Drop volume prediction of a bioscientific printing 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/biotech company

Objective:

- Building predictive model for target feature (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

01

Step 1

Loading data and deleting features that are not useful

02

Step 2

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

03

Step 3

Dealing with outliers and NaN values by interpolation

04

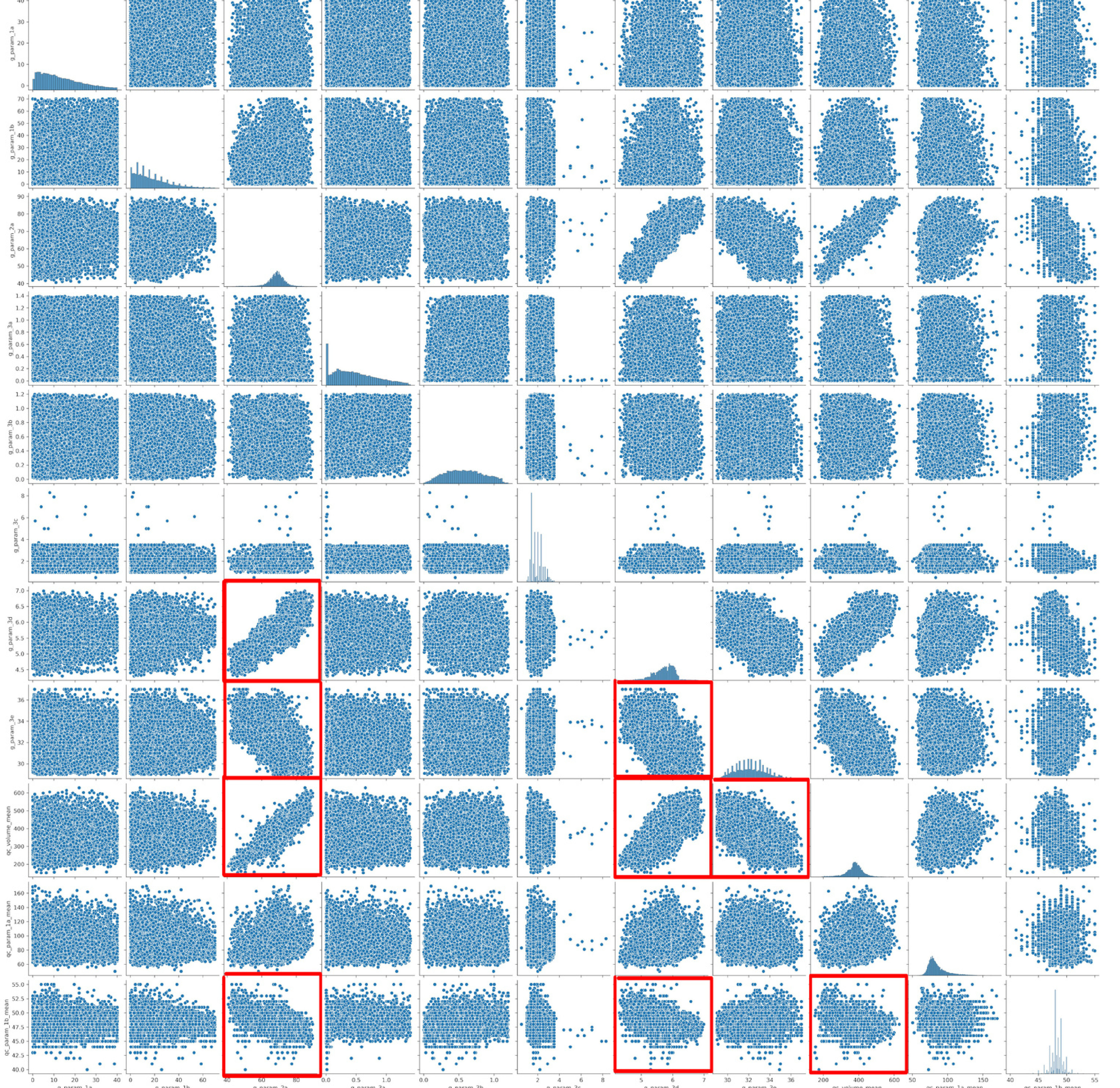
Step 4

Splitting categorical and numerical data, focusing on numerical for modeling

Feature Investigation

Investigating correlations between numerical features

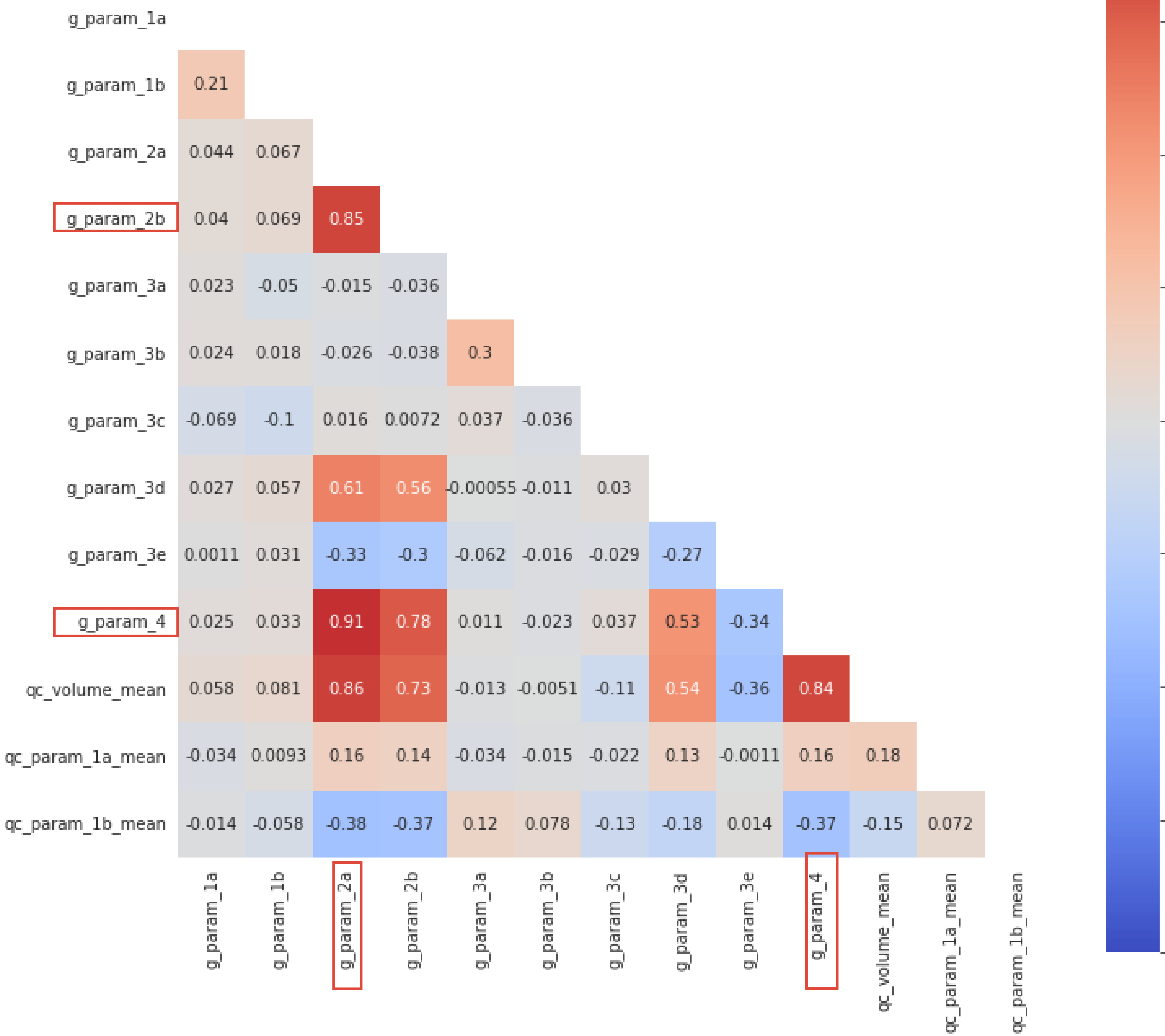
Likely linear relationship between target feature (qc_volume_mean) and other variables



