Отчёт по проекту по дисциплине "Основы машинного обучения"

Предсказание доли переработанного пластика в данном регионе в данный год

Добрецова Елизавета, группа 5030102/00101

Январь 2024

Описание задачи и примеры

В ходе выполнения проекта были составлены и решены две задачи с семантическим разрывом.

1. Задача безусловного предсказания

   На основе тренировочной выборки троек вида (регион, год, доля переработанного пластика из выброшенного в этом регионе в этом году пластика) предсказать, какая доля выбрасываемого пластика будет переработана в конкретный год в конкретном регионе.
   Набор регионов следующий: Северная и Южная Америки без США, Азия без Китая и Индии, Китай, Европа, Индия, Ближний Восток и Северная Африка, Океания, Африка ниже Сахары, США.

Описание задачи и примеры

- На вход подается пара (год, регион), на выход - доля перерабатываемого пластика, предсказанная безусловно на основе тренировочной выборки.

2. Задача предсказания на основе признаков

   Теперь на вход подаётся год и набор из 8 значений экономических признаков региона. По признакам модель машинного обучения предсказывает параметры зависимости доли переработанного пластика от года. На выход программа подаёт долю переработанного пластика, предсказанную для данного года в регионе с данными значениями признаков.

Описание задачи и примеры

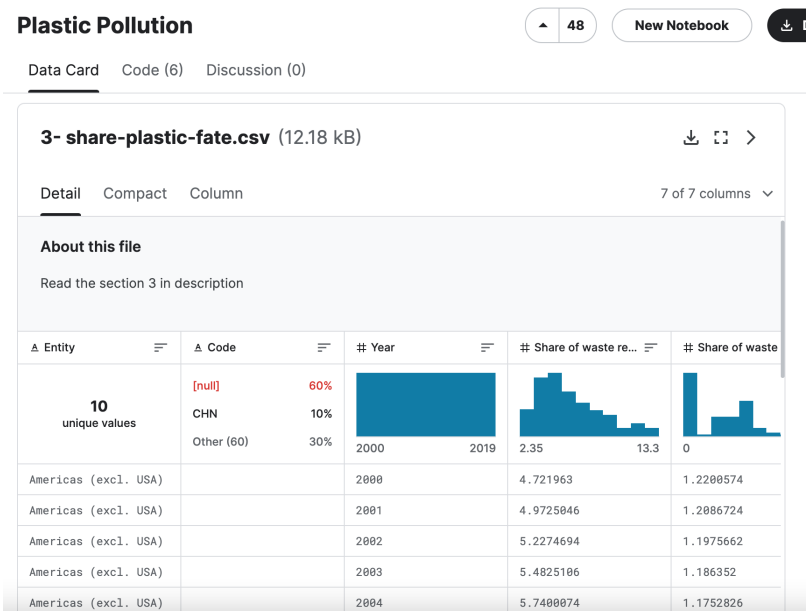
prediction_from_previous_years

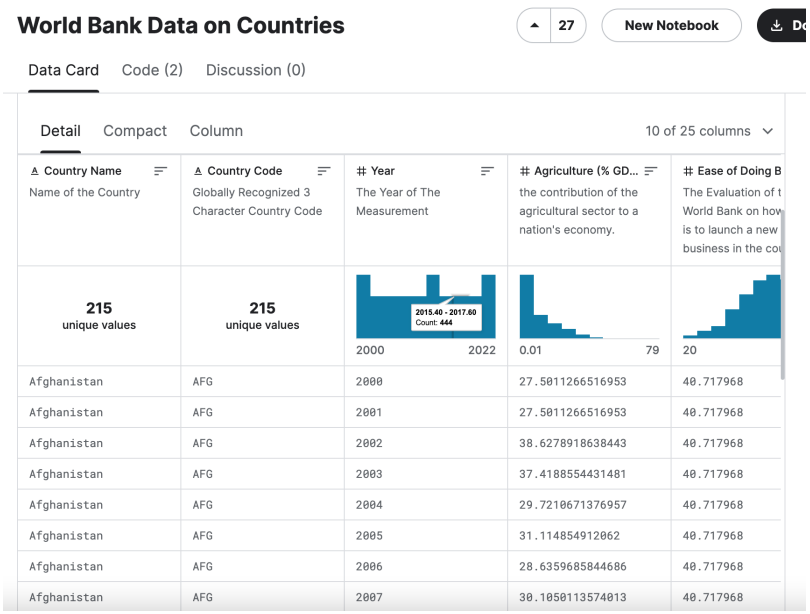
```
/Users/paks/Desktop/Polytech/ML/machine_learning/venv/bin/python /Users/paks/Desktop/Polyte
Input GDP:
2.130520e+12
Input R&D:
7.834763e+09
Input Population:
1.185010e+09
Input Land:
2.347501e+07
Input Export:
5.760199e+11
Input Education Expenditure:
2.947490e+11
Input Health Expenditure:
1.016391e+11
Input Net Trade:
4.081170e+10
Input a year:
2019
/Users/paks/Desktop/Polytech/ML/machine_learning/src/prediction_from_previous_years.py:62:
