Target parameter prediction of a bioscientific device based on its geometrical parameters

Project by: Paul Kollhof & Aykut Avci

Agenda

- Introduction & Objective
- Data Overview
- Data Processing
- Model Presentation
- Conclusion & Outlook

Introduction & Objective

Introduction:

- Anonymized manufacturing data of anonymized life science company

Objective:

- Building predictive model for target parameter (mainly based on geometrical parameters)

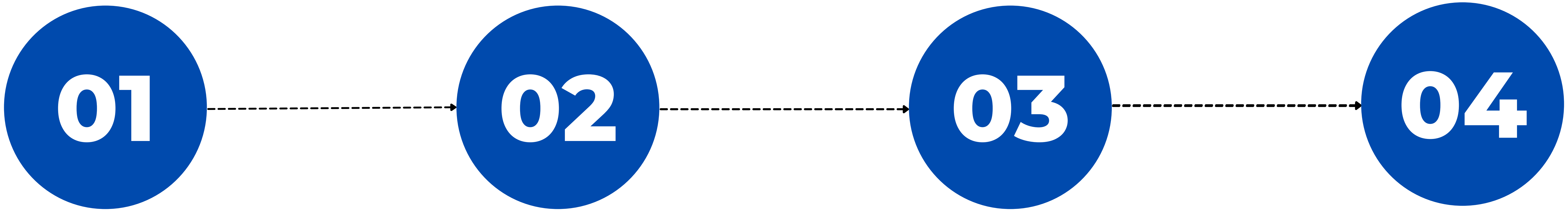
Data Overview

Data Source:

- Variety of numerical and categorical features
- Gathered throughout a multi-step manufacturing process
- Various manual and automatized data inputs
- Centralized in SQL Database
- Raw Data: ~110k sample & 195 Features
- Cleaned Data: ~23k sample & ~14 Features

Data Processing

Processing raw data to improve data quality to build accurate predictive model



Step 1

Loading data and deleting features that are not useful

Step 2

Standardizing and anonymizing column names deleting rows without entries and converting string type data into numeric ones, initial EDA

