Машинное обучение для оптимизации энергозатрат производственного процесса

Специализация «Data Scientist»

Александр Иванов
Senior Software Engineer

Александр Иванов

Senior Software Engineer at Wizata

О спикере

- Разработка программного обеспечения
- Опыт 7+ лет
- Azure, AWS clouds, IoT SOA, Microservices
- C#, .Net, Python, JS, etc

Аккаунты в соцсетях:

linkedin.com/in/alexdotnet

t.me/AlexDotNet

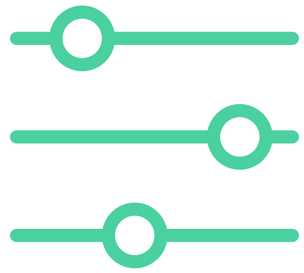
План проекта

1. Что такое Industry 4.0
2. Анализ исходных данных
3. Обучение и сравнение алгоритмов
4. Анализ полученных результатов
5. Выводы и заключение
6. Список источников

Industry 4.0

Что и зачем?

1



1st

Steam-based Machines

18th Century

2nd

Electrical Energy-based Mass Production

19th - 20th Century

3rd

Computer and Internet-based Knowledge

Late 20th Century

The 4th Industrial Revolution

Artificial Intelligence
Information Technology

Intelligence + Information

A.I.

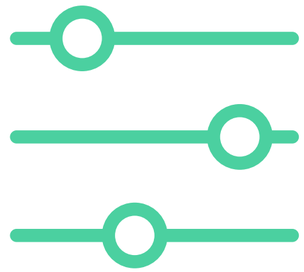
Big Data
IoT
Cloud

Early 21st Century
The Confluence and Convergence of Emerging Technologies

Анализ исходных данных

Exploratory Data Analysis (EDA)

2



Признаки:

Контекстные:

- selenium
- carbon
- manganese
- silicium
- dinitrogen

Управляющие:

- temperature_zone1
- temperature_zone2
- temperature_zone3
- temperature_zone4
- temperature_zone5
- temperature_zone6
- oxygen_zone123
- oxygen_zone456
- hydrogen_zone123
- hydrogen_zone456

Target

Бинарная классификация:

- 0 – неэффективное, высокое энергопотребление
- 1 – эффективное, низкое энергопотребление

