НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
им. В.А. ТРАПЕЗНИКОВА РАН

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)

МАТЕРИАЛЫ
XX Международной конференции
имени А. Ф. Терпугова
1–5 декабря 2021 г.

ТОМСК
Издательство Томского
государственного университета
2022

УДК 519
ББК 22.17
И74
Информационные технологии и математическое моделирование (ИТММ-2021): Материалы XX Международной конференции имени А. Ф. Терпугова (1–5 декабря 2021 г.). — Томск: Издательство Томского государственного университета, 2022. — 391 с.

ISBN 978–5–907572–20–1

Сборник содержит избранные материалы XX Международной конференции имени А.Ф. Терпугова по следующим направлениям: теория массового обслуживания и ее приложения, интеллектуальный анализ данных и визуализация, информационные технологии и программная инженерия, математическое и компьютерное моделирование технологических процессов.

Для специалистов в области информационных технологий и математического моделирования.

УДК 519
ББК 22.17

Р е д к о л л е г и я:
А.А. Назаров, доктор технических наук, профессор
С.П. Моисеева, доктор физико-математических наук, профессор
А.Н. Моисеев, доктор физико-математических наук, доцент

Конференция проведена при поддержке
международного научно-методического центра
Томского государственного университета по математике,
информатике и цифровым технологиям в рамках
федерального проекта «Кадры для цифровой экономики»
национальной программы
«Цифровая экономика в Российской Федерации»

ISBN 978–5–907572–20–1

© Авторы. Текст, 2022
© Томский государственный университет. Оформление. Дизайн, 2022

NATIONAL RESEARCH TOMSK STATE UNIVERSITY

PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA

V.A. TRAPEZNIKOV INSTITUTE OF CONTROL
SCIENCES OF RUSSIAN ACADEMY OF SCIENCES

INFORMATIONAL TECHNOLOGIES AND MATHEMATICAL MODELLING (ITMM-2021)

PROCEEDINGS
of the 20th International Conference
named after A. F. Terpugov
2021 December, 1–5

TOMSK
Tomsk State
University Publishing
2022

UDC 519
LBC 22.17
I60
Informational technologies and mathematical modelling (ITMM-2021): Proceedings of the 20th International Conference named after A. F. Terpugov (2021 December, 1–5). — Tomsk: Tomsk State University Publishing, 2021. — 391 p.

ISBN 978–5–907572–20–1

This volume presents selected papers from the XIX International Conference named after A.F. Terpugov. The papers are devoted to new results in the following areas: queuing theory and its applications, data mining and visualization, information technology and software engineering, mathematical and computer modeling of technological processes.

UDC 519
LBC 22.17

E d i t o r s:
A.A. Nazarov, Doctor of Technical Sciences, Professor,
S.P. Moiseeva, Doctor of Physical and Mathematical Sciences, Professor,
A.N. Moiseev, Doctor of Physical and Mathematical Sciences, Associate Professor.

The conference was supported by
International Computer Science
Continues Professional Development Center
of the Federal project "Human Resources for the Digital Economy"
of the National program
"Digital Economy of the Russian Federation"

ISBN 978–5–907572–20–1

© Authors. Text, 2022
© Tomsk State University
Publishing. Design, 2022

Методы анализа и визуализации данных

ПРОГНОЗИРОВАНИЕ ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ДАННЫХ, ОТОБРАННЫХ С ПОМОЩЬЮ РАСЧЕТА КОЭФФИЦИЕНТОВ WOE И IV

Д. Д. Бугакова, Е. Ю. Лисовская

Национальный исследовательский
Томский государственный университет, г. Томск, Россия

В данной работе рассматриваются основные методы машинного обучения для прогнозирования платежеспособности клиентов банка и методы оценивания качества их работы на предоставленных данных. Для прогнозирования целевого признака в данной работе будут рассмотрены такие методы как: логистическая регрессия, случайный лес, метод ближайших соседей и метод опорных векторов. Для сравнения работы методов будут применены метрики accuracy, recall, precision, F_1, AUC ROC, AUC PR, индекс Джини. На основе значений метрик сделан вывод о том, что рассматриваемые методы примерно одинаково хорошо работают.

