Predicting Performance of Integrated Circuits using Regression Analysis

DS6372: Project 2

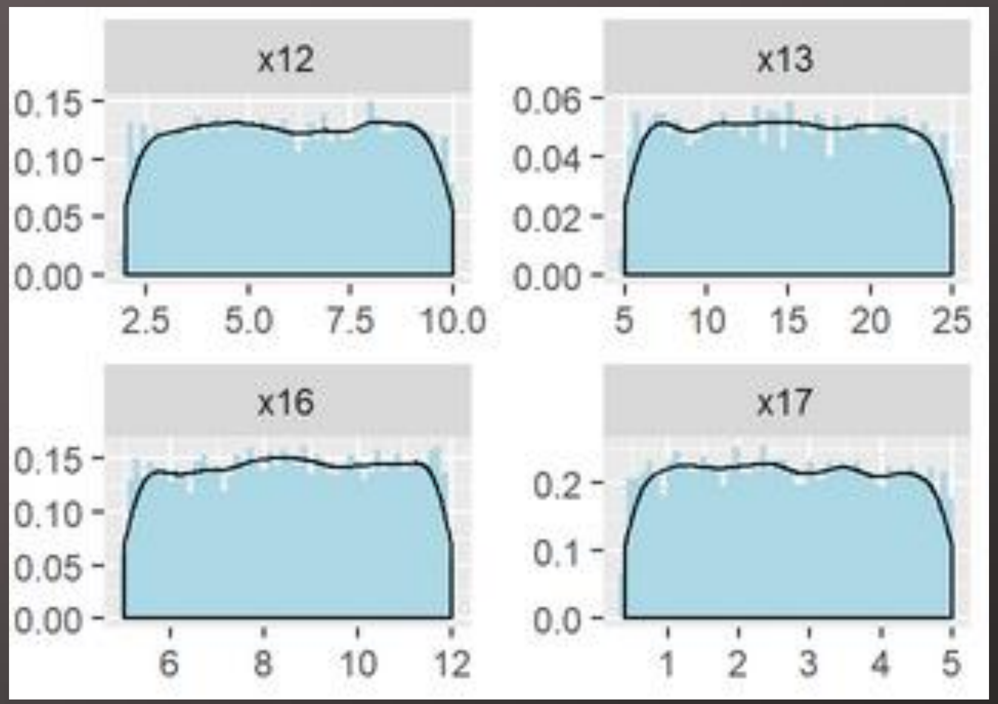
Nikhil Gupta, Max Moro, Joanna Duran

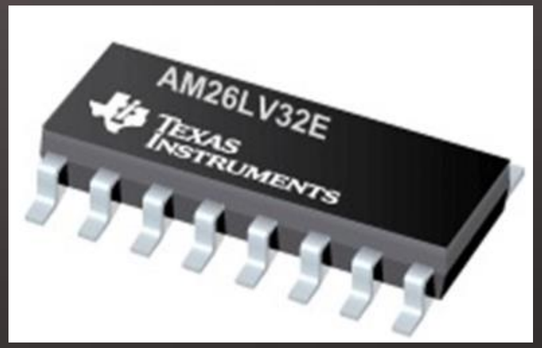
AGENDA

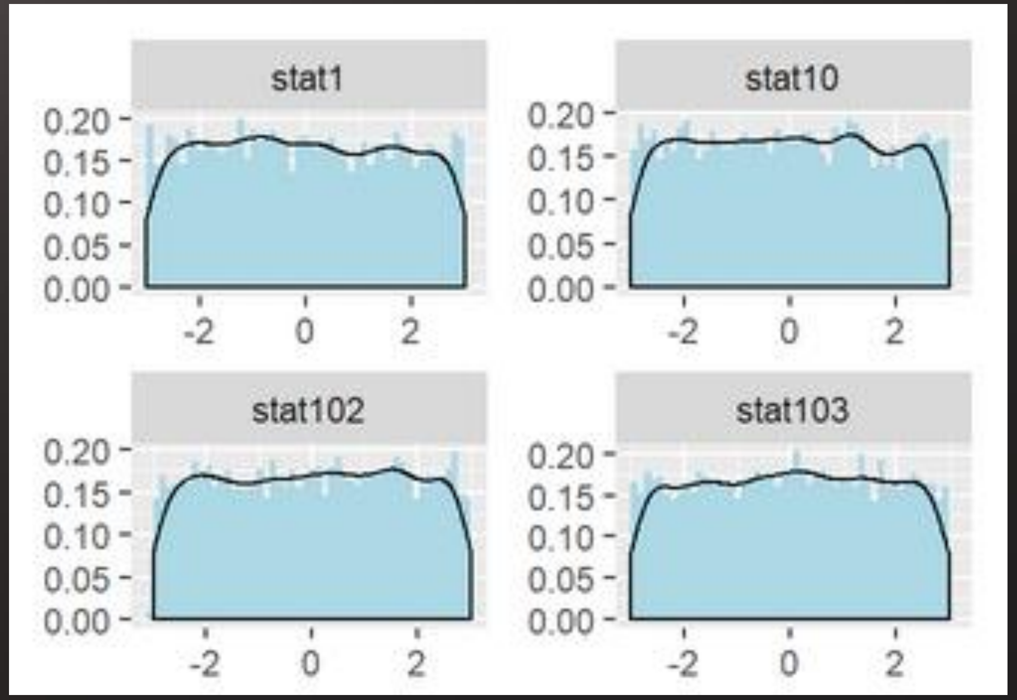
INTRODUCTION

- Data has been sponsored and approved for use by Texas Instruments Inc.

- 10,000 rows capturing the performance of a circuit under various conditions.
- 240 features
  - x1 - x23: values that can be controlled by engineers to tune the performance
    - Wide range of values, some ranging from to 100, others are in the Nano or Micro range
  - stat1 – stat217: process variation parameters beyond human control
    - Range from -3 to 3, with mean of 0
    - Values represent sigma variation around mean
  - Output variable to target: y3 (most critical)



