УДК 333.33

ПОСТРОЕНИЕ СКОРИНГОВОЙ МОДЕЛИ ДЛЯ ОПРЕДЕЛЕНИЯ КРЕДИТОСПОСОБНОСТИ ПОТЕНЦИАЛЬНОГО ЗАЁМЩИКА

Расторгуев Л.М.

студент,

Чувашский государственный университет имени И. Н. Ульянова,

Чебоксары, Россия

Микишанина Е.А.

Доцент кафедры актуарной и финансовой математики,

Чувашский государственный университет имени И. Н. Ульянова,

Чебоксары, Россия

Аннотация

Цель данной работы – построение инструмента классификации (скоринговой модели), который будет указывать является ли заёмщик кредитоспособным, т.е. надёжным. При построении лучшей модели классификации был проведён сравнительный анализ разных методов для определения наиболее подходящего для данной задачи. В этих целях были использованы классические методы машинного обучения (ML): логистическая регрессия, дерево решений, k-ближайших соседей (kNN). А также более современные ансамблевые методы бэггинга и бустинга, куда входят случайный лес вместе с CatBoost и XGBoost соответственно. В итоге более усовершенствованные методы, основанные на деревьях решений, показали самые лучшие результаты. По выбранным критериям

они показали самые высокие результаты качества. По этой причине данные методы бустинга, показавшие лучшие результаты, можно использовать в банках для снижения риска невозврата кредита заёмщиком.

Ключевые слова: скоринг, благонадёжность заёмщика, Machine Learning, AUC ROC, задача классификации, CatBoost.

CONSTRUCTION OF A SCORING MODEL TO DETERMINE THE CREDITWORTHINESS OF A POTENTIAL BORROWER

Rastorguev L.M.

Student,

Chuvash State University,

Cheboksary, Russia

Mikishanina E.A.

Associate Professor of the Department of Actuarial and Financial Mathematics,

Chuvash State University,

Cheboksary, Russia

Abstract

The purpose of this work is to build a classification tool (scoring model), which will indicate whether the borrower is creditworthy, i.e. reliable. In order to build the best classification model a comparative analysis of different methods was conducted to determine the most appropriate one for this task. For this purpose, classical machine learning (ML) methods were used: logistic regression, decision tree, k nearest neighbors (kNN). As well as more advanced ensemble bagging

and boosting methods, which include random forest along with CatBoost and XGBoost, respectively. As a result, the more advanced decision-tree based methods showed the best results. For the selected criteria, they showed the highest quality results. For this reason, these best performing boosting methods can be used in banks to reduce the risk of a borrower defaulting on a loan.

Keywords: scoring, borrower reliability, Machine Learning, AUC ROC, classification task, CatBoost.

Важным финансовым инструментом являются кредиты. В частности, ими нередко пользуются и физические лица. Займы могут браться как для личных целей, так и являются хорошим начальным капиталом при достижении целей. При этом существуют явные риски для обоих сторон – для банков и заемщиков. Один из главных рисков – невозврат кредита (дефолта) [1]. Очевидным решением данной проблемы является предсказывание таких случаев. Такие задачи относятся к скорингу заявки (application-scoring) [3, 12]. Этот вид скоринга позволяет получить оценку кредитоспособности потенциального заёмщика, тем самым позволив решить кредитору, предоставлять кредит или нет. Рассматриваемая задача относится к классу задач бинарной классификации.

Существует множество методов и моделей проведения оценки заёмщков. Со временем эти методы устаревают, а потому совершенствуются. В последнее время стало популярным применение методов машинного обучения (machine learning, ML), а также нейронных сетей (neural networks). Огромные объемы информации, большой поток данных, скорость принятия решений заставляют перейти на новые модели. Машинное обучение хорошо справляется с такого рода требованиями.

В 21 веке кредиты пользуются большой популярностью. Это лишь укрепляет важность и своевременность решения вопросов из сферы кредитования. Одновременно с этим актуальность проблемы невозврата кредитов также подтверждает статистика ЦБ РФ по задолженностям по федеральным округам. На графике (рис. 1) можно наблюдать положительную тенденцию роста по всем округам.

