



Bank of Russia

МОДЕЛЬ ВЕРОЯТНОСТИ ДЕФОЛТА РОССИЙСКИХ КОМПАНИЙ С ИСПОЛЬЗОВАНИЕМ ТРАНЗАКЦИОННЫХ ДАННЫХ

Андрей Шевелев¹

Глеб Бузанов²

(1) Департамент исследований и прогнозирования

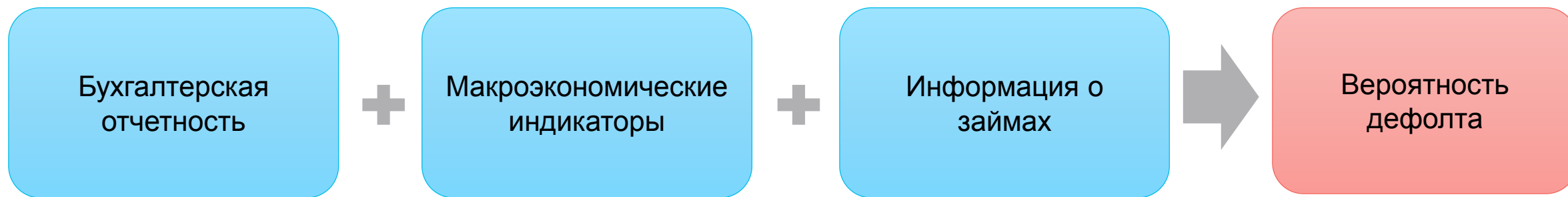
(2) Департамент финансовой стабильности

01.12.2021



Глобальной целью данного проекта является совершенствование существующих моделей Банка России для прогнозирования вероятности дефолта российских компаний с точки зрения качества и частоты принятия решений с использованием транзакционных данных Платежной системы Банка России

Целью данной работы является исследование полезности данных Платежной системы Банка России (ПСБР) для улучшения существующих моделей вероятности дефолта российских компаний



Минусы:

- Редкая публикация данных
- Задержка в публикации данных
- Нет учета связи между агентами

Что можно улучшить?

ПСБР - платежная система, через которую реализуется денежно-кредитная и бюджетная политика Российской Федерации

- Переводы через счета, открытые в Банке России
- ~10 млн транзакций в день

Платежные данные фирмы могут использоваться в качестве информации об изменении ее состояния:

- Платежи по НДС ~ выручка организации
- Платежи по налогу на прибыль ~ прибыль организации
- Платежи по НДФЛ ~ фонд оплаты труда
- Платежи партнеров ~ риск контрагентов
- И т.п.

Ключевые моменты ПСБР:

- Покрытие основной части платежей компаний
- Платежный граф
- Ежедневные данные



A. Khandani, A. Kim, A. Lo (2010): Consumer Credit Risk Models via Machine-Learning Algorithms

- Использование ежемесячных агрегированных транзакционных данных для оценки кредитного риска

H. Kvammea, N. Sellereiteb, K. Aasb, S. Sjursen (2018): Predicting Mortgage Default using Convolutional Neural Networks

- Использование ежедневных транзакционных данных для прогнозирования дефолтов по ипотеке

D. Babaev, M. Savchenko, A. Tuzhilin, and D. Umerenkov (2019): E.T.-RNN: Applying Deep Learning to Credit Loan Applications

V. Shumovskaia, K. Fedyanin, I. Sukharev, D. Berestnev, and M. Panov (2020): Linking Bank Clients using Graph Neural Networks Powered by Rich Transactional Data

- Использование ежедневных транзакционных данных для скоринга предотвращения мошенничества

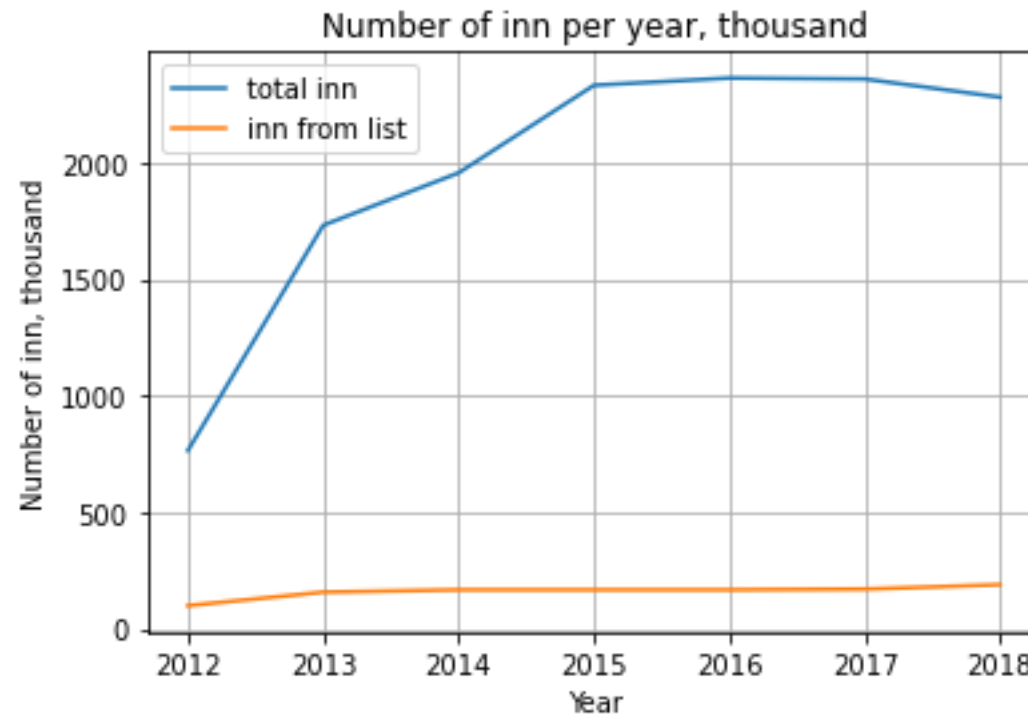
2018												2019												Class in 2018
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
																								not def
															>=90									def
																								X
														>=90					>=90					def
																	>=90							def
																				✓				not def
																		>=90				✓		def

