PM 2.5 Prediction in Beijing

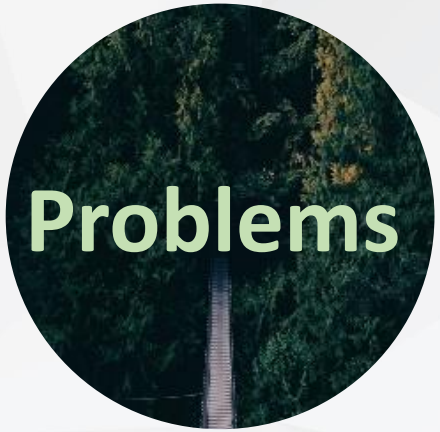
Team Da Da Da

CONTENTS

Part One

Context & Data Preparation

Context



- Trigger health issues
- Business problems

- Study the relationship
- The outcome optimized models help oversee and cut down

2 Data Wrangling

Credible Data Source

- https://www.kaggle.com/uciml/pm25-data-for-five-chinese-cities
- *"Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating", Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015)*

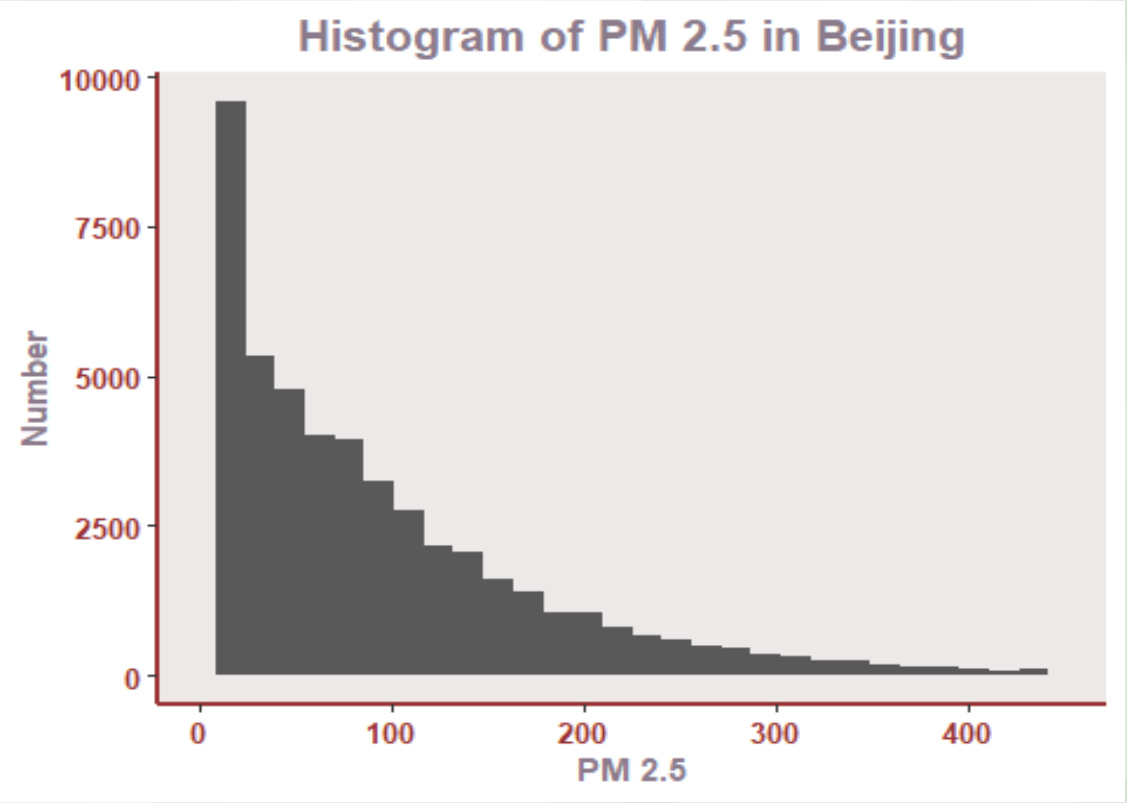
Variables	Description
Year	
PM_US.Post	PM2.5 (ug/m^3)
HUMI	Humidity (%)
DEWP	Dew Point (Celsius Degree)
TEMP	Temperature (Celsius Degree)
PRES	Pressure (Pa)
seasonadj	Season
Iws	Wind Speed (m/s)
precipitation	Precipitation (mm)
SW	Wind from southwest direction
SE	Wind from southeast direction
NW	Wind from northwest direction
NE	Wind from northeast direction

