|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | ***«*Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

НАПРАВЛЕНИЕ ПОДГОТОВКИ: 09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА\_\_\_

**ОТЧЕТ**

**по лабораторной работе №** \_\_2**\_\_**

**Дисциплина: Прикладной анализ данных**

**Название лабораторной работы:** Задача бинарной классификации

Студент гр. ИУ6-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  А.М. Латышев

(Подпись, дата) (И.О. Фамилия)

Преподаватель  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О. Фамилия)

Москва, 2024

**Цель работы:**

Поиск решения задачи бинарной классификации

**Задание:**

Вариант 12

Вы решаете задачу бинарной классификации на таргет Ybin – бинарная переменная, где 1 обозначает регион с выраженностью демографических трендов.

Этапы:

1. Разделение данных на обучающую (85%) и тестовую часть (15%) случайным образом (можно сделать более корректным методом – разделить в такой пропорции с сохранением распределения таргета в каждой подвыборке – желательно, но не обязательно).
2. Нормирование (масштабирование) исходных данных. Обратите внимание, что данные для нормализации (масштабирования) рассчитываются только на основе обучающей выборки.
3. С помощью библиотеки sklearn сделать fit-predict модели kNN. Перебрать по сетке параметр числа соседей с целью определения наилучшего на тестовой выборке.
4. С помощью библиотеки sklearn сделать fit-predict модели логистической регрессии. Перебрать по сетке параметр регуляризации с целью определения наилучшего на тестовой выборке.
5. С помощью библиотеки sklearn сделать fit-predict модели дерева решений. Перебрать по сетке параметр глубины дерева с целью определения наилучшего на тестовой выборке. Дополнительно (желательно, но не обязательно): с помощью библиотеки sklearn сделать fit-predict модели случайного леса. Перебрать по сетке параметр глубины дерева с целью определения наилучшего на тестовой выборке.
6. Сравнить качество всех моделей на обучающей и тестовой выборке отдельно по метрикам Accuracy, ROC-AUC, Precision, Recall, F1-мера. Обратите внимание, что 4 из 5 метрик требуют определения порога отсечения по вероятности. В качестве эвристики предлагается взять его как среднее значение полученных вероятностей (желательно, но не обязательно: подобрать по сетке такой порог, при котором precision и recall примерно уравниваются).
7. Проанализировать различие в качестве между моделями. Определить на основе метрик модели, в которых сильно выражено переобучение.
8. Сравнить полученную важность признаков в модели логистической регрессии, в модели деревьев решений и в случайном лесе (для древесных моделей это можно сделать с помощью ключа feature\_importances у обученной модели). Проинтерпретировать полученную важность признаков.

**Ход работы:**

1. Исходные данные были разделены на обучающую и тестовую выборки в пропорции 85% к 15%.



Рисунок 1 – Исходные данные и разделение на выборки

1. Было выполнено нормирование (масштабирование) данных на основе обучающей выборки.

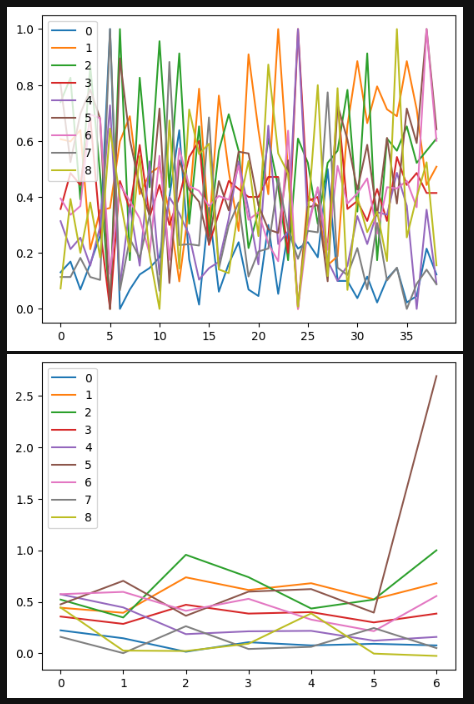


Рисунок 2 – Графики нормированных данных

1. Была построена модель kNN с использованием библиотеки sklearn, перебраны параметры числа соседей и выбраны наилучшие параметры на тестовой выборке.

