ЭКОНОМИКА И МЕНЕДЖМЕНТ

DOI 10.24412/1829-0450-2025-2-10-23
УДК 336.7

Поступила: 08.03.2025г.
Сдана на рецензию: 10.03.2025г.
Подписана к печати: 25.05.2025г.

ПОСТРОЕНИЕ СКОРИНГОВОЙ МОДЕЛИ НА ОСНОВЕ НОВОГО АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ

Г.Г. Аракелян

Армянский государственный экономический университет
g.arakelyan0393@gmail.com
ORCID: 0000-0001-6217-9681

АННОТАЦИЯ

Банковская система, являясь важнейшим сектором экономики, часто сталкивается с разными рисками. Основным из таких рисков является риск возможных потерь при кредитовании, поэтому финансовым организациям важно правильно оценивать кредитоспособность заемщика.

Данная статья посвящена решению задачи бинарной классификации путем разработки нового алгоритма машинного обучения. В ходе исследования на основе нового метода и реальных данных кредитования была построена скоринговая модель для оценки кредитоспособности заемщиков, которая возвращает скор баллы в диапазоне 300–850. Были оценены такие качественные показатели полученной модели как ROC AUC, PR AUC, Precision, Recall, F1score.

Результаты исследования показывают, что применение построенной модели на практике может снизить уровень кредитного риска и защитить финансовые организации от существенных потерь.

Ключевые слова: кредитный риск, машинное обучение, кредитоспособность, скоринг, бинарная классификация.

Введение

Каждое кредитное учреждение, в том числе и банки, осуществляют свою деятельность в нестабильной среде и, не имея полной информации о ней, могут понести финансовые потери. Одним из основных источников потерь являются кредитные риски, эффективное управление которыми является залогом формирования здорового кредитного портфеля.

В 1997г. Базельский комитет по банковскому надзору назвал кредитный риск «основным видом финансового риска», с которым сталкиваются финансовые институты в своей деятельности. Этот факт отражает прокатившуюся

по всему миру в 1980–1990-х годах волну корпоративных банкротств, ставших результатом кредитного риска [1].

Для оценки кредитного риска при кредитовании физических лиц часто используются различные математические модели, в том числе скоринговые модели. На данный момент в РА при кредитовании часто используется скоринговая модель FICO Score, которая разработана компанией Fair Isaac Services Limited [2]. Данная модель и ее аналоги представляют собой «черный ящик», и кредитные учреждения часто не могут точно определить причину формирования того или иного скор-балла. Данная скоринговая модель построена на основании тех же весовых коэффициентов, которые используются в других странах, из-за чего она полностью не описывает социально-экономическую ситуацию в РА. Учитывая вышесказанное, кредитные учреждения с целью точной оценки и управления кредитными рисками собирают информацию о своих заемщиках и предоставленных им кредитах и строят различные математические модели на основе этой информации. Используя такие модели, кредиторы принимают решения об утверждении или отклонении кредитных заявок. Исторически для решения подобных задач использовалась модель логистической регрессии. Однако в последнее время с увеличением вычислительных мощностей современных компьютеров часто для создания подобных решений используются такие эффективные алгоритмы машинного обучения, как нейронные сети, XGBoost, LightGBM и т.д.

Целью данного исследования является изучение поведения заемщиков на основании реальных данных и построение нового алгоритма машинного обучения.

Результаты исследования могут быть использованы кредитными учреждениями для оптимального управления кредитными рисками и снижения финансовых потерь.

Основная часть

В рамках исследования использовались различные методы анализа данных. Так для выявления закономерностей и связей в данных использовался корреляционный анализ.

Корреляция показывает статистическую взаимосвязь двух или нескольких случайных величин [3].

Для обработки категориальных переменных был использован “One Hot Encoding” метод, который представляет из себя метод преобразования категориальных переменных в двоичный формат. При его использовании создаются новые двоичные столбцы (1 указывает на наличие данной категории, а 0 указывает на ее отсутствие) для каждой категории в исходной переменной [4].

Для обработки числовых переменных использовался “Standard Scaler”-метод, который работает по принципу нормализации данных. Метод преобразует распределение каждой функции так, чтобы среднее значение было равно нулю, а стандартное отклонение – единице [5].

