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

Article in European Journal of Operational Research
May 2015

DOI: 10.1016/j.ejor.2015.05.030

Stefan Lessmann
Humboldt-Universität zu Berlin

Hsin-Vonn Seow
University of Nottingham, Malaysia Campus

Bart Baesens
University of Southampton

DS Workshop, 23.05.2019
Афанасьев Сергей

TABLE 1: ANALYSIS OF CLASSIFIER COMPARISONS IN RETAIL CREDIT SCORING

Retail credit scoring study (in chronological order)	Data*				Classifiers**				Evaluation***			
	No. of data sets	Observations/v ariables per data set		No. of classifier	ANN	SVM	ENS	S-ENS	TM	AUC	*H*	ST
(Baesens, et al., 2003)	8	4,875	21	17	X	X			X	X		P
(Malhotra & Malhotra, 2003)	1	1,078	6	2	X				X			P
(Atish & Jerrold, 2004)	2	610	16	5	X				X	X		P
(He, et al., 2004)	1	5,000	65	4	X				X			
(Lee & Chen, 2005)	1	510	18	5	X				X			
(Hand, et al., 2005)	1	1,000	20	4	X		X					
(Ong, et al., 2005)	2	845	17	6	X				X			
(West, et al., 2005)	2	845	19	4	X		X		X			P
(Y.-M. Huang, et al., 2006)	1	10,000	n.a.	10	X				X			
(Lee, et al., 2006)	1	8,000	9	5	X				X			
(S.-T. Li, et al., 2006)	1	600	17	2	X	X			X			P
(Xiao, et al., 2006)	3	972	17	13	X	X	X		X			P
(C.-L. Huang, et al., 2007)	2	845	19	4		X			X			F
(Yang, 2007)	2	16,817	85	3		X			X			
(H. Abdou, et al., 2008)	1	581	20	6	X				X			A
(Sinha & Zhao, 2008)	1	220	13	7	X	X			X	X		A
(C.-F. Tsai & Wu, 2008)	3	793	16	3	X		X		X			P
(Xu, et al., 2009)	1	690	15	4		X			X			
(Yu, et al., 2008)	1	653	13	7			X	X	X			
(H. A. Abdou, 2009)	1	1,262	25	3					X			
(Bellotti & Crook, 2009)	1	25,000	34	4		X				X		
(Chen, et al., 2009)	1	2,000	15	5		X			X			
(Nanni & Lumini, 2009)	3	793	16	16	X	X	X		X	X		
(Šušteršič, et al., 2009)	1	581	84	2	X				X			
(M.-C. Tsai, et al., 2009)	1	1,877	14	4	X				X			Q
(Yu, et al., 2009)	3	959	16	10	X	X	X		X	X		P
(J. Zhang, et al., 2009)	1	1,000	102	4					X			
(Hsieh & Hung, 2010)	1	1,000	20	4	X	X	X			X		
(Martens, et al., 2010)	1	1,000	20	4		X			X			
(Twala, 2010)	2	845	18	5			X		X			
(Yu, et al., 2010)	1	1,225	14	8	X	X	X		X			P
(D. Zhang, et al., 2010)	2	845	17	11	X	X	X		X			
(Zhou, et al., 2010)	2	1,113	17	25	X	X	X	X	X			
(J. Li, et al., 2011)	2	845	17	11		X			X			
(Finlay, 2011)	2	104,649	47	18	X		X		X			P
(Ping & Yongheng, 2011)	2	845	17	4	X	X			X			
(Wang, et al., 2011)	3	643	17	13	X	X	X		X			
(Yap, et al., 2011)	1	2,765	4	3					X			
(Yu, et al., 2011)	2	845	17	23	X	X			X			
(Akkoc, 2012)	1	2,000	11	4	X				X	X		
(Brown & Mues, 2012)	5	2,582	30	9	X	X	X			X		F/P
(Hens & Tiwari, 2012)	2	845	19	4		X			X			
(S. Li, et al., 2012)	2	672	15	5		X	X		X			
(Marqués, et al., 2012a)	4	836	20	35	X	X	X		X			F/P
(Marqués, et al., 2012b)	4	836	20	17	X	X	X		X	X		F/P
(Kruppa, et al., 2013)	1	65,524	17	5			X			X		
(Abellán & Mantas, 2014)	3	793	16	5	X		X			X		A
(C.-F. Tsai, 2014)	3	793	16	21	X		X		X			F/P
Mean / counts	1.9	6,167	24	7.8	30	24	18	3	40	10	0	17

* We report the mean of observations and independent variables for studies that employ multiple data sets. Eight studies mix retail and corporate credit data. Table 1 considers the retail data sets only.