  tan += feature * linear_regression_coefficients.loc["Tangent"][i]
/Users/paks/Desktop/Polytech/ML/machine_learning/src/prediction_from_previous_years.py:63:
  bias += feature * linear_regression_coefficients.loc["Bias"][i]
tan =  0.17157996635451392 ,bias =  4.010931453313901
According to prediction with features, in 2019 the share of plastic recycled was 100 %

Process finished with exit code 0
```

Как собирались данные (1)

Проект выполнен на основе датасетов, взятых с сайта Kaggle:

- `https://www.kaggle.com/datasets/imtkaggleteam/plastic-pollution/?select=3-+share-plastic-fate.csv`
- `https://www.kaggle.com/datasets/yusufglcan/country-data?select=Countries.csv`





Как собирались данные (2)

Как можно видеть, первый датасет содержит доли переработанного пластика в конкретном регионе в конкретный год, а второй - описанные ранее признаки, только не по регионам, а по странам. Соответственно, первой задачей стало выровнять датасеты, т. е. собрать страны в регионы. Это делалось с помощью библиотеки Pandas языка Python в Jupyter Notebook на основе данных из Википедии (`https://en.wikipedia.org/wiki/North_Africa`, `https://en.wikipedia.org/wiki/Middle_East`). Вручную были исправлены несколько ошибок в данных: так, Армения, Турция, Российская Федерация и ещё несколько стран были отнесены к Азии вместо Европы, а Кипр - к Европе вместо Азии. После этого два датасета были слиты в один табличного типа.

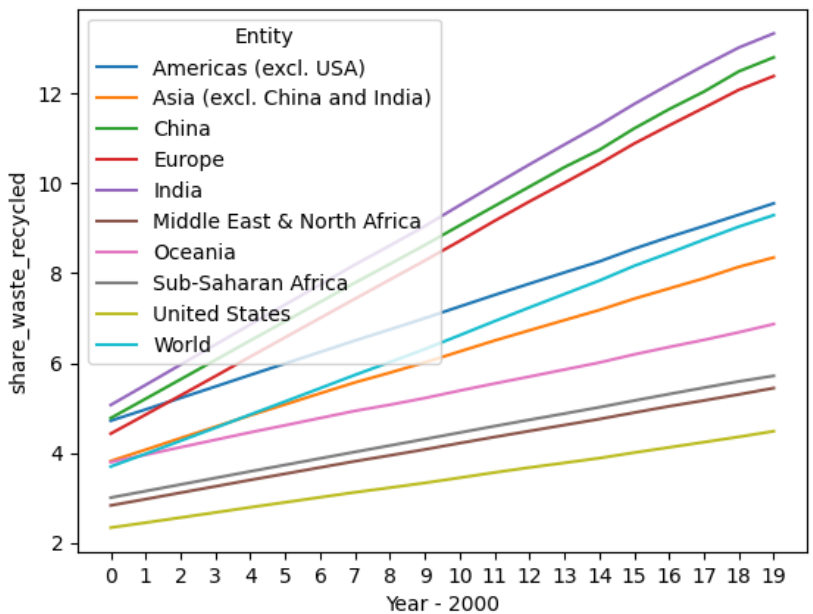
Как собирались данные (3)

В таблице из датасета со странами ("World Bank Data") содержались признаки как имеющие относительные (выраженные в процентах) значения, так и абсолютные. При выравнивании датасетов были взяты только имеющие абсолютные значения признаки: ВВП, расходы на научные исследования, население, площадь региона, экспорт, импорт, расходы на образование, расходы на здравоохранение, объём торговли. Они были просуммированы по значениям для всех стран каждого региона.

