DAIMLER

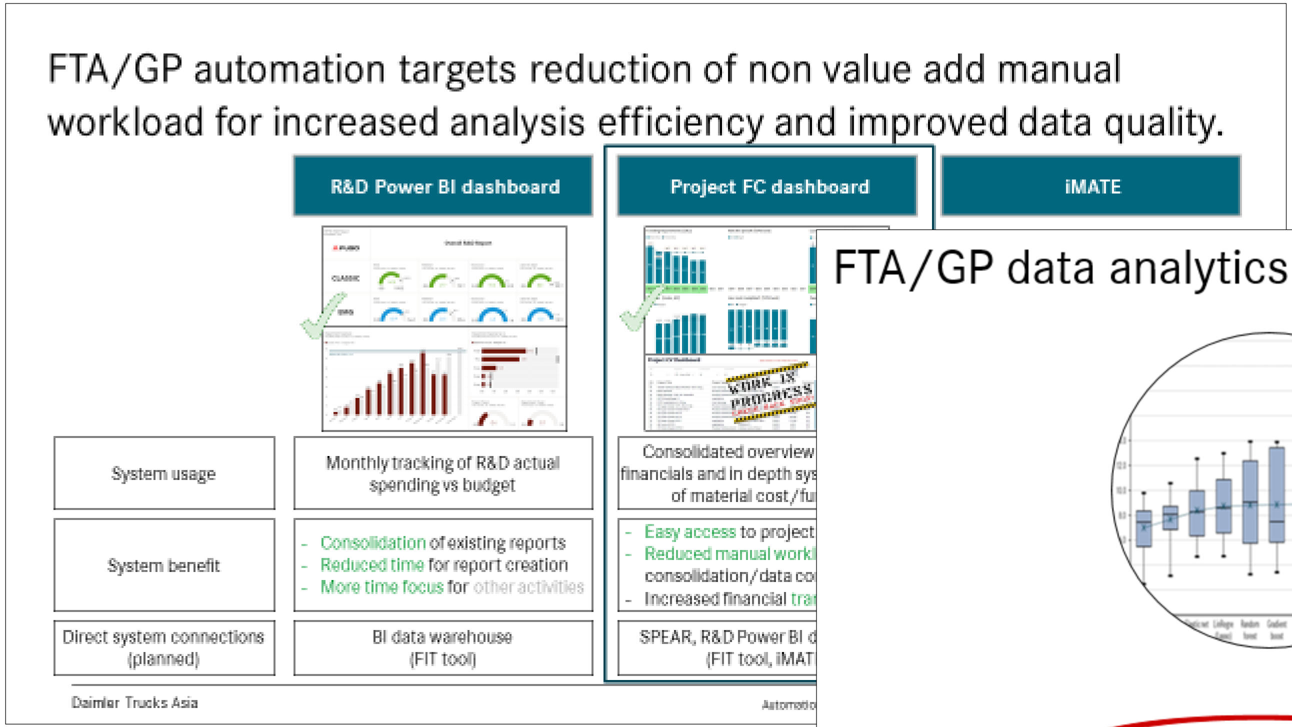
Product Funding Prediction using Python Machine Learning

13 April 2021

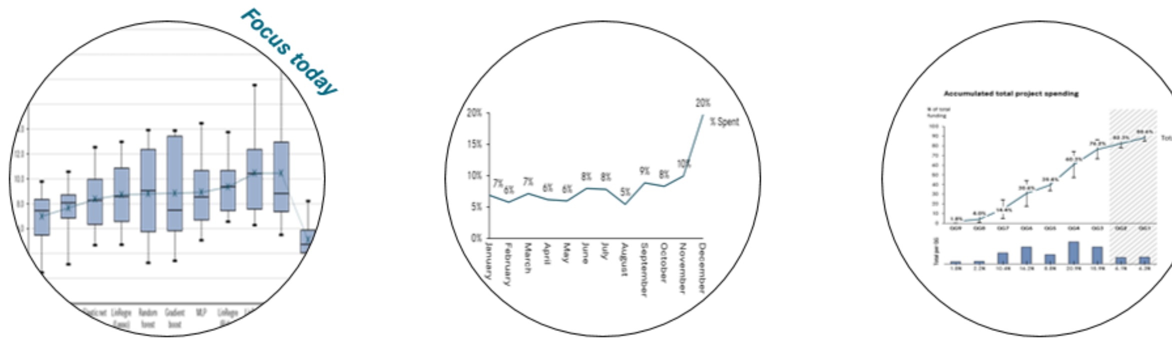
FTA/GP

Daimler Trucks Asia

In addition to report automation and iMate tool, FTA/GP focuses on data prediction using machine learning algorithms