** Abbreviations have the following meaning: ANN=Artificial neural network, SVM=Support vector machine, ENS=Ensemble classifier, S-ENS=Selective Ensemble (e.g., Partalas, et al., 2010).

*** Abbreviations have the following meaning: TM=Threshold metric (e.g., classification error, true positive rate, costs, etc.), AUC=Area under receiver operating characteristics curve, *H*=H-measure (Hand, 2009), ST=Statistical hypothesis testing. We use the following codes to report the type of statistical test used for classifier comparisons: P=Pairwise comparison (e.g., paired *t*-test), A=Analysis of variance, F=Friedman test, F/P=Friedman test together with post-hoc test (e.g., Demšar, 2006), Q=Press's Q statistic.

Figure 1: Classifier development and evaluation process

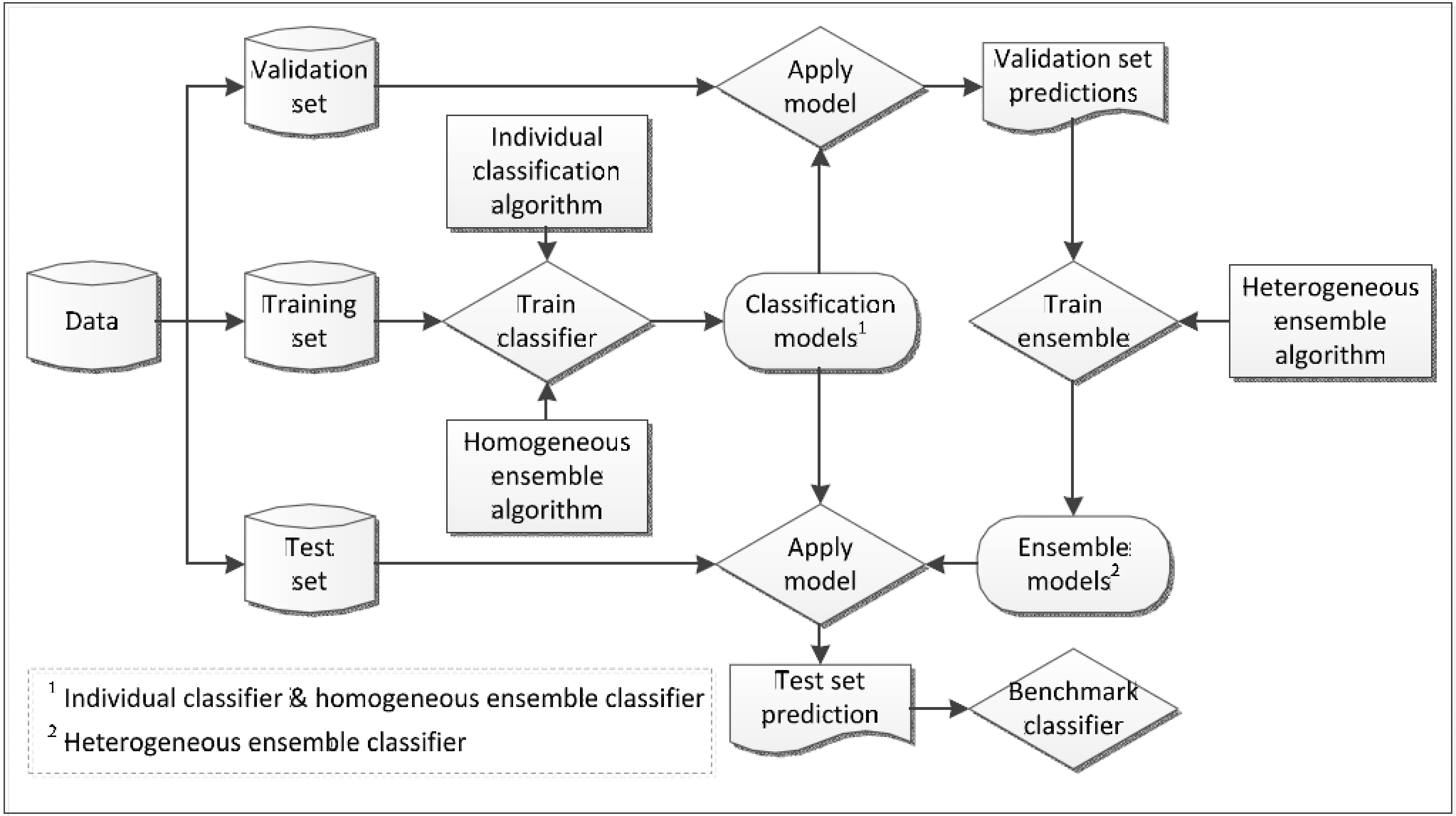


TABLE 2: CLASSIFICATION ALGORITHMS CONSIDERED IN THE BENCHMARKING STUDY

	BM selection	Classification algorithm	Acronym	Models
Individual classifier (16 algorithms and 933 models in total)	n.a.	Bayesian Network	B-Net	4
		CART	CART	10
		Extreme learning machine	ELM	120
		Kernalized ELM	ELM-K	200
		k-nearest neighbor	kNN	22
		J4.8	J4.8	36
		Linear discriminant analysis[1]	LDA	1
		Linear support vector machine	SVM-L	29
		Logistic regression[1]	LR	1
		Multilayer perceptron artificial neural network	ANN	171
		Naive Bayes	NB	1
		Quadratic discriminant analysis[1]	QDA	1
		Radial basis function neural network	RbfNN	5
		Regularized logistic regression	LR-R	27
		SVM with radial basis kernel function	SVM- Rbf	300
		Voted perceptron	VP	5
	Classification models from individual classifiers		**16**	**933**
Homogenous ensembles	n.a.	Alternating decision tree	ADT	5
		Bagged decision trees	Bag	9
		Bagged MLP	BagNN	4
		Boosted decision trees	Boost	48
		Logistic model tree	LMT	1
		Random forest	RF	30
		Rotation forest	RotFor	25
		Stochastic gradient boosting	SGB	9
	Classification models from homogeneous ensembles		**8**	**131**

	BM selection	Classification algorithm	Acronym	Models