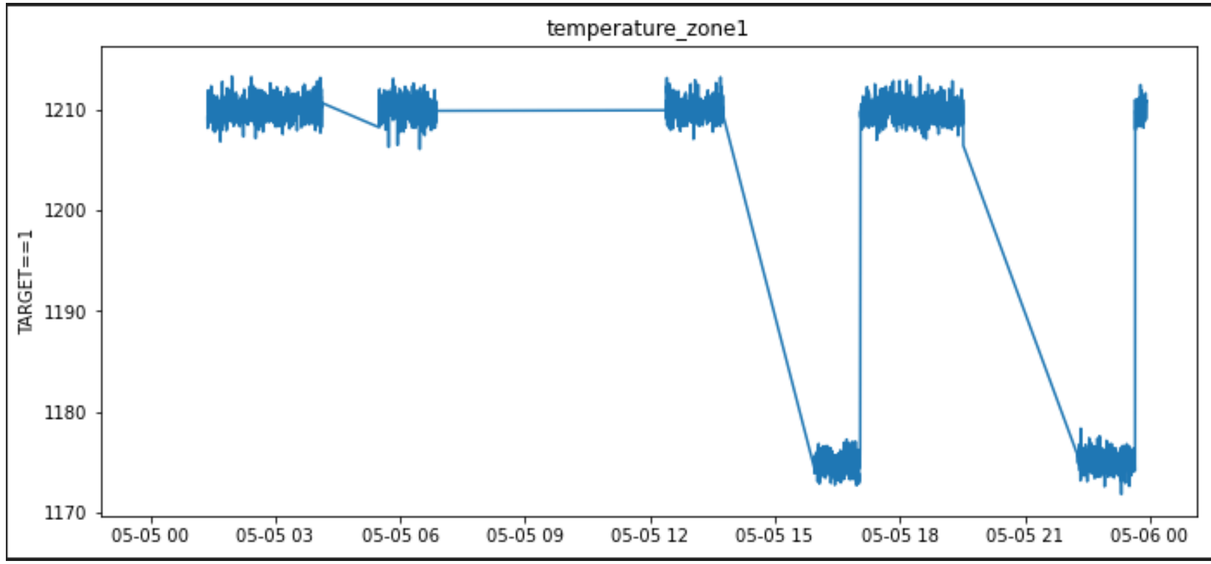
Dataset info

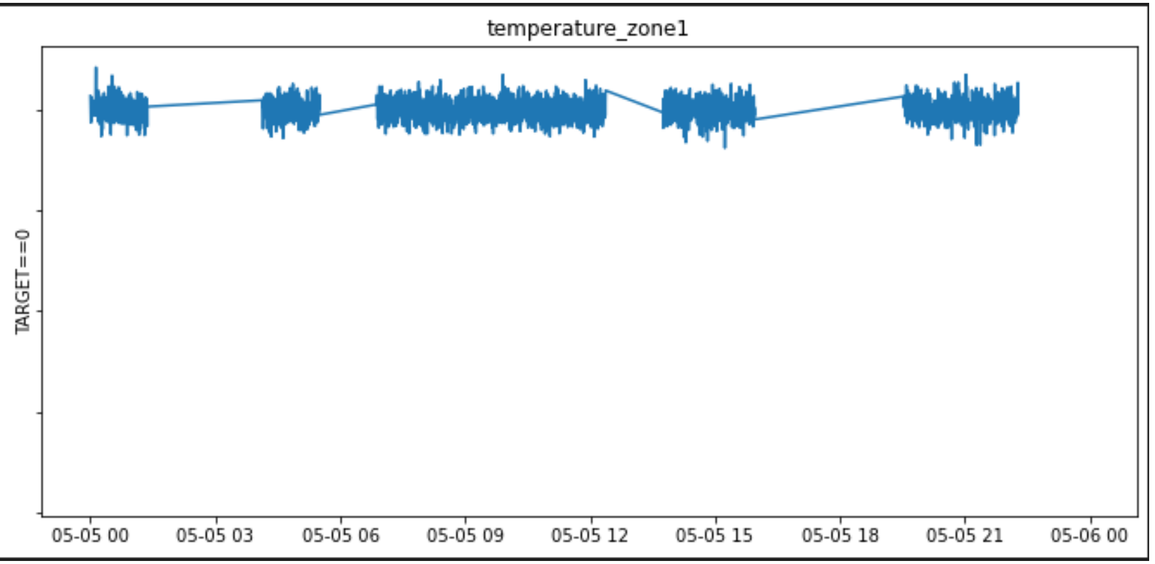
- Outliers – ok
- NaN values – ok
- DataTypes – ok
- Sorted - ok

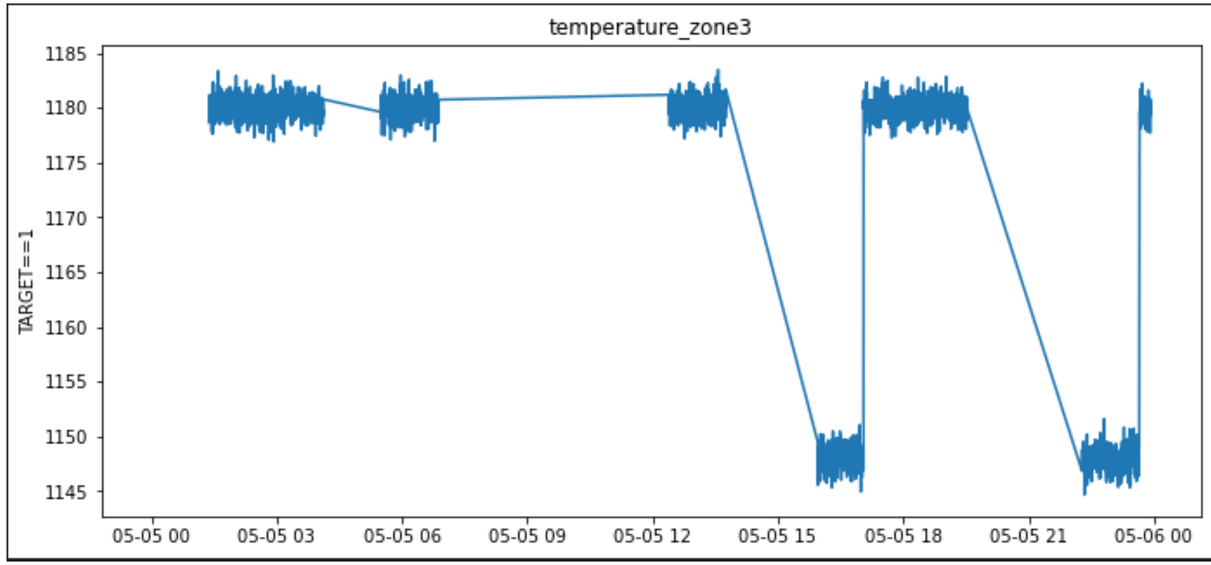
```
1  # Check data for gaps, nulls and datatypes. Looks all ok, no preparations needed
2  df.info()
```