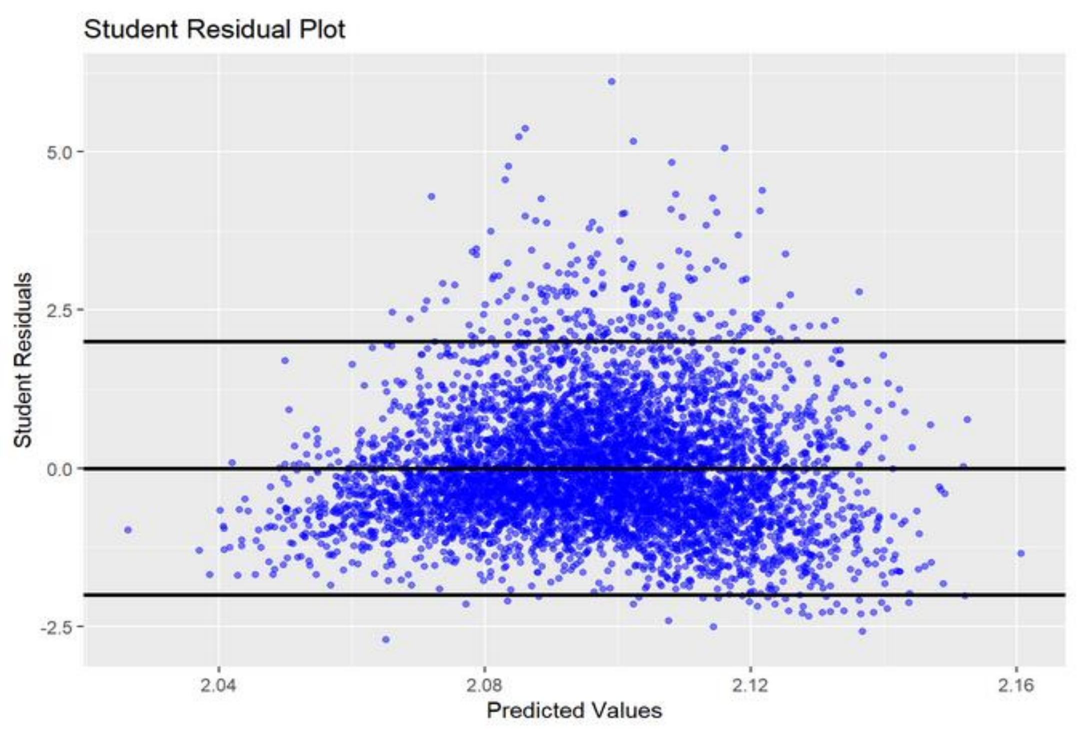


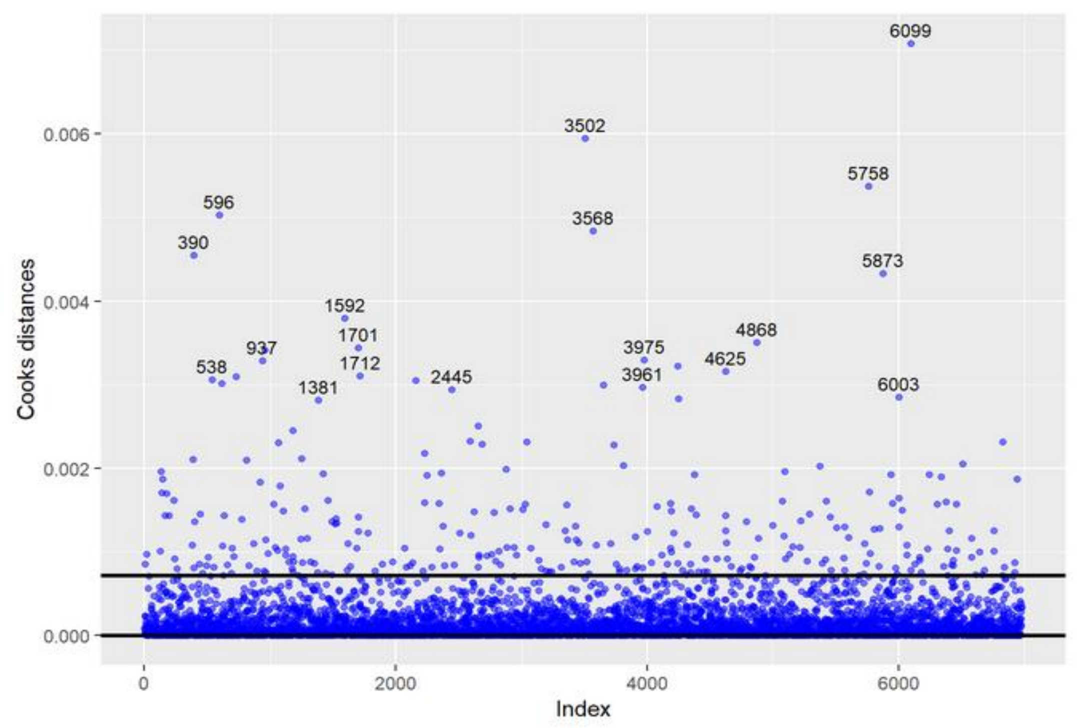


- Data Collection Process
  - Engineer controlled variables were uniformly randomly selected
  - Statistical features were uniformly randomly sampled

PROJECT 1

Synopsis



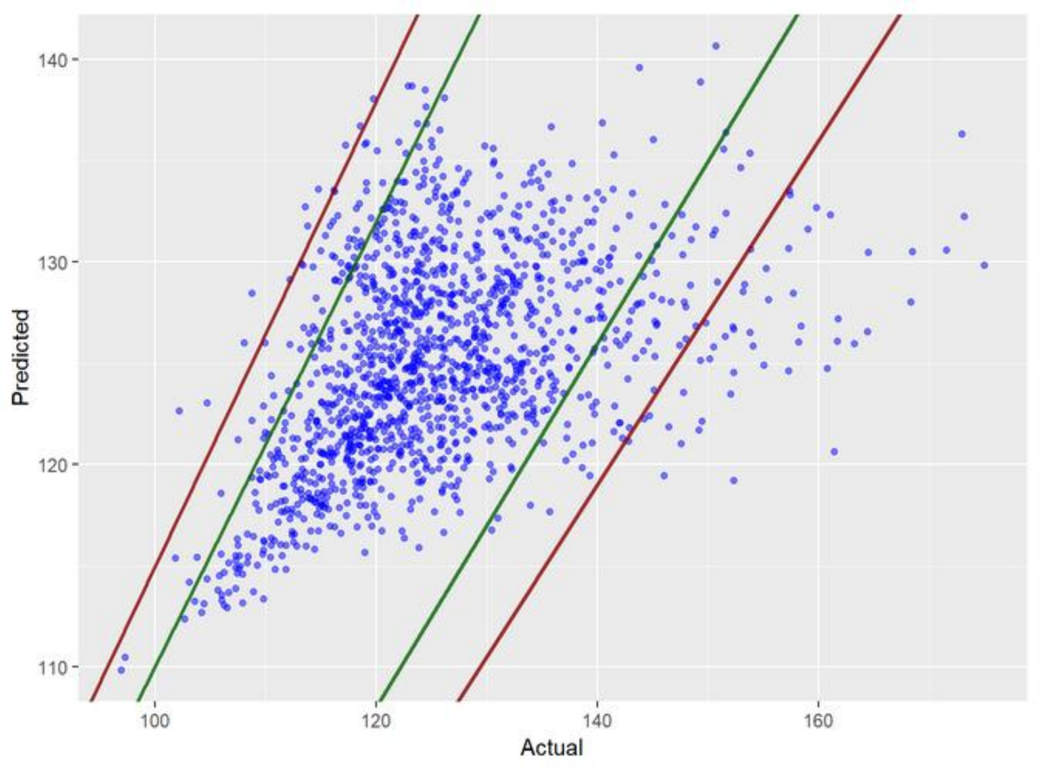


- Cleaned data
- Exploratory data analysis on the target variable y3 and transformed variable to "y3.log"
- After running the analysis, the correlation table and VIF values confirmed that there is no multicollinearity.
- None of the features were highly correlated to the output, hence we had low prediction capability.
- The full model analysis was performed. The adjusted R^2 is 0.2275 which is low and expected.
- Full model suffers from assumption violations
- The influence statistics were also analyzed and from the Cook's D plot we notice that there are 288 points beyond the 4/n line). But there is no justification to remove them.
- Could not perform 2-way interaction of variables.

PROJECT 1

Synopsis - Continued

Algorithm	# Variables	R^2	RMSE (Train)	MAE (Train)	RMSE (Test)
Full Model	240	0.238	0.0314		
Forward Selection	10	0.233	0.0316	0.0241	0.0321
Backward Elimination	10	0.233	0.0316	0.0241	0.0321
Stepwise Selection	10	0.233	0.0316	0.0241	0.0321
LASSO	44	0.232	0.0316	0.0242	0.0320
LARS	39	0.232	0.0316	0.0242	0.0320



- The variable selection techniques that were used were
  - Forward Selection
  - Backward Elimination
  - Stepwise Selection
  - LASSO
  - LARS
- We found that all techniques give essentially the same fit statistics. The best model is Backward Elimination based on low # predictors.
- In analyzing the residuals, we conclude that the final model suffers from same assumption violations as the full model.

PROJECT 2

- **GOAL**
  - Improve the model fit from Project 1 utilizing Principal Component Regression, Linear/Quadratic Discriminate Analysis or Logistic Regression.
  - Target accuracy of ±10% is desired, but ±15% would be acceptable.

- **APPROACH 1:**
  - Principal Component Regression with intelligent feature engineering and selection

- **APPROACH 2:**
  - Clustering based on High Cook's D and using the cluster as a predictor in the regression model

RESEARCH METHODOLOGIES

- ***Modeling Approach 1***
  - Performing intelligent feature engineering using semiconductor domain expertise, which included theoretical equations from semiconductor.
  - Created variables then performed 2-way interaction of all the variables.
  - Dimensionality reduction using Principal Component Analysis (PCA).
  - The reduced Principal Components (PCs) were then used to build a linear regression model.
- ***Modeling Approach 2***
  - From project 1 observations as high leverage or low leverage.
  - This data was then used to train various classification models including Linear/Quadratic Discriminate Analysis (LDA/QDA) and Logistic Regression models.
  - This class variable could then be used in the eventual linear regression model to improve predictions.