В результате исследования был разработан алгоритм машинного обучения для решения задачи бинарной классификации. Рассмотрим шаги для реализации указанного алгоритма.

1. Генеральную совокупность данных (S), в зависимости от зависимой переменной (y), нужно разделить на две подвыборки. Первая подвыборка (S_0) должна содержать все записи, у которых зависимая переменная имеет значение 0, а вторая подвыборка (S_1) должна содержать все записи, у которых зависимая переменная имеет значение 1. Имеет место:

$$S_0 = \{x_i \in S \,|y_i = 0\}, S_1 = \{x_i \in S \,|y_i = 1\},$$

где x_i – вектор переменных для i-го наблюдения.

$$S = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2j} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{31} & x_{32} & \ldots & x_{ij} & y_i \end{matrix} \; y \in (0,1),$$

$$S_0 = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{31} & x_{32} & \cdots & x_{ij} \end{matrix} \; y = 0,$$

$$S_1 = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{31} & x_{32} & \cdots & x_{ij} \end{matrix} \; y = 1$$

2. Для каждой подвыборки из пункта 1 необходимо найти центроиды, согласно ниже приведенным формулам:

$$C_0 = \left(\frac{1}{s_0}\sum_{i=1}^{n} x_{i1}, \frac{1}{s_0}\sum_{i=1}^{n} x_{i2}, \ldots, \frac{1}{s_0}\sum_{i=1}^{n} x_{ij}\right),$$

$$C_1 = \left(\frac{1}{s_1}\sum_{i=1}^{n} x_{i1}, \frac{1}{s_1}\sum_{i=1}^{n} x_{i2}, \ldots, \frac{1}{s_1}\sum_{i=1}^{n} x_{ij}\right),$$

где C_0 и C_1 – центроиды выборок, а s_0 и s_1 – количество наблюдений в соответствующих выборках.

3. Рассчитываем расстояния от точки, которую нужно классифицировать, до центроидов каждого класса и до заранее выбранных N ближайших соседей. Для простоты расчетов в рамках данной статьи для расчета расстояния используется формула Евклида:

$$p = (p_1, p_2, \dots, p_n)\ \ q = (q_1, q_2, \dots, q_n)$$

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2},$$

где p и q – две точки в n мерном пространстве.

4. На основе расчетов третьего пункта рассчитывается влияние каждого центроида и каждого ближайщего соседа, согласно формуле:

$$Impact = \frac{1}{d(p,q)^k + e},$$

где $d(p,q)$ – расстояние между точками p и q, k – коэффициент воздействия, e – рациональное число.

5. Рассчитывается вероятность принадлежности новой точки, которую нужно классифицировать, к классу 0 и 1, согласно ниже указанным формулам:

$$Impact\ (0) = Impact\ (C_0) + \sum_{i=1}^{N0} Impact(d0_i)\,, N0 \in N$$

$$Impact\ (1) = Impact\ (C_1) + \sum_{i=1}^{N1} Impact(d1_i)\,, N1 \in N$$

$$Probability(0) = \frac{Impact\ (0)}{Impact\ (0) + Impact\ (1)},$$

$$Probability(1) = \frac{Impact\ (1)}{Impact\ (0) + Impact\ (1)},$$

где $Impact\ (C_0)$ и $Impact\ (C_1)$ – воздействие центроидов в классах 0 и 1, соответственно, $Impact(d0_i)$ и $Impact(d1_i)$ – воздействие ближайщих N соседей в классах 0 и 1, соответственно, $Probability(0)$ и $Probability(1)$ – вероятности принадлежности новой точки к классу 0 и 1, соответственно.

6. В целях улучшения качества алгоритма можно применить также ансамблевые методы. Тем самым можно изначально найти оптимальные пара-

метры k и N, затем создать m – количество моделей с указанными параметрами, которые отдельно обучаются на разных наборах данных. В данном случае в качестве конечной вероятности применяется усредненная вероятность.

На Рис. 1 представлена диаграмма выше указанного алгоритма.