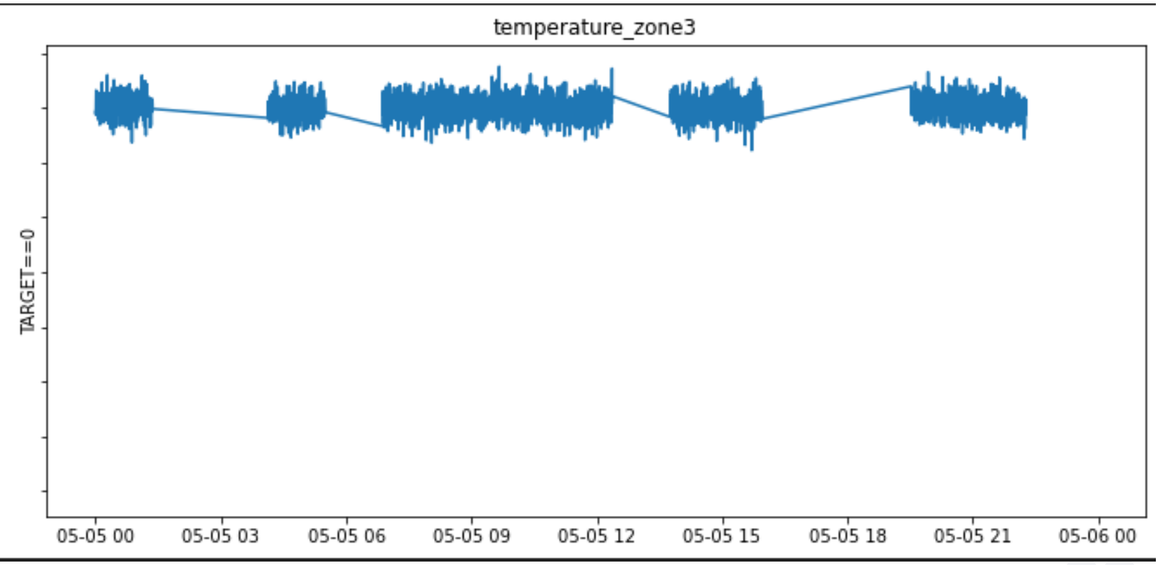
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 5742 entries, 2022-05-05 00:00:00 to 2022-05-05 23:55:15
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   selenium           5742 non-null   float64
 1   carbon             5742 non-null   float64
 2   manganese          5742 non-null   float64
 3   silicium           5742 non-null   float64
