# Прогнозирование банкротств предприятий с помощью методов машинного обучения.

**Abstract**

Использование моделей прогнозирования банкротства предприятий для управления инвестиционными рисками лежит в основе управленческой деятельности. В большинстве исследований моделей банкротств используется традиционные методы статистики (например, дискриминантный анализ и логистическая регрессия). В данном исследовании основной акцент делается на методах машинного обучения для прогнозирования банкротства за один, два, три и четыре года до события и сравнивается их эффективность с результатами дискриминантного анализа, логистической регрессии и нейронных сетей.

# 1 Введение

Инвестиционные риски является основной проблемой для банков и инвесторов и заставляет их проверять предприятия и контролируют их финансовую платежеспособность. Как следствие, существует острая потребность в аналитических инструментах, способных предсказать банкротство среди предприятий, в которые инвесторы готовы вкладывать свои инвестиции. Оценки прогноза банкротства предприятий является не простой задачей финансового анализа, что делает ее актуальной областью исследования. Волатильность и непредсказуемость рынков создают большую неопределенность и существенно усложняют решение этой задачи.

В большинстве исследований прогноз банкротства рассматривается как проблема бинарной классификации. Целевая переменная моделей обычно принимает два значения: 0 (платежеспособное предприятие) и 1 (предприятие банкрот). Традиционно для оценки банкротства используются дискриминантный анализ [1-5], регрессионный анализ [5-7]. Альтман использовал линейный дискриминантный анализ (ЛДА) для того, чтобы различать предприятия банкротов от не банкротов [1]. ЛДА применяет линейную комбинацию показателей для определения рейтинга предприятия. Эта оценка затем применяется для разделения предприятий на банкротов и не банкротов. В работе [2] ЛДА используется для прогнозирования риска банкротства российских предприятий. Применимость ЛDA и квадратичного дискриминантного анализа исследована в работах [4, 5]. В работе [5] используется логистический регрессионный анализ к оценке банкротства, но в отличие от модели Альтмана [1] модель Олсона определяет вероятность банкротства. Однако, несмотря на относительную легкость проведения дискриминантного анализа и логистической регрессии, использование этих моделей показывает плохую способность к обобщению и низкую точность прогнозирования [8]. Поэтому в дальнейшем многие исследователи используют методы машинного обучения, такие как нейронные сети [9-12], метод опорных векторов [13-14], бустинговые методы [15-16].

Традиционный способ построения моделей базируется на классификации данных имеющих один период. При этом модель определяет вероятность банкротства, используя расстояние в пространстве финансовых показателей между финансовым положением данного предприятия и стандартной ситуацией финансового банкротства. Однако для инвесторов способность модели правильно прогнозировать судьбу компаний в среднесрочной перспективе имеет важное значение, так как эти модели используются для принятия решения о том, следует ли вкладывать инвестиции, исходя из вероятности того, что предприятие не обанкротится. И та как большинство инвестиций являются долгосрочными, то модель должна уметь оценивать платежеспособность компании в течение всего срока действия инвестиций.

Тем не менее, было сделано много попыток для улучшения точности модели на среднесрочном горизонте (5 лет). Были исследованы различные способы построения моделей. Ряд авторов использовали данные, измеренные в течение нескольких последовательных лет в сочетании с традиционными методами классификации. Ряд исследователей использовали методы, в которых модель опирается на набор правил, разработанных с использованием таких методов, как багинговые или бустинговые методы (Huang, Tang, Lee, & Chang, 2012; Sun, Jia & & Li, 2011) или методы, в которых правила строятся с использованием различных методов классификации (Geng et al., 2015). Другие использовали модели с несколькими правилами, где каждое правило априори специализировалось в определенной области пространства принятия решений (du Jardin, 2015). Все эти работы показывают, что многопериодные модели, как правило, предлагают краткосрочные прогнозы (1 год), которые являются более точными, чем те, которые были достигнуты при использовании однопериодных моделей, но без предоставления более качественных среднесрочных прогнозов. Тот же вывод справедлив для моделей, которые основаны на наборе правил: на краткосрочном горизонте они более точны, чем модели с одним правилом, но не на среднесрочном горизонте. Тем не менее, модели, которые априори специализируются по регионам пространства принятия решений, обеспечивают лучшие среднесрочные прогнозы, чем традиционные модели с одним правилом, но не за пределами 3-летнего горизонта и без повышения точности модели на 1-летнем горизонте.