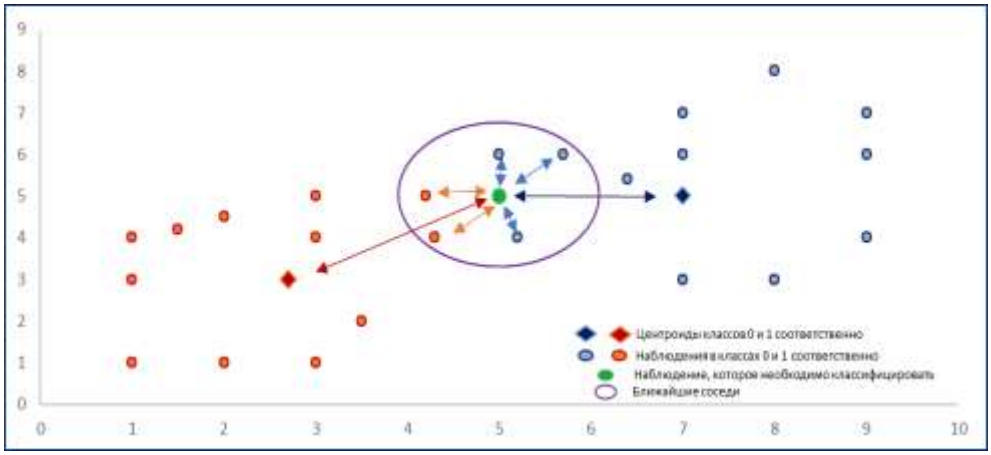

Рисунок 1. Схема алгоритма классификации.

К преимуществам указанного алгоритма можно отнести:

- алгоритм прост в реализации и результаты прогнозирования легко интерпретируемы;
- алгоритм рассчитывает расстояния до центроидов классов и ближайших соседей: это позволяет одновременно учитывать обобщенное положение классов в признаковом пространстве и локальную структуру данных: такая комбинация повышает точность классификации при сложном распределении объектов в пространстве признаков;
- за счет использования центроидов, алгоритм становится менее чувствительным к единичным выбросам и зашумленным наблюдениям, что часто свойственно реальным наборам данных, благодаря использованию ближайших соседей компенсируются потенциальные потери точности, которые возникают при использовании исключительно усредненной информации;
- за счет комбинации центроидов и ближайших соседей улучшается обобщающая способность модели и снижается вероятность переобучения.

В рамках данного исследования использовались реальные данные о кредитах «Юнибанк» ОАО. В табличных данных имеются 13 переменных. Выборка содержит 4603 наблюдений. Рассмотрим независимые переменные.

1. **Семейное положение**. В исходных данных эта переменная имела значения M1, M2, M3 и M4. Каждое значение было изменено на «Женат (замужем)», «Вдовец (вдова)», «Разведен (разведена)» и «Одинокий (одинокая)», соответственно.

2. **Образование**. В исходных данных эта переменная имела значения E1, E2, E3 и E4. Каждое значение было изменено на «Ученая степень», «Высшее образование», «Среднее профессиональное» и «Средняя школа», соответственно.

3. **Наличие имущества**. В исходных данных эта переменная имела значения P1, P2, P3 и P4. Каждое значение было изменено на «Наличие недвижимости», «Наличие движимого имущества», «Наличие недвижимости и движимого имущества» и «Отсутствие», соответственно.

4. **Пол**. В исходных данных эта переменная имела значения S1 и S2. Каждое значение было изменено на «Женский» и «Мужской», соответственно.

5. **Возраст заемщика**. Числовая переменная, которая содержит значения от 20 до 65.

6. **Количество дней просрочки за последние 12 месяцев**. Числовая переменная, которая содержит значения от 0 до 1600.

7. **Количество просрочек**. Числовая переменная, которая содержит значения от 1 до 67. Здесь имеются пропущенные данные. На основании полученной от банка информации стало ясно, что у данных клиентов нет просроченной задолженности. Поэтому пропущенные значения были изменены на 0.

8. **Количество изменений в классах риска**. Числовая переменная, которая содержит значения от 0 до 28.

9. **Кредитная нагрузка**. Числовая переменная, которая содержит значения от 0 до 2 984 303. Данные представлены в драмах.

10. **Длительность кредитной истории**. Числовая переменная, которая содержит значения от 0 до 488. Данные представлены в днях.

