NATIONAL RESEARCH UNIVERSITY

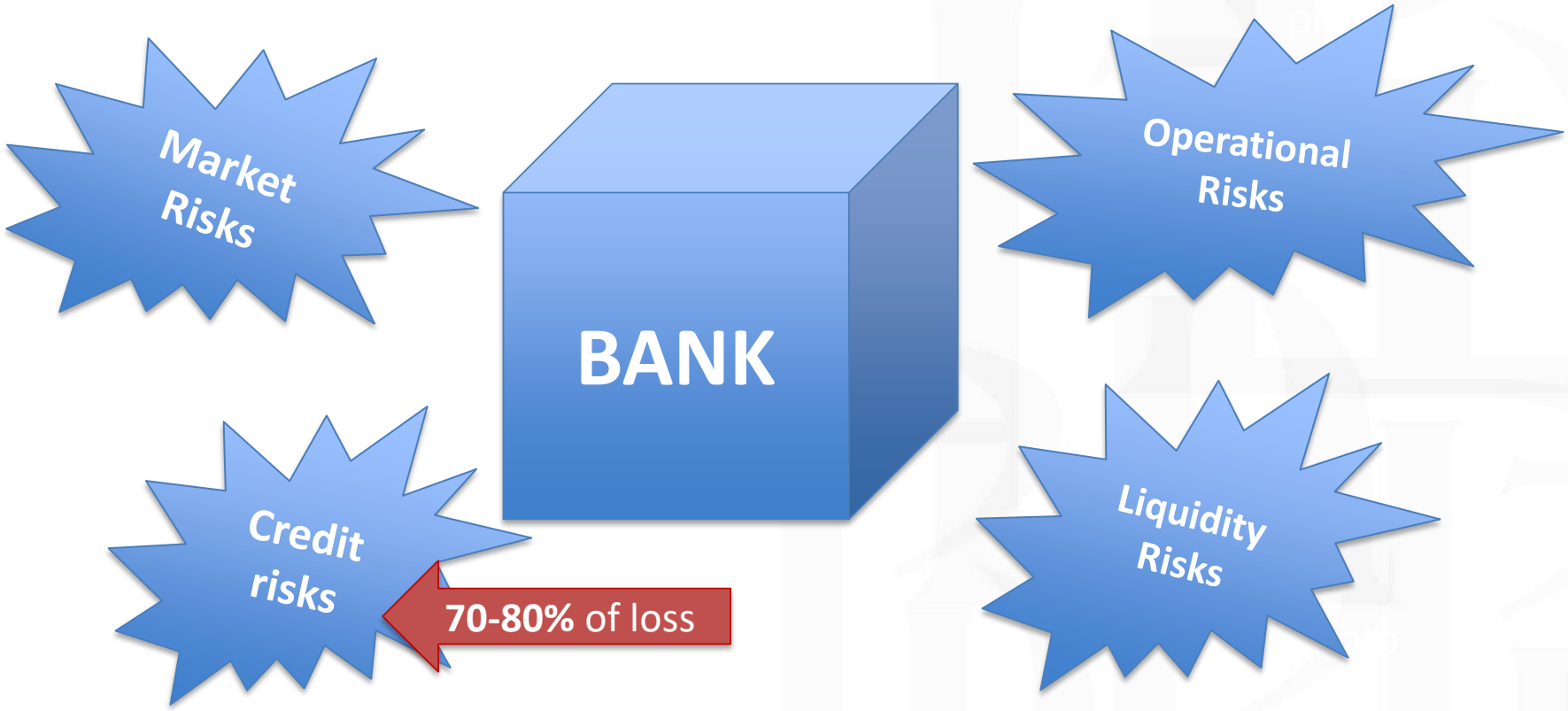
Query-based classification with interval pattern structures: application to credit scoring