FTA/GP data analytics activities as the next step in finance digitization

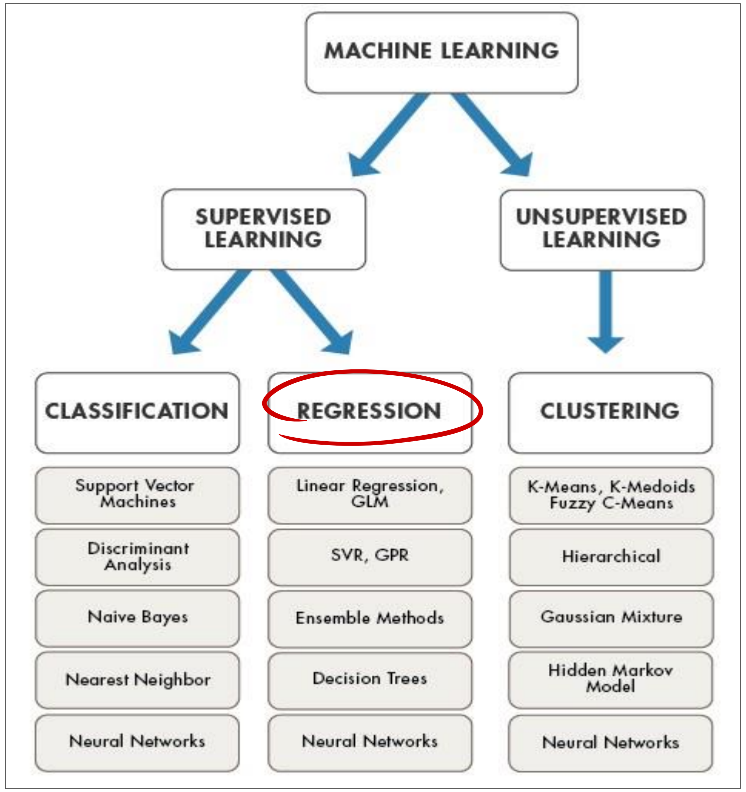


	Project funding prediction	Time series-based periodic forecasting	Milestone-based funding cycle prediction
Usage	Predictors-based funding estimation tool based on regression models.	Seasonality-adjusted auto-regression model. Based on historical data & rolling commitment	Project funding spending prediction. Based on historical data & project milestones.
Benefits	Early funding estimation with low-effort parametric input	Improved monthly/YE R&D forecast	Improved accuracy of the yearly budget planning for projects

Machine learning provides multiple finance applications. Regression models best suited as base for automated prediction tools

“ <u>Machine learning:</u> ”