11. **Максимально погашенные кредиты**. Числовая переменная, которая содержит значения от 0 до 25 131 048. Данные представлены в драмах.

12. **Сумма контракта**. Числовая переменная, которая содержит значения от 30 000 до 2,300,000. Данные представлены в драмах РА. Здесь имеются выбросы. Это – клиенты, которым банк выдал кредит на сумму более 1,000,000 драмов РА. Таких наблюдений 7. Было подсчитано, что удаление данных наблюдений не оказывает существенного влияния на качество моделей, поэтому при создании моделей эти наблюдения не удалялись из набора данных.

13. **Дефолт**. Бинарная переменная (1 – дефолт, 0 – нет дефолта). На основании полученной от банка информации стало ясно, что дефолтом считается наличие хотя бы одного раза просроченных обязательств на 90 дней и более за последний год.

Исследование данных и моделирование осуществлялись с применением программы Excel, языка программирования Python и его соответствующих библиотек.

В Табл. 1 представлены описательные статистики для числовых переменных. Отсюда можно увидеть, что пропущенных значений не имеется.

Таблица 1. Описательные статистики.

Показатель	Сумма контракта	Возраст	Кол-во дней просрочки за последние 12 месяцев	Кол-во просрочек	Кол-во изменений в классах риска	Кред. нагрузка	Длительность кредитной истории	Макс. погаш. кредиты	Дефолт
count	4 603	4 603	4 603	4 603	4 603	4 603	4 603	4 603	4 603
mean	254229,24	37,54	4,03	2,71	0,55	471551,5	59,39	332959,85	0,43
std	144770,86	12,09	38,54	4,28	1,46	537158,02	61,27	920340,39	0,49
min	30000	20	0	0	0	0	0	0	0
25%	181900	27	0	0	0	0	13	0	0
50%	200000	36	0	1	0	300000	40	150000	0
75%	377050	48	0	4	0	825100	86	376000	1
max	2300000	65	1600	67	28	2984303	488	25131048	1

В Табл. 2 представлена корреляционная матрица для числовых переменных, откуда видно, что имеется умеренная связь между переменными «Количество изменений в классах риска» и «Количество просрочек», а также «Длительность кредитной истории» и «Кредитная нагрузка».

Таблица 2. Корреляционная матрица.

	Сумма контракта	Возраст	Кол-во дней просрочки за последние 12 месяцев	Кол-во просрочек	Кол-во изменений в классах риска	Кред. нагрузка	Длительность кредит. истории	Макс. погаш. кредиты	Дефолт
Сумма контракта	1,00	0,07	0,04	0,01	0,01	0,07	0,10	0,13	0,10
Возраст	0,07	1,00	0,01	0,05	0,08	0,14	0,21	0,12	-0,08
Кол-во дней просрочки за последние 12 месяцев	0,04	0,01	1,00	0,34	0,38	-0,01	0,09	0,14	0,08

	Сумма контракта	Возраст	Кол-во дней просрочки за последние 12 месяцев	Кол-во просрочек	Кол-во изменений в классах риска	Кред. нагрузка	Длительность кредит. истории	Макс. погаш. кредиты	Дефолт
Кол-во просрочек	0,01	0,05	0,34	1,00	0,57	0,05	0,35	0,18	0,04
Кол-во изменений в классах риска	0,01	0,08	0,38	0,57	1,00	0,00	0,22	0,12	-0,05
Кред. нагрузка	0,07	0,14	-0,01	0,05	0,00	1,00	0,50	0,17	0,15
Длительность кред. истории	0,10	0,21	0,09	0,35	0,22	0,50	1,00	0,26	-0,05
Макс. погаш. кредиты	0,13	0,12	0,14	0,18	0,12	0,17	0,26	1,00	0,00
Дефолт	0,10	-0,08	0,08	0,04	-0,05	0,15	-0,05	0,00	1,00

В Табл. 3 представлены данные по группам для каждой переменной в зависимости от наличия дефолта. Дефолт больше всего сконцентрирован среди клиентов женского пола (46,83%), клиентов в возрастной группе 20–25 (50,49%), клиентов, у которых семейное положение в группе «Вдовец (вдова)» (54,14%), клиентов со средне профессиональным образованием (42,83%), клиентов, у которых имеются недвижимое и движимое имущества (42,67%), клиентов, у которых кредитная нагрузка больше 300,000 драм (51,74%), а также клиентов, у которых количество дней просрочки за последние 12 месяцев больше 30 дней (87,04%).