Heterogeneous ensembles	n.a.	Simple average ensemble	AvgS	1
		Weighted average ensemble	AvgW	1
		Stacking	Stack	6
	Static direct	Complementary measure	CompM	4
		Ensemble pruning via reinforcement learning	EPVRL	4
		GASEN	GASEN	4
		Hill-climbing ensemble selection	HCES	12
		HCES with bootstrap sampling	HCES-Bag	16
		Matching pursuit optimization ensemble	MPOE	1
		Top-*T* ensemble	Top-*T*	12
	Static indirect	Clustering using compound error	CuCE	1
		k-Means clustering	k-Means	1
		Kappa pruning	KaPru	4
		Margin distance minimization	MDM	4
		Uncertainty weighted accuracy	UWA	4
	Dynamic	Probabilistic model for classifier competence	PMCC	1
		k-nearest oracle	kNORA	1
	Classification models from heterogeneous ensembles		**17**	**77**
Overall number of classification algorithms and models			**41**	**1141**

[1] To overcome problems associated with multicollinearity in high-dimensional data sets, we use correlation-based feature selection (Hall, 2000) to reduce the variable set prior to building a classification model.

Обычные классификаторы

	BM selection	Classification algorithm	Acronym	Models
Individual classifier (16 algorithms and 933 models in total)	n.a.	Bayesian Network	B-Net	4
		CART	CART	10
		Extreme learning machine	ELM	120
		Kernalized ELM	ELM-K	200
		k-nearest neighbor	kNN	22
		J4.8	J4.8	36
		Linear discriminant analysis[1]	LDA	1
		Linear support vector machine	SVM-L	29
		Logistic regression[1]	LR	1
		Multilayer perceptron artificial neural network	ANN	171
		Naive Bayes	NB	1
		Quadratic discriminant analysis[1]	QDA	1
		Radial basis function neural network	RbfNN	5
		Regularized logistic regression	LR-R	27
		SVM with radial basis kernel function	SVM- Rbf	300
		Voted perceptron	VP	5
Classification models from individual classifiers			**16**	**933**

[1] To overcome problems associated with multicollinearity in high-dimensional data sets, we use correlation-based feature selection (Hall, 2000) to reduce the variable set prior to building a classification model.

