Прогнозирование почасового потребления электроэнергии

#Draft

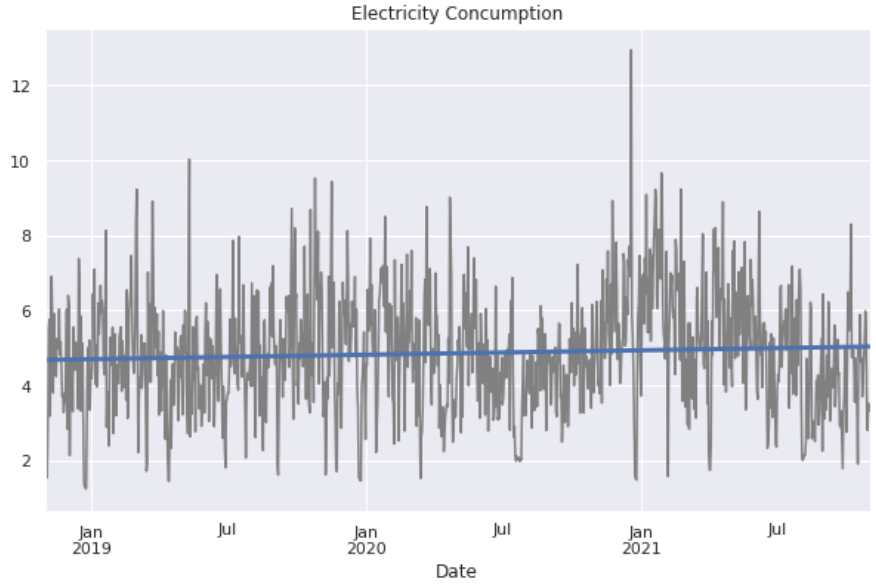
Степан Васильев

Исходные данные

- Датасет о потреблении электроэнергии в студенческом кампусе одного из университетов Дании

