Team #6

Course Project
26/10/2023

Prediction of Mechanical Properties of Steels

Team Project on the course “Introduction to Data Science” by **Kamil Garifullin**, **Pavel Bartenev**, and **Viktoriia Zinkovich**

Team VPK

V **Viktoriia Zinkovich**
Data Science
MS-1

P **Pavel Bartenev**
Data Science
MS-1

K **Kamil Garifullin**
Data Science
MS-1

Problem

Calculating the mechanical properties of steels requires **expensive experiments**

So our aim is to make a model that will predict the **properties of steels** and **reduce the cost** of steel research

Namely, based on the dataset with chemical composition of steels (%C, %Si, %Mn, %P, %S...) we want **to predict its tensile strength**

Dataset
Dataset
Dataset
Dataset
Dataset

Data pre-processing: search for outliers, missing data

Dataset
Dataset
Dataset
Dataset

Data

```
df = pd.read_csv('Steels_kaggle.csv')
```

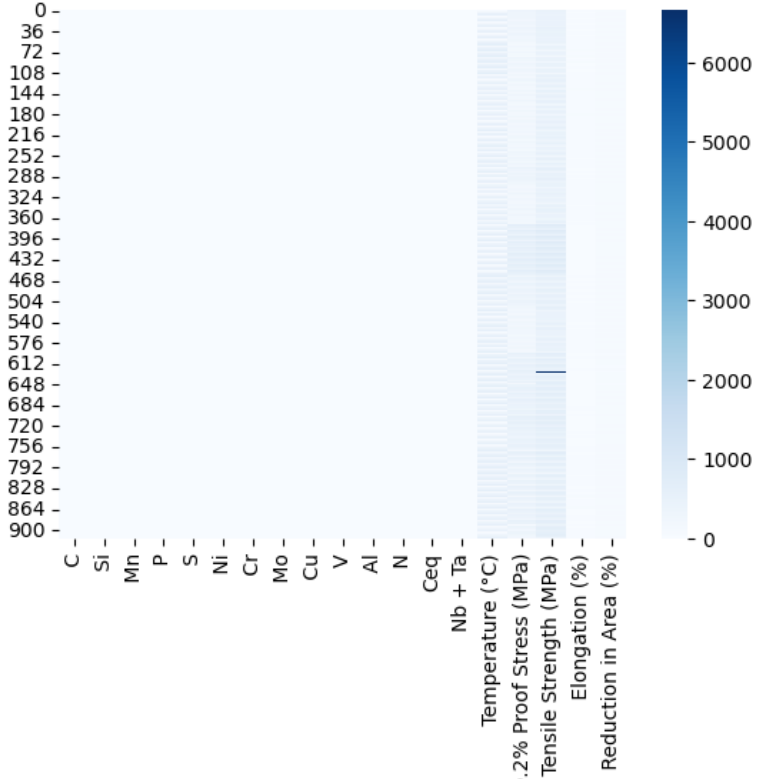
	Alloy code	C	Si	Mn	P	S	Ni	Cr	Mo	Cu	V	Al	N	Ceq	Nb + Ta	Temperature (°C)	0.2% Proof Stress (MPa)	Tensile Strength (MPa)	Elongation (%)	Reduction in Area (%)
0	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	27	342	490	30	71
1	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	100	338	454	27	72
2	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	200	337	465	23	69
3	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	300	346	495	21	70
4	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	400	316	489	26	79