Step 3

Dealing with outliers and NaN values by interpolation

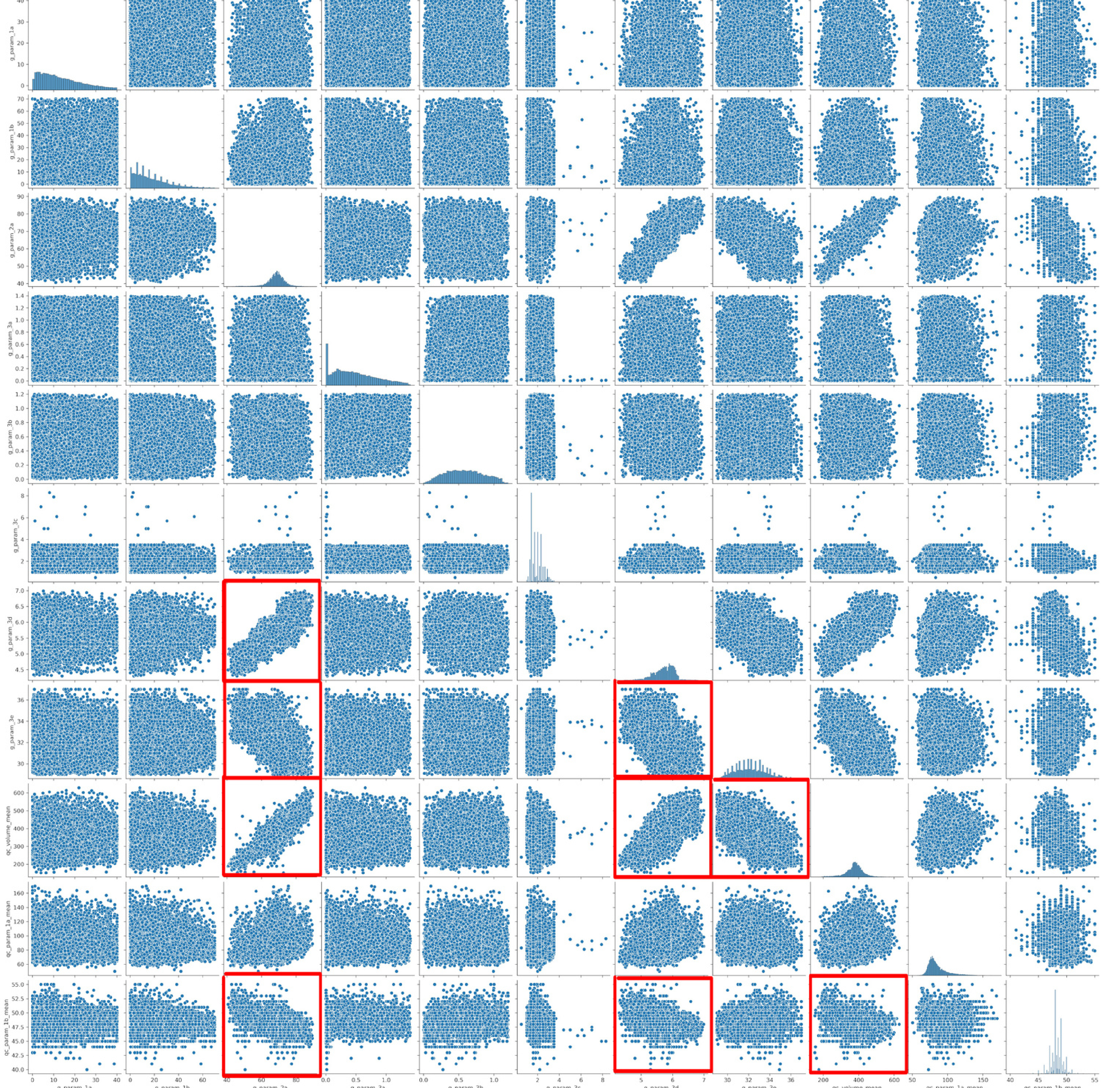
Step 4

Splitting categorical and numerical data, focusing on numerical for modeling

