Prescriptive Analytics

For Response variation reduction

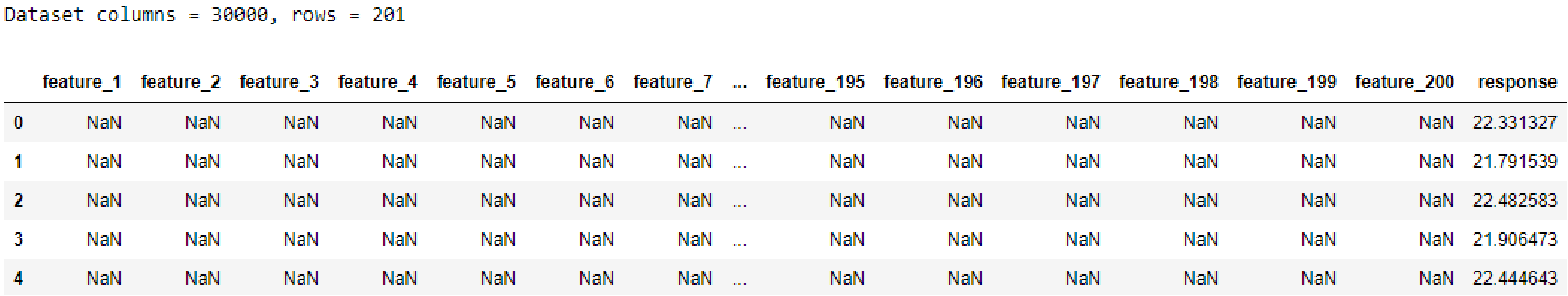
stephen.leo87@gmail.com

Problem Statement

- Customer Specs for 'response'
  - Upper Control Limit = 24
  - Lower Control Limit = 21
- Current Performance
  - ~10% of 'response' falls outside the control limits and cannot be sold to the customer
- Questions to Answer with data:
  - Which features have the most impact on response's variation?
  - How much should the variation in these features be tightened to ensure response variation meets customer specs?
- Approach:
  - 1st question to be answered with a linear model for interpretability
  - 2nd question to be answered with Monte-Carlo simulation
- Challenge:
  - Features are sampled measurements to control cost and hence have ~80% missing data

Dataset Analysis

- 201 columns
  - 1 Response column
  - 200 Feature columns
- 30K rows
  - 80:20 split between Train and Test
  - First 24K rows used to Train model
  - Last 6K rows used to Test model
  - Simulate Training model on historical data and predicting on future data
  - Many missing values observed



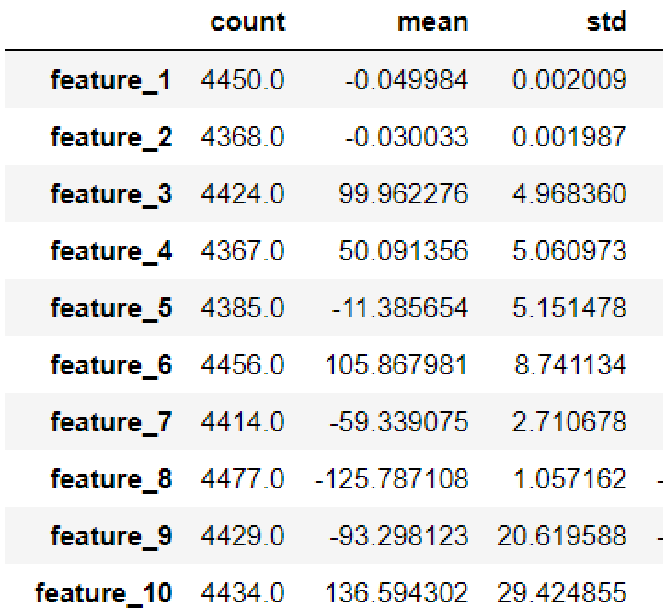
```
Dataset columns = 30000, rows = 201
```

	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	...	feature_195	feature_196	feature_197	feature_198	feature_199	feature_200	response
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	22.331327
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	21.791539
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	22.482583
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	21.906473
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	22.444643

Features Analysis

- 200 features available.
- Approximately 80% of values missing for each feature.
- Probably due to sampling of measurements.
- Different features have different magnitude of measurement.

