**Прогнозирование банкротства банков на основе финансовых и макроэкономических показателей.**

Банковская система занимает центральную позицию в современной экономике, что требует особо строгого регулирования данной сферы деятельности. Снижение доверия к банковской системе приводит к сокращению частных сбережений и неэффективности распределения средств, что негативно отражается на состоянии экономики в целом. Среди существующих теоретических и практических проблем управления коммерческим банком, проблема предсказания кризисной ситуации и банкротства банка является важной бизнес-задачей в финансовой сфере. События мирового финансового кризиса начавшегося в декабре 2007 г., наглядно демонстрируют необходимость использования кризисно-прогнозных моделей.

В настоящее время, когда экономика России и банковская система в т.ч., работают в условиях санкций, что значительно усложняет работу многих отраслей экономики, ставит новые вопросы и требует пересмотра подходов к их решению, использование моделей, предсказывающих потенциальные случаи банкротства, обретает важность в обнаружении проблем, улучшении принятия решений о выделении средств и снижении рисков, а также в повышении эффективности использования ресурсов и предотвращении негативных последствий для экономики в целом.

Актуальность этой бизнес-задачи обусловлена несколькими факторами.

Во-первых, устойчивая и надежная банковская система является ключевым элементом для поддержания стабильности экономики. Прогнозирование банкротства банков помогает выявлять финансовые учреждения, которые могут столкнуться с проблемами, чтобы принять своевременные меры и предотвратить потенциальные последствия для финансовой системы в целом.

Во-вторых, прогнозирование банкротства позволяет банкам и инвесторам принимать решения о выделении кредитов и инвестиций с учетом рисков. Это помогает улучшить эффективность использования ресурсов, а также минимизировать потери, связанные с банкротством и несостоятельностью.

В-третьих, с развитием технологий и возросшей доступностью больших объемов данных, возможности прогнозирования банкротства становятся все более точными. С использованием аналитических методов и алгоритмов машинного обучения мы можем выявить скрытые паттерны и тренды в данных, что позволяет делать более достоверные прогнозы.

Информация о возможных банкротствах может быть полезна также для регуляторных органов и государственных органов, которые могут использовать эти данные для укрепления нормативно-правовой базы и принятия мер по предупреждению и управлению финансовыми кризисами.

Учитывая вышеперечисленное, можно выделить четырех основных стейкхолдеров проекта:

* **Банки:** Информация, полученная из модели, может быть полезной для банков, позволяя им оценить свою финансовую устойчивость, идентифицировать факторы, влияющие на риски банкротства и разработать стратегии по их снижению.
* **Инвесторы:** Инвесторы могут использовать прогнозы модели для оценки рисков вкладывания средств в определенные банки. Это поможет им принимать информированные решения и регулировать свои инвестиционные портфели.
* **Регуляторные органы:** Регуляторные органы, такие как Банк России, могут использовать модель для оценки финансовой стабильности банковской системы, выявления банков, находящихся под угрозой банкротства, и принятия соответствующих мер по предотвращению системных рисков.
* **Аналитики и исследователи:** Прогнозная модель может быть полезной для аналитиков и исследователей, которые занимаются изучением финансовой стабильности и рисков банковской системы. Они могут использовать модель для проведения аналитических исследований, оценки различных сценариев и разработки стратегий с целью минимизации рисков.

В целом, эта модель может быть полезной для многих заинтересованных сторон, которые желают оценить потенциальные риски финансовой нестабильности банков и принять необходимые меры для снижения этих рисков.

Описание данных

В работе использовался набор данных из открытого источника найденный на сайте [kaggle.com](https://www.kaggle.com/datasets/johnds2/bankdefaultsinrussia). Данный набор включает в себя ежемесячные финансовые показатели банков, данные по банковской системе и макроэкономическим показателям для всех российских банков за период с 2010 по 2017 год.

Данные представлены в количестве 45 переменных.

Анализируя типы данных переменных и специфику, можно разделить переменные на количественные и качественные, качественные показатели представлены тремя переменными, а все остальные можно отнести к количественным, вот некоторые из них:

Качественные переменные:

* ***license***: Иными словами уникальный номер банка, это разрешение, выданное Центральным Банком Российской Федерации, позволяющее банку предоставлять услуги по приему вкладов и выдаче кредитов, а также проводить другие операции, связанные с финансовыми операциями и платежами.
* ***msk\_spb***: Московское или Санкт-Петербургское расположение банка.
* ***default***: Показатель дефолта или вероятности дефолта.