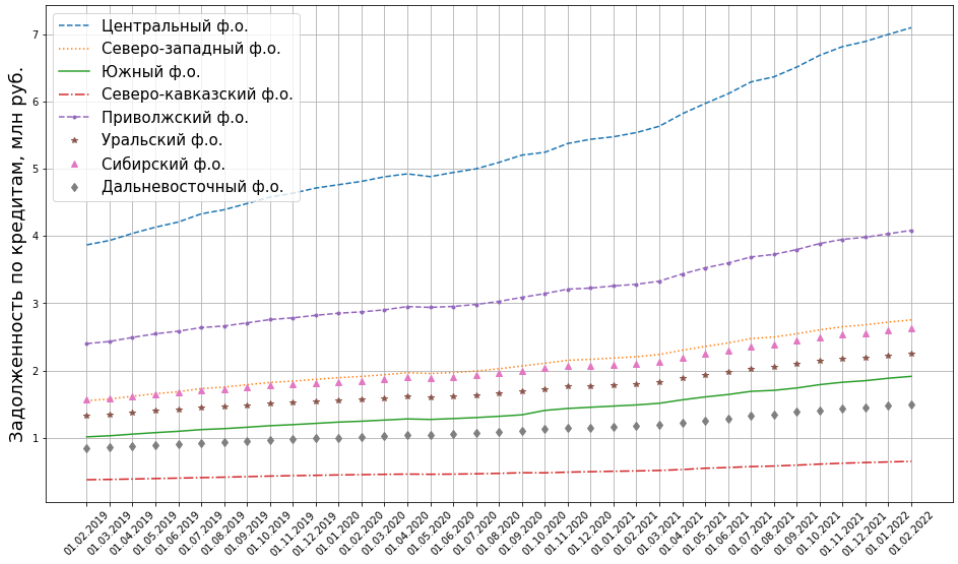

Рис. 1. Задолженность по кредитам за 2019 – 2022 годы по данным ЦБ РФ (составлено автором)

А если смотреть в целом по стране, то становится очевидно, что задолженность только растет. Эти задолженности в будущем могут как погаситься, так и обратиться в дефолт. Данную пессимистичную картину может изменить отказ в выдаче кредита или изменение условий кредитования для неблагонадежных заемщиков [4].

Целью работы является построение скоринговой модели на основе некоторых методов машинного обучения для решения задачи классификации заемщиков.

Для примера построения модели классификации был взят размеченный набор данных (Data set, датасет) из открытого источника

Kaggle. Набор имеет достаточно большой объем и состоит из 74 000 наблюдений. Также в нем 18 признаков – анкетных данных заемщика, и итоговый результат по заявке на кредит, который отображается в виде категориального столбца – был дефолт или нет. Датасет состоит из категориальных, числовых и текстовых видов данных. Как и в большинстве задач скоринга заявки датасет является несбалансированным: 87% и 13% составляют от общей численности составляют мажоритарный и положительный миноритарный классы соответственно. Положительный класс соответствует неблагонадёжным заёмщикам.

Данные являются «сырыми» и нуждаются в предобработке (Data Preparation [5, 10]). Были удалены выбросы, заполнены пропуски, а также из-за ограничений некоторых методов на типы признаков перекодированы категориальные признаки в числовые. Для этого использовались LabelEncoder и OneHotEncoder [5, 10].

Из-за очевидных различий в величинах единиц измерения признаков были применены z-масштабирование и нормализация по методу минимакс.

Как и в любой другой задаче классификации основой для всех критериев оценок (метрик) качества построенных моделей использовалась матрица ошибок (confusion matrix) [5, 10, 13].

После предобработки данных следует определиться с тем, какие модели будут строиться и по каким метрикам следует оценивать их оптимальность и проводить обучение. В роли основных были взяты метрики устойчивые к дисбалансу классов – полнота (Recall) и площадь под кривой ошибок (AUC ROC) [13]. Дополнительно отслеживались метрики точность (precision) и F1-мера [13]. Выбор Recall обусловлен тем, что цель классификации – правильная классификация всех случаев дефолта. Поэтому важно минимизировать ошибку 2-го рода, которую и характеризует Recall.

