МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

МЕТОДЫ КЛАССИФИКАЦИИ И ИХ ПРИЛОЖЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Каревой Ксении Александровны

Научный руководитель
зав.каф., д.ф.-м.н., доцент _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

ВВЕДЕНИЕ

Задачи классификации занимают центральное место в современной науке о данных и прикладной математике. Под классификацией понимается отнесение объектов или явлений к одному из заранее определённых классов на основе набора признаков. Область её применения охватывает самые разные отрасли — от финансового сектора и медицинской диагностики до метеорологии и систем технического зрения.

Цель работы — провести сравнительный анализ методов классификации и исследовать их практическую применимость в двух предметных областях: кредитном скоринге в банковской сфере и классификации форм атмосферной циркуляции по методу Вангенгейма–Гирса.

Для достижения поставленной цели необходимо решить следующие задачи:

— Рассмотреть теоретические основы методов классификации в машинном обучении, выявить ключевые различия между ними с точки зрения математического аппарата и вычислительной сложности;
— Реализовать и обучить пять выбранных моделей классификации на данных кредитного скоринга;
— Провести сравнительный анализ эффективности моделей и разработать рекомендации по выбору оптимальной модели;
— Изучить метод Вангенгейма–Гирса и разработать экспертную систему, реализующую классификацию форм атмосферной циркуляции в автоматическом режиме;
— Оценить работоспособность и точность экспертной системы на реальных метеорологических данных.

Практическая значимость работы заключается в разработке инструментов для двух прикладных задач: системы кредитного скоринга на основе пяти алгоритмов машинного обучения и экспертной системы для автоматической классификации типов атмосферной циркуляции по данным реанализа ERA5.

Объект исследования — методы классификации: экспертные системы, статистические методы и алгоритмы машинного обучения.

Предмет исследования — применение логистической регрессии, дерева решений, случайного леса, градиентного бустинга и нейронных сетей в задаче кредитного скоринга, а также разработка продукционной экспертной системы для классификации форм атмосферной циркуляции по методу Вангенгейма–Гирса.

Основное содержание работы. Первый раздел

Первый раздел формирует теоретическую базу, необходимую для последующего анализа и практической реализации. Рассматриваются три парадигмы, в рамках которых могут решаться задачи классификации: экспертные системы, статистические методы и методы машинного обучения.

Задача классификации формулируется следующим образом: имеется пространство объектов X и конечное множество классов $Y = \{1, 2, \ldots, K\}$; требуется по обучающей выборке $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, где $x_i \in X$, $y_i \in Y$, построить решающую функцию $f : X \to Y$, минимизирующую ожидаемую ошибку классификации на новых объектах. При $K = 2$ задача называется бинарной классификацией, при $K > 2$ — многоклассовой.

Экспертные системы (ЭС) воспроизводят процесс рассуждения специалиста в узкой предметной области. Архитектура ЭС включает базу знаний (продукционные правила вида «ЕСЛИ <условие> ТО <заключение>»), рабочую память и механизм вывода. Достоинства ЭС — прозрачность решений и способность работать с малыми выборками; недостатки — трудоёмкость формализации знаний и неспособность к самообучению.

Статистические методы основаны на байесовском подходе. Оптимальный байесовский классификатор имеет вид:

$$f^*(x) = \arg\max_k P(Y = k \mid X = x) = \arg\max_k p(x \mid Y = k) \cdot P(Y = k).$$

Линейный дискриминантный анализ (LDA) предполагает нормальность распределений в классах с общей матрицей ковариаций Σ:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln P(Y = k)$$

Также рассмотрено шесть алгоритмов машинного обучения. Логисти-

ческая регрессия является линейным классификатором с хорошей интерпретируемостью коэффициентов.

Метод опорных векторов (SVM) ищет разделяющую гиперплоскость с максимальным зазором, допуская нелинейные границы через ядровой трюк.

Дерево решений рекурсивно разбивает пространство признаков, используя критерии примеси Джини или энтропию.