Однородные ансамбли

	BM selection	Classification algorithm	Acronym	Models
Homogenous ensembles	n.a.	Alternating decision tree	ADT	5
		Bagged decision trees	Bag	9
		Bagged MLP	BagNN	4
		Boosted decision trees	Boost	48
		Logistic model tree	LMT	1
		Random forest	RF	30
		Rotation forest	RotFor	25
		Stochastic gradient boosting	SGB	9
	Classification models from homogeneous ensembles		**8**	**131**

Разнородные ансамбли

	BM selection	Classification algorithm	Acronym	Models
Heterogeneous ensembles	n.a.	Simple average ensemble	AvgS	1
		Weighted average ensemble	AvgW	1
		Stacking	Stack	6
	Static direct	Complementary measure	CompM	4
		Ensemble pruning via reinforcement learning	EPVRL	4
		GASEN	GASEN	4
		Hill-climbing ensemble selection	HCES	12
		HCES with bootstrap sampling	HCES-Bag	16
		Matching pursuit optimization ensemble	MPOE	1
		Top-*T* ensemble	Top-*T*	12
	Static indirect	Clustering using compound error	CuCE	1
		k-Means clustering	k-Means	1
		Kappa pruning	KaPru	4
		Margin distance minimization	MDM	4
		Uncertainty weighted accuracy	UWA	4
	Dynamic	Probabilistic model for classifier competence	PMCC	1
		k-nearest oracle	kNORA	1
	Classification models from heterogeneous ensembles		**17**	**77**
Overall number of classification algorithms and models			**41**	**1141**

Name	Cases	Independent variables	Prior default rate	Nx2 cross-validation	Source
AC	690	14	.445	10	(Lichman, 2013)
GC	1,000	20	.300	10	(Lichman, 2013)
Th02	1,225	17	.264	10	(Thomas, et al., 2002)[6]
Bene 1	3,123	27	.667	10	(Baesens, et al., 2003)
Bene 2	7,190	28	.300	5	(Baesens, et al., 2003)
UK	30,000	14	.040	5	(Baesens, et al., 2003)
PAK	50,000	37	.261	5	http://sede.neurotech.com.br/PAKDD2010/
GMC	150,000	12	.067	3	http://www.kaggle.com/c/GiveMeSomeCredit

1. AC – Australian credit
2. GC – German credit
3. Th02 – data set from Thomas, et al. (2002)
4. Bene-1 – used in Baesens, et al. (2003), were collected from major financial institution in the Benelux
5. Bene-2 – used in Baesens, et al. (2003) , were collected from major financial institution in the Benelux
6. UK – used in Baesens, et al. (2003), were collected from major financial institution in the UK
7. PAK – have been provided by financial institution for the 2010 PAKDD data mining challenge
8. GMC – have been provided by financial institution for the “Give me some credit” Kaggle competition.

Рэнкинг обычных классификаторов

TABLE 4: AVERAGE CLASSIFIER RANKS ACROSS DATA SETS FOR DIFFERENT PERFORMANCE MEASURES

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.