Ключевые слова: *Логистическая регрессия, случайный лес, метод опорных векторов, метод ближайших соседей, метрики качества.*

Введение

В наше время область машинного обучения набирает большую популярность в сфере бизнеса, финансов, сфере услуг, промышленности. В частности, в банковском деле для прогнозирования платежеспособности клиента. Своевременный анализ кредитного потенциала заемщика поможет предотвратить невозврат кредитных средств и избежать банкротства банковской организации.

1. Обзор исследуемых данных

Для того, чтобы на практике посмотреть работу алгоритмов воспользуемся набором данных с сайта kaggle.com [1] о клиентах банка «Тинькофф», для которого предлагается по данным из анкеты с использованием алгоритмов машинного обучения предсказать факт наличия дефолта. Набор данных содержит информацию о 205296 клиентах и 17

признаках, 5 количественных (возраст, скоринговый балл, количество обращений в банк, доход, количество отказанных заявок), 7 категориальных (количество связей с другими клиентами, уровень образования, пол, количество лет, которое заявитель является клиентом банка, тип жилплощади, должность, регион), 4 бинарных признаках (наличие автомобиля, иномарки, дохода выше среднего, загранпаспорта) и 1 целевого признака (наличие дефолта у клиента).

2. Отбор признаков

Перед обучением модели на данных предварительно их нужно обработать. Обработка данных состоит из двух этапов: первичная обработка данных и отбор информативных признаков. В данной работе отбор признаков производился с помощью расчета коэффициентов WoE (Weight of Evidence) для признаков с последующей оценкой предсказательной силы отобранных факторов с помощью расчета коэффициента IV (information value) [4]. Подробное описание предварительной обработки данных выходит за рамки данной работы. Результатом применения предложенных методов являются, следующие отобранные признаки: Жилплощадь: студия, Жилплощадь: дом, Должность: начальник, Балл: [-2.387; -2.116], Балл: [-2.116; -1.865], Балл: [-1.865; -1.566], Балл: больше -1.566, Связь с клиентами: 2, Связь с клиентами: 3, Связь с клиентами: более 3.

3. Подготовка данных для обучения моделей

Будем условно называть клиентов «плохими», если значение целевого признака – дефолт, иначе «хорошими». Для построения моделей сначала нужно разделить выборку на тренировочную, на которой модель будет обучаться и тестовую, на которой мы будем проверять качество моделей, в отношении 80/20 (Таблица 1).

Таблица 1

Распределение данных в выборках

Тип выборки	Количество «хороших» клиентов	Количество «плохих» клиентов	Доля «плохих» клиентов
Исходная	144631	18748	11,48 %
Тренировочная	115652	15051	11,52 %
Тестовая	28979	3697	11,31 %

4. Метрики для оценки качества модели

Работа модели может быть охарактеризована с помощью таких критериев качества как: ошибки первого и второго рода, *accuracy*, *recall*, *precision*, F_1, *AUC ROC*, *AUC PR*, индекс Джини. Перед началом рассмотрения метрик введем важное понятие матрицы ошибок (Рис. 1).

Рис. 1. Матрица ошибок

Метрика *accuracy* общая для всех классов и не применима в задачах с несбалансированной выборкой, как и в рассматриваемой задаче [3]

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}.$$

Для правильной оценки качества работы алгоритмов нужно использовать метрики *recall*, *precision* [3]:

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN}.$$

Recall, показывает какая доля объектов, положительного класса предсказала модель из всех объектов положительного класса. *Precision* показывает какая доля объектов, которую модель предсказала как положительную действительно является положительной.