Для классификации были выбраны такие уже классические методы как логистическая регрессия (Logistic Regression), Дерево решений (Decision Tree), метод K-ближайших соседей (K-nearest neighbor, kNN) [8]. А также в добавок к ним были применены более современные ансамблевые методы, основанные на комбинировании деревьев решений [8, 2]. Модель из беггинга (Bagging) – случайный лес (Random Forest) [8, 2]. Модели из бустинга (Boosting) – CatBoost [11] и XGBoost [7]. Последние 2 модели самые молодые и имеют большой потенциал.

Для обучения и анализа качества модели датасет необходимо разделить на обучающую (train) и тестовую (test). Таким образом, 80% данных идет на обучение модели, а 20% прогоняются по обученной модели и сравнивают результаты с истинными. Однако из-за дисбаланса классов невозможно их хорошо разделить. Поэтому были синтезированы новые наблюдения для малого класса по принципу соседства, т.е. использован инструмент SMOTE (Способ передискретизации синтезированных меньшинств) [6, 9].

Весь процесс обработки и анализа данных, а также построение моделей был проведён на языке программирования Python. В работе использовались следующие библиотеки: numpy, pandas, matplotlib, seaborn, datetime, sklearn, imblearn, catboost, xgboost.

После предобработки данных на тепловой карте (рис. 2), показывающей корреляцию, можно обнаружить всего лишь 3 значения корреляции превышающих по модулю отметку 0,5. Примечательно, что ни один из признаков не имеет сильной корреляционной связи с итоговым признаком default. Наибольшую корреляцию с этим признаком имеют признаки score_bki и sna с результатами 0,18 и 0,13 соответственно. Это подтверждает предположение о слабой связи между анкетными данными и позволяет сделать вывод, что нет единственного признака, от которого в большей степени, чем от остальных зависело бы решение о выдаче кредита.

Т.е. решение принимается, основываясь на показателях совокупности признаков.