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26328 entries, 0 to 26327
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Datetime              26328 non-null  object
 1   kWh                   26328 non-null  float64
 2   hour                  26328 non-null  float64
 3   day_of_month          26328 non-null  float64
 4   day_of_week           26328 non-null  float64
 5   month                 26328 non-null  float64
 6   is_weekend            26328 non-null  int64
 7   pressure_at_sea       26323 non-null  float64
 8   precip_dur_past10min  26323 non-null  float64
 9   wind_dir              26323 non-null  float64
 10  wind_speed            26323 non-null  float64
 11  temp_dew              26323 non-null  float64
 12  pressure              26323 non-null  float64
 13  visib_mean_last10min  26323 non-null  float64
 14  temp_dry              26323 non-null  float64
 15  humidity              26323 non-null  float64
 16  cloud_cover           26323 non-null  float64
 17  visibility            26323 non-null  float64
dtypes: float64(16), int64(1), object(1)
memory usage: 3.6+ MB
```



Мотивация

- Решение задач прогнозирования потребления электроэнергии значительно упрощает планирование бюджета организации.

- Прогнозы могут быть также использованы для расчета необходимой к установке локальной генерации на базе возобновляемых источников энергии.

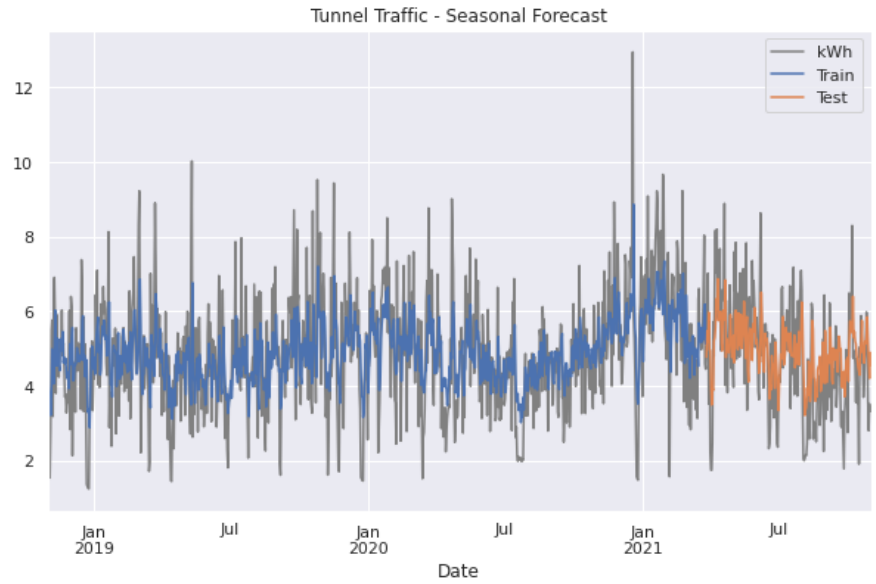
Задачи и цели

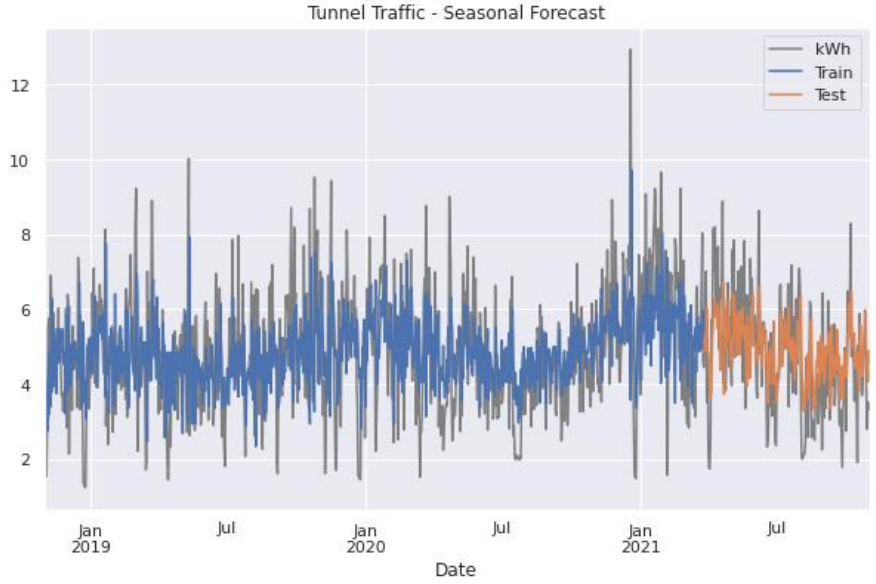
- Провести разведочный анализ датасета.

- Применить различные методы для прогнозирования: LinearRegression, XGBRegressor, ARIMA, SARIMA, ARIMAX. Сравнить модели по метрикам MSE, RMSE.

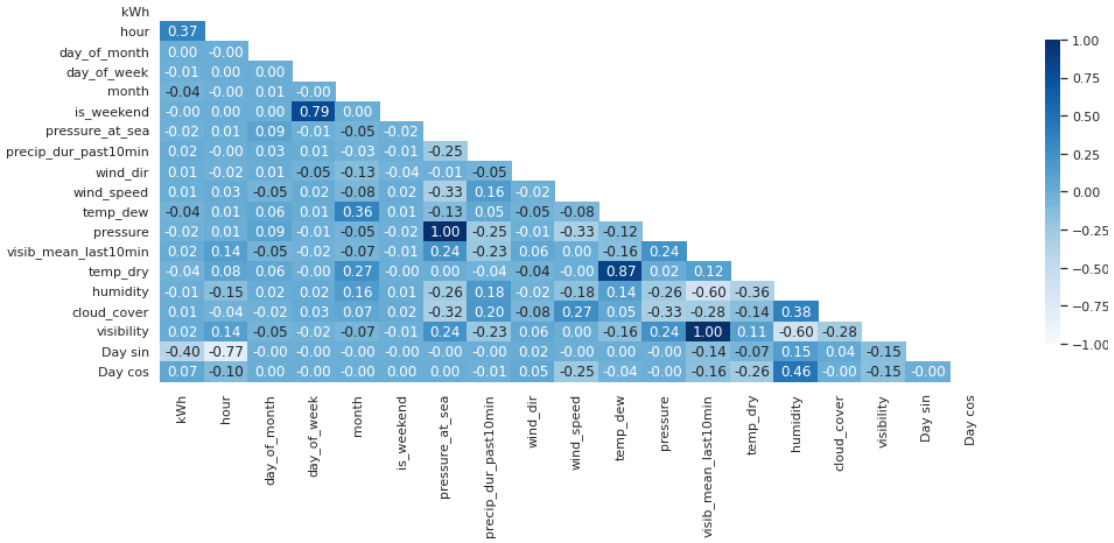
- Применить несколько вариантов искажения датасета и посмотреть, как изменится прогноз.

Задачи и цели





EDA



Тепловая карта корреляции признаков основана на коэффициенте корреляции Пирсона. Значения ниже 0,8 и выше -0,8 свидетельствуют об отсутствии корреляции.

Построение моделей

Результаты ADF-теста показывают, что целевой временной ряд стационарен.

