ҚАЗАҚСТАН РЕСПУБЛИКАСЫ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ

«Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ» КЕАҚ

Студенттер мен жас ғалымдардың
«ǴYLYM JÁNE BILIM - 2024»
XIX Халықаралық ғылыми конференциясының
БАЯНДАМАЛАР ЖИНАҒЫ

СБОРНИК МАТЕРИАЛОВ
XIX Международной научной конференции
студентов и молодых ученых
«ǴYLYM JÁNE BILIM - 2024»

PROCEEDINGS
of the XIX International Scientific Conference
for students and young scholars
«ǴYLYM JÁNE BILIM - 2024»

2024
Астана

УДК 001
ББК 72
G99

«ǴYLYM JÁNE BILIM – 2024» студенттер мен жас ғалымдардың XIX Халықаралық ғылыми конференциясы = XIX Международная научная конференция студентов и молодых ученых «ǴYLYM JÁNE BILIM – 2024» = The XIX International Scientific Conference for students and young scholars «ǴYLYM JÁNE BILIM – 2024». – Астана: – 7478 б. - қазақша, орысша, ағылшынша.

ISBN 978-601-7697-07-5

Жинаққа студенттердің, магистранттардың, докоранттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов имолодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 001
ББК 72
G99

ISBN 978-601-7697-07-5

Кез келген өңірде, кез келген жағдайда жолды білмеу үлкен қиындықтарға әкелуі мүмкін. Ал ол жағдай қауіп төндіретін болса, адам өмірі қыл үстінде тұрады. Сондықтан осындай қосымшаларды жасауға үлкен назар аударылғаны маңызды, өйткені бұл тек пайдалы құрал ғана емес, адам өмірін сақтап қалуға үлкен көмегін тигізетін жәрдемші. Адам өміріне маңызды өнім шығару кез келген салада басты қағида, маңызды орынды алатын ой, сондықтан, осындай идеяларға көңіл бөлінсе, еліміздің болашағы жарқын болмақ. ГАЖ қосымшаларын талдай отырып, туристерге арналған ГАЖ қосымшасы әзірленді.

Қолданылған әдебиеттер тізімі

1. https://shorturl.at/dzBG5
2. https://apps.apple.com/ru/app/2%D0%B3%D0%B8%D1%81-%D0%BA%D0%B0%D1%80%D1%82%D1%8B-%D0%BD%D0%B0%D0%B2%D0%B8%D0%B3%D0%B0%D1%82%D0%BE%D1%80-D0%B4%D1%80%D1%83%D0%B7%D1%8C%D1%8F/id481627348
3. https://forbes.kz/leader/top-30_mobilnyih_prilojeniy_kazahstana_-_2023_1695006170/
4. https://forbes.kz/news/2023/02/23/newsid_295748
5. https://apps.apple.com/ru/app/%D0%BA%D0%B0%D1%80%D1%82%D1%8B/id91505676
6. https://apps.apple.com/ru/app/%D1%8F%D0%BD%D0%B4%D0%B5%D0%BA%D1%81-%D0%BA%D0%B0%D1%80%D1%82%D1%8B-%D0%B8-%D0%BD%D0%B0%D0%B2%D0%B8%D0%B3%D0%B0%D1%82%D0%BE%D1%80/id313877526
7. https://apps.apple.com/us/app/google-maps/id585027354
8. https://revvy.ai/blog/tpost/vu752d4ez1-yandeks-karti-2gis-google-maps-chto-effe
9. https://skillbox.ru/media/design/chto_takoe_figma/
10. https://habr.com/ru/articles/596183/
11. https://docs.expo.dev/get-started/expo-go/

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ В КРЕДИТНОМ СКОРИНГЕ

Байтемиров Мадияр Ерланович
madiyar5155@mail.ru
Студент факультета информационных технологий, ЕНУ им. Л.Н.Гумилева, Астана, Казахстан Научный руководитель – А. Муханова

Аннотация. Данное исследование посвящено применению методов машинного обучения в кредитном скоринге с целью разработки эффективных моделей прогнозирования. Акцент делается на точности предсказания вероятности дефолта и высокой интерпретируемости моделей для обоснованного принятия банковских решений. Исследование основано на анализе данных из набора Home Credit с Kaggle.com и проведении экспериментов над различными моделями классификации, такими как lgb.LGBMClassifier, LogisticRegression, и LinearDiscriminantAnalysis. Результаты исследования могут быть полезны как для финансовых учреждений, так и для области управления рисками и персональных финансов.