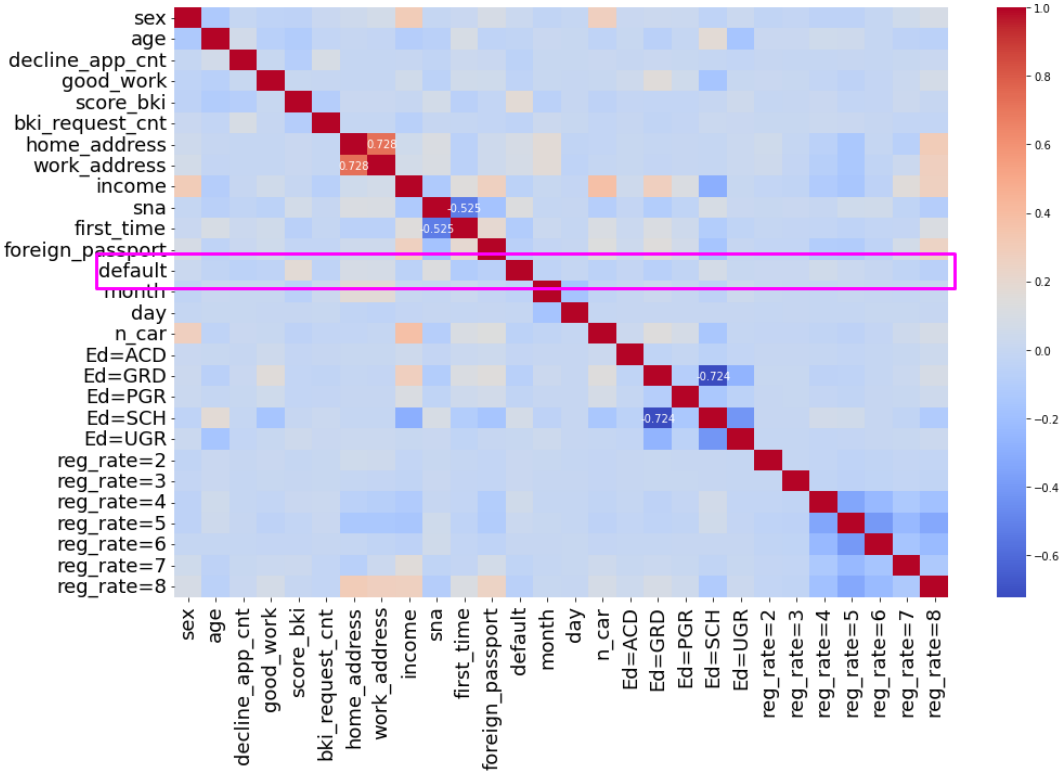

Рис. 2. Тепловая карта, отражающая корреляцию (составлено автором)

Чтобы иметь представление о плотности разброса наблюдений, обратимся к графику (рис. 3) по двум важным входным числовым признакам - скоринговый балл по данным из БКИ (score_bki) и доход (income). Из графика видно, что данные находятся в плотном «облаке». Исходя из этого графика и информации о слабой корреляционной связи можно сделать предположение, что модели, которые разделяют классы используя кривые и плоскости, например, линейная регрессия или SVM, будут существенно уступать в предсказательной силе ансамблевым моделям.

После оптимизации параметров выбранные модели были обучены по метрике AUC ROC. Были получены результаты, указанные в таблице 1.

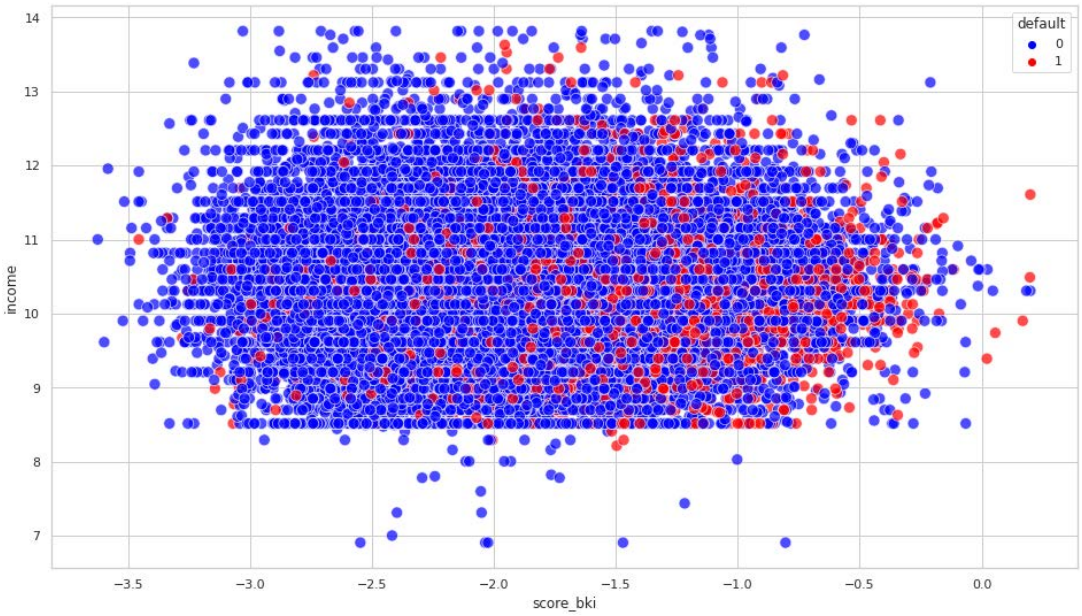

Рис. 3. График по признакам income и score_bki (составлено автором)

Таблица 1 – Построенные модели и их результаты качества по отслеживаемым метрикам (составлено автором)

		Метрики			
		Recall	precison	F1	AUC ROC
Модели	LR	0,568	0,218	0,315	0,694
	kNN	0,417	0,179	0,251	0,602
	DT	0,652	0,179	0,280	0,622
	RF	0,335	0,249	0,285	0,679
	CatBoost	0,633	0,243	0,351	0,735
	XGBoost	0,614	0,247	0,352	0,733

В идеале значения всех четырех отслеживаемых метрик должны стремиться к единице, что будет указывать на абсолютную предсказательную силу, надежность модели. Стоит также отметить, что значение метрики recall в бинарной классификации совпадает с долей правильно классифицированных значений положительного класса.

