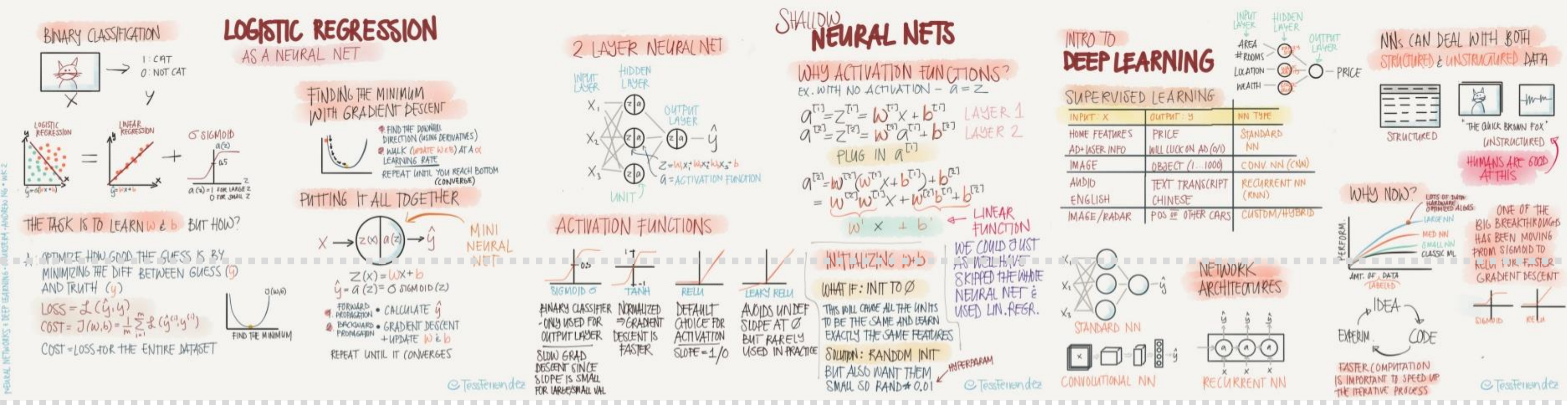


Data Science в скоринге

Что предлагает Science?

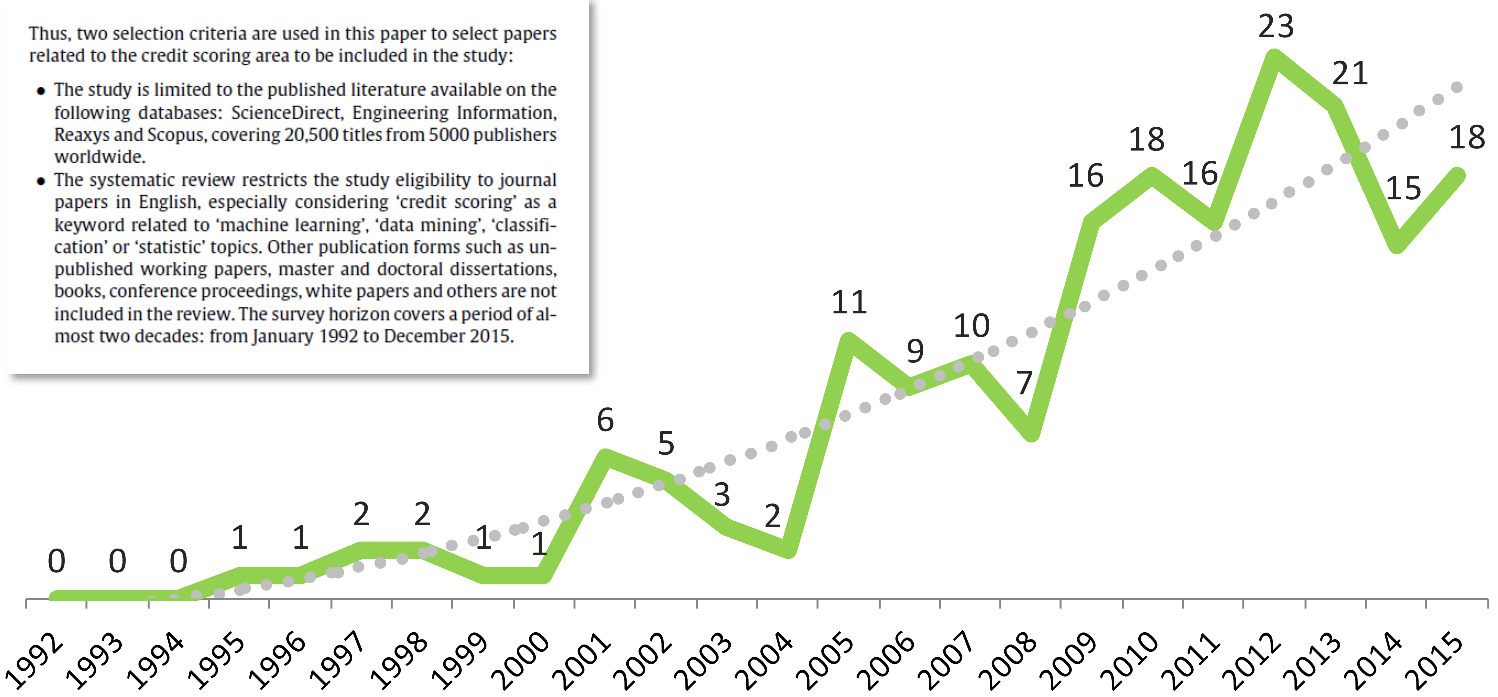
Афанасьев Сергей
КБ «Ренессанс Кредит»

1 августа 2019 г.
Москва

Растет количество статей по скорингу

Thus, two selection criteria are used in this paper to select papers related to the credit scoring area to be included in the study:

- The study is limited to the published literature available on the following databases: ScienceDirect, Engineering Information, Reaxys and Scopus, covering 20,500 titles from 5000 publishers worldwide.
- The systematic review restricts the study eligibility to journal papers in English, especially considering 'credit scoring' as a keyword related to 'machine learning', 'data mining', 'classification' or 'statistic' topics. Other publication forms such as unpublished working papers, master and doctoral dissertations, books, conference proceedings, white papers and others are not included in the review. The survey horizon covers a period of almost two decades: from January 1992 to December 2015.



Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison." Surveys in Operations Research and Management Science 21.2 (2016): 117-134.

Тематики научных статей (1992-2015)

Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison." Surveys in Operations Research and Management Science 21.2 (2016): 117-134.

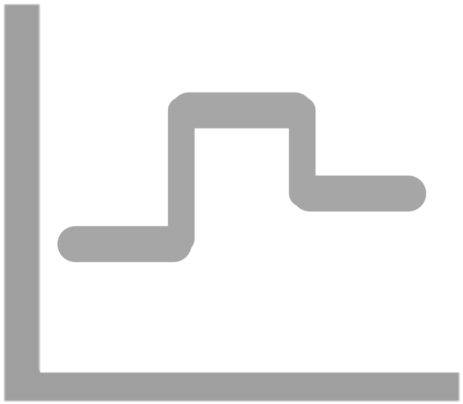
CRISP-DM & ML-Pipeline

Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

Modeling:

- Feature Extraction
- Feature Preprocessing
- Feature Selection
- Model Selection
- Hyper-Parameter Tuning
- Model Validation

1. Биннинг переменных



Биннинг в кредитном скоринге

Задача кредитного скоринга:

- x_i — заемщики
- $y_i \in \{-1(bad), +1(good)\}$

Бинаризация признаков $f_i(x)$:

$$b_{jk}(x) = [f_j(x) \in D_{jk}]$$

Возраст	до 25	**5**
	25-40	**10**
	40-50	**15**
	50 и больше	**10**
Собственность	владелец	**20**
	совладелец	**15**
	съемщик	**10**
	другое	**5**
Работа	руководитель	**15**
	менеджер	**10**
	служащий	**5**
	другое	**0**
Стаж	1/безработный	**0**
	1...3	**5**
	3...10	**10**
	10 и более	**15**

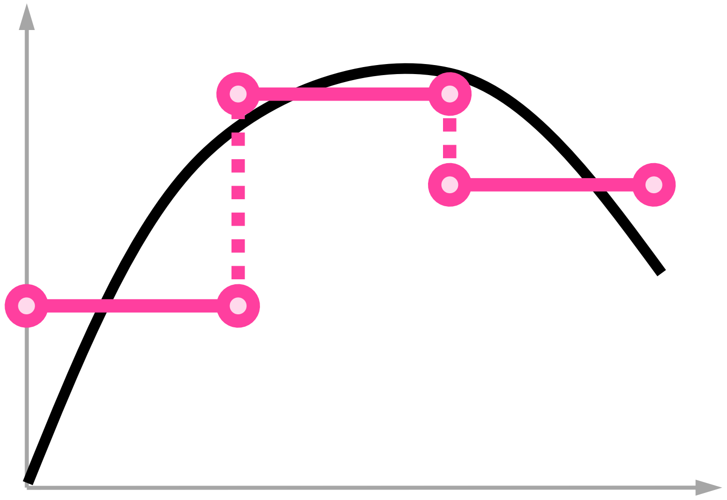
Coursera, К.В. Воронцов

Зачем нужен биннинг?

Интерпретируется и прост в применении

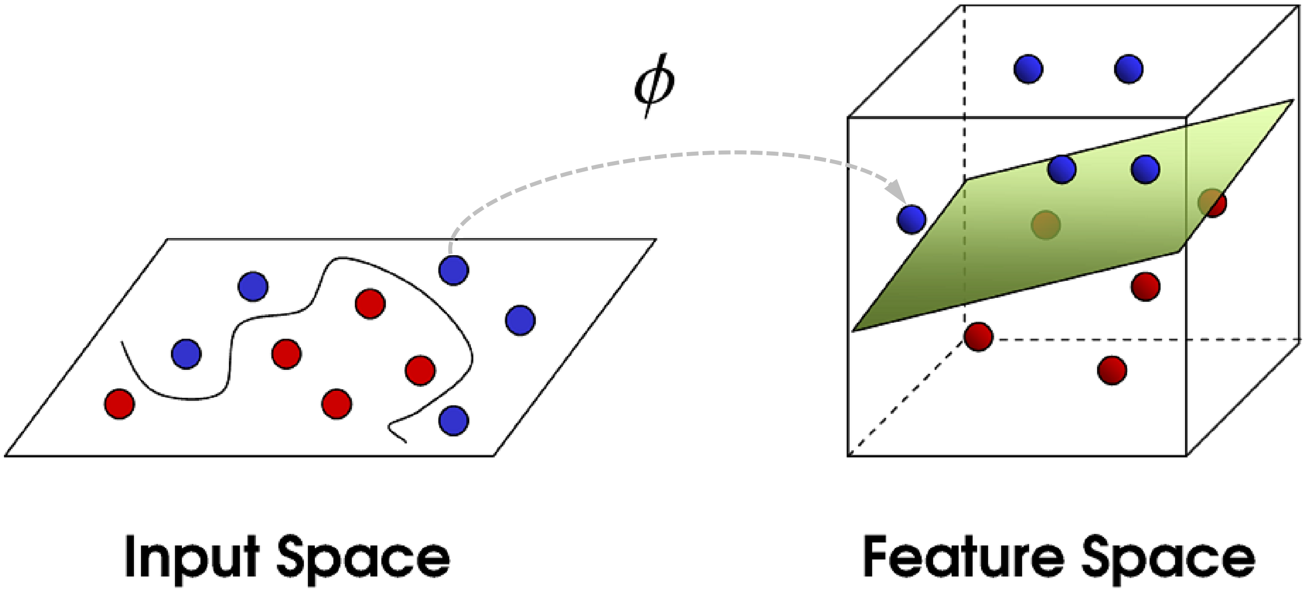
Возраст	до 25	**5**
	25-40	**10**
	40-50	**15**
	50 и больше	**10**
Собственность	владелец	**20**
	совладелец	**15**
	съемщик	**10**
	другое	**5**
Работа	руководитель	**15**
	менеджер	**10**
	служащий	**5**
	другое	**0**
Стаж	1/безработный	**0**
	1...3	**5**
	3...10	**10**
	10 и более	**15**

Описывает нелинейные зависимости



Бинить или не бинить?