Mean missing = 81.39%
Min missing = 65.23%
Max missing = 81.98%



	count	mean	std
feature_1	4450.0	-0.049984	0.002009
feature_2	4368.0	-0.030033	0.001987
feature_3	4424.0	99.962276	4.968360
feature_4	4367.0	50.091356	5.060973
feature_5	4385.0	-11.385654	5.151478
feature_6	4456.0	105.867981	8.741134
feature_7	4414.0	-59.339075	2.710678
feature_8	4477.0	-125.787108	1.057162
feature_9	4429.0	-93.298123	20.619588
feature_10	4434.0	136.594302	29.424855

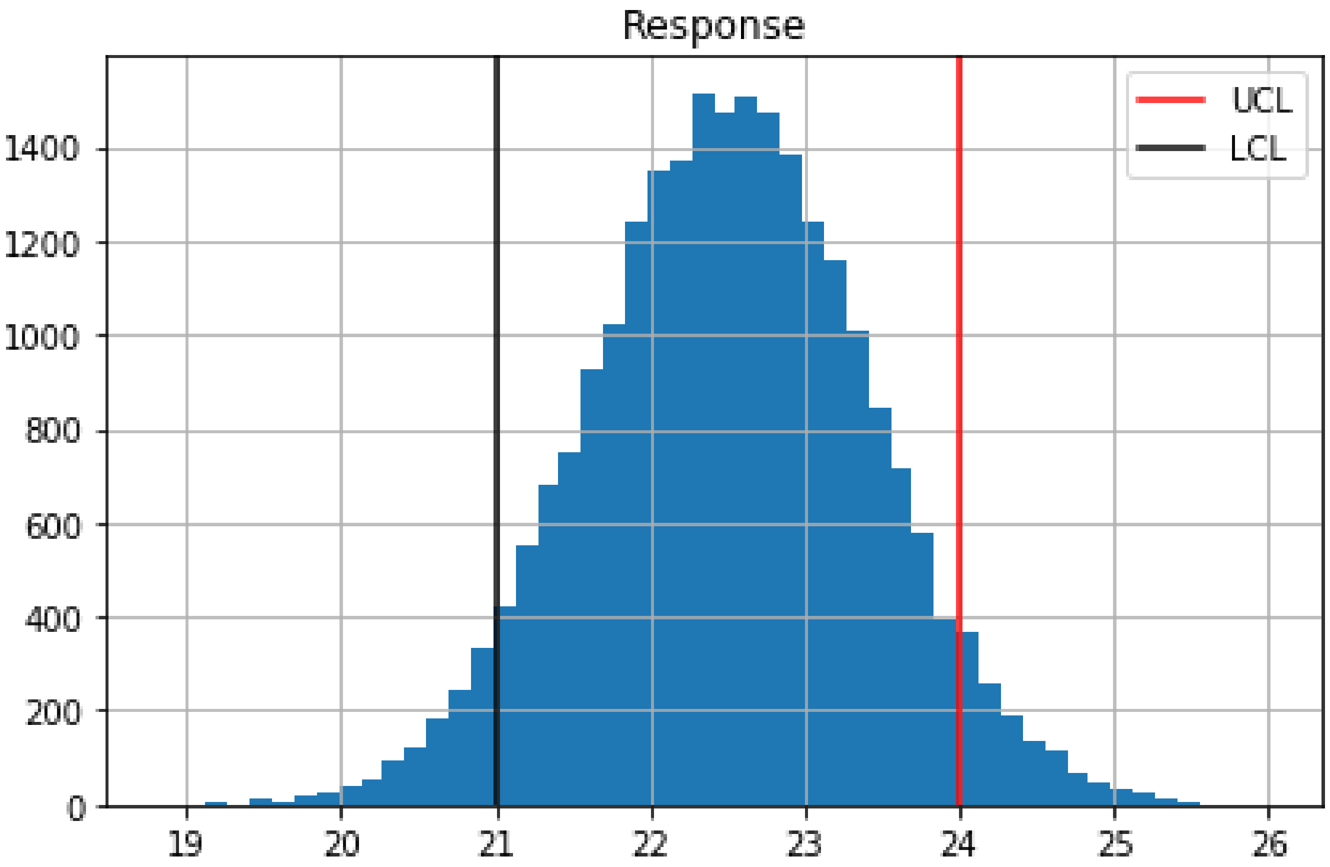
Response Analysis

- Response has normal distribution.
- Approximately 10% of Response falls outside customer specs

Customer Specs:
LCL = 21
UCL = 24

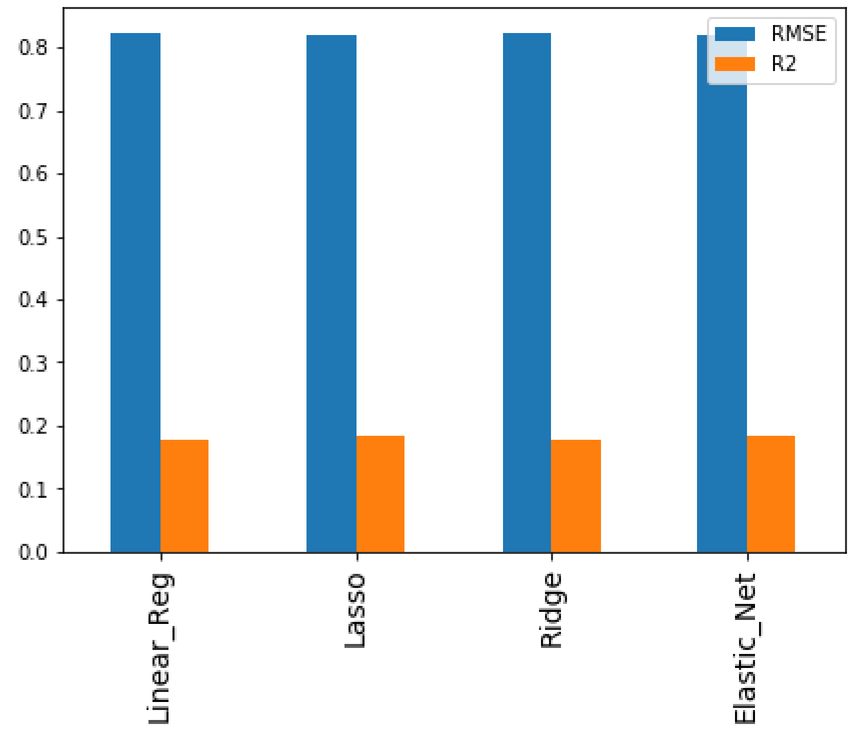
Hence,
Desired response stddev = (UCL-LCL)/6 = 0.5