Heat Map

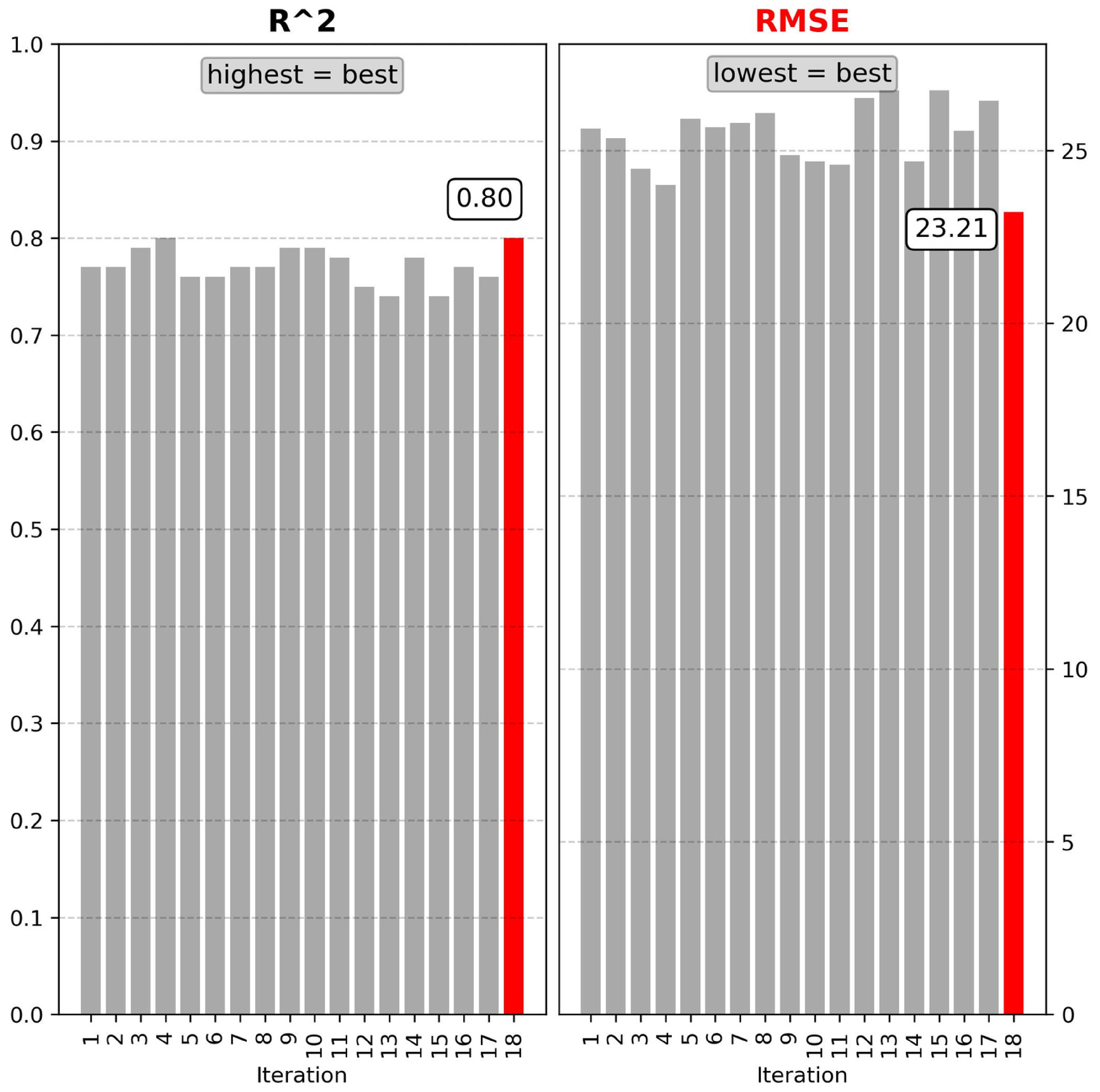
Investigating feature correlations

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

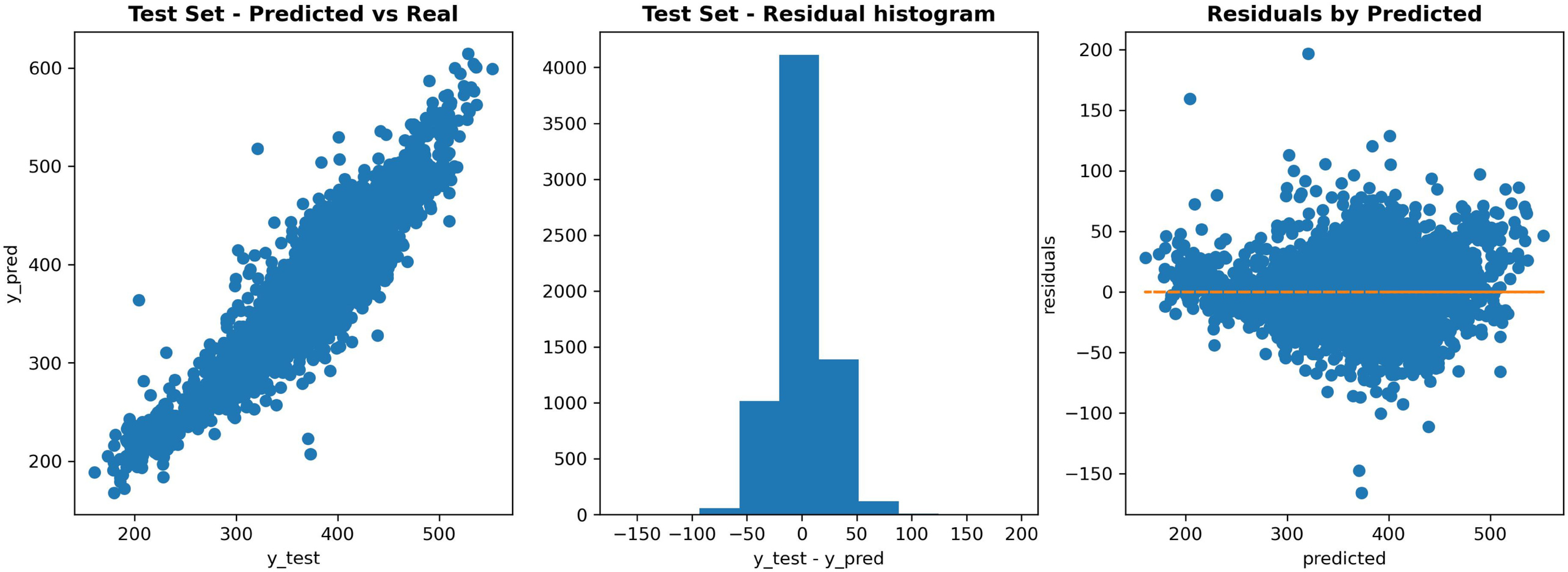


Lin. Reg. Model Construction & Performance

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



Evaluating Model Performance