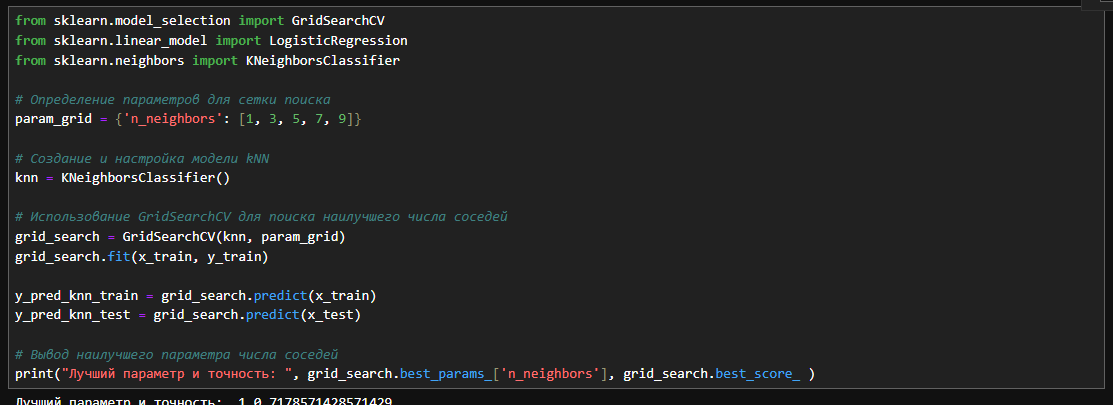
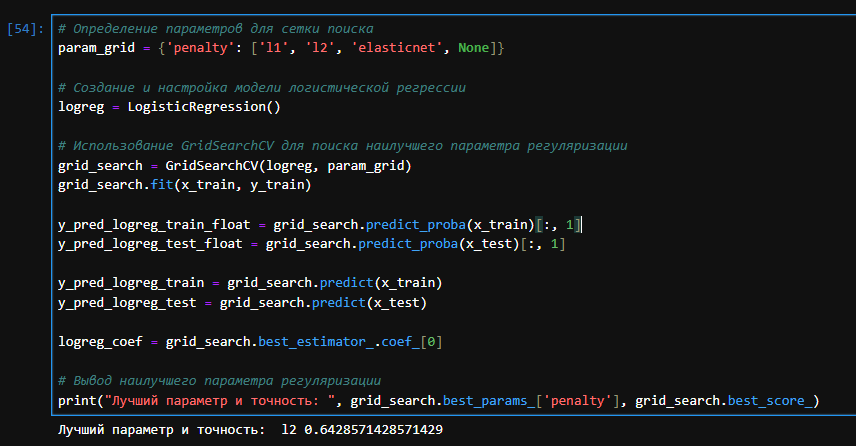


Рисунок 3 – Модель knn

1. Была построена модель логистической регрессии с использованием библиотеки sklearn, перебраны параметры регуляризации и выбраны наилучшие параметры на тестовой выборке.

  
Рисунок 4 – Модель логистической регрессии

1. Была построена модель дерева решений с использованием библиотеки sklearn, перебраны параметры глубины дерева и выбраны наилучшие параметры на тестовой выборке. Также была построена модель случайного леса, перебраны параметры глубины дерева и выбраны наилучшие параметры на тестовой выборке.

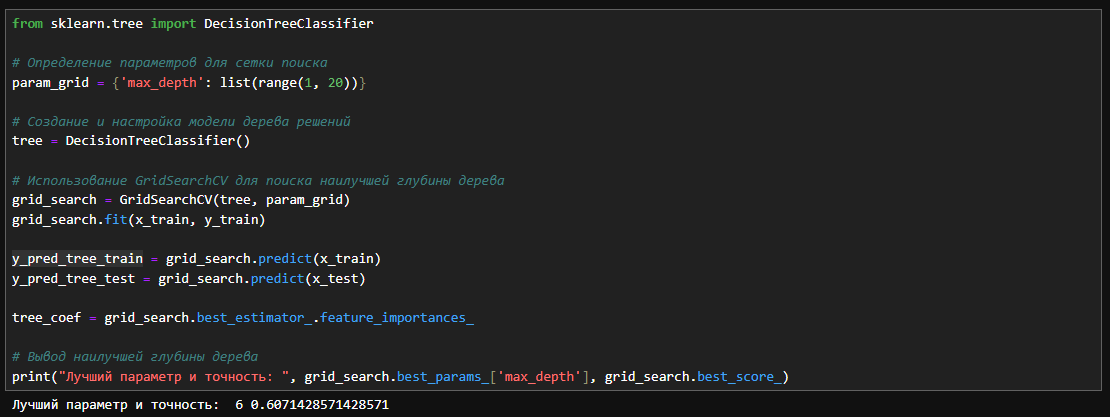


Рисунок 5 – Модель дерева решений

1. Было проведено сравнение качества всех моделей на обучающей и тестовой выборках по метрикам Accuracy, ROC-AUC, Precision, Recall и F1-мера. Были определены пороги отсечения по вероятности для каждой модели.