>= 90 days payment overdue

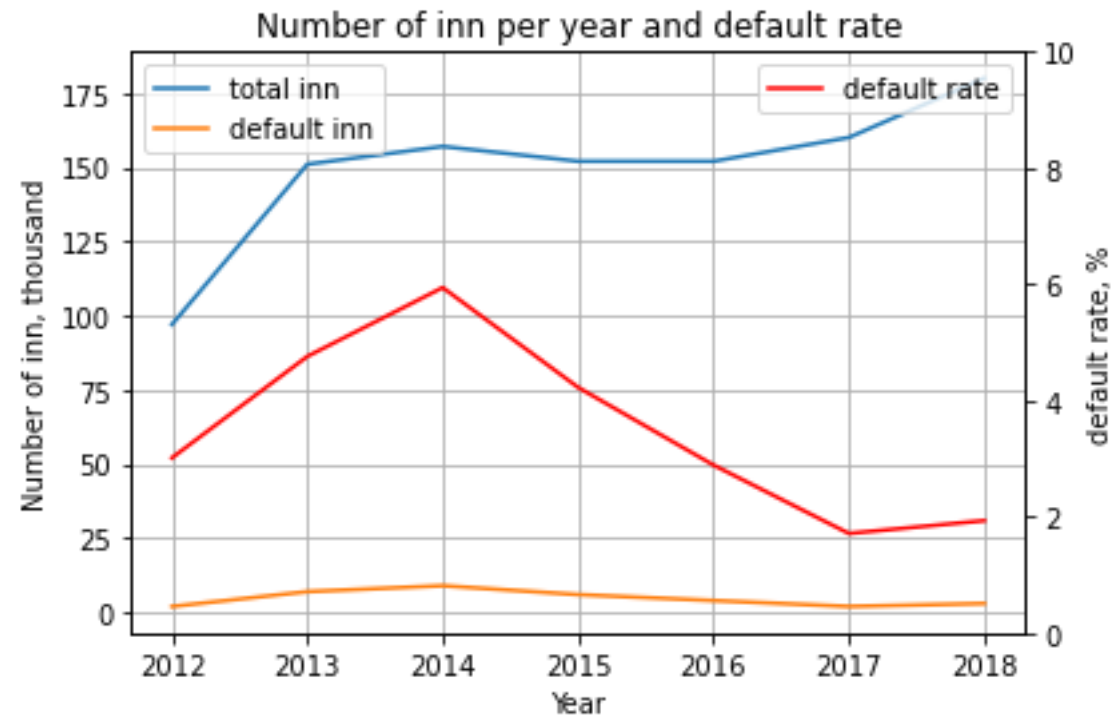
✓ successful loan

- Задолженность более 90 дней [Basel II (III) IRB]

- Бухгалтерская отчетность с 2012 по 2018
- Данные дефолтов за 2013-2019 (до 2018 - БКИ, после – 0409303 форма)
- Более 2 млн уникальных ИНН компаний в год
- ~150 тыс. компаний имею заем в следующем году после отчетного года

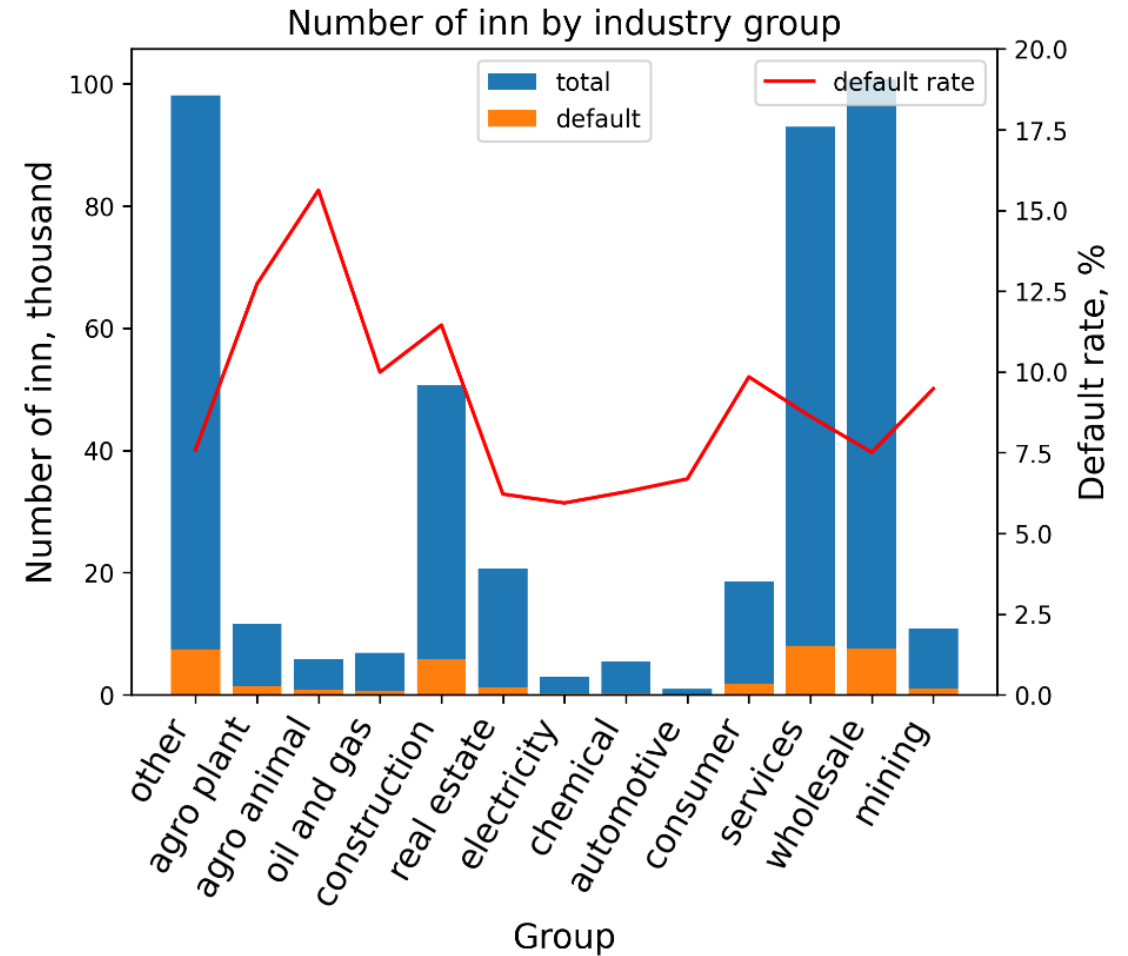


- 100-170 тыс. уникальных ИНН
- уровень дефолтов 2-6%





№	Отрасли
1	АПК растениеводство
2	АПК животноводство
3	Нефтегазовая отрасль
4	Строительство
5	Операции с недвижимым имуществом
6	Электроэнергетика и коммунальный сектор
7	Химическая промышленность
8	Автомобилестроение
9	Потребительский сектор
10	Сфера услуг
11	Оптовая торговля (за исключением торговли топливом и полезными ископаемыми)
12	Металлургия и горная добыча
-1	Прочее





#	Показатель
K1	Коэффициент текущей ликвидности
K2	Рентабельность активов
K3	Валовая рентабельность
K4	Рентабельность по чистой прибыли
K5	Рентабельность по операционной прибыли
K6	Коэффициент автономии
K7	Долг/Прибыль от продаж
K8	Коэффициент покрытия процентов
K9	Коэффициент процентного бремени
K10	Коэффициент соотношения собственных и заемных средств
K11	Коэффициент обеспеченности собственными оборотными средствами
K12	Коэффициент налогового бремени
K13	Краткосрочный долг/Выручка
K14	Оборачиваемость дебиторской задолженности
K15	Оборачиваемость кредиторской задолженности