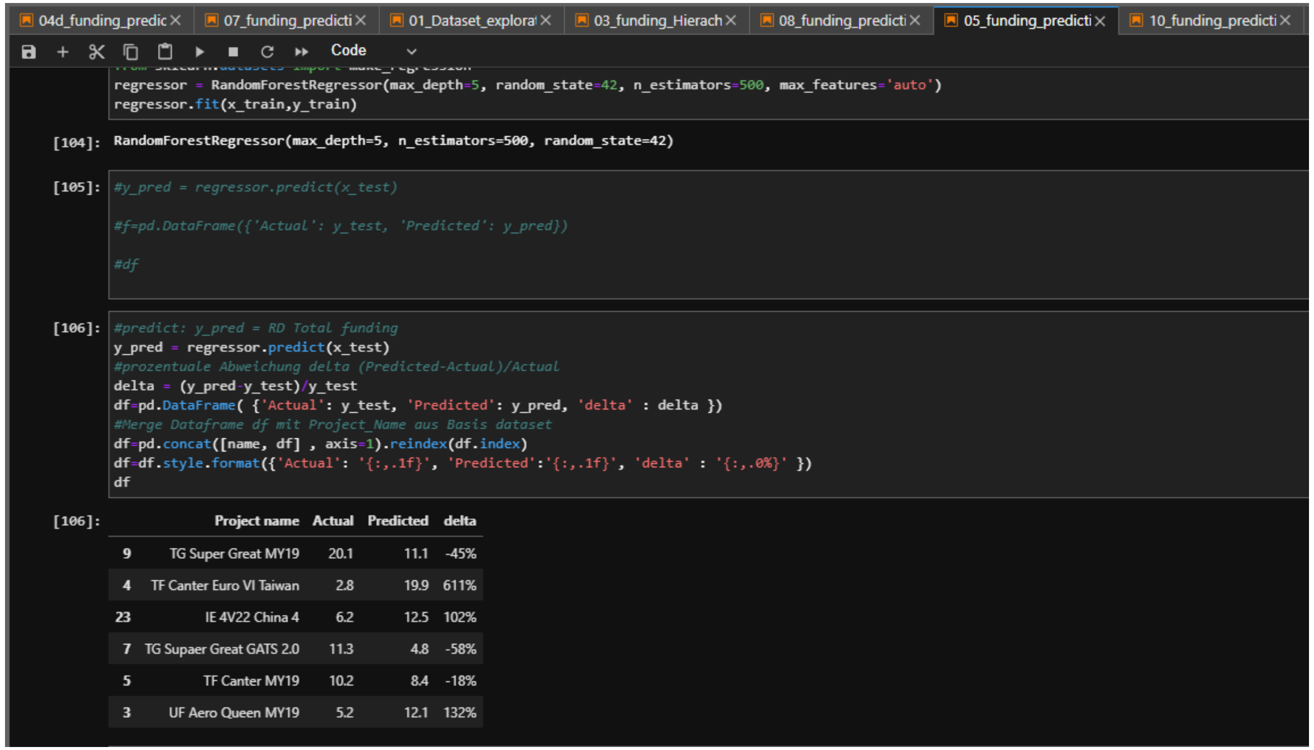
Utilization of computer systems that are able to learn and adapt by recognizing patterns in data without following explicit instructions



<u>Tested models (10)</u>: Linear regression, Lasso regression, Ridge regression, Elastic net, Random forest, Gradient boost, XGBoost, MLP, SVM, Stacking

Used tools:

python™ scikit learn jupyter NumPy matplotlib pandas



Funding prediction activity follows standard machine learning process

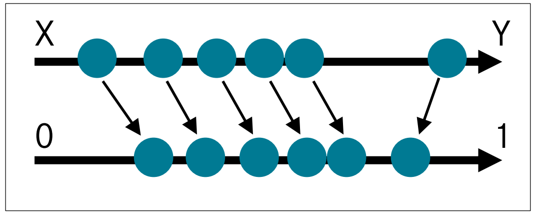
1. Data preparation

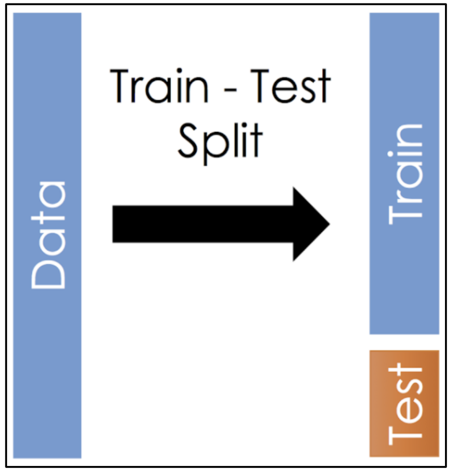
Dataset:
# of projects: 31
Selected data