Нелинейности в размерности

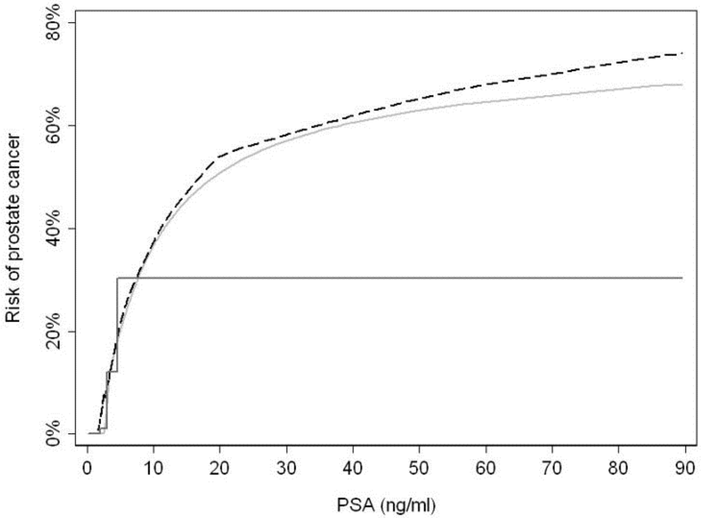


15+1 причина против биннинга

Problems Caused by Categorizing Continuous Variables

1. Optimum decisions are made by applying a utility function to a predicted value (e.g., predicted risk). At the decision point, one can solve for the personalized cutpoint for predicted risk that optimizes the decision. Dichotomization on independent variables is completely at odds with making optimal decisions. To make an optimal decision, the cutpoint for a predictor would necessarily be a function of the continuous values of all the other predictors, as shown here in Section 18.3.1.
2. Loss of power and loss of precision of estimated means, odds, hazards, etc. Dichotomization of a predictor requires the researcher to add a new predictor to the mix to make up for the lost information.
3. Categorization assumes that the relationship between the predictor and the response is flat within intervals; this assumption is far less reasonable than a linearity assumption in most cases
4. Researchers seldom agree on the choice of cutpoint, thus there is a severe interpretation problem. One study may provide an odds ratio for comparing BMI > 30 with BMI <= 30, another for comparing BMI > 28 with BMI <= 28. Neither of these has a good definition and they have different meanings.
5. Categorization of continuous variables using percentiles is particularly hazardous. The percentiles are usually estimated from the data at hand, are estimated with sampling error, and do not relate to percentiles of the same variable in a population. Percentiling a variable is declaring to readers that how similar a person is to other persons is as important as how the physical characteristics of the measurement predict outcomes. For example, it is common to group the continuous variable BMI into quantile intervals. BMI has a smooth relationship with every outcome studied, and relates to outcome according to anatomy and physiology and not according to how many subjects have a similar BMI.
6. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required. The needed dummy variables will spend more degrees of freedom than will fitting a smooth relationship, hence power and precision will suffer. And because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding.
7. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed
8. Categorization only seems to yield interpretable estimates such as odds ratios. For example, suppose one computes the odds ratio for stroke for persons with a systolic blood pressure > 160 mmHg compared to persons with a blood pressure <= 160 mmHg. The interpretation of the resulting odds ratio will depend on the exact distribution of blood pressures in the sample (the proportion of subjects > 170, > 180, etc.). On the other hand, if blood pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for **exact** settings of the predictor, e.g., the odds ratio for 200 mmHg compared to 120 mmHg.
9. When the risk of stroke is being assessed for a new subject with a known blood pressure (say 162), the subject does not report to her physician "my blood pressure exceeds 160" but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
10. If cutpoints are determined in a way that is not blinded to the response variable, calculation of *P* -values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely invalid. For example, if cutpoints are chosen by trial and error in a way that utilizes the response, even informally, ordinary *P* -values will be too small and confidence intervals will not have the claimed coverage probabilities. The correct Monte-Carlo simulations must take into account both multiplicities and uncertainty in the choice of cutpoints. For example, if a cutpoint is chosen that minimizes the *P* -value and the resulting *P* -value is 0.05, the true type I error can easily be above 0.5; see here
11. Likewise, categorization that is not blinded to the response variable results in biased effect estimates (see this and this)
12. "Optimal" cutpoints do not replicate over studies. Hollander, Sauerbrei, and Schumacher (see here) state that "... the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the `optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were in included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology."
13. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations (see this)
14. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.
15. A better approach that maximizes power and that only assumes a smooth relationship is to use a restricted cubic spline (regression spline; piecewise cubic polynomial) function for predictors that are not known to predict linearly. Use of flexible parametric approaches such as this allows standard inference techniques (*P* -values, confidence limits) to be used

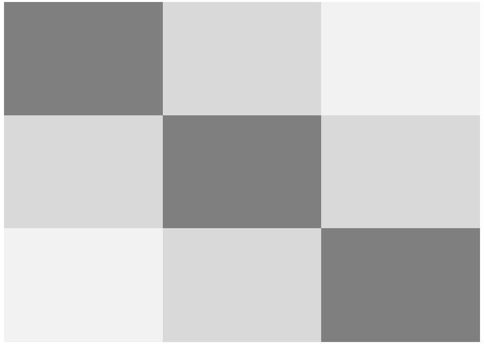
Prostate cancer risk by PSA (black dashed line), with predicted risks using either cubic splines (light gray solid line) or quartiles (dark gray solid line).



http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous

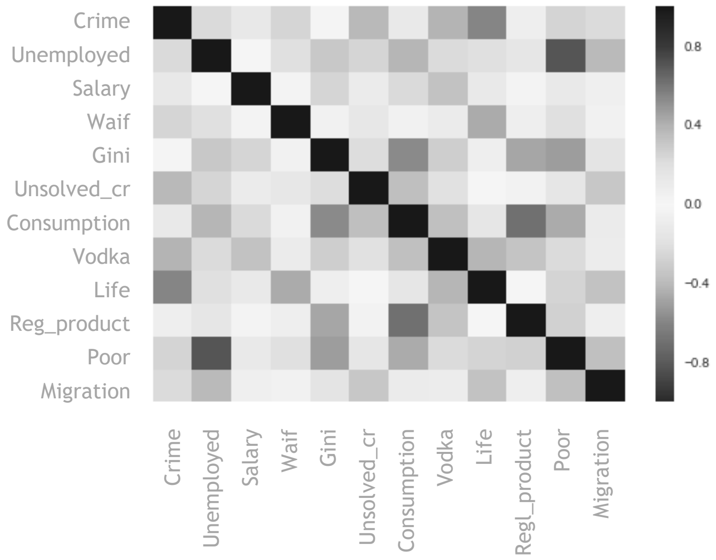
https://doi.org/10.1186/1471-2288-12-21

2. Отбор переменных



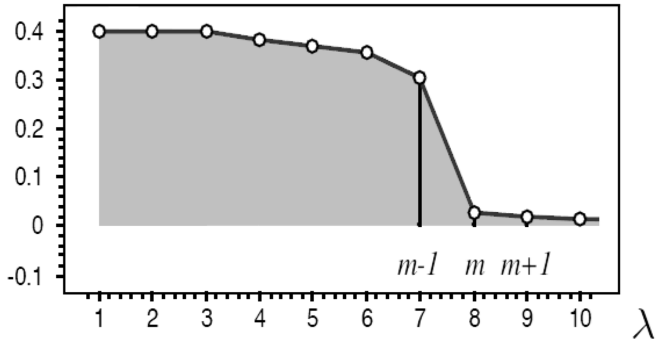
не очень удобно

Матрица корреляций



- Чувствительность к качеству данных (выбросы, ошибки)
- Не учитываются сложные взаимосвязи
- Ошибки при интерпретации

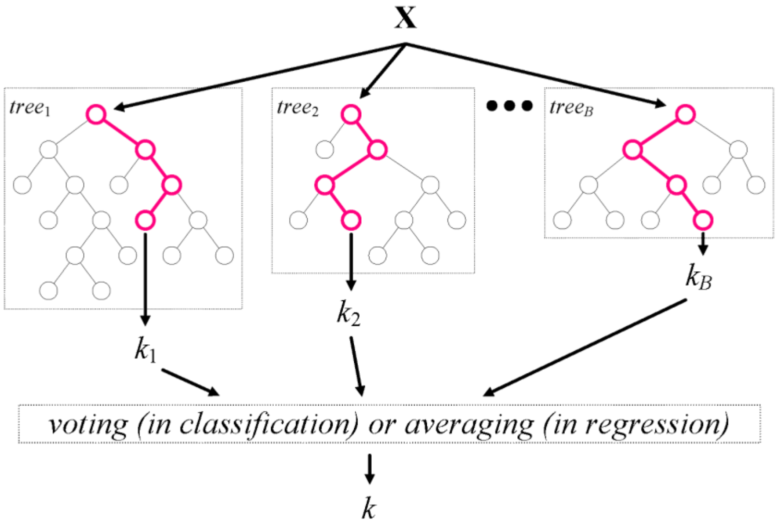
Метод главных компонент (PCA)



$$E_m = \frac{\|GU^{\mathsf{T}} - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leqslant \varepsilon$$

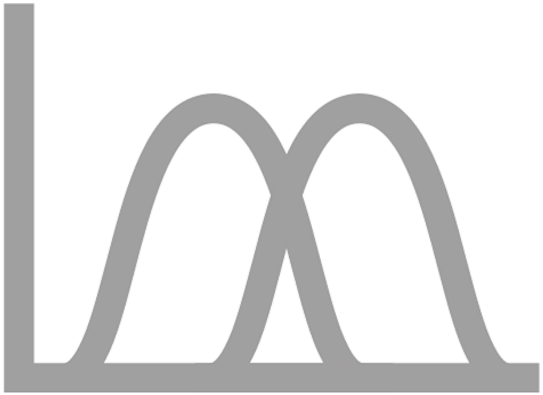
- Обучение «без учителя» – не учитывается целевая переменная
- Главные компоненты не всегда самые информативные
- Чувствительность к масштабу
- Проблема выбора порога

New Approach by Random Forest



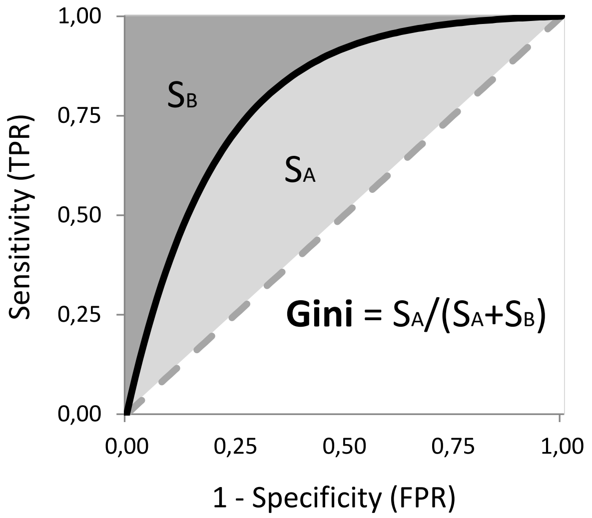
- Высокое качество отбора переменных
- Вычислительная сложность
- Переобучение