S.Kuznetsov, Y.Kashnitskiy, A.Masyutin
School of Computer Sciences

Risk Management in Retail Banking

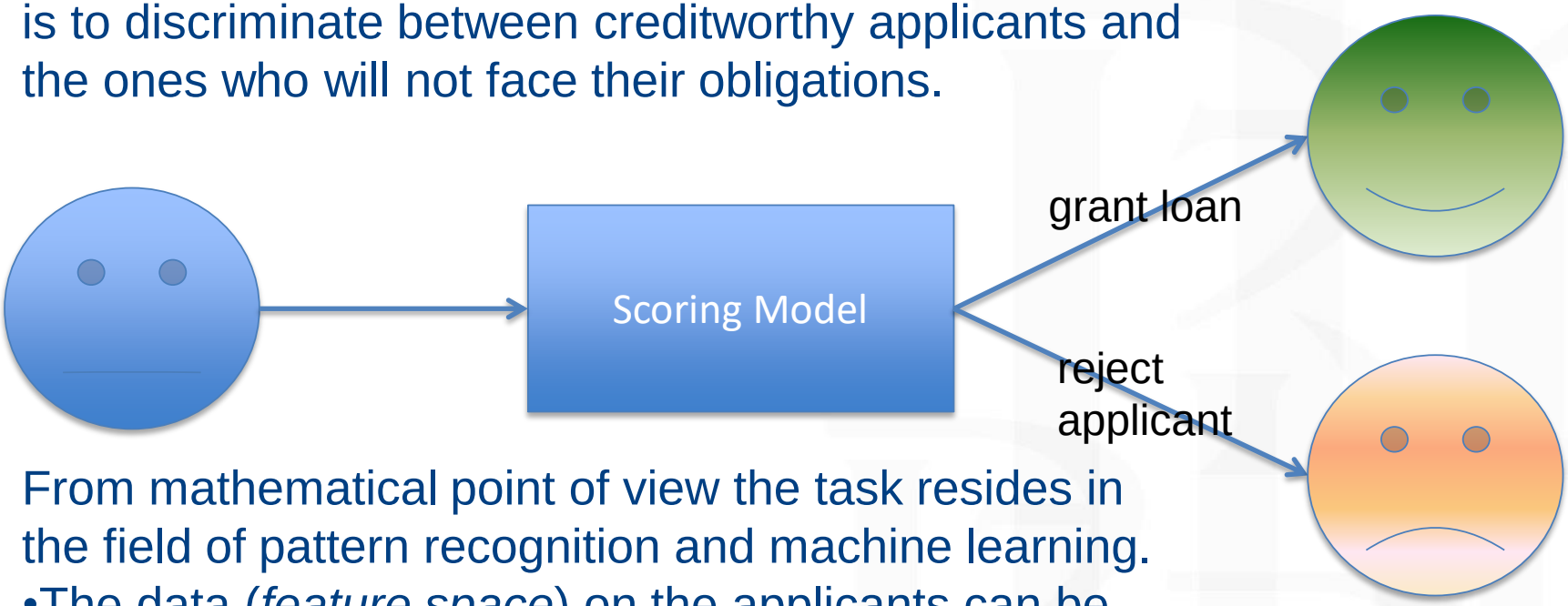
Banks and credit organizations take risks in order to make profit.

- The higher the risk the higher is return.
- The balance between the risks and return is necessary for staying afloat.



Scoring is a Key Tool to Manage Credit Risk in Retail Banking

Classification Task in Credit Scoring
is to discriminate between creditworthy applicants and the ones who will not face their obligations.



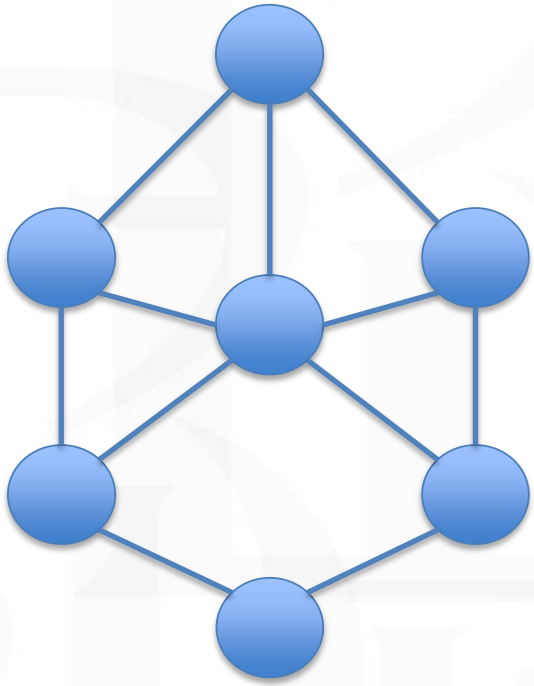
From mathematical point of view the task resides in the field of pattern recognition and machine learning.

