***Модель***

Для обобщения исторических данных и оценки состояния участников банковской системы мы использовали следующую модель. Мы предполагаем, что есть функциональная зависимость между отдельными показателями деятельности кредитных организаций, используемых для расчета обязательных нормативов (раздел 2 формы 0409135 «Информация об обязательных нормативах» ЦБ РФ [ссылка на форму]), и вероятностью отзыва лицензии или ликвидации кредитной организации. Математически эта зависимость может быть записана в следующем виде:

где L – показывает, находится ли кредитная организация в «зоне риска» по отзыву лицензии или ликвидации (значение 1), или не находится (значение 0), f – некоторая функция, ai – значения показателей деятельности организации. Так как конкретный вид такой функции неизвестен и не может быть получен известными методами, мы используем машинное обучение для аппроксимации этой функции. Фактически, мы сводим задачу к бинарной классификации – классической задаче анализа данных, где метки – это наличие или отсутствие риска отзыва лицензии у кредитной организации, а признаки – показатели деятельности из описанной выше формы ЦБ.

Для данных за 2019 год [1 ссылка на данные ЦБ] мы выделили ряд самых важных показателей, которые приведены в таблице 1.

|  |  |
| --- | --- |
| Код показателя в соответствии с Указанием Банка  России от 08.10.2018 № 4927-У | Расшифровка кода показателя |
| Ар1.0 |  |
| Ар2.0 |  |
| Ар4.0 |  |
| Ариск0 |  |
| Кф |  |
| Лат |  |
| Овт |  |
| ПК0 |  |
| ВР |  |

Таблица 1. Коды показателей деятельности кредитных организаций, выбранных в качестве признаков для модели машинного обучения

В качестве модели машинного обучения была выбрана популярная реализация алгоритма градиентного бустинга решающих деревьев XGBoost [2]. Эта реализация обладает рядом достоинств: продвинутой регуляризацией, инвариантностью относительно масштаба признаков и возможность параллельного обучения модели.

***Обучение модели***

Для обучения модели мы использовали исторические данные за 2019 год для 378 кредитных организаций, из которых для 30 известно, что в течении 2019 года была отозвана лицензия или организация была ликвидирована (из выборки также были исключены 50 крупнейших по объему средства кредитные организации (объяснить зачем)). Значения брались за февраль, апрель, июль, октябрь указанного года, общий размер датасета составил 1735 записей. Выбранные нами гиперпараметры модели, отличающиеся от значений по умолчанию, приведены в таблице 2.

|  |  |
| --- | --- |
| Гиперпараметр модели | Значение |
| Количество решающих деревьев | 1000 |
| Глубина решающего дерева | 5 |

Таблица 2. Значение гиперпараметров классификатора XGBoost, отличающиеся от значений по умолчанию

Так как кредитных организаций с отозванной лицензией (или ликвидированных) в исходной выборке намного меньше, что может сказаться на качестве обучения модели, мы воспользовались техникой случайного отбора элементов мажоритарного класса (англ. undersampling) [3]. В итоге в обучающей выборке отношение элементов миноритарного класса к мажоритарному составило один к трём.

Для валидации модели использовалась отложенная выборка размером в 20% от исходной. Для оценки качества классификации мы рассчитывали общую точность, то есть долю записей, для которых классификатор верно указал класс, а также для каждого класса вычислялись точность – отношение верного срабатывания к общему числу срабатываний, полнота – отношение верных срабатываний к общему числу истинных объектов и F1-мера – гармоническое средние между точностью и полнотой [4]. Также была построена кривая ошибок, которая показывает соотношение специфичности и чувствительности модели при различных значениях порога отсечения [5]. Модель считается хорошей, если площадь под кривой больше 0,7 [6]. Результаты обучения модели для 2019 года представлены в таблице 3 и на рисунке 1.

|  |  |  |
| --- | --- | --- |
| Параметр оценки качества  классификации | Значение | Количество записей в отложенной выборке |
| Общая точность классификации | 91% | 69 |
| Точность для класса «Возможен отзыв лицензии или ликвидация кредитной организации» | 82% | 13 |
| Полнота для класса «Возможен отзыв лицензии или ликвидация кредитной организации» | 69% | 13 |
| F1-мера для класса «Возможен отзыв лицензии или ликвидация кредитной организации» | 75% | 13 |
| Точность для класса «Маловероятен отзыв лицензии или ликвидация кредитной организации» | 93% | 56 |
| Полнота для класса «Маловероятен отзыв лицензии или ликвидация кредитной организации» | 96% | 56 |
| F1-мера для класса «Маловероятен отзыв лицензии или ликвидация кредитной организации» | 95% | 56 |

Таблица 3. Значения параметров оценки качества классификации обученной модели на отложенной выборке кредитных организаций за 2019 год