3. Метрики качества

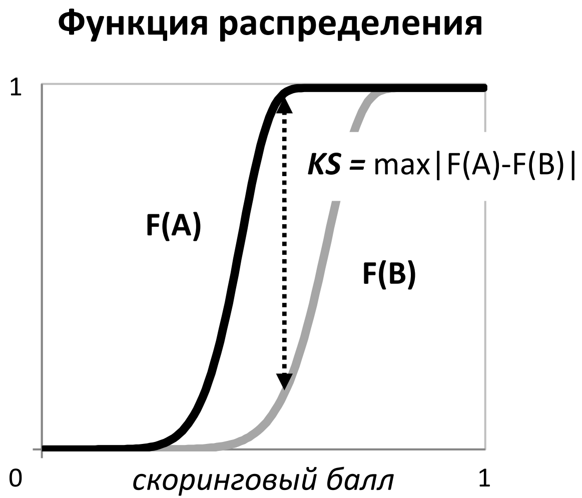


Отраслевой стандарт

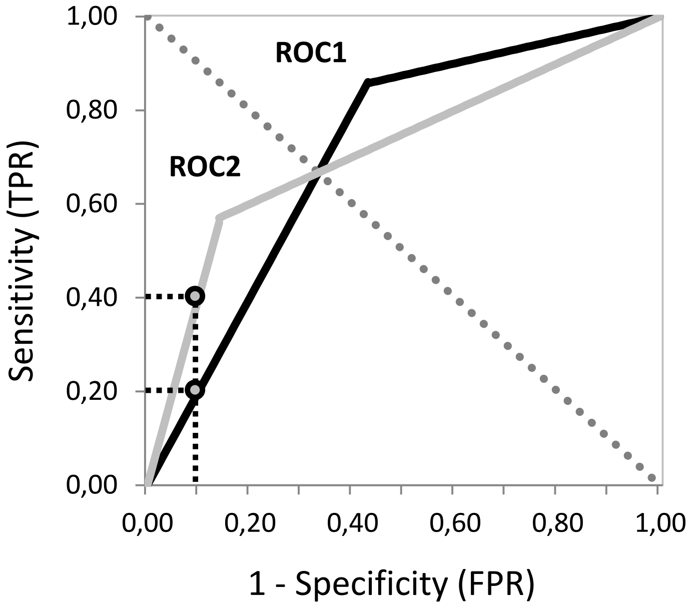
Gini Index



KS statistic



Проблемы метрики Gini



- Является интегральной метрикой с артефактами (см. рисунок)
- Плохо работает на несбалансированных выборках
- Не отражает финансовый результат

Метрики бывают разные

Пороговые метрики качества*

Метрика	Формула
ACC	(TP + TN) / (TP + TN + FN + FP)
ERR	(FP + FN) / (TP + TN + FN + FP)
PPCR	(TP + FP) / (TP + TN + FN + FP)
TNR	TN / (TN + FP)
REC, SN, TPR	TP / (TP + FN) = 1 – FNR
bACC	0,5 (TNR + TPR)
SP	TN / (TN + FP) = 1 – FPR
FPR	FP / (TN + FP) = 1 – SP
FNR	FN / (TP + FN) = 1 – SN
LRP	SN / (1 – SP) = (1 – FNR) / FPR
LRN	(1 – SN) / SP = FNR / (1 – FPR)
PREC, PPV	TP / (TP + FP)
FDR	FP / (TP + FP) = 1 – PPV
NPV	TN / (TN + FN)
FOR	FN / (TN + FN) = 1 – NPV
$F_{0,5}$	1,25 × PREC × REC / (0,25 × PREC + REC)
F_1	2 × PREC × REC / (PREC + REC)
F_2	5 × PREC × REC / (4 × PREC + REC)
MCC	(TP × TN – FP × FN) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))$^{1/2}$
LIFT	PREC / (TP + FN) / (TP + TN + FN + FP)

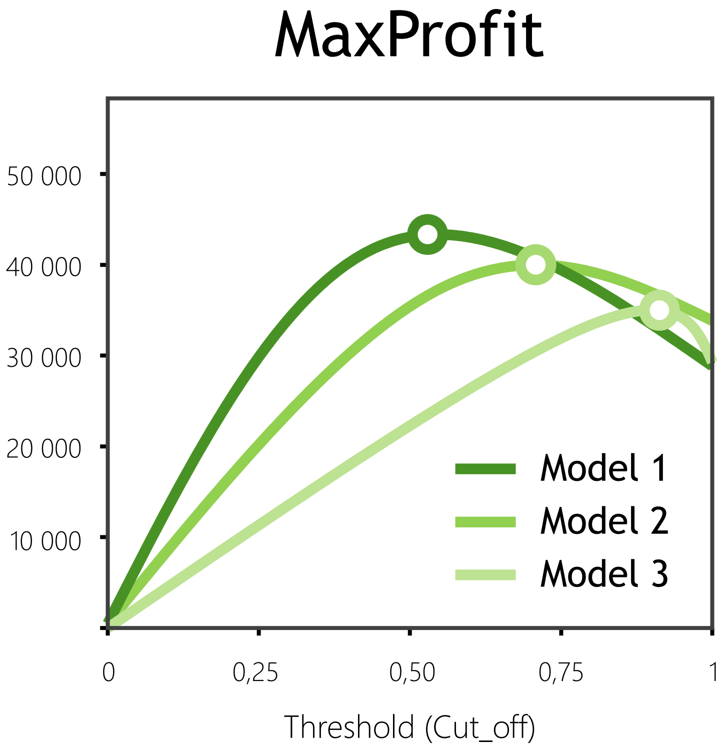
* ACC – accuracy; ERR – error rate; PPCR – predicted positive condition rate; TNR – true negative rate; REC – recall; SN – sensitivity; TPR – true positive rate; bACC – balanced accuracy; SP – specificity; FPR – false positive rate; FNR – false negative rate; LRP – likelihood ratio positive; LRN – likelihood ratio negative; PREC – precision; PPV – positive predictive value; FDR – false discovery rate; NPV – negative predictive value; FOR – false omission rate; *F* – *F*-score; MCC – Matthews correlation coefficient; LIFT – concentration increase; TP – true positives; TN – true negatives; FP – false positives; FN – false negatives.

Метрики качества, не зависящие от порога**

Метрика	Анализируемая кривая	Методика вычисления
AUC-ROC	ROC-кривая	Площадь под ROC-кривой
Gini	ROC-кривая	Gini = 2 × AUC_ROC – 1
AUC-CROC	CROC-кривая[1]	Площадь под CROC-кривой
AUC-PR, AP	PR-кривая	Площадь под PR-кривой
AUC LIFT	LIFT-кривая[2]	Площадь под LIFT-кривой
NAP	PR-кривая	NAP = AP/(1 – d) – d/(1 – d)
KS	Функции распределения	KS = max$\lvert F(a_i) - F(b_i)\rvert$
S-test	Плотности распределения	S = $\Sigma\lvert a_i - b_i\rvert/2$
Chi	Плотности распределения	Chi = $\Sigma((a_i - b_i)$^(0,5))/$(a_i + b_i)$
T-test	Плотности распределения	T = (E(a_i) – E(b_i))/((var(a_i)/N_a) + (var(b_i)/N_b)^(0,5))
MAD-test[3]	Плотности распределения	MAD = $\Sigma\lvert a_i - b_i\rvert/K$, $1 \le i \le K$
AD-test[4]	Плотности распределения	AD = (Σ(N_i × Z($N_a + N_b - N_a$ × i))^(0,5))/(i × Z($N_a + N_b - i$))/ ($N_a \times N_b$), $1 \le i \le N_a + N_b$
KLD	Плотности распределения	KLD(a_i‖b_i) = Σa_i × log(a_i/b_i)
JSD	Плотности распределения	JSD(a_i‖b_i) = (KLD(a_i‖($a_i + b_i$)/2) + KLD(b_i‖($a_i + b_i$)/2))/2

** AUC-ROC – area under curve ROC; Gini – Gini index; AUC-CROC – area under curve concentrated ROC; AUC-PR – area under curve PR; AP – average precision; AUC LIFT – area under curve LIFT; NAP – normalized average precision; KS – Kolmogorov-Smirnov test; Chi – Pearson's chi-squared test; T-test – Welch's t-test, unequal variances t-test; MAD – mean absolute deviation; AD-test – Anderson-Darling test; KLD – Kullback-Leibler divergence; JSD – Jensen-Shannon divergence.

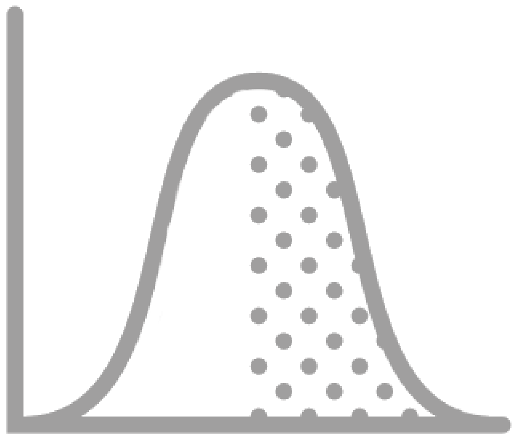
MaxProfit — максимизирует прибыль



$$\boldsymbol{P} = (1 - \mathrm{RR})\left(\frac{t_0}{1 - t_0} \times \mathrm{TN} - \mathrm{FN}\right) \times s$$

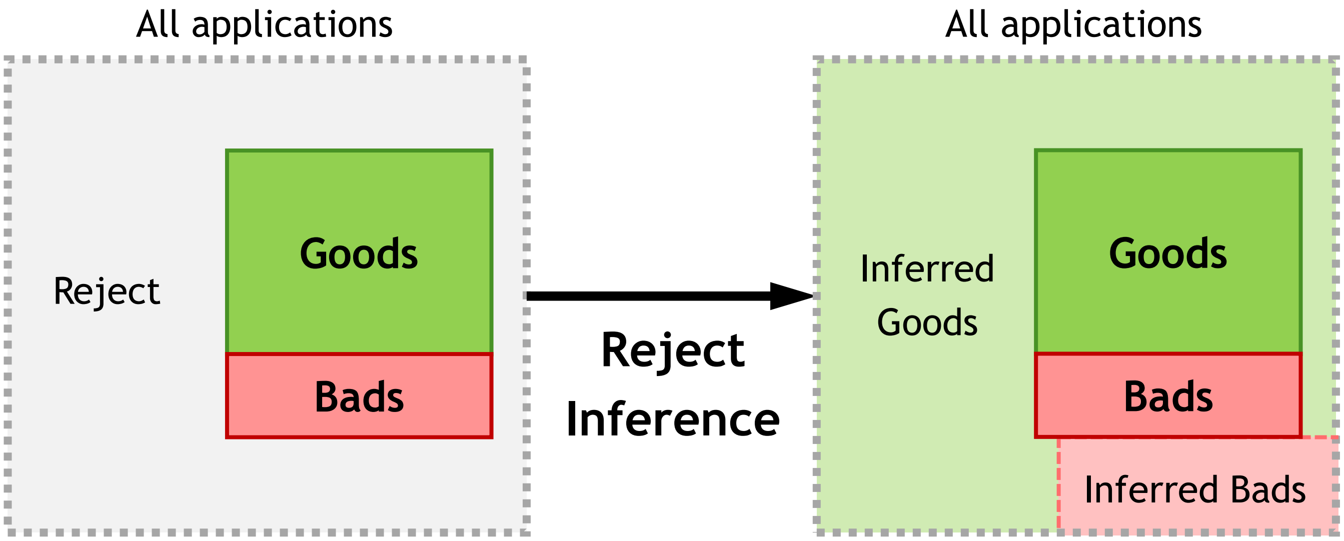
- Отражает финансовый результат
- Не чувствителен к дисбалансу
- Подбирает оптимальный порог