	Input: R&D change content					
Project name	Ext. Facelift	Int. Facelift & Telematics	New E/E structure	New Safety features/Safety regulation	New emission Standard	Optional: other expected, significant funding drivers
TA Fighter MY19(Vehicle&Engine)	0	0.5	0.5	0.5	0.5	0.25
4P10 Euro VI Step D	0	0	0.25	0.25	0.5	0
UB Aero Star MY19	0	0.25	0.25	0.25	0	0
UF Aero Queen MY19	0.25	0.5	0.25	0.5	0	0
TF Canter Euro VI Taiwan	0	0	0	0.25	0.25	0
TF Canter MY19	0	0	0	0.25	0.25	0
TF Canter MY20	1	0.5	0	0.25	0	0
TG Supaer Great GATS 2.0	0	0	0	0	0	0.5
TG Super Great 12.8L to AUS/NZ	0.25	0	0	0	0	0.25
TG Super Great MY19	0	0.25	0.25	0.5	0	0
TG Super Great FX (6x6) MY21	0.25	0.25	0.25	0	0.25	0.5
TD Colt Diesel EuroIV	0	0	0	0	0.25	0.25
TF Canter Euro V Emerging Markets	0	0	0	0.25	0.25	0.25
TF Canter Euro VI Step E	0.5	0.25	0	0.75	0.75	0
A Fighter for the World (Vehicle&Engine)	0	0	0	0.5	0.25	0.5
UG Rosa MY21	0.0	0.25	0.25	0.50	0.25	0.50
TF Canter MY21 (W0)	0	0	0	0	0.25	0.25
TF Canter MY21 (W1)	0	1	0.5	0.25	0	1
TG Super Great MY21	0.25	0	0.25	0.5	0	0.25
TA Fighter MY21 (W1&2)	0	0	0	0	0	1
TA Fighter MY21 (W3)	0	0	0	0	0	1
TG Super Great 10.7L short-cab	0	0	0	0	0	0.25
IE 4M50 China 4	0	0	0	0	1	0
IE 4V22 China 4	0	0	0	0	1	0
UF Black Hawk	1	0.75	0.5	0.5	0.25	0
UB Sumo MY17	0	0	0	0	0.25	0
eCanter 1.0	0	0.25	0	0	1	0
eCanter 1.1	0	0	0	0.25	0	0
eCanter 2.1	0	1	0.5	0.5	1	1.0
eAxle	0	0	0	0	1	0
CATL Battery & Charging	0	0	0	0	1	0