915
steels

20
features

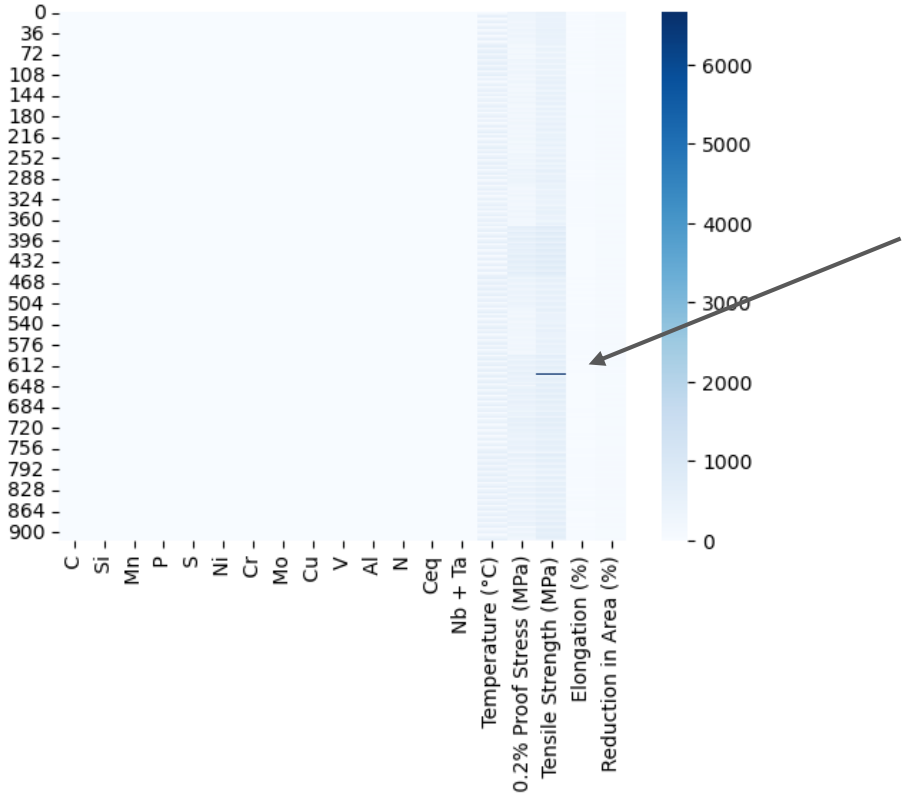
12
elements

Outliers




```
sns.heatmap(df[cols], cmap=color)
```

Outliers

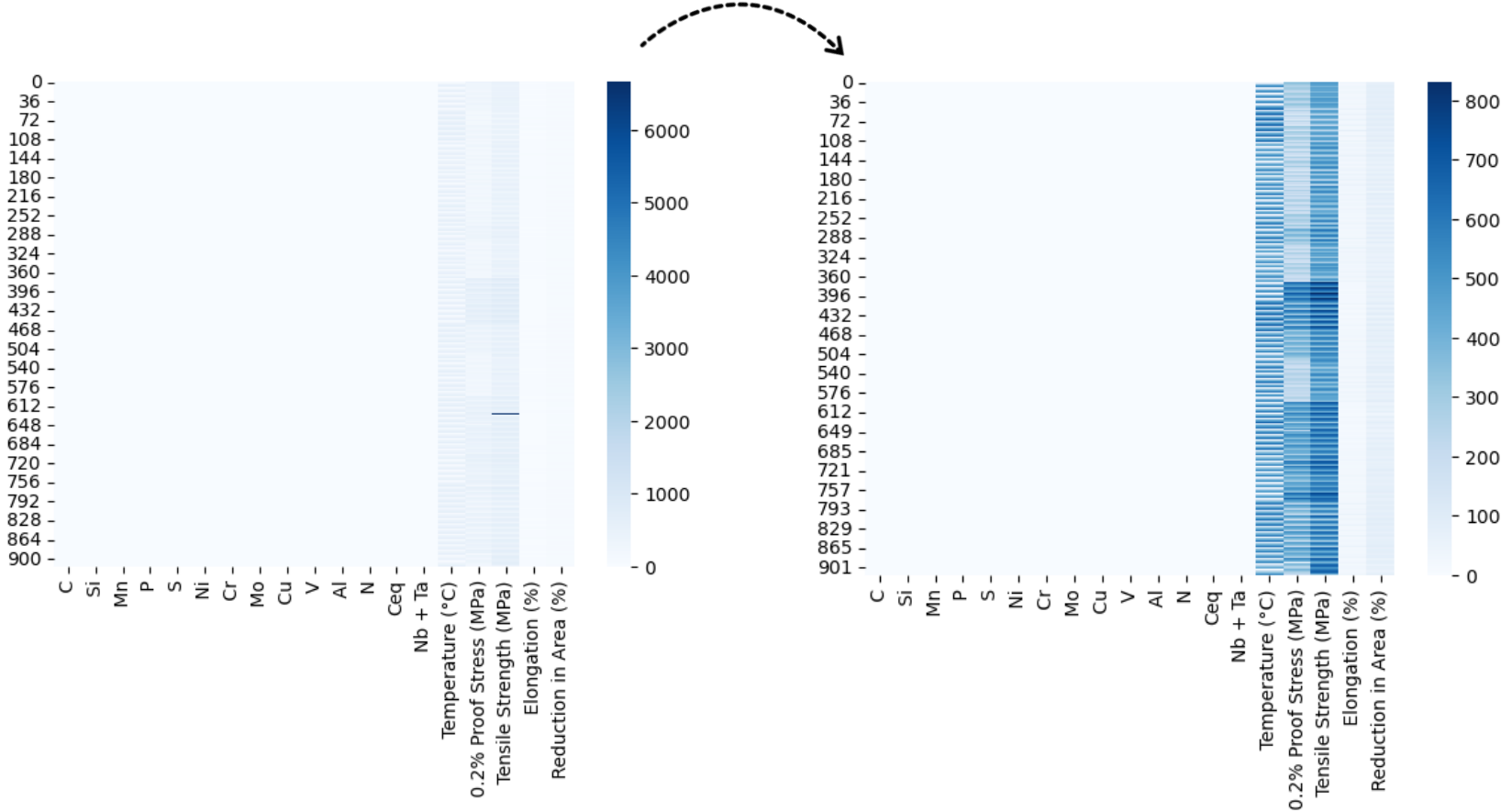



```
sns.heatmap(df[cols], cmap=color)
```

outlier with a **tensile strength** value of **6000 MPa**

Outliers

exclude the outlier from the data



Pre-processing

1. Constant values

Removed columns that contain the same number in all rows

2. NaNs

Removed columns that contain unknown values

3. Categorical

Processed categorical columns with one-hot encoding (i.e. code of the alloy)

```
1. df = df.loc[:, df.nunique() != 1]
```

```
1. columns_with_nan = df.columns[df.isnull().any()].tolist()
```

```
1. df = pd.get_dummies(df, columns=categorical_columns, drop_first=True)
```