Проанализировав таблицу, можно сделать вывод, что практически все модели показали хороший результат по выбранной для оптимизации

метрике AUC ROC. Однако если мы обратимся ко 2 основной метрике recall, то можем выделить как хорошие модели LR и DT, а также выделить модели, показавшие самые высокие результаты качества классификации CatBoost и XGBoost.

Отсюда мы подтвердили наше предположение о качестве ансамблевых моделей. Модели бустинга показали лучшие результаты правильно классифицировав большую часть положительного класса, при этом показали самые высокие значения AUC ROC.

Если анализировать дополнительные метрики можно увидеть, что ни одна из них не перешагнула отметку 0,5. Это достаточно просто объясняется тем, что F1-мера зависит от recall и precision. А последняя метрика отвечает за ошибку 1-го рода. Опираясь на здравый смысл, мы делаем акцент не на неё, а на ошибку 2-го рода, известное также как «упущенное событие», т.к. важнее предупредить случай выдачи кредита неблагонадёжному заёмщику, чем не выдача кредита благонадёжному заёмщику. Всё из-за того, что некредитоспособный заёмщик несёт за собой большие риски и убытки. Поэтому при анализе мы рассматриваем основные метрики, а дополнительные носят вспомогательную функцию.

Дополнительно следует остановиться на модели Catboost, которая показала в среднем наилучшие результаты. Его матрица ошибок (рис. 4) и ROC-кривая (рис. 5) выглядят следующим образом:

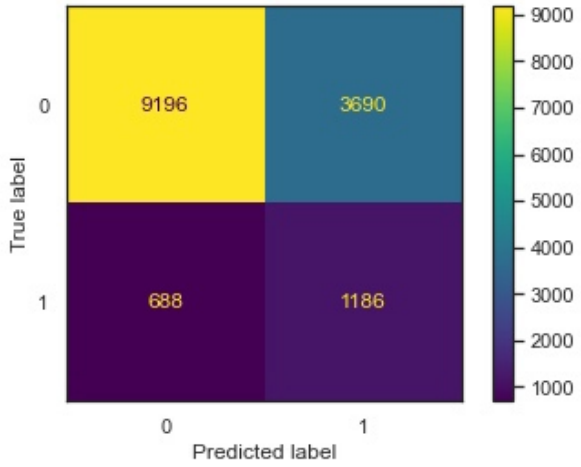

Рис. 4. Матрица ошибок для модели CatBoost (составлено автором)

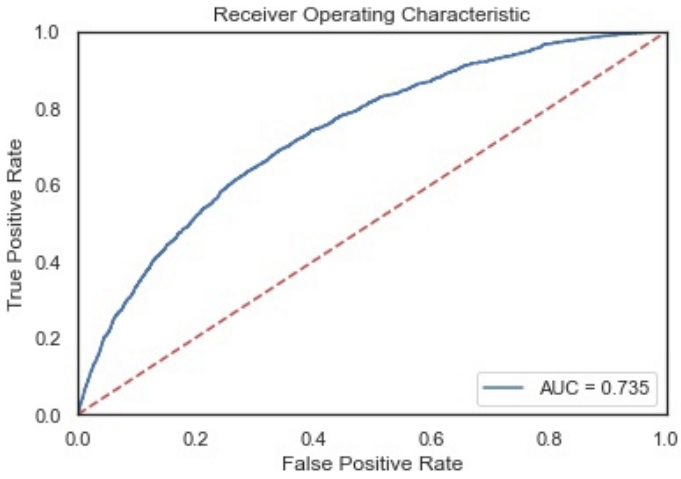

Рис. 5. ROC-кривая для модели Catboost (составлено автором)

Как можно увидеть, модель действительно определила большинство значений положительного класса. ROC-кривая также проходит выше диагональной линии, которая соответствует случайной классификации. Модель можно определить, как хорошо классифицирующую. Можно также предположить, что качество модели было бы ещё выше, если бы входной датасет обладал бы большим числом признаков.

