УДК 004.855.5

Оценка кредитоспособности заёмщика при помощи методов машинного обучения

Р.Р. Янбеков

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, пр. Большевиков, 22, корп. 1, Санкт-Петербург, 193232, Россия

E-mail: yanbekov14@mail.ru

Аннотация. Построение моделей машинного обучения является на сегодняшний день одной из самых популярных и современных областей человеческой деятельности на стыке информационных технологий, математического анализа и статистики. В связи с быстрым развитием информационных технологий и ростом конкуренции на рынке кредитных услуг, построение автоматизированной модели оценки надёжности заемщика является актуальной задачей. В статье представлен сравнительный анализ четырех моделей машинного обучения для решения данной проблемы.

Ключевые слова: кредитоспособность, машинное обучение, кредитный скоринг, бинарная классификация, искусственный интеллект

Assessing the creditworthiness of a borrower using machine learning methods

R.R. Yanbekov

The Bonch-Bruevich Saint Petersburg State University of Telecommunications, 22 Bolsheviks pr., St. Petersburg, 193232, Russia

E-mail: yanbekov14@mail.ru

Abstract. The construction of machine learning models is today one of the most popular and modern areas of human activity at the intersection of information technology, mathematical analysis and statistics. Due to the rapid development of information technology and the growth of competition in the credit services market, the construction of an automated model for assessing the reliability of a borrower is an urgent task. The article presents a comparative analysis of four machine learning models to solve this problem.

Keywords: creditworthiness, machine learning, credit scoring, binary classification, artificial intelligence

ISBN 978-5-6045165-2-2

1. Введение

Одним из важнейших факторов при принятии банком решения о выдаче кредита и условиях кредитного соглашения является оценка кредитоспособности потенциального заемщика, т. е. его способности полностью исполнить взятые на себя обязательства, своевременно погасив кредит и начисленные проценты.

Оценка кредитоспособности заемщика является комплексным целенаправленным процессом, осуществляемым с учетом анализа многих параметров субъекта, качественных и количественных показателей. Многие банки при оценке надежности заемщиков используют скоринговые системы, которые дают возможность быстро принять решение о возможности предоставления кредита. Они позволяют оценить кредитоспособность заемщика, основываясь на статистических методах [1].

Ежедневно банки получают тысячи заявок на получение кредитов и кредитных карт. Ручная обработка таких заявок может привести к человеческим ошибкам и занимает сравнительно много времени. Методы машинного обучения позволяют в значительной степени автоматизировать эти процессы.

2. Постановка задачи (Цель исследования)

Рассматриваемая задача прогнозирования надежности заемщика представляет собой задачу классификации. Классификация — задача разделения множества наблюдений или объектов на группы, называемые классами, на основе анализа их формального описания. При классификации каждая единица наблюдения относится определенной группе или номинальной категории на основе некоторого качественного свойства.

Целью работы является построение модели, которая наиболее точно классифицирует заемщиков на платежеспособных или ненадежных, имея лишь сведения об 11 его признаках. В работе рассмотрены такие алгоритмы машинного обучения как логистическая регрессия, решающие деревья, случайный лес и метод опорных векторов.

2.1. Исходные данные

Имеется набор данных, который содержит информацию о 614 заемщиках. Каждый клиент описывается 12 признаками такими как: семейный статус, пол, количество детей, наличие высшего образования, доход и другие. Целевым признаком является статус платежеспособности клиента. Обучение будет осуществляться на основе 491 заемщиках, далее модель будет тестироваться на оставшихся 123 клиентах, не имея данных об их статусе.

Прогноз будет осуществляться на основе 12 признаков, описывающих клиента [2]:

- Gender – Пол бинарная переменная мужской или женский;

ISBN 978-5-6045165-2-2

- Married – Семейное положение бинарная переменная в браке или нет;
- Dependents – Количество детей
- Education – Наличие высшего образования бинарная переменная есть или нет;
- Self_Employed – Вид занятости бинарная переменная самозанятый или нет;
- ApplicantIncome – Доход заемщика;
- CoapplicantIncome – Доход супруга клиента;
- Loan_Amount – Сумма займа;
- Loan_Amount_Term – Срок займа;
- Credit_History – Кредитная история бинарная переменная удовлетворяет или нет;
- Property_Area – Область проживания номинативная переменная город, посёлок или деревня
- Loan_Status – Статус заёмщика

2.2. Предобработка данных

Имеющиеся пропуски в данных были заменены на моду признака для категориальных переменных и на среднее значение для числовых. Все категориальные данные были преобразованы в числовое значения, где каждому классу соответствовало единственное число. Количественные переменные были стандартизированы, то есть все исходные значения набора данных были приведены к набору значений из распределения с нулевым средним и стандартным отклонением, равным 1.