Training
Training
Training
Models Training

Used models, results of training, accuracy of predictions

Training
Training
Training
Training

Used Models

We used **three different types** of models to compare prediction results

1st

Linear Regression

2nd

Random Forest Regression

3rd

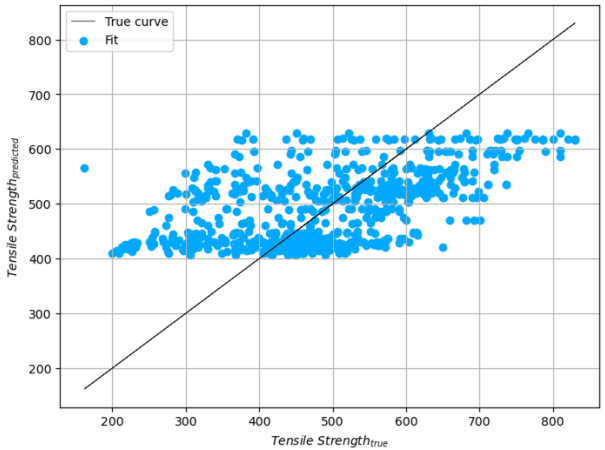
Catboost

Models Training

concentrations

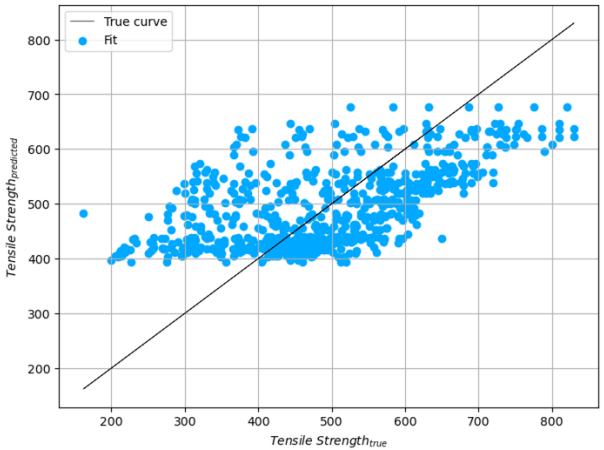
Concentrations: Train

Linear Regression



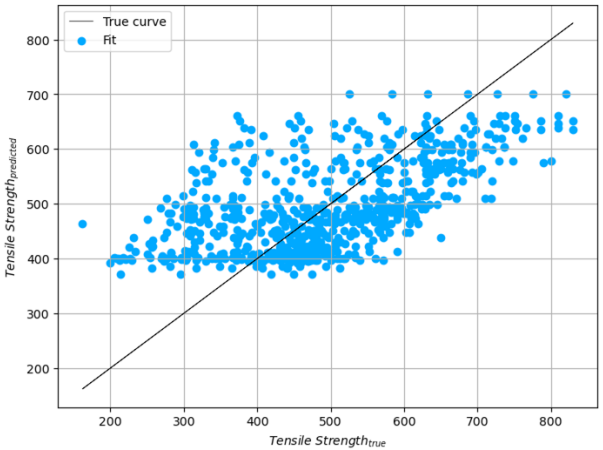
$\mathrm{RMSE} = 107.56$

Random Forest Regression



$\mathrm{RMSE} = 103.29$

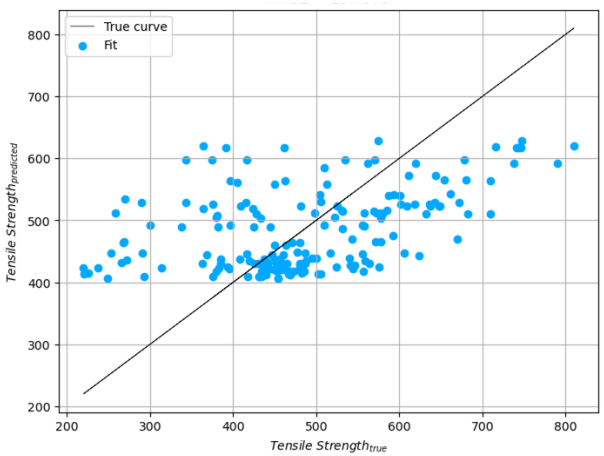
Catboost



$\mathrm{RMSE} = 104.42$

Concentrations: Test

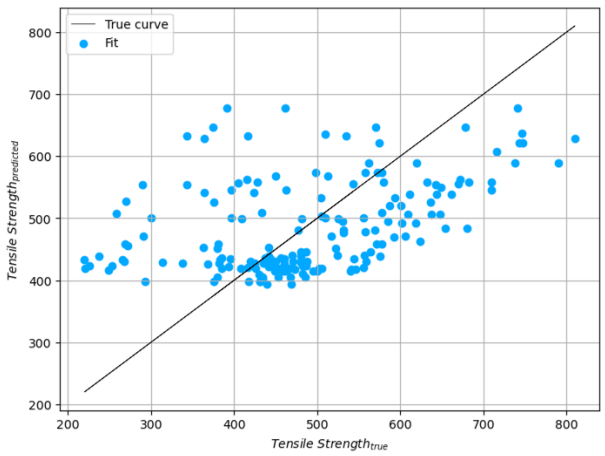
Linear Regression



$$\mathrm{RMSE} = 107.07$$

$$\mathrm{R}^2 = 0.21$$

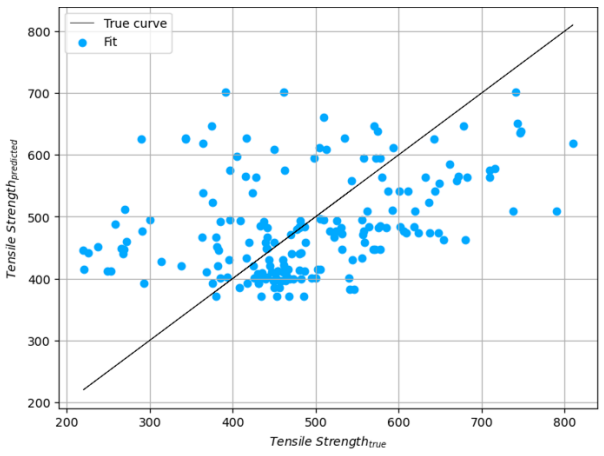
Random Forest Regression



$$\mathrm{RMSE} = 110.44$$

$$\mathrm{R}^2 = 0.16$$

Catboost



$$\mathrm{RMSE} = 117.66$$

$$\mathrm{R}^2 = 0.04$$

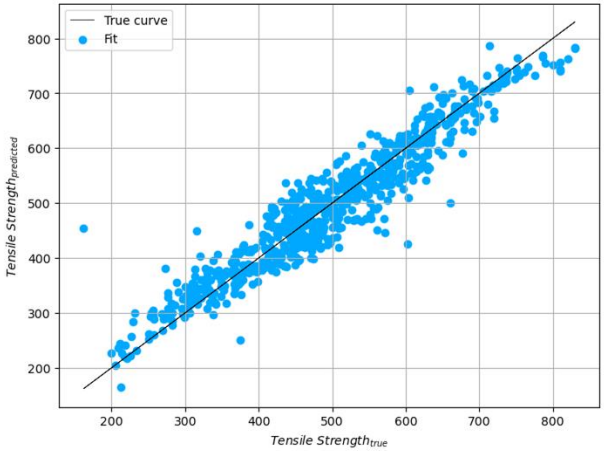
Models Training

original dataset

Let's train models not only on concentrations, but also on all the features remaining in the dataset(e.g. reduction in area, elongation, temperature)