Ключевые слова: Нейронная сеть, машинное обучение, data science, анализ данных, математическая модель.

Введение. В современном мире финансовых отношений кредитный скоринг стал ключевым инструментом для оценки кредитоспособности заемщиков. С целью минимизации рисков и обеспечения финансовой устойчивости банковской системы, анализ данных стал

неотъемлемой частью процесса выдачи кредитов. В этом контексте внимание исследователей все более фокусируется на создании эффективных моделей прогнозирования, способных точно выявлять потенциальные случаи дефолта [1].

Данный исследовательский вопрос не только актуален для банков и кредитных организаций, но и имеет широкие практические применения, например, в области управления рисками и принятия решений в сфере персональных финансов [2]. Обзор предыдущих исследований в области кредитного скоринга позволяет выявить существующие методы и проблемы, что служит фундаментом для предложения новых подходов и улучшений.

Таким образом, исследование призвано внести вклад в область кредитного скоринга, предоставив новые перспективы на основе современных методов машинного обучения и данных, доступных в сфере финансов [3]. В следующих разделах статьи подробно рассмотрим используемый набор данных, методологию проведения исследования, а также результаты и обсуждение, направленные на достижение поставленных целей [4].

1. **Данные**. Для построения моделей было взять набор данных Home Credit с kaggle.com. В наборе данных есть колонка TARGET [5]. Это целевая переменная (1 - клиент с трудностями с оплатой: у него/нее была просрочка платежа, 0 - все остальные случаи).

2. **Модели и методы**. В ходе исследовательской работы, было решено провести эксперименты над 6 моделями классификации.

lgb.LGBMClassifier - это классификатор градиентного бустинга, реализованный в библиотеке LightGBM. LightGBM (Light Gradient Boosting Machine) - это быстрый и эффективный алгоритм градиентного бустинга, предназначенный для обработки больших объемов данных и использования распределенных вычислений [6].

LogisticRegression - это классификатор логистической регрессии в библиотеке scikit-learn. Логистическая регрессия - это метод бинарной классификации, который предсказывает вероятность принадлежности объекта к одному из двух классов. Он использует логистическую функцию для преобразования взвешенной суммы признаков в вероятность [7].

LinearDiscriminantAnalysis - это метод линейного дискриминантного анализа (LDA) в библиотеке scikit-learn. LDA - это метод, используемый в статистике и машинном обучении для поиска комбинаций признаков, которые лучше всего разделяют два или более класса [8].

Decision Tree Classifier - это классификатор дерева решений в библиотеке scikit-learn. Дерево решений - это алгоритм машинного обучения, который принимает решения на основе серии вопросов о признаках данных. Он разбивает данные на подгруппы, пока не достигнет условия останова или достигнет максимальной глубины [9].

Gradient Boosting Classifier - это классификатор, реализующий метод градиентного бустинга для задачи классификации в библиотеке scikit-learn. Градиентный бустинг - это метод ансамблевого обучения, который строит ансамбль слабых моделей (обычно деревьев решений) и комбинирует их для получения более точной и устойчивой модели.

XGBClassifier - это классификатор градиентного бустинга, реализованный в библиотеке XGBoost (Extreme Gradient Boosting). XGBoost представляет собой высокоэффективный и мощный алгоритм градиентного бустинга, который применяется для задач классификации и регрессии [10].

3. Метрики. Для оценки качества моделей в задачах классификации используются различные метрики. В данной работе, будут использованы одни из наиболее распространенных метрик классификации:

1. **Accuracy** (Точность): Показывает долю правильных предсказаний по отношению к общему количеству наблюдений.

2. **Precision** (Точность): Показывает долю правильно предсказанных положительных классов относительно всех предсказанных положительных классов.

3. **Recall** (Полнота): Показывает долю правильно предсказанных положительных классов относительно всех истинных положительных классов.

4. **F1 Score**: Сбалансированная метрика, объединяющая Precision и Recall. F1 Score близок к 1, если и Precision, и Recall высоки.