```
Augmented Dickey-Fuller Test:
ADF test statistic     -1.365825e+01
p-value                 1.538983e-25
# lags used             4.800000e+01
# observations          2.625500e+04
critical value (1%)    -3.430599e+00
critical value (5%)    -2.861650e+00
critical value (10%)   -2.566829e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Выбор лучших моделей для прогнозирования.

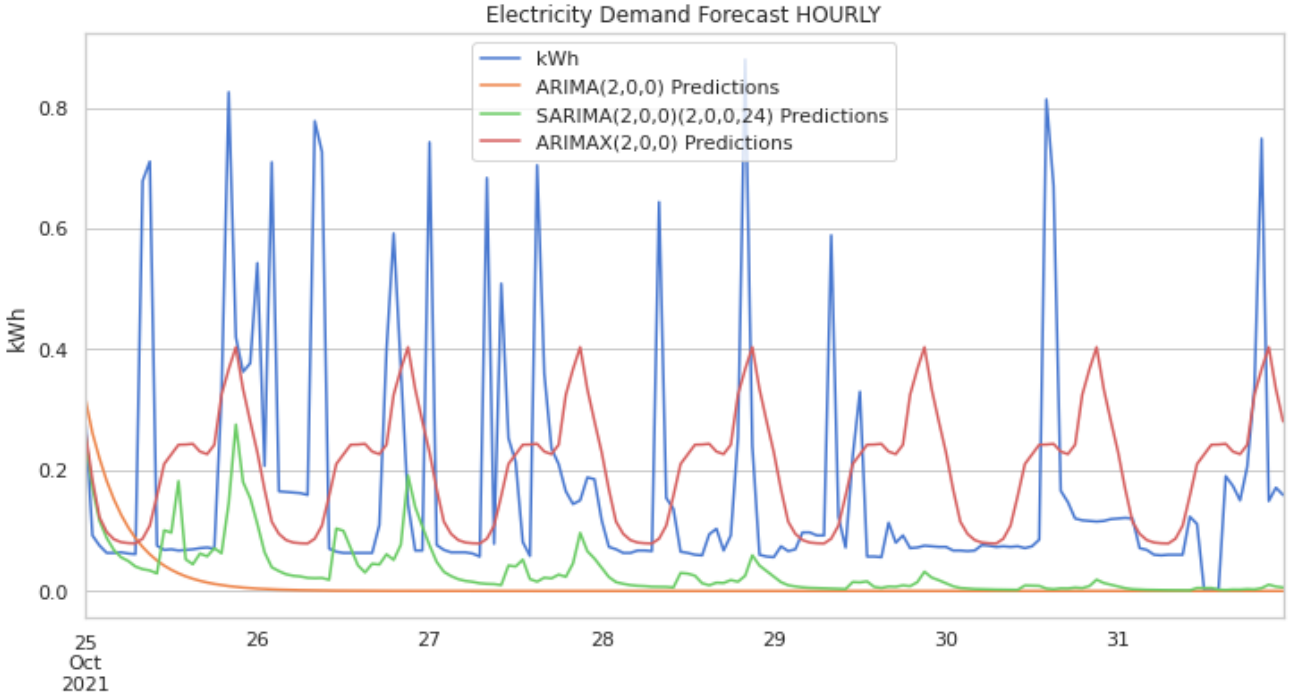
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	744
Model:	SARIMAX(2, 0, 0)x(2, 0, 0, 24)	**Log Likelihood**	122.285
Date:	Mon, 27 Jun 2022	**AIC**	-232.570
Time:	07:34:43	**BIC**	-204.898
Sample:	0	**HQIC**	-221.903
	- 744		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0930	0.018	5.118	0.000	0.057	0.129
ar.L1	0.2494	0.028	8.968	0.000	0.195	0.304
ar.L2	0.1221	0.029	4.208	0.000	0.065	0.179
ar.S.L24	0.1232	0.036	3.439	0.001	0.053	0.193
ar.S.L48	0.1039	0.034	3.014	0.003	0.036	0.171
sigma2	0.0421	0.001	31.525	0.000	0.039	0.045