Feature Investigation

Investigating correlations between numerical features

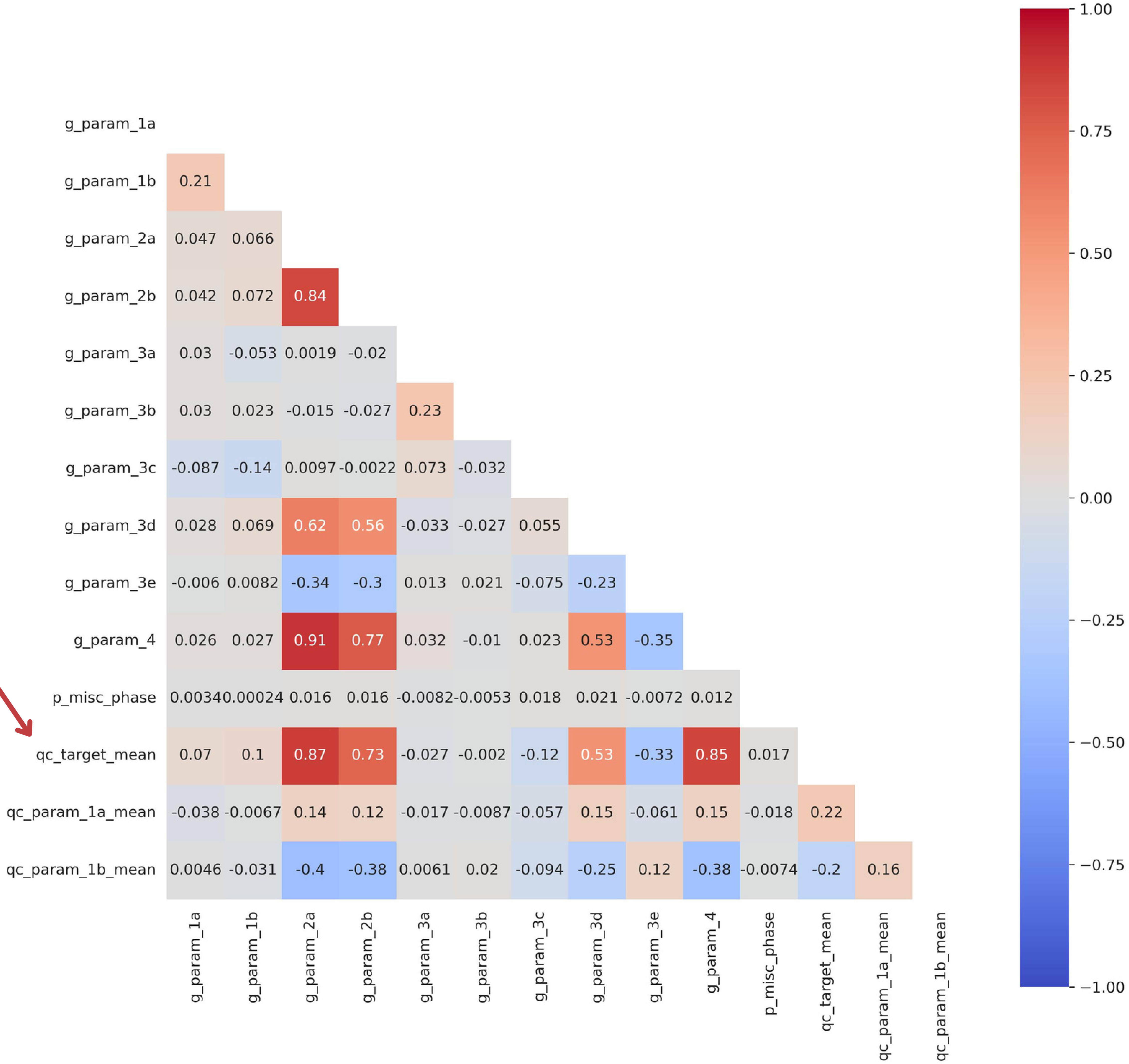
Likely linear relationship between target feature (qc_target_mean) and other variables



