Министерство образования Российской Федерации
Научно-Исследовательский Университет Высшая Школа Экономики

Факультет Компьютерных Наук
Факультет Экономических Наук

Credit Approval

Мирзоева Алина и Шевченко Артём, БЭАД223

Научно-учебная лаборатория методов анализа больших данных

Научно-Исследовательский Семинар, _____Т.А. Рамазян

Москва, 2023

5 ноября 2023 г.

Содержание

1 Введение

Кредитный скоринг - это важный инструмент в финансовой сфере, который позволяет банкам и финансовым учреждениям оценивать кредитоспособность заемщиков на основе различных финансовых и личных данных. 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 попытаемся определить модель машинного обучения, которая наиболее точно предсказывает целесообразность выдачи кредита заемщику.

2 Задача

Наш датасет содержит информацию о решениях банка относительно выдачи кредитных карт, где каждая заявка подвергается определению к одному из классов : "одобрен"или "не одобрен". Главная задача банка заключается в том, чтобы с достаточно высокой точностью определять надежных и ненадежных заемщиков, чтобы минимизировать потери от невозврата кредитов и одновременно избегать большого количества упущенной прибыли, связанной с отказом хорошим кандидатам.

Таким образом, 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.

3 Методология

Для построения различных моделей машинного обучения в своем исследовании мы использовали библиотеку scikit-learn на языке программирования Python. В нашем проекте мы исследовали модель линейной, логистической регрессии и классификации с помощью метода случайного леса. Особенности каждой из этих моделей будут описаны чуть позже, а пока разберемся с инструментами оценивания точности моделей. В нашей работе основными инструментами оценивания точности модели будут показатели, получаемые из **Матрицы ошибок**. Для начала объясним структуру матрицы ошибок:

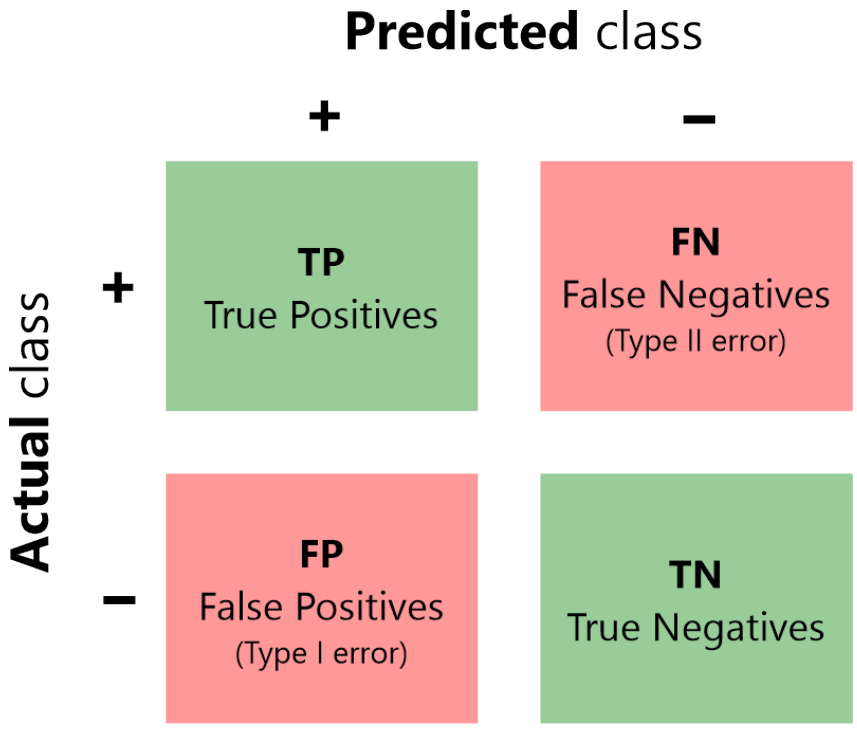

Рис. 1: Матрица ошибок

Пусть:
0 — объекты относятся к нулевому классу, (в нашей модели - отказ в выдаче займа)
1 — объекты относятся к классу 1 (в нашей задаче - одобренный займ).

Матрица ошибок - матрица, в нашем случае размера 2 на 2, такая что столбцы означают то, к какому классу объект относится в реальности, а строки - к какому признаку относится объект по прогнозу модели. В ячейках на пересечении обычно записываются числа, означающие количество объектов на валидационной выборке с одной из 4 возможных комбинаций.

С помощью матрицы ошибок мы можем ввести несколько следующих метрик:
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 относится к классу 1, были правильно определены моделью.

Иными словами, в условиях нашей задачи:
Precision - сколько мы выдали кредитов тем людям, которым должны были выдать в реальности. Тут учитывается, что, мы могли выдать кредиты тем, кто в последствии будет недобросовестным заемщиком (то есть в реальности принадлежит классу 0). Из-за этого банк может понести убытки.

Recall - сколько мы выдали кредитов тем, кто должен был их получить, с учетом того, что мы отсеяли какую-то часть добросовестных заемщиков (представителей класса 1) и как банк упустили выгоду. Recall - способность нашей модели определять класс 1.

Также в проекте мы использовали показатель *"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:

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

Где:

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

MSE измеряет средний квадрат разницы между фактическими и предсказанными значениями. Чем меньше значение MSE, тем лучше модель соответствует данным, поскольку это означает, что предсказанные значения ближе к фактическим значениям.

4 Исследуемые модели машинного обучения

4.1 Линейная регрессия

Описание алгоритма линейной регрессии.

4.2 Логистическая регрессия

Описание логистической регрессии.

4.3 Модель классификации

Так как 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.

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. Чуть более подробно, чем в ноутбуке с кодом остановимся на алгоритме:

- 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 предсказанием случайного леса.

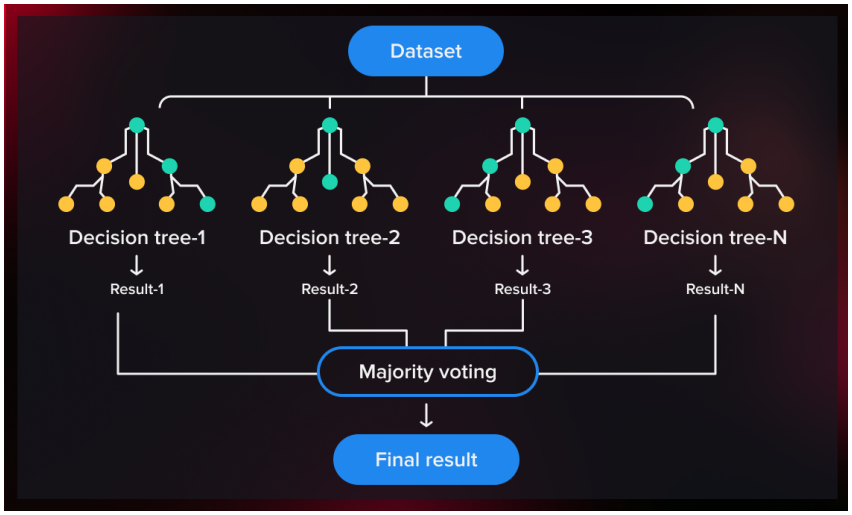

Рис. 2: Алгоритм случайного леса

5 Результаты

6 Выводы

Сделать выводы