4. Reject Inference



Проблема смещения оценок

Reject Inference – процедура включения в выборку отклоненных заявок с целью корректировки смещения скоринговых оценок



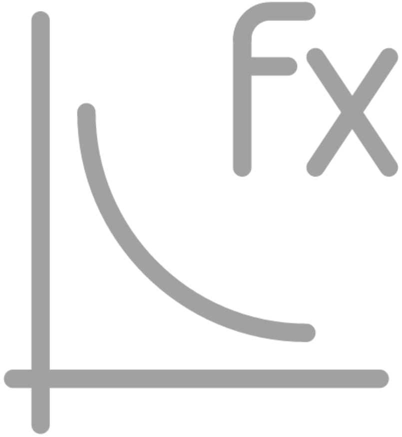
Методы Reject Inference

		Качество	Стоимость
Неявные методы	Hard cutoff	●○○○○	БЕСПЛАТНО
	Fuzzy	●●○○○	БЕСПЛАТНО
	Semi-Supervised	●●●○○	БЕСПЛАТНО
A/B test	Triggers/batch	●●●●○	НЕДОРОГО
	Open Gate	●●●●●	50 000 000 руб.*

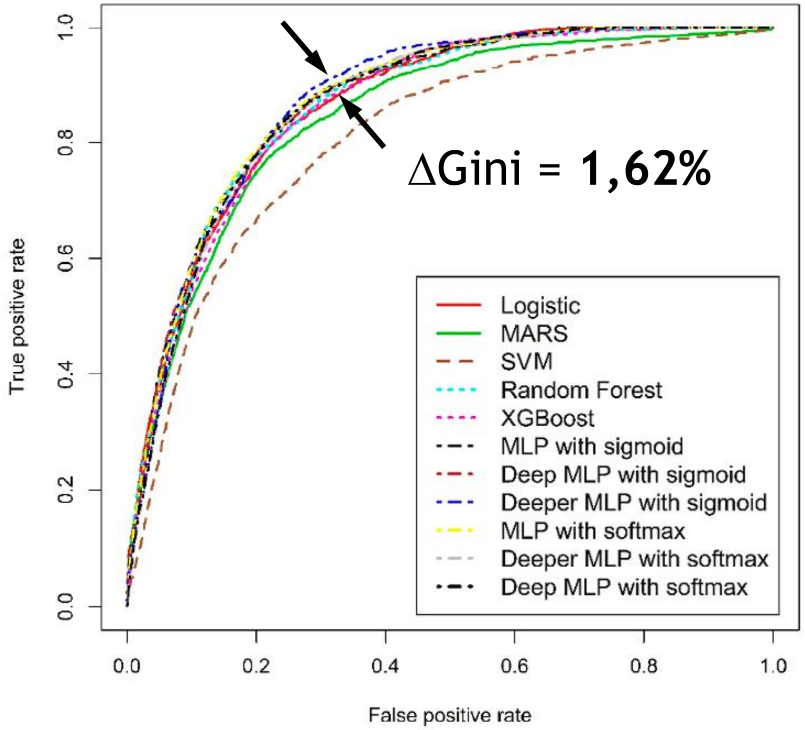
* Стоимость указана за 10 000 заявок

5.

ML
Алгоритмы



Сравнение алгоритмов (2019)



Log Regression – один из лучших алгоритмов по метрике Gini

Machine Learning Models	AUC	H-Measure
Logistic	0.8667	0.4151
MARS	0.8462	0.3868
SVM	0.8083	0.3097
RF	0.8682	0.4214
XGBoost	0.8633	0.3987
MLP with sigmoid	0.8726	0.4256
Deep MLP with sigmoid	0.8718	0.4233
Deeper MLP with sigmoid	**0.8748**	0.4298
MLP with softmax	0.8742	**0.4311**
Deeper MLP with softmax	0.8664	0.4126
Deep MLP with softmax	0.8682	0.4172

Lkhagvadorj Munkhdalai, et al. "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments." (2019)

Сравнение алгоритмов (2015)

Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research

Stefan Lessmann[a,*], Bart Baesens[bc], Hsin-Vonn Seow[d], Lyn C. Thomas[c]

[a] *School of Business and Economics, Humboldt-University of Berlin*

[b] *Department of Decision Sciences & Information Management, Catholic University of Leuven*

[c] *School of Management, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom*

[d] *Nottingham University Business School, University of Nottingham-Malaysia Campus*

Abstract

Many years have passed since Baesens et al. published their benchmarking study of classification algorithms in credit scoring [Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635.]. The interest in prediction methods for scorecard development is unbroken. However, there have been several advancements including novel learning methods, performance measures and techniques to reliably compare different classifiers, which the credit scoring literature does not reflect. To close these research gaps, we update the study of Baesens et al. and compare several novel classification algorithms to the state-of-the-art in credit scoring. In addition, we examine the extent to which the assessment of alternative scorecards differs across established and novel indicators of predictive accuracy. Finally, we explore whether more accurate classifiers are managerial meaningful. Our study provides valuable insight for professionals and academics in credit scoring. It helps practitioners to stay abreast of technical advancements in predictive modeling. From an academic point of view, the study provides an independent assessment of recent scoring methods and offers a new baseline to which future approaches can be compared.

Keywords: Data Mining, Credit Scoring, OR in banking, Forecasting benchmark

В исследовании сравнивались:

- **41** алгоритм
- на **8** выборках (портфелях)
- по **6** метрикам

Lessmann Stefan, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." (2015)

Рэнкинг обычных классификаторов

Classifier family	BM selection	Classifier	AUC	PCC	BS	H	PG	KS	AvgR	High score
Individual classifier	n.a.	ANN	*16.2* (.000)	*18.6* (.000)	27.5 (.000)	*17.9* (.000)	*14.9* (.020)	*17.6* (.000)	*18.8*	*14*
		B-Net	27.8 (.000)	26.8 (.000)	20.4 (.000)	28.3 (.000)	23.7 (.000)	26.2 (.000)	25.5	30
		CART	36.5 (.000)	32.8 (.000)	35.9 (.000)	36.3 (.000)	25.7 (.000)	34.1 (.000)	33.6	38
		ELM	30.1 (.000)	29.8 (.000)	35.9 (.000)	30.6 (.000)	27.0 (.000)	27.9 (.000)	30.2	36
		ELM-K	20.6 (.000)	19.9 (.000)	36.8 (.000)	19.0 (.000)	23.0 (.000)	20.6 (.000)	23.3	26
		J4.8	36.9 (.000)	34.2 (.000)	34.3 (.000)	35.4 (.000)	35.7 (.000)	32.5 (.000)	34.8	39
		k-NN	29.3 (.000)	30.1 (.000)	27.2 (.000)	30.0 (.000)	26.6 (.000)	30.5 (.000)	29.0	34
		LDA	21.8 (.000)	20.9 (.000)	16.7 (.000)	20.5 (.000)	24.8 (.000)	21.9 (.000)	21.1	20
		LR	20.1 (.000)	19.9 (.000)	*13.3* (.000)	19.0 (.000)	23.1 (.000)	20.4 (.000)	19.3	16
		LR-R	22.5 (.000)	22.0 (.000)	34.6 (.000)	22.5 (.000)	21.4 (.000)	21.4 (.000)	24.1	28
		NB	30.1 (.000)	29.9 (.000)	23.8 (.000)	29.3 (.000)	22.2 (.000)	29.1 (.000)	27.4	33
		RbfNN	31.4 (.000)	31.7 (.000)	28.0 (.000)	31.9 (.000)	24.1 (.000)	31.7 (.000)	29.8	35
		QDA	27.0 (.000)	26.4 (.000)	22.6 (.000)	26.4 (.000)	23.6 (.000)	27.3 (.000)	25.5	31
		SVM-L	21.7 (.000)	23.0 (.000)	31.8 (.000)	22.6 (.000)	19.7 (.000)	21.7 (.000)	23.4	27
		SVM-Rbf	20.5 (.000)	22.2 (.000)	31.8 (.000)	22.0 (.000)	21.7 (.000)	21.3 (.000)	23.2	25
		VP	37.8 (.000)	36.4 (.000)	31.4 (.000)	37.8 (.000)	34.6 (.000)	37.6 (.000)	35.9	40