# 2 Методология

# 3 Выбор данных.

## 3.1 Набор данных.

В данной статье, данные были взяты с сайта UCI machine learning repository. Изначально они были извлечены с ресурса Emerging Markets Information Services (EMIS, securities.com), который представляет собой базу данных, содержащую информацию о развивающихся рынках по всему миру. Обанкротившиеся компании были проанализированы в период 2000-2012 годов, в то время как все еще действующие компании были оценены с 2007 по 2013 год Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Основные метаданные готового набора данных представлены в таблице 1.

Таблица 1. Основные характеристики набора данных.

|  |  |
| --- | --- |
| Название | Описание |
| Характеристика набора данных | Многофакторный |
| Характеристика показателей | Реальные |
| Тип задачи | Классификация |
| Количество объектов | 10503 |
| Количество показателей | 64 |
| Пропущенные значения | Да |
| Сфера деятельности | Бизнес/Экономика |

В каждом наборе данных содержатся 64 финансовых показателя + класс прогнозирования (0 – не банкрот, 1 – банкрот)

*1st year* – данные содержат финансовые показатели с 1-го года прогнозируемого периода и метку соответствующего класса, которая указывает на статус банкротства через 5 лет.

*2nd year* – данные содержат финансовые показатели за 2-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 4 года.

*3rd year* – данные содержат финансовые показатели за 3-ий год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 3 года.

*4th year* – данные содержат финансовые показатели за 4-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 2 года.

*5th year* – данные содержат финансовые показатели за 5-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 1 год.

Расшифровка финансовых показателей финансовых показателей представлена в таблице 2.

Таблица 2 – Описание финансовых показателей

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Описание** | **ID** | **Описание** |
| X1 | Чистая прибыль/Суммарные активы | X33 | Операционные расходы/ Краткосрочные обязательства |
| X2 | Сумма всех обязательств/Суммарные активы | X34 | Операционные расходы/ Суммарные обязательства |
| X3 | Текущие активы/ Суммарные активы | X35 | Прибыль от продаж/ Суммарные активы |
| X4 | Текущие активы/ Краткосрочные обязательства | X36 | Суммарные продажи/ Суммарные активы |
| X5 | [(Наличные ДС + краткосрочные инвестиции + дебиторская задолженность - Краткосрочные обязательства)/ (операционные расходы - амортизация)] | X37 | (Текущие активы – запасы)/ Долгосрочные обязательства |
| X6 | Нераспределенная прибыль/ Суммарные активы | X38 | (Постоянный капитал)/ Суммарные активы |
| X7 | Прибыль до налогов и процентов/ Суммарные активы | X39 | Прибыль от продаж/ продажи |
| X8 | Балансовая стоимость капитала/ Суммарные обязательства | X40 | (Текущие активы – запасы – дебиторская задолженность)/ Краткосрочные обязательства |
| X9 | Продажи/ Суммарные активы | X41 | Суммарные обязательства/(( Прибыль от операционной деятельности + амортизация) \* (12/365)) |
| X10 | Капитал/ Суммарные активы | X42 | Прибыль от операционной деятельности/ Продажи |
| X11 | (Валовая прибыль + иные активы и транзакции + финансовые затраты)/ Суммарные активы | X43 | ротация дебиторской задолженности + товарооборот (в днях) |
| X12 | Валовая прибыль/ Краткосрочные обязательства | X44 | (дебиторская задолженность \* 365) / Продажи |
| X13 | (Валовая прибыль + амортизация) / Продажи | X45 | Чистая прибыль/ запасы |
| X14 | (Валовая прибыль + выплаты по кредиту)/ Суммарные активы | X46 | (Текущие активы - запасы) /Краткосрочные обязательства |
| X15 | (Суммарные обязательства \* 365)/ (Валовая прибыль + амортизация) | X47 | (Запасы \* 365)/ Издержки проданной продукции |
| X16 | (Валовая прибыль + амортизация)/ Суммарные обязательства | X48 | прибыль до вычета процентов, налогов и амортизации (Прибыль от операционной деятельности – амортизация) / Суммарные активы |
| X17 | Суммарные активы/ суммарные обязательства | X49 | прибыль до вычета процентов, налогов и амортизации (Прибыль от операционной деятельности – амортизация) / Продажи |
| X18 | Валовая прибыль/ Суммарные активы | X50 | Текущие активы/ Суммарные обязательства |
| X19 | Валовая прибыль/ Продажи | X51 | Краткосрочные обязательства/ Суммарные активы |
| X20 | (Инвентарь \* 365) / Продажи | X52 | (Краткосрочные обязательства\*365)/ Издержки на проданную продукцию |
| X21 | (продажи (n) / продажи (n-1)) | X53 | Капитал/ Основные средства |
| X22 | Прибыль от операционной деятельности/ Суммарные активы | X54 | Постоянный капитал/ основные средства |
| X23 | Чистая прибыль/ Продажи | X55 | Оборотный капитал |
| X24 | Валовая прибыль (за 3 года)/ | X56 | (Продажи – стоимость изготовления проданной продукции) / Продажи |
| X25 | (Капитал – акционерский капитал)/ Суммарный капитал | X57 | (Текущие активы - запасы – краткосрочные обязательства)  / (Продажи – Валовая прибыль - амортизация) |
| X26 | (Чистая прибыль + амортизация)/ Суммарные обязательства | X58 | Суммарные издержки/ Суммарные продажи |
| X27 | (Прибыль от операционной деятельности)/ Финансовые расходы | X59 | Долгосрочные обязательства/ Капитал |
| X28 | Оборотный капитал/ основные средства | X60 | Продажи/ Запасы |
| X29 | Логарифм от суммарных активов | X61 | Продажи/ Дебиторская задолженность |
| X30 | (Суммарные обязательства – наличные ДС)/продажи | X62 | (Краткосрочные обязательства \* 365)/ продажи |
| X31 | (Валовая прибыль + выплаты по кредитам)/ продажи | X63 | Продажи/ Краткосрочные обязательства |
| X32 | (Текущие обязательства \* 365)/ затраты на проданную продукцию | X64 | Продажи/ Основной капитал |