Количественные переменные (интервальные):

* ***date***: Дата, к которой относятся финансовые и макроэкономические показатели.

Количественные переменные:

* ***net\_assets***: Чистые активы банка.
* ***ROA***: Рентабельность активов.
* ***INF\_SA***: Уровень инфляции.
* ***NX\_growth***: Рост чистого экспорта.
* и др.

В наборе данных без предварительной обработки аномалий представлена информация по 1006 банкам, из которых у 323 банков была отозвана лицензия. Это означает, что примерно 32% банков в наборе данных обанкротились в заданном периоде. Количество наблюдений - 72439. Распределение количества записей по каждому банку обозначает данные за период с 2010-02-01 по 2017-12-01. Это позволяет проводить анализ финансовых показателей банков за восьмилетний период.

В процессе подготовки данных для построения моделей, были изменены названия переменных и приведены к единому стилю, так как некоторые названия были на кириллице, в то время как большинство применяли латинские символы. А данные в переменной ‘date’ были приведены к типу ‘datetime64’ вместо ‘object’.

При использовании диаграммы разброса, описывающей распределение количества записей по каждому банку, было определено медианное значение - 80, но также можно увидеть неравномерное распределение записей, о чем говорит смещенная к верхней границе медиана и разница длины верхнего и нижнего усов диаграммы. Исходя из чего было принято решение оставить только те банки, количество записей по которым, более 6, т.е. банки, которые сдавали отчетность на протяжении полугода и более. Количество наблюдений сократилось до 70419. Количество банков сдавших отчетность 6 и более раз: 990, их них 766 банков располагаются в г. Москва или в г. Санкт-Петербург, количество дефолтных банков уменьшилось на один и равно 322, их них 220 банков располагаются в г. Москва или в г. Санкт-Петербург. Учитывая объем банков расположенных в данных регионах, принято решение включить переменную ‘msk\_spb’ в признаки обучения модели.

Основным показателем, характеризующим финансовое состояние банка, является показатель “чистые активы” - та часть активов, которая обеспечена собственным капиталом. Однако величина ‘чистые активы’ отличается от размера показателя ‘собственный капитал’ согласно Положению Банка России №395-П. Различие заключается в том, что в расчет собственного капитала входят субординированные кредиты, которые не входят в состав чистых активов. Исходя из этого, дополнительные 12 показателей, характеризующие финансовое состояние банка, были рассчитаны в отношении к ‘чистым активам’. Поэтому значение данной переменной не может быть равным 0. После проверки данных на это условие, объем наблюдений сократился до 68068, общее количество банков стало 982, а количество банков с отозванной лицензией сократилось до 311.

Список дополнительных показателей:

* ***'liquidity\_ratio'***: отношение ликвидных активов к чистым активам; показывает долю активов банка, которые могут быть быстро преобразованы в наличные средства для покрытия текущих или краткосрочных обязательств. Служит индикатором финансовой ликвидности и способности банка удовлетворить свои обязательства в коротком сроке.
* ***'retail\_deposit\_ratio'***: отношение депозитов ФЛ к чистым активам; показывает долю депозитов от физических лиц в общей структуре активов банка. Отражает уровень привлечения розничных клиентов и зависимость банка от розничных депозитов.
* ***'organization\_deposit\_ratio'***: отношение депозитов ЮЛ к чистым активам; показывает долю депозитов от юридических лиц в общей структуре активов банка. Отражает уровень привлечения корпоративных клиентов и зависимость банка от корпоративных депозитов.
* ***'credit\_portf\_ratio'***: отношение кредитного портфеля к чистым активам; показывает долю кредитов, предоставленных банком, в общей структуре его активов. Указывает на важность кредитного портфеля в деятельности банка и его роль в генерации доходов и рисков.
* ***'retail\_credit\_ratio'***: отношение кредитов ФЛ к чистым активам; показывает долю кредитов, выданных физическим лицам, в общей структуре активов банка. Отражает направленность банка на предоставление кредитов населению и его экспозицию к розничным кредитным рискам.
* ***'organization\_credit\_ratio'***: отношение кредитов ЮЛ к чистым активам; показывает долю кредитов, выданных юридическим лицам, в общей структуре активов банка. Отражает направленность банка на предоставление кредитов корпоративным клиентам и его экспозицию к корпоративным кредитным рискам.
* ***'reserv\_credit\_portf\_ratio'***: отношение процента резерва кредитов к общей сумме кредитного портфеля; показывает долю средств, выделенных банком на формирование резервов, для покрытия возможных кредитных потерь. Отражает уровень осторожности и способность банка управлять кредитным риском.
* ***'reserv\_credit\_net\_assets'***: отношение процента резерва кредитов к чистым активам; показывает долю сформированных резервов относительно общих активов банка. Выступает показателем уровня финансовой стабильности и готовности банка к снижению степени риска и потенциальным убыткам.
* ***'stocks\_net\_assets'***: доля акций в структуре активов банка.
* ***‘bond\_net\_assets’***: доля облигаций в структуре активов банка.
* ***'loans\_ratio'***: отношение заемных средств к общим активам банка. Коэффициент кредитования - этот показатель может отражать уровень риска, связанного с заемными средствами.
* ***'foreign\_assets\_ratio'***: доля иностранных активов от общих активов.