Ниже представлена визуализация обучения модели по метрике recall (рис. 6). На графике видно, что значение метрики recall на обучающем наборе постоянно возрастает, а на тестовом после некоторого момента

начинает убывать. Мы можем наблюдать количество деревьев, после которого происходит переобучение модели. Момент переобучения обозначен точкой. В этой точке метрика достигает своего оптимального значения. Здесь это значение recall, равное 0,633. После этого дерева остальная часть обрезается, т.к. она начинает подстраивать модель под тренировочные данные, тем самым модель перестает быть универсальной.

Аналогичный график при оптимизации по AUC ROC (рис. 7). Если для метрики recall хватило 272 деревьев, то здесь остановка происходит на 415 дереве. Хоть после 200 дерева рост качества не очень сильный, но всё же имеется незначительный рост. Через примерно 200 деревьев метрика принимает своё максимальное значение.

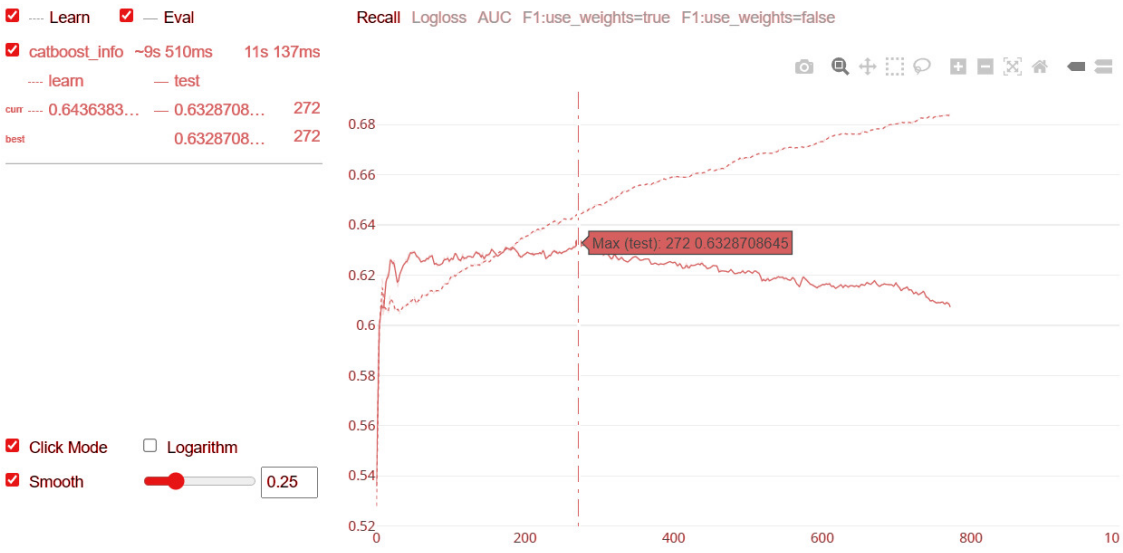

Рис. 6. График обучения модели CatBoost по метрике recall (составлено автором)