Случайный лес объединяет B деревьев с двойной рандомизацией, формируя итоговое предсказание голосованием. Градиентный бустинг строит ансамбль последовательно.

Рассмотрены метрики: точность (precision), полнота (recall), F-мера и AUC-ROC. Методологически корректная оценка предполагает стратифицированную кросс-валидацию с раздельной настройкой гиперпараметров и финальной оценкой на отложенной тестовой выборке.

Логистическая регрессия и LDA оптимальны при линейной разделимости классов. SVM эффективен в пространствах высокой размерности. Ансамблевые методы устраняют нестабильность деревьев за счёт усреднения. Нейронные сети превосходят остальные методы при больших объёмах данных, однако наименее интерпретируемы.

Второй раздел

Второй раздел посвящён практическому применению пяти алгоритмов классификации к задаче прогнозирования дефолта по кредиту. Описывается набор данных, проводится разведочный анализ, выполняется обучение и сравнение моделей, разрабатываются практические рекомендации.

Использован набор данных о кредитных запросах, содержащий 32 581 запись с 12 переменными. Целевая переменная — loan_status (0 — нет дефолта, 1 — дефолт). Выявлены некорректные значения (возраст 144 года, стаж 123 года), устранённые фильтрацией диапазонов. Пропущенные значения заполнены медианой (числовые признаки) и модой (категориальные). Категориальные столбцы преобразованы методом Label Encoding. Из-за значительной правосторонней асимметрии к признаку дохода применено логарифмическое преобразование. Структура набора данных представлена на рисунке 1.