Также при обучении модели существуют ошибки I-го и II-го рода *False Positive* и *False Negative*. В рассматриваемой задаче ошибку I-го рода можно интерпретировать как коммерческий риск, связанный с отказом кредитоспособным клиентам. Ошибка II-го рода характеризует кредитный риск, связанный с количеством некредитоспособных клиентов, классифицированных как кредитоспособных. Если *recall* и *precision* являются одинаково значимыми для задачи, используется F_1-мера (среднее гармоническое двух метрик *recall* и *precision*) [3]:

$$F_1 = \frac{2 * precision * recall}{precision + recall}.$$

ROC-кривая – график, показывающий зависимость между верно классифицируемыми объектами положительного класса (TPR) и ложно положительно классифицируемыми объектами негативного класса (FPR) [2]

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}.$$

Метрика $ROC\ AUC$ (Area Under Curve) измеряет площадь под кривой ROC (Рис. 2), чем сильнее крутизна ROC-кривой, тем больше площадь под ней и тем лучше работает модель [2].

На основе метрики $ROC\ AUC$ можно вычислить другую метрику – индекс Джини

$$Gini = 2 * (ROC\ AUC - 0.5),$$

чем выше индекс Джини, тем лучше дискриминирующая способность модели.

PR-кривая – график, построенный в координатах *recall* и *precision*. Площадь под PR-кривой ($AUCPR$) лучше использовать для задач с несбалансированной выборкой (Рис. 3).

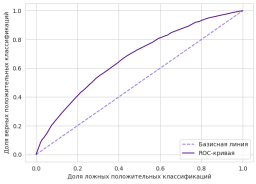

Рис. 2. График ROC-кривой

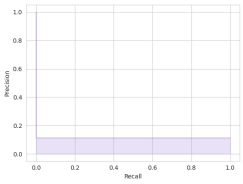

Рис. 3. График PR-кривой

5. Построение моделей

Построим базовые модели для всех алгоритмов. Для удобства обозначений пронумеруем модели. Модель 1 – Логистическая регрессия, Модель 2 – Метод ближайших соседей, Модель 3 – Случайный лес, Модель 4 – Метод опорных векторов. Модели были реализованы с помощью библиотек Python (LogisticRegression, KNeighborsClassifier, RandomForestClassifier, SVC). Результаты работы базовых моделей показывают, что модели не особо сильно отличаются друг от друга своей предсказательной способностью, поэтому для каждой модели нужно подобрать параметры, которые будут улучшать их (Таблица 2). Также для

моделей 1, 3 и 4 был применен метод балансировки [3, 2]. Коэффициент регуляризации логистической регрессии получился слишком большой, модель могла переобучиться, поэтому нужно проверить с помощью кросс-валидации на 10 фолдах. После проведения кросс-валидации, выяснилось, что сильный коэффициент регуляризации почти никак не повлиял на предсказательную способность модели, значения метрик изменились совсем немного.

Таблица 2

Подобранные параметры

Модель	Гиперпараметр	Значение
Модель 1	Коэффициент регуляризации	0,0064281
Модель 2	Количество соседей	3
Модель 3	Число деревьев	267
	Число признаков	log2
	Глубина дерева	1200
	Число объектов в листьях	4
Модель 4	Коэффициент регуляризации	1
	Гамма	1
	Вид ядра	гауссово ядро

Подобранные гиперпараметры и применение метода балансировки к некоторым алгоритмам значительно улучшили предсказательную способность моделей (Рис. 4, Таблица 3).

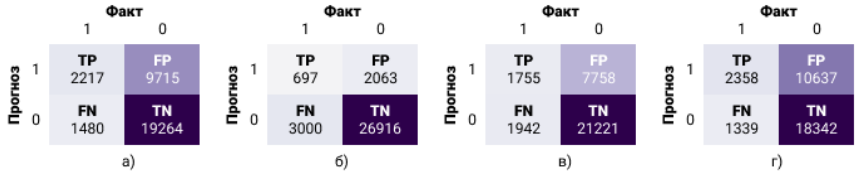

Рис. 4. Матрица ошибок моделей с подобранными гиперпараметрами: а) Модель 1, б) Модель 2, в) Модель 3, г) Модель 4

Таблица 3

Значения метрик базовых моделей после подбора гипперпараметров

Метрика	Модель 1	Модель 2	Модель 3	Модель 4
precision	0,168052	0,125000	0,174900	0,174773
recall	0,667027	0,000541	0,602380	0,602651
F_1	0,268467	0,001077	0,271089	0,270964
AUC PR	0,149768	0,113148	0,1503433	-

Заключение

Применение той или иной модели зависит от конкретной задачи. В задачах прогноза нелинейные модели, такие как случайный лес и метод опорных векторов показывают лучшие результаты.