EXPLORATORY DATA ANALYSIS

Data Prep

- 3020 NA values for y3

Output

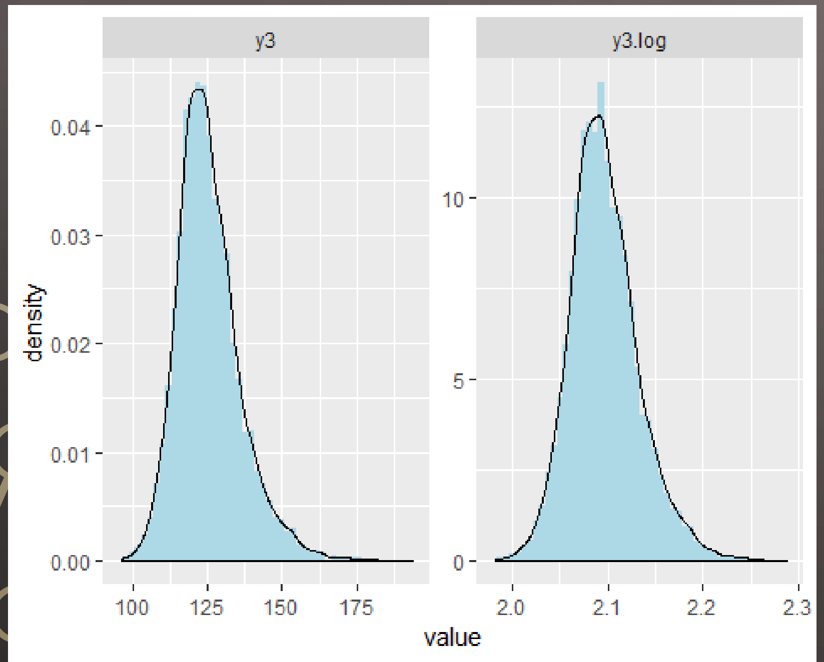
- Right skewness
- Log transformed variable “y3.log”

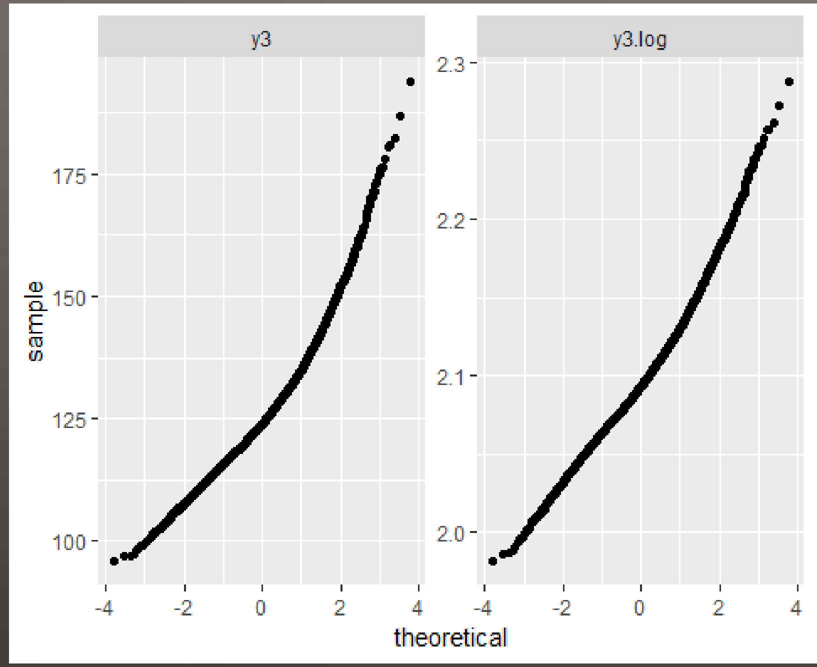
Input Predictors

- Multicollinearity
- All VIF values were < 10

Feature Engineering

- Derived from semiconductor theory
- 41 features created





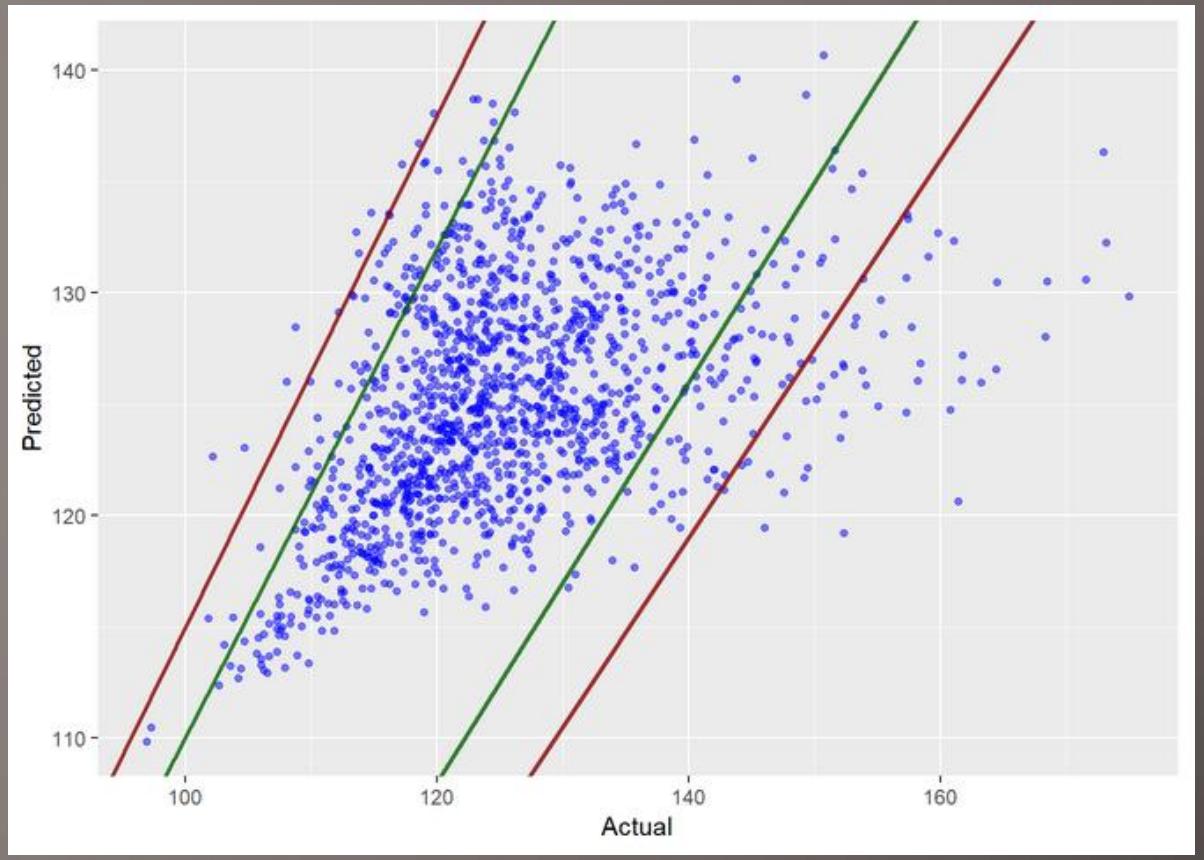

```
##     Variables       VIF
## 1     stat204 1.063870
## 2     stat175 1.063370
## 3      stat66 1.062060
## 4     stat105 1.062008
## 5          x6 1.061394
## 6       stat2 1.061388
## 7      stat14 1.061212
## 8          x7 1.060532
## 9     stat216 1.060477
## 10    stat142 1.060190
## 11    stat154 1.059695
## 12     stat32 1.059608
## 13    stat141 1.059564
## 14    stat138 1.059507
## 15     stat73 1.059386
```

```
data$x2byx1 = data$x2/data$x1
data$x6byx5 = data$x6/data$x5
data$x9byx7 = data$x9/data$x7
data$x10byx8 = data$x10/data$x8
data$x14byx12 = data$x14/data$x12
data$x1sqinv = 1/(data$x1)^2
data$x5sqinv = 1/(data$x5)^2
data$x7sqinv = 1/(data$x7)^2
data$x8sqinv = 1/(data$x8)^2
data$x1log = log(data$x1)
data$x20log = log(data$x20)
data$x21log = log(data$x21)
data$x22log = log(data$x22)
data$x23log = log(data$x23)
data$x11log = log(data$x11)
```