	Entity	Year	GDP	R&D	Population	Land	Export	Import	Education Expenditure	Health Expenditure	Net Trade	share_waste_recycled
0	Americas (excl. USA)	2000	3.047337e+12	2.558195e+10	552084912.0	30715742.0	8.104982e+11	7.829520e+11	4.593228e+11	2.061849e+11	2.754618e+10	4.721963
1	Americas (excl. USA)	2001	2.992335e+12	2.599439e+10	559828034.0	30715742.0	7.853328e+11	7.592970e+11	4.572757e+11	2.075397e+11	2.603575e+10	4.972505
2	Americas (excl. USA)	2002	2.785549e+12	2.470977e+10	567415146.0	30715742.0	7.809170e+11	7.255198e+11	4.270469e+11	1.928021e+11	5.539720e+10	5.227469
3	Americas (excl. USA)	2003	2.962284e+12	2.809765e+10	574740675.0	30715742.0	8.501986e+11	7.722607e+11	4.721862e+11	2.122359e+11	7.793788e+10	5.482511
4	Americas (excl. USA)	2004	3.405703e+12	3.251802e+10	581958944.0	30715742.0	1.006211e+12	9.045789e+11	5.146562e+11	2.459280e+11	1.016325e+11	5.740007
...	...	...	...	...	...	...	...	...	...	...	...	...
175	United States	2015	1.820602e+13	5.065024e+11	320738994.0	9831510.0	2.268651e+12	2.794850e+12	2.466122e+12	3.000551e+12	-5.261990e+11	4.012979
176	United States	2016	1.869511e+13	5.320348e+11	323071755.0	9831510.0	2.232110e+12	2.738359e+12	2.459194e+12	3.139498e+12	-5.062490e+11	4.129148
177	United States	2017	1.947734e+13	5.640091e+11	325122128.0	9831510.0	2.388260e+12	2.924994e+12	2.700468e+12	3.265947e+12	-5.367340e+11	4.245349
178	United States	2018	2.053306e+13	6.161601e+11	326838199.0	9831510.0	2.538089e+12	3.131166e+12	2.699313e+12	3.416894e+12	-5.930770e+11	4.363442
179	United States	2019	2.138098e+13	6.769409e+11	328329953.0	9831510.0	2.538450e+12	3.117235e+12	2.860892e+12	3.565593e+12	-5.787850e+11	4.488883

Задача безусловного предсказания

После получения таблицы с данными по регионам на основе данных из третьего и последнего столбцов были построены графики доли переработанного пластика от года для каждого из регионов. Видно, что на графиках практически прямые линии, следовательно, задача безусловного предсказания свелась к нахождению для каждой прямой двух чисел: тангенсы угла наклона и ординаты точки пересечения с осью ординат.