Рассмотренные в работе алгоритмы показали хорошую предсказательную способность для использования в задачах прогнозирования платежеспособности клиентов банка. В данной работе по метрике F_1 лучше всего себя проявила модель случайного леса. Несмотря на достоинства метода ближайших соседей (простая реализация, хорошая интерпретация, настраивание гиперпараметра), он показал не очень хорошие результаты на данных, по сравнению с другими методами, вероятно из-за проблемы несбалансированности данных.

СПИСОК ЛИТЕРАТУРЫ

1. `https://www.kaggle.com/c/fintech-credit-scoring.` — Kaggle. 2021.
2. *Замятин А. В.* Интеллектуальный анализ данных. Томск: Издательский Дом Томского государственного университета, 2020. 119 с.
3. *Миркин Б. Г.* Введение в анализ данных: учебник и практикум. Москва: Издательство Юрайт университета, 2019. 174 с.
4. *Шунина Ю. С.* Критерии качества работы классификаторов. // Вестник Ульяновского государственного технического университета. 2015. № 2. С. 67–70.

Бугакова Дарья Дмитриевна — студент института прикладной математики и компьютерных наук. E-mail: *bugashka17@inbox.ru*

Лисовская Екатерина Юрьевна — к.ф.-м.н., доцент кафедры теории вероятностей и математической статистики института прикладной математики и компьютерных наук. E-mail: *ekaterina_lisovs@mail.ru*

СОДЕРЖАНИЕ

Информационные технологии и программная инженерия .. 5
Гилин С. В. Задача автоматического распознавания зданий в водоохранных зонах на спутниковых снимках 6
Зоркин А. С., Змеев Д. О. Гибридный алгоритм поиска академического плагиата исходного кода с использованием парсера ANTLR . 13
Саринова А. Ж., Дунаев П. А., Бекбаева А. М. Дискретно-косинусное преобразование для сжатия гиперспектральных изображений в фитосанитарном контроле зерновых культур 19
Шарапов С. Ф. Обзор способов разработки клиентских веб-приложений и преимущества использования Генератора Статичных Сайтов 24
Моделирование телекоммуникационных сетей связи 29
Ashurmetova N., Sopin E. Response time analysis in fog computing system with threshold-based offloading mechanism 30
Ivanova N. M., Vishnevsky V. M. Applications of k-out-of-n:G system and machine learning methods on reliability analysis of unmanned high-altitude module 36
Копать Д. Я. Асимптотический анализ G-сети с ненадёжными многолинейными системами обслуживания 42
Методы анализа и визуализации данных 49
Бугакова Д. Д., Лисовская Е. Ю. Прогнозирование платежеспособности клиентов банка с использованием методов машинного обучения на данных, отобранных с помощью расчета коэффициентов WoE и IV 50
Бугакова Д. Д., Лисовская Е. Ю., Баймеева Г. В. Основные этапы обработки и методы отбора признаков для дальнейшего прогнозирования платежеспособности клиентов банка 56
Математическая теория телетрафика и теория массового обслуживания 63
Anilkumar M. P., Jose K. P. Discrete Time Queue with Self Interruption Resulting Reduced Priority 64
Kuki A., Bérczes T., Sztrik J. Modeling Two-Way Communication Systems with Catastrophic Breakdowns 70
Bulinskaya E. Limit behavior and stability of applied probability systems ... 76
Morozov E., Rogozin S. Stability analysis of classical retrials: a revised regenerative proof 82

Научное издание

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
(ИТММ-2021)

МАТЕРИАЛЫ
XX Международной конференции
имени А. Ф. Терпугова
1–5 декабря 2021 г.