К этим показателям, так же были добавлены две переменные, которые уже были рассчитаны в наборе данных, это 'ROA' - рентабельность активов (отношение прибыли к общим активам) и 'ROE' - рентабельность собственного капитала (отношение прибыли к собственному капиталу).

Таблица 1 - Результаты непараметрического теста для независимых выборок с использованием U-критерия Манн-Уитни

| **№** | **Финансовый показатель** | **p-value** | **Нулевая гипотеза** |
| --- | --- | --- | --- |
| 1 | ROA | 0.001 | Отклоняем |
| 2 | ROE | 0.006 | Отклоняем |
| 3 | liquidity\_ratio | 0.002 | Отклоняем |
| 4 | retail\_deposit\_ratio | 0.030 | Отклоняем |
| 5 | organization\_deposit\_ratio | 0.456 | Не отклоняем |
| 6 | credit\_portf\_ratio | 0.105 | Не отклоняем |
| 7 | retail\_credit\_ratio | 0.000 | Отклоняем |
| 8 | organization\_credit\_ratio | 0.028 | Отклоняем |
| 9 | reserv\_credit\_portf\_ratio | 0.000 | Отклоняем |
| 10 | reserv\_credit\_net\_assets | 0.000 | Отклоняем |
| 11 | stocks\_net\_assets | 0.000 | Отклоняем |
| 12 | bond\_net\_assets | 0.483 | Не отклоняем |
| 13 | loans\_ratio | 0.000 | Отклоняем |
| 14 | foreign\_assets\_ratio | 0.000 | Отклоняем |

По результатам непараметрического теста для независимых выборок с использованием U-критерия Манн-Уитни нулевая гипотеза о равенстве средних значений всех показателей, характеризующих финансовое состояние банка, для действующих банков и банков, у которых была отозвана лицензия, была отвергнута на уровне значимости p-value > 0.05, по 11 переменным, за исключением: ‘organization\_deposit\_ratio’, ‘credit\_portf\_ratio’, ‘bond\_net\_assets’. Остальные переменные считаются показателями, характеризующими различия между функционирующими банками и банками, которые обанкротились.

Основываясь на эти результаты можно сделать вывод, что большинство показателей имеют статистически значимые различия между действующими банками и банками с отозванной лицензией. Учитывая результаты теста, можно отметить эти различия и использовать значимые показатели для построения модели.

В качестве макроэкономических признаков и данных по банковской системе для обучения модели также были выбраны следующие переменные:

* 'NX growth': Рост чистого экспорта
* 'INF\_SA': Уровень инфляции
* 'usd\_rub\_std\_diff': Разница в стандартном отклонении курса USD/RUB.
* 'usd\_rub\_return': Доходность USD/RUB.
* ‘micex\_std’: Стандартное отклонение индекса MICEX.
* ‘miacr\_std’: Стандартное отклонение индекса MIACR.
* ‘miacr\_amount’: Абсолютное значение индекса MIACR.
* ‘net\_foreign\_assets\_diff’: Разница в чистых иностранных активах.
* ‘net\_gov\_debt\_diff’: Разница в государственном долге.
* ‘other\_fin\_debt\_diff’: Разница в долгах от других финансовых институтов.
* ‘retail\_debt\_SA\_DETREND\_diff’: Разница в розничном долге.
* ‘stocks\_capital\_diff’: Разница в стоимости акций и капитала.
* ‘i\_retail\_spread\_diff’: Разница между доходностью розничного кредита и ставкой по розничным депозитам.