 4   dinitrogen         5742 non-null   int64  
 5   temperature_zone1  5742 non-null   float64
 6   temperature_zone2  5742 non-null   float64
 7   temperature_zone3  5742 non-null   float64
 8   temperature_zone4  5742 non-null   float64
 9   temperature_zone5  5742 non-null   float64
 10  temperature_zone6  5742 non-null   float64
 11  oxygen_zone123     5742 non-null   float64
 12  oxygen_zone456     5742 non-null   float64
 13  hydrogen_zone123   5742 non-null   float64
 14  hydrogen_zone456   5742 non-null   float64
 15  TARGET             5742 non-null   int64  
dtypes: float64(14), int64(2)
memory usage: 762.6 KB
```

Зависимость зон от таргета: 1 в 1

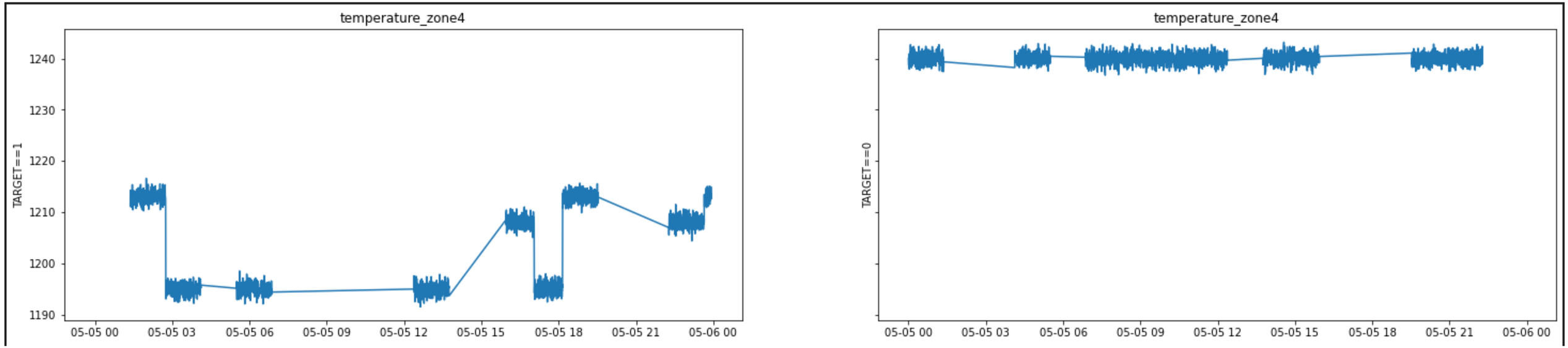


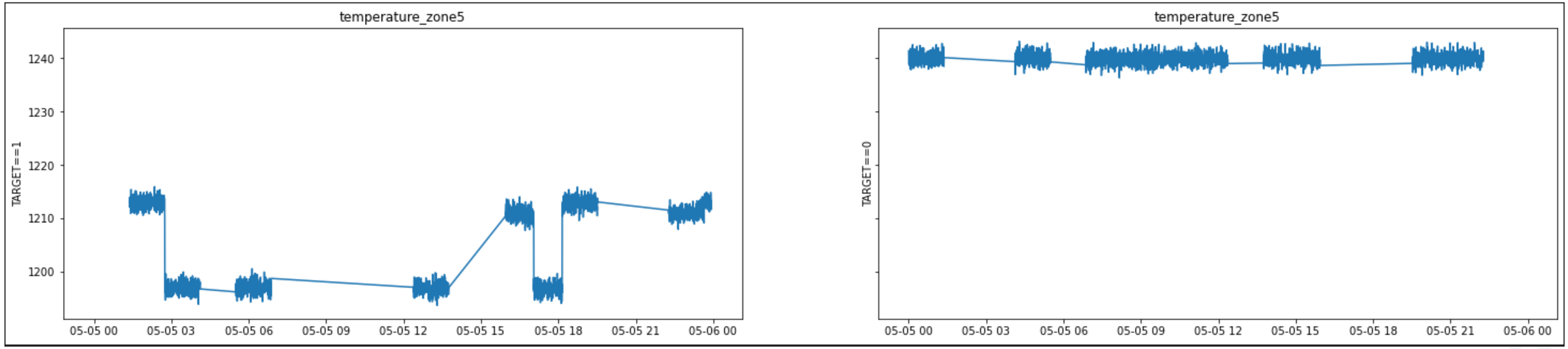




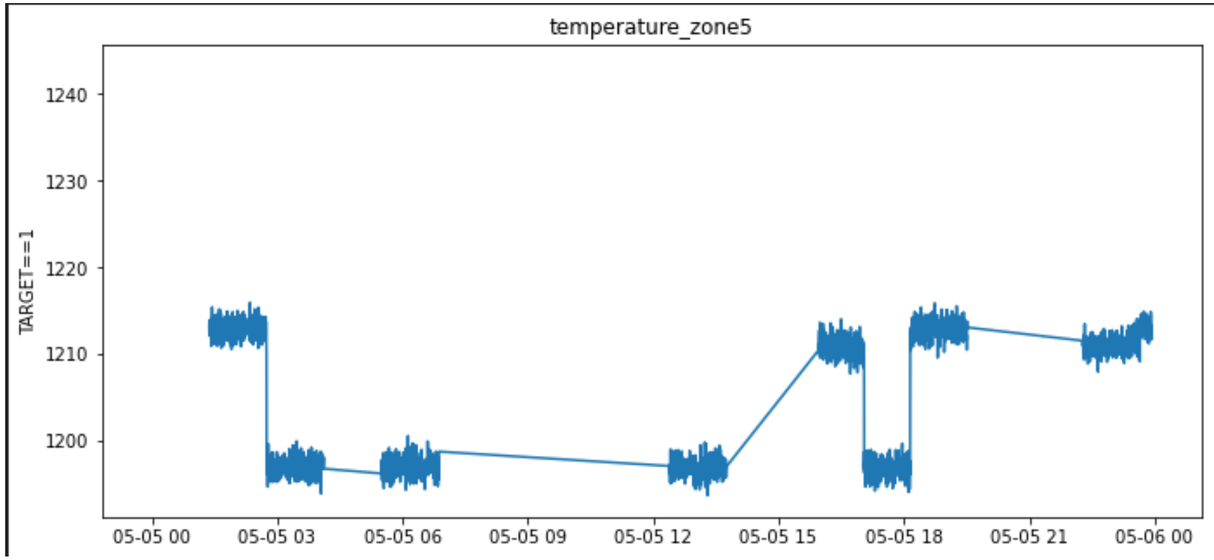