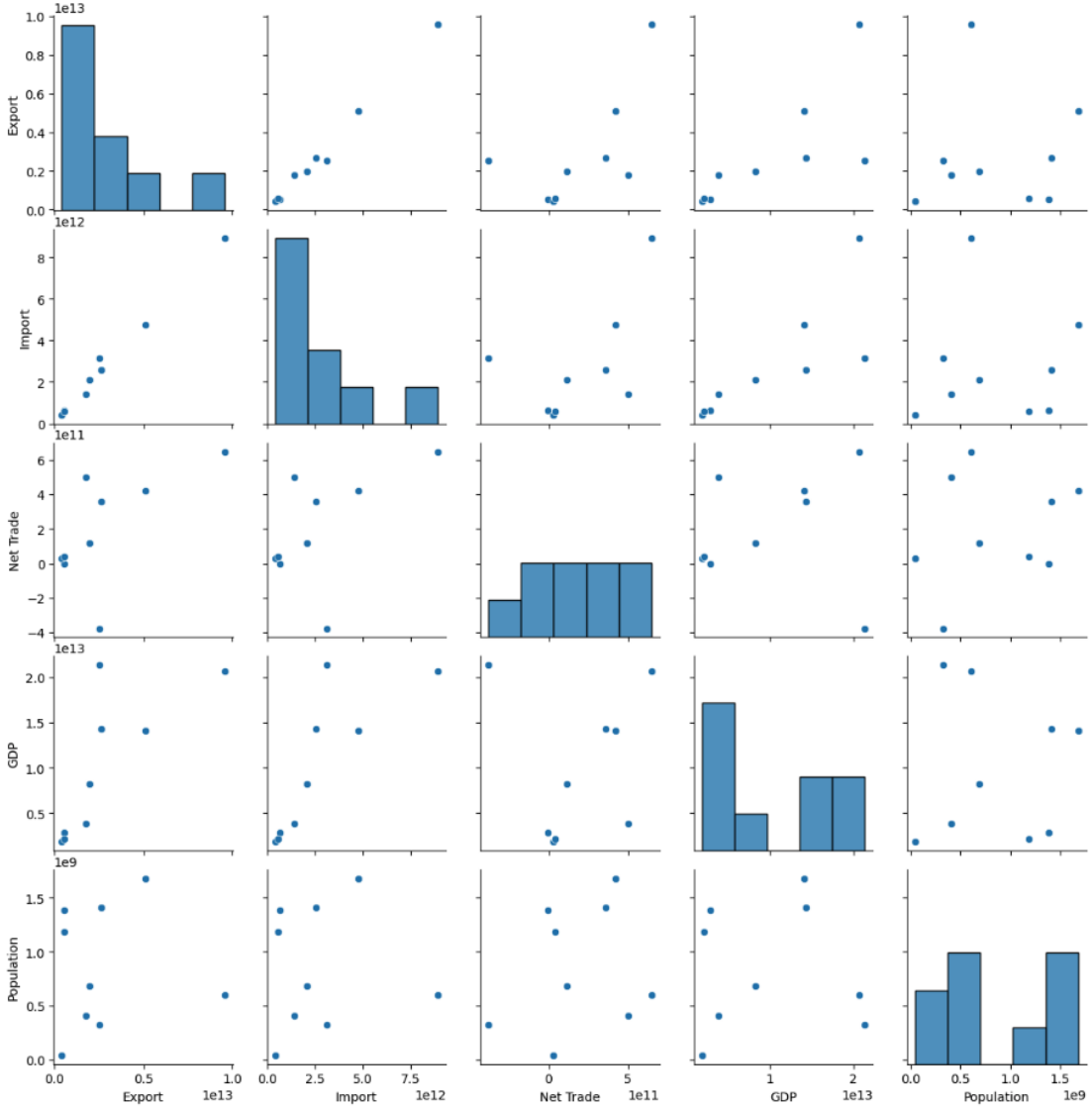
Задача предсказания на основе признаков

Более интересной оказалась задача предсказания доли переработанного пластика в конкретный год при известных значениях всех перечисленных выше признаков. В качестве первого шага к её решению средствами библиотеки Seaborn языка Python были построены графики зависимостей значений признаков друг от друга с целью выявить корреляции между ними и таким образом уменьшить размерность пространства признаков. Были взяты максимальные значения по каждому из регионов.

На рисунке на следующем слайде видно, что коррелируют между собой такие признаки как экспорт и импорт, следовательно, один из них можно исключить, тем самым уменьшив количество признаков до 8.

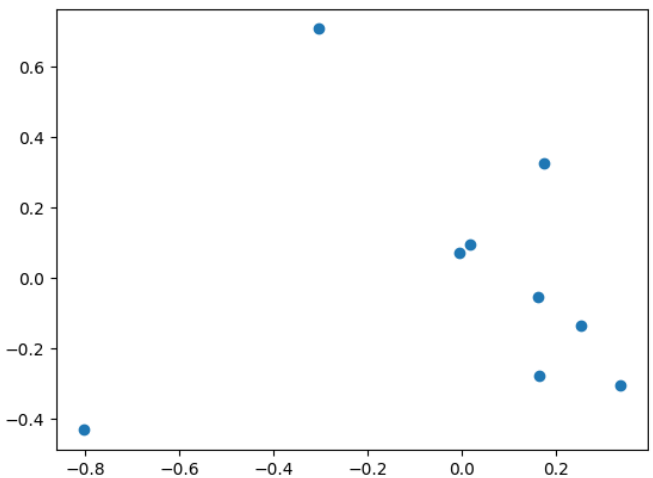
Для удобства чтения не все признаки показаны на рисунке, но визуально было проверено, что попарная корреляция между оставшимися недостаточно велика, чтобы выбросить ещё что-то.