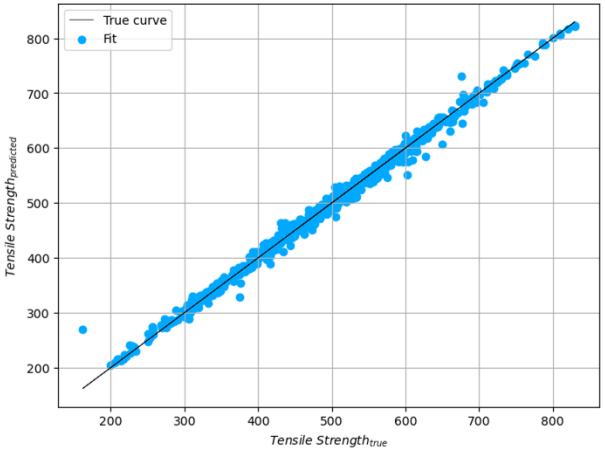
Original Dataset: Train

Linear Regression



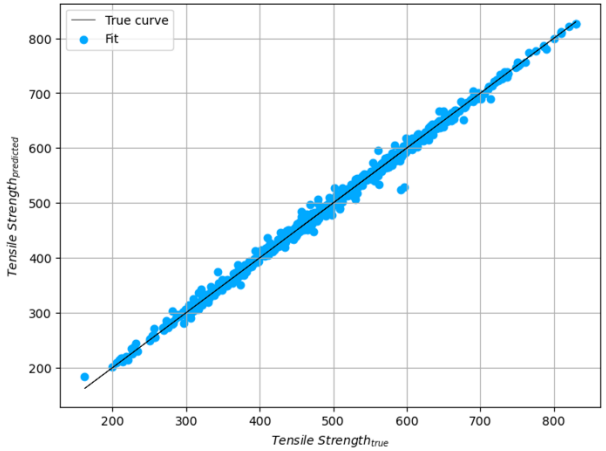
$\mathrm{RMSE} = 38.39$

Random Forest Regression



$\mathrm{RMSE} = 10.13$

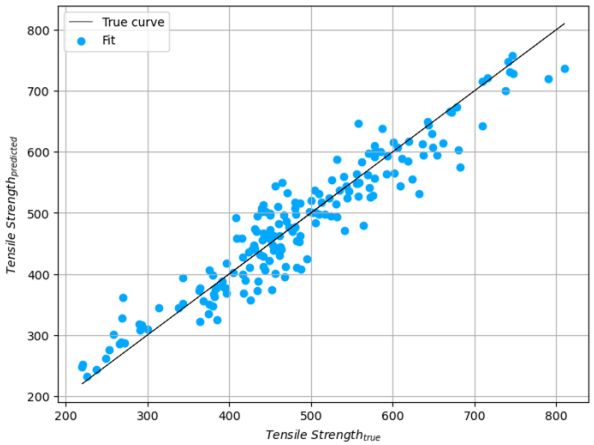
Catboost



$\mathrm{RMSE} = 8.11$

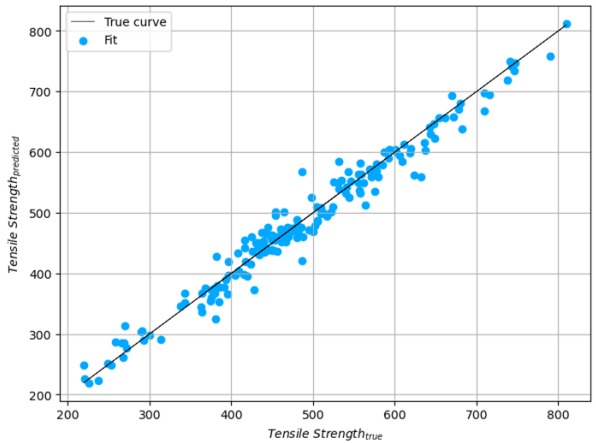
Original Dataset: Test

Linear Regression



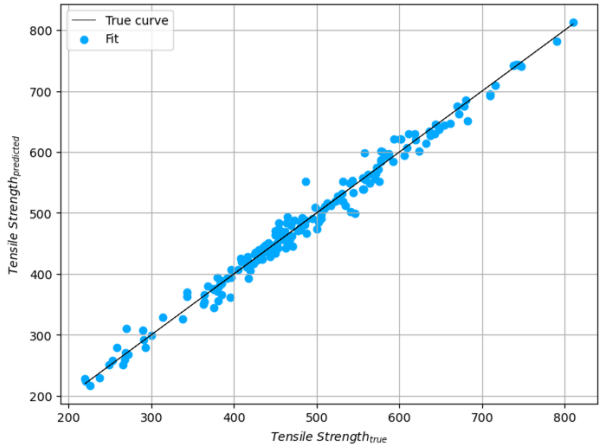
$$\mathrm{RMSE} = 38.51$$
$$\mathrm{R}^2 = 0.90$$

Random Forest Regression



$$\mathrm{RMSE} = 21.77$$
$$\mathrm{R}^2 = 0.97$$

Catboost



$$\mathrm{RMSE} = 14.53$$
$$\mathrm{R}^2 = 0.99$$

Models Training

magpie database

Data set expansion

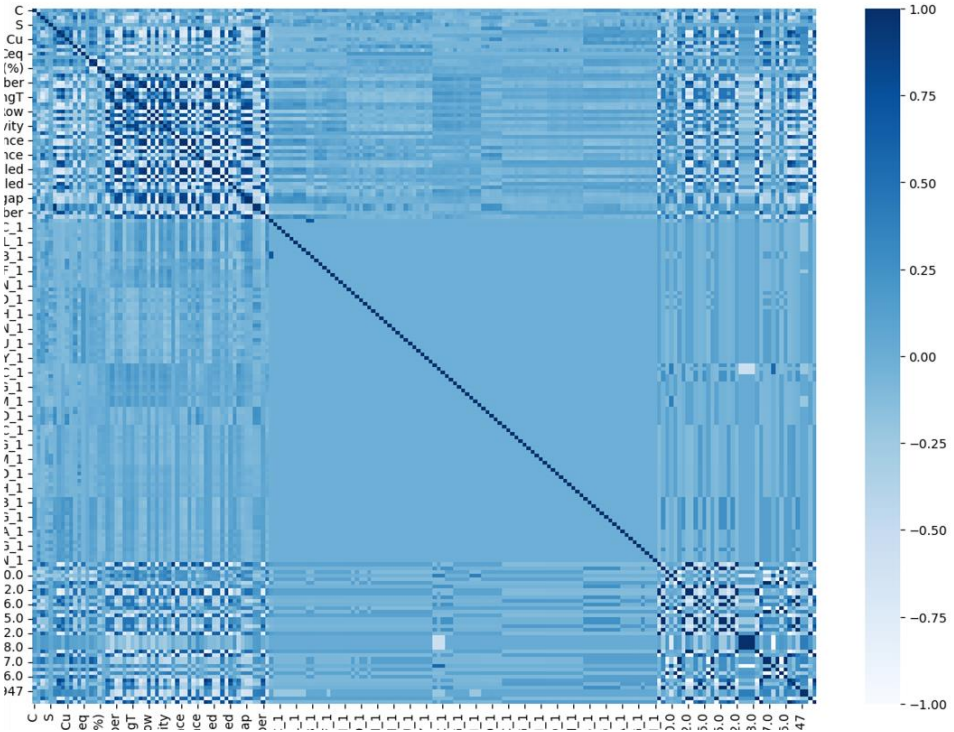
1. MAGPIE

"Materials Aggregated Property Prediction Engine”

Tool and dataset used in materials science and informatics for predicting materials properties based on the elemental composition of a material.

- maximum MeltingTemperature
- minimum CovalentRadius
- mean CovalentRadius
- maximum Electronegativity