- The data (*feature space*) on the applicants can be represented by a numerical context.
- The *target attribute* is a flag whether the borrower defaulted on his/her obligations or not.

FCA in Classification Tasks

Formal Concept Analysis
provides a framework to represent the relations between objects and their attributes.
There are several algorithms to use this framework such as JSM classification algorithm:

$$(\bigvee_{i=1}^{p}[\delta(g_{test}) \sqsubseteq h_i^+]) \wedge \neg(\bigvee_{j=1}^{n}[\delta(g_{test}) \sqsubseteq h_j^-])$$



where h is hypothesis and g_{test} is unlabeled object, $\delta(g_{test})$ is its description.
Basically, the classification can be based on several *concepts* that are relevant for the unlabeled object.
The concept-based learning model for standard object-attribute representation (i.e., formal contexts) is naturally extended to pattern structures.

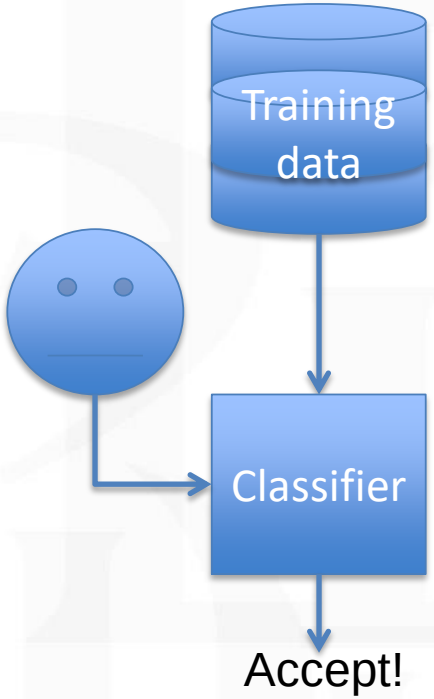
Query-based Classification versus Training Based Methods

Training Based Methods

Learning Step:The training data is used to construct a model which relates the feature variables.

Test Step:The trained model is used to predict the class variable for test instances.

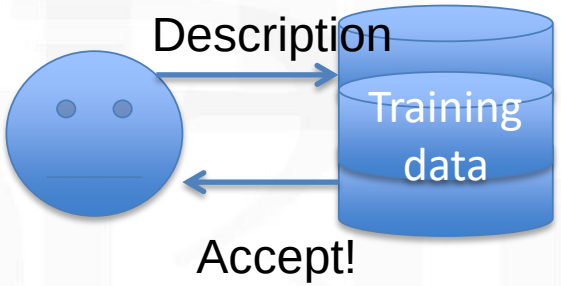
In case of pattern structures the learning step implies the full lattice construction, which can be NP-hard in general case.



Query based Classification

No Learning Step: the test object is processed online in a kNN manner. No prior model is designed.

Test Step: the test object is classified based on its description.



Definitions

Let us give several definitions:

Let G be a set (of objects), let $(\mathrm{D}, \sqcap)$ be a meet-semi-lattice (of all possible object descriptions) and let δ: $\mathrm{G} \to \mathrm{D}$ be a mapping. Then $(\mathrm{G}, \underline{D}, \delta)$, where $\underline{D} = (\mathrm{D}, \sqcap)$, is called a *pattern structure*

A pattern structure gives rise to the following derivation operator:

$$A^\diamond = \prod_{g \in A} \delta(g) \qquad \text{for } A \in G,$$

$$d^\diamond = \{g \in G \mid d \sqsubseteq \delta(g)\} \qquad \text{for } d \in (D, \sqcap).$$

Instead of strict notion of hypothesis we introduce alpha-weak premises:

A pattern $c \in \mathrm{D}$ is an
α - weak positive premise (classifier) iff:

$$\frac{||c^\diamond \cap G_-||}{||G_-||} \leq \alpha \; and \; \exists A \subseteq G_+ : c \sqsubseteq A^\diamond$$

Voting Schemes Based on Premises

The classification of a test object is based on a voting scheme among premises. In most general case voting scheme F is a mapping:

$$F(g_{test}, h_1^+, ..., h_p^+, h_1^-, ..., h_n^-) \rightarrow [-1, 1, \emptyset]$$

where g_{test} is the test object with unknown class, h_i^+ is a positive premise $\forall i = \overline{1, p}$ and h_j^- is a negative premise $\forall j = \overline{1, n}$, -1 is a label for negative class, and 1 is a label for positive class (i.e. defaulters).