Heat Map

Investigating feature correlations

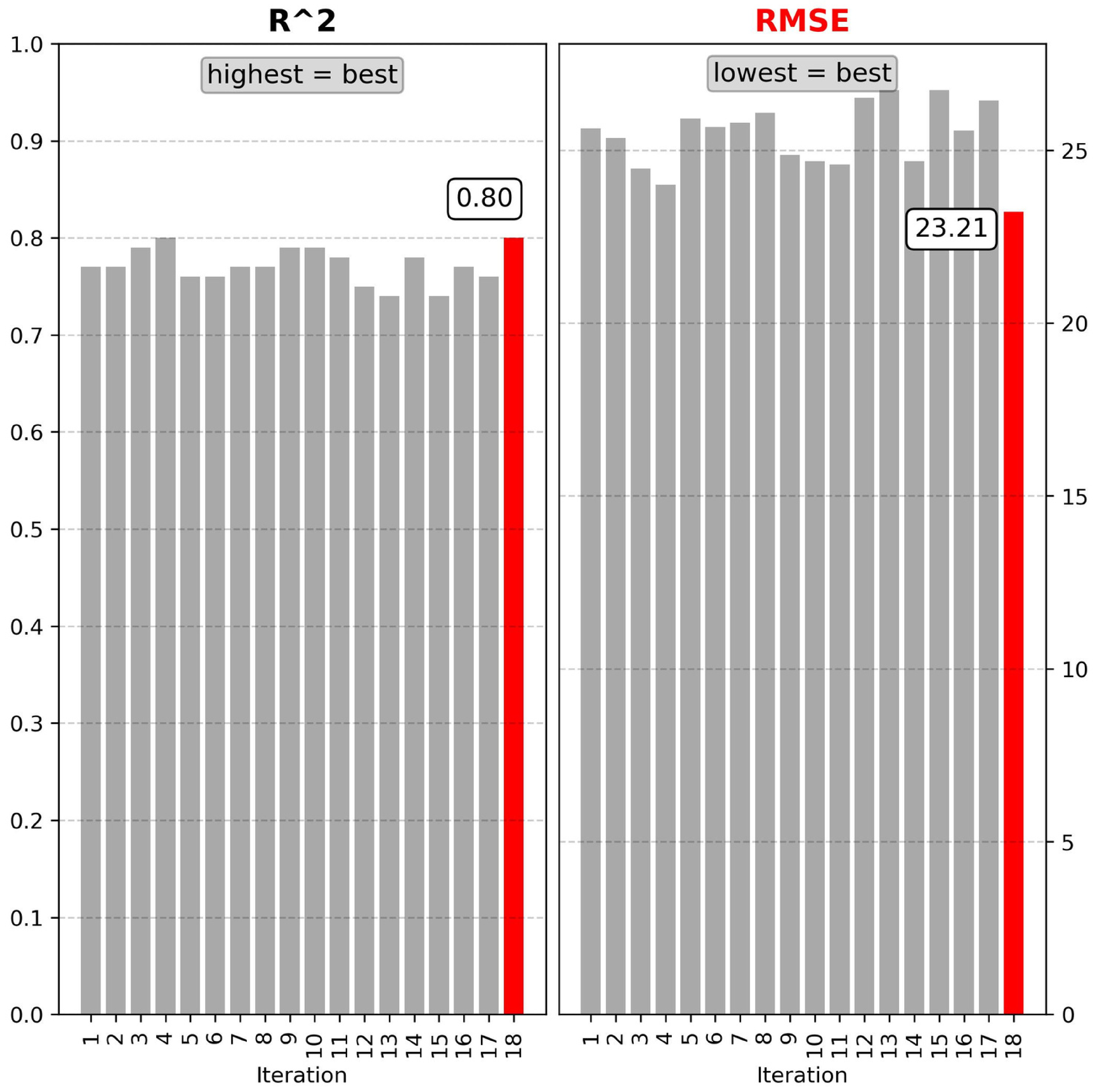
- Dropping features highly correlated to target feature
- Lots of features with (very) low correlation



