

ATOMIC HACK

CONTINUE



НАША КОМАНДА



Магеррамов
Ешгин



Георгий
Андреянов

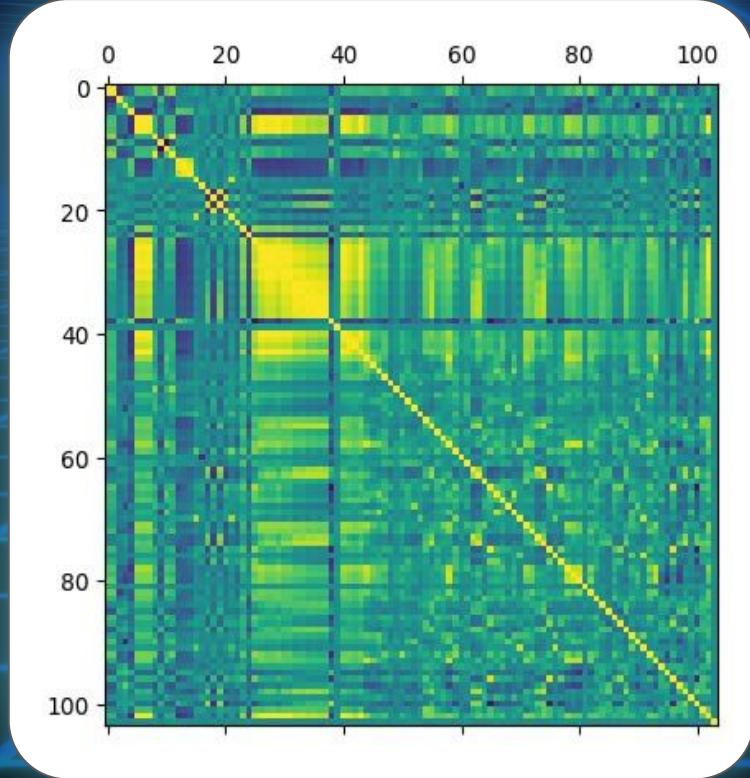


Когановский
Григорий



Носов
Артём

Анализ исходных данных



Корреляционная
матрица
признаков

Предобработка данных



```
# Detect useful features
```

```
useless_features +=  
list(set(remove_collinear_features(useful_features_dataset,0.9)))  
  
useful_features = list(set(features) - set(useless_features))
```

Предобработка данных



```
# Detect useless features

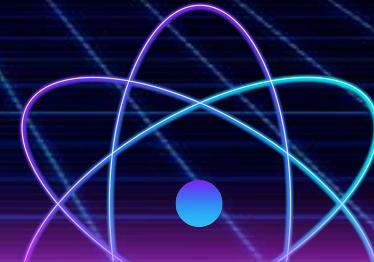
unique_values_dict = {}
for col in features_dataset:
    unique_values_dict[col] = features_dataset[col].unique()

useless_features = []
for key, value in unique_values_dict.items():
    if len(value) == 1:
        useless_features.append(key)
```

Преобразование данных



```
# scale data
scaler = StandardScaler()
y_train = scaler.fit_transform(y_train)
y_test = scaler.transform(y_test)
```



Построение и исследование модели



```
# fit → predict
from catboost import CatBoostRegressor
model = CatBoostRegressor()
model.fit(X_train, y_train)
pred = model.predict(X_test)
```

Результаты

RMSE = 0.00958

Альтернативные решения



RandomForest 10000 деревьев, RMSE=1.11

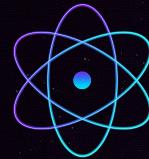


Baseline 200 эпох, RMSE=0.57



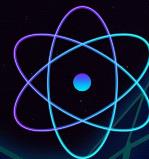
Линейная регрессия без предобработки,
RMSE=395266.85

Горизонты развития



Графовые представления молекул.

Например, в виде матрицы связности



Использование других моделей.

Например, графовых нейронных сетей

Q/A