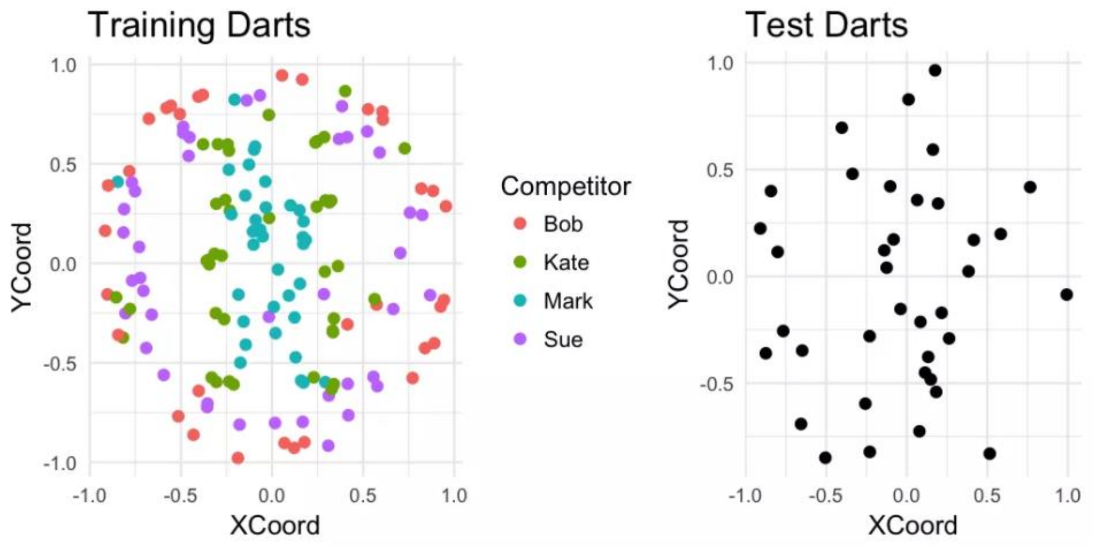
Рэнкинг разнородных ансамблей

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
Friedman χ^2_{40}			2775.1 (.000)	2076.3 (.000)	3514.4 (.000)	2671.7 (.000)	1462.3 (.000)	2202.6 (.000)		

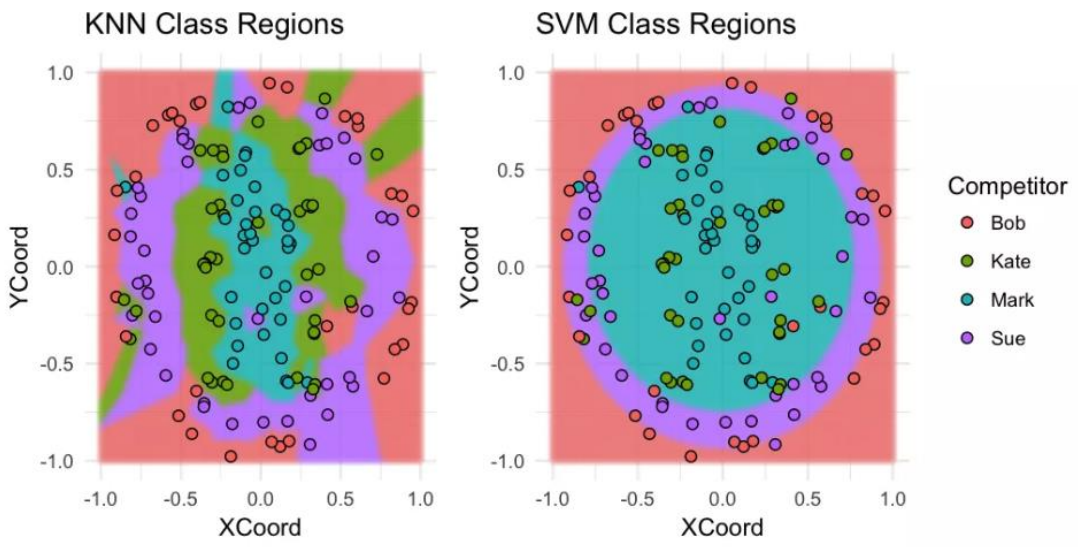
Почему стекинг улучшает результаты?

Предположим, что четыре человека бросают разом 187 дротиков в доску для дартса. Для 150 из них мы знаем, кто бросил каждый дротик и куда он попал. По остальным мы только видим, где приземлился дротик.

Задача: угадать, кто бросил каждый из немаркированных дротиков, исходя из места их попадания.



SVM хорошо справляется с классификацией бросков Боба и бросков Сью, но плохо дифференцирует броски Кейт и броски Марка. Модель k-ближайших соседей наоборот – хорошо классифицирует броски Кейт и броски Марка, но плохо справляется с бросками Боба и Сью. Стекинг этих моделей, вероятно, будет плодотворным.



***Ben Gorman.** A Kaggler's Guide to Model Stacking in Practice. 12.27.2016*
http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

Table 5: Full-Pairwise Comparison of Selected Classifiers

	AvgR	Adjusted p-values of pairwise comparisons		
		ANN	LR	RF
ANN	2.44			
LR	3.02	.000		
RF	2.53	.167	.000	
HCES-Bag	**2.01**	.000	.000	.000
Friedman χ^2_3	216.2	.000		