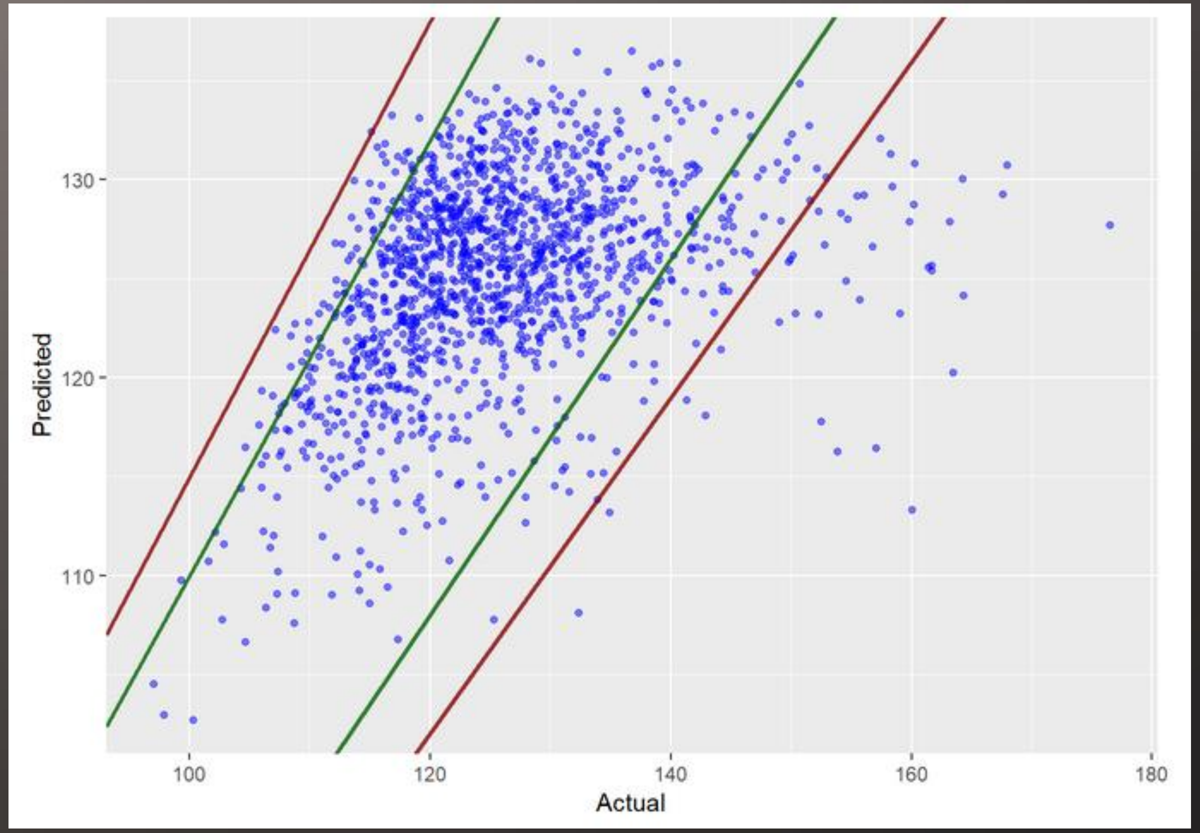
MODEL RESULTS (MODEL 1)

- Tried taking 2-way interactions of only the engineer-controlled variables (x1 – x23) amongst themselves along with the standalone statistical variables (Model ID 1).
- Then performed Principal Component Analysis (PCA) and Regression (PCR).
- However, did not improve the model fit compared to the best model from project 1.

ID #	Model (with PCA)	# VARS	Train R^2	Test RMSE
1	*Full Model (2-way EC + Stat)*	164	0.266	9.45



Predicted vs. Actual Values (Project 1 best model)

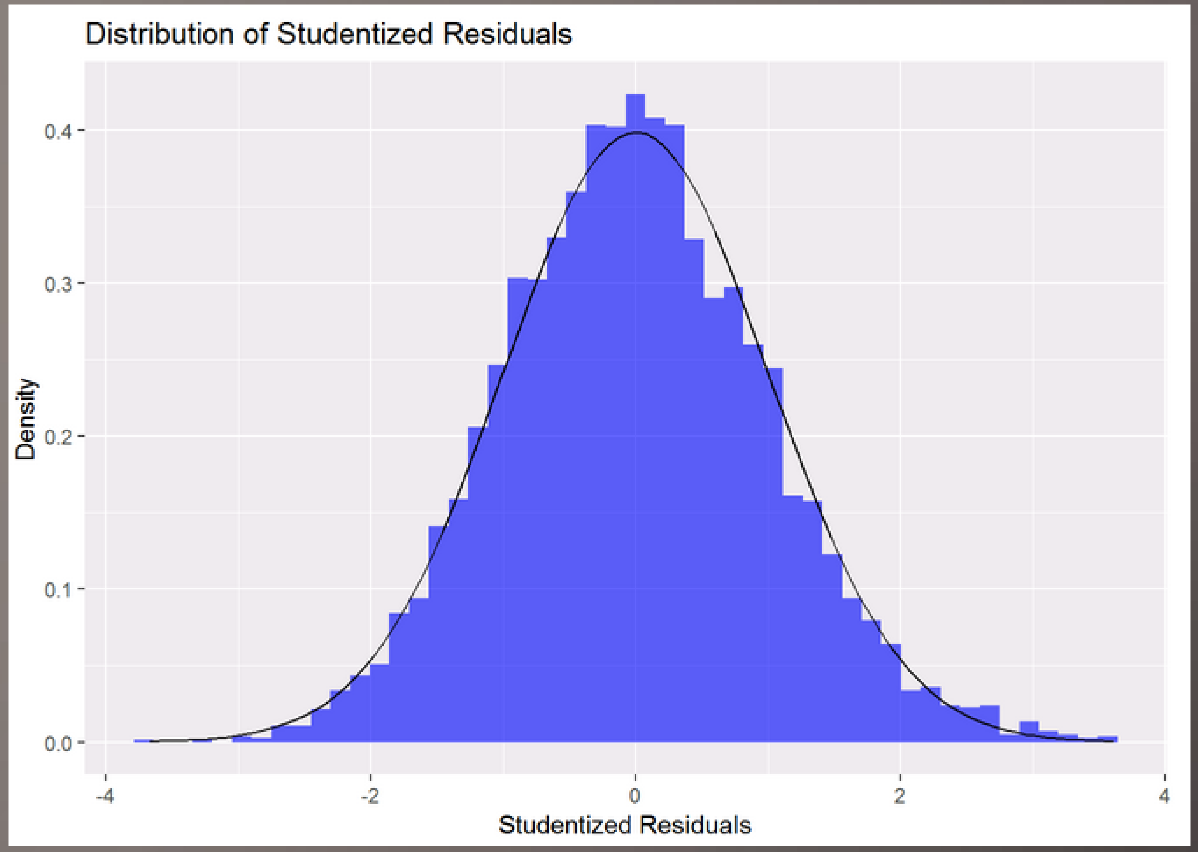


Predicted vs. Actual Values (ID 1 - Full Model 2-way EC + Statistical variables)