2. Pre-processing

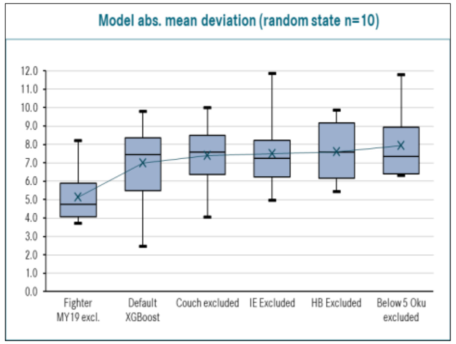
Min-max scaling





3. Modeling

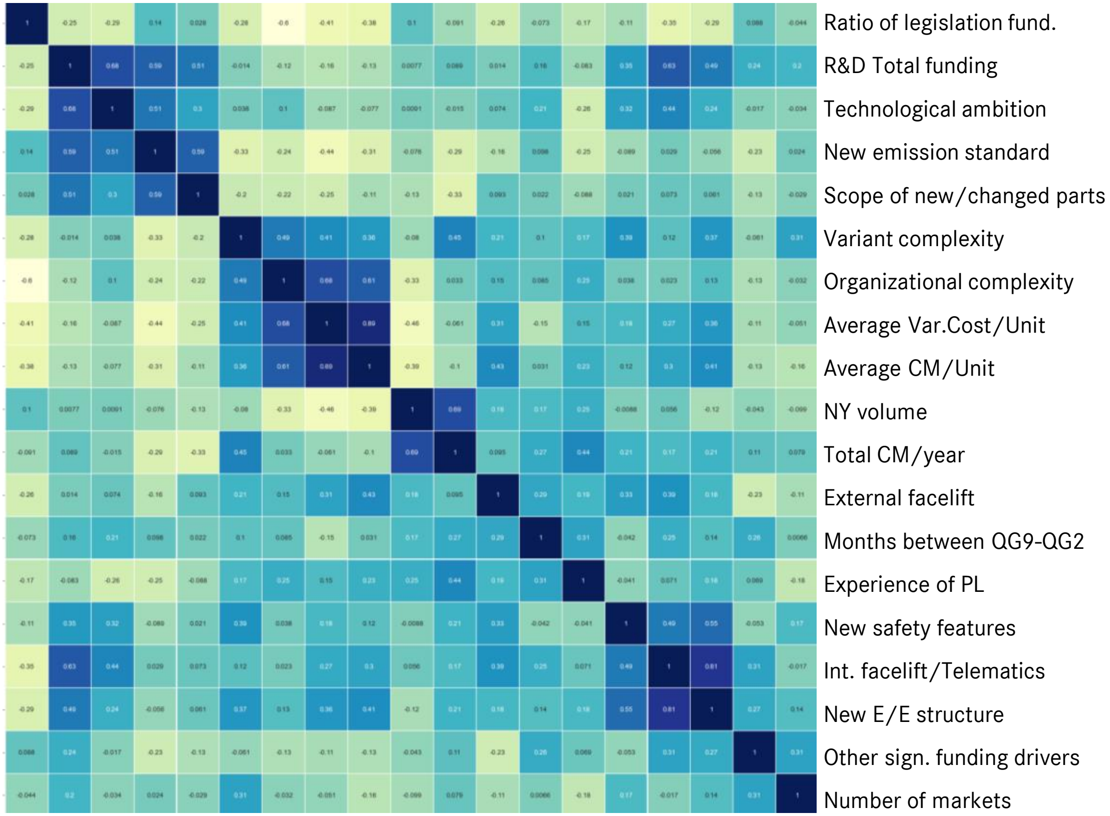
4. Validation



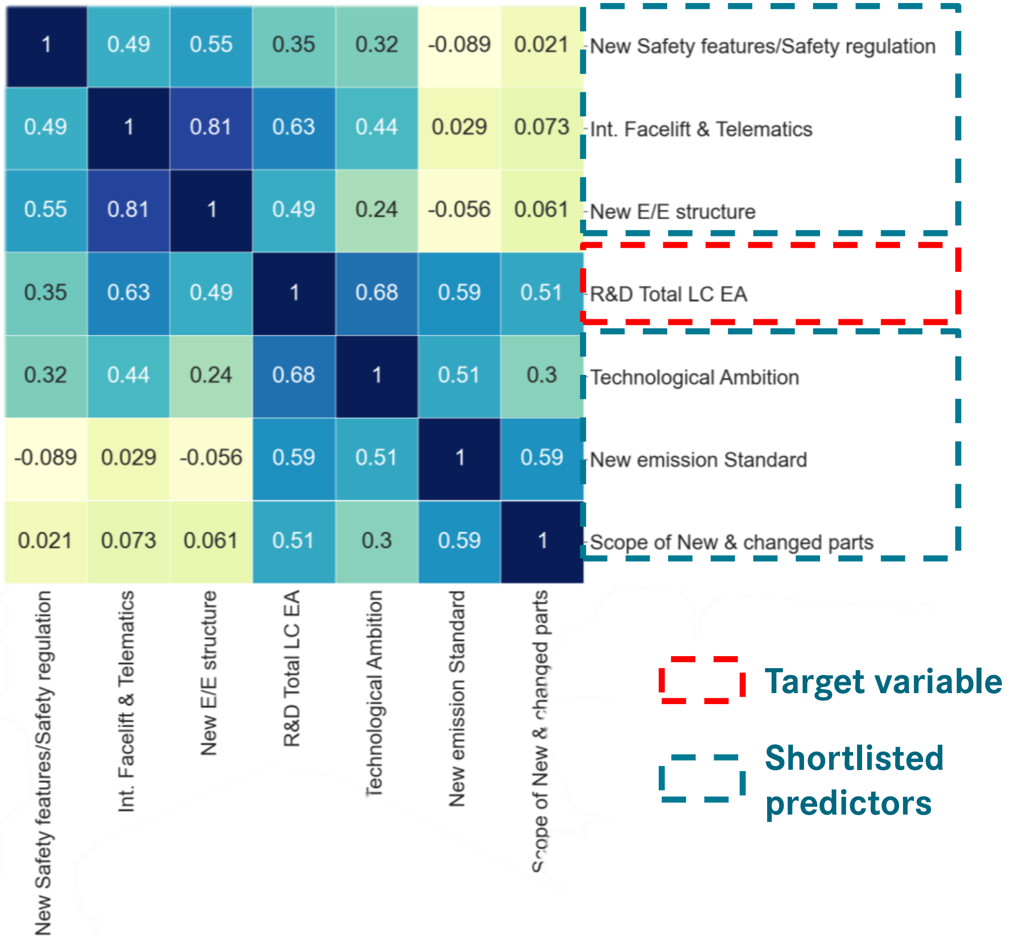
During data exploration 6 most promising predictors out of 18 have been shortlisted to be used in prediction models

Correlation heatmaps (Pearson)