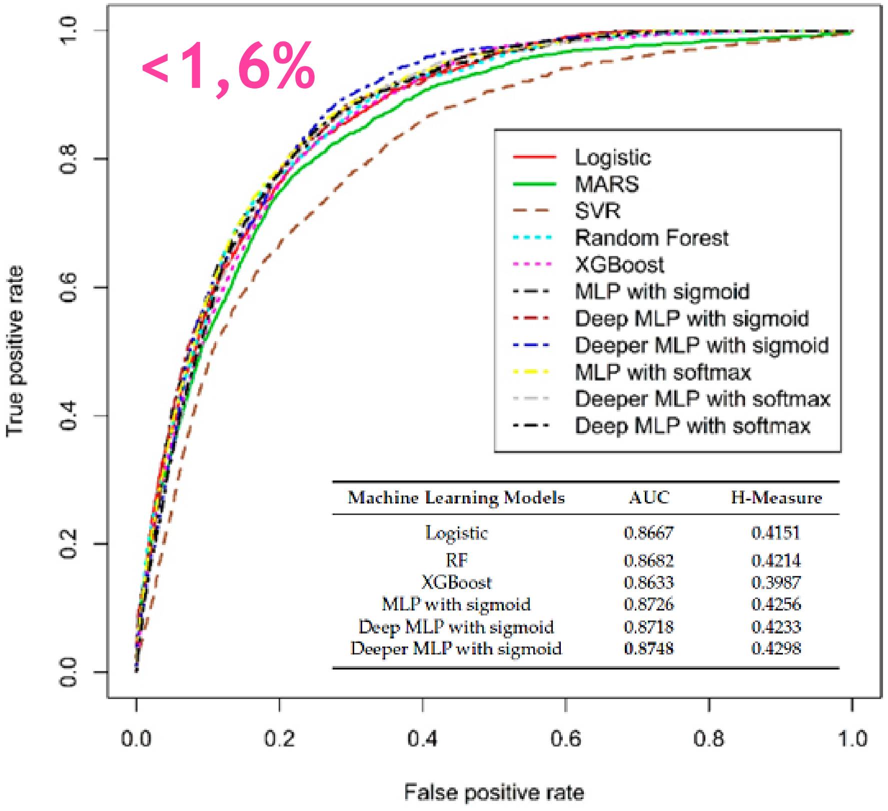
1. ANN – Multilayer Perceptron Artificial Neural Network
2. LR – Logistic Regression
3. RF – Random Forest
4. HCES-Bag – Hill-Climbing Ensemble Selection with Bootstrap Sampling

Table 6: Correlation of Classifier Rankings Across Performance Measures

	AUC	PCC	BS	H	PG	KS
AUC	1.00					
PCC	.88	1.00				
BS	$\boxed{.54}$	.54	1.00			
H	.93	.91	.56	1.00		
PG	$\boxed{.79}$	.72	$\boxed{.51}$	.76	1.00	
KS	.92	.89	.54	.91	.79	1.00

1. AUC – Area Under Curve ROC
2. PCC – Percentage Correctly Classified
3. BS – Brier Score
4. H – H-measure (Hand)
5. PG – Partial Gini Index
6. KS – Kolmogorov-Smirnov statistic