Особенностью данных, на которых строились модели, является их несбалансированность. Это означает, что число успешно функционирующих банков в течении периода наблюдений существенно превышает число банков, у которых была отозвана лицензия. И поскольку масштаб макроэкономических переменных отличается от масштаба финансовых показателей было принято решение использовать нелинейные модели RandomForest и XGBoost, которые в свою очередь не чувствительны к масштабированию и одинаково хорошо обрабатывают как непрерывные, так и дискретные признаки.

RandomForest принимает случайное подмножество функций и объектов для каждого дерева, а затем усредняет их для принятия решения. Это позволяет модели быть менее склонной к переобучению и предоставлять более устойчивые прогнозы на новых данных.

XGBoost использует регуляризацию и добавляет комбинированные базовые алгоритмы, что делает его менее склонным к переобучению. Кроме того, он имеет возможность контролировать сложность модели в процессе обучения.

Для обучения данных моделей предлагается два варианта группировки данных.

**Группа I:** в первом варианте используется группировка по банку, т.е по его номеру лицензии, для остальных показателей рассчитывается среднее значение за рассматриваемый период, а для обозначения региона и дефолта, которые являются бинарными переменными указывается его максимальное значение. Таким образом набор данных содержит количество наблюдений равное количеству банков.

**Группа II:** во втором варианте данные сгруппированы по банку и по году, и имеют псевдопанельную структуру, т.е. каждый банк в каждый год считается отдельным наблюдением. В результате группировки, набор данных состоит из 6283 наблюдений.

Целевая переменная в обоих случаях группировки одинаковая и принимает значение “1”, если у банка была отозвана лицензия, и “0”, если банк продолжает функционировать.

Тестовая выборка в данном случае определяется параметром `test\_size’ и равна 0.2 для всех типов построенных моделей. В этом случае, 20% данных будет выделено для тестового набора, в то время как оставшиеся 80% данных будут использованы для обучения модели. Такое разделение позволяет проверить эффективность модели на новых данных и оценить её обобщающую способность. Общий размер выборки 197 наблюдений.

Для наглядности сравнения эффективности моделей, поместим полученные значения в таблицы, по две таблицы на каждый вариант группировки данных.

Таблица 2 - Результаты моделей на обучающей выборке. Группа I

|  | Модель 1. RandomForest | Модель 2. XGBoost |
| --- | --- | --- |
| Время работы, мс | 618.3769 | 297.6963 |
| Accuracy | 0.8071 | 0.7969 |
| AUC-ROC | 0.74103 | 0.7580 |

Таблица 3 - Отчет о классификации. Группа I

| Классы | Модель 1. RandomForest | | Модель 2. XGBoost | |
| --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 1 |
| Precision | 0.83 | 0.72 | 0.86 | 0.66 |
| Recall | 0.91 | 0.58 | 0.86 | 0.66 |
| F1-score | 0.87 | 0.64 | 0.86 | 0.66 |
| Support (кол-во наблюдений) | 138 | 59 | 138 | 59 |

Accuracy это показатель точности модели, который показывает долю правильных предсказаний. В данном случае для обеих моделей точность составляет около 80%, что означает, что модели правильно классифицировали примерно 80% наблюдений.

Значения AUC-ROC тоже практически одинаковые, однако все же модель XGBoost с показателем 0.7580 имеет преимущество перед RandomForest с показателем 0.7410 на 0,017, и это указывает на то, что модель имеет приемлемую способность различать между классами и достаточно хорошо предсказывает правильные и ошибочные классификации. А также модель XGBoost имеет более высокие значения F1-score, что указывает на хорошую точность и полноту модели в предсказании второго класса.

Таким образом, модель XGBoost показала лучшие результаты в прогнозировании вероятности банкротства банка, хотя стоит учитывать, что значения модели RandomForest достаточно близки, поэтому в целом их производительность можно оценить как хорошую.