The voting scheme is built upon weighting function $\omega()$, aggregation operator $A()$ and comparing operator $\otimes$:

$$F(\omega(\cdot), A(\cdot), \otimes) =$$
$$= (A_{i=1}^{p}[\omega(h_i^+)]) \otimes (A_{j=1}^{n}[\omega(h_j^-)])$$

Query Based Classification Algorithm

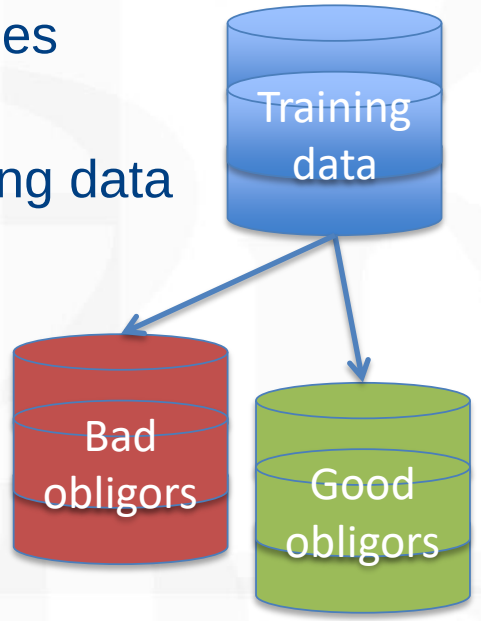
The algorithm is designed to be run iteratively and includes randomization.

Inputs: positive and negative contexts (two parts of training data with defaulters and non-defaulters).

Voting scheme:

$$F(g_{test}, h_1^+, ..., h_p^+, h_1^-, ..., h_n^-) =$$

$$= (\sum_{i=1}^{p}[\delta(g_{test}) \sqsubseteq h_i^+]) \otimes (\sum_{j=1}^{n}[\delta(g_{test}) \sqsubseteq h_j^-])$$



There are three parameters within the procedure:

- *Number of iterations*
- *Subsample size* - percentage of the context randomly used for intersection with the test object (parameter)
- *Alpha-threshold* - is the maximum allowable percentage of the opposite context for that the premise is not falsified

Query Based Classification Algorithm

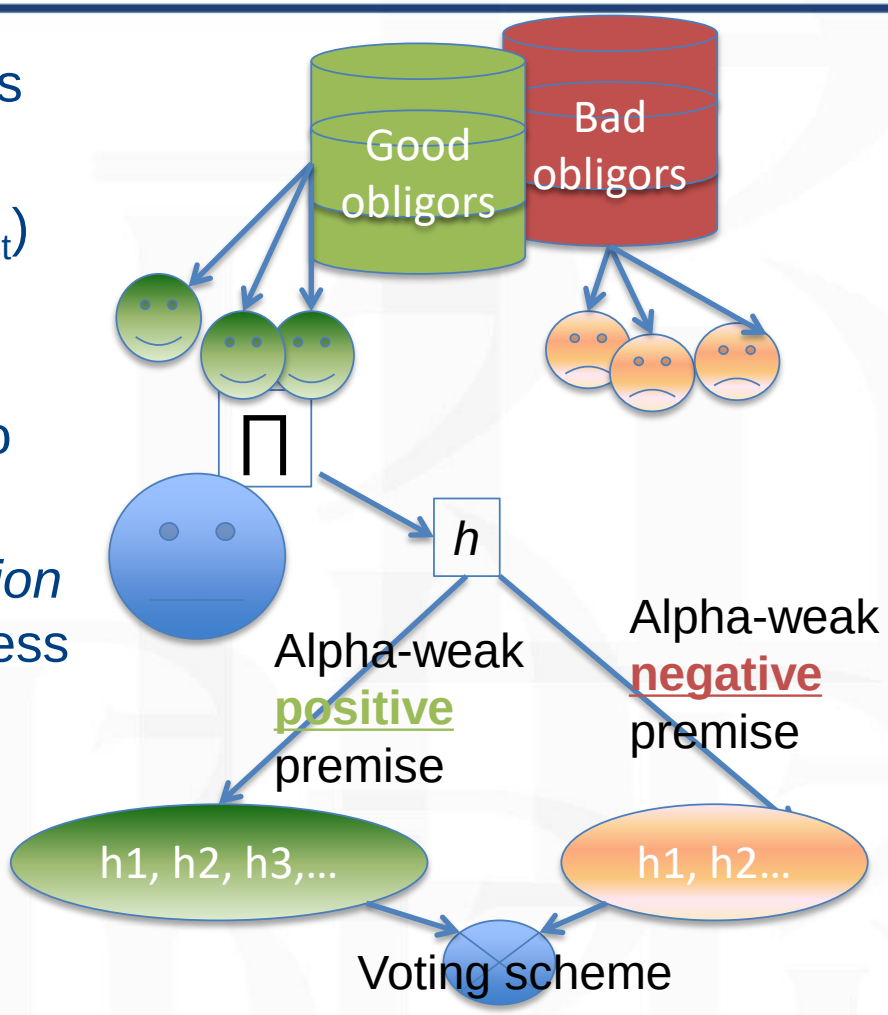
Step 1. Extract *subsample size* of objects from the positive (negative) context.