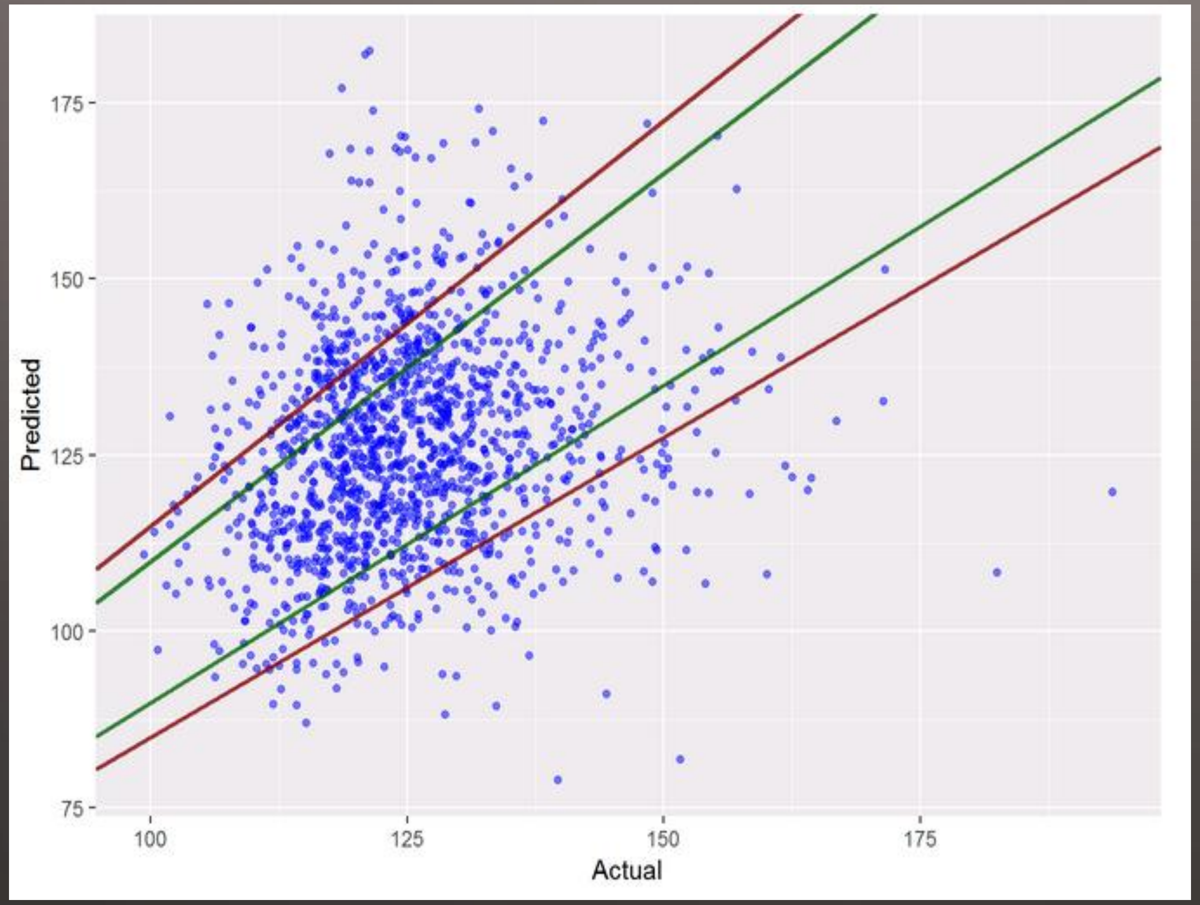
MODEL RESULTS (MODEL 2)

- 2-way interaction of all 240 predictors followed by PCA and PCR
- Improvements to assumption violations: The residuals were normally distributed and showed equal variance
- However, the model was overfitting as the test RMSE saw a degradation

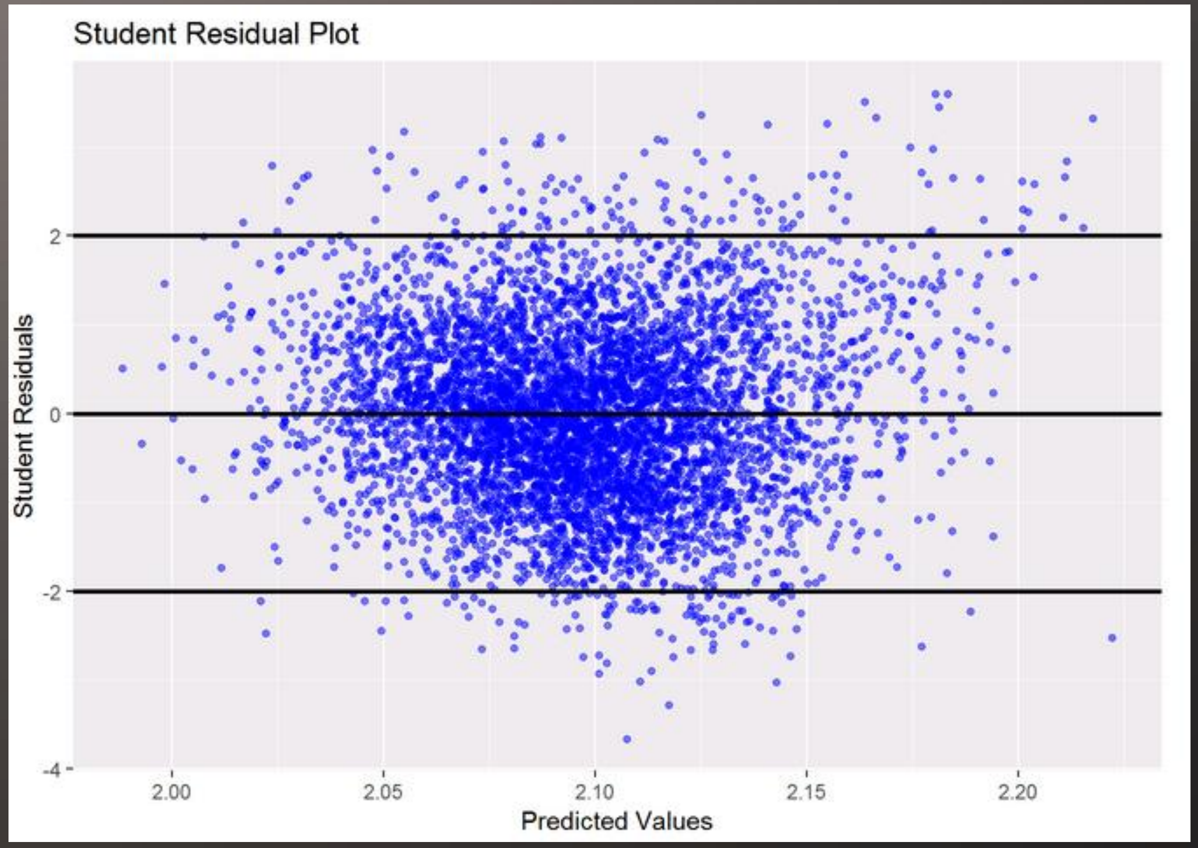
ID #	Model (with PCA)	# VARS	Train R^2	Test RMSE
1	*Full Model (2-way EC + Stat)*	164	0.266	9.45
2	*Full Model (2-way All)*	3,834	0.786	16.76



Histogram of Residuals



Predicted vs. Actual Values (ID 2 - Full Model 2-way interaction of all variables)