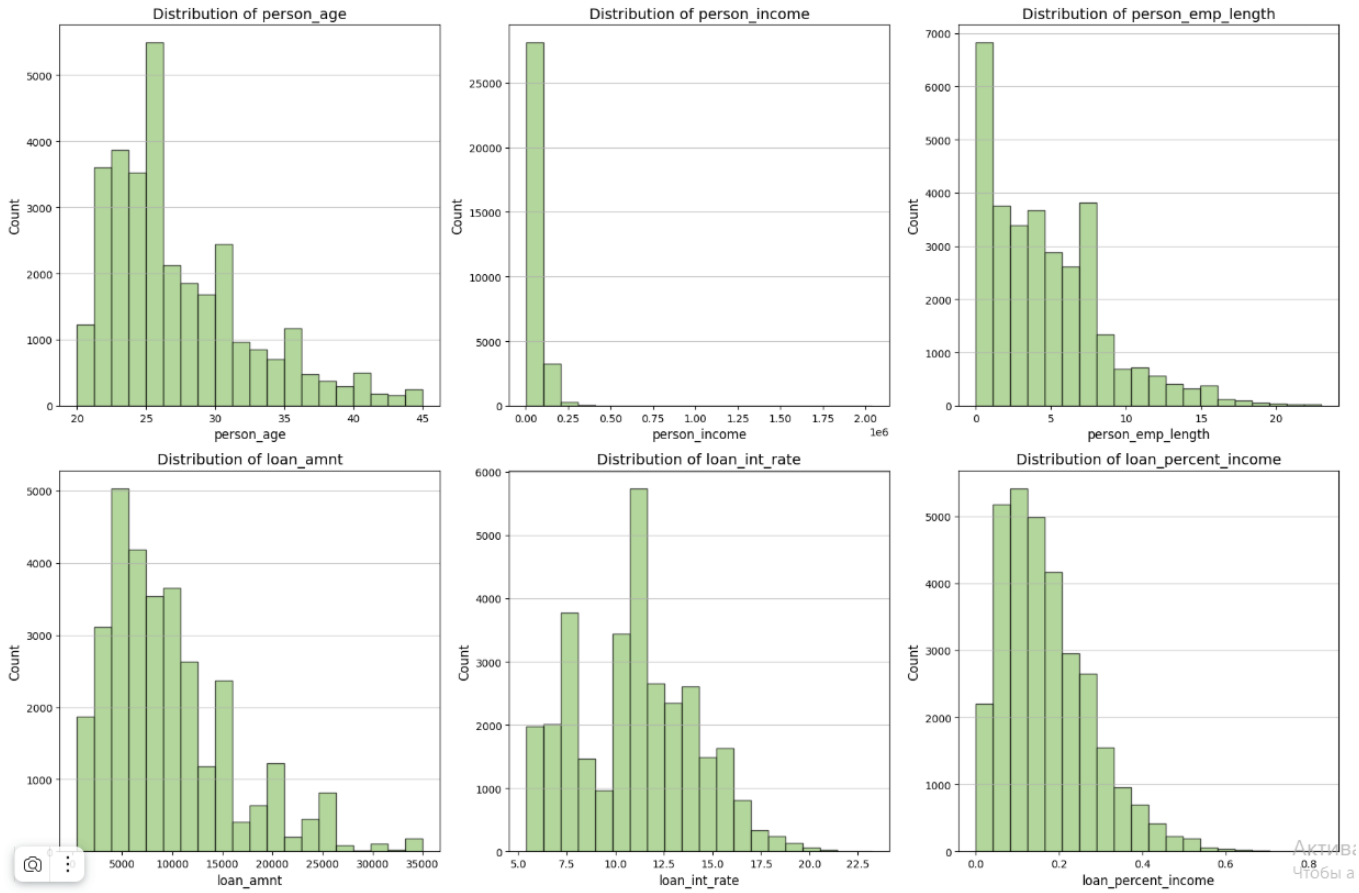

Рисунок 1 – Сетка гистограмм

Разведочный анализ показал, что большинство заёмщиков — молодые люди 22–28 лет с небольшим стажем и умеренными доходами. Обнаружена сильная положительная корреляция между длиной кредитной истории и возрастом, а также между процентной ставкой и кредитным рейтингом. Для снижения влияния мультиколлинеарности на логистическую регрессию применена L2-регуляризация (Ridge). Целевая переменная несбалансирована; для устранения дисбаланса применён метод SMOTE исключительно на обучающей выборке, что позволило достичь равного соотношения классов (50/50) при сохранении исходного распределения в тестовой части. Диаграмма распределения после применения метода балансировки представлена на рисунке 2.

Нейронная сеть реализована средствами TensorFlow/Keras. Архитектура: три скрытых слоя $64 \to 32 \to 16$ нейронов с активацией ReLU, пакетной нормализацией и слоями Dropout; выходной слой — сигмоида. Суммарное число параметров: 3 841, из которых 3 649 обучаемых. Для предотвращения

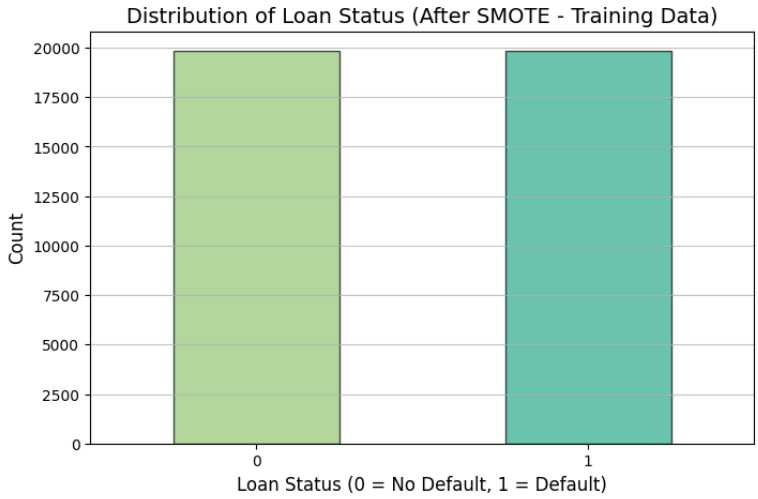

Рисунок 2 – Диаграмма распределения после применения SMOTE

переобучения применены L2-регуляризация и ранняя остановка (patience = 15) с восстановлением весов лучшей эпохи. Обучение завершилось на 58-й эпохе (лучшая — 43-я). Результат получившейся архитектуры представлен на рисунке 3.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	832
batch_normalization (BatchNormalization)	(None, 64)	256
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2,080
batch_normalization_1 (BatchNormalization)	(None, 32)	128
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 16)	528
dropout_2 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 1)	17