Step 2. Intersect with the test object (g_{test}) and get a premise *h.*

Step 3. If *h* is an *alpha-threshold*-weak positive (negative) premise then add *h* to positive (negative) premises set.

Repeat Step 1 – Step 3 *number of iteration* times. These steps can be called a process of premises mining.

Final step. Having acquired the sets of positive and negative premises, use the voting scheme in order to produce the prediction (or margin).



Data and Experiments Results

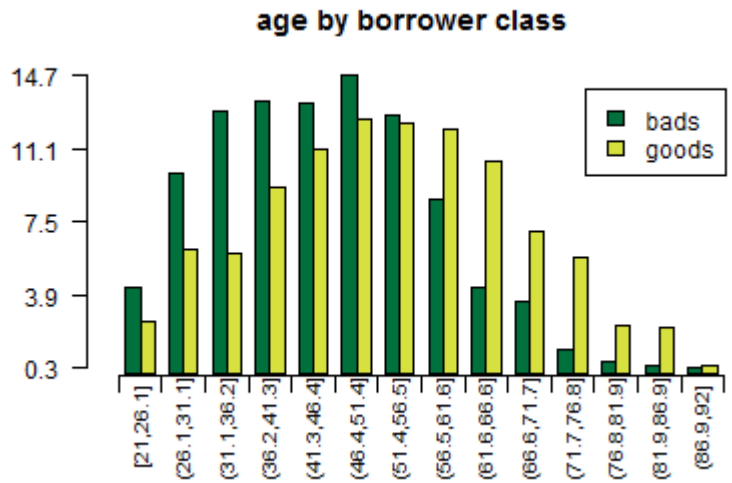
- Open dataset devoted to the credit scoring. We considered the “Give Me Some Credit” contest held in 2012. (https://www.kaggle.com/c/GiveMeSomeCredit). The validation process requires calculation of performance metrics (ROC AUC and Gini coefficient) of the model based on the data sample that was retrieved from the same distribution but was not used to develop the model itself. All observations were randomly extracted from the contest dataset.
- The data represent the customers and their metrics assessed on the date of loan application. Each context consists of 1000 objects. The test dataset consists of 300 objects.
- Feature set represents various metrics such as loan amount, term, rate, payment-to-income ratio, age of the borrower, undocumented-to-documented income, credit history metrics etc.