- Доступна с 2015 года
- ~ 10 млн транзакций в день

pd_main

- oper_dt
- pmt_payment_doc_hk
- inn_in
- inn_out
- amt
- acc_in
- acc_out
- kbk_cd
- pmt_type_kor_cd

pd_reestr

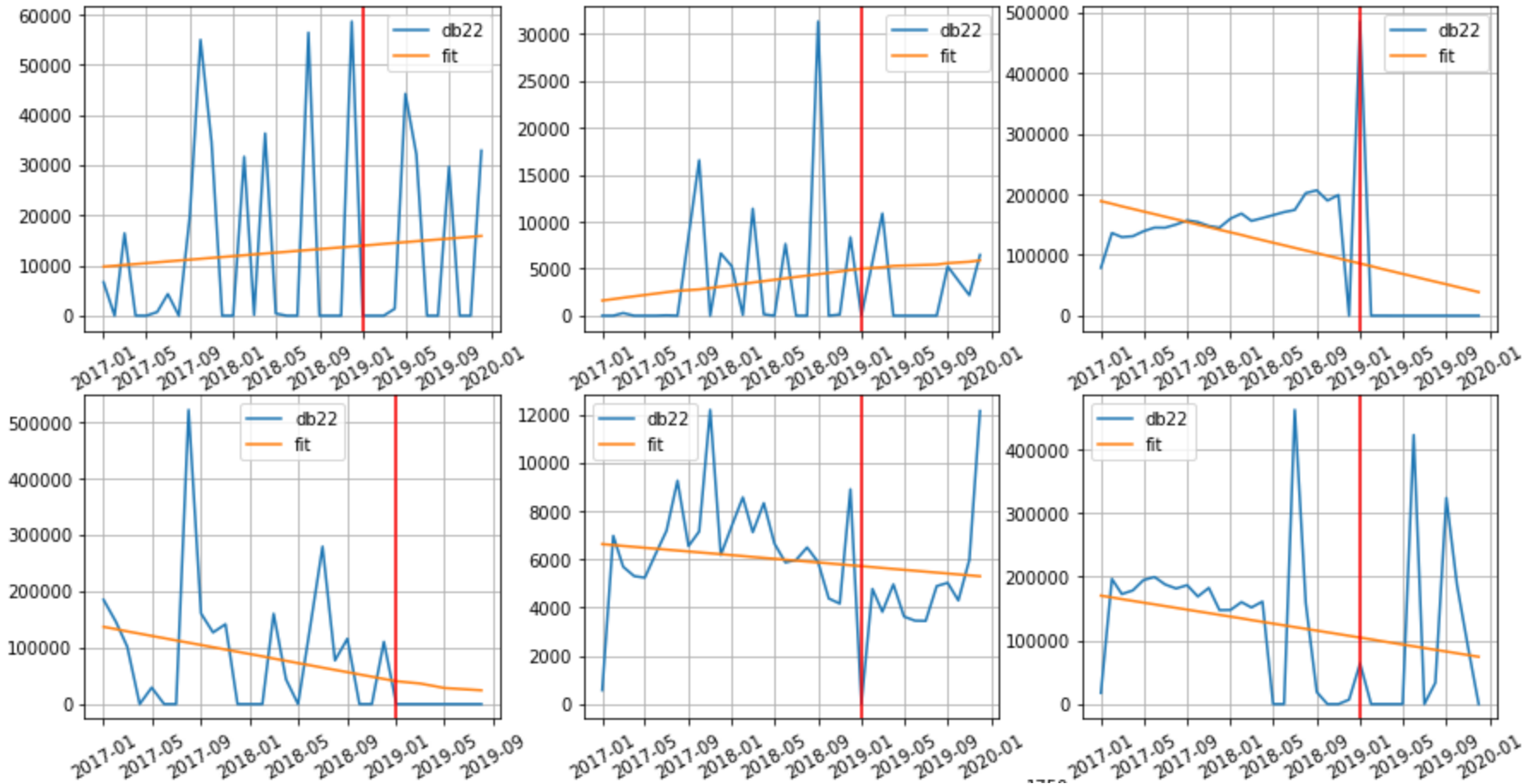
- oper_dt
- reestr_sqn
- pmt_payment_doc_hk
- inn_in
- inn_out
- trx_amt
- acc_in
- acc_out
- trx_nm

pd_status

- oper_dt
- pmt_payment_doc_hk
- pd_status_cd



Пример данных ПСБР: social security payments



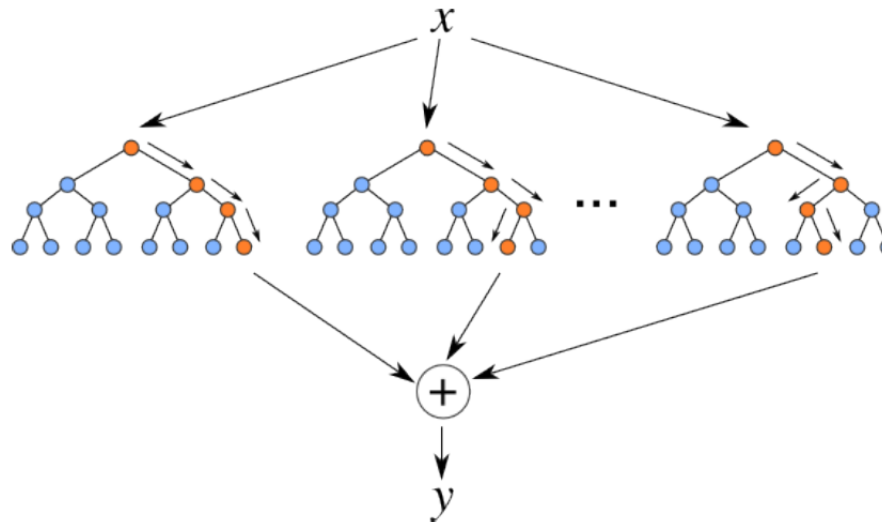


$$\min_{w,c} \left(\frac{1}{2} w^T w + C \sum_{i=1}^n \log \left(\exp \left(-y_i (X_i^T w + c) \right) + 1 \right) \right)$$