Total params: 3,841 (15.00 KB)

Trainable params: 3,649 (14.25 KB)

Non-trainable params: 192 (768.00 B)

Рисунок 3 – Архитектура нейронной сети

Итоговые показатели всех пяти моделей представлены на рисунке 4.

По совокупности метрик выделяются три группы. Случайный лес демонстрирует наилучшие значения F1, Precision и Recall при тестовой точности 0.906. Вторую группу образуют нейронная сеть и градиентный бустинг с

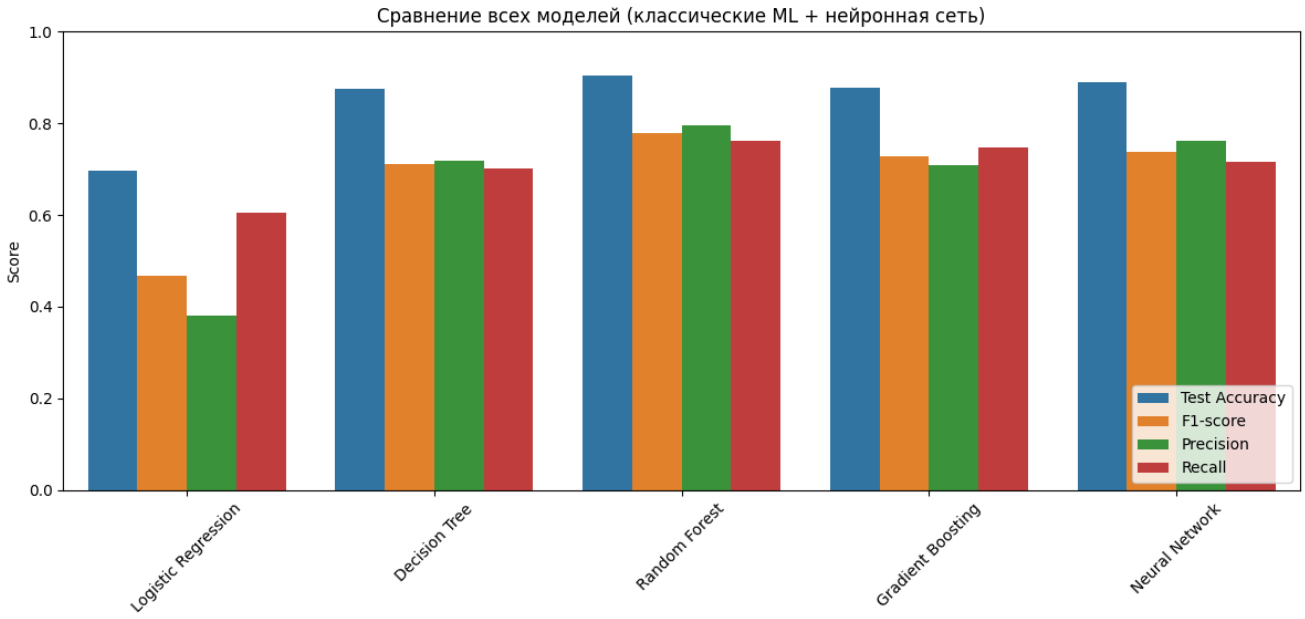

Рисунок 4 – Сравнительная гистограмма для всех моделей

минимальным разрывом train/test. Дерево решений и логистическая регрессия уступают ансамблевым методам по точности, однако сохраняют ценность ввиду интерпретируемости.

