ОЦЕНКА ВЛИЯНИЯ АНКЕТНЫХ ДАННЫХ ЗАЕМЩИКА ПОТРЕБИТЕЛЬСКОГО КРЕДИТА НА ДЕФОЛТ

Губин Е.И.[1], Таячинов Д.А.[2]
[1]Томский политехнический университет, к.ф.- м.н., доцент
[2]Томский политехнический университет, ИШИТР, студент группы 8К22, e-mail: dat59@tpu.ru

Аннотация

Потребительский кредит является наиболее популярным банковским продуктом, но и наиболее рискованным в смысле риска дефолта. Поэтому оценка влияния портрета заемщика (его анкетных данных) для оценки платежеспособности клиента, позволяет предсказать вероятность возникновения просрочек по кредитам. В данной работе исследуется задача классификации с использованием метода машинного обучения.

Ключевые слова: кредитный риск, машинное обучение.

Введение

Целью данной работы является построение модели классификации, способной дать оценку влияния каждого признака (атрибута) на возможный дефолт заемщика потребительского кредита. Были использованы методы анализа, изложенные в работе «Статистические методы анализа в биологии и медицине» [1].

Описание алгоритма

Для решения поставленной задачи был применен алгоритм машинного обучения. Исходным набором данных служил датасет, содержащий информацию о 1500 заемщиках и 8 характеристиках анкетных данных. Эти данные включали в себя информацию о поле, типе второго документа, роде профессиональной деятельности, уровне образования, наличии автомобиля, доходе, запрашиваемой сумме кредита, возрасте и количестве дней просрочки.

В процессе анализа данных [2] выявлено, что датасет содержит выбросы, пропущенные значения и признаки, не имеющие значения для модели. Пропущенные числовые значения были заменены медианными, а остальные - наиболее часто встречающимися. Для обработки выбросов применен метод межквартильного размаха (IQR). Строки с просрочкой более 90 дней были помечены как 1, а строки с просрочкой менее 90 дней - как 0.

Поскольку задача сводилась к классификации, для работы выбран метод «случайного леса» из пакета sklearn. ensemble.

RandomForestClassifier, который позволяет оценить важность каждого признака с помощью функции feature_importance [3]. Для прогнозирования случае регрессии мы можем усреднить эти результаты, чтобы получить окончательный прогноз. В классификации применяется стратегия «мягкого голосования». Это означает, что каждый алгоритм выдает «мягкий» прогноз, вычисляя вероятности для каждого класса. Эти вероятности с помощью случайного леса алгоритм сначала выдает прогноз для каждого дерева в лесу. В затем усредняются по всем деревьям, и прогнозируется класс с наибольшей вероятностью. Как и дерево решений, случайный лес позволяет вычислить важности признаков, которые рассчитываются путем агрегирования значений важности по всем деревьям леса. Как правило, важности признаков, вычисленные случайным лесом, являются более надежным показателем, чем важности, вычисленные одним деревом [2].

В таблице 1 показаны (по мере возрастания) признаки (анкетные данные), которые наиболее сильно влияют на дефолт. Как мы можем увидеть, возраст является наиболее значимым признаком.