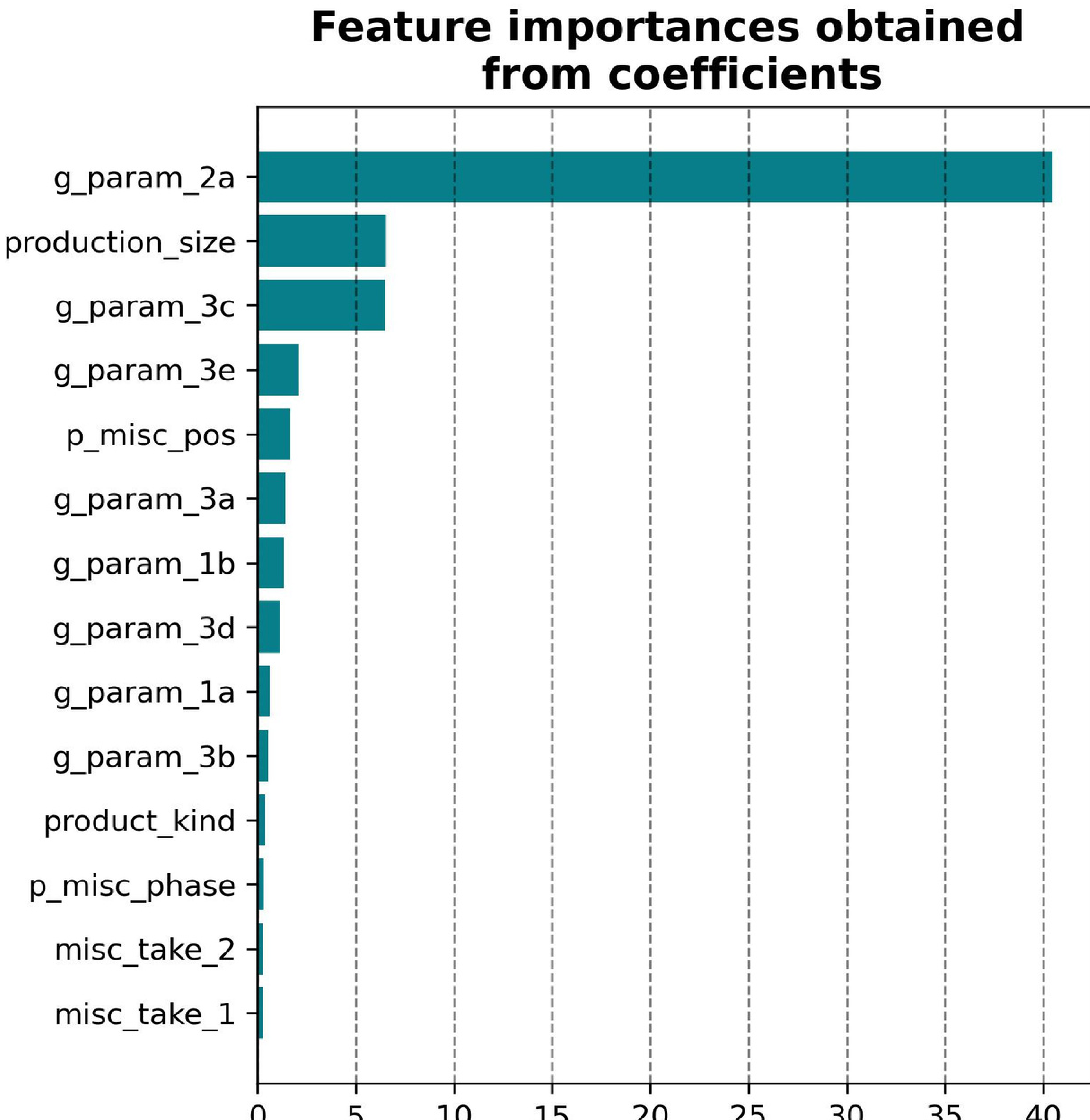


Model Presentation

Numerical data is used to build a linear regression model based on linear correlation in pairplot

Drop Volume Prediction Function:

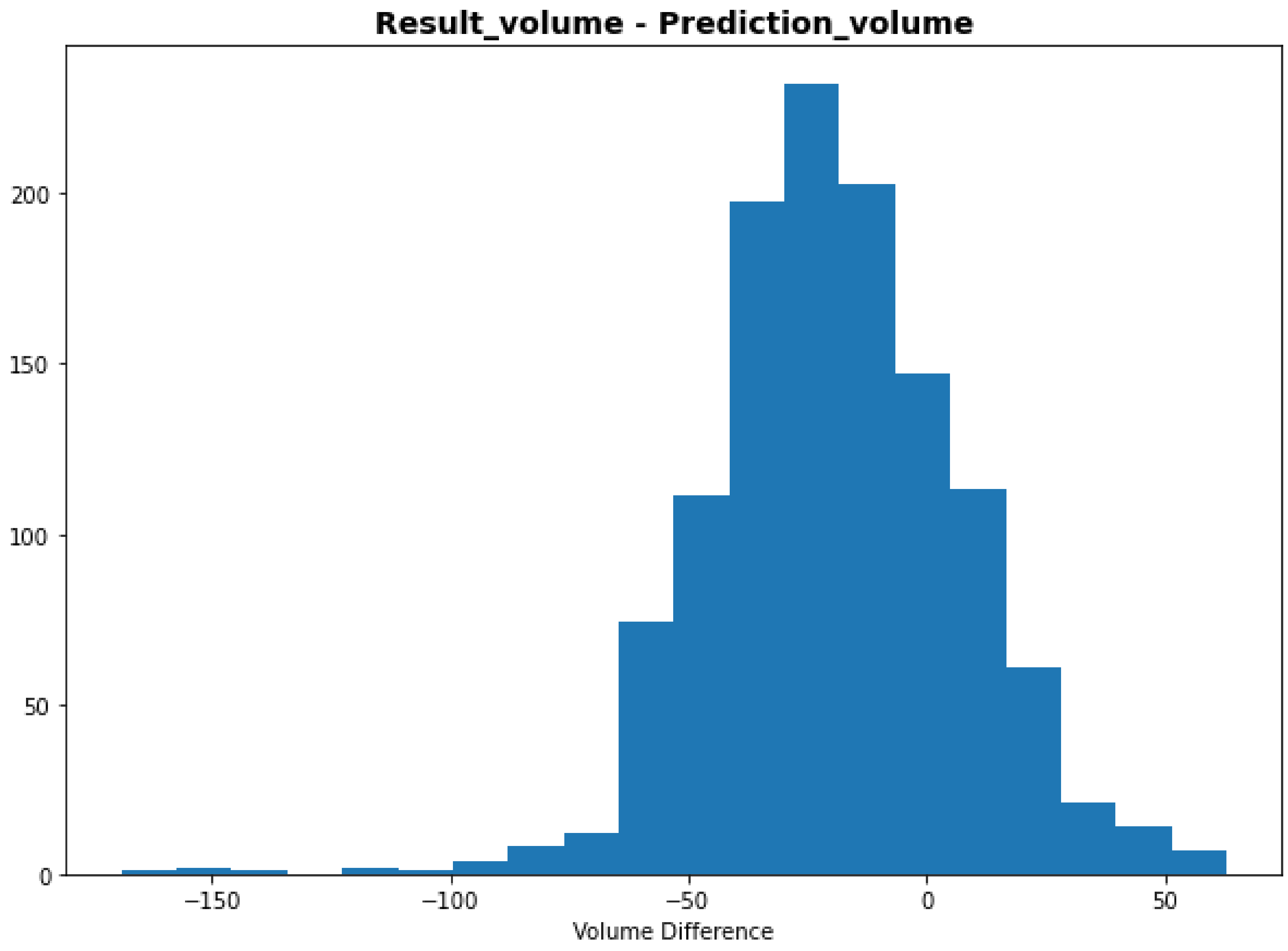
qc_volume_mean =
0.63 * g_param_1a + 1.35 * g_param_1b + 40.47 * g_param_2a -1.44 * g_param_3a + 0.53 * g_param_3b - 6.52 * g_param_3c - 1.16 * g_param_3d - 2.12 * g_param_3e + 0.32 * p_misc_phase + 0.41 * product_kind + 6.53 * production_size + 0.28 * misc_take_1 + 0.28 * misc_take_2 + 1.67 * p_misc_pos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