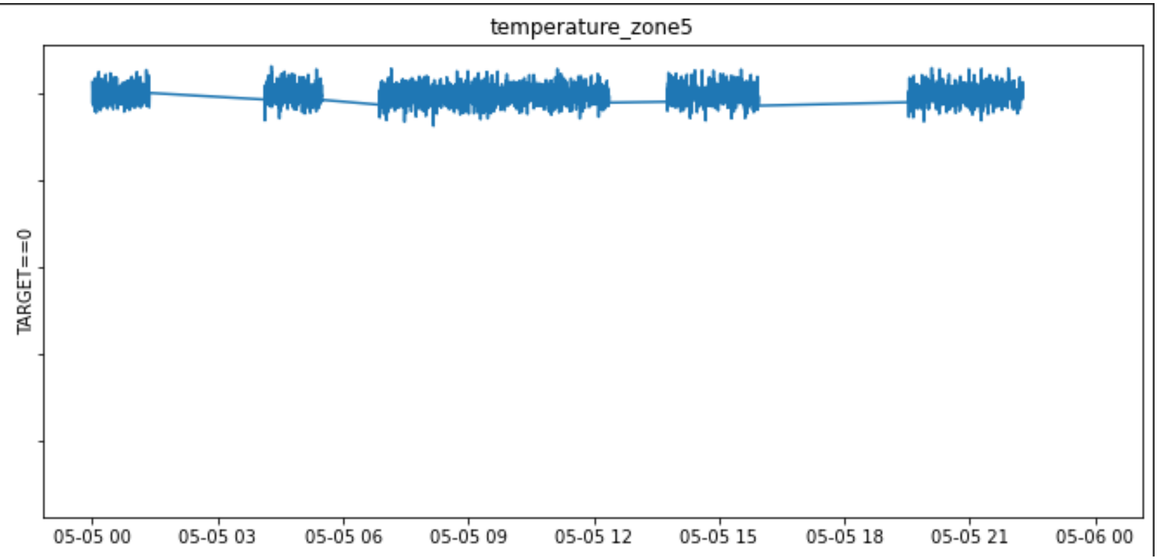
Зависимость зон от таргета: 1 в 1

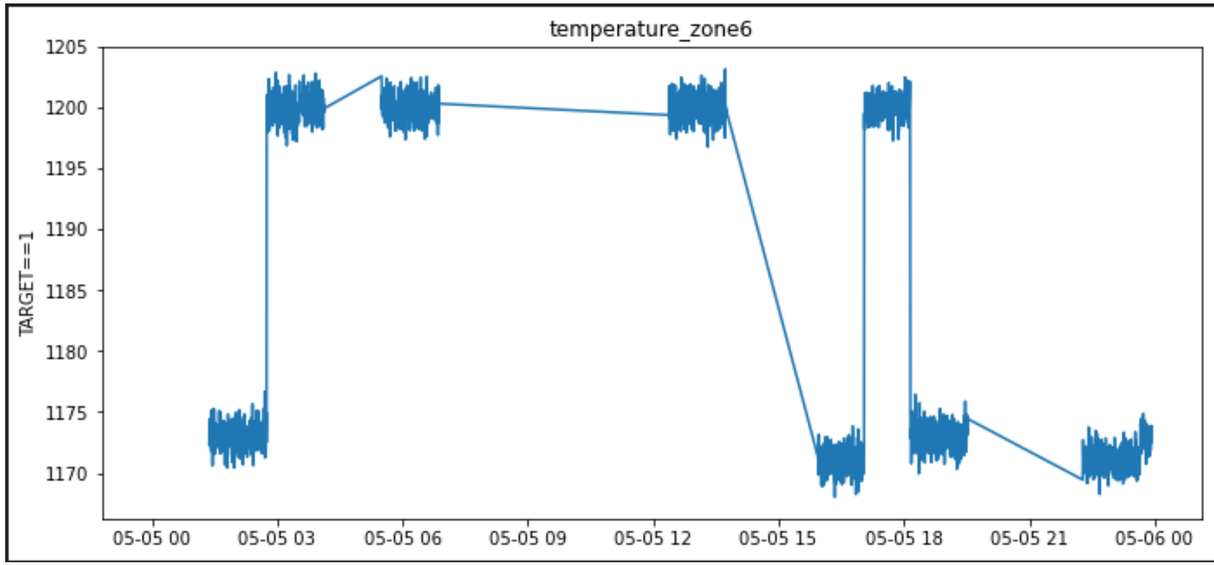


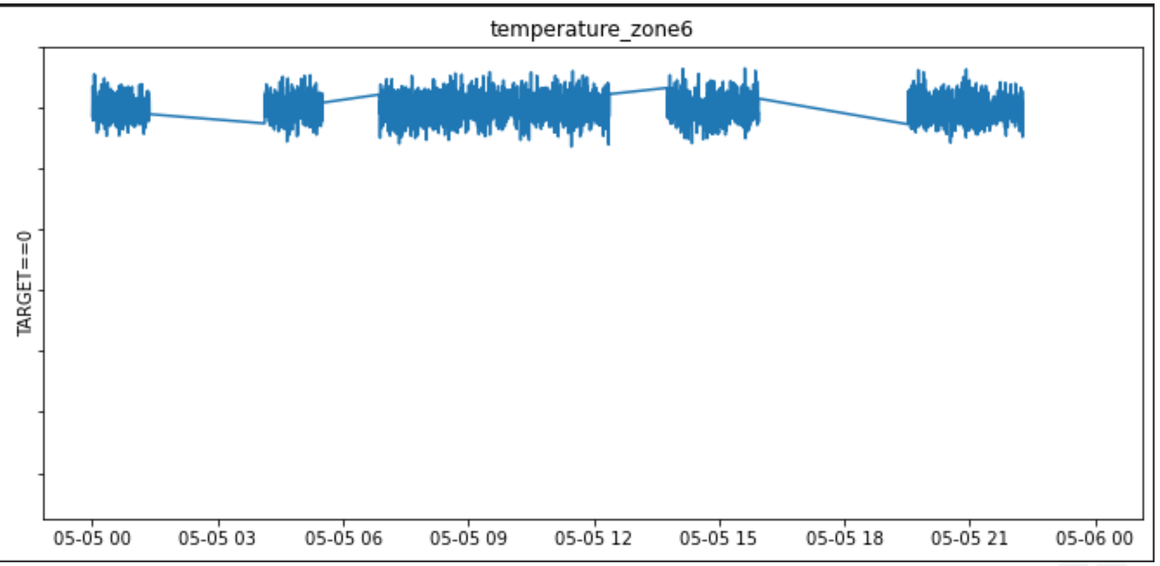


Зависимость зон от таргета: обратная



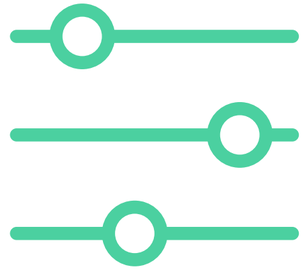






Обучение алгоритмов

3



Логистическая регрессия (L1 регуляризация)

L1 отбрасывает ненужные признаки,
путём установки нулевых весов не значимым признакам