Magpie Dataset: Correlation



915

steels

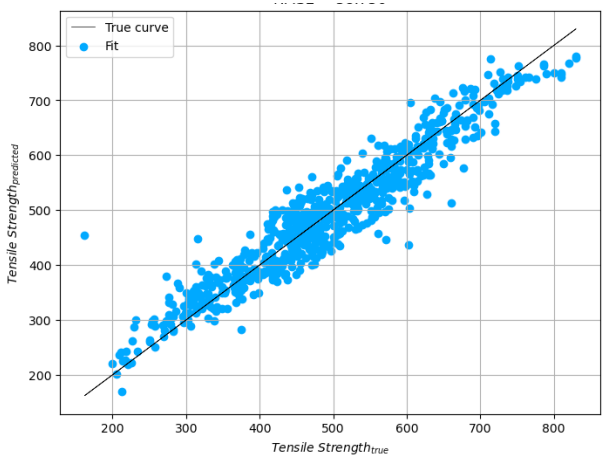
192

features

Magpie Dataset: Train

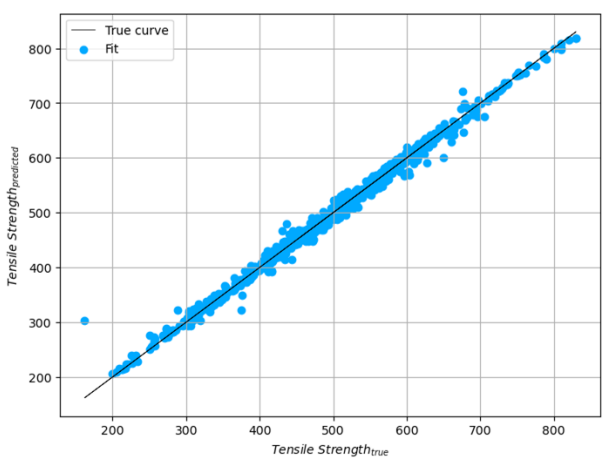
original dataset + magpie

Linear Regression



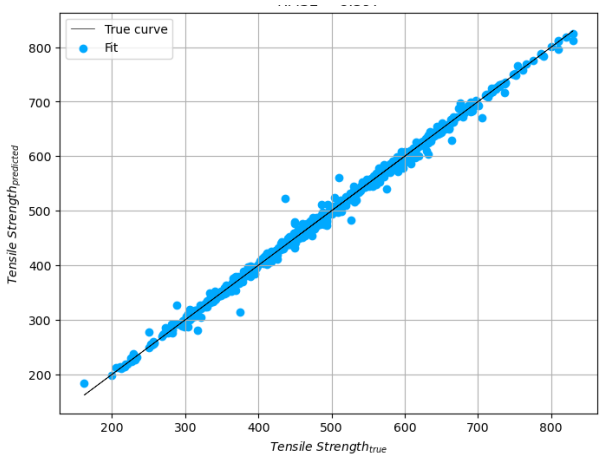
$\mathrm{RMSE} = 38.76$

Random Forest Regression



$\mathrm{RMSE} = 10.97$

Catboost

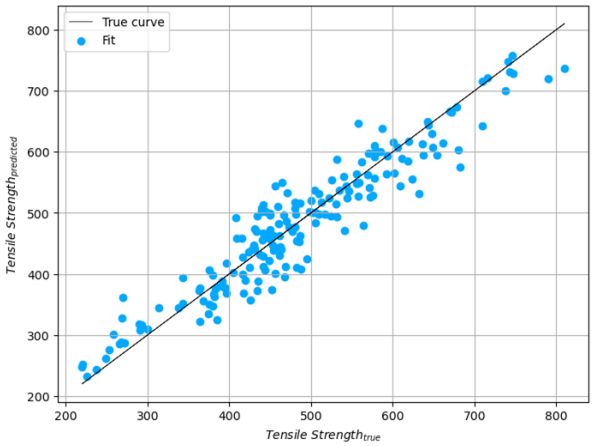


$\mathrm{RMSE} = 8.40$

Magpie Dataset: Test

original dataset + magpie

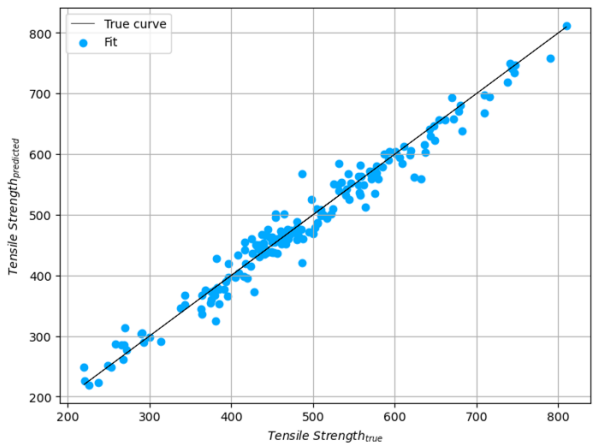
Linear Regression



$\mathrm{RMSE} = 39.36$

$R^2 = 0.89$

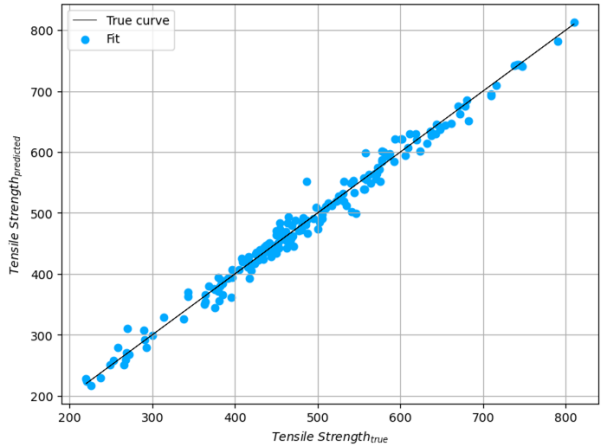
Random Forest Regression



$\mathrm{RMSE} = 21.61$

$R^2 = 0.97$

Catboost



$\mathrm{RMSE} = 16.20$

$R^2 = 0.98$

Models Training

megnet database

Data set expansion

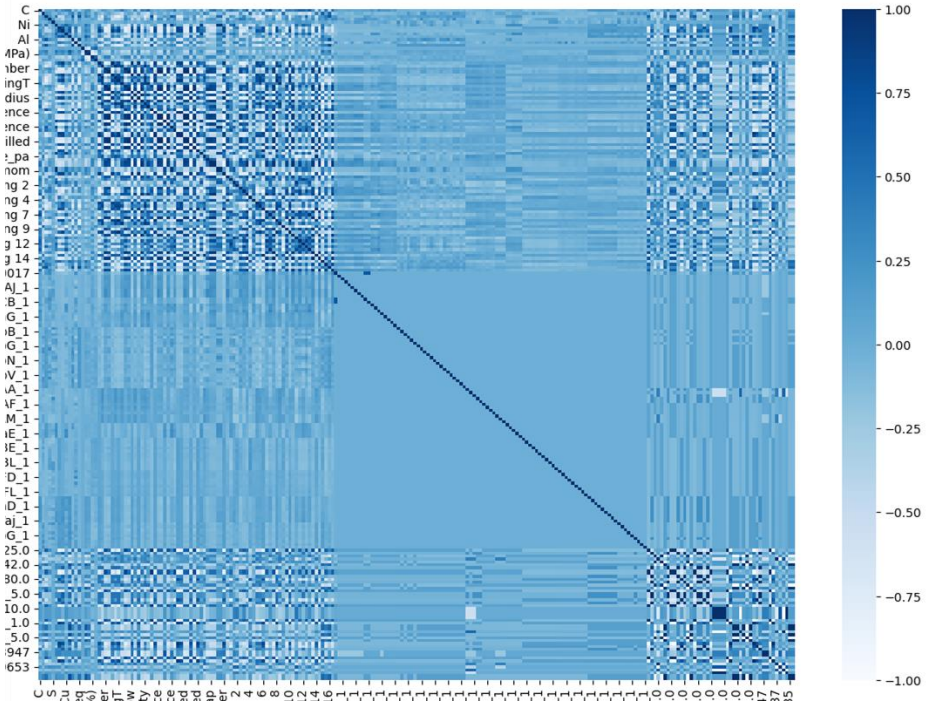
2. MEGNET embeddings

"The MatErials Graph Network ”

- Embeddings

Megnet provides element’s embeddings that encode useful chemical information that can be transferred learned to develop models with smaller datasets.