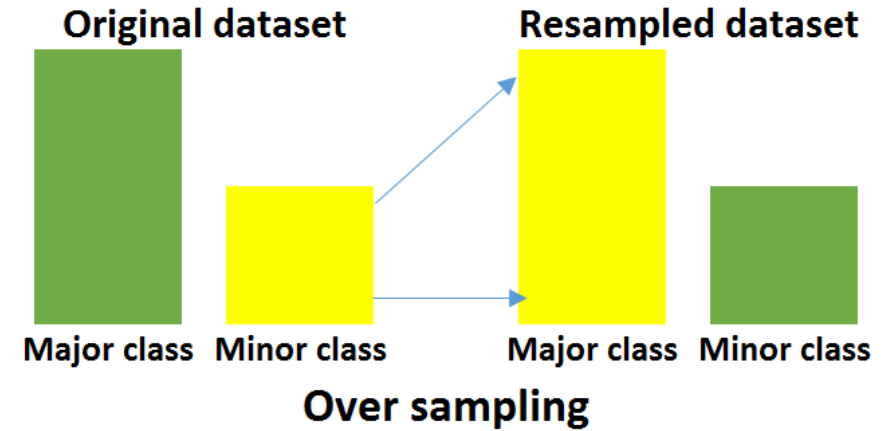
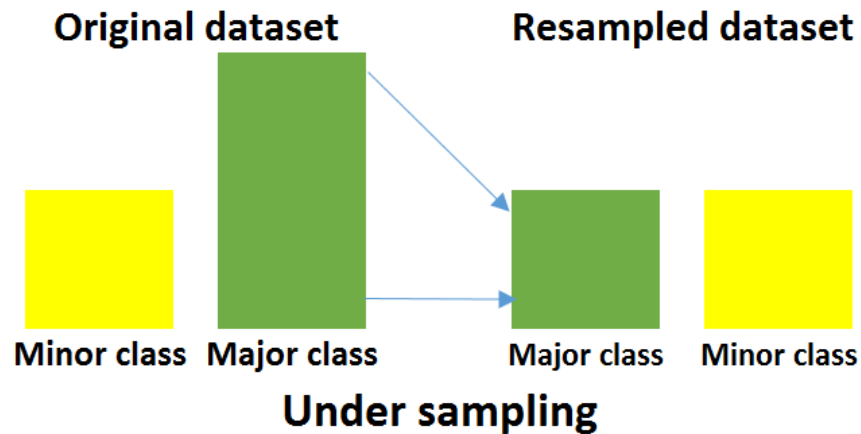
- L2 регуляризация
- p – количество показателей
- $w = (w_0, \dots, w_p)$ – веса модели
- $X = (x_1, \dots, x_p)$ – входные данные
- C – коэффициент обратной силы нормализации
- y_i – таргет

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$$

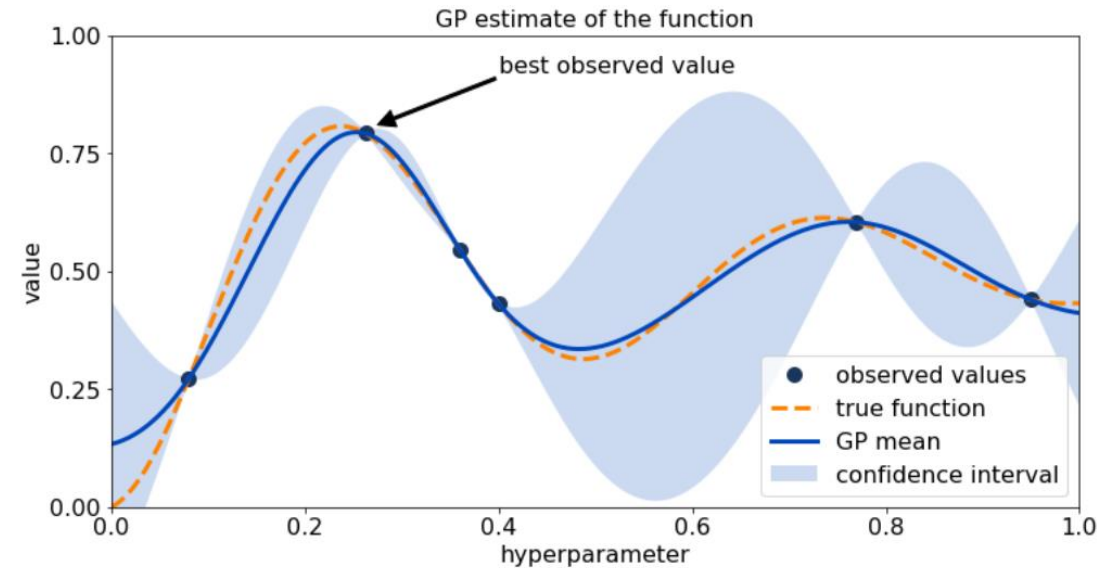
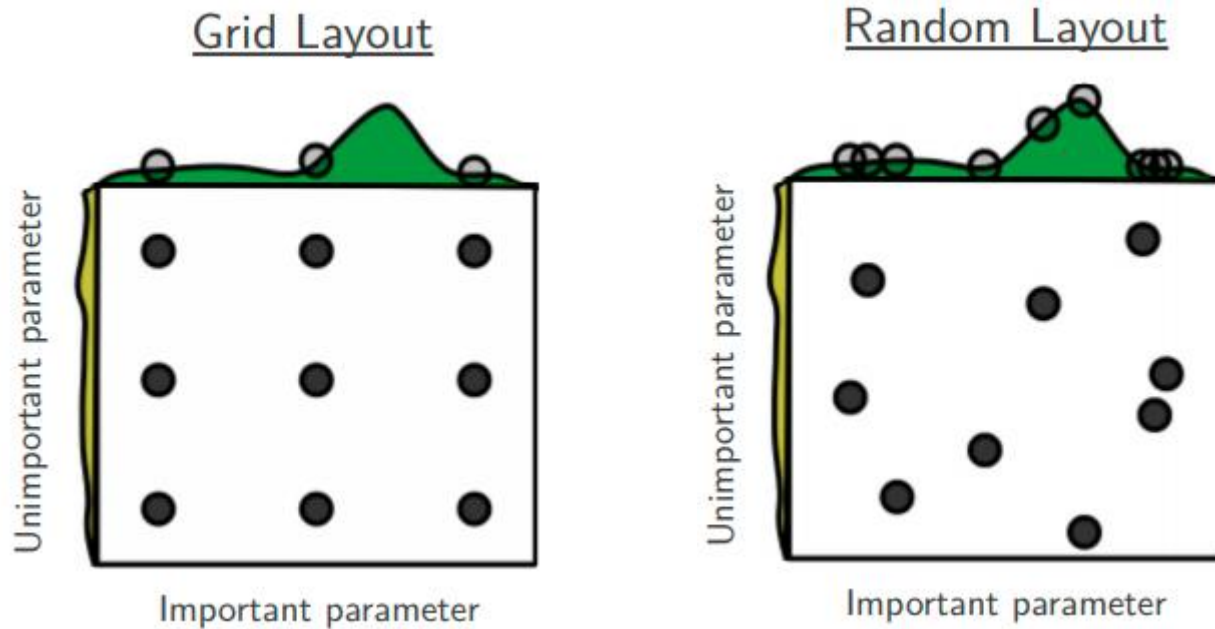
- $H(Q_m)$ – критерий информативности
- $\bar{y}_m$ – полученные оценки
- y – таргет



- Random Under Sampling
- Random Over Sampling
- weighted likelihood function by King and Zeng (2001) для Logistic Regression
- adjust weights Chen et al (2003) для Random Forest