```python
# Try to use different coefficients
C = [10, 1, .1, .001]

for c in C:
    clf = LogisticRegression(penalty='l1', C=c, solver='liblinear')
    clf.fit(X_train, y_train)
    print('C:', c)
    print('Coefficient of each feature:', clf.coef_)
    print('Training accuracy:', clf.score(X_train_std, y_train))
    print('Test accuracy:', clf.score(X_test_std, y_test))
    print('')

    predictions = clf.predict_proba(X_test_std)
    print('Predictions:')
    print(predictions)

```

Логистическая регрессия: Результат

- Training accuracy: 0.8601642199552127
- Test accuracy: 0.8572257690075 45

Не плохо… Но попробуем улучшить с помощью дерева решений

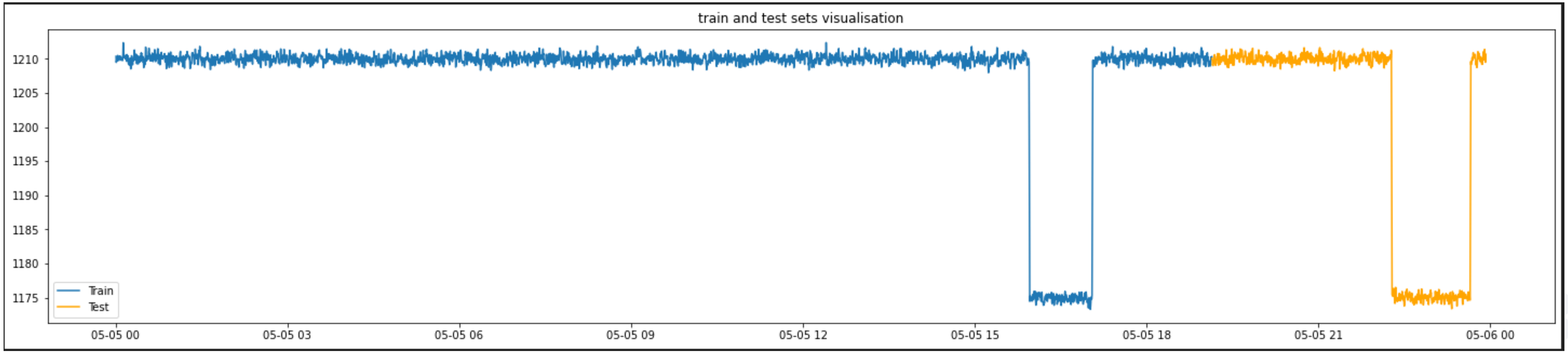
DecisionTreeClassifier

Почему дерево?

Аналогично IF...ELSE разберёт всё на правила, которые можно будет извлечь из полученных данных, в результате увидим успешную комбинацию параметров зон

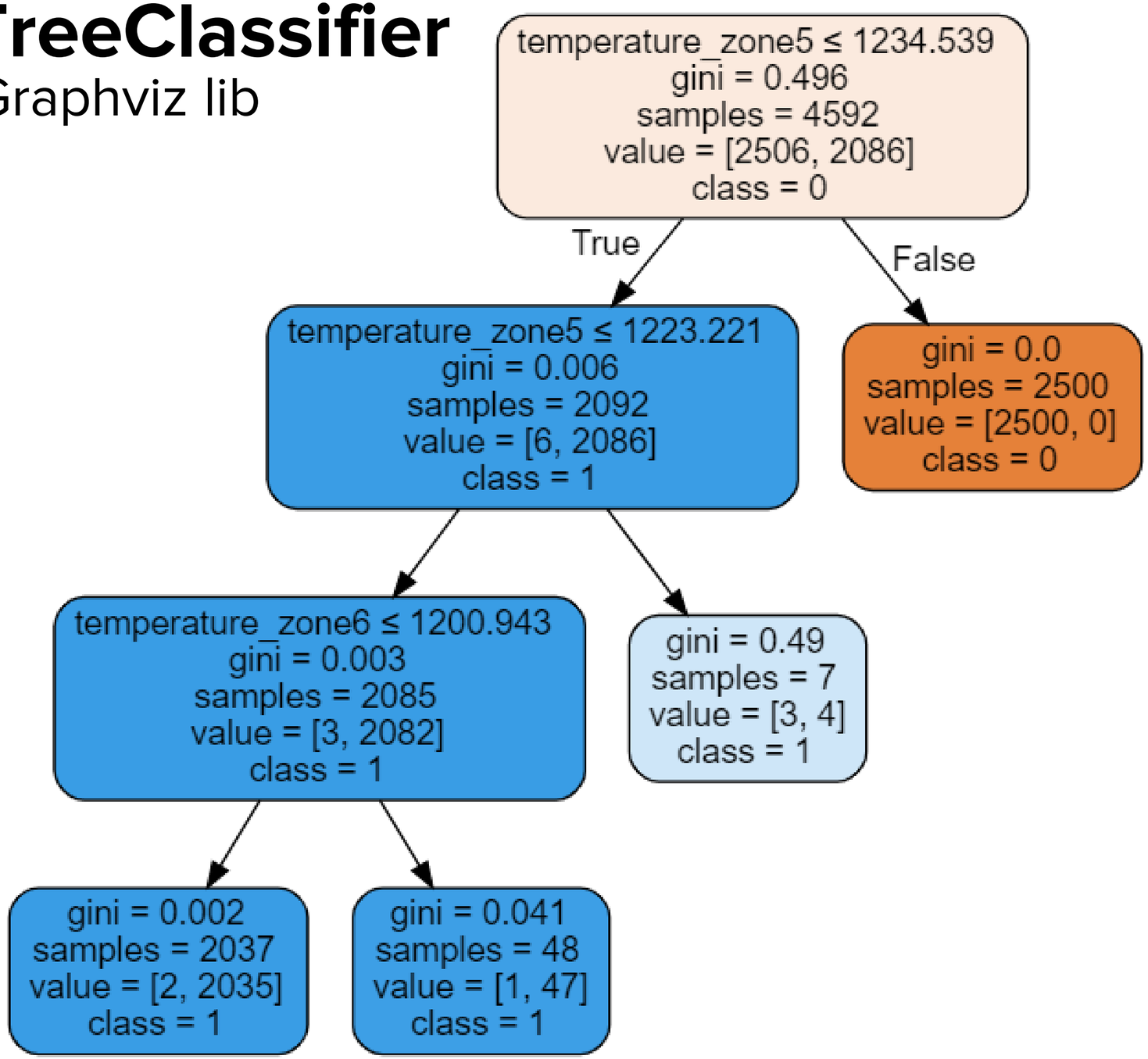
DecisionTreeClassifier

- StandardScaler не нужен
- Трейн и тест делим по времени (80% - train)
- Среднее значение с окном 3 вместо чистых значений
- Обучаем DecisionTreeClassifier на этих значениях



DecisionTreeClassifier

plotting with Graphviz lib



DecisionTreeClassifier

textual representation