Megnet Dataset: Correlation



915

steels

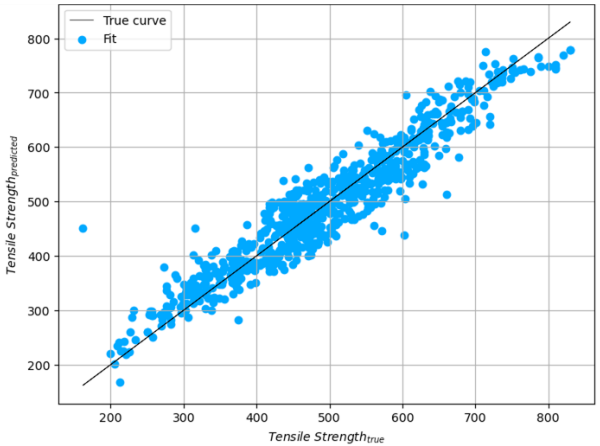
230

features

Megnet Dataset: Train

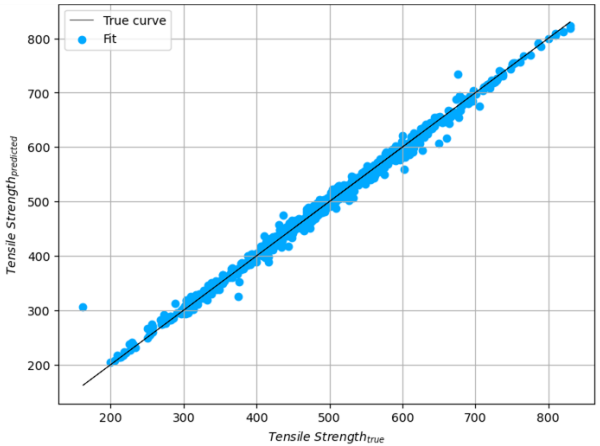
original dataset + magpie + megnet

Linear Regression



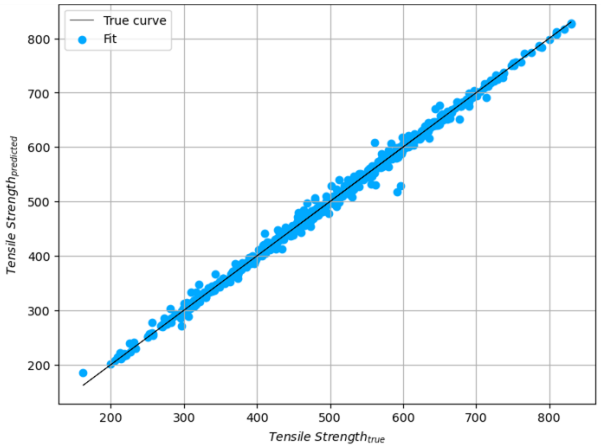
RMSE = 38.64

Random Forest Regression



RMSE = 10.63

Catboost

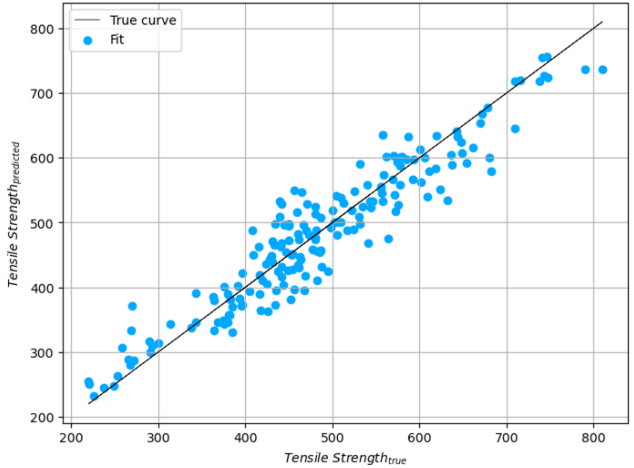


RMSE = 8.24

Megnet Dataset: Test

original dataset + magpie + megnet

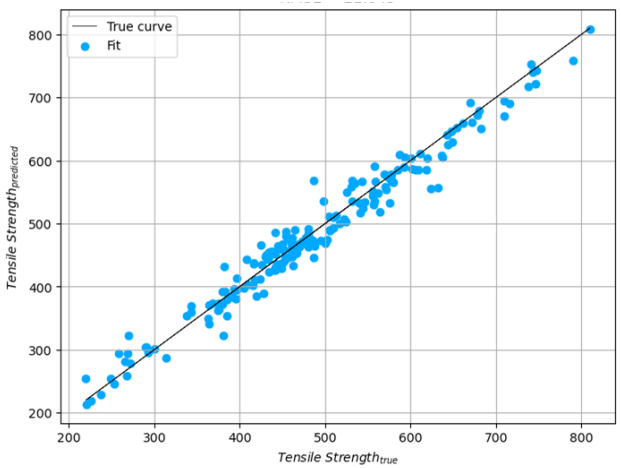
Linear Regression



RMSE = 39.08

$R^2 = 0.89$

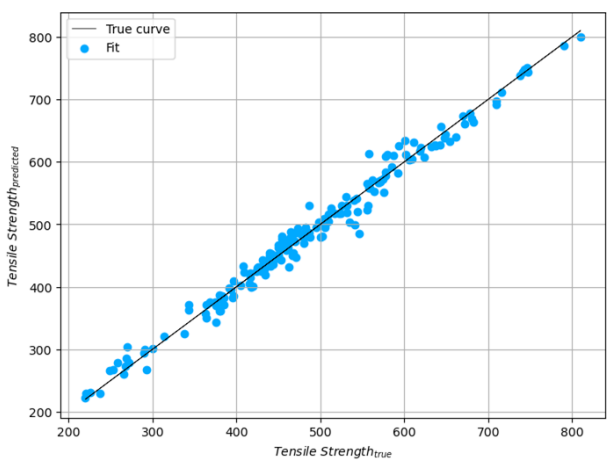
Random Forest Regression



RMSE = 21.84

$R^2 = 0.97$

Catboost



RMSE = 15.10

$R^2 = 0.98$

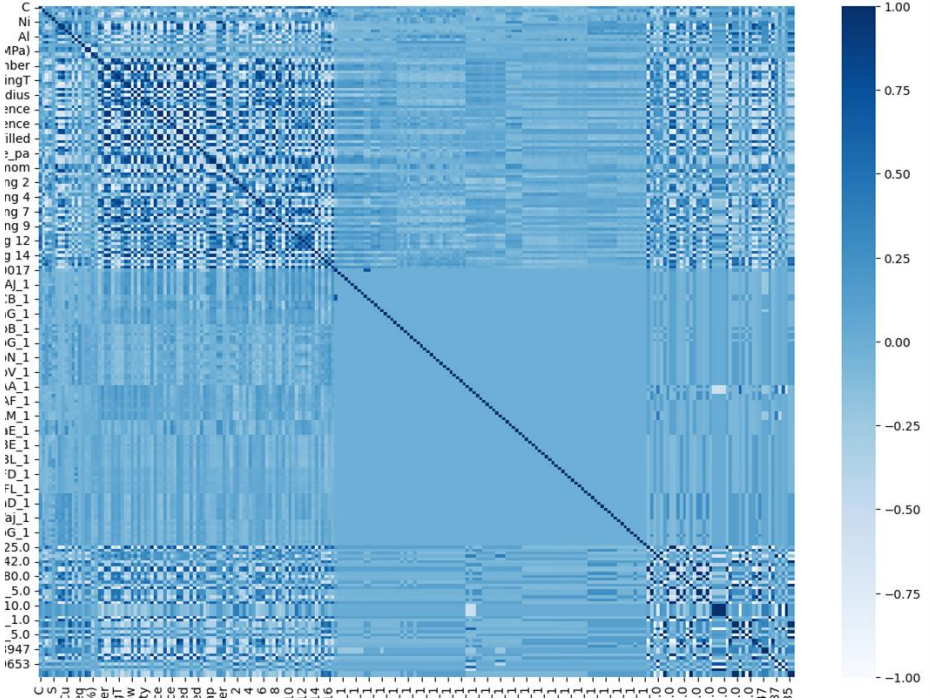
Models Training

top features

Finally, we train the model that performed best in the previous sections - **Catboost** with selection of the most important features