Рэнкинг однородных ансамблей

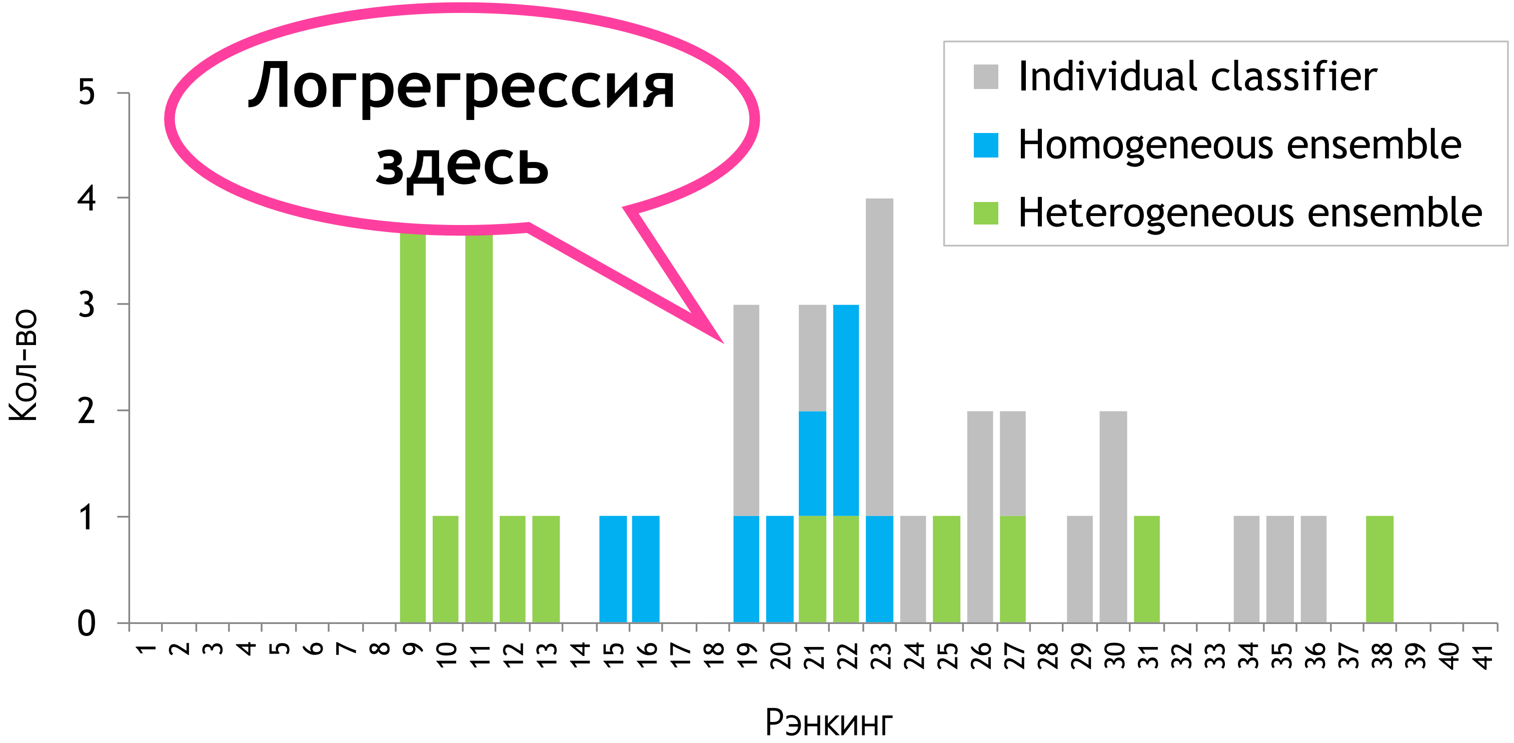
Classifier family	BM selection	Classifier	AUC		PCC		BS		H		PG		KS		AvgR	High score
Homogeneous ensemble	n.a.	ADT	22.0	(.000)	18.8	(.000)	19.0	(.000)	21.7	(.000)	19.4	(.000)	20.0	(.000)	20.2	17
		Bag	25.1	(.000)	22.6	(.000)	18.3	(.000)	23.5	(.000)	25.2	(.000)	24.7	(.000)	23.2	24
		BagNN	15.4	(.000)	17.3	(.000)	12.6	(.000)	16.5	(.000)	*15.0*	(.020)	16.6	(.000)	15.6	13
		Boost	16.9	(.000)	16.7	(.000)	25.2	(.000)	18.2	(.000)	19.2	(.000)	18.1	(.000)	19.0	15
		LMT	22.9	(.000)	23.4	(.000)	15.6	(.000)	25.1	(.000)	20.1	(.000)	22.9	(.000)	21.7	22
		RF	*14.7*	(.000)	*14.3*	(.039)	*12.6*	(.000)	*12.8*	(.004)	19.4	(.000)	*15.3*	(.000)	*14.8*	*12*
		RotFor	22.8	(.000)	21.9	(.000)	23.0	(.000)	21.1	(.000)	21.6	(.000)	22.9	(.000)	22.2	23
		SGB	21.0	(.000)	19.9	(.000)	20.8	(.000)	21.2	(.000)	22.5	(.000)	20.8	(.000)	21.0	19

Bold face indicates the best classifier (lowest average rank) per performance measure. Italic script highlights classifiers that perform best in their family (e.g., best individual classifier, best homogeneous ensemble, etc.). Values in brackets give the adjusted *p*-value corresponding to a pairwise comparison of the row classifier to the best classifier (per performance measure). An underscore indicates that *p*-values are significant at the 5% level. To account for the total number of pairwise comparisons, we adjust *p*-values using the *Rom*-procedure (García, et al., 2010). Prior to conducting multiple comparisons, we employ the Friedman test to verify that at least two classifiers perform significantly different (e.g., Demšar, 2006). The last row shows the corresponding χ^2 and *p*-values.

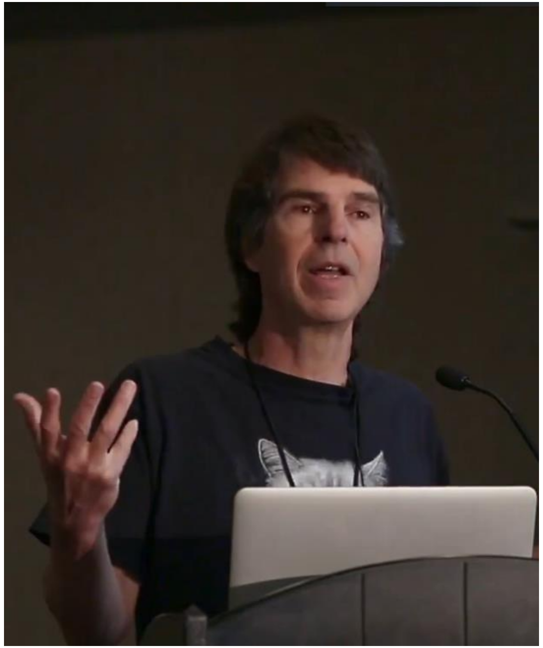
Рэнкинг разнородных ансамблей

Classifier family	BM selection	Classifier	AUC		PCC		BS		H		PG		KS		AvgR	High score
Heterogeneous ensemble	none	AvgS	8.7	(.795)	*10.8*	(.812)	*6.6*	(.628)	9.2	(.556)	12.0	(.420)	9.2	(.513)	9.4	4
		AvgW	***7.3***	(/)	12.6	(.578)	7.9	(.628)	**7.3**	(/)	**10.2**	(/)	**7.9**	(/)	*8.9*	*2*
		Stack	30.6	(.000)	26.6	(.000)	37.4	(.000)	29.6	(.000)	30.7	(.000)	29.5	(.000)	30.7	37
	Static direct	CompM	18.3	(.000)	15.3	(.004)	36.5	(.000)	17.2	(.000)	20.0	(.000)	18.2	(.000)	20.9	18
		EPVRL	8.2	(.795)	10.8	(.812)	6.8	(.628)	9.3	(.556)	13.7	(.125)	11.0	(.226)	10.0	5
		GASEN	8.6	(.795)	10.6	(.812)	6.5	(.628)	9.0	(.556)	*11.4*	(.420)	*9.0*	(.513)	9.2	3
		HCES	10.9	(.191)	11.7	(.812)	7.5	(.628)	10.2	(.449)	14.8	(.020)	13.1	(.010)	11.4	9
		HCES-Bag	*7.7*	(.795)	**9.7**	(/)	**5.8**	(/)	*8.2*	(.559)	12.5	(.420)	9.2	(.513)	**8.8**	*1*
		MPOE	9.9	(.637)	10.1	(.812)	9.4	(.126)	9.9	(.524)	15.1	(.018)	10.9	(.226)	10.9	6
		Top-T	8.7	(.795)	11.3	(.812)	10.0	(.055)	9.8	(.524)	14.8	(.020)	12.3	(.048)	11.2	8
	Static indirect	CuCE	10.0	(.637)	12.0	(.812)	10.1	(.050)	10.8	(.220)	*12.1*	(.420)	11.2	(.226)	*11.0*	*7*
		k-Means	12.6	(.008)	13.6	(.118)	*9.8*	(.073)	11.2	(.109)	14.9	(.020)	12.0	(.077)	12.4	10
		KaPru	27.7	(.000)	25.3	(.000)	15.7	(.000)	28.1	(.000)	25.1	(.000)	25.4	(.000)	24.5	29
		MDM	24.4	(.000)	24.0	(.000)	11.6	(.002)	23.7	(.000)	21.7	(.000)	23.7	(.000)	21.5	21
		UWA	*9.3*	(.795)	*11.8*	(.812)	19.5	(.000)	*10.1*	(.453)	14.3	(.049)	*10.9*	(.226)	12.7	11
	Dyna-mic	kNORA	*27.1*	(.000)	*26.7*	(.000)	*28.1*	(.000)	*28.1*	(.000)	*23.4*	(.000)	*25.9*	(.000)	*26.6*	*32*
		PMCC	40.1	(.000)	38.6	(.000)	32.9	(.000)	39.5	(.000)	39.9	(.000)	38.8	(.000)	38.3	41

Рэнкинг алгоритмов



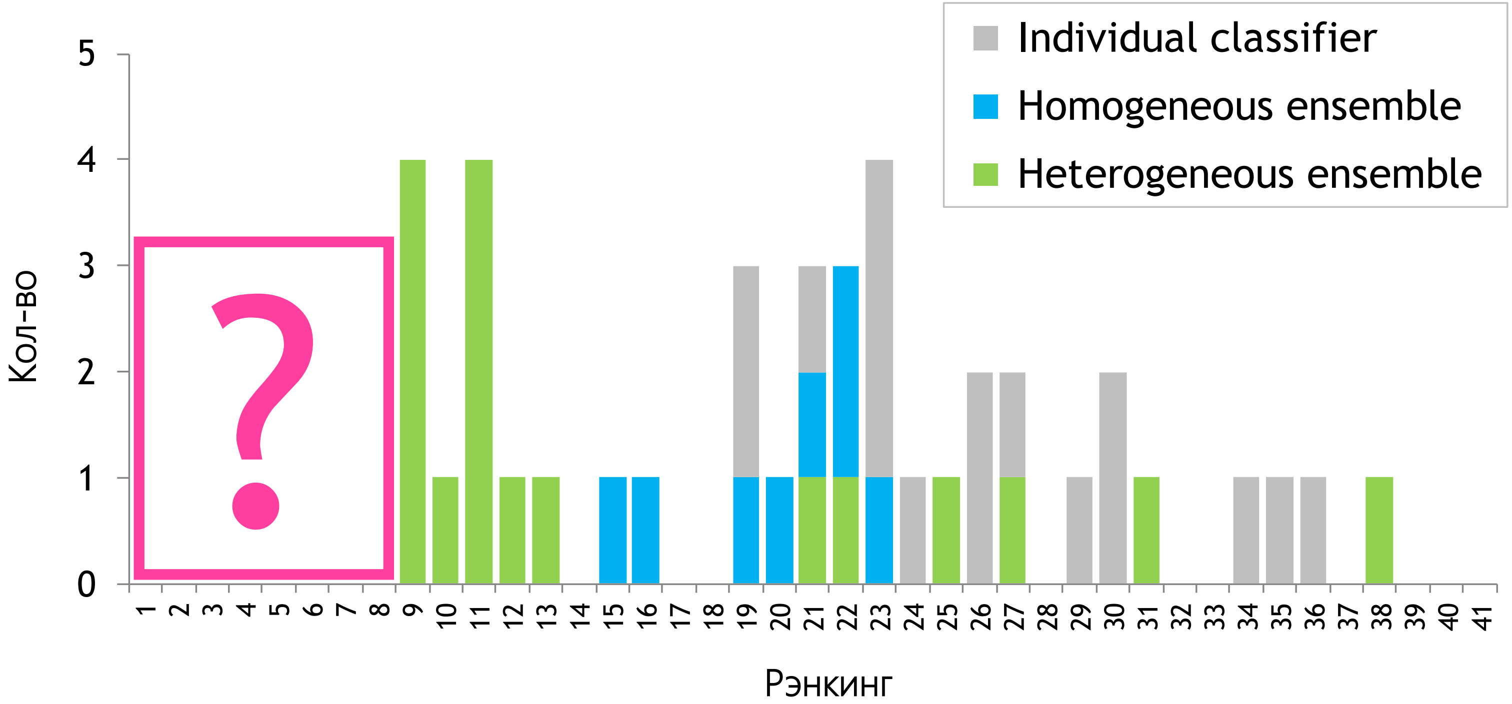
Теорема о бесплатных завтраках



В среднем по всем возможным порождающим определениям у любого алгоритма классификации частота ошибок классификации ранее не наблюдавшихся примеров одинакова. Самый изощренный алгоритм, который мы только можем придумать, в среднем (по всем возможным задачам) дает такое же качество, как простейшее предсказание: все точки принадлежат одному классу.