Studentized Residuals vs. Predicted Values

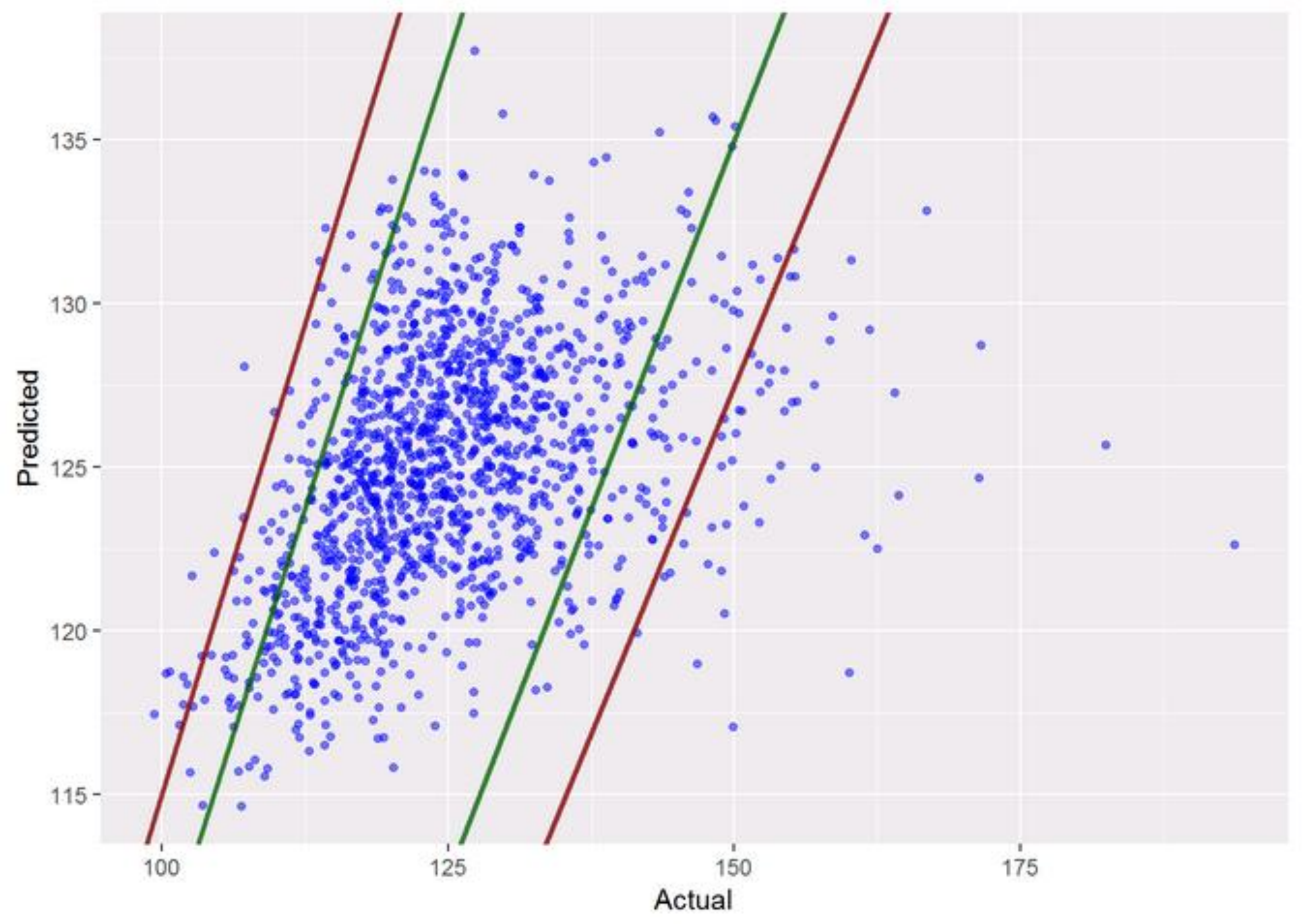
MODEL RESULTS (MODELS 3 AND 4)

- Using LASSO (ID 3) and LARS (ID4) for regularization, we were able to reduce overfitting
- Two important conclusions

  (1) Need to consider 2-way interaction of all variables, not just the engineer-controlled variables

  (2) Variable selection and regularizations techniques were necessary to prevent overfitting.

ID #	Model (with PCA)	# VARS	Train R^2	Test RMSE
1	*Full Model (2-way EC + Stat)*	164	0.266	9.45
2	*Full Model (2-way All)*	3,834	0.786	16.76
3	*LASSO (2-way All)*	3,834	0.139	9.91
4	*LARS (2-way All)*	3,834	0.139	9.91



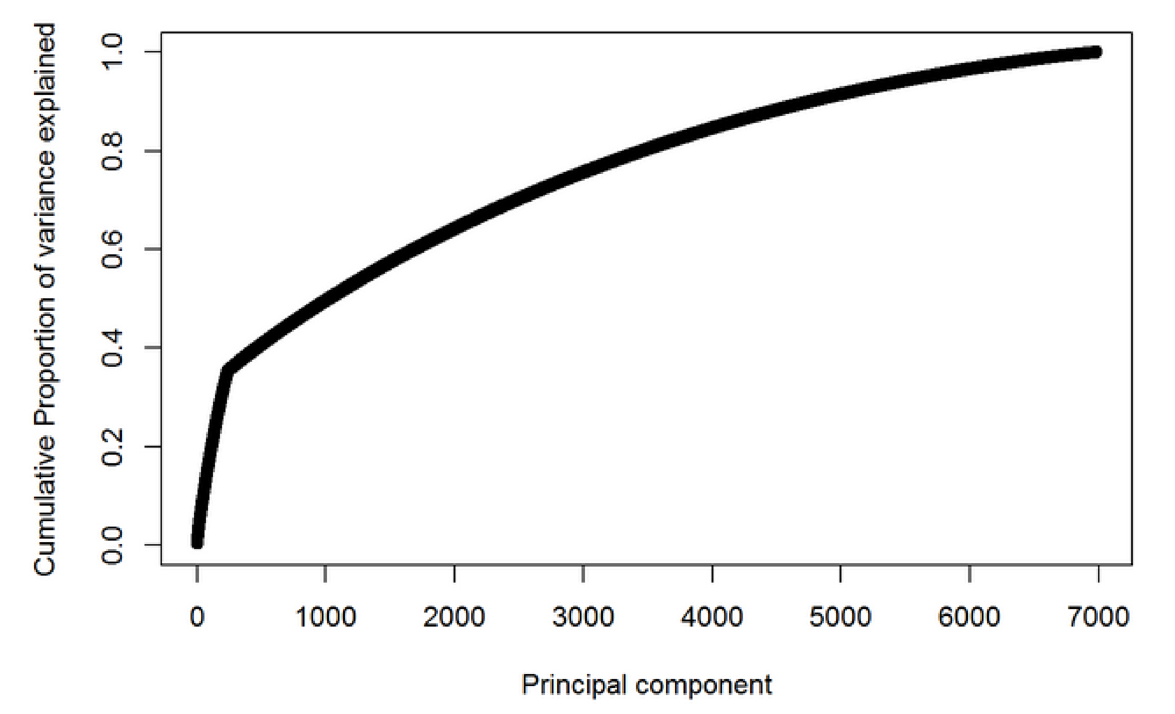
Predicted vs. Actual Values (ID 3 - LASSO 2-WAY ALL)

FURTHER IMPROVEMENTS

- Used the intelligent feature engineered variables along with the original variables.
- New dataset showed considerable multicollinearity since many features were derived from the others.
- Performed a 2-way interaction of all the features with each other which increased predictors from 240 to 39,340.
- PCA transformed the correlated inputs into a set of non-correlated linear combinations which resolved the multicollinearity issue.
- 3,455 principal components accounted for 80% of the variability in the predictors

##	Variables	VIF
## 1	x16log	3898.657885
## 2	x5log	2548.241629
## 3	x20log	1933.938073
## 4	x16	1794.335885
## 5	x11	1617.835412
## 6	x11log	1617.529796
## 7	x8log	1616.618846
## 8	x5	1181.352484
## 9	x20	899.832193
## 10	x8	760.784111
## 11	x7log	551.537736
## 12	x1log	539.431432
## 13	x16sqinv	449.246006
## 14	x13log	379.929844
## 15	x12log	371.951963
## 16	x18log	316.819341
## 17	x5sqinv	296.949745
## 18	x7	267.551246
## 19	x1	259.322495
## 20	x20sqinv	226.688625
## 21	x8sqinv	190.358246
## 22	x13	186.168906
## 23	x12	185.412904
## 24	x22log	176.081153
## 25	x18	158.571673
## 26	x22	89.101209
## 27	x7sqinv	67.107315
## 28	x1sqinv	65.708942
## 29	x21	57.729060
## 30	x2	48.327951
## 31	x13sqinv	46.881321
## 32	x21log	46.673228
## 33	x12sqinv	45.232723
## 34	x2log	43.040667
## 35	x18sqinv	40.950701
## 36	x23	38.225244
## 37	x23log	35.673242
## 38	x6	34.791262
## 39	x17	28.762744
## 40	x21byx20	23.217635
## 41	x22sqinv	23.209779
## 42	x6log	22.007646
## 43	x17byx16	21.586372
## 44	x6byx5	20.627254
## 45	x9	20.143701
## 46	x19	19.326069
## 47	x10	19.180606
## 48	x14	17.547345
## 49	x2byx1	17.031845
## 50	x10byx8	15.083700