Response Mean: 22.51, StdDev: 0.91
Below LCL(21): 4.92%, Above UCL(24): 5.01%



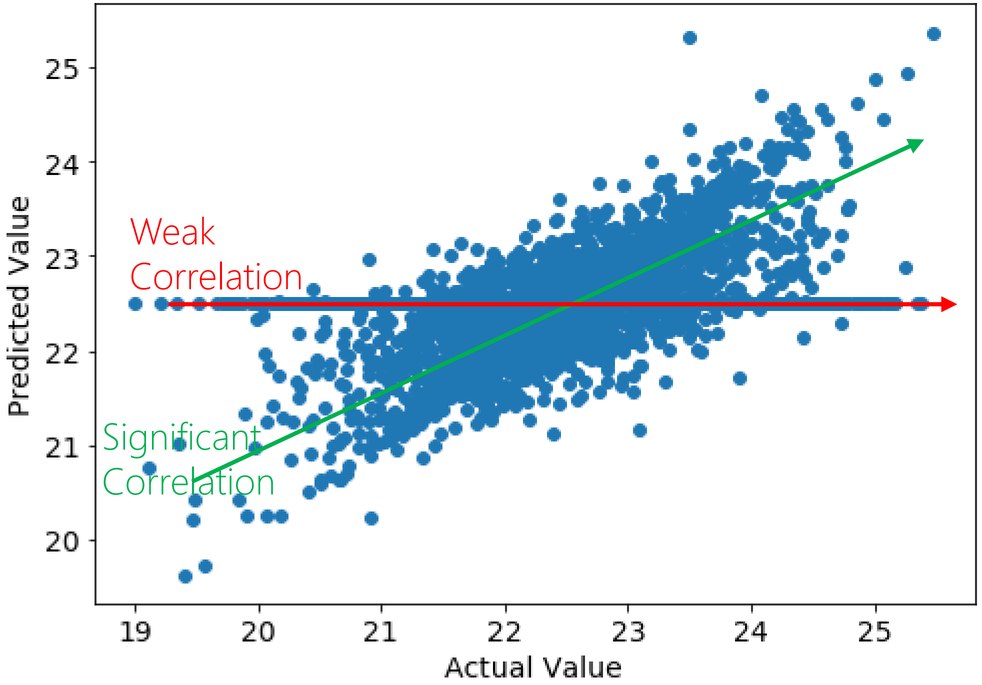
Initial Model

- Missing data filled with column mean.
- Choose Linear Models to have interpretable results
- All Models have high error (high RMSE)
- All Models have low correlation between model's predicted value and actual value (low R2)



Model Investigation

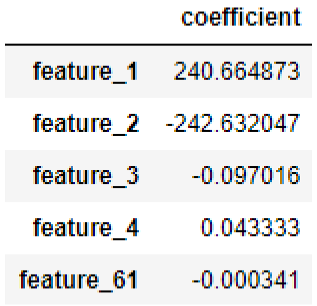
- Investigate further into 1 model to find reason for high error and low correlation
- Plot the actual correlation between model's predicted value vs actual value
- Correlation looks significant for many points
- However, there is a second distribution that has weak correlation
- This is an artifact of filling missing data with the column mean.
- Let's change the prediction to only predict the response when the row has values for all the columns used by the model.