18 predictors that can be objectively attributed to a project



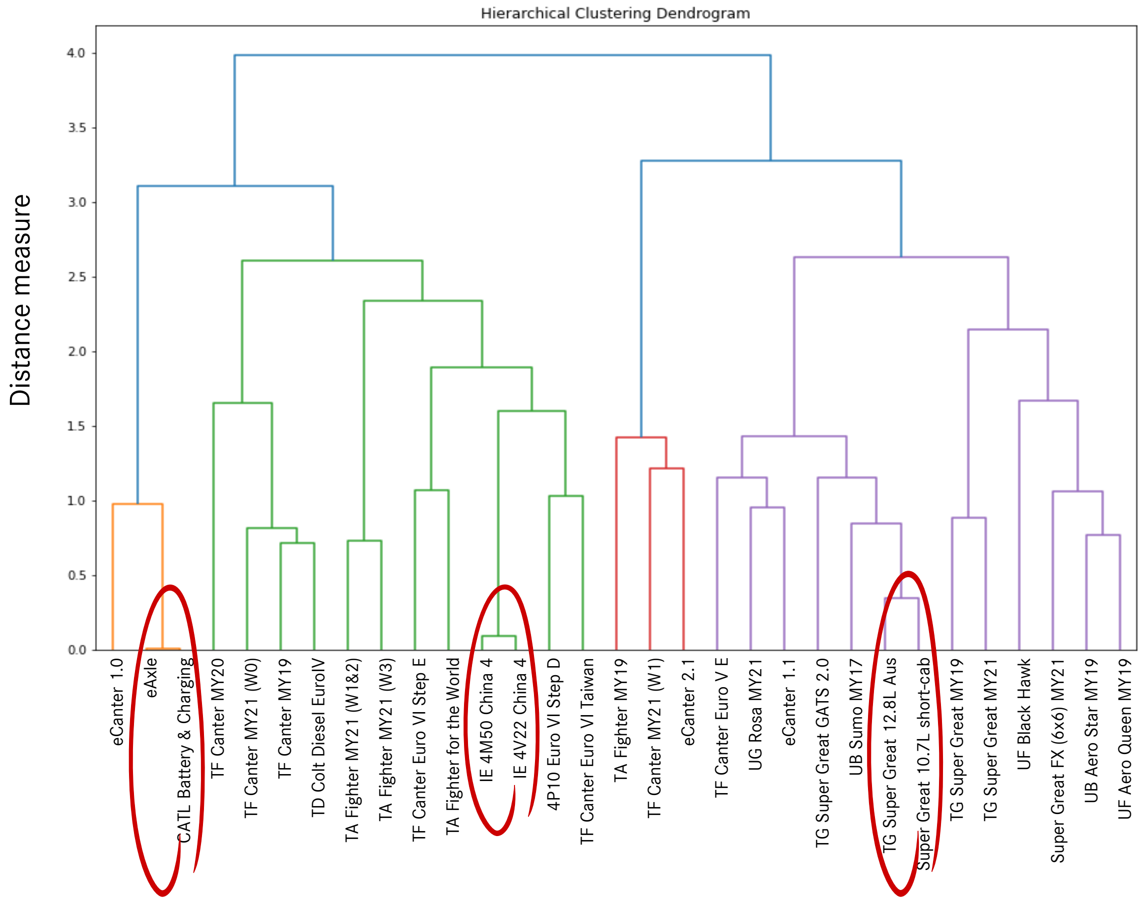
6 predictors with high levels of correlation to R&D funding



Correlation over r=**0.35** considered moderate/strong

Heterogenic nature of dataset, limited sample size and input scarcity are main challenges that have to be addressed to achieve prediction quality

Hierarchical agglomerative clustering (ward)



Cluster analysis shows high connection point for majority of projects ➡ **Highly heterogenic dataset**

