Analysis of German Credit Data

Jan Kasperek & Kevin Nyamai

Seminar of Applied Statistics

Université de Neuchâtel

Purpose of this Project

- The goal of this project is to set up a model to classify credit applicants according to their credit risk (good vs bad), based on a dataset of applicants in Germany.
- The end goal is to maximize profit from the bank's perspective hence the need for a decision rule for loan approval.

Business Understanding Phase

- Understanding of Target/Response Variable:
  - Credit Risk- The probability of a financial loss resulting from a borrower's failure to repay a loan
    - Good credit risk - An applicant with a bad credit risk is likely to repay the loan within time
    - Bad credit risk - An applicant with a bad credit risk is not likely to repay the loan within time
- Determinants of Credit Risk : Demographic and socio-economic profile
- Profit maximization : cost reduction and prediction accuracy.

Data Understanding Phase

- Data Description
  - Raw dataset contains 1000 observations across 32 variables with no missing values
    - 1 ID variable
    - Binary response variable: (Good (700 observations) vs Bad credit (300 observations)
    - 5 numerical predictors, from which 3 can be considered continuous (Age, credit amount, duration of credit in months)
    - 25 categorical predictors

Data Understanding Phase

- The predictors can broadly be classified under the "**5 Cs**" described in credit management in addition to **demographic** factors as below:

Character

This is an estimation of the applicant's general financial trustworthiness and credibility.

Capacity

This is the ability to repay a loan assessed by cash flow and debt to income ratio

Collateral

This is an asset owned by the borrower that can be used to serve as security loan in the event of failing to repay the loan.

Conditions

These are the circumstances under which the loan is to be issued.

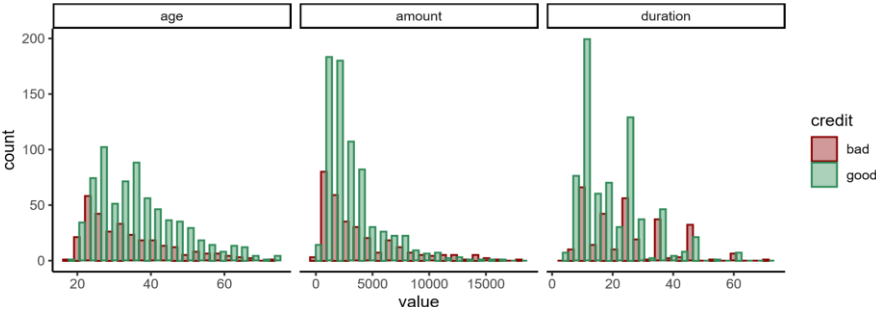
Capital

These are the circumstances under which the loan is to be issued.

The demographic factors include, among others, information regarding age, marriage/relationship status and whether the applicant owns a house or rents.

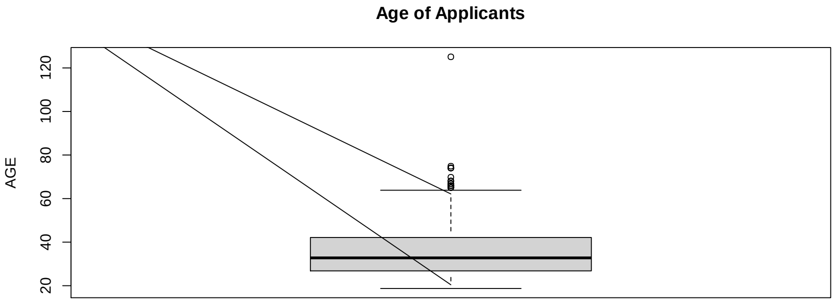
Data Preparation Phase

- 3 predictors can be considered continuous i.e., age, credit amount and duration of credit in months. The predictor *amount* has a highly skewed distribution with a long right tail.



Data Preparation Phase

- 3 implausible observations are removed
  - One where age is 125 years
  - *Education* , where factor level is '-1' which not defined in the data description
  - *Guarantor*, where factor level is '2' , which is not defined in the data description



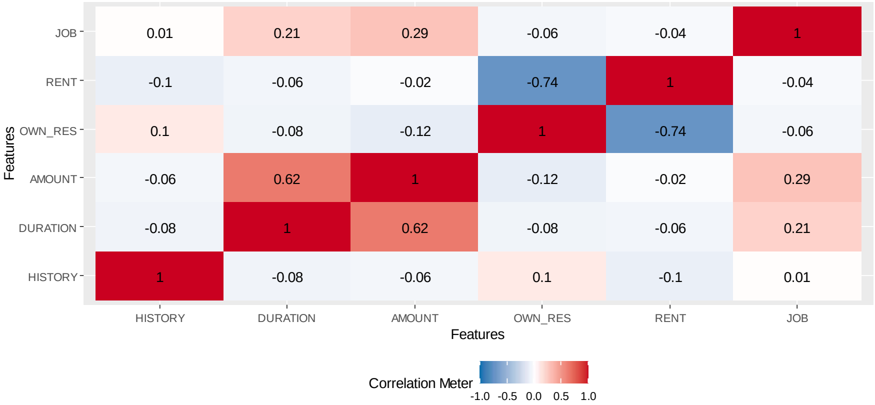
Data Preparation

- The variable *foreign* (whether an applicant is a foreign worker) is deleted because:
  - There are only a few cases (3.6%)
  - Moreover, determining applicants credit risk using their nationality could be illegal (or become illegal in the future) due to discrimination