Top features

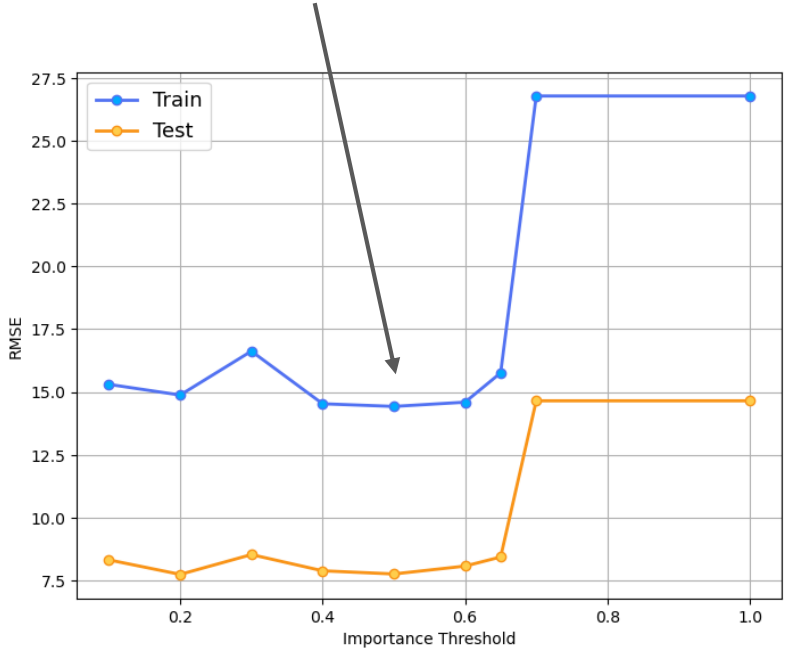
reminder

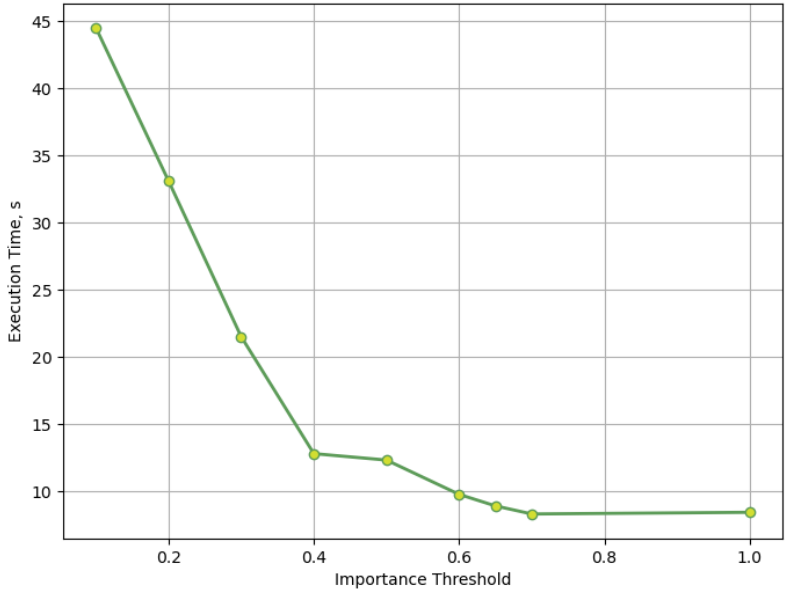



```
thresholds = [0.1, 0.2, ..., 10]

for threshold in thresholds:
          X = data[importance >
threshold]
          model = CatboostModel(train,
test)

print(time,
      number_of_features,
      RMSE_train,
      RMSE_test
     )
```