Excerpt of the dataset

31 projects **6 shortlisted predictors**

Project name	Ext. Facelift	Int. Facelift & Telematics	New E/E structure	New Safety features/Safety regulation	New emission Standard	Optional: other expected, significant funding drivers
TA Fighter MY19(Vehicle&Engine)	0.00	0.50	0.50	0.50	0.50	0.25
4P10 Euro VI Step D	0.00	0.00	0.25	0.25	0.50	0.00
UB Aero Star MY19	0.00	0.25	0.25	0.25	0.00	0.00
UF Aero Queen MY19	0.25	0.50	0.25	0.50	0.00	0.00
TF Canter Euro VI Taiwan	0.00	0.00	0.00	0.25	0.25	0.00
TF Canter MY19	0.00	0.00	0.00	0.25	0.25	0.00
TF Canter MY20	1.00	0.50	0.00	0.25	0.00	0.00
TG Supaer Great GATS 2.0	0.00	0.00	0.00	0.00	0.00	0.50
TG Super Great 12.8L to AUS/NZ	0.25	0.00	0.00	0.00	0.00	0.25
TG Super Great MY19	0.00	0.25	0.25	0.50	0.00	0.00
TG Super Great FX (6x6) MY21	0.25	0.25	0.25	0.00	0.25	0.50
TD Colt Diesel EuroIV	0.00	0.00	0.00	0.00	0.25	0.25
TF Canter Euro V Emerging Markets	0.00	0.00	0.00	0.25	0.25	0.25
TF Canter Euro VI Step E	0.50	0.25	0.00	0.75	0.75	0.00
TA Fighter for the World (Vehicle&Engine)	0.00	0.00	0.00	0.50	0.25	0.50
UG Rosa MY21	0.00	0.25	0.25	0.50	0.25	0.50
TF Canter MY21 (W0)	0.00	0.00	0.00	0.00	0.25	0.25
TF Canter MY21 (W1)	0.00	1.00	0.50	0.25	0.00	1.00
TG Super Great MY21	0.25	0.00	0.25	0.50	0.00	0.25
TA Fighter MY21 (W1&2)	0.00	0.00	0.00	0.00	0.00	1.00
TA Fighter MY21 (W3)	0.00	0.00	0.00	0.00	0.00	1.00
TG Super Great 10.7L short-cab	0.00	0.00	0.00	0.00	0.00	0.25
IE 4M50 China 4	0.00	0.00	0.00	0.00	1.00	0.00
IE 4V22 China 4	0.00	0.00	0.00	0.00	1.00	0.00
UF Black Hawk	1.00	0.75	0.50	0.50	0.25	0.00
UB Sumo MY17	0.00	0.00	0.00	0.00	0.25	0.00
eCanter 1.0	0.00	0.25	0.00	0.00	1.00	0.00
eCanter 1.1	0.00	0.00	0.00	0.25	0.00	0.00
eCanter 2.1	0.00	1.00	0.50	0.50	1.00	1.00
eAxle	0.00	0.00	0.00	0.00	1.00	0.00
CATL Battery & Charging	0.00	0.00	0.00	0.00	1.00	0.00

Small sample size (n=31) and heterogenic dataset ➡ **Careful selection of modeling approach required**

Python Demo

04d_fun ✕ | 07_fundi ✕ | 01_Data ✕ | 03_fundi ✕ | New_fur ✕ | 08_fundi ✕ | 05_fundi ✕ | 10_fundi ✕ | 09_fundi ✕ | 06_fundi

Code | Python 3

```python
#dataset=dataset[dataset['Component project'] != 1]
#dataset=dataset[dataset['LDT'] != 1]
#dataset=dataset[dataset['NY volume TARGET'] > 500 ]
#dataset=dataset[dataset['NY volume TARGET'] < 50000 ]
#dataset=dataset[dataset['IE'] != 1]
dataset=dataset[dataset['Project name'] != 'TA Fighter MY19(Vehicle&Engine)']
#dataset=dataset[dataset['IE'] != 1]
#dataset=dataset[dataset['LB'] != 1]
#dataset=dataset[dataset['HB'] != 1]
#dataset=dataset[dataset['Project name'] != 'TA Fighter MY19(Vehicle&Engine)']
#dataset=dataset[dataset['TA Classic?'] == 1]
dataset=dataset[dataset['R&D Total LC EA'].notna()]
dataset=dataset[dataset['Total funding LC EA'] > 3]
dataset=dataset[dataset['Total funding LC EA'] < 150]
y=dataset.iloc[:,12]
#x=dataset.iloc[:,np.r_ [5:9, 24:29, 30:36]]
x=dataset.iloc[:,np.r_ [25:29,31:33]]
#Scale all features of dataset to 0 to 1 range to avoid distortions during PCA and clustering.
#Convert np array back to a panda DataFrame and add back column header cut off by sklearn preprocessing
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
xs = scaler.fit_transform(x)
x= pd.DataFrame(xs, columns=x.columns)
x=x.fillna(0)
```