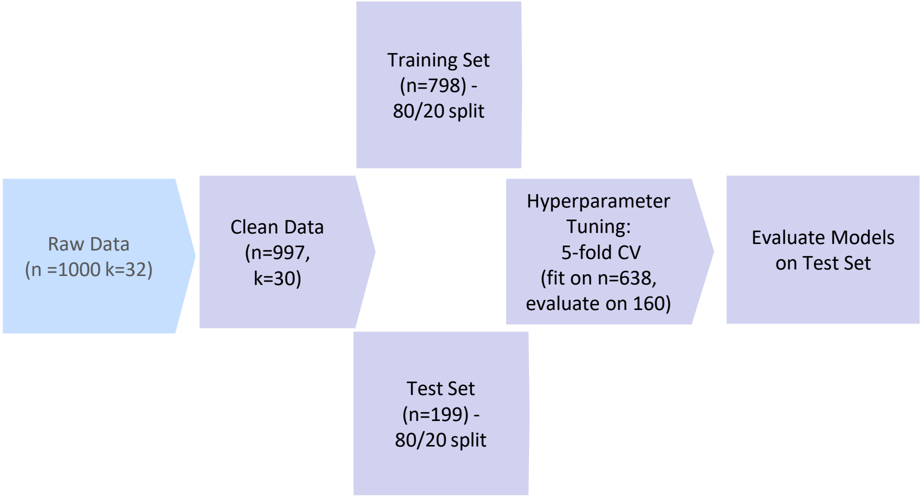
- The data kept as raw as possible:
  - Avoiding many assumptions on predictors
  - All models considered can do variable selection themselves.
  - Assumption of no data leakage issues with the dataset.

Data Preparation

- High correlations not a problem for modelling:
  - There are no perfectly collinear predictors in the dataset due to always leaving one category in decomposing categorical variables into dummies.
  - Only a few variables have |r| > 0.5: duration of credit with credit amount (r = 0.62) and whether the applicant rents with whether he owns a residence (r = −0.74)



Modelling Phase



Modelling Phase- Model Choice

- Models chosen are:
  - Elastic Net Logistic Regression (*ElasticNetLogit*)
  - Random Forests
  - Bayesian Additive Regression Trees (*BART*)
  - Multivariate Adaptive Regression Splines (*MARS*)

- Choice of Model due to nature of the dataset, computational considerations and implementation in *R*
  - The data is not very big but has a high number of variables, many of which binary.
  - Considering very complex models for this dataset, the risk of overfitting is high, and these models would be sensitive to small changes in the hyperparameters.
  - The models chosen work well without extensive tuning
  - This keeps computational complexity at a tolerable level
  - Models are implemented in the R package *tidymodels*

Modelling Phase- Model Choice

Elastic Net Logistic Regression (Zou and Hastie 2005)

- Penalized logistic regression with a mixture of the Ridge and LASSO penalties

- Can also reduce to pure Ridge or pure LASSO, depending on the mixture coefficient

- Predicted probabilities obtained the same way as in standard logistic regression

Random Forests (Breiman 2001)

- Ensembles of decision trees fitted on different bootstrap samples.

- Avoids the overfitting problem of single decision trees

- Known for high prediction accuracy while still having a tolerable amount of model complexity.

- Predicted probabilities can be obtained by averaging the class predictions across all trees.

Modelling Phase- Models

BART
(Chipman, George and McCulloch 2010)

- more complex tree ensemble model

- Main difference to Random Forests: each tree is edited iteratively to fit the yet unexplained variation in the target variable (similar to boosting).

- the way in which the single trees are changed over the iterations is regularized by a Bayesian model, the prior of which is determined by tuning parameters.

- To obtain predictions from each tree in the ensemble, the trees gotten after burn-in are averaged

MARS
(Friedman 2001)

- Non-parametric regression spline model which builds piecewise linear functions.

- Constructed in a stepwise fashion and includes a pruning procedure to avoid overfitting.

- Tries all possible products between the predictors up to some degree *d*, where *d* = 1 yields a purely additive model.