Таблица 3. Группы имеющихся данных в зависимости от наличия дефолта.

Группа	Признак дефолта		Всего	Дефолт (%)
	Нет	Да		
Пол				
Мужской	1,492	948	2,440	38,85%
Женский	1,150	1,013	2,163	46,83%
Возраст				
1. 20–25	457	466	923	50,49%
2. 26–35	760	589	1,349	43,66%
3. 36–50	880	558	1,438	38,80%
4. 51+	545	348	893	38,97%
Семейное положение				

Группа	Признак дефолта		Всего	Дефолт (%)
	Нет	Да		
Пол				
Разведен (разведена)	46	13	59	22,03%
Женат (замужем)	2,139	1,428	3,567	40,03%
Одинокий (одинокая)	42	30	72	41,67%
Вдовец (вдова)	415	490	905	54,14%
Образование				
Ученая степень	343	255	598	42,64%
Среднее образование	3	2	5	40,00%
Высшее образование	626	453	1,079	41,98%
Среднее профессиональное	1,670	1,251	2,921	42,83%
Наличие имущества				
Наличие недвижимого и движимого имущества	133	99	232	42,67%
Наличие недвижимого имущества	153	122	275	44,36%
Наличие движимого имущества	611	464	1,075	43,16%
Отсутствие имущества	1,745	1,276	3,021	42,24%
Количество просрочек				
0	1,349	849	2,198	38,63%
1 +	1,293	1,112	2,405	46,24%
Кредитная нагрузка				
0	1,050	424	1,474	28,77%
1–300,000	482	347	829	41,86%
300,001+	1,110	1,190	2,300	51,74%
Количество дней просрочки за последние 12 месяцев				
0–30	2,635	1,914	4,549	42,08%
31+	7	47	54	87,04%
Количество изменений в классах риска				
0	1,989	1,595	3,584	44,50%
1+	653	366	1,019	35,92%
Длительность кредитной истории (день)				
0–270	2,612	1,941	4,553	42,63%
271–365	24	15	39	38,46%
366+	6	5	11	45,45%
Максимально погашенные кредиты				

Группа	Признак дефолта		Всего	Дефолт (%)
	Нет	Да		
Пол				
0–350,000	1,943	1,464	3,407	42,97%
350,001+	699	497	1,196	41,56%

После анализа данных была создана модель, согласно выше указанному алгоритму, с применением ансамблевого метода. Были оценены качественные показатели полученной модели, а полученные вероятности были приведены в скор-балы в диапазоне от 300 до 850.

В целях моделирования были осуществлены следующие действия.

1. В целях оценки качества построенной модели имеющиеся данные были разделены на обучающую (80% от общего объема данных) и тестовую (20% от общего объема данных) выборки. Это – стандартная практика, позволяющая объективно оценить способность модели обобщать информацию на новых, ранее не виденных данных.

2. Для некатегориальных переменных (кроме «Количество изменений в классах риска») осуществлена стандартизация с применением инструмента Standard Scaler из библиотеки Sklearn [6] языка программирования Python, который позволяет преобразовать данные таким образом, что они приобретают нулевое среднее значение и единичное стандартное отклонение. Чтобы избежать утечки данных, данное действие было осуществлено отдельно для каждой полученной выборки из пункта 1.

3. С применение инструмента get_dummies из библиотеки Pandas [7] языка программирования Python все категориальные переменные были преобразованы в числовой формат. Это позволяет включить такие переменные в модель без нарушения ее математических допущений. В данном случае для каждого уникального значения категориального признака создается отдельный бинарный столбец, что делает данные пригодными для разработанного алгоритма.

Модель согласно выше указанному алгоритму была создана с применением языка программирования Python. Была проведена оптимизация гиперпараметров модели. Так оптимальным количеством ближайщих N соседей является 12 (диапазон поиска: 2–15), значением коэффициента k является 0.0001 (диапазон поиска: 0.0001–0.2001), а количество моделей в ансамбле составляет 300.