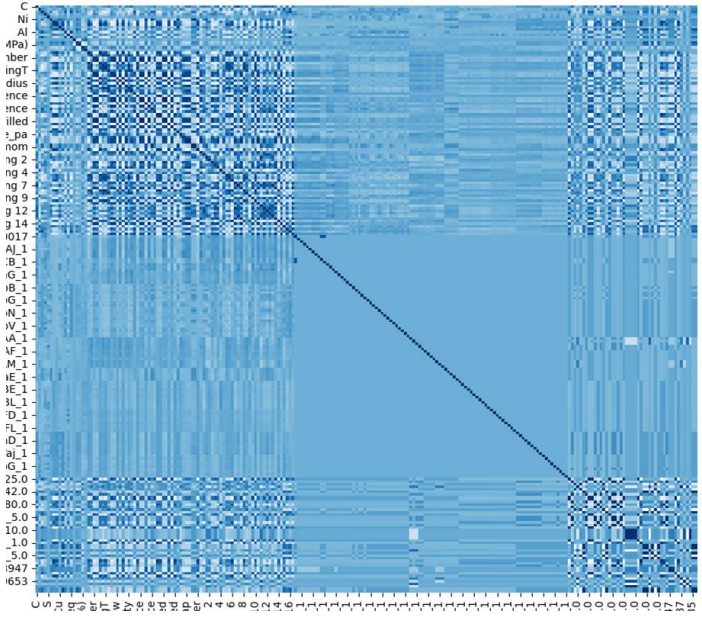
Top features

Model with **importance threshold = 0.5** is the best

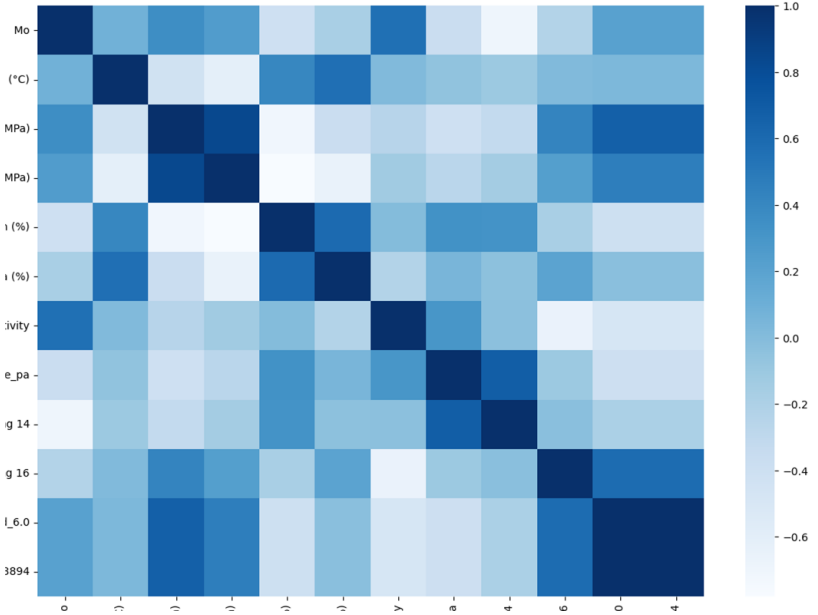




Top features



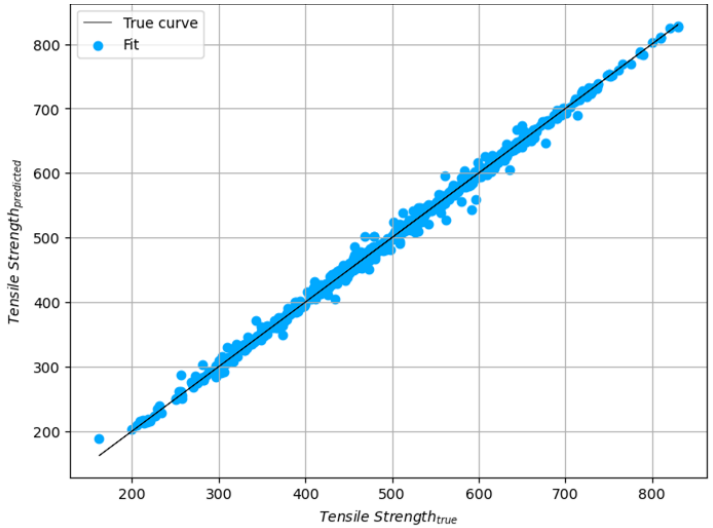
230 features



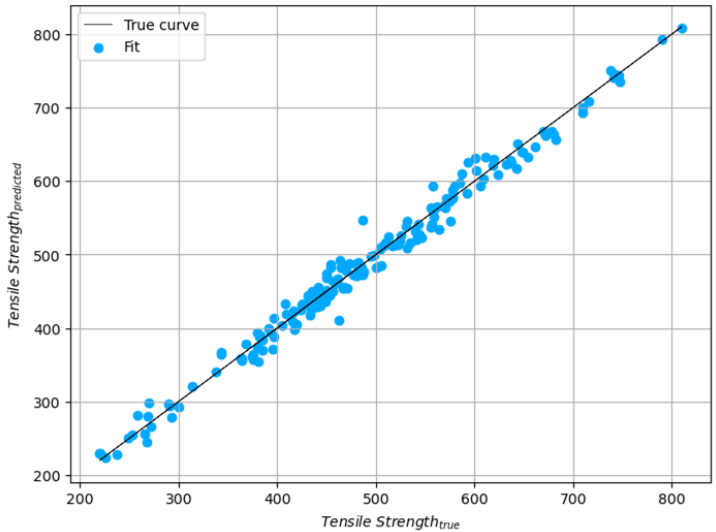
12 features

Top 12 Features: Catboost

original dataset + magpie + megnet



RMSE = 7.48

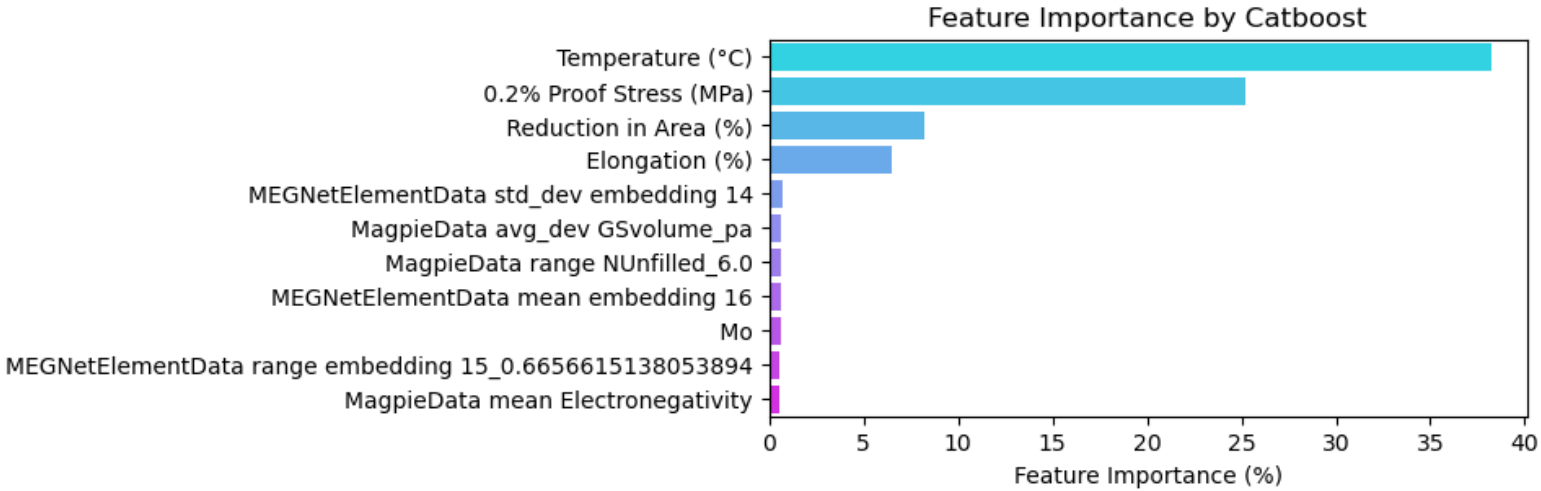


RMSE = 14.42

$R^2 = 0.99$

Top 12 Features: Catboost

original dataset + magpie + megnet



6
original dataset

3
magpie

3
megnet

Conclusion

Conclusion

Conclusion

Conclusion

Conclusion

Conclusion

Conclusion

Results: RMSE

	Concentrations	Original Data	Original + Magpie	Original + Magpie +
Linear Regression	107.075	38.517		
Random Forest Regressor	110.441	21.77		
Catboost	117.657	14.535		
Catboost Top 12 features	-	-	-	14.423

Results: R^2

	Concentrations	Original Data	Original + Magpie	Original + Magpie + Megnet
Linear Regression	0.21	0.9	0.89	0.89
Random Forest Regressor	0.16	0.97	0.97	0.97
Catboost	0.045	0.99	0.98	0.98
Catboost Top 12 features	-	-	-	0.99

Conclusion

1. The developed model avoids expensive and time-consuming experiments

2. The best model turned out to be catboost trained on the top 12 features using two open databases (Megnet & Magpie)

13
models

0.99
R^2 score

Questions?

	Concentrations	Original Data	Original + Magpie	Original + Magpie + Megnet
Linear Regression	107.075	38.517	39.364	39.078
Random Forest Regressor	110.441	21.77	21.611	21.845
Catboost	117.657	14.535	16.119	15.098
Catboost Top 12 features	-	-	-	14.423

RMSE for different models

Viktoriia Zinkovich
Viktoriia.Zinkovich@skoltech.ru
Data Science

Pavel Bartenev
Pavel.Bartenev@skoltech.ru
Data Science

Kamil Garifullin
Kamil.Garifullin@skoltech.ru
Data Science