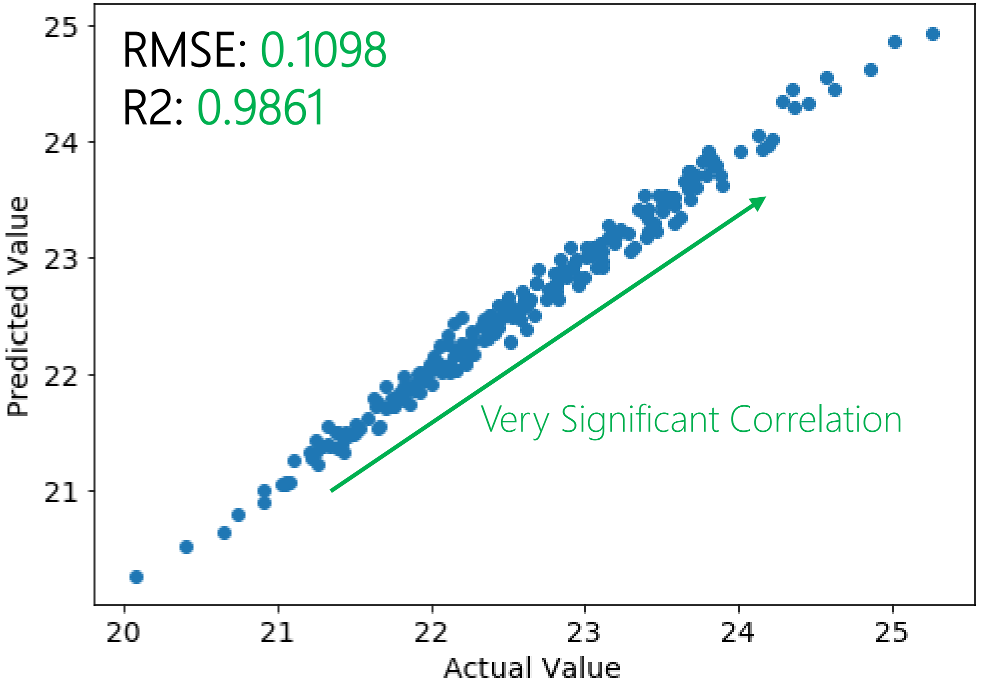
Lasso Model

- Lasso model has automatic feature selection and only uses features that are important for prediction.
- The model picks up 5 features as important. Coefficients are the Model's estimate of rate of change of response per unit change in feature.
- For these 5 features, drop the rows that have missing values and plot predicted vs actual again
- Very high correlation and low error observed
- Hence the model is very good at predicting those rows that have measurements for all 5 important features
- **Recommendation:** Ensure measurement sampling can measure all 5 features as much as possible

Important Features



	coefficient
feature_1	240.664873
feature_2	-242.632047
feature_3	-0.097016
feature_4	0.043333
feature_61	-0.000341