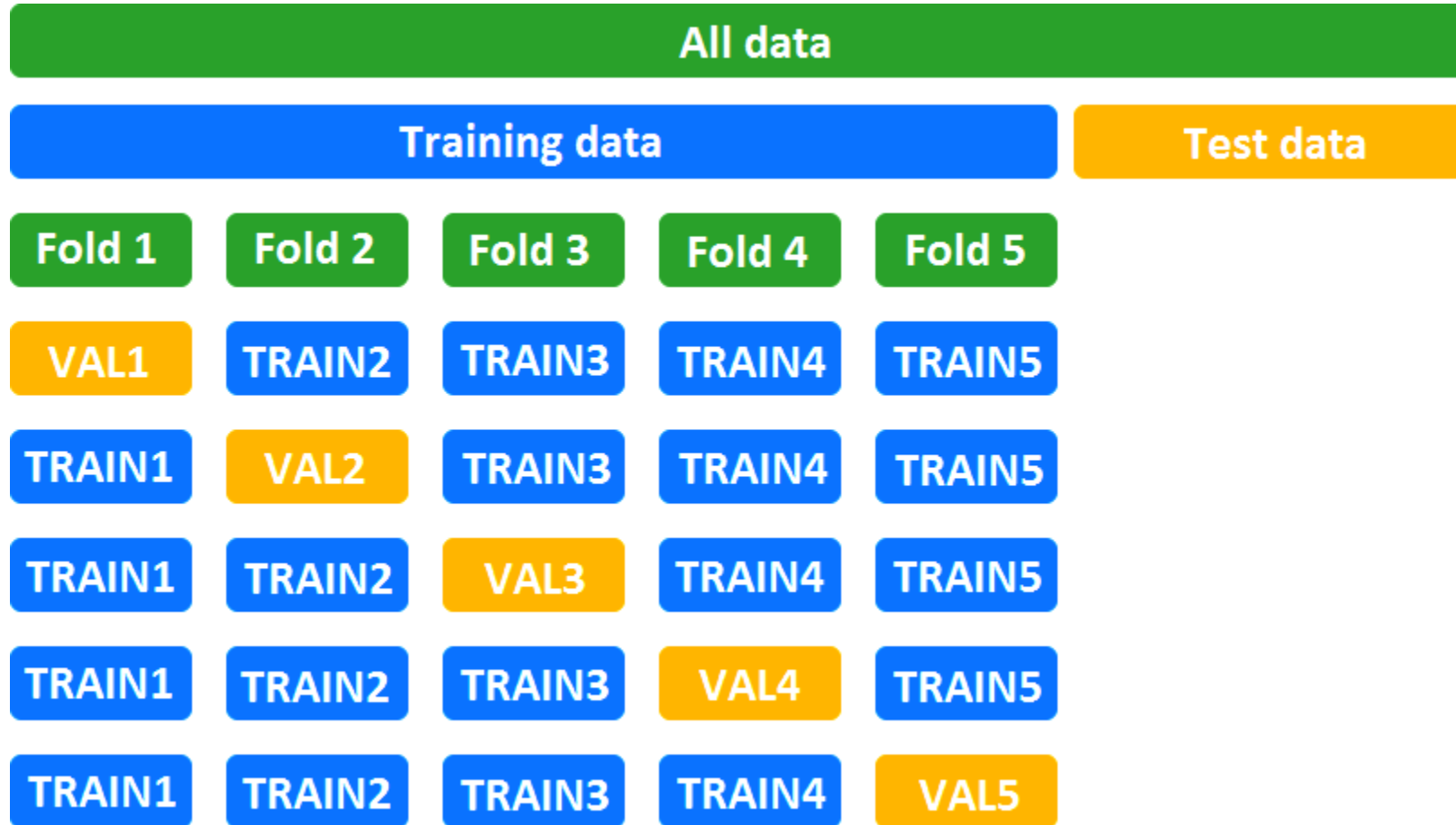


- Grid Search
- Random Search (Bergstra and Bengio (2012))
- Bayesian optimization (Dewancker et al (2016))