Top 50 variables by VIF show multicollinearity

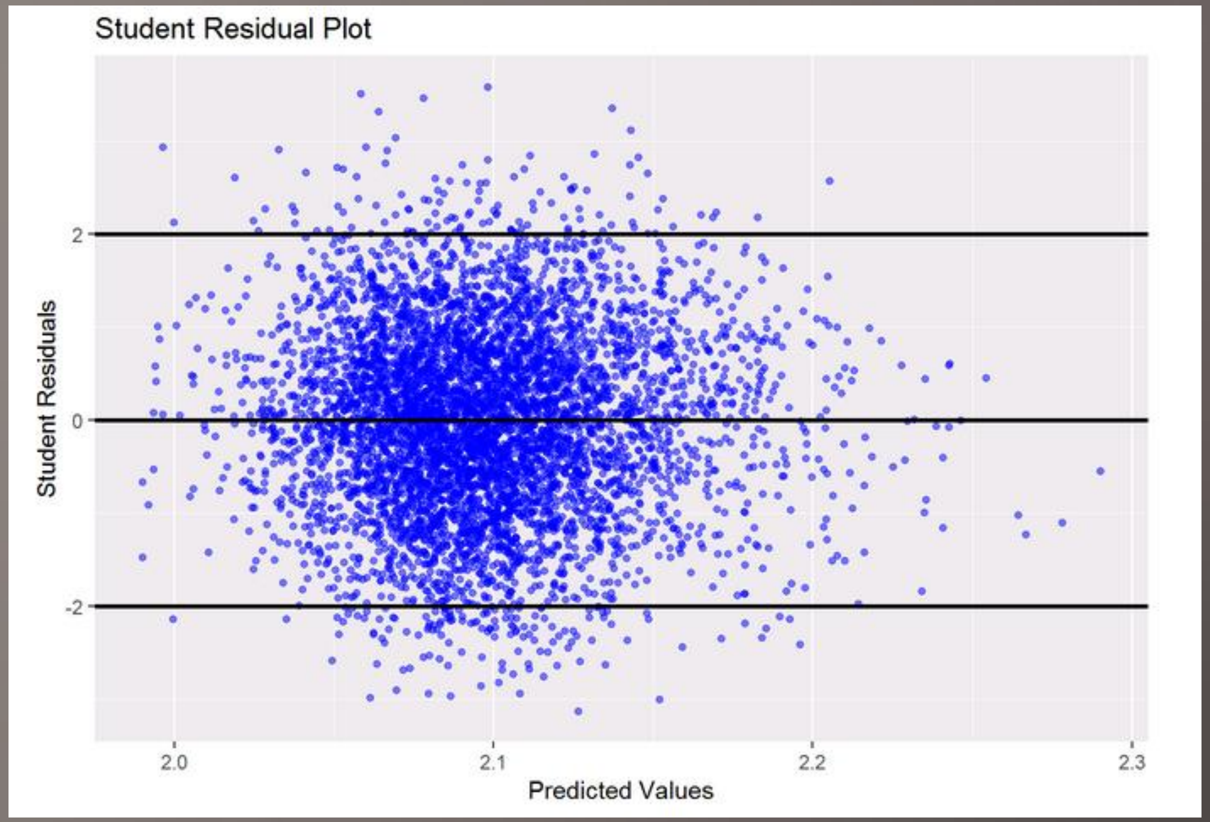


Scree plot showing cumulative variance explained by the first 7000 principal components

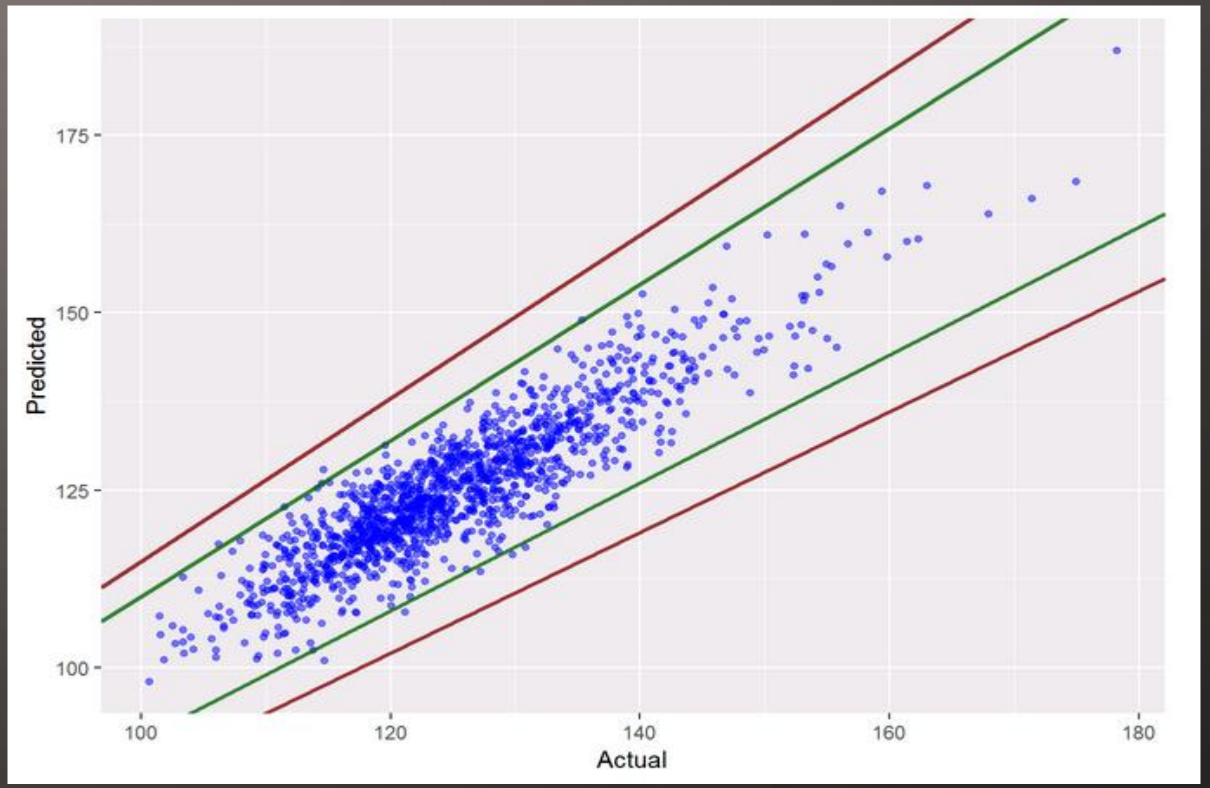
MODEL RESULTS (MODEL 5)

- Full model with these 3455 PCs and log(y3) as the output
- The residuals were normally distributed as shown by the histogram and the studentized residuals also show almost constant variance,
- The model exhibited very good fit statistics with an adjusted R^2 of 0.938 (R^2 = 0.976).
- Test RMSE also showed improvement (to 4.59) pointing to a good fitting model

ID #	Model (with PCA)	# VARS	Train R^2	Test RMSE
1	*Full Model (2-way EC + Stat)*	164	0.266	9.45
2	*Full Model (2-way All)*	3,834	0.786	16.76
3	*LASSO (2-way All)*	3,834	0.139	9.91
4	*LARS (2-way All)*	3,834	0.139	9.91
5	*Full Model (Feat Eng. 2-way All)*	3,455	0.976	4.59