Практические рекомендации: случайный лес рекомендуется в качестве основной производственной модели; градиентный бустинг — в сценариях, где критична стабильность оценок риска между циклами обновления; нейронная сеть — как дополнительный оценщик при больших объёмах данных (требует методов объяснимости SHAP для регуляторного соответствия). Пороговое значение следует устанавливать по кривой precision-recall с учётом асимметрии стоимостей ошибок. Переобучение моделей рекомендуется проводить не реже одного раза в квартал.

Третий раздел

Третий раздел посвящён разработке и реализации экспертной системы для автоматической классификации форм атмосферной циркуляции по методу Вангенгейма–Гирса на основе данных реанализа ERA5.

Метод классификации, разработанный Г. Я. Вангенгеймом в 1930-е годы и усовершенствованный А. А. Гирсом в 1960–70-е, выделяет три обобщённые формы атмосферной циркуляции над Евро-Атлантическим сектором. Физическое основание — структура поля геопотенциала на уровне 500 гПа; ключевая диагностическая изолиния — 536 геопотенциальных дам (дам). Ка-

талог форм циркуляции непрерывно ведётся с 1899 года.

Три типа циркуляции: тип W (западный) — субширотное положение изолинии без выраженных меридиональных отклонений; тип E (восточный) — ложбина над Атлантикой, гребень над материком, суровые зимы в Европе; тип C (меридиональный) — гребень над Атлантикой, ложбина над материком, активные вторжения арктических масс. Схематичное изображение трёх форм представлено на рисунке 7.

Использованы данные реанализа ERA5 (Copernicus Climate Data Store, ECMWF): геопотенциал на уровне 500 гПа, январь 1940–1962 гг., 0:00 UTC, итого 713 суточных полей. Центральный элемент системы — три эталонных контура, задающих «идеальную» конфигурацию изолинии 536 дам для каждого типа. Координаты опорных точек (20–30 точек на тип) были вручную определены в Excel на основе анализа синоптических закономерностей и экспортированы в CSV-файлы. Критерии: для W — субгоризонтальное положение; для E — отрицательная разность средних широт западной и восточной частей сектора; для C — положительная разность аналогичных зон. Между опорными точками применяется линейная интерполяция. Эталонные изолинии представлены на рисунке 5.