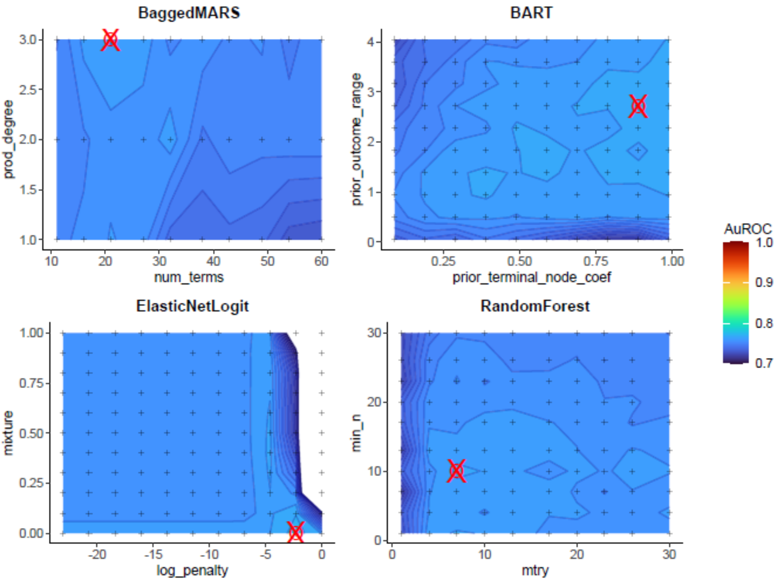
- To get predicted probabilities, the model is passed through a logistic regression

- Additionally, we use Bagging to reduce variance

Modelling Phase- Hyperparameter Tuning

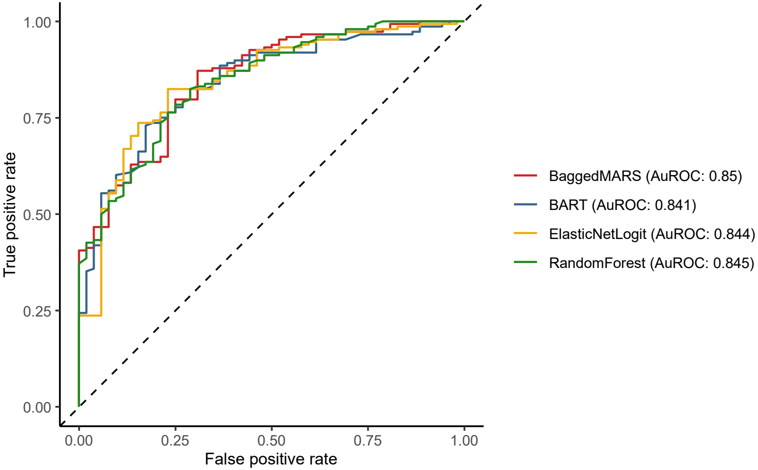
model	hyperparameter	explanation	range	package
BaggedMARS	*num_terms*	max. number of terms before pruning	1-3	*earth*
	prod_degree	max. degree of interaction	10-60	
BART	*prior_terminal_node_coef*	prior on probability that tree node is terminal	0.09-0.99	*dbarts*
	prior_outcome_range	prior on range of predicted outcomes	0.05-4.05	
ElasticNetLogit	*mixture*	mixture of Ridge (0) and LASSO (1) penalty	0-1	*glmnet*
	log_penalty	Natural Log of regularization parameter	(-25)-0	
RandomForest	*mtry*	# of predictors considered for tree splits	1-30	*randomForest*
	min_n	min. # of observations in terminal nodes	1-30	

Model Evaluation Phase

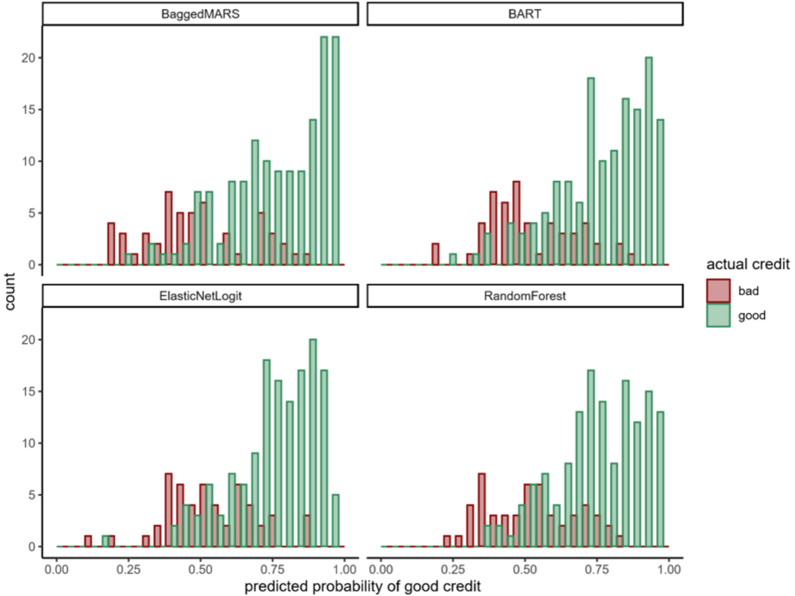


For better visibility, surface areas with an AuROC < 0.7 are not displayed.