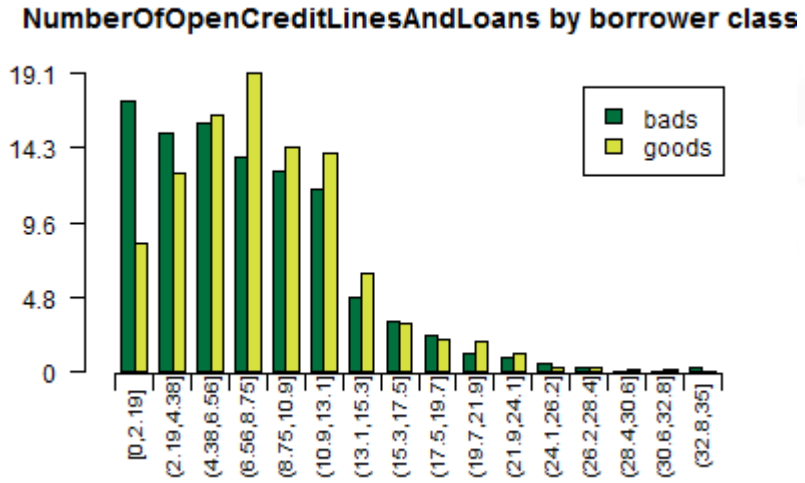
“Give Me Some Credit” contest held in 2012

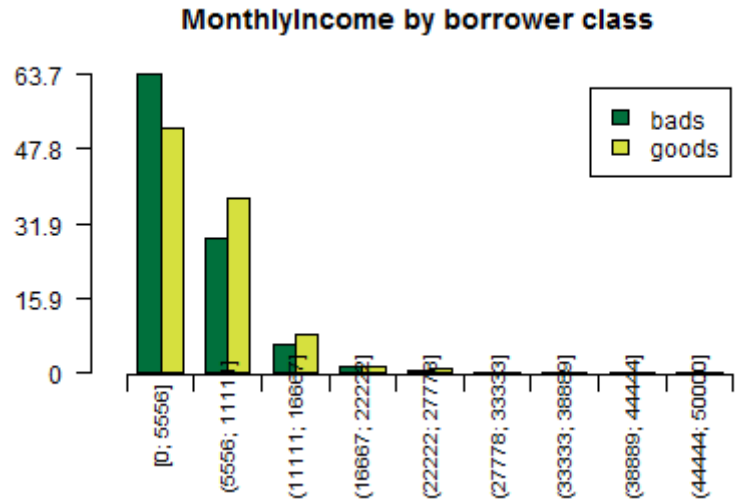
Kaggle Data Description

Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthly gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Examples of one-factor distributions by target class





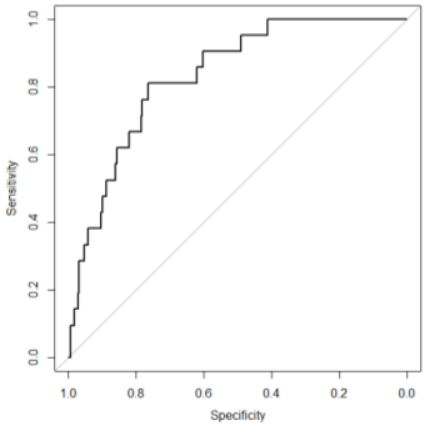


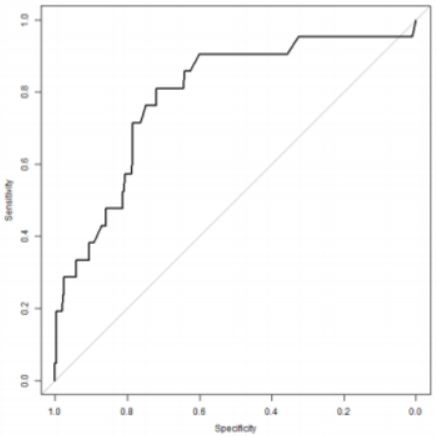
- One factor distributions show that the target variable is mixed within range of factors, although they are promising to make some discrimination between good and bad obligors.

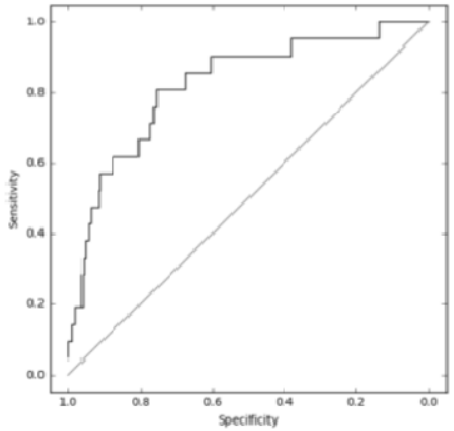
Data and Experiments Results

- According to performed calculations the model with the highest Gini (70.8%) on the validation sample was obtained with XGBoost algorithm.
- Classical scorecard adopted in bank has shown moderate discriminatory power with 58.1% Gini.
- Query based algorithm reached accuracy of 66.3% in terms of Gini

Metric	Query based classification	Scorecard	XGboost
Gini on validation dataset	0.6624	0.5806	0.708







Conclusion

- When dealing with large numerical datasets, query based classification may be preferable to classification based on full set of hypotheses extracted from lattice, since it requires less time and memory resources

- We considered query based classification for Kaggle open data set devoted to problem of credit scoring

- The classification accuracy of the algorithm was compared both to conventionally adopted models used in the bank and to black-box models.