В качестве наилучшего порога классификации (Threshold) было выбрано 0,4355 значение, при котором Accuracy составляет 0,72095, Precision составляет 0,7257, Recall составляет 0,7209, а F1– мера [8] составляет 0,7224. На Рис. 2 представлена матрица ошибок [9], согласно которой модель идентифицирует 70,4% дефолтных клиентов (266/378), однако при этом модель ошибочно

классифицирует 26,7% недефолтных клиентов как дефолтных (145/543). Выше приведенные показатели рассчитаны для тестовых данных.

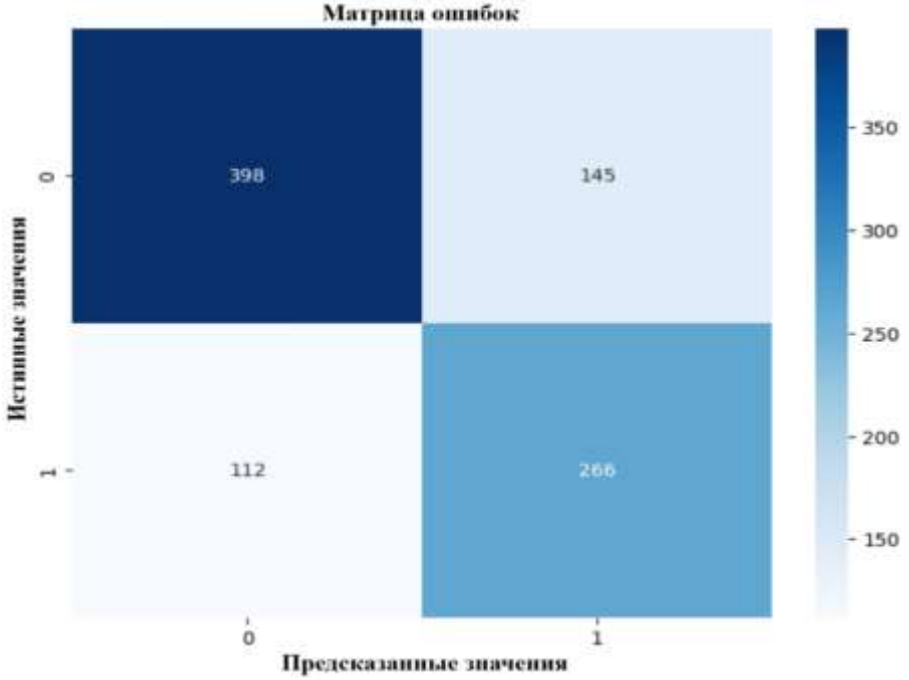

Рисунок 2. Матрица ошибок для тестовых данных.

Полученные вероятности тестовых данных для класса 0 были приведены в скор-баллы, согласно ниже указанной формуле:

$$Score = \left(\ln\left(\frac{P(y=0)}{1-P(y=0)}\right) * \frac{850-300}{|Percentile_1| + |Percentile_{99}|}\right) + \left(850 - \frac{850-300}{|Percentile_1| + |Percentile_{99}|} * Percentile_1\right),$$

где $Score$ – скор-балл, $P(y=0)$ – вероятность принадлежности наблюдения к классу 0, $Percentile_1$ и $Percentile_{99}$ – 1-ый и 99-ый персентили для прологарифмированного ряда.

На Рис. 3 изображены кривые ROC и PR [10] для полученной модели, а на Рис. 4 изображено распределение полученных скор баллов.