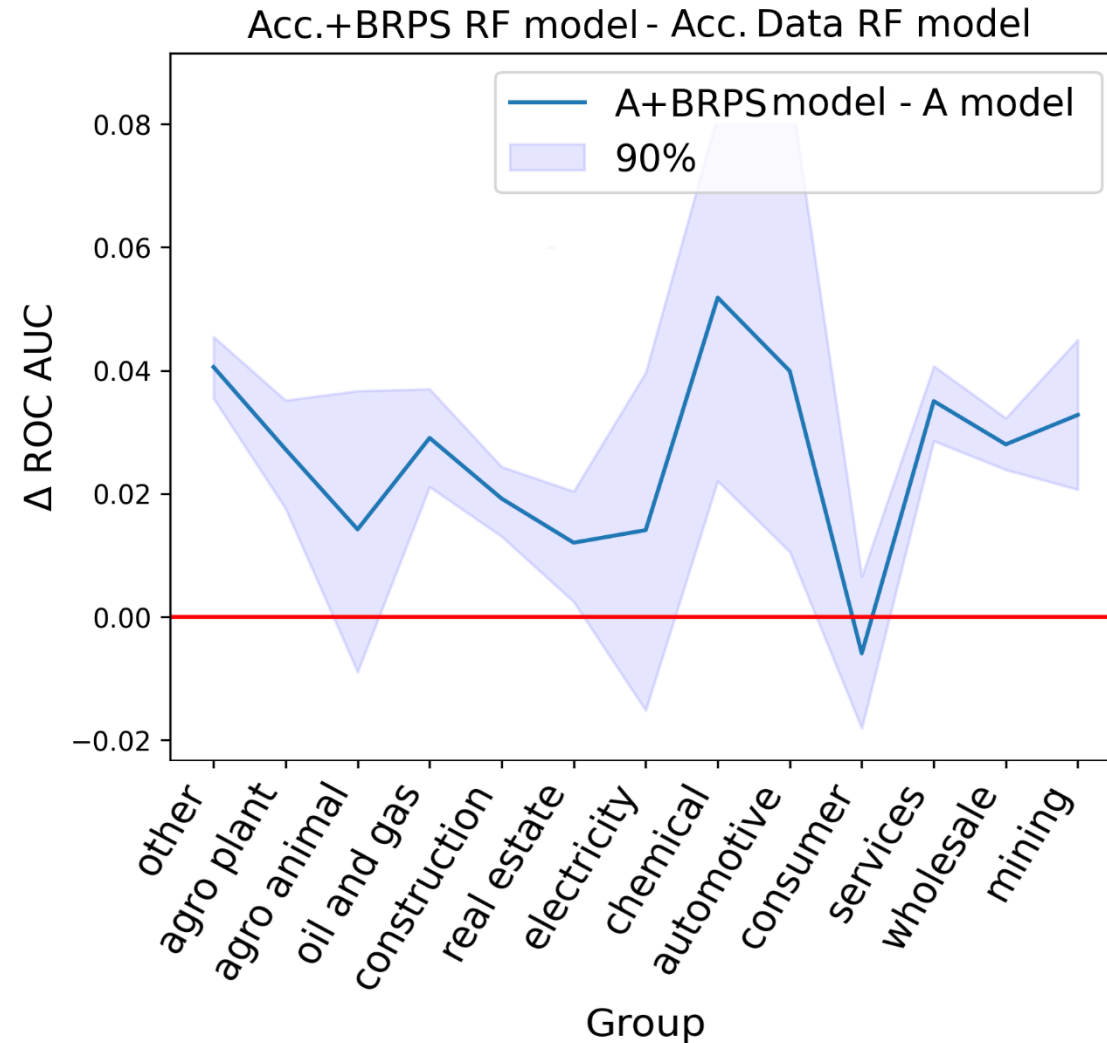


Source: <https://arxiv.org/abs/1911.02501>

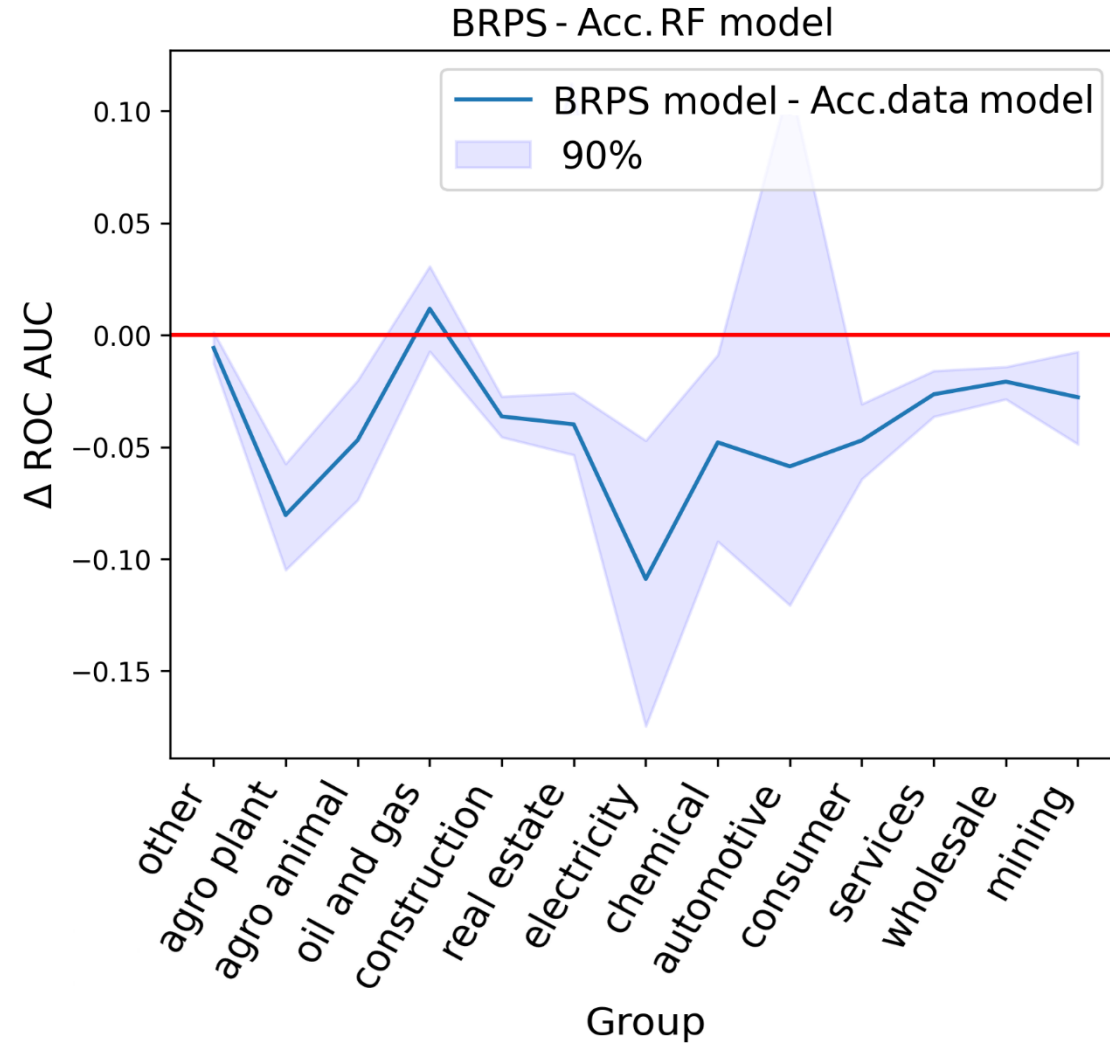
- Stratified k-Fold cross-validation



- Hyper parameters optimization
- Weighted criterion for finding splits



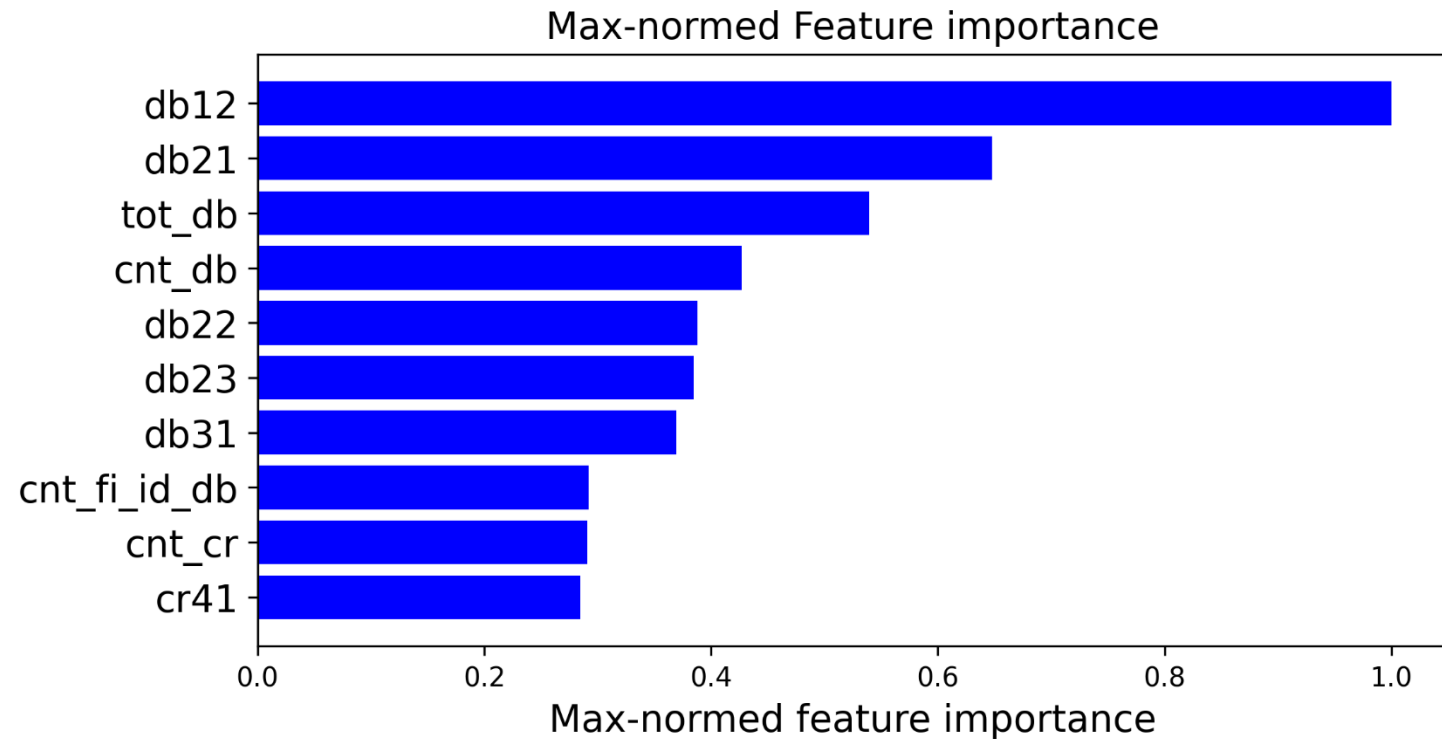