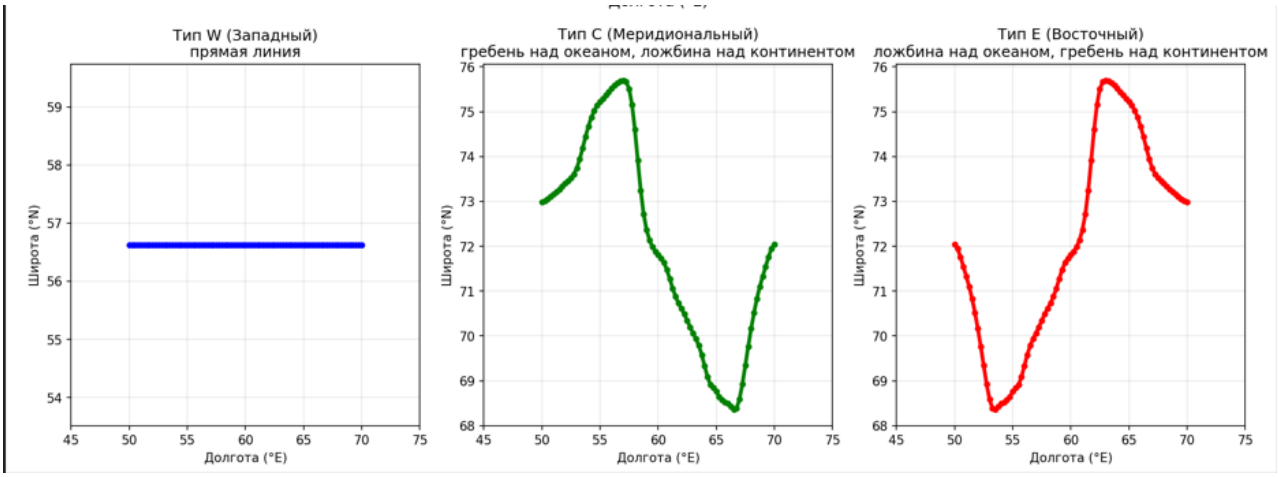

Рисунок 5 – Эталонные изолинии для трёх типов циркуляции

Алгоритм классификации реализует пятиэтапный конвейер обработки: загрузка и конвертация данных; выделение изолинии 536 дам; вычисление признаков — амплитуда, угол наклона, разность средних широт в зонах, положение гребня и ложбины; сравнение с эталонами по взвешенной мере сход-

ства; присвоение типа или метки «U».

Мера сходства вычисляется по формуле:

$$S(T) = 0.5 \cdot \text{distance\_score} + 0.2 \cdot \text{correlation} + 0.2 \cdot \text{slope\_score} + 0.1 \cdot \text{bend\_score},$$

где distance_score — нормированное отклонение широт (допуск 12°), correlation — коэффициент Пирсона между интерполированными координатами, slope_score и bend_score — оценки сходства наклона и изгиба изолиний соответственно.

Алгоритм обработал 713 суточных бланков. Западный тип W закономерно преобладает в январе — зимний атлантический штормовой трек наиболее активен именно в этот период. Типы E и C встречаются реже и связаны с блокирующими ситуациями, что согласуется с известными климатологическими оценками. Результат работы программы представлен на рисунке 6.

Дата	Точек	Тип	Лучш.	Оценка	W	C	E
1940-01-01	340	W	W	0.519	0.519	-0.098	0.383
1940-01-02	242	W	W	0.393	0.393	-0.109	0.366
1940-01-03	538	U	E	0.238	0.173	-0.172	0.238
1940-01-04	215	W	W	0.359	0.359	-0.036	0.169
1940-01-05	308	E	E	0.444	0.241	-0.126	0.444
1940-01-06	315	E	E	0.362	0.159	-0.129	0.362
1940-01-07	264	U	W	0.277	0.277	-0.075	0.187
1940-01-08	234	E	E	0.340	0.185	-0.116	0.340
1940-01-09	339	U	C	0.107	0.064	0.107	-0.044

Рисунок 6 – Результат работы программы

Главное преимущество программы — полная интерпретируемость: каждое решение прослеживается до физически осмысленных компонентов меры сходства. Система не требует размеченного обучающего набора, что делает её применимой к архивным данным любой длины. Ограничения: качество классификации зависит от точности эталонов; веса компонентов подобраны эвристически; атипичные ситуации не дифференцируются за пределами трёх категорий. Перспективным направлением является гибридный подход — использование вычисленных признаков как входа для обучаемого классифика-

тора (SVM или случайный лес) на малом размеченном подмножестве исторического каталога.

Заключение

В рамках данной работы проведён сравнительный анализ подходов к решению задач классификации и исследована их практическая применимость в двух принципиально разных предметных областях — банковском кредитном скоринге и классификации форм атмосферной циркуляции. Было показано, что универсального метода классификации, одинаково эффективного для любых задач, не существует: каждый подход обладает собственными преимуществами и ограничениями, проявляющимися в зависимости от структуры данных, объёма выборки и требований к интерпретируемости.

Основные результаты работы:

— Изучены теоретические основы шести методов классификации;

— На данных кредитного скоринга обучены и оптимизированы пять моделей;

— Разработана экспертная система для классификации форм атмосферной циркуляции по методу Вангенгейма–Гирса;

— Показано, что методы машинного обучения предпочтительны при больших объёмах данных и сложных зависимостях, тогда как экспертные системы более эффективны при хорошей формализуемости знаний и ограниченных данных.

Проведённое исследование подтвердило: методы классификации не конкурируют друг с другом в абсолютном смысле, а взаимно дополняют друг друга, обеспечивая наиболее эффективное решение при грамотном выборе метода под конкретную задачу.