Оптимальный порог отсечения для Логистической регрессии оказался равен 0.35998

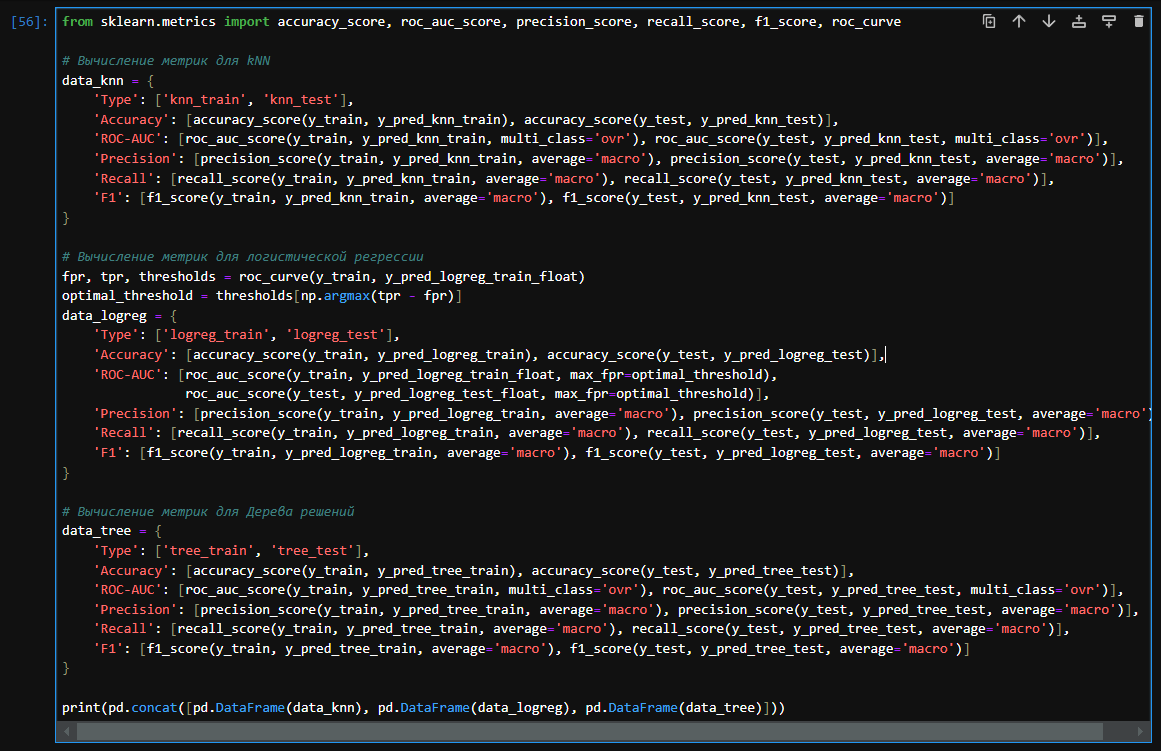


Рисунок 6 – Подсчет и формирование результатов сравнений

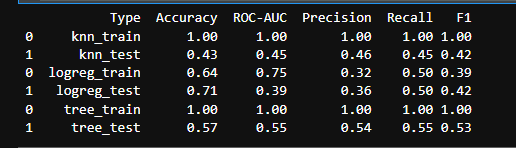


Рисунок 7 – Метрики по моделям

1. Было проведено анализ различий в качестве моделей и определены модели с переобучением на основе метрик.

Переобучение модели происходит, когда ее показатели на обучающей выборке значительно превышают результаты на тестовой. В данной работе для определения переобучения был установлен порог в 0,5. По этому критерию модель логистической регрессии имела явные проблемы с точностью на тестовой выборке, в то время как kNN демонстрировала стабильно низкие показатели. Наилучшей моделью среди всех оказалась модель дерева решений, которая имела высокие и стабильные метрики как на обучающей, так и на тестовой выборках, хотя recall указывал на возможное переобучение.

1. Была выполнена оценка важности признаков в моделях логистической регрессии, деревьев решений с помощью ключа feature\_importances. Была проинтерпретирована полученная важность признаков.

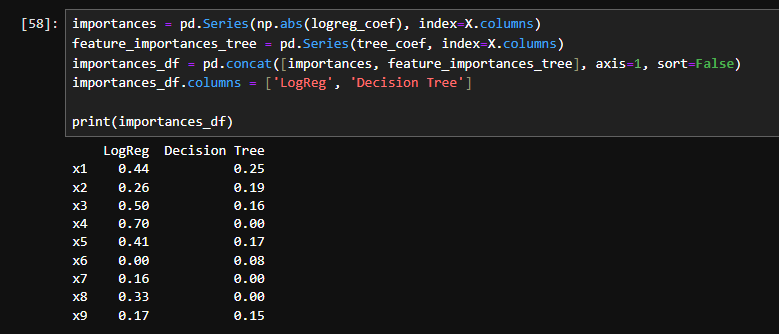


Рисунок 8 – Важность признаков для моделей

**Вывод:**

В ходе выполнения лабораторной работы были изучены методы построения моделей бинарной классификации, таких как дерево решений, логистическая регрессия и kNN. Также полученные в ходе работы модели были проанализированы на предмет их переобучения.