## 3.2 Предварительная обработка данных

Используемый для исследования набор данных имеет 2 проблемы:

* Неполноценность данных (большое количество пропущенных значений);
* Несбалансированность классов.

Для решения первой проблемы использовалась техника замены пропущенных величин с помощью среднего арифметического значения.

Так как данные являются несбалансированными (успешных предприятий намного больше, чем предприятий банкротов), для исправления дисбаланса была использована техника Synthetic Minority Oversampling Technique (SMOTE). Класс меньшинства дополняется новыми объектами, путем взятия каждого примера выборки и введения искусственных примеров вдоль отрезков, соединяющих между любой или все k-ближайших соседей. В зависимости от того какое количество дополнения набора данных необходимо, объекты из k-ближайших соседей выбираются случайным образом Nitech V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002).

# 4 Эксперимент.

## 4.1 Цель эксперимента

Целью данного эксперимента является определение лучшей по точности прогнозирования модели и оценка важности финансовых показателей. Мы приняли во внимание следующие традиционные модели и алгоритмы машинного обучения в задачах классификации:

* Linear Discriminant Analysis (LDA),
* Logistic Regression (LR),
* Decision Tree Classifier (DT),
* Random Forest Classifier (RDF),
* Extreme gradient boosting classifier (XGboost),
* Natural gradient boosting classifier (Ngboost).

Для каждой модели были определены оптимальные гиперпараметры с помощью техник Grid Search и Random Grid Search. Также использовалась 10-ти блочная перекрестная проверка.

Для предварительной обработки данных использовались следующие библиотеки языка программирования python: imblearn, sklearn, numpy и pandas. При тестировании различных алгоритмов также были использованы такие библиотеки как: sklearn и xgboost.

## 4.2 Результаты исследования

Результаты эксперимента представлены в таблицах 3,4,5,6,7. В них содержится следующая информация: среднее значение точности модели (**MN**), стандартное отклонение (**STD**), точность прогнозирования до сих пор действующих предприятий (**N**), точность прогнозирования предприятий банкротов (**B**).

Таблица 3 – результаты прогнозирования по первому году.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название модели | **MN** | **STD** | **N** | **B** |
| **LDA** | 0.961 | 0.009 | 0.969 | 0.455 |
| **LR** | 0.961 | 0.010 | 0.972 | 0.010 |
| **DT** | 0.962 | 0.006 | **0.982** | 0.487 |
| **RF** | 0.977 | 0.005 | 0.981 | 0.860 |
| **XGboost** | **0.980** | 0.005 | 0.981 | **0.894** |

Таблица 4 – результаты прогнозирования по второму году.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название модели | **MN** | **STD** | **N** | **B** |
| **LDA** | 0.956 | 0.008 | 0.967 | 0.100 |
| **LR** | 0.957 | 0.008 | 0.963 | 0.010 |
| **DT** | 0.951 | 0.008 | 0.975 | 0.411 |
| **RF** | 0.971 | 0.007 | **0.976** | 0.630 |
| **XGboost** | **0.974** | 0.008 | **0.976** | **0.914** |