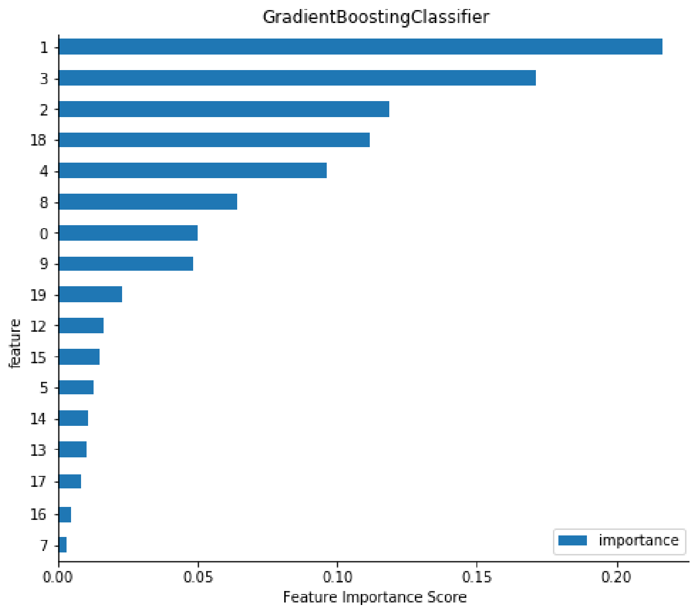
David H. Wolpert, 1996

Бесплатных завтраков не бывает?



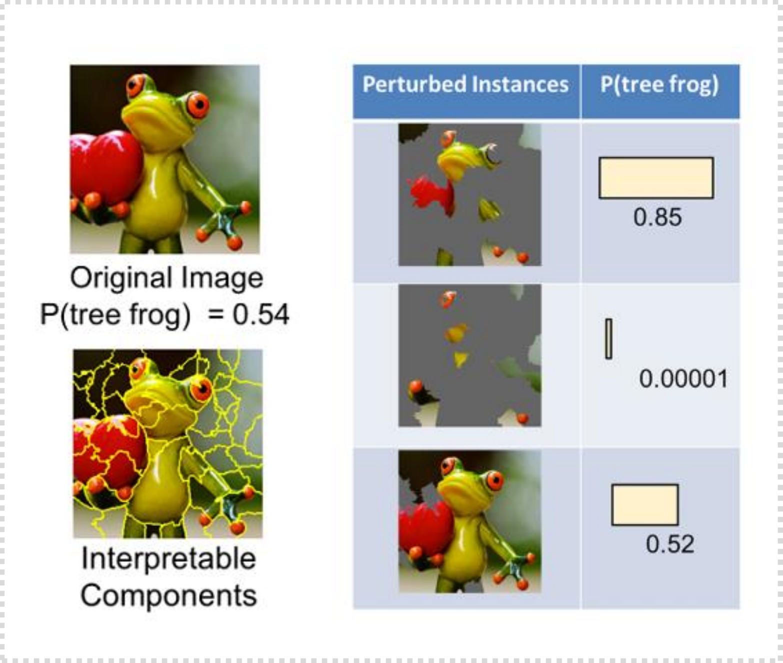
6. Интер-претаторы

Feature Importance



Ансамблевые алгоритмы на основе деревьев решений (Random Forest, Gradient Boosting и др.) позволяют оценить важность каждого признака через показатель Feature Importance.

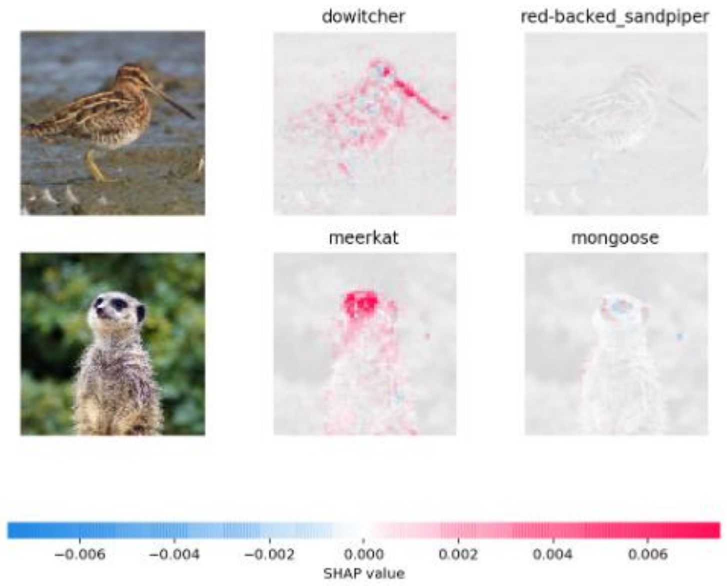
LIME



LIME использует подход интерпретации сложных моделей через простые модели:

Простая линейная модель аппроксимируют функцию сложной модели путем локального подбора линейных моделей к перестановкам исходного обучающего набора.

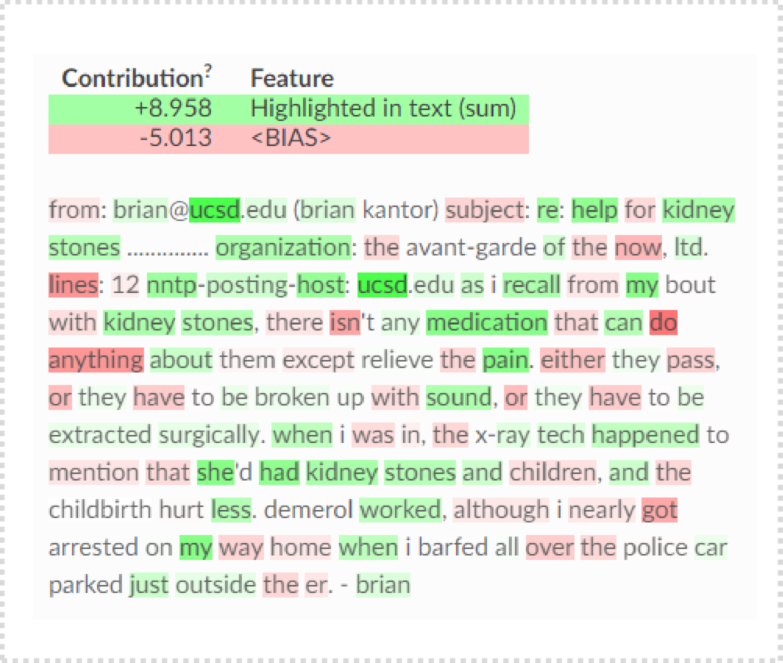
SHAP



SHAP (SHapley Additive exPlanations) - это унифицированный подход для объяснения результатов сложных моделей. SHAP связывает теорию игр с локальными объяснениями, объединяя несколько предыдущих методов и представляя аддитивный метод атрибуции признаков, основанный на ожиданиях.

https://github.com/slundberg/shap#citations

Eli5



Обширная библиотека, содержащая различные алгоритмы для интерпретации моделей:

- Permutation Importance
- LIME
- TextExplainer

https://eli5.readthedocs.io/en/latest/

7. Кодирование переменных

Энкодеры бывают не только WoE

Ordinal — convert string labels to integer values 1 through k. Ordinal.

OneHot — one column for each value to compare vs. all other values. Nominal, ordinal.

Binary — convert each integer to binary digits. Each binary digit gets one column. Some info loss but fewer dimensions. Ordinal.

BaseN — Ordinal, Binary, or higher encoding. Nominal, ordinal. Doesn't add much functionality. Probably avoid.

Hashing — Like OneHot but fewer dimensions, some info loss due to collisions. Nominal, ordinal.

Helmert (reverse) — The mean of the dependent variable for a level is compared to the mean of the dependent variable over all previous levels.

Sum — compares the mean of the dependent variable for a given level to the overall mean of the dependent variable over all the levels.

Backward Difference — the mean of the dependent variable for a level is compared with the mean of the dependent variable for the prior level.

Polynomial — orthogonal polynomial contrasts. The coefficients taken on by polynomial coding for k=4 levels are the linear, quadratic, and cubic trends in the categorical variable.

Target — use the mean of the DV, must take steps to avoid overfitting/ response leakage. Nominal, ordinal. For classification tasks.

LeaveOneOut — similar to target but avoids contamination. Nominal, ordinal. For classification tasks.

Weight of Evidence — added in v1.3. Not documented in the docs as of April 11, 2019. The method is explained in this post.

James-Stein — forthcoming in v1.4. Described in the code here.

M-estimator — forthcoming in v1.4. Described in the code here. Simplified target encoder.

https://towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of-3-6dca2f71b159

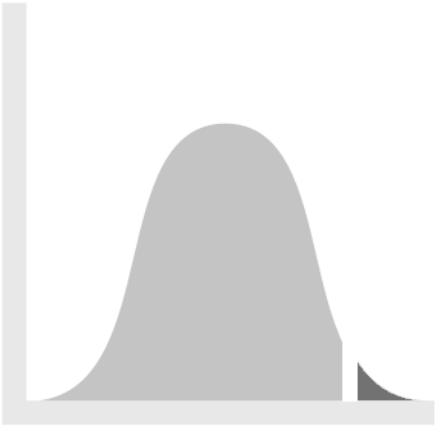
Практические результаты

Table 1.3 ROC AUC scores for Single Validation

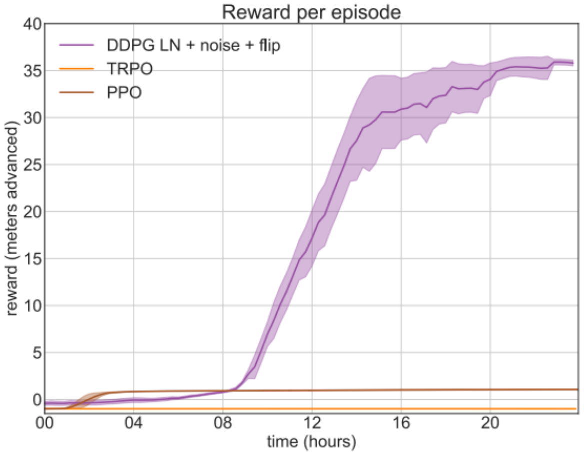
	telecom	adult	employee	credit	mortgages	promotion	kick	kdd_upselling	taxi	poverty_A	poverty_B	poverty_C
BackwardDifferenceEncoder	0.8382	0.9293	0.7569	0.7595	0.6894	0.9064				0.7323	0.6151	0.7108
CatBoostEncoder	0.8392	0.9292	**0.8498**	0.7594	0.6951	0.8918	0.7901	**0.8654**	0.5844	**0.7429**	**0.6902**	0.7333
FrequencyEncoder	0.8392	0.9293	0.8138	0.7592	0.6937	0.9055	**0.7902**	0.8634	0.582	0.7302	0.6128	0.7195
HelmertEncoder	**0.8404**	0.9297	0.8344	0.7597	**0.7027**	**0.9083**				0.7297	0.6374	0.7196
JamesSteinEncoder	0.8388	0.9292	0.7817	0.7597	0.667	0.9053	0.5835	0.726	0.5898	0.7303	0.6764	0.7217
LeaveOneOutEncoder	0.5	0.5182	0.6121	0.4997	0.5	0.5403	0.4682	0.5	0.5	0.5103	0.5	0.4959
MEstimateEncoder	0.8394	0.929	0.7353	0.7593	0.6957	0.9054	0.5877	0.5953	0.5946	0.7302	0.6493	0.7076
OrdinalEncoder	**0.8404**	**0.9299**	0.8274	0.7585	0.6917	0.9078	0.7809	0.8465	**0.6034**	0.7337	0.6635	**0.742**
SumEncoder	**0.8404**	0.929	0.8053	0.7593	0.6944	0.9073				0.7355	0.6206	0.7372
TargetEncoder	0.8388	0.9293	0.815	**0.7599**	0.6702	0.9057	0.7042	0.713	0.5894	0.7292	0.6742	0.7207
WOEEncoder	0.8393	0.9294	0.8325	**0.7599**	0.6801	0.9056	0.7172	0.8391	0.5903	0.7279	0.6737	0.7224