```
|--- temperature_zone5 <= 1234.54
|   |--- temperature_zone5 <= 1223.22
|   |   |--- temperature_zone6 <= 1200.94
|   |   |   |--- class: 1
|   |   |--- temperature_zone6 >  1200.94
|   |   |   |--- class: 1
|   |--- temperature_zone5 >  1223.22
|   |   |--- class: 1
|--- temperature_zone5 >  1234.54
|   |--- class: 0
```

DecisionTreeClassifier

classification_report

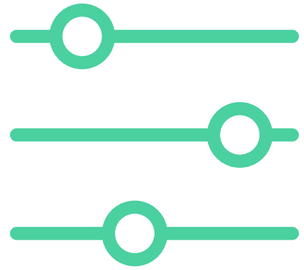
```
Classification report:
Accuracy:  0.9982578397212544
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       658
           1       1.00      1.00      1.00       490

    accuracy                           1.00      1148
   macro avg       1.00      1.00      1.00      1148
weighted avg       1.00      1.00      1.00      1148
```

Анализ полученных результатов

4



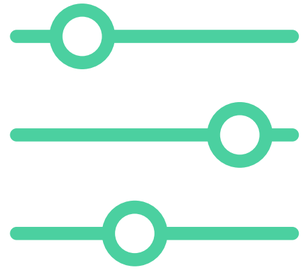
Результат

- На целевую переменную имеет влияние только 5я зона

- Для эффективного использования энергозатрат необходимо:
  - Температура зоны 5 должна быть ниже значения 1234.54 но выше значения 1223.22

Выводы

5



В итоге

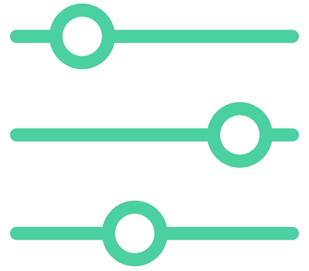
- Модель DecisionTreeClassifier подходит лучше чем LogisticRegression для задач оптимизации энергопотребления

- Было бы хорошо добавить метрики качества производимого продукта, т.к. нахождение оптимальных значений энергопотребления ничего нам не говорит о качестве продукта на выходе. Информация возможно уже заложена в Таргет, но мы об этом ничего не знаем.

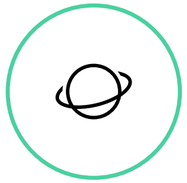
- Возможно, не хватило данных для выявления дополнительных кейсов.

Заключение

6

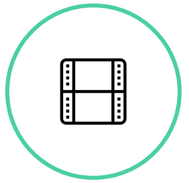


В ходе работы был получен ряд результатов:



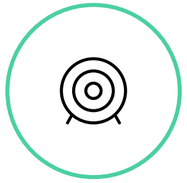
Произведён анализ исходных данных

Произведено сравнение результатов модели логистической регрессии и дерева решений

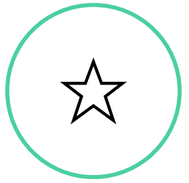


Произведен отбор наиболее значимых признаков

Реализована модель машинного обучения, которая выявила определённые закономерности в данных



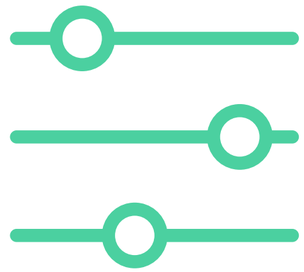
Получено достаточно информации которая позволит оптимизировать энергозатраты производства



Необходимо учитывать метрики качества продукции при оптимизации энергозатрат

Список источников

6



1. kontron.com: THE FUTURE IS NOW - ARTIFICIAL INTELLIGENCE FOR INDUSTRY 4.0
2. parametrix.com: The 4th Industrial Revolution: Where are we in history?
3. sap.com: Что такое «Индустрия 4.0»?
4. machinelearning.ru: Воронцов К. Лекции по логическим алгоритмам классификации.
5. machinelearningknowledge.ai: Decision Tree Classifier
6. scikit-learn.org: LogisticRegression
7. scikit-learn.org: DecisionTreeClassifier

Спасибо за внимание!

Вопросы?

Александр Иванов
Senior Software Engineer

in linkedin.com/in/alexdotnet
t.me/AlexDotNet