Таблица 1

Важность признаков

Признак	**Важность**
Возраст	0.238120
Доход	0.202871
Сумма кредита	0.139631

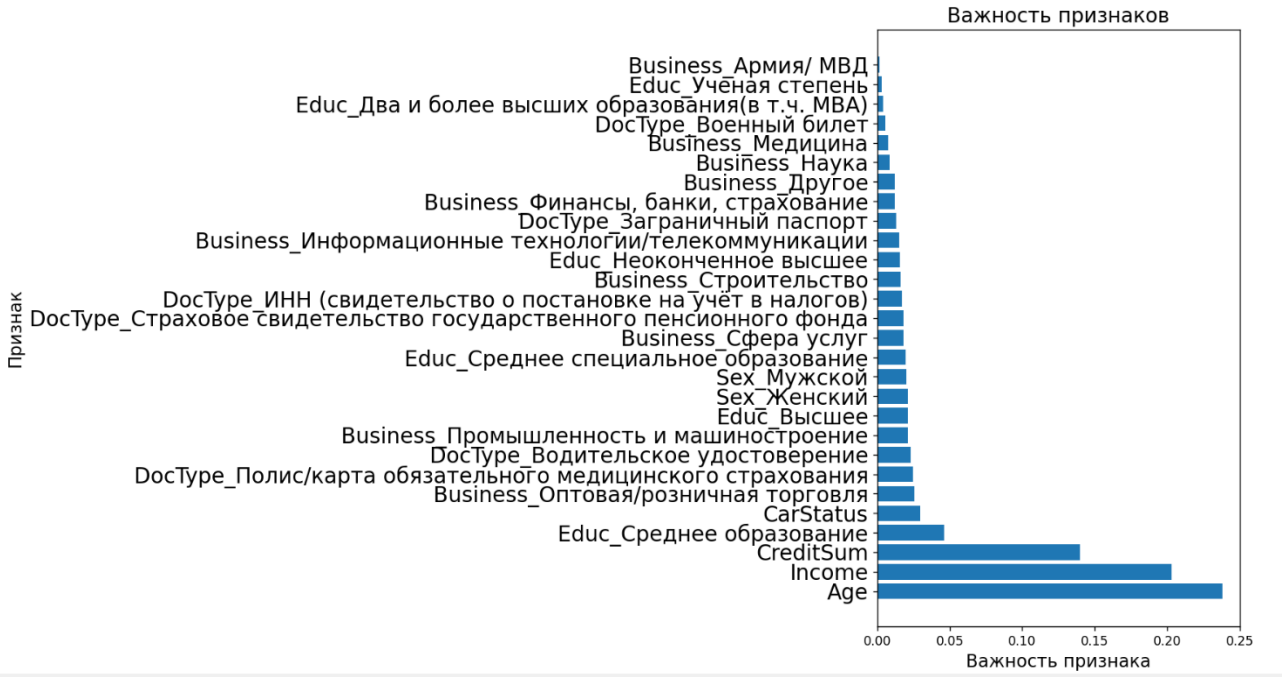

Рис. 1. Важность признаков для предсказания дней просрочки

Показатели метрик классификатора [4] приведены в таблице 2, где 0 – клиент имеет менее 90 дней просрочки, 1 – клиент имеет более 90 дней просрочки. Точность модели составила 0.89, что является не плохим результатом. На рис. 2 представлен график зависимости Precision от Recall.

Таблица 2

Показатели метрик классификатора

	precision	recall	F1-score	support
0	0.89	0.98	0.93	262
1	0.60	0.15	0.24	39
Accuracy			0.89	301
Macro avg	0.74	0.57	0.59	301
Weighted avg	0.85	0.88	0.84	301

Recall показывает, способность обнаружить данный класс, а Precision – отличить его от объектов другого класса. Гармоническое среднее между двумя этими метриками показывает F1-score. Из таблицы можно сделать вывод, что для класса 0 F1 мера является очень высокой, что также показано на рис. 2. Support отражает количество истинных образцов для каждого класса в тестовом наборе данных. Другими словами, это количество объектов в тестовом наборе, которые принадлежат определенному классу.

Синяя кривая на рис. 2 соответствует классу 0 (клиент считается «хорошим»), и при увеличении полноты она демонстрирует высокие значения точности, что свидетельствует о правильном распознавании «хороших» клиентов. Однако, для оранжевой кривой, представляющей класс 1 (клиент считается «плохим»), с увеличением полноты точность снижается. Это объясняется тем, что обучающий набор данных содержит только небольшую часть записей для класса 1 (с просрочкой более 90 дней), что снижает точность прогнозирования этого класса.

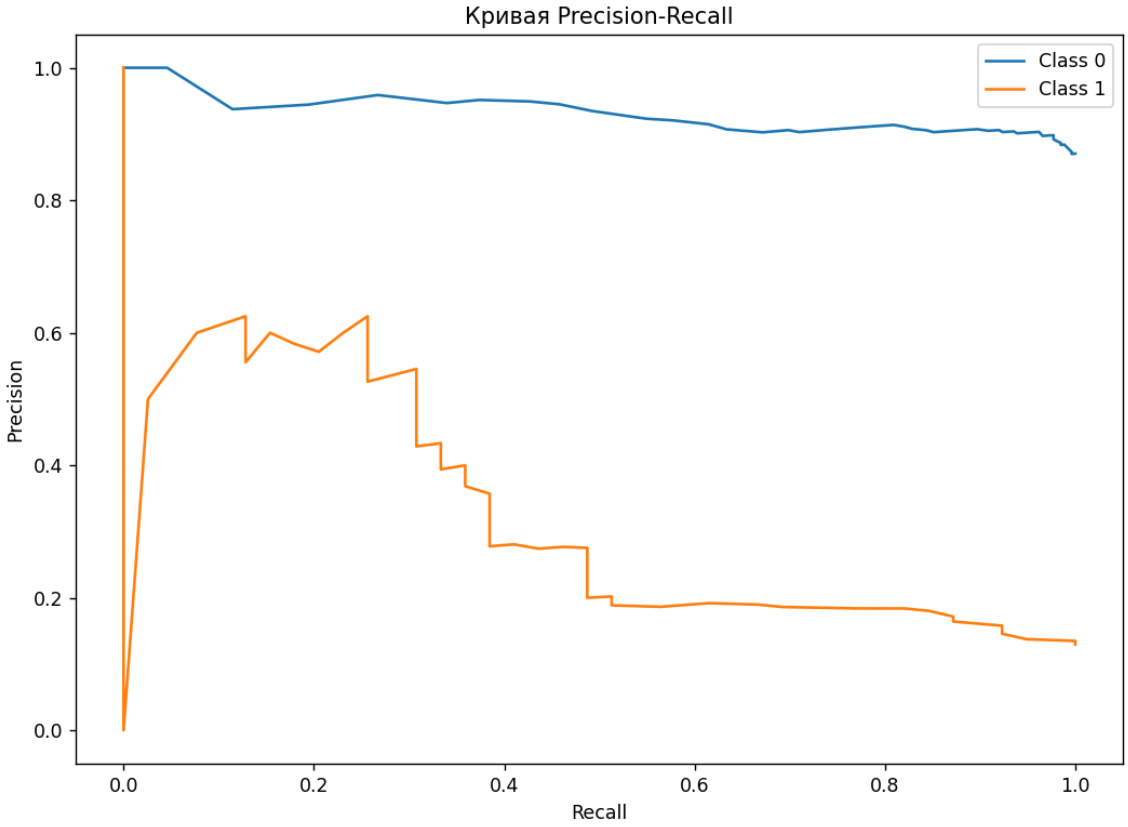

Рис. 2. Кривые Precision-Recall

Заключение

На данном этапе разработки алгоритма наилучшие метрики качества имеет класс 0, так как данные содержат подавляющее большинство записей для этого класса. Для улучшения метрик класса 1 в будущем следует дополнить исходный набор записями о классе 1 (более 90 дней просрочки).

В результате работы было установлено, что наибольший вклад в оценку дефолта заемщика потребительского кредита вносят такие признаки как доход, сумма запрашиваемого кредита и возраст клиента. Таким образом, можно сделать вывод, что при оценке рисков следует уделять внимание этим признакам. Полученная модель может быть использована для оценки рисков в финансовом секторе для других банковских продуктов.

Список использованных источников

1. Статистические методы анализа в биологии и медицине – Беспалов А.Ф., Беляев А.Н // ИФМиБ: Казань.
2. Aurélien Géron Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow // O'Reilly Media: Inc., 1005 Gravenstein Highway North, Sebastopol. – CA 95472. – 2019 Second Edition. – С. 200-201.
3. Документация Scikit-learn. [Электронный ресурс]. – URL: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
4. Андреас Мюллер, Сара Гвидо Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными – М: 2016. – 2017. – С. 100-106, 147-180.