Credit approval

Степанова Елена, Мария Пахолкова

Содержание

1 Введение

Задача кредитного скоринга заключается в том, чтобы на основе некоторых признаков разделить заявителей на два типа: заявители с хорошей кредитной историей (кредит одобряется) и плохой (кредит не одобряется). Для финансовых институтов точность классификации очень важна, именно поэтому используются различные модели, которые позволяют увеличивать эффективность прогноза.

2 Предобработка данных

Необходимо провести предварительную обработку датасета, так как исходные данные содержат пропуски.

Самый простой способ - удалить строки с пропусками, однако это может привести к потере большого количества данных, поэтому мы рассмотрим другой способ. Для столбцов с буквенными и символьными значениями (категориальные признаки) заменим пропуски наиболее популярными значениями, для столбцов с числовыми значениями - средними. Также убедимся, что нет повторяющихся строк.

3 Разведочный анализ данных (EDA)

Данный этап заключается в выявлении общих тенденций, закономерностей и связей между переменными. Разведочный анализ проходит в несколько этапов:

3.1 Распределение по переменным:

На данном этапе мы рассматриваем распределение по различным переменным. Для этого рассмотрим один из примеров. Возьмем столбец A1 и посмотрим на распределение значений в нем. Можно заметить, что в столбце встречаются значения a и b, мы можем посчитать их количество и процентное соотношение, так как мы знаем общее количество значений. Аналогично поступим с другими столбцами.

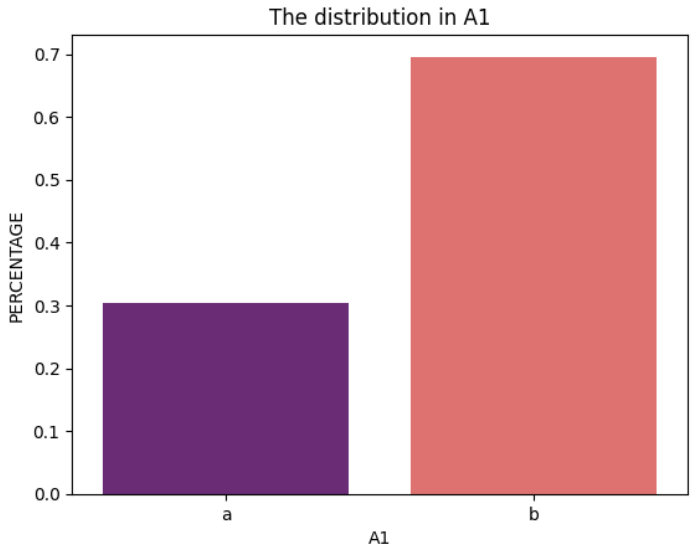

Рис. 1: The distribution in A1

Еще несколько примеров представлены в colab в соответсвующем разделе.

3.2 Корреляция между переменными:

На данном этапе мы хотим узнать, взаимосвязаны ли между собой столбцы. Для этого мы возьмем столбцы с числовыми значениями и посчитаем коэффициент корреляции попарно с помощью метода numpy.corrcoef. Возможны 3 варианта:

- **Положительная корреляция:**
  Если одна переменная увеличивается, другая также увеличивается. Коэффициент корреляции находится в диапазоне от 0 до 1.
- **Отрицательная корреляция:**
  Если одна переменная увеличивается, другая уменьшается. Коэффициент корреляции находится в диапазоне от 0 до -1.
- **Нулевая корреляция:**
  Отсутствие линейной зависимости между переменными. Коэффициент корреляции близок к 0.

Например, можно заметить небольшую положительную корреляцию между столбцами A2 и A8 (коэффициент корреляции между A2 и A8: 0.3927872276643495).

4 Применение различных моделей

Задача кредитного одобрения является классической задачей классификации, т. е. получения категориального ответа на основе некоторых признаков. В нашей задаче мы имеем два варианта

ответа: «кредит одобрен» и «кредит не одобрен». В нашей работе мы использовали основные модели, которые применяются при решении задачи кредитного скоринга.