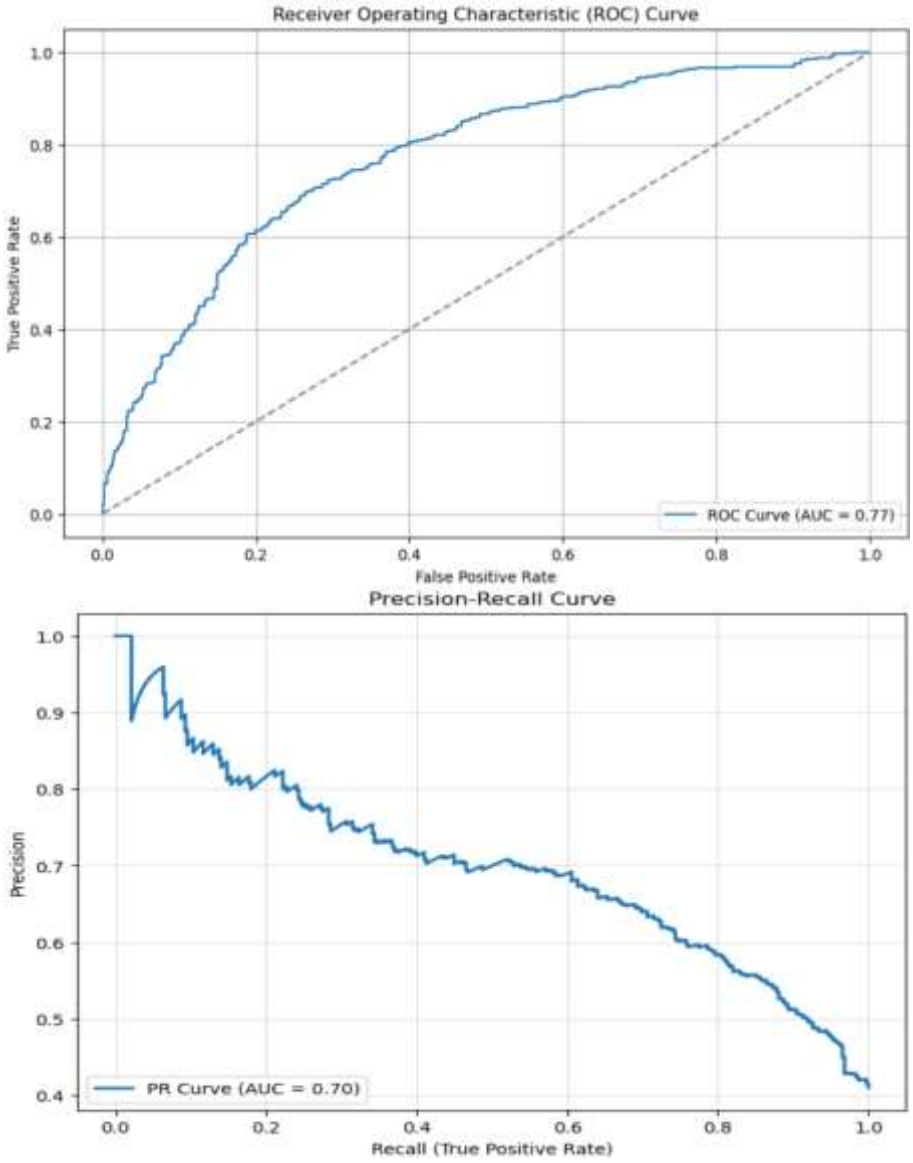

Рисунок 3. ROC и PR-кривые.

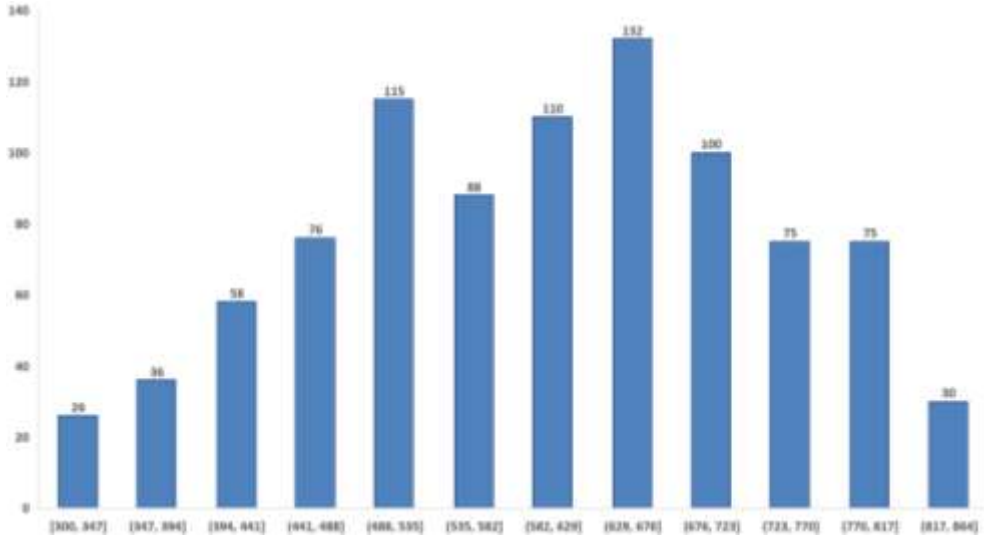

Рисунок 4. Распределение скор баллов на тестовой выборке.