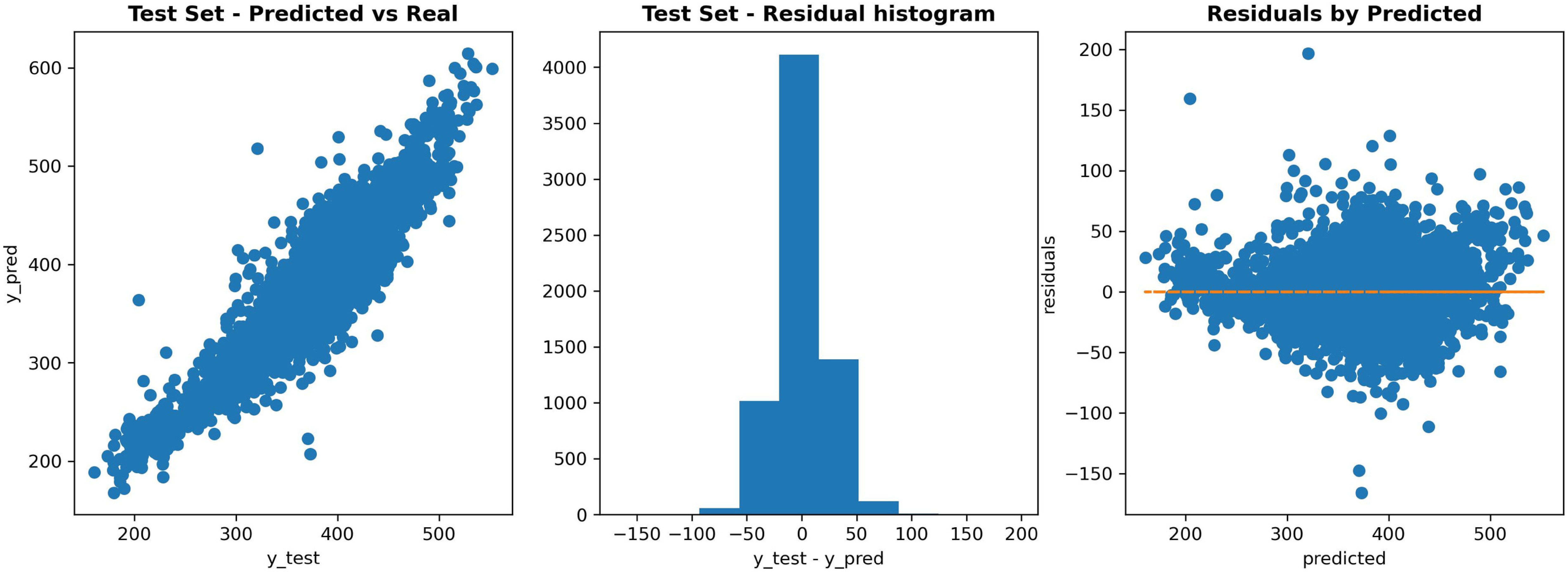
	g_param_1a	g_param_1b	g_param_2a	g_param_2b	g_param_3a	g_param_3b	g_param_3c	g_param_3d	g_param_3e	g_param_4	p_misc_phase	qc_target_mean	qc_param_1a_mean	qc_param_1b_mean
g_param_1a														
g_param_1b	0.21													
g_param_2a	0.047	0.066												
g_param_2b	0.042	0.072	0.84											
g_param_3a	0.03	-0.053	0.0019	-0.02										
g_param_3b	0.03	0.023	-0.015	-0.027	0.23									
g_param_3c	-0.087	-0.14	0.0097	-0.0022	0.073	-0.032								
g_param_3d	0.028	0.069	0.62	0.56	-0.033	-0.027	0.055							
g_param_3e	-0.006	0.0082	-0.34	-0.3	0.013	0.021	-0.075	-0.23						
g_param_4	0.026	0.027	0.91	0.77	0.032	-0.01	0.023	0.53	-0.35					
p_misc_phase	0.0034	0.00024	0.016	0.016	-0.0082	-0.0053	0.018	0.021	-0.0072	0.012				
qc_target_mean	0.07	0.1	0.87	0.73	-0.027	-0.002	-0.12	0.53	-0.33	0.85	0.017			
qc_param_1a_mean	-0.038	-0.0067	0.14	0.12	-0.017	-0.0087	-0.057	0.15	-0.061	0.15	-0.018	0.22		
qc_param_1b_mean	0.0046	-0.031	-0.4	-0.38	0.0061	0.02	-0.094	-0.25	0.12	-0.38	-0.0074	-0.2	0.16	