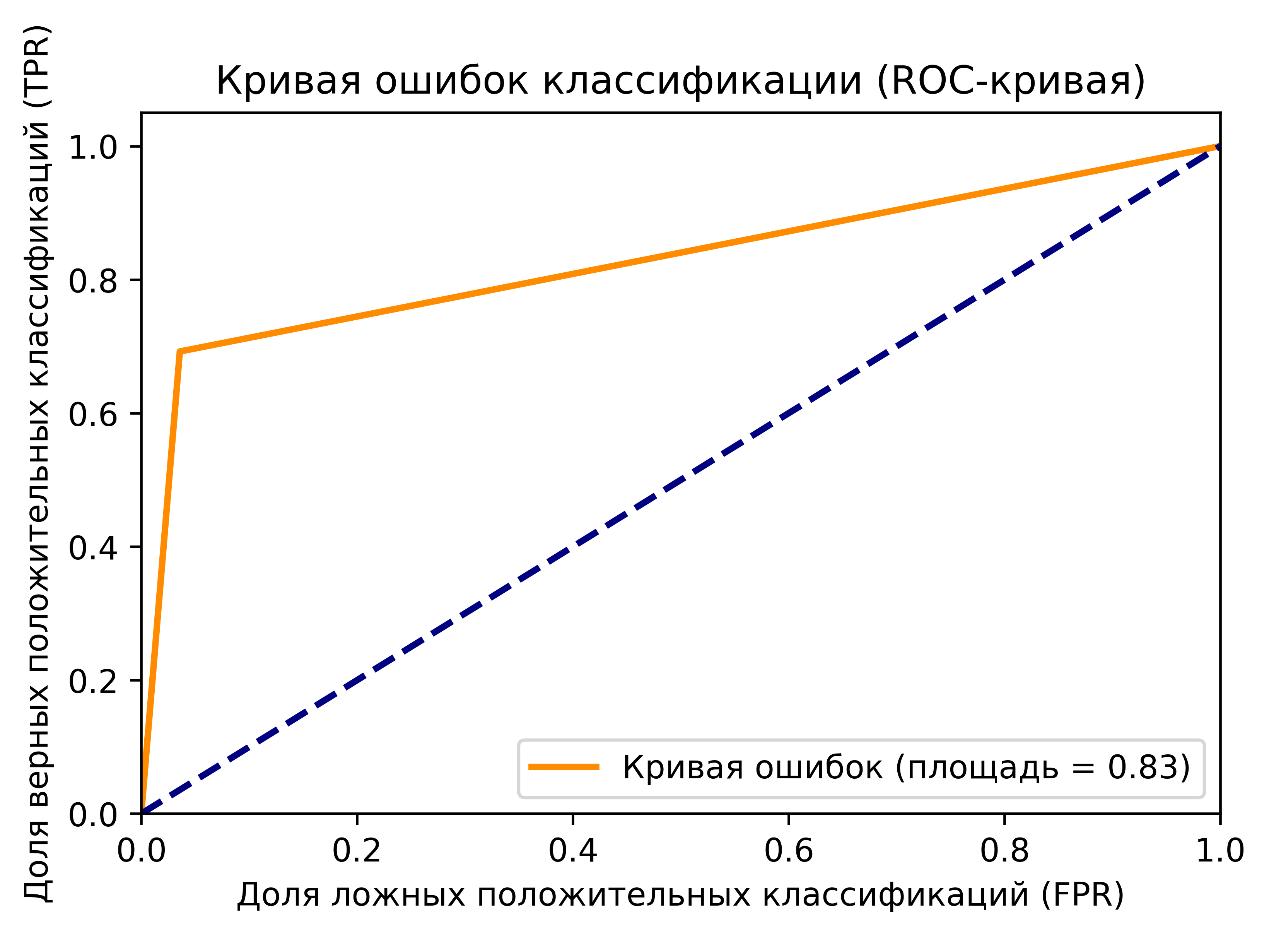


Рисунок 1. Кривая ошибок классификации по отложенной выборке для данных о кредитных организациях за 2019 год

Как видно из таблицы 3 и рисунка 1, построенная модель обладает хорошим общим качеством классификации (общая точность – 91% и площадь под кривой ошибок – 0.83), и что самое важное – высокой точностью для класса «Возможен отзыв лицензии или ликвидация кредитной организации» – 82%. Все это позволяет применять полученную модель для анализа вероятности попадания кредитной организации под внимание регулятора и возможности отзыва лицензии.

***Применение модели***

Оценив качество построенной модели за 2019 год, мы применили её для анализа состояния участников банковской системы на основе данных, уже опубликованных в 2020 году. Также как и для 2019 года, мы исключили из рассмотрения 50 крупнейших кредитных организаций. В результате мы получили следующее: из 325 рассмотренных кредитных организаций, 43 находятся под угрозой отзыва лицензии, что составляет примерно 11% от общего количества участников банковской системы с учетом 50 крупнейших.

(Вот здесь может быть кусок с информацией, о который ты мне писала в личку по регн, кто там под слиянием, кто аннулирован, что-то вроде:

Так, в этот список попали: коммерческий банка «Центрально-Азиатский» (рег. № 3037), у которого в июле 2020 года была аннулирована лицензия,

и так далее несколько примеров можно привести, чтобы показать, что мы не ерунду посчитали )

***Вывод***

Таким образом, в нашей работе мы показали возможность построения модели машинного обучения для анализа состояния конкретных кредитных организаций по показателям их деятельности и оценки вероятности отзыва лицензии. Модель была обучена на актуальных данных за 2019 год, было получено высокое качество классификации. В работе были также получены оценки кредитных организаций по показателям за 2020 год и указаны несколько примеров точных оценок вероятности отзыва лицензии. Модели, построенные таким образом могут быть использованы в деятельности ЦБ РФ для оценок риска банковского сектора.

Литература

1. Ссылка на данные ЦБ

2. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – С. 785-794.

3. Zheng A., Casari A. Feature engineering for machine learning: principles and techniques for data scientists. – " O'Reilly Media, Inc.", 2018.

4. Grus J. Data science from scratch: first principles with python. – O'Reilly Media, 2019.

5. Narkhede S. Understanding AUC-ROC Curve //Towards Data Science. – 2018. – Т. 26.

6. Логистическая регрессия и ROC-анализ — математический аппарат // Материалы онлайн-конференции по продвинутой аналитике Loginom days 2020 [Электронный ресурс]. Режим доступа: https://loginom.ru/blog/logistic-regression-roc-auc, свободный.