Gini



Error Costs

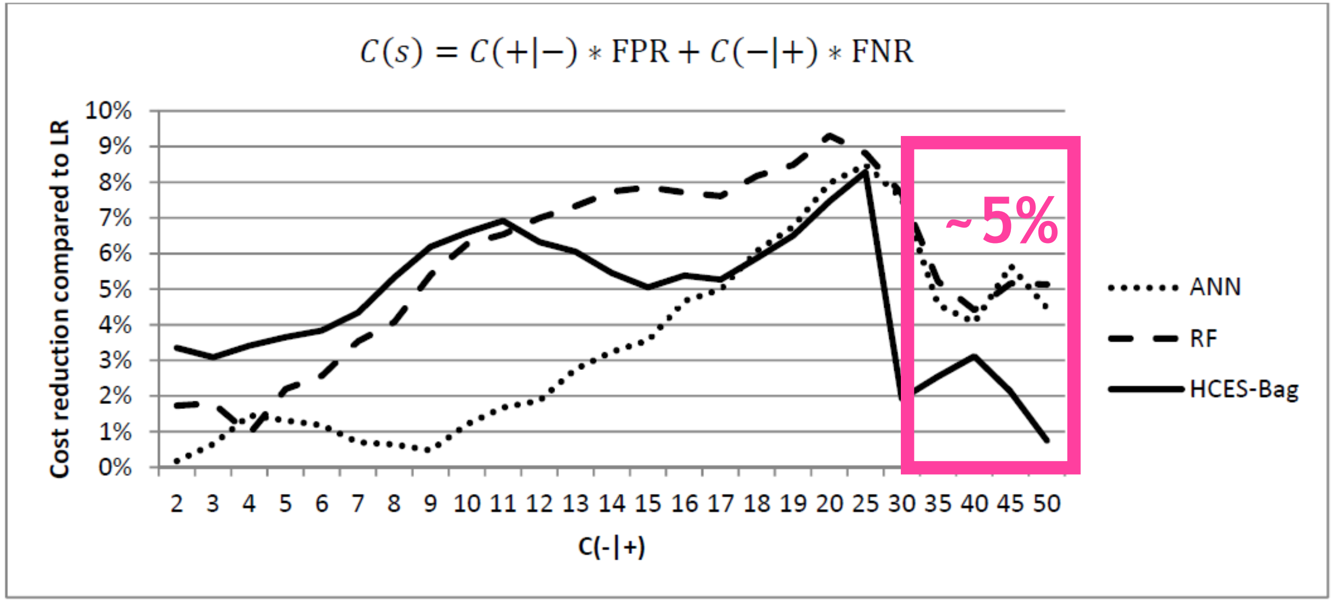
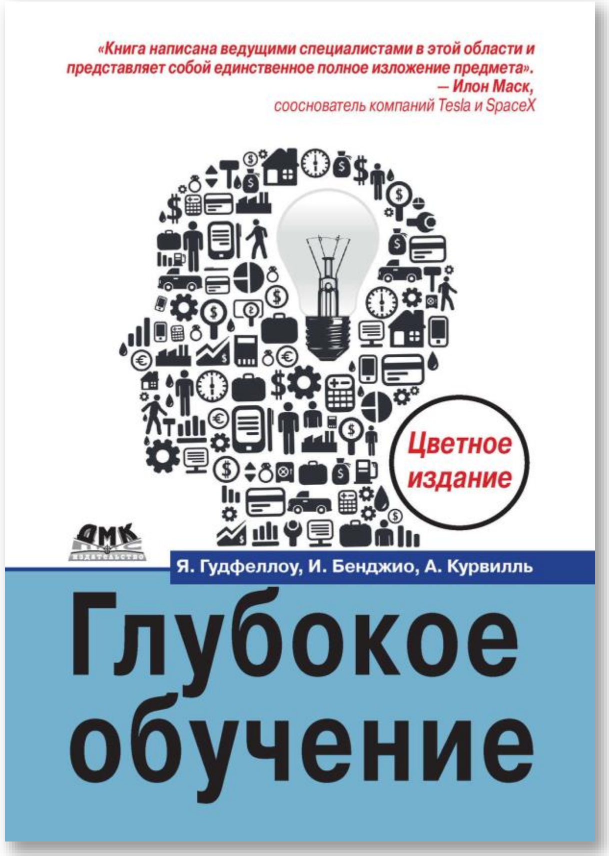


Figure 2: Expected percentage reduction in error costs compared to LR across different settings for $C(-|+)$ assuming $C(+|-) = 1$ and using a Bayes optimal threshold.

Logistic Regression – один из лучших алгоритмов по метрике Gini (левый график)
Но по бизнес-метрикам разница уже существенная (правый график)

– С точки зрения академической науки, когда я начинал заниматься машинным обучением 25 лет назад, тот научный коллектив, в который я пришел еще студентом, в общем-то жил с полной уверенностью, и она была основана на примерно 30-летнем опыте предыдущих исследований, что **задачу можно решать любым методом**.