Таблица 5 – результаты прогнозирования по третьему году.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название модели | **MN** | **STD** | **N** | **B** |
| **LDA** | 0.948 | 0.11 | 0.964 | 0.125 |
| **LR** | 0.950 | 0.009 | 0.963 | 0.013 |
| **DT** | 0.951 | 0.009 | **0.972** | 0.479 |
| **RF** | 0.965 | 0.010 | 0.9973 | 0.646 |
| **XGboost** | **0.970** | 0.008 | **0.972** | **0.904** |

Таблица 6 – результаты прогнозирования по четвертому году.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название модели | **MN** | **STD** | **N** | **B** |
| **LDA** | 0.941 | 0.009 | 0.962 | 0.138 |
| **LR** | 0.944 | 0.007 | 0.959 | 0.101 |
| **DT** | 0.944 | 0.008 | 0.968 | 0.475 |
| **RF** | 0.960 | 0.005 | **0.970** | 0.732 |
| **XGboost** | **0.965** | 0.007 | 0.969 | **0.899** |

Таблица 7 – результаты прогнозирования по пятому году.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название модели | **MN** | **STD** | **N** | **B** |
| **LDA** | 0.930 | 0.011 | 0.9535 | 0.376 |
| **LR** | 0.929 | 0.011 | 0.958 | 0.436 |
| **DT** | 0.942 | 0.011 | 0.970 | 0.584 |
| **RF** | 0.963 | 0.006 | **0.971** | 0.805 |
| **XGboost** | **0.966** | 0.006 | 0.970 | **0.874** |

По результатам эксперимента, можно сделать вывод о том, что в среднем алгоритм **XGboost** показывает лучший результат по всем показателям. К тому же стоит заметить, что прогнозирование предприятий банкротов у данной модели намного превосходит все остальные модели.

## 4.2 Оценка важности показателей

# Ссылки

1. E.I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” The journal of finance, vol. 23, no. 4, pp. 589–609, 1968

2. L. Lugovskaya, “Predicting default of Russian SMEs on the basis of financial and nonfinancial variables,” Journal of Financial Services Marketing, vol. 14, no. 4, pp. 301–313, 2010

3. Deakin, E. (1972). A Discriminant Analysis of predictors of Business Failure. Journal of Accounting Research, 10(1), 167-179.

4. Antunesa F., Ribeiroa B., Pereirab F. Probabilistic modeling and visualization for bankruptcy prediction Applied Soft Computing 60 (2017) 831–843

5. J.A. Ohlson, “Financial ratios and the probabilistic prediction of bankruptcy,” Journal of accounting research, vol. 18, no. 1, pp. 109–131, 1980

6. D. Martin, “Early warning of bank failure: A logit regression approach,” Journal of Banking & Finance, vol. 1, no. 3, pp. 249–276, 1977

7. Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 15, 757–770.

8. Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman’s and Ohlson’s models. *Review of Accounting Studies, 1* (4), 267–284.

9. R.L Wilson, R. Sharda, “Bankruptcy prediction using neural networks,” Decision support systems, vol. 11, no. 5, pp. 545–557, 1994

10. K.Y. Tam, M.Y. Kiang, “Managerial applications of neural networks: the case of bank failure predictions,” Management science, vol. 38, no. 7, pp. 926–947, 1992

11. E.I. Altman, G. Marco, F. Varetto, “Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience),” Journal of Banking & Finance, vol. 18, no. 3, pp. 505–529,1994

12. F. Ciampi, C. Vallini, N. Gordini, (2009) “Using Artificial Neural Networks Analysis for Small Enterprise Default Prediction Modeling: Statistical Evidence from Italian Firms, Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research (ABER), vol. 1, pp. 126, 2009

13. L. Wei, J. Li, Z. Chen, “Credit risk evaluation using support vector machine with mixture of kernel,” Proceedings of the 7th International Conference on Computational Science. Lecture Notes in Computational Science and Engineering, vol. 4488, pp 431–438, 2007

14. W.K. Härdle, Y.J. Lee, D. Schäfer, “The default risk of firms examined with smooth support vector machines,” Discussion papers, German Institute for Economic Research, no. 757, pp. 1–30, 2007

15. M. Zieba, S. K. Tomczak, J. M. Tomczak Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction Expert Systems with Applications 58 (2016) 93–101

16. Y. Xia, C. Liu, Y. Li, N. Liu A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring Expert Systems with Applications 78 (2017) 225–241

17. Huang, S.-C., Tang, Y.-C., Lee, C.-W., & Chang, M.-J. (2012). Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. Expert Systems with Applications, 39, 3855–3861

18. Sun, J., Jia, M.-Y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. Expert Systems with Applications, 38, 9305–9312.

19. Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research, 241, 236–247.

20. du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. European Journal of Operational Research, 242, 286–303.

21. Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93-101.