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	7074.65
Prob(Q):	0.98	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.66	**Skew:**	3.10
Prob(H) (two-sided):	0.00	**Kurtosis:**	16.78

Результаты прогноза на неделю

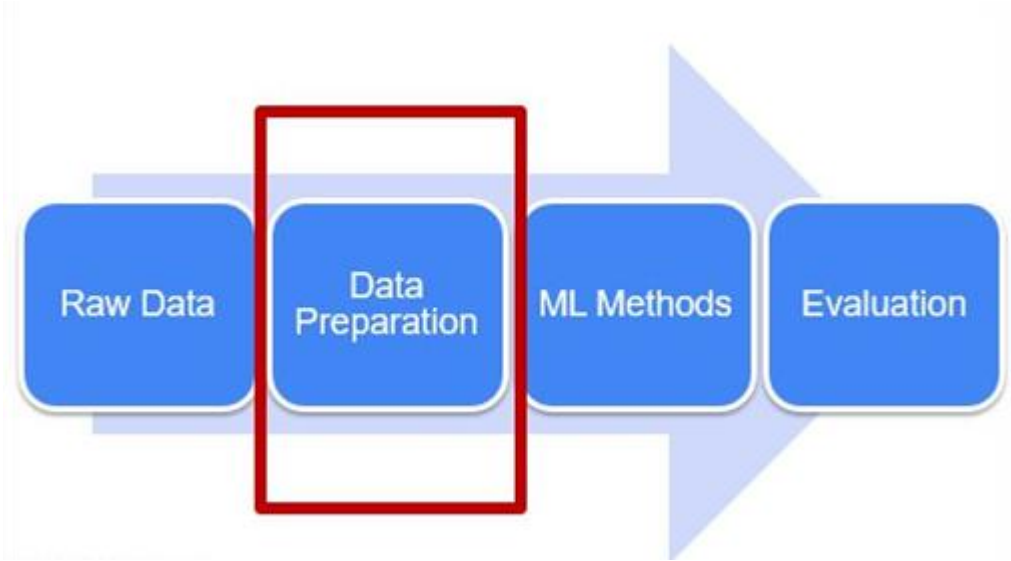



```
ARIMA(2,0,0) MSE Error: 0.06903767671
ARIMA(2,0,0) RMSE Error: 0.2627502173
ARIMAX(2,0,0)  MSE Error: 0.04401179512
ARIMAX(2,0,0)  RMSE Error: 0.2097898833
SARIMA(2,0,2)(2,0,0,24) MSE Error: 0.05949651117
SARIMA(2,0,2)(2,0,0,24) RMSE Error: 0.2439190668
```

Модель SARIMA показала самый точный прогноз электропотребления

Исследование искажений данных

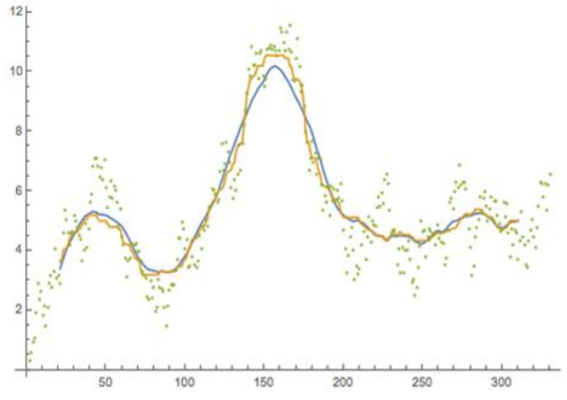
- Реальные данные очень редко бывают идеальными.
- Исследовано два варианта искажения датасета: добавление шума SNR=0.2,



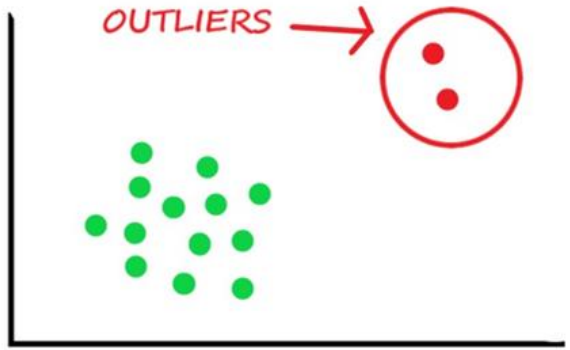
Наиболее важной задачей машинного обучения является подготовка данных

Искажения данных

Шум



Выбросы



Пропущенные значения

Добавление шума

Аддитивный белый гауссовский шум (AWGN)
с заданным отношением сигнала к шуму (SNR)

Итерация по целевому признаку kWh:
- Вычисление «мощности» сигнала;
- Добавление шума с мощностью, соответствующей определенному SNR.