Studentized Residuals vs. Predicted Values

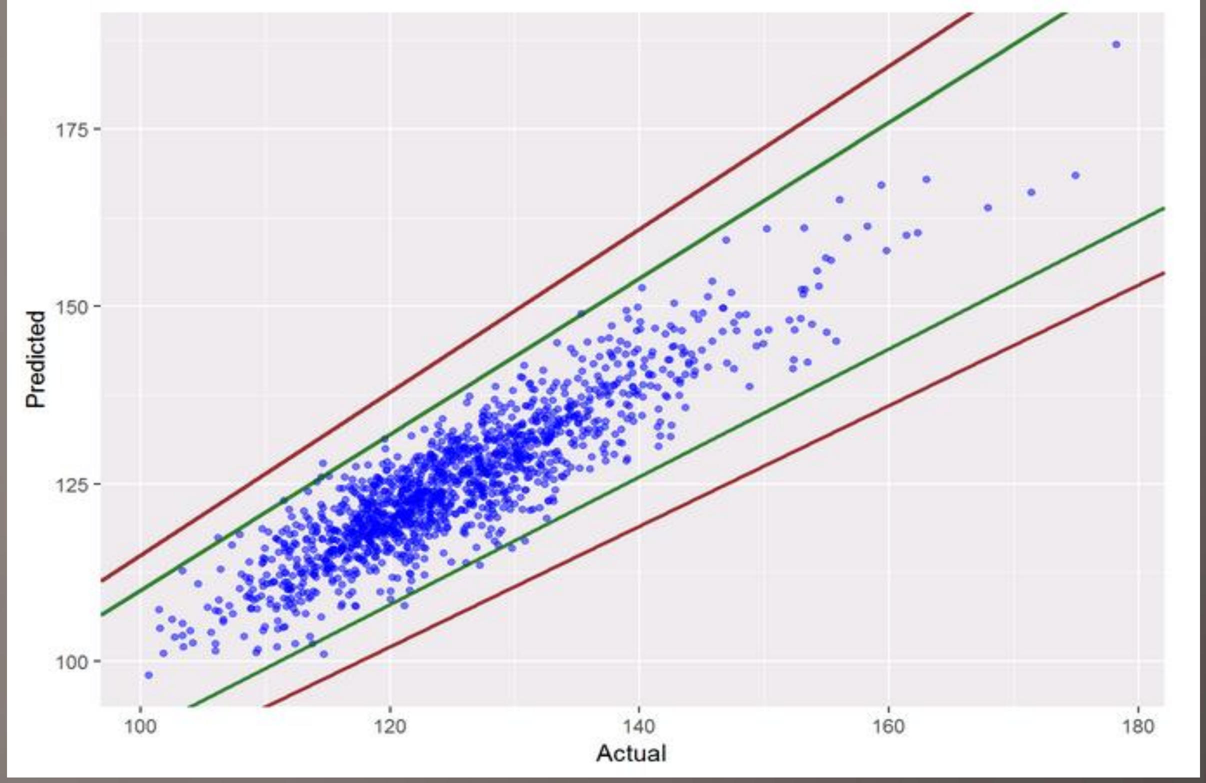


Predicted vs. Actual Values (ID 5 - Full Model Feat Eng. 2-way All)

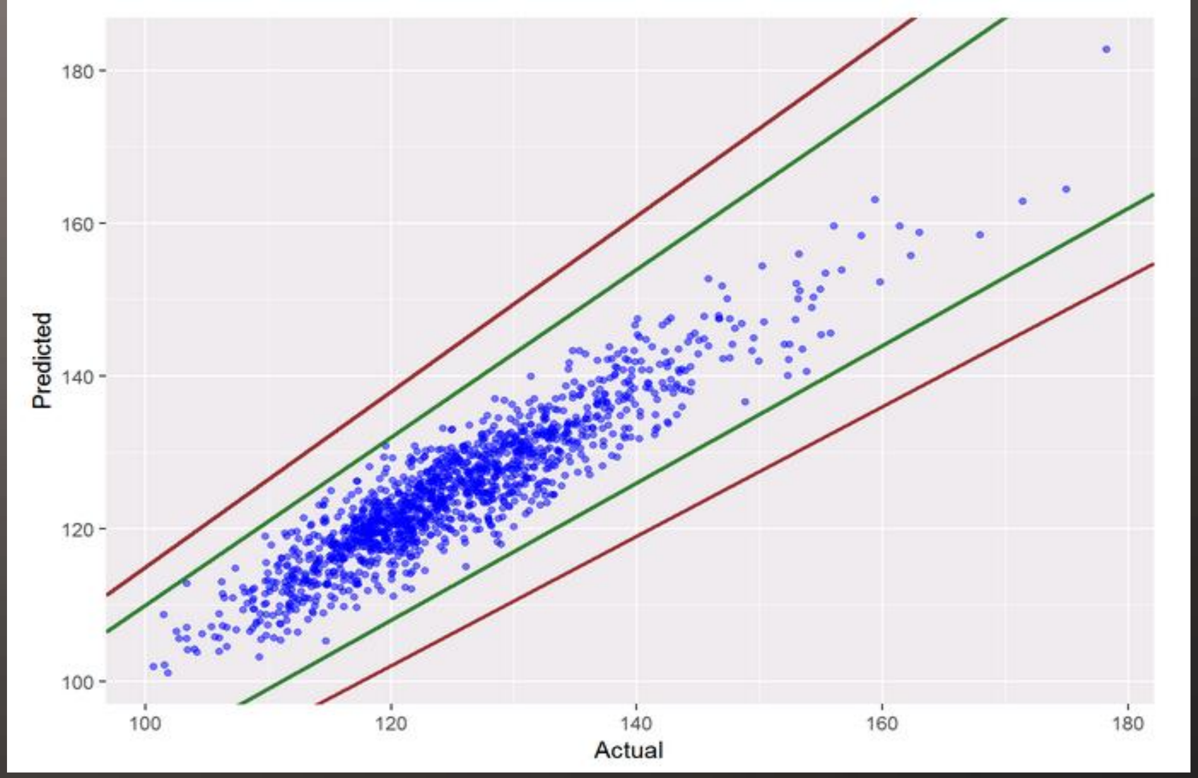
MODEL RESULTS (MODEL 6 AND 7)

- Variable selection and regularization on the full model to remove any scope of overfitting.
- LASSO & LARS:
  - Both these models provided similar performance metrics on the tests set
  - Both offered some improvement in the test RMSE compared to the full model
- Given Test RMSE, we choose the LASSO model as our final model.

ID #	Model (with PCA)	# VARS	Train R^2	Test RMSE
1	*Full Model (2-way EC + Stat)*	164	0.266	9.45
2	*Full Model (2-way All)*	3,834	0.786	16.76
3	*LASSO (2-way All)*	3,834	0.139	9.91
4	*LARS (2-way All)*	3,834	0.139	9.91
5	*Full Model (Feat Eng. 2-way All)*	3,455	0.976	4.59
6	*LASSO (Feat Eng. 2-way All)*	3,455	0.864	3.84
7	*LARS (Feat Eng. 2-way All)*	3,455	0.864	3.85



Predicted vs. Actual Values
(ID 5 - Full Model Feat Eng. 2-way All)



Predicted vs. Actual Values
(ID 6 - LASSO Feat Eng. 2-way All)

DISCUSSION - APPROACH 1

- We were able to achieve the prediction goal set out for this project
- We recognize the method has some issues in terms of scalability.
- Performing PCA and variable selection is extremely computation intensive.
  - Model development failed to run on a laptop with 4 multithreaded cores with 16GB of RAM.
  - Model ran on more powerful machine (32 cores and 256 GB of memory) and even then, it took over 6 hours of compute time (for PCA, LASSO and LARS combined) with RAM usage peaking at over 60GB.
  - Stepwise variable selection was not performed since it ran for more than 7 days on this machine (without completing).
- Suggestion:
  - Use cloud computing platform such as Amazon Web Service, or algorithms that allow parallelization

DISCUSSION- APPROACH 2

- **Clustering based on High Cook's D and using the cluster as a predictor in the regression model**
  - From project 1, we classified observations as high leverage or low leverage.
  - This data was then used to train various classification models including Linear/Quadratic Discriminate Analysis (LDA/QDA) and Logistic Regression models
  - Issue:
    - Could not obtain good accuracy of classification with any approach
    - More importantly, labels (influence statistic) were a subjective choice
    - Identification can itself vary and depend on the fit of the original model.
  - A good exercise in theory, this would not be of practical significance when implementing a predictive system at scale.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.TRAIN
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into highcd

From highcd	0	1	Total
0	2520 100.00	0 0.00	2520 100.00
1	126 100.00	0 0.00	126 100.00
Total	2646 100.00	0 0.00	2646 100.00
Priors	0.95	0.05	

Error Count Estimates for highcd

	0	1	Total
Rate	0.0000	1.0000	0.0500
Priors	0.9500	0.0500	

Cross Validation Results from QDA showing Specificity = 0

CONCLUSIONS

- Even though semiconductor physics offers a highly non-linear design space, can still use linear regression techniques to obtain a reasonable model fit
- The key to this process is incorporating intelligent domain specific feature engineering
- Challenges: Computation requirements
- Future Recommendations
  - Other non-parametric tree-based models (such as Random Forest, XGBoost) or Artificial Neural Networks could also be trained on this dataset