Lin. Reg. Model Construction & Performance

- 18 different model iterations tested
- Best model: numerical + categorical combination (18)
- R2 = 0.80
- RMSE = 23.21



Evaluating Model Performance

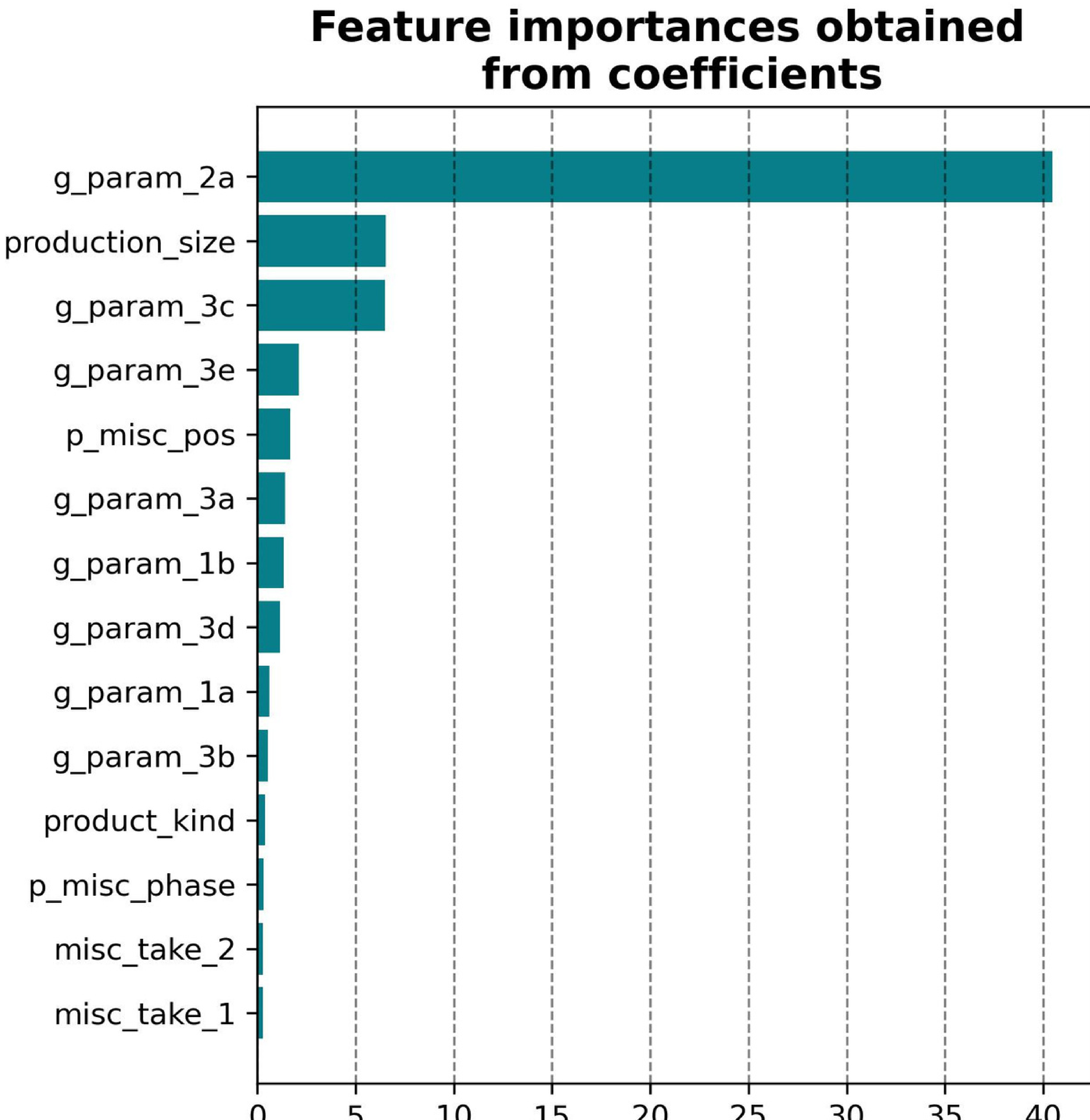


Model Presentation

Numerical & categorical data is used to build a linear regression model

Target Parameter Prediction Function:

qc_target_mean =
0.63 * g_param_1a + 1.35 * g_param_1b +
40.47 * g_param_2a -1.44 * g_param_3a +
0.53 * g_param_3b - 6.52 * g_param_3c -
1.16 * g_param_3d - 2.12 * g_param_3e +
0.32 * p_misc_phase + 0.41 * product_kind +
6.53 * production_size + 0.28 * misc_take_1 +
0.28 * misc_take_2 + 1.67 * p_misc_pos +
377.21