Редактор *В.Г. Лихачева*
Компьютерная верстка *Д.В. Семенова, Е.Ю. Лисовская, О.Д. Лизюра*
Дизайн обложки *Л.Д. Кривцовой*

Отпечатано на оборудовании
Издательства Томского государственного университета
634050, г. Томск, пр. Ленина, 36.
Тел. 8+(382-2)–52-98-49
Сайт: http://publish.tsu.ru
E-mail: rio.tsu@mail.ru

Подписано к печати 12.09.2022 г.
Формат $60 \times 84^1/_{16}$. Бумага для офисной техники. Гарнитура «Times».
Печ. л. 24.5. Усл. печ. л. 22.7. Тираж 500 экз. Заказ № 5145.

ISBN 978-5-907572-20-1

elibrary.ru ПРОГНОЗИРОВАНИЕ ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННО... ★ 918 отзывов

НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА
eLIBRARY.RU

вход

IP-адрес компьютера:
92.63.69.5
Название организации:
Национальный исследовательский Томский государственный университет
Имя пользователя:
Пароль:
Вход
Запомнить меня
Правила доступа
Регистрация
Забыли пароль?
Вход через Вашу организацию

ИНФОРМАЦИЯ О ПУБЛИКАЦИИ

РОССИЙСКИЙ ИНДЕКС НАУЧНОГО ЦИТИРОВАНИЯ
Science Index

eLIBRARY ID: 49546400 EDN: FXPCCY

ПРОГНОЗИРОВАНИЕ ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ДАННЫХ, ОТОБРАННЫХ С ПОМОЩЬЮ РАСЧЕТА КОЭФФИЦИЕНТОВ WOE И IV

БУГАКОВА Д. Д.[1], ЛИСОВСКАЯ Е. Ю.[1]

[1] Национальный исследовательский Томский государственный университет

Тип: статья в сборнике трудов конференции Язык: русский Год издания: 2022
Страницы: 50-55

ИСТОЧНИК:

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)
материалы XX Международной конференции имени А. Ф. Терпугова. Томск, 2022
Издательство: Национальный исследовательский Томский государственный университет

КОНФЕРЕНЦИЯ:

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)
Томск, 01–05 декабря 2021 года
Организаторы: Национальный исследовательский Томский государственный университет

КЛЮЧЕВЫЕ СЛОВА:

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, СЛУЧАЙНЫЙ ЛЕС, МЕТОД ОПОРНЫХ ВЕКТОРОВ, МЕТОД БЛИЖАЙШИХ СОСЕДЕЙ, МЕТРИКИ КАЧЕСТВА

АННОТАЦИЯ:

В данной работе рассматриваются основные методы машинного обучения для прогнозирования платежеспособности клиентов банка и методы оценивания качества их работы на предоставленных данных. Для прогнозирования целевого признака в данной работе будут рассмотрены такие методы как: логистическая регрессия, случайный лес, метод ближайших соседей и метод опорных векторов. Для сравнения работы методов будут применены метрики accuracy, recall, precision, F1, AUC ROC, AUC PR, индекс Джини. На основе значений метрик сделан вывод о том, что рассматриваемые методы примерно одинаково хорошо работают.

БИБЛИОМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ:

Входит в РИНЦ®: да
Входит в ядро РИНЦ®: нет
Норм. цитируемость по направлению:
Цитирований в РИНЦ®: 0
Цитирований из ядра РИНЦ®: 0
Дециль в рейтинге по направлению:

ИНСТРУМЕНТЫ

Содержание сборника
Следующая публикация
Предыдущая публикация

Загрузить:
Полный текст (PDF)
Отправить публикацию по электронной почте

Добавить публикацию в подборку
Новая подборка

Редактировать Вашу заметку к публикации
Обсудить эту публикацию с другими читателями

Показать все публикации этих авторов
Найти близкие по тематике публикации