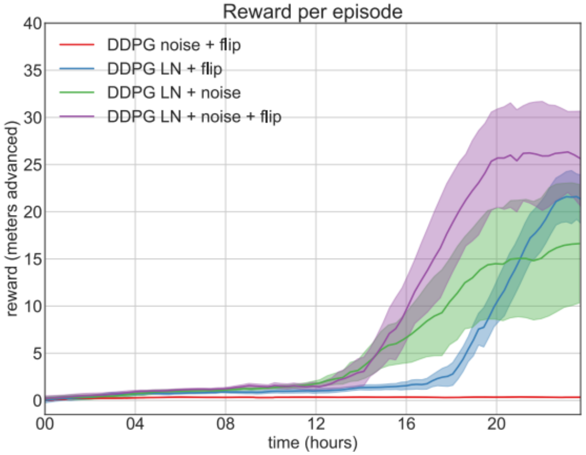
https://github.com/DenisVorotyntsev/CategoricalEncodingBenchmark/blob/master/README.md

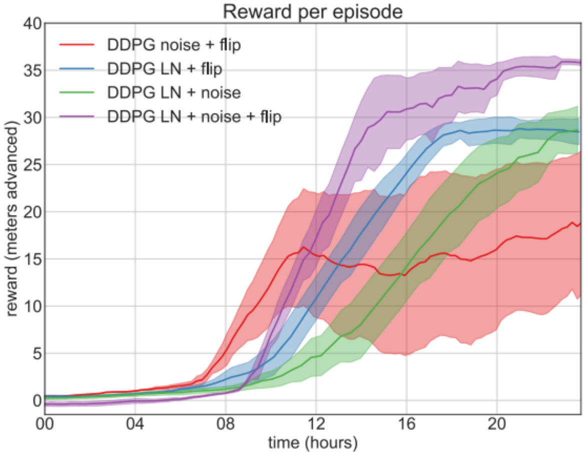
8. Bootstrap и стат. тесты



Bootstrap (пример)

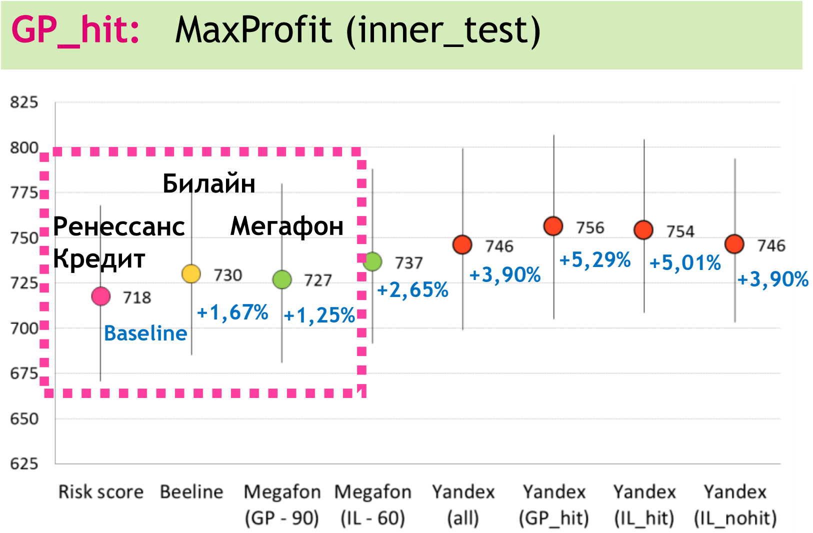


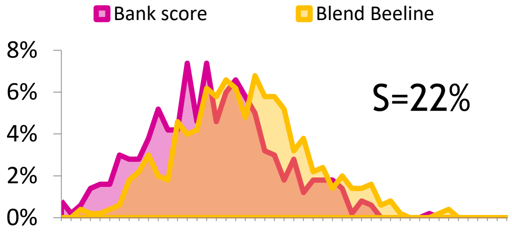


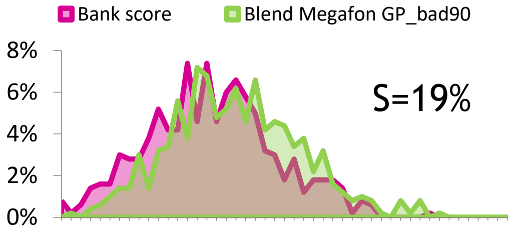


Bootstrap – обучение одной модели на случайных подвыборках много раз (~100-500). На выходе получается распределение оценок, по которому можно проверить статистическую значимость полученных результатов.

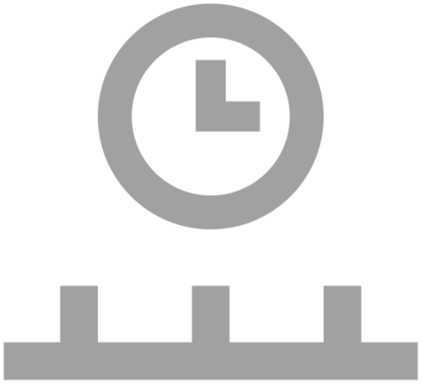
Bootstrap для Blend-моделей

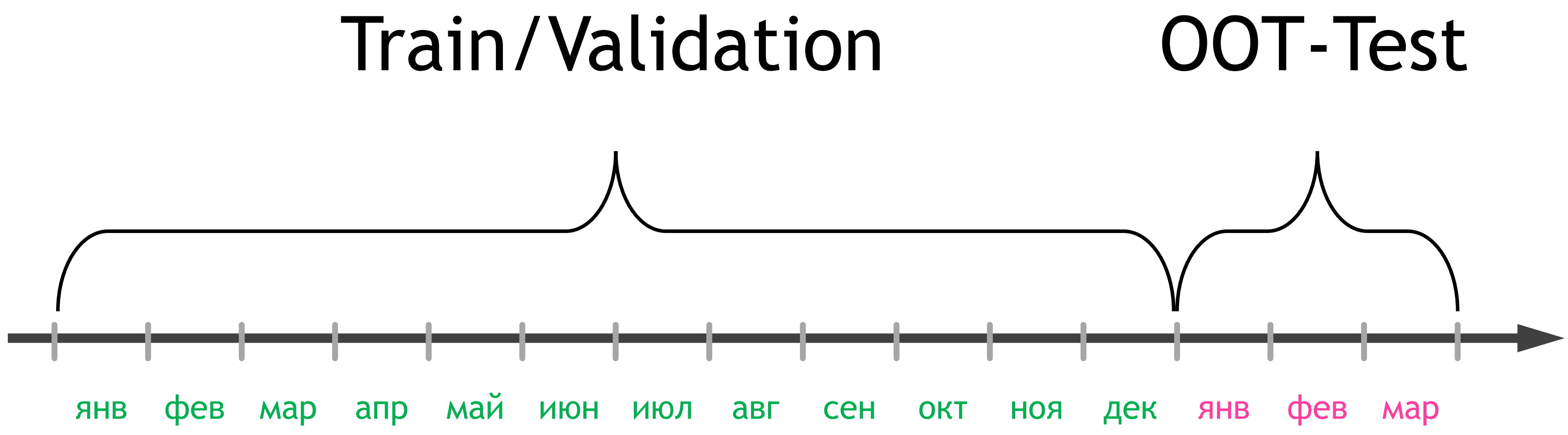






9. Out-of-Time выборки

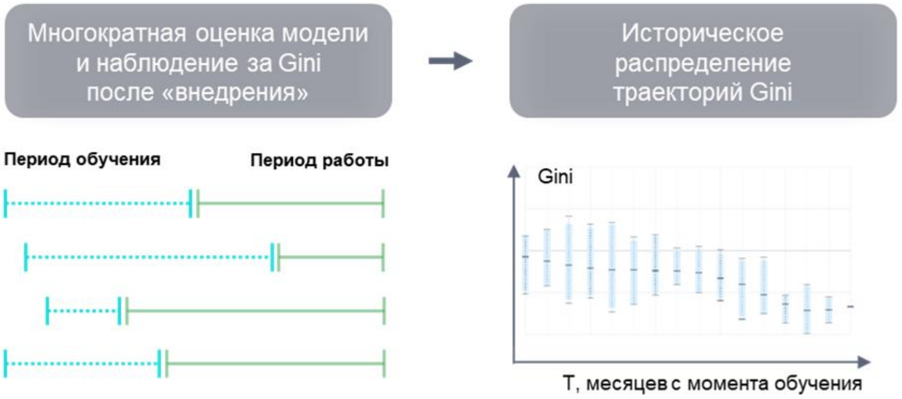



Train/Validation
OOT-Test
янв
фев
мар
апр
май
июн
июл
авг
сен
окт
ноя
дек
янв
фев
мар

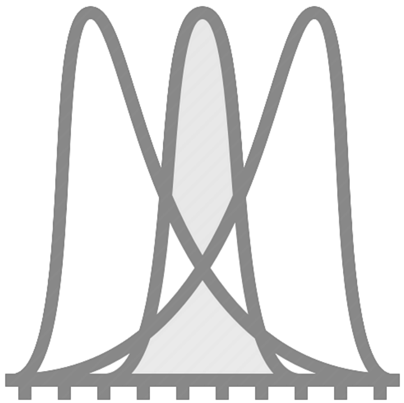
Сколько моделей не внедряется из-за низкого качества на ООТ?

- Много
- Мало
- Ни одной

Для оценки стабильности Gini можно строить отдельную модель



10. Family-Wise Error Rate



Family-Wise Error Rate

Поправка на множественную проверку гипотез (multiple comparisons, multiplicity, multiple testing problem) — способ устрания эффекта множественных сравнений, возникающего при необходимости построения семейства статистических выводов. Для устранения этого эффекта было разработано несколько подходов.