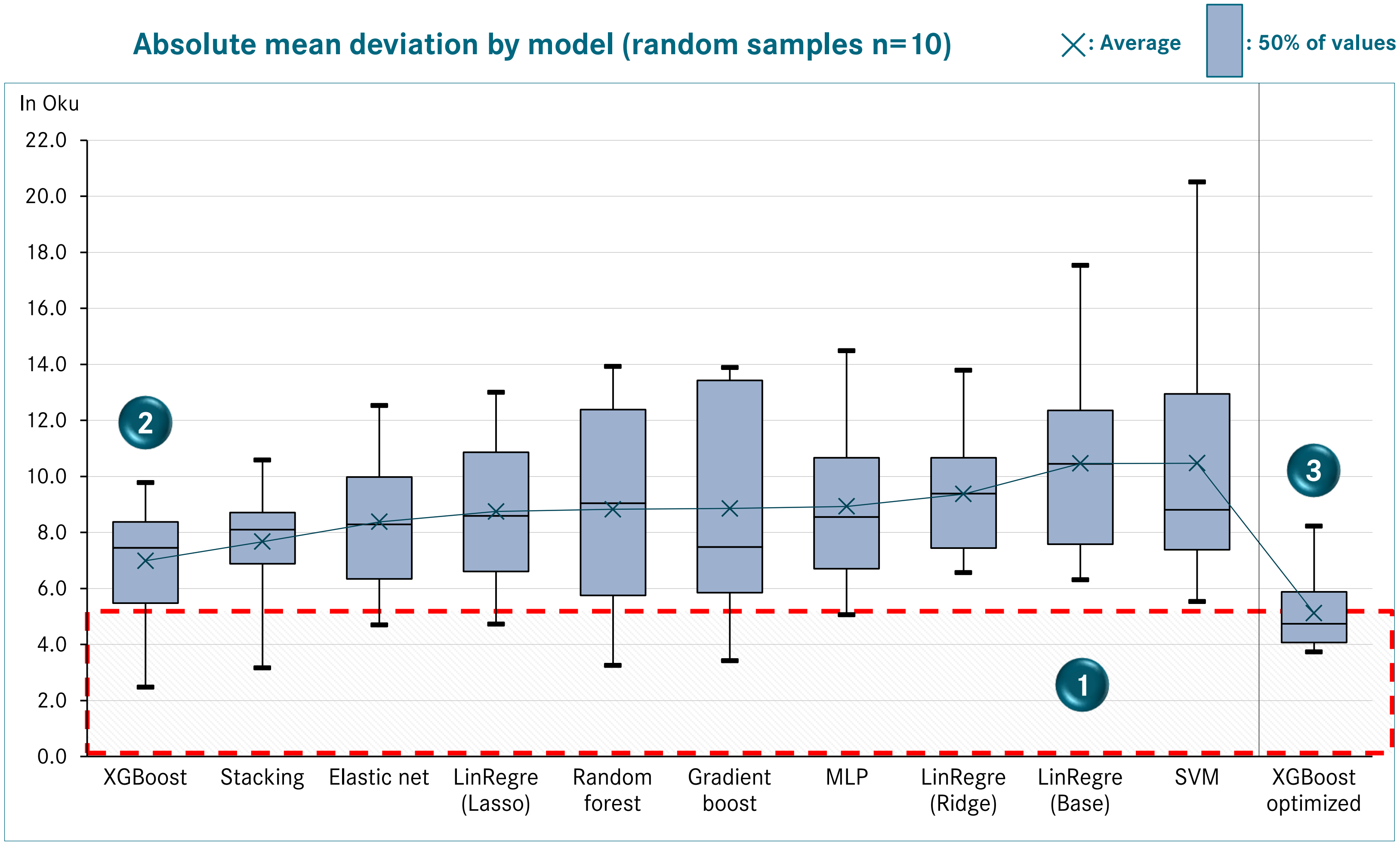
```
[499]: print (dataset.shape)
(29, 71)

[507]: y

[507]: 1      13.80
       2       8.30
       3       7.50
```

XGBoost is the best model with average prediction accuracy close to ±5 Oku. Further optimization required to improve avg. accuracy and spread

Absolute mean deviation by model (random samples n=10)



Testing and validation approach

- R&D funding chosen as testing variable
- Model has been trained by 80% of random projects and tested on 20% of random projects
- Validation repeated x10 with random sets of projects
- Results represent average deviation of 6 random test projects (20%) from actual funding target

Key takeaways

1. Prediction corridor aspiration within 5 Oku of actual
2. XGBoost performs best among all tested projects even before optimization, but doesn't achieve the target prediction accuracy
3. XGBoost optimization with exclusion of just one outlier improves the prediction accuracy by 2 Oku and just reaches the target prediction range

- [Not shown] Prediction accuracy for total funding comparable with R&D funding

Promising results. Further optimization of dataset & model to achieve target range