Задача предсказания на основе признаков. Визуализация данных



Задача предсказания на основе признаков. Визуализация данных

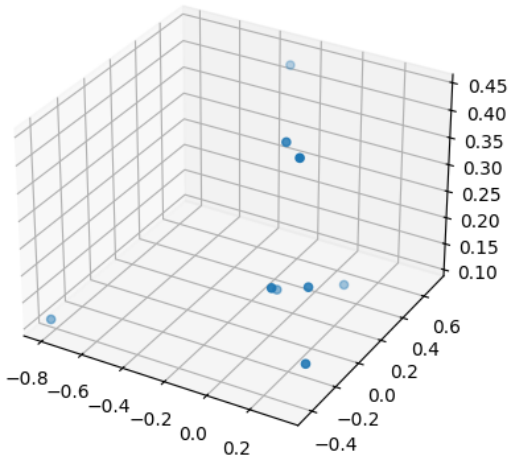
В качестве модели машинного обучения для решения задачи была выбрана модель линейной регрессии.
Для визуализации данных с помощью метода PCA размерность пространства признаков была понижена до двух (доля объяснённой дисперсии первыми двумя компонентами $\approx 70\%$).



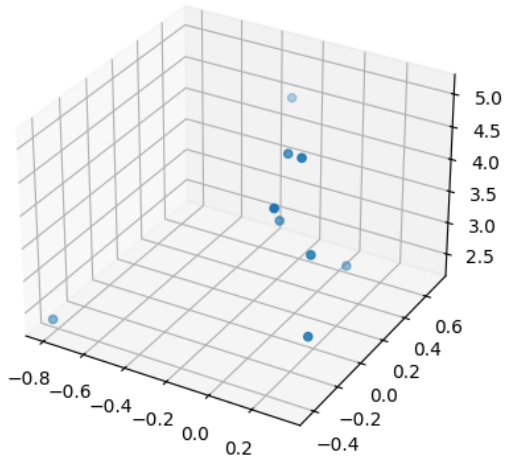
Principal components

Задача предсказания на основе признаков. Линейный регрессор

Затем были построены два трехмерных графика - тангенса и свободного члена от главных компонент.