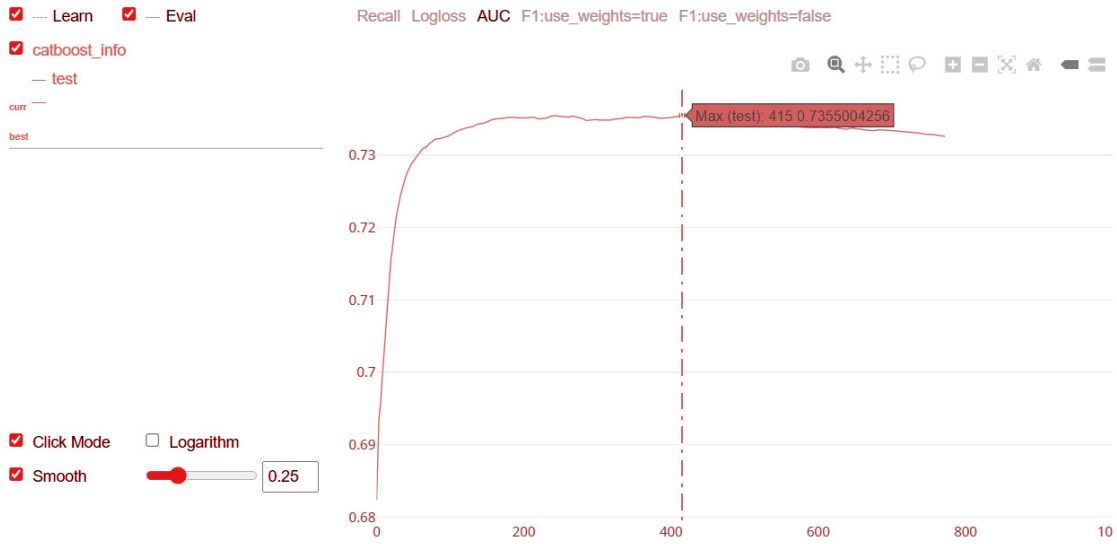

Рис. 7. График обучения модели CatBoost по метрике AUC ROC (составлено автором)

Таким образом, на практическом примере была продемонстрирована эффективность использования методов машинного обучения в банковской сфере, а именно в задаче оценки благонадежности заемщика. Использование таких методов помогает избежать убытков и дефолтов, которые возникают из-за некредитоспособности клиентов. Как правило, построение моделей происходит поэтапно, поэтому дополнительно были рассмотрены основные моменты подготовки и балансировки данных, а также выбора метрик и обучения по ним моделей. Рассмотрели некоторые популярные методы, которые уже используются или могут в будущем использоваться в сфере кредитования. Кроме того, было выяснено, что ансамблевые модели бустинга обладают высоким потенциалом. При плотной концетрации данных эти модели могут хорошо их классифицировать. Как итог, из построенных по открытому набору данных скоринговых моделей классификации потенциальных заёмщиков лучшей оказалась модель на основе CatBoost с итоговыми метриками recall и AUC ROC равными 0.633 и 0.735 соответственно.

Библиографический список:

1. Гобарева Я. Л., Городецкая О. Ю., Еременко И. А. Современные инновационные технологии в банковской сфере // Банковские услуги. 2018. № 6. С. 24-32.

2. Канищев И. С. Оценка точности работы ансамблевых алгоритмов для кредитного скоринга // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2021. № 6. С. 102-108.

3. Клейнер Г. Б., Коробов Д. С. История современного кредитного скоринга // Проблемы региональной экономики. 2012. № 17. С. 49-62.

4. Мадера А. Г. Прогнозирование кредитной благонадежности заемщика // Финансы и кредит. 2013. №12 (540).

5. Brownlee J. Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Machine Learning Mastery, 2020.

6. Fernández A. [et al.] SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary. // AI Access Foundation. 2018. Vol. 61. P. 863–905.

7. Gass S.I., John F. Magee, Assad A. Profiles in Operations Research// International Series in Operations Research & Management Science. 2011. Vol. 147.

8. Gupta P., Sehgal N. K. Introduction to Machine Learning in the Cloud with Python: Concepts and Practices. Springer International Publishing, 2021.

9. Han H., Wang WY., Mao BH. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. // ICIC 2005: Advances in Intelligent Computing. 2005. Vol 3644. P. 878-887.

10. Leon A., Aguirre L. Data Processing with Optimus: Supercharge Big Data Preparation Tasks for Analytics and Machine Learning with Optimus Using Dask and PySpark. Packt Publishing, 2021.

11. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. CatBoost: Unbiased Boosting with Categorical Features. // In Proceedings of the 32nd International Conference on Neural Information Processing Systems: NIPS'18. 2018. P. 6639–6649.

12. Sánchez J. F. M., Lechuga G. P. Assessment of a credit scoring system for Popular Bank Savings and credit // Contaduría y Administración. 2016. P. 391–417.

13. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation // In AI 2006: Advances in Artificial Intelligence. 2006. Vol. 4304. P. 1015–1021.

Оригинальность 96%