Testing Model's Predictive Power

Comparing actual and predicted volumes

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



Learnings

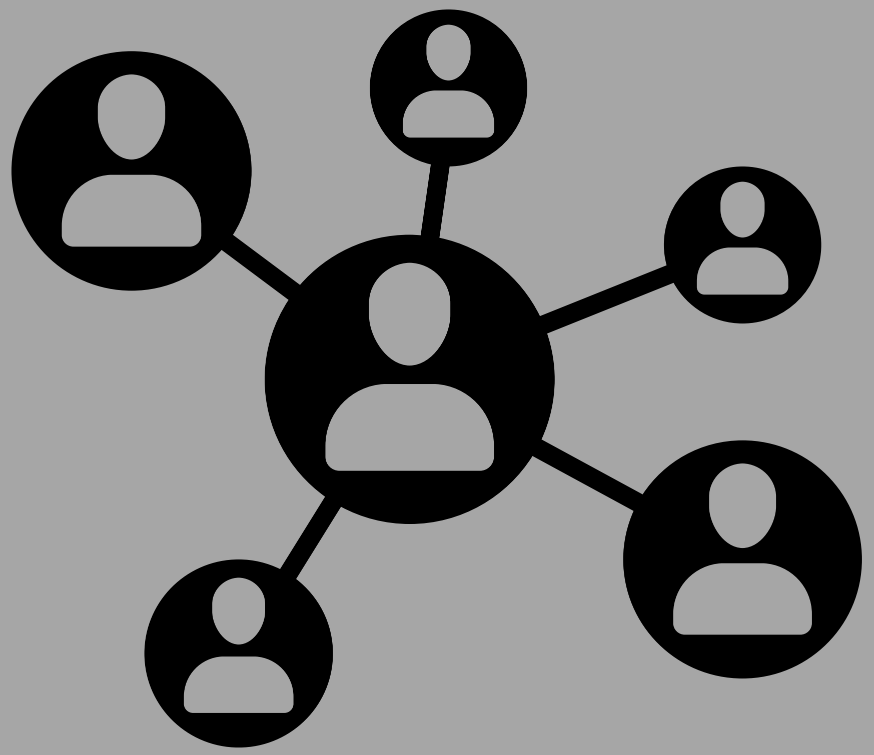
m

- 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 volume 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