Avg.
ROC AUC:
0.72 vs 0.76



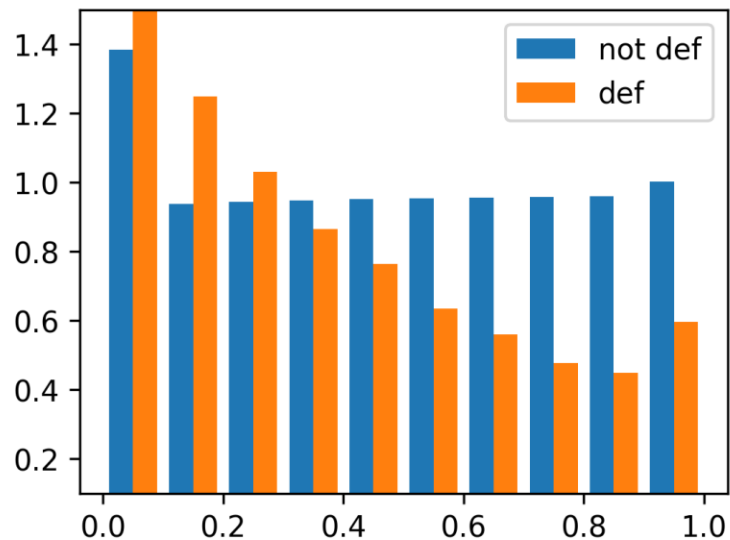
Avg.
ROC AUC:
0.72 vs 0.70

Most important:

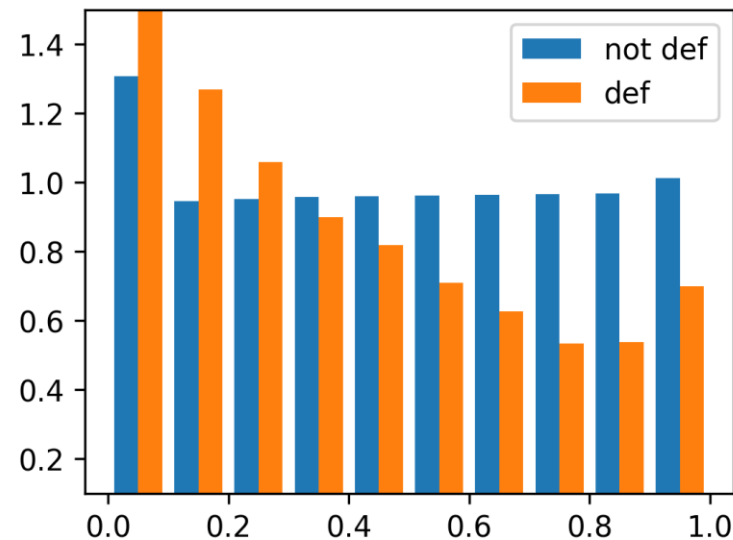
- Платежи НДФЛ
- Платежи в пенсионный фонд
- Оборот по списанию
- Количество операций по списанию
- Платежи в ФСС