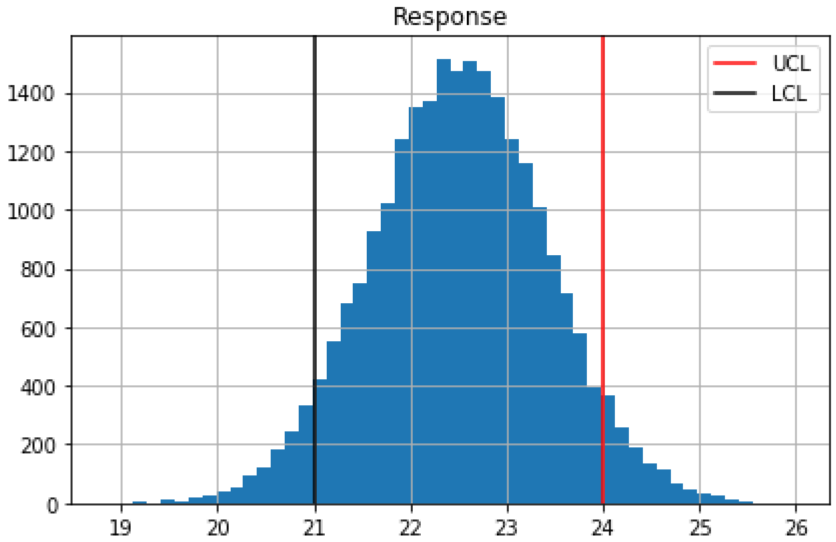


RMSE: 0.1098
R2: 0.9861

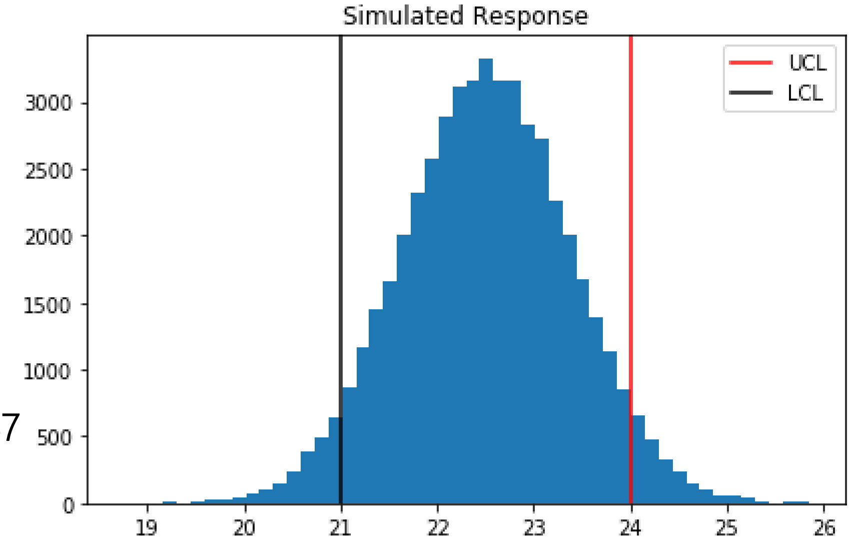
Very Significant Correlation

Monte-Carlo Simulation

Response Mean: 22.51, StdDev: 0.91
Below LCL(21): 4.92%, Above UCL(24): 5.01%



- Run simulation 50K times
- Each time, randomly choose a value from each of the 5 important features and calculate simulated response using the model parameters.
- Compare final simulated response to actual response and see that they are quite similar which indicates simulation is successful.

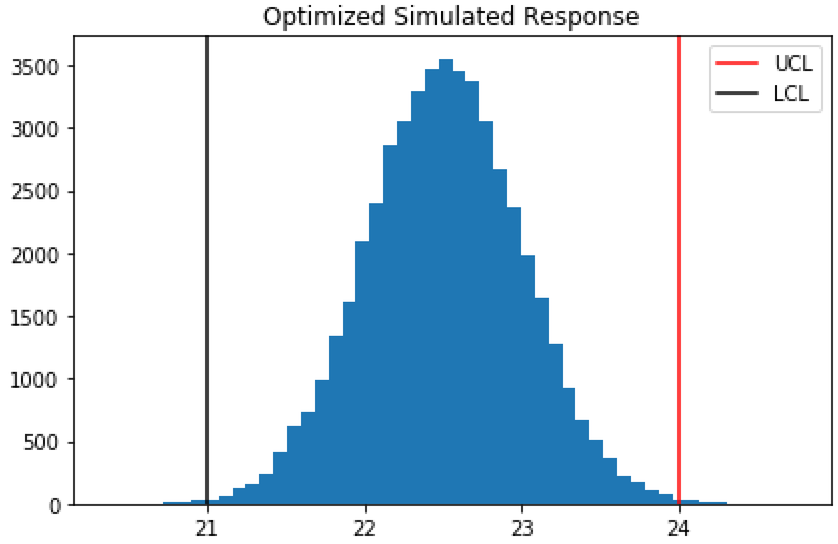


Simulated Response Mean: 22.51, StdDev: 0.87
Below LCL(21): 4.19%, Above UCL(24): 4.29%

Optimization

- Run optimization by reducing the stddev of each of the 5 important features and calculate stddev of simulated response.
- Stop when the stddev of simulated response is lower than the customer specs (0.5)
- StdDev of each feature needs to divide by: 1.80 to get Final Simulated Response StdDev: 0.49.
- The desired std dev for each of the important features are shown in the table. If these conditions are met, only 0.2% of response will fall outside customer specs.

Optimized Simulated Response Mean: 22.51, StdDev: 0.49
Below LCL(21): 0.11%, Above UCL(24): 0.12%



	Current StdDev	Desired StdDev
feature_1	0.002009	0.001116
feature_2	0.001987	0.001104
feature_3	4.96836	2.7602
feature_4	5.060973	2.811652
feature_61	4.053497	2.251943

Conclusion & Recommendations

- Conclusions:
  - Response can be predicted with low error stddev of 0.1 using 5 features: feature_1,2,3,4,61.
  - Desired stddev of important features are shown in the table. Each stddev needs to divide by 1.8 of current value.
  - Response not meeting customer specs can be drastically reduced from ~10% to 0.2% if all features meet the desired stddev values.
- Recommendations:
  - Ensure measurement can measure all 5 important features as much as possible
  - Control 5 important features within desired stddev values

Optimized Simulated Response:
Below LCL(21): 0.11%, Above UCL(24): 0.12%

	Current StdDev	Desired StdDev
feature_1	0.002009	0.001116
feature_2	0.001987	0.001104
feature_3	4.96836	2.7602
feature_4	5.060973	2.811652
feature_61	4.053497	2.251943

Thank You