3. Методы и материалы исследования

Модель строится на основе анализа 4 методов: дерево решений, случайный лес, логистическая регрессия, метод опорных векторов.

Дерево решений — это метод представления решающих правил в иерархической структуре, состоящей из элементов двух типов — узлов и листьев. В узлах находятся решающие правила и производится проверка соответствия примеров этому правилу по какому-либо атрибуту обучающего множества.

Случайный лес — это алгоритм классификации, основанный на принципе использования ансамбля нескольких решающих деревьев для достижения ими большей точности. Классификаторы (решающие деревья) обучаются независимо друг от друга. Затем классификаторы независимо друг от друга делают предсказания о входном элементе, и класс, за который проголосовало больше всего классификаторов, становится предсказанием итогового классификатора.

Логистическая регрессия — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам. Данный алгоритм

ISBN 978-5-6045165-2-2

классификации использует сигмоидную функцию в качестве функции активации и позволяет дать вероятностную оценку принадлежности объекта каждому классу.

Метод опорных векторов — алгоритм ищет точки на графике, которые расположены непосредственно к линии разделения ближе всего. Эти точки называются опорными векторами. Затем, алгоритм вычисляет расстояние между опорными векторами и разделяющей плоскостью. Это расстояние называется зазором. Основная цель алгоритма — максимизировать расстояние зазора. Лучшей гиперплоскостью считается такая гиперплоскость, для которой этот зазор является максимально большим [3].

4. Полученные результаты

Моделирование производилось на платформе IPython с использованием библиотек scikit-learn, numpy и pandas. Для оценки качества модели использовалась точность на тестовой выборке и на кросс-валидации. Точность представляет собой количество верно классифицированных клиентов по отношению к общему числу. Кросс-валидация — это метод формирования, обучающего и тестового множеств для обучения аналитической модели в условиях недостаточности исходных данных или неравномерного представления классов. В основе метода лежит разделение исходного множества данных на k примерно равных блоков, например, k = 5. Затем на k − 1, т.е. на 4-х блоках, производится обучение модели, а 5-й блок используется для тестирования. Для оценки качества кросс-валидация осуществлялась на 5 частях и в итоге бралось усредненное значение [4].

Таблица 1. Результат исследования.

Алгоритм	Точность	Точность на кросс-валидации
Логистическая регрессия	0.84	0.81
Решающие деревья	0.72	0.7
Случайный лес	0.83	0.77
Метод опорных векторов	0.82	0.8

5. Выводы

В результате построения моделей были получены метрики в соответствии с таблицей 1. Все алгоритмы показали высокий результат точности на тестовой выборке, однако наивысшую точность показала модель с логистической регрессией. Из этого следует, что она является наиболее предпочтительной для решения данной задачи.

Таким образом, проведенное исследование показало возможность применения методов машинного обучения на основе накопленных нами данных. В результате чего был выявлен наилучший метод для оценки надежности плательщика.

ISBN 978-5-6045165-2-2

Список литературы

1. Клементьев, В.А. Применение скоринговых систем в банковском кредитовании физических лиц / В.А. Клементьев // Вестник УГТУ. – 2009. – С. 58-62.
2. Kaggle: сайт. – 2018. – URL: https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset (дата обращения: 12.07.2021).
3. Жерон, О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow / О. Жерон. – Спб.: ООО "Альфа-книга", 2018. – 688 с.
4. Вьюгин, В.В. Математические основы теории машинного обучения и прогнозирования / В.В Вьюгин. – Москва: 2013. – 387 с.

ISBN 978-5-6045165-2-2