Посмотрим второй вариант группировки данных, и сравним как использование псевдопанельной структуры данных повлияет на качество и производительность этих моделей. В данном случае были построены по две модели каждого типа с учетом корректировок. В случае с моделью RandomForest дополнительно применялось использование сокращенных компонент для создания нового набора признаков, в результате чего вторая модель строилась на меньшем количестве признаков, использование PCA для перехода от 25 переменных к 20 компонентам позволяет объяснить 96% дисперсии данных, а другие 5 компонент объясняют менее 4% дисперсии, а это значит, что от них мы можем отказаться. Тестовая выборка для модели увеличена до 0.3.

А в случае с моделью XGBoost дополнительно применялось создание сетки гиперпараметров для модели XGBoost с использованием метода перекрестной проверки GridSearchCV. Для улучшения производительности модели.

Общий размер выборки 1885 наблюдений.

Таблица 4 - Результаты моделей на обучающей выборке. Группа II

|  | Модель 1. RandomForest | Модель 1.2 RandomForest | Модель 2.  XGBoost | Модель 2.2  XGBoost |
| --- | --- | --- | --- | --- |
| Время работы, мс | 2326.5745 | 3285.4714 | 3800.8186 | 4838.7212 |
| Accuracy | 0.9538 | 0.9533 | 0.9570 | 0.9570 |
| AUC-ROC | 0.5433 | 0.5216 | 0.6391 | 0.6305 |

Таблица 5 - Отчет о классификации. Группа II

| Классы | Модель 1. RandomForest | | Модель 1.2 RandomForest | | Модель 2. XGBoost | | Модель 2.2  XGBoost | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | **1** | **0** | **1** | **0** | **1** | **0** | **1** |
| Precision | 0.96 | 0.57 | 0.95 | 0.50 | 0.97 | 0.59 | 0.97 | 0.57 |
| Recall | 1 | 0.09 | 1 | 0.02 | 0.99 | 0.29 | 0.99 | 0.27 |
| F1-score | 0.98 | 0.16 | 0.98 | 0.04 | 0.98 | 0.39 | 0.98 | 0.37 |
| Support | 1796 | 89 | 1796 | 89 | 1198 | 59 | 1198 | 59 |

В случае с группировкой данных, где каждый банк в каждый год считается отдельным наблюдением, аccuracy составляет около 0.95, а значит точность выше, чем в первой группе. Но если посмотреть на значения F1-score и Recall для второго класса, до будет понятно, что такой хороший процент точности предсказания получается, за счет первого класса, в то время как по второму классу, показатели очень низкие. Так и значения AUC-ROC значительно ниже для моделей RandomForest 0.5433 и 0.5216 соответственно, и попадают в диапазон 0,5 < AUC < 0,6, т.е. модель работает удовлетворительно, хотя стоит отметить, что использование метода главных компонент, не улучшило производительность модели.

Так для моделей XGBoost AUC-ROC значения выше 0.6391 и 0.6305 соответственно, и попадают в диапазон 0,6 ≤ AUC <0,8, т.е. модель работает хорошо, при чем использование создания сетки гиперпараметров не повысило качество модели, а наоборот немного его снизило.

Заключение

Проведенный мной анализ дает основание считать, что все модели показали свою работоспособность при прогнозировании вероятности отзыва лицензии у банка. Однако, модели первой группы были значительно эффективнее. И модель XGBoost является самой удачной для выполнения этой задачи, т.к. она дала самое высокое значение ROC-AUC - метрики, при хорошей точности в 0.8, а также имеет самое быстрое время работы кода в миллисекундах: 297.6963520050049, что более, чем в два раза быстрее обучения модели RandomForest на тех же данных. И в отличии от RandomForest, модель XGBoost не имеет проблем с переобучением, т.к. результаты модели RandomForest могут меняться в пределах небольшого интервала +-5%, что потребует дополнительного решения данной проблемы.

И все же, дополнительный анализ требуется для оценки того, какое число показателей, по которым имеются значимые различия в группах, необходимо и достаточно, чтобы иметь убедительное доказательство того, что ошибки классификации модели объясняются неэкономическими причинами.

В работе предложена концепция одновременного учета финансовых показателей банков, показателей банковской системы, а также макроэкономических показателей, наблюдения по которым фиксировались на ежемесячной основе, в качестве признаков для обучения модели, которая прогнозирует вероятность банкротства банка. Отличительной характеристикой работы, является также период наблюдений, т.к. эти годы были переломными для российской банковской системы, ведь около половины всех банков обанкротились, что с точки зрения обучения моделей, является хорошим подспорьем в улучшении классификации банков-банкротов.