NATIONAL RESEARCH UNIVERSITY

Thank you for your attention!

Appendix (feature importance)

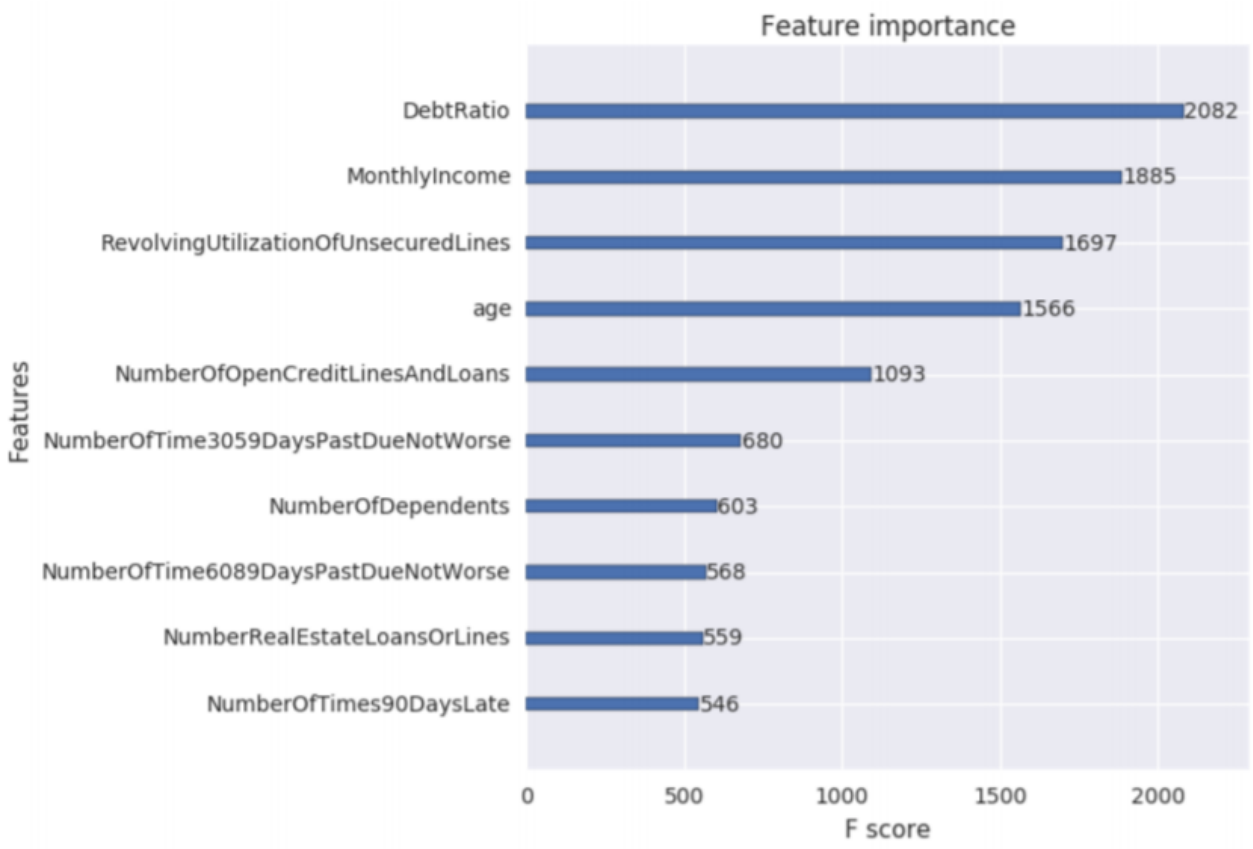


Fig. 4. Feature importance estimated with Xgboost