Пропуск значений в датасете и заполнение пропусков

Missing Completely at Random (MCAR)

$$P(R|D^m, D^o) = P(R)$$

- Итерация по всем строкам и столбцам
- Пропуск записей с вероятностью 0.25

Пропуск значений в датасете и заполнение пропусков

Missing Not at Random (MNAR)

$$P(R|D^m, D^o) = P(R|D^m)$$

- Записи отбрасываются, если они меньше (больше) медианного значения

- Заполнение пропусков нулями.
- Заполнение пропусков средними значениями.
- Заполнение пропусков медианными значениями.

Пропуск значений в датасете и заполнение пропусков

Исходные данные

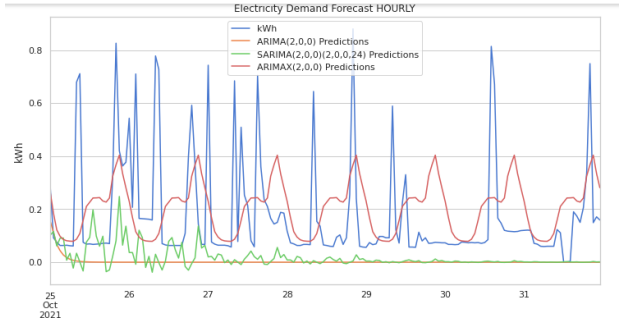
```
ARIMA(2,0,0) MSE Error: 0.06903767671
ARIMA(2,0,0) RMSE Error: 0.2627502173
ARIMAX(2,0,0)  MSE Error: 0.04401179512
ARIMAX(2,0,0)  RMSE Error: 0.2097898833
SARIMA(2,0,2)(2,0,0,24) MSE Error: 0.05949651117
SARIMA(2,0,2)(2,0,0,24) RMSE Error: 0.2439190668
```

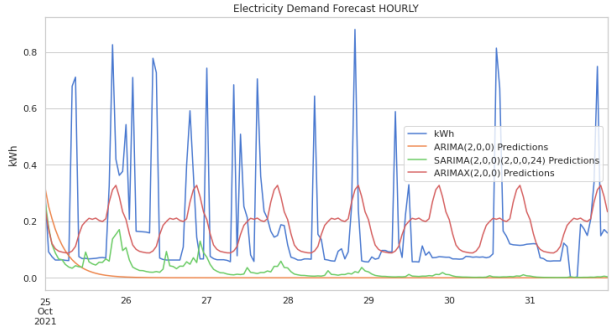
Данные с шумом

```
ARIMA(2,0,0) MSE Error: 0.07015438496
ARIMA(2,0,0) RMSE Error: 0.2648667306
ARIMAX(2,0,0)  MSE Error: 0.04401179512
ARIMAX(2,0,0)  RMSE Error: 0.2097898833
SARIMA(2,0,2)(2,0,0,24) MSE Error: 0.06568838964
SARIMA(2,0,2)(2,0,0,24) RMSE Error: 0.2562974632
```

Данные с пропущенными значениями

```
ARIMA(2,0,0) MSE Error: 0.06920046347
ARIMA(2,0,0) RMSE Error: 0.2630598097
ARIMAX(2,0,0)  MSE Error: 0.04047380516
ARIMAX(2,0,0)  RMSE Error: 0.2011810258
SARIMA(2,0,2)(2,0,0,24) MSE Error: 0.06068002744
SARIMA(2,0,2)(2,0,0,24) RMSE Error: 0.2463331635
```





Выводы на основе текущих результатов

- Сезонная модель ARIMA показывает лучшую эффективность при прогнозировании. Рассмотренный датасет представляет собой потребление электричества в студенческом кампусе и, соответственно, имеет сильную зависимость от ежедневной, недельной, годовой сезонности.

- Использованные методы показали себя достаточно робастными в рамках выполненных искажений. Было бы интересно проверить различные уровни шума и методы пропуска значений в датасете.