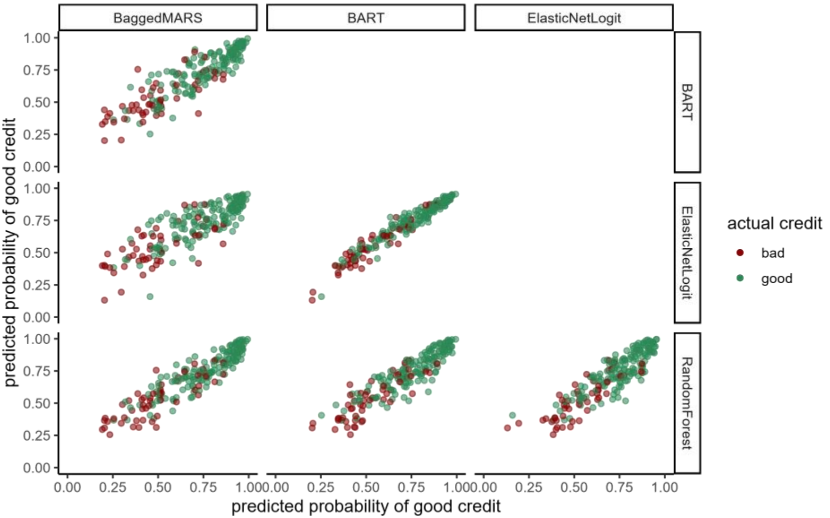
Model Evaluation Phase- ROC Curves



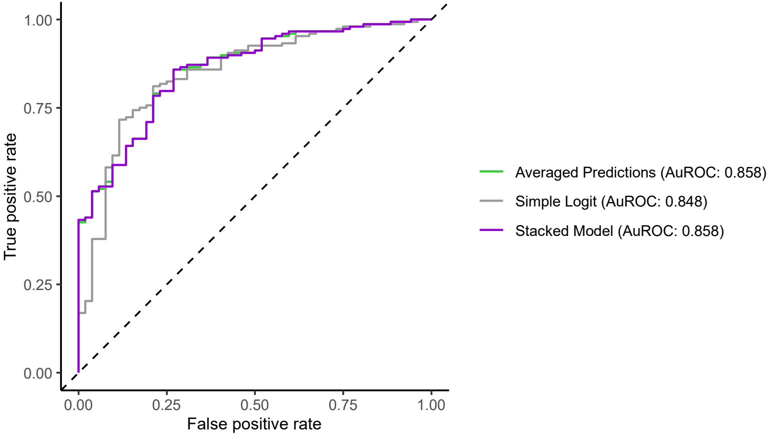
Model Evaluation Phase- Distributions of predictions on test set



Model Evaluation Phase- Predictions on Test Set



Modeling Alternatives - ROC curves of ensemble models and simple logistic regression



Deployment Phase

- Elastic Net Logistic chosen based on:
  - High predictive accuracy and easy implementation.
  - Lowest running time
  - Lower variance than simple logistic regression
  - Alternative: Model Averaging and Stacking (slightly better, but high computational load)

- Profit maximization and cost evaluation.
  - Model eases manual intensive work of checking individual applications.
  - Mistaking applicants with good credit for bad credit is not as costly as misclassifying individuals with truly bad credit
  - The cutoff for bad credit can be set more leniently while high for good credit
  - Updating model (e.g. yearly) to avoid **model drift**.

Deployment Phase

- Deployment:
  - Model as a prediction tool may not be necessary to assess applicant's credit worthiness.

  - Used congruently with Schufa (German: *Schutzgemeinschaft für allgemeine Kreditsicherung*) score
    - The score measures the probability with which an individual honours their bill, credits and contracts.
    - The score however does not encompass other socio-economic and demographic factors hence can not be used as standalone tool as it is only based on credit history and bill payments
    - if the costs of using the SCHUFA system exceed the costs of building, using and maintaining our model, our model could be useful.

References

Breiman, Leo. 2001. “Random Forests.” Machine Learning 45: 5–32.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. “BART: Bayesian additive regression trees.” The Annals of Applied Statistics 4 (1): 266–98.

Friedman, Jerome H. 1991. “Multivariate Adaptive Regression Splines.” The Annals of Statistics 19 (1): 1–67. 10

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. “A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis.” Journal of Clinical Epidemiology 49 (12): 1373–79.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” Journal of the Royal Statistical Society 67 (2): 301–20.