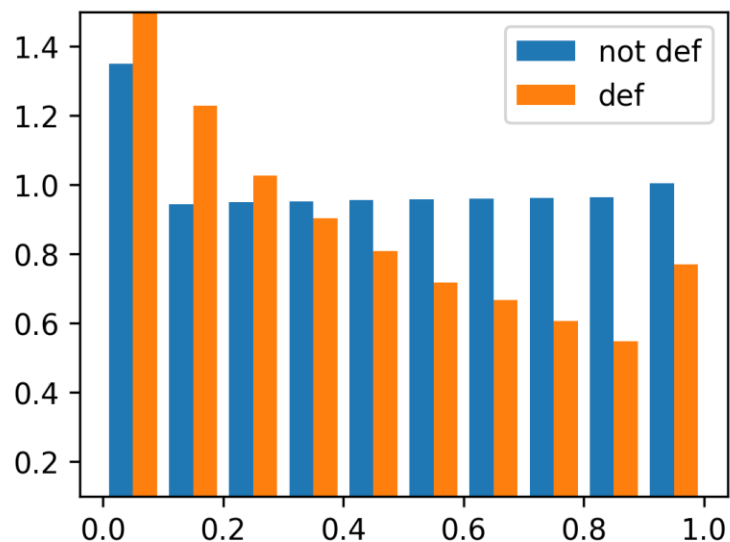
db12 / assets



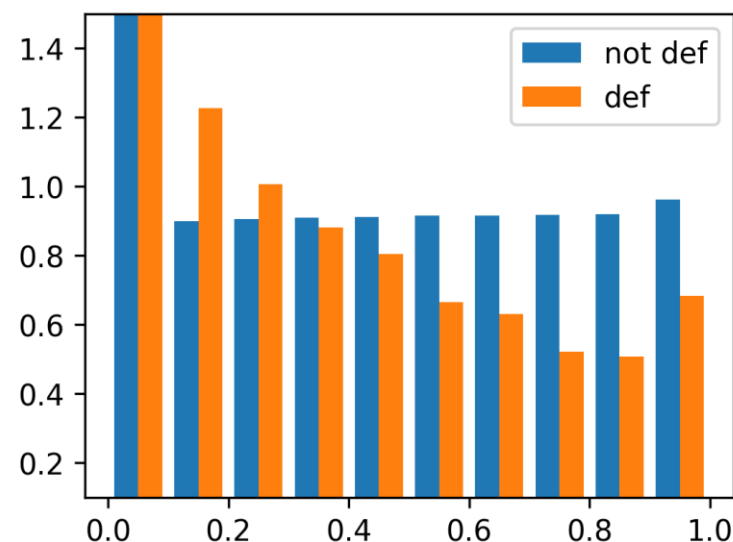
db21 / assets



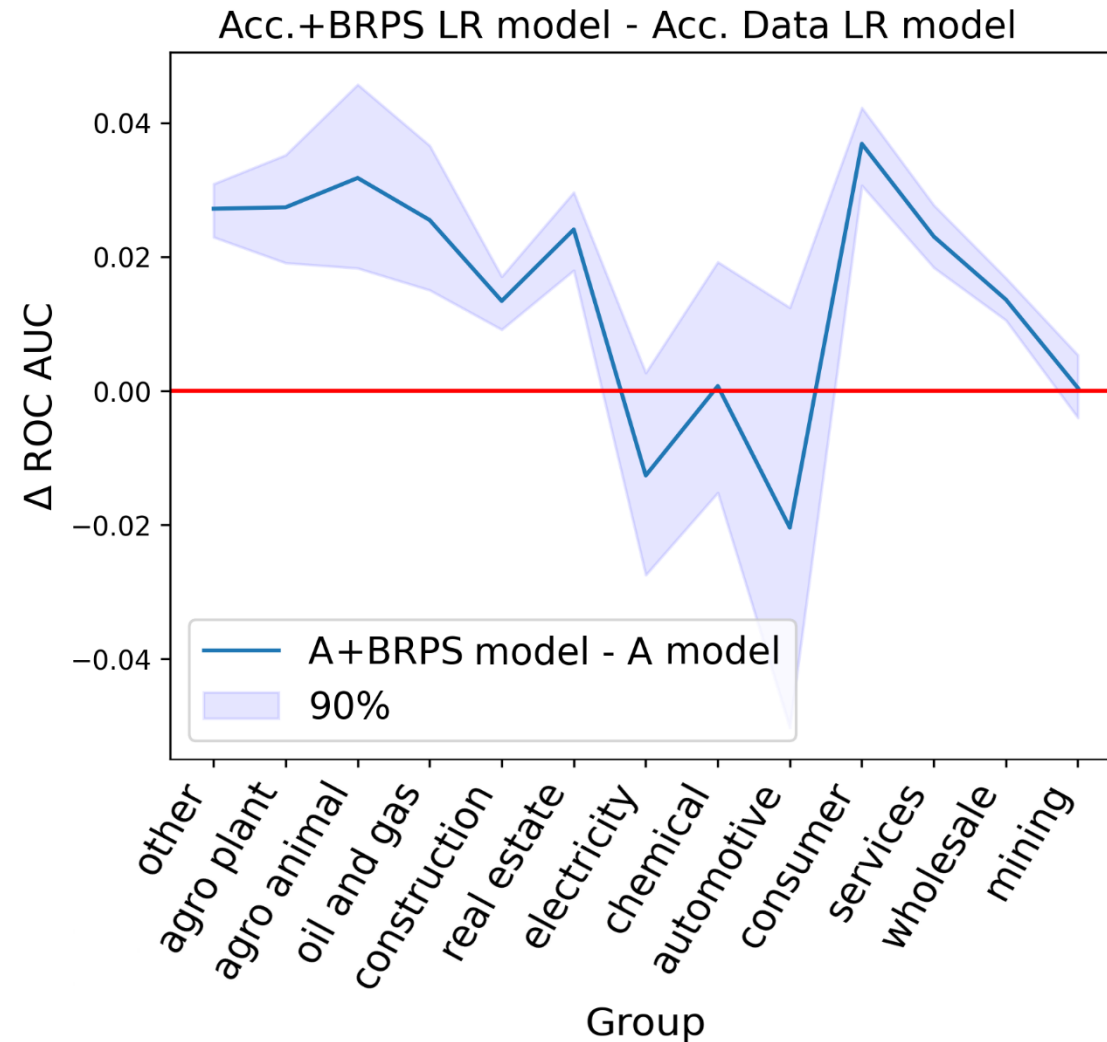
db22 / assets



db23 / assets



- L2 регуляризация
- Weighted likelihood



Combined
score:
0.74 vs 0.76

Результаты:

- Данные ПСБР улучшают качество прогнозирования вероятности дефолта компаний
- Главным источником дополнительной информации являются платежи связанные с фондом оплаты труда
- Модели на данных ПСБР можно использовать для прогнозирования вероятности дефолта компаний для получения более ранних оценок

Дальнейшие направления:

- Переход к ежемесячным оценкам по транзакционным данным
- Использование кросс-валидации с учетом времени и компаний
- Graph NN и другие NN



Bank of Russia

BACKUP SLIDES

- Altman E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609
- Babaev D., Savchenko M., Tuzhilin A., Umerenkov D. E.T.-RNN: Applying Deep Learning to Credit Loan Applications // In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3292500.3330693>
- Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research 13 (2012) 281-305
- Beaver W. (1966). Financial Ratios as Predictors of Failures. Journal of Accounting Research, 4(3) (Supplement), 71-102.
- Beaver W. (1968). Market prices, financial ratios and prediction of failure. Journal of Accounting Research, 6(2), 179-192.
- Breiman L. Random Forests. Machine Learning 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- Byrd R.H., Lu P. and Nocedal J. A Limited Memory Algorithm for Bound Constrained Optimization, (1995), SIAM Journal on Scientific and Statistical Computing , 16, 5, pp. 1190-1208
- Chakraborty C., Joseph A. Machine Learning at Central Banks // Bank of England Working Paper No. 674. 2007. 89p.
- Chen C., Liaw A., Breiman L. Using Random Forest to Learn Imbalanced Data // Research article, UC Berkeley (2003), pp. 1-12
- Demeshev B., Tikhonova A. (2014). Default prediction for Russian companies: Intersectoral comparison // HSE Economic Journal, Vol. 18, No. 3, pp. 359—386. (In Russian)
- Dewancker I, McCourt M., Clark S. Bayesian Optimization for Machine Learning : A Practical Guidebook // arXiv.org > cs > arXiv:1612.04858
- Ganganwar V. An overview of classification algorithms for imbalanced datasets // International Journal of Emerging Technology and Advanced Engineering. 2012, 2(4), 42-47.
- Haibo He, Yunqian Ma Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons, Jun 7, 2013 - Technology & Engineering - 216 p.

- Hastie T, Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York : Springer, c2009. 745 p.
- Jackson R., Wood A. The performance of insolvency prediction and credit risk models in the UK: A comparative study // The British Accounting Review Volume 45, Issue 3, September 2013, Pages 183-202
- Karminsky A.M., Burekhin R.N. (2019) Comparative analysis of methods for forecasting bankruptcies of Russian construction companies. Business Informatics, vol. 13, no 3, pp. 52–66. DOI: 10.17323/1998-0663.2019.3.52.66
- King G., Zeng L. (2001). Logistic regression in rare events data. Political Analysis, Vol. 9, No. 2, pp. 137-163.
- Louppe G., “Understanding Random Forests: From Theory to Practice”, PhD Thesis, U. of Liege, 2014.
- Merton R.C. (1974). On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance, 29(2), 449-470.
- Mogilat A. Bankruptcy in Russian real sector: Basic tendencies and financial indicators of a typical bankrupt // Nauchnye Trudy INP RAN, No. 13. 2015. pp. 156—186. (In Russian)
- Mogilat A. Modelling financial distress of Russian industrial companies, or What bankruptcy analysis can tell // Voprosy Ekonomiki. 2019;(3):101-118. (In Russ.) <https://doi.org/10.32609/0042-8736-2019-3-101-118>
- Odom M.D., Sharda R. (1990, June). A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on neural networks (pp. 163-168). IEEE.
- Ohlson J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 109-131.
- Shibitov D., Mamedli M. (2019, August). The finer points of model comparison in machine learning: forecasting based on Russian banks' data. Retrieved from Bank of Russia Working Paper Series No. 43: http://www.cbr.ru/content/document/file/87572/wp43_e.pdf
- Sonak A., Patankar R. A. A survey on methods to handle imbalance dataset // Int. J. Comput. Sci. Mobile Comput. 2015. 4(11), 338-343.
- Strobl, C., Boulesteix, AL., Zeileis, A. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007). <https://doi.org/10.1186/1471-2105-8-25>