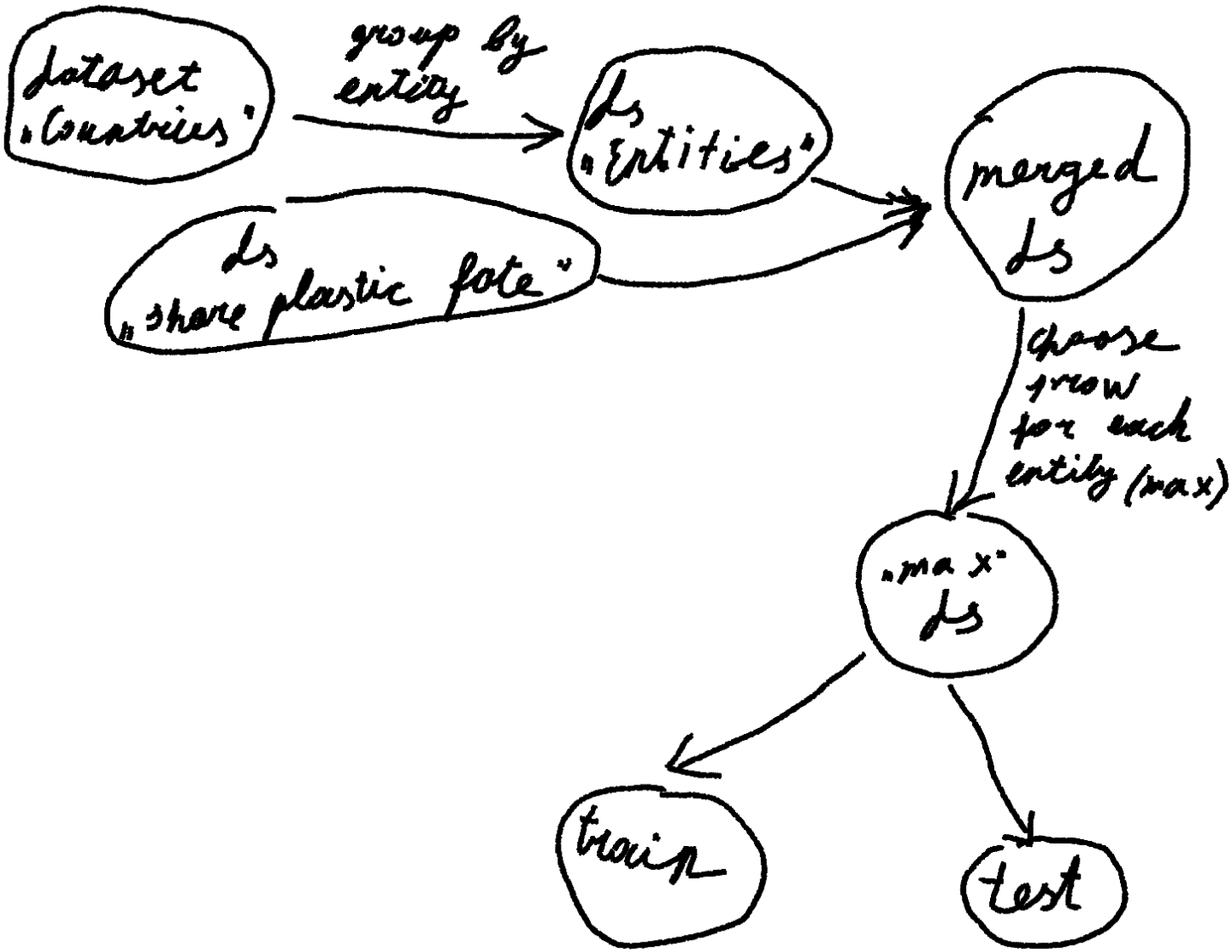


Tangent on principal components



Bias on principal components

Пайплайн


dataset "Countries"
group by entity
ds "Entities"
ds "share plastic fate"
merged ds
choose 1 row for each entity (max)
"max" ds
train
test

Линейный регрессор

Мы получили выборку из 9 точек (по одной для каждой части света). Для малого количества точек не имеют смысла сложные модели, поэтому снова выбрана линейная регрессия. Для обучения модели линейной регрессии 6 из этих точек (Америка без США, Азия без Китая и Индии, Европа, Ближний Восток и Северная Африка, Океания, США) составили обучающую выборку, а три оставшихся (Китай, Индия, Африка ниже Сахары) - тестовую. Разделение осуществлялось случайным образом.
Модель предсказывала значения тангенса (*tan*) и свободного члена (*bias*) для прямой, а искомое значение программа считала по формуле $share = tan \cdot year + bias$.

Линейный регрессор. Результаты

В качестве метрики качества предсказанного значения использовались суммы квадратов отклонений предсказанных значений тангенса и угла наклона от истинных.

```
[34]: metric_tan = sum((predicted_tans - test_df["Tan"]) ** 2)
      metric_tan

[34]: 0.36047065189647515

[35]: metric_biases = sum((predicted_biases - test_df["Bias"]) ** 2)
      metric_biases

[35]: 16.24745408891496
```

Можно видеть, что выбранная модель достаточно хорошо предсказывает значение угла наклона линии регрессии, но плохо предсказывает значение ординаты точки пересечения прямой с осью ординат. Следовательно, для уточнения предсказания стоит знать значение доли переработанного пластика в один конкретный год, а узнать это значение используя рассмотренные признаки не представляется возможным.

Спасибо за внимание!

Код программы доступен в репозитории
github.com/paxwritescode/machine_learning