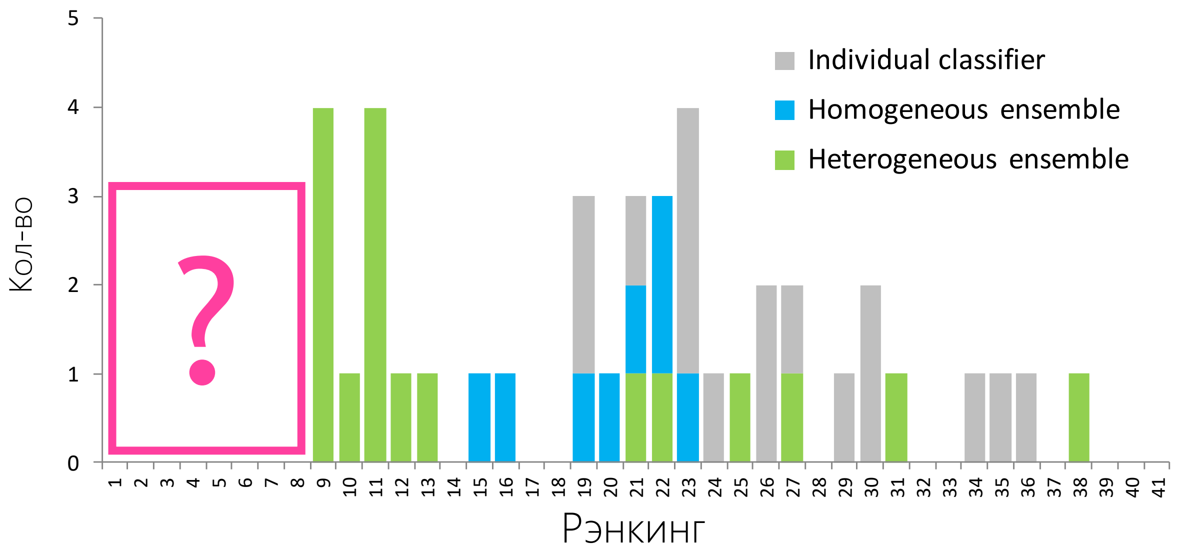
Теорема о бесплатных завтраках

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.

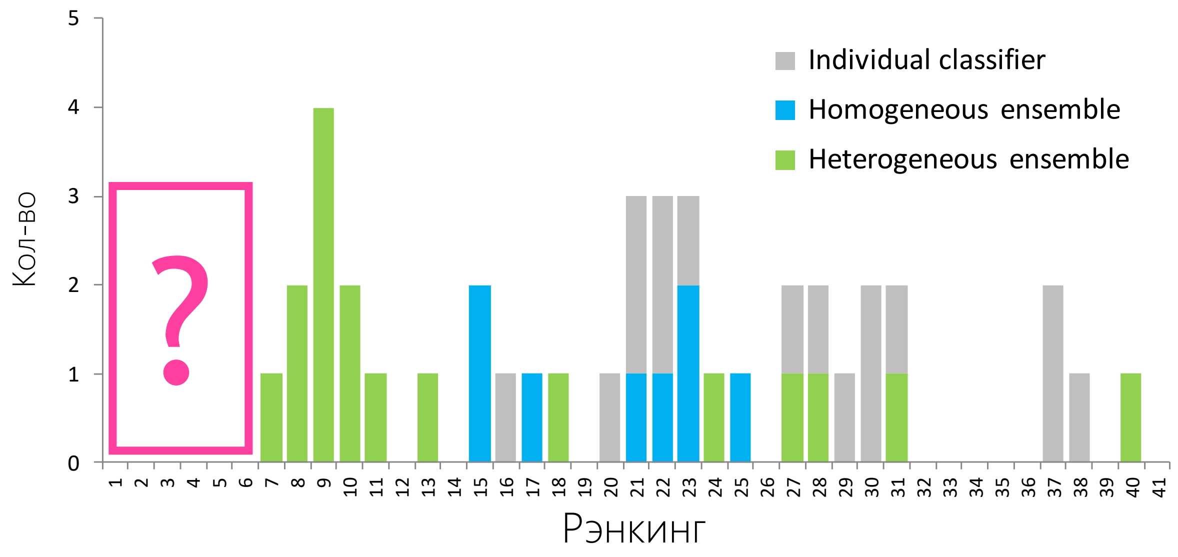
Wolpert, 1996

Бесплатных завтраков не бывает

Средний рэнкинг по всем метрикам



Средний рэнкинг по AUC ROC



Некоторые выводы:

1. **Logistic Regression** – один из лучших **не** ансамблевых алгоритмов.

2. **Ансамбли** на неоднородных моделях круто работают не только на Каггле, но и в задачах скоринга. Однако внедрять ансамбли в онлайне очень сложно и дорого.

3. Теорема «**О бесплатных завтраках**» работает – не существует одного универсального алгоритма под разные датасеты и разные метрики. Нужно экспериментировать.