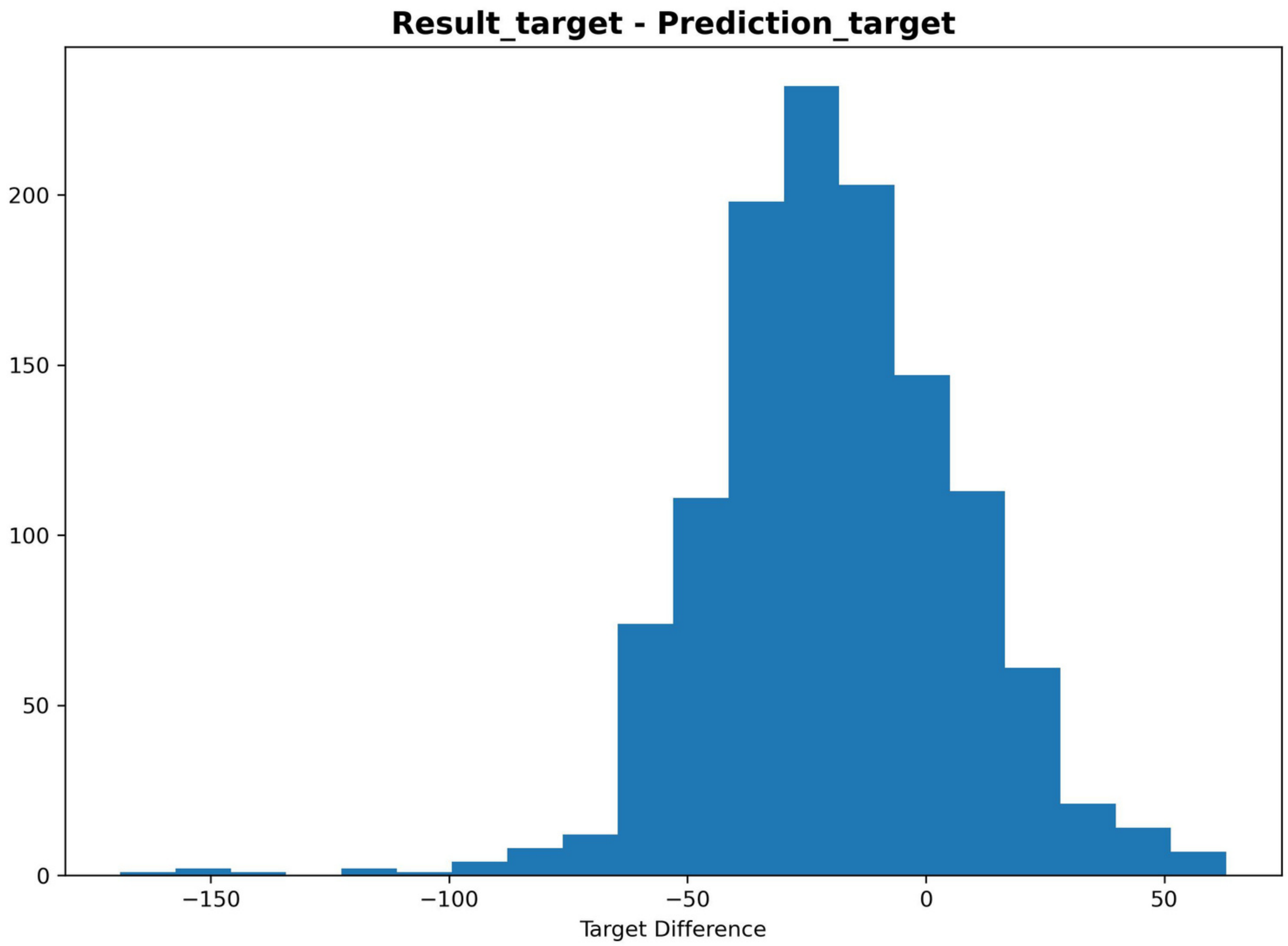


Feature importances obtained from coefficients

Testing Model's Predictive Power

Comparing actual and predicted target parameter

- Target parameter comparison (actual vs. predicted) on previously unused data
- Tested approx. 1K values
- Model overestimates target parameter by 20 units on average (~5.5%)



Learnings

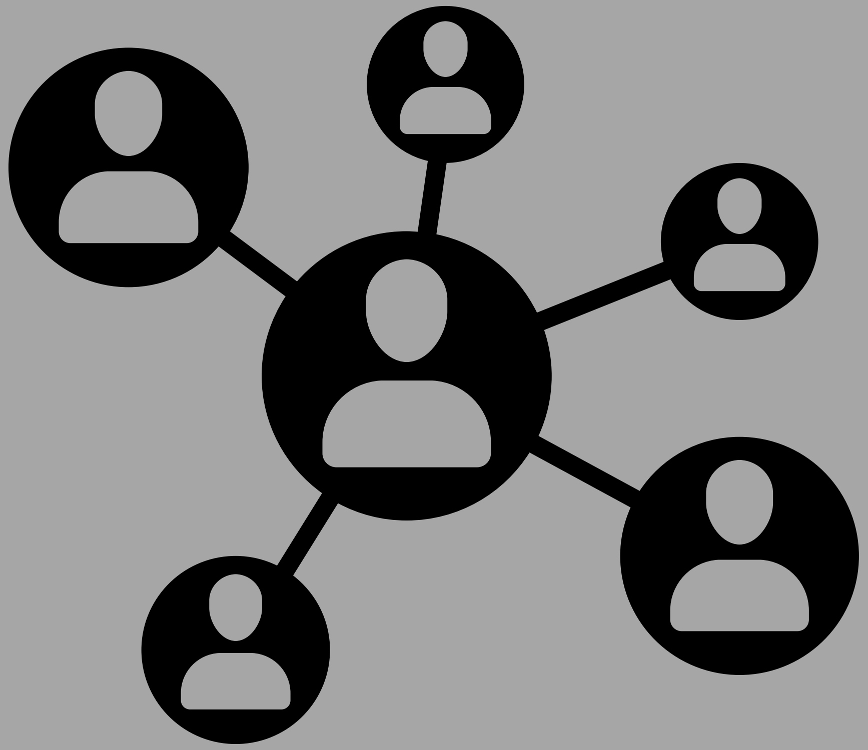
m

- Data cleaning & processing takes ton of time (GIGO)
- Clean data = easy life
- Dirty data = hard life
- Good interpolation needs to be feature-specific

Conclusion

- Solid prediction model was created
- R2 = 0.80 / RMSE = 23.21
- Model might be more potent when including other non-geometric features (makes it less representative though)
- Model overestimates target parameter by 20 units on average (~5.5%)

Outlook



- **Use different regression model(s) to possible achieve better predictive model**
  - Ridge / Lasso / Elastic Net
  - Non-linear

- **Use more sophisticated interpolation methods**

- **Introduce predictive model into manufacturing process to enhance**:
  - product quality
  - product prioritization
  - overall yield

THANK YOU