Попарвка Бонферрони

$$\mathrm{FWER} = P(V \geq 1) = P\left\{\bigcup_{i=1}^{m}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m} P\left(p_i \leq \frac{\alpha}{m}\right) \leq m\frac{\alpha}{m} = \alpha$$

Метод Шидака

$$\mathrm{P}(T_1 \leq t_1, \ldots, T_m \leq t_m) \geq \prod_{i=1}^{m} \mathrm{P}(T_i \leq t_i), \forall t$$

Метод Шидака-Холма

$$\alpha_1 = 1 - (1-\alpha)^{1/m} \ldots \alpha_i = 1 - (1-\alpha)^{\frac{1}{m-i+1}} \ldots \alpha_m = \alpha$$

Примеры из практики (правила AFS)

Группа	№ Правила	30+mob3 ценз	% 30+mob3	Hit rate
1	1	1095	1,6%	2,4%
1	2	199	2,5%	0,4%
1	3	3255	1,0%	7,2%
1	5	2785	1,0%	6,2%
1	6	6355	0,7%	14,1%
1	8	3097	1,5%	6,9%
1	9	531	3,0%	1,2%
2	10	1637	1,0%	3,6%
2	11	2613	1,0%	5,8%
2	12	3943	1,3%	8,8%
2	13	116	1,7%	0,3%
2	14	739	1,8%	1,6%
2	15	296	2,0%	0,7%
2	16	1327	1,7%	3,0%
2	17	961	1,2%	2,1%
2	18	3832	1,4%	8,5%
3	19	6048	1,6%	13,4%
3	20	1219	1,2%	2,7%
3	21	934	1,3%	2,1%
3	22	1107	1,3%	2,5%
3	23	1403	1,3%	3,1%
3	24	2518	1,2%	5,6%
3	25	3577	1,2%	8,0%
3	26	936	1,5%	2,1%
3	27	1486	1,7%	3,3%
3	28	1943	1,5%	4,3%
3	29	2759	1,4%	6,1%
3	30	11405	1,2%	25,4%
4	31	3653	2,2%	8,1%
4	32	3738	2,2%	8,3%
4	33	6912	1,7%	15,4%
4	34	4487	1,8%	10,0%
4	35	1531	3,4%	3,4%

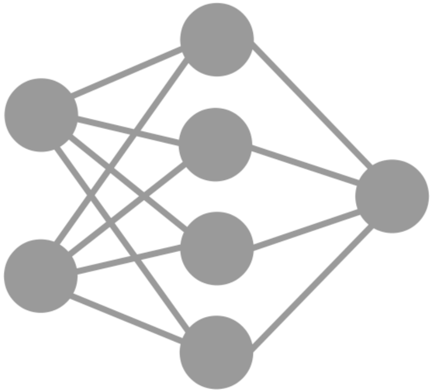
Группа	№ Правила	30+mob3 ценз	% 30+mob3	Hit rate
4	36	3146	1,0%	7,0%
4	37	13491	1,7%	30,0%
4	38	13760	1,8%	30,6%
5	39	323	3,4%	0,7%
5	40	546	2,7%	1,2%
5	41	776	2,6%	1,7%
5	42	1382	1,5%	3,1%
5	43	2243	1,3%	5,0%
5	44	3027	1,5%	6,7%
5	45	539	1,7%	1,2%
5	46	886	1,6%	2,0%
5	47	1273	1,4%	2,8%
6	48	109	5,5%	0,2%
6	49	140	5,0%	0,3%
6	50	31	6,5%	0,1%
6	51	34	5,9%	0,1%
6	52	252	0,4%	0,6%
6	53	340	0,3%	0,8%
6	54	102	2,9%	0,2%
6	55	128	1,6%	0,3%
6	56	1264	1,5%	2,8%
6	57	1705	1,3%	3,8%
6	58	688	1,6%	1,5%
6	59	807	1,5%	1,8%
6	60	385	1,8%	0,9%
6	61	503	1,6%	1,1%
6	62	755	1,9%	1,7%
6	63	2261	1,3%	5,0%
6	64	1602	1,8%	3,6%
6	65	2138	1,5%	4,8%
6	66	940	1,8%	2,1%
6	67	1212	1,7%	2,7%
7	68	1933	2,2%	4,3%

Группа	№ Правила	30+mob3 ценз	% 30+mob3	Hit rate
7	69	2553	2,2%	5,7%
7	70	3186	2,0%	7,1%
7	71	1929	2,2%	4,3%
7	72	2544	2,2%	5,7%
7	73	3176	2,0%	7,1%
7	74	3186	2,0%	7,1%
7	75	4108	1,8%	9,1%
7	76	4800	1,7%	10,7%
7	77	1020	2,4%	2,3%
6	78	7390	1,5%	16,4%
6	79	3551	1,6%	7,9%
6	80	11060	1,4%	24,6%
6	81	6047	1,4%	13,4%
6	82	3428	1,7%	7,6%
6	83	14239	1,2%	31,7%
6	84	9605	1,3%	21,4%
6	85	14497	1,3%	32,2%
6	86	11211	1,4%	24,9%
6	87	13298	1,4%	29,6%
6	88	11231	1,4%	25,0%
6	89	2805	1,7%	6,2%
6	90	2009	1,6%	4,5%
6	91	1191	1,8%	2,6%
6	92	906	1,9%	2,0%
8	93	4879	1,0%	10,8%
8	94	7543	1,1%	16,8%
8	95	8304	1,2%	18,5%
8	96	12050	1,1%	26,8%
9	98	4249	0,5%	9,4%
9	99	230	0,9%	0,5%
9	100	220	0,9%	0,5%
9	101	481	0,4%	1,1%
9	102	774	1,0%	1,7%

Группа	№ Правила	30+mob3 ценз	% 30+mob3	Hit rate
9	103	820	2,0%	1,8%
9	104	389	1,0%	0,9%
9	105	1188	2,9%	2,6%
10	106	[illegible]	[illegible]	[illegible]
10	109	510	2,9%	1,1%
10	110	1054	3,3%	2,3%
11	113	99	6,1%	0,2%
11	114	35	11,4%	0,1%
11	115	37	2,7%	0,1%
11	116	30	3,3%	0,1%
11	117	131	4,6%	0,3%
11	118	51	9,8%	0,1%
11	119	49	0,0%	0,1%
11	120	44	2,3%	0,1%
11	121	245	7,3%	0,5%
11	122	108	6,5%	0,2%
11	123	99	9,1%	0,2%
11	124	83	7,2%	0,2%
11	125	1020	2,9%	2,3%
11	126	496	2,6%	1,1%
11	127	409	3,9%	0,9%
11	128	312	3,2%	0,7%
11	129	[illegible]	[illegible]	[illegible]
11	130	2648	2,5%	5,9%
11	131	1944	2,0%	4,3%
11	132	1628	2,3%	3,6%
12	133	418	1,4%	0,9%
12	134	3520	1,0%	7,8%
13	135	13	7,7%	0,0%
13	136	134	1,5%	0,3%
13	137	93	0,0%	0,2%
13	138	196	1,0%	0,4%
13	139	0	#ДЕЛ/0!	0,0%

неэффективные правила	дефолт по правилу ниже таргета
низкоэффективные правила	дефолт по правилу выше таргета, но ниже 3*таргет
высокоэффективные правила	дефолт по правилу выше 3*таргет

11. Feature engineering



NLP — типы задач

Syntax	Semantics	Speech	Discourse
Grammar induction	Lexical semantics	Speech recognition	Automatic summarization
Lemmatization	Machine translation	Speech segmentation	Coreference resolution
Morphological segmentation	Named entity recognition	Text-to-speech	Discourse analysis
Part-of-speech tagging	Natural language generation		
Parsing	Natural language understanding		
Sentence breaking	Optical character recognition		
Stemming	Question answering		
Word segmentation	Recognizing Textual entailment		
Terminology extraction	Relationship extraction		
	Sentiment analysis		
	Topic segmentation		
	Word sense disambiguation		

Natural Language Processing

Выявление оффшорных компаний по публикациям в СМИ

ГАЗПРОМБАНК

В Газпромбанке используют методы NLP для построения графа связей российских компаний с оффшорными юр. лицами при рассмотрении заявок на корпоративные кредитные линии.

Результаты:

1) Посредством транзитивного замыкания в графе российских компаний через иностранные юр. лица сняты трансграничные ограничения;
2) Поиск не только юридических, но и экономических связей;
3) Topic и Sentiment разметка дают дополнительное понимание контекста взаимосвязи.

Задачу можно разделить на следующие этапы:

1. Идентификация российских компаний в новостной ленте
2. Идентификация иностранных компаний в новостной ленте
3. Sentiment analysis разметка новостного корпуса
4. Разметка новостного корпуса тематиками
5. Расчет статистических метрик силы взаимосвязанности компаний в графе
6. Отсев статистически недостоверных ребер путем «калибровки» на достоверно связанный граф

***2017.** Антифрод в банке. MSB Evens.*

Computer Vision

Open Source фотобиометрия

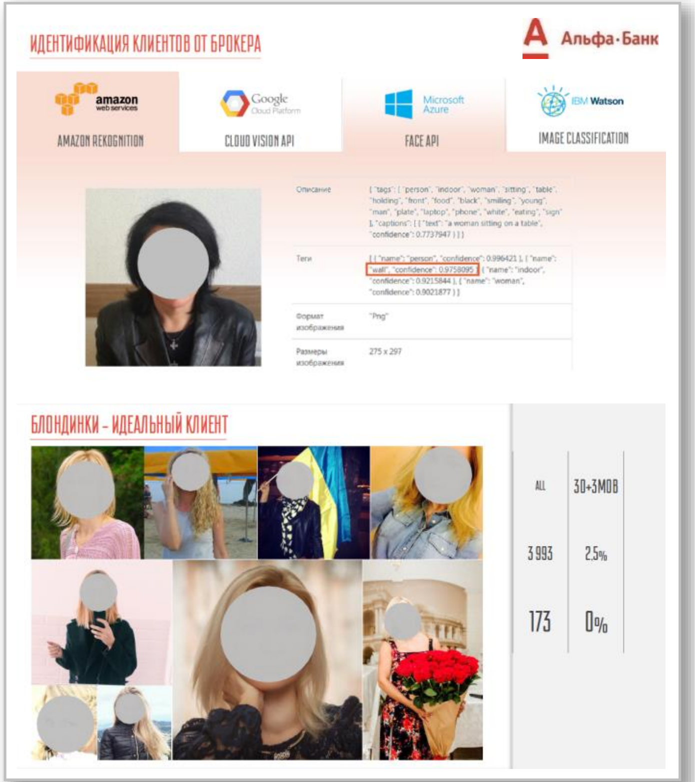
Специалисты из украинского Альфа-Банка использовали **бесплатные** Open Source технологии от Microsoft и Google для обработки фотографий клиентов.

Из фото клиента извлекалась информация:

1) Эмоции на лице;
2) Различные объекты (машины, пляж, дети и т.п.);
3) Задний фон

Извлеченную информацию использовали:

1. Как дополнительные предикторы в скоринге: например, выяснилось, что **лучшие клиенты** – это **блондинки**, а клиенты, сфотографированные на фоне дорогих авто – чаще допускают просрочки;
2. Для выявления **«черных брокеров»**: выявлялись одинаковые предметы на фоне разных фотографий клиентов, сделанных якобы в разных ТТ.



2017. Антифрод в банке. MSB Evens.

РЕЗЮМЕ

1. **Биннинг количественных переменных** – нужен или нет?

2. **Отбор переменных** – RF вместо Cross-Correlation Matrix & PCA

3. **Метрики качества** – MaxProfit вместо (или вместе) Gini и K-S

4. **Reject Inference** – методы борьбы со смещением скоринговых оценок

5. **ML-алгоритмы** – что может быть лучше логистической регрессии?

6. **Интерпретаторы сложных алгоритмов** – LIME, SHAP, Eli5

7. **Encoding** – WoE, Target, James-Stein, Hashing, Helmert, M-estimator, etc.

8. **Bootstrap и стат.тесты** – t-test, Friedman, Q-statistic, etc.

9. **Out-of-Time подход** – признан излишне консервативным

10. **Family-wise error rate** – проверка статистической значимости правил

11. **Feature engineering** – классические методы, NLP, Computer Vision, etc.

... **И многое другое, чего мы пока не знаем**

Психология

64% психологических экспериментов не воспроизводятся

Социология

75% экспериментов в социальной психологии не воспроизводятся

Data Science

Какой % работ по скорингу не реплицируются?

Пока не посчитали

Спасибо за внимание!

Афанасьев Сергей

Исполнительный директор
Начальник управления статистического анализа

КБ «Ренессанс Кредит»