Основные используемые библиотеки:

- pandas
- seaborn
- numpy
- matplotlib
- sklearn (включает все алгоритмы и инструменты, которые нужны для задачи классификации)

Приступим непосредственно к применению моделей. Для этого нужно разделить данные на обучающие и тестовые, используем функцию train_test_split().

4.1 Дерево принятия решений:

В данной модели данные разбиваются на более мелкие подмножества на основе различных критериев и образуют иерархическую структуру. Решающее дерево включает в себя элементы двух типов - узлы (содержат правила принятия решения) и листья (последние узлы, в которых принимаются решения).

Создаем модель решающего дерева с помощью функции DecisionTreeClassifier() и обучаем нашу модель с помощью функции fit(). Модель построена. С помощью функции predict() проверяем ее работу на тестовых данных.

Теперь мы можем оценить точность полученных результатов. Для этого используем функцию metrics.accuracy_score() и сравниваем результаты, которые были изначально представлены в таблице, с теми, которые мы получили в результате применения модели. Получаем следующую точность: 0.8265895953757225. Также можем построить матрицу неточностей confusion_matrix(), которая показывает сколько объектов было классицировано правильно и сколько неправильно по каждому классу.

4.2 Логистическая регрессия:

Данная модель используется для получения некоторого прогноза с помощью определения зависимости между переменными, где одна из переменных категориально зависима, а другие - независимы. В результате получается моделирование вероятности ограниченного числа результатов (в нашем случае - двух).

Для начала создаем модель логистической регрессии LogisticRegression(). Далее делаем то же самое, что и в предыдущем пункте - обучаем и тестируем модель, проверяем точность полученных результатов, строим матрицу неточностей и выводим отчёт о классификации. Используя функцию metrics.accuracy_score() получаем следующую точность: 0.8497109826589595.

4.3 Метод опорных векторов:

В данной модели между различными группами точек рисуется разделительная линия. Основная идея заключается в том, что каждый объект находится в одном из двух классов, и мы можем провести линию между этими классами. Далее ищутся точки, которые находятся ближе всего к этой линии (опорные вектора), и строится максимально удаленная гиперплоскость. В итоге мы получаем разбиение точек на классы.

Далее все аналогично предыдущим пунктам - строим модель с помощью функции SVC(). Делаем то же самое, что и в предыдущем пункте - обучаем и тестируем модель, проверяем точность полученных результатов, строим матрицу неточностей и выводим отчёт о классификации. Используя функцию metrics.accuracy_score() получаем следующую точность: 0.7167630057803468.

5 Результаты

Оценим работу построенных моделей и сделаем выводы об их эффективности. Для этого можно использовать матрицу неточностей. В данной матрице столбцы представляют собой прогнозируемые классы, а стоки - фактические классы. Таким образом, получаем 4 ячейки: истинно-положительные объекты, ложно-положительные объекты, истинно-отрицательные объекты, ложно-отрицательные объекты. Таким образом, определяется, правильно или неправильно объекты были отнесены к одному из классов. Рассмотрим матрицы неточностей, которые получились в построенных моделях:

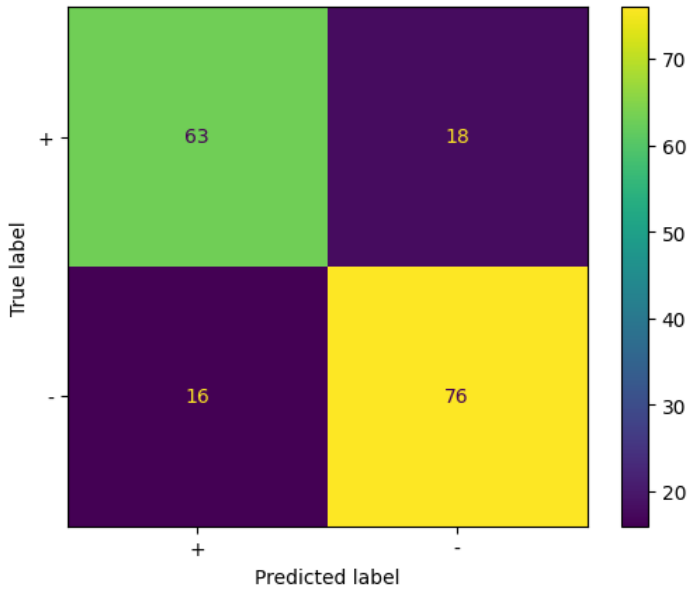

Рис. 2: Матрица неточностей для решающего дерева

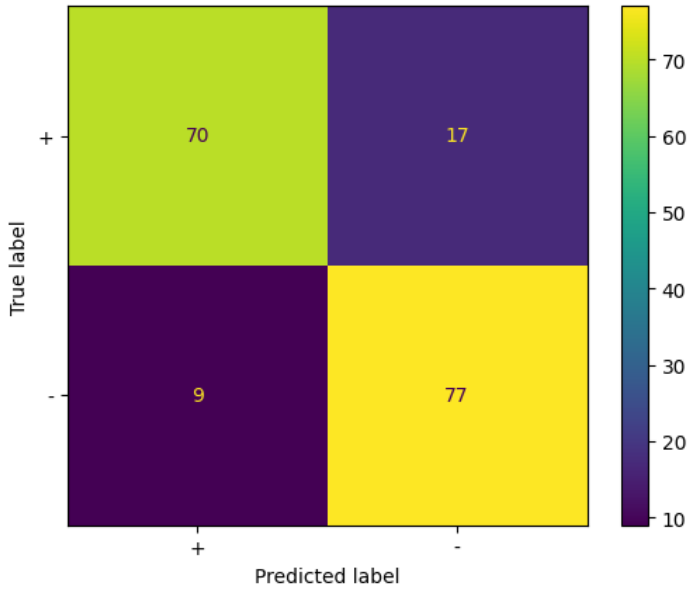

Рис. 3: Матрица неточностей для логистической регрессии

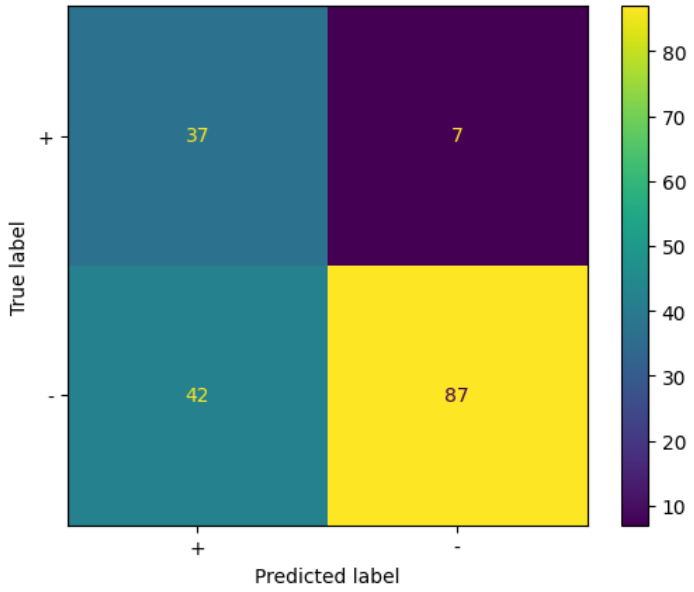

Рис. 4: Матрица неточностей для метода опорных векторов

Видим, что в методе **логистичнской регрессии** наибольшее количество объектов определено истинно положительно или истинно отрицательно, и наименьшее количество - ложно положительно или ложно отрицательно. Из этого можно сделать вывод, что из трех построенных выше моделей **логистичнская регрессия** дает наиболее приближенный к истинному результат.