Заключение

Для любой кредитной организации важно своевременное выявление и управление кредитными рисками. Для управления данным видом риска важно внедрить систему принятия решений. Целью данного исследования была разработка нового алгоритма бинарной классификации для выявления дефолтных клиентов кредитных организаций. На основе полученных результатов можно сделать вывод, что комбинация идей, лежавших в основе классических моделей машинного обучения, дает возможность создания новых алгоритмов машинного обучения.

К основным преимуществам данного исследования следует отнести разработку гибридного алшоритма, который сочетает сильные стороны анализа ближайших соседей и центроидов. Разработанный алгоритм повышает устойчивость классификации к шуму и выбросам данных, а также снижает риск переобучения модели. Данный алгоритм дает возможность оптимально сдерживать кредитный риск, и, как следствие, снижать убытки по возможным потерям по ссудам.

ЛИТЕРАТУРА

1. Энциклопедия финансового риск-менеджмента. Под ред. А.А. Лобанова и А.В. Чугунова. М: «Альпина Паблишер», 2003. С. 323.
2. ՖՊՅԿՈ սրոր: URL: https://acra.am/սրոր/
3. Корреляция (Correlation). URL: https://wiki.loginom.ru/articles/correlation.html
4. One Hot Encoding in Machine Learning URL: https://www.geeksforgeeks.org/ml-one-hot-encoding/
5. What is StandardScaler? URL: https://www.geeksforgeeks.org/what-is-standardscaler/

6. Sklearn User Guide. URL: https://scikit-learn.org/stable/user_guide.html
7. Pandas User Guide. URL: https://pandas.pydata.org/docs/user_guide/index.html
8. Evaluation Metrics for Classification. URL: https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d
9. What is A Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained URL: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning
10. Understanding AUC–ROC and Precision-Recall Curves. URL: https://medium.com/@data.science.enthusiast/auc-roc-curve-ae9180eaf4f7

REFERENCES

1. Encyclopedia of financial risk management. Under red. A. Lobanova and A. Chugunova. – M: "Alpina Publisher", 2003. P. 323.
2. FICO score. URL: https://acra.am/upnp/
3. Correlation. URL: https://wiki.loginom.ru/articles/correlation.html
4. One Hot Encoding in Machine Learning. URL: https://www.geeksforgeeks.org/ml-one-hot-encoding/
5. What is StandardScaler? URL: https://www.geeksforgeeks.org/what-is-standardscaler/
6. Sklearn User Guide. URL: https://scikit-learn.org/stable/user_guide.html
7. Pandas User Guide. URL: https://pandas.pydata.org/docs/user_guide/index.html
8. Evaluation Metrics for Classification. URL: https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d
9. What is A Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained URL: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning
10. Understanding AUC – ROC and Precision-Recall Curves. URL: https://medium.com/@data.science.enthusiast/auc-roc-curve-ae9180eaf4f7

CONSTRUCTION OF A SCORING MODEL BASED ON A NEW MACHINE LEARNING ALGORITHM

G. Arakelyan
Armenian State University of Economics

ABSTRACT

The banking system, being the most important sector of the economy, often faces various risks. The main one of these risks is the risk of possible losses when lending. As a result, it is important for financial institutions to correctly assess the creditworthiness of the borrower.

This article is devoted to solving the problem of binary classification by developing a new machine learning algorithm. During the study, based on the new method and real lending data, a scoring model was built to assess the creditworthiness of borrowers, which returns scores in the range of 300–850. Such qualitative indicators of the resulting model as ROC AUC, PR AUC, Precision, Recall, F1 score were evaluated.

The results of the study show that the application of the constructed model in practice can reduce the level of credit risk and protect financial institutions from significant losses.

Keywords: credit risk, machine learning, creditworthiness, scoring, binary classification.