* Oxford University Press, Oxford.