5. **ROC-AUC** (Receiver Operating Characteristic - Area Under the Curve): Метрика, измеряющая площадь под кривой ROC. Оценивает способность модели различать между классами.

6. **Specificity** (Специфичность): Показывает долю правильно предсказанных отрицательных классов относительно всех истинных отрицательных классов.

TN (True Negative), TP (True Positive), FN (False Negative), и FP (False Positive) - это четыре элемента матрицы ошибок (confusion matrix), которая используется для оценки производительности модели классификации. Давайте рассмотрим каждый из них:

1. TN (True Negative): Количество отрицательных примеров, которые модель правильно классифицировала как отрицательные.
2. TP (True Positive): Количество положительных примеров, которые модель правильно классифицировала как положительные.
3. FN (False Negative): Количество положительных примеров, которые модель неправильно классифицировала как отрицательные.
4. FP (False Positive): Количество отрицательных примеров, которые модель неправильно классифицировала как положительные.

4. Эксперименты и сравнение моделей. Так как задача заключается в бинарной классификации (две категории), то округление предсказанных вероятностей (y_pred) обычно является распространенной практикой.

Округление обычно выполняется с использованием порога 0,5: вероятности, равные или превышающие 0,5, округляются до 1 (положительный класс), а вероятности ниже 0,5 округляются до 0 (отрицательный класс).

Таким образом, возвращая вероятности принадлежности к классам для каждого объекта. Предсказанные метки классов для каждого объекта округляются до 0 либо 1.

Таблица №1. Метрики на обучающей выборке (Train).

Название модели	roc_auc	Accuracy	Precision	Recall	Specificity	F1	TN	TP	FN	FP	Порог
LogisticRegression	0,499996	0,919138	0	0	0,919144	0	141323	0	1	12432	0,5
LinearDiscriminantAnalysis	0,507455	0,919093	0,016409	0,490385	0,920256	0,031756	141112	204	212	12228	0,5
GradientBoostingClassifier	0,507555	0,919008	0,016731	0,475973	0,920271	0,032326	141095	208	229	12224	0,5
LGBMClassifier	0,509736	0,918566	0,02204	0,430141	0,920598	0,041931	140961	274	363	12158	0,5
XGBClassifier	0,510097	0,919229	0,02204	0,51215	0,920651	0,042261	141063	274	261	12158	0,5
DecisionTreeClassifier	**0,540398**	0,851401	0,169402	0,14397	0,925781	0,155654	128802	2106	12522	10326	0,5

Таблица №2. Метрики на тестовой выборке (Test).

Название модели	roc_auc	Accuracy	Precision	Recall	Specificity	F1	TN	TP	FN	FP	Порог
LogisticRegression	0,499996	0,919138	0	0	0,919144	0	141323	0	1	12432	0,5
LinearDiscriminantAnalysis	0,507455	0,919093	0,016409	0,490385	0,920256	0,031756	141112	204	212	12228	0,5
GradientBoostingClassifier	0,507555	0,919008	0,016731	0,475973	0,920271	0,032326	141095	208	229	12224	0,5
LGBMClassifier	0,509736	0,918566	0,02204	0,430141	0,920598	0,041931	140961	274	363	12158	0,5
XGBClassifier	0,510097	0,919229	0,02204	0,51215	0,920651	0,042261	141063	274	261	12158	0,5
DecisionTreeClassifier	**0,540398**	0,851401	0,169402	0,14397	0,925781	0,155654	128802	2106	12522	10326	0,5

Как видно по таблицам 1-2, в основном были правильно предсказаны отрицательные классы. Это заметно по Specificity.

Но здесь нужно настроить порог решения в соответствии с требованиями или найти компромисс между ложноположительными и ложноотрицательными результатами. Поэтому, было решено экспериментировать с различными значениями порога и наблюдать, как это влияет на метрики, такие как precision, recall и F1 score.

Таблица №3. Эксперименты на обучающей выборке (Train).

Таблица №3. Эксперименты на обучающей выборке (Train).

Название модели	roc_auc	Accuracy	Precision	Recall	Specificity	F1	TN	TP	FN	FP	Порог
LogisticRegression	0,590898	0,584469	0,598564	0,11183	0,943092	0,188451	82447	7418	58915	4975	0,08
LinearDiscriminantAnalysis	0,674329	0,683464	0,663439	0,155956	0,958717	0,252546	96864	8222	44498	4171	0,08
GradientBoostingClassifier	0,685263	0,700862	0,666667	0,164828	0,960137	0,264308	99499	8262	41863	4131	0,08
XGBClassifier	0,712004	0,698865	0,727669	0,17361	0,96685	0,280336	98436	9018	42926	3375	0,08
LGBMClassifier	0,712356	0,736061	0,684096	0,187795	0,963954	0,294692	104695	8478	36667	3915	0,09
DecisionTreeClassifier	**1**	1	1	1	1	1	141362	12393	0	0	0,005

Таблица №4. Эксперименты на тестевой выборке (Test).

Название модели	roc_auc	Accuracy	Precision	Recall	Specificity	F1	TN	TP	FN	FP	Порог
DecisionTreeClassifier	0,540398	0,851401	0,169402	0,14397	0,925781	0,155654	128802	2106	12522	10326	0,005
LogisticRegression	0,594331	0,618519	0,565476	0,116617	0,942208	0,193358	88071	7030	53253	5402	0,085
LinearDiscriminantAnalysis	0,672446	0,660703	0,686454	0,150233	0,959794	0,246516	93053	8534	48271	3898	0,075
GradientBoostingClassifier	0,679462	0,697065	0,658462	0,162041	0,958872	0,260079	98992	8186	42332	4246	0,08
LGBMClassifier	0,682054	0,689222	0,673504	0,160716	0,960072	0,259507	97599	8373	43725	4059	0,08
XGBClassifier	**0,683342**	0,669203	**0,700209**	0,155892	**0,961937**	0,255009	94189	8705	47135	3727	0,075

Результаты. На таблицах 1 и 3, метрики модели Decision Tree Classifier на тренировочных и на тестовых данных, остались без изменений. Это связано с тем, что дерево не возвращает предсказания в виде вероятности, а возвращает его целым числом.

По таблице 1-2: Accuracy в моделях высок на начальных таблицах с порогом равным 0.5, так как в исследуемых данных значительно превышается количество одного класса, над другим. Модели предсказывают хорошо данные плохих клиентов, но плохо предсказывают хороших. Поэтому следует ориентироваться на остальные метрики.

По таблице 3-4: Основные метрики значительно увеличились у всех моделей, кроме решающего дерева, что говорит о положительном влиянии правильного подбора порогового значения.

Вывод. В процессе исследования были рассмотрены 6 различных моделей классификации на данных кредитного скоринга. Каждая из моделей предоставила свои уникальные результаты и показатели. Анализ производительности моделей осуществлялся с учетом различных метрик, таких как точность, полнота, F1-мера, а также кривые ROC-AUC.

Основываясь на проведенных экспериментах, было выявлено, что XGBClassifier, например, показала наилучшие результаты с точки зрения roc_auc, Precision, F1 score и Specificity.

Также было выявлено, улучшение предсказательной способности моделей с различными значениями порога при округлении вероятности принадлежности к классам для каждого объекта. Это влияет на метрики, такие как roc_auc, precision, recall, Specificity и F1 score. Поэтому рекомендуется найти компромисс между ложноположительными и ложноотрицательными результатами.

Список использованной литературы

[1] Nicolas Suhadolnik & Jo Ueyama & Sergio Da Silva, (2023). "Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach," JRFM, MDPI, vol. 16(12), pages 1 - 21, November.

[2] Tatsat, H., Puri, S., & Lookabaugh, B. (2021). Machine Learning and Data Science Blueprints for Finance: From Building Trading Strategies to Robo-Advisors Using Python.

[3] Boguslauskas, V., & Mileris, R. (2009). Estimation of credit risks by artificial neural networks models. Izinerine Ekonomika-Engerrring Economics, 4, 7–14.

[4] Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation.

Quarterly Review of Economics and Finance, 48, 733–755.

[5] Website: https://www.kaggle.com/competitions/home-credit-default-risk/data

[6]Website: https://thecleverprogrammer.com/2021/01/15/lightgbm-in-machine-learning/

[7]Website:https://mlcheatsheet.readthedocs.io/en/latest/logistic_regression.html

[8]Website: https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

[9] Website: https://scikit-learn.org/stable/modules/tree.html

[10]Website:https://www.apmonitor.com/pds/index.php/Main/XGBoostClassifier