Credit Approval

Шевченко Артём и Мирзоева Алина, БЭАД223

НИС, ВШЭ

7 ноября 2023 г.

Введение

В данном проекте объектом исследования является датасэт с данными о заявках по выдаче кредитных карт. Датасэт содержит зашифрованные данные из анкеты потенциального заемщика. В своем исследовании мы попыпытаемся определить модель машинного обучения, которая наиболее точно предсказывает целесообразность выдачи кредита заемщику.

Постановка задачи

Наша цель заключается в разработке эффективной системы классификации, которая может точно отличать между "кредитоспособными"и "некредитоспособными"клиентами на основе предоставленных векторов признаков. Это позволит банку принимать обоснованные решения при выдаче кредитов, минимизируя риски и максимизируя прибыль.

Использованные методы

Для построения различных моделей использовали библиотеку scikit-learn на языке программирования Python. Основными инструментами оценивания точности модели будут показатели, получаемые из Матрицы ошибок. Также в проекте мы использовали показатель *"Среднеквадратичная ошибка"*.

Матрица ошибок

Accuracy:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

— доля исходов, которую наша модель определила правильно.

Precision:

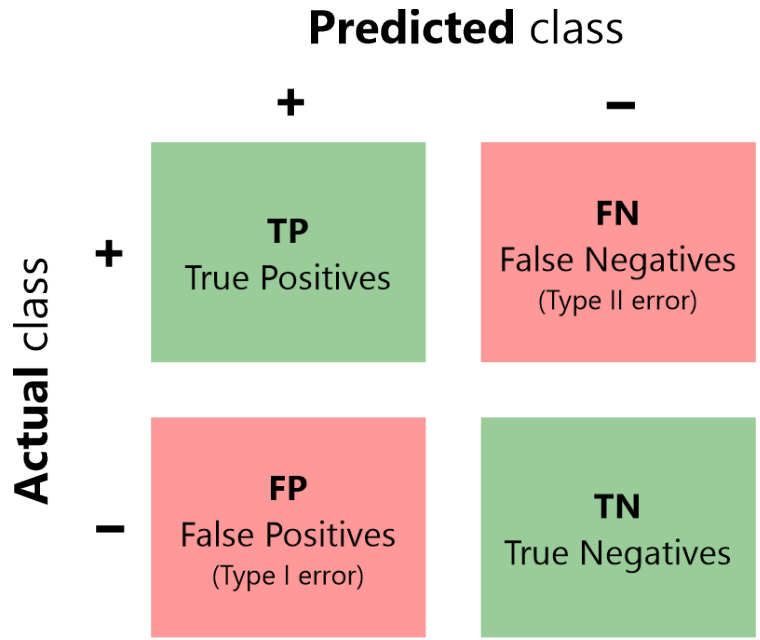
$$\frac{TP}{TP + FP}$$

— какая доля от предсказанных объектов класса 1 и в реальности является объектом класса 1.

Recall:

$$\frac{TP}{TP + FN}$$

— какая доля от тех значений, которые в реальности должны относиться к классу 1, были правильно определены моделью.



Среднеквадратичная ошибка

Среднеквадратичная ошибка (Mean Squared Error, MSE) используется для измерения средней квадратичной разницы между фактическими значениями и предсказанными значениями в регрессионных моделях.

Среднеквадратичная ошибка вычисляется по формуле:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Где:

- n - количество наблюдений в выборке.

Среднеквадратичная ошибка

Среднеквадратичная ошибка (Mean Squared Error, MSE) используется для измерения средней квадратичной разницы между фактическими значениями и предсказанными значениями в регрессионных моделях.

Среднеквадратичная ошибка вычисляется по формуле:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Где:

- n - количество наблюдений в выборке.
- y_i - фактическое значение переменной i.

Среднеквадратичная ошибка

Среднеквадратичная ошибка (Mean Squared Error, MSE) используется для измерения средней квадратичной разницы между фактическими значениями и предсказанными значениями в регрессионных моделях.

Среднеквадратичная ошибка вычисляется по формуле:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Где:

- n - количество наблюдений в выборке.
- y_i - фактическое значение переменной i.
- $\hat{y}_i$ - предсказанное значение переменной i.

Множественная линейная регрессия

В контексте задачи скоринга

Логистическая регрессия

В контексте задачи скоринга

Метод случайных лесов

Задача классификации – получение категориального ответа на основе набора признаков. Классификацию можно производить с помощью множества алгоритмов, но мы рассмотрим один из самых простых - метод случайного леса.

Алгоритм

- Создание выборок: Для построения каждого дерева случайным образом выбираются с повторениями примеры из исходных данных.
- Построение деревьев решений: Для каждой из выборок строится отдельное решающее дерево. Построение дерева включает в себя разбиение данных на узлы (по одному признаку в каждом узле), выбор лучшего разбиения и продолжение этого процесса до достижения необходимой глубины дерева.
- Прогнозирование: Когда все деревья построены, они могут быть использованы для прогнозирования новых данных. Для задачи классификации, каждое дерево голосует за класс объекта, и класс с наибольшим числом голосов становится предсказание

В контексте кредитного скоринга

Наша таргетная переменная представляет собой бинарный категориальный признак ("одобрен"или "не одобрен"кредит), то задачу предсказания таких ответов можно отнести к задаче классификации в машинном обучении.

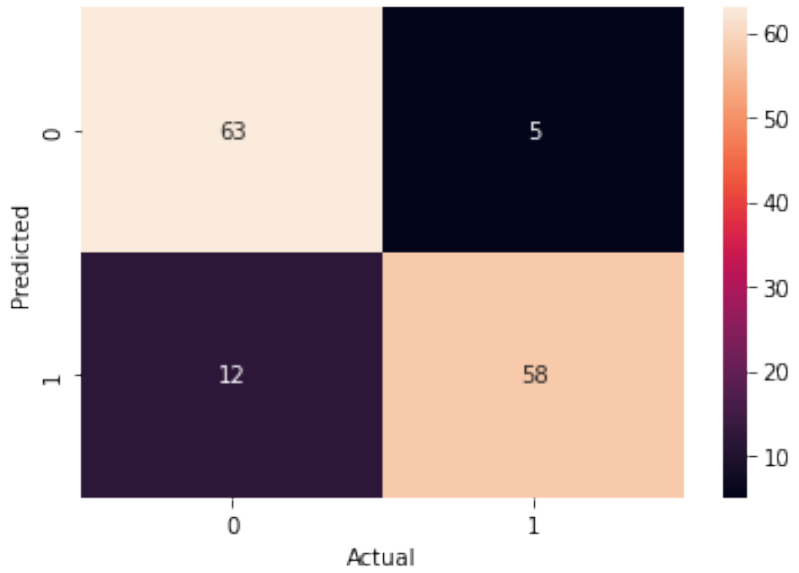
Полученные результаты

Полученные результаты

Полученные результаты

Метод случайных лесов:

- MSE = 0.123
- Precision = 0.921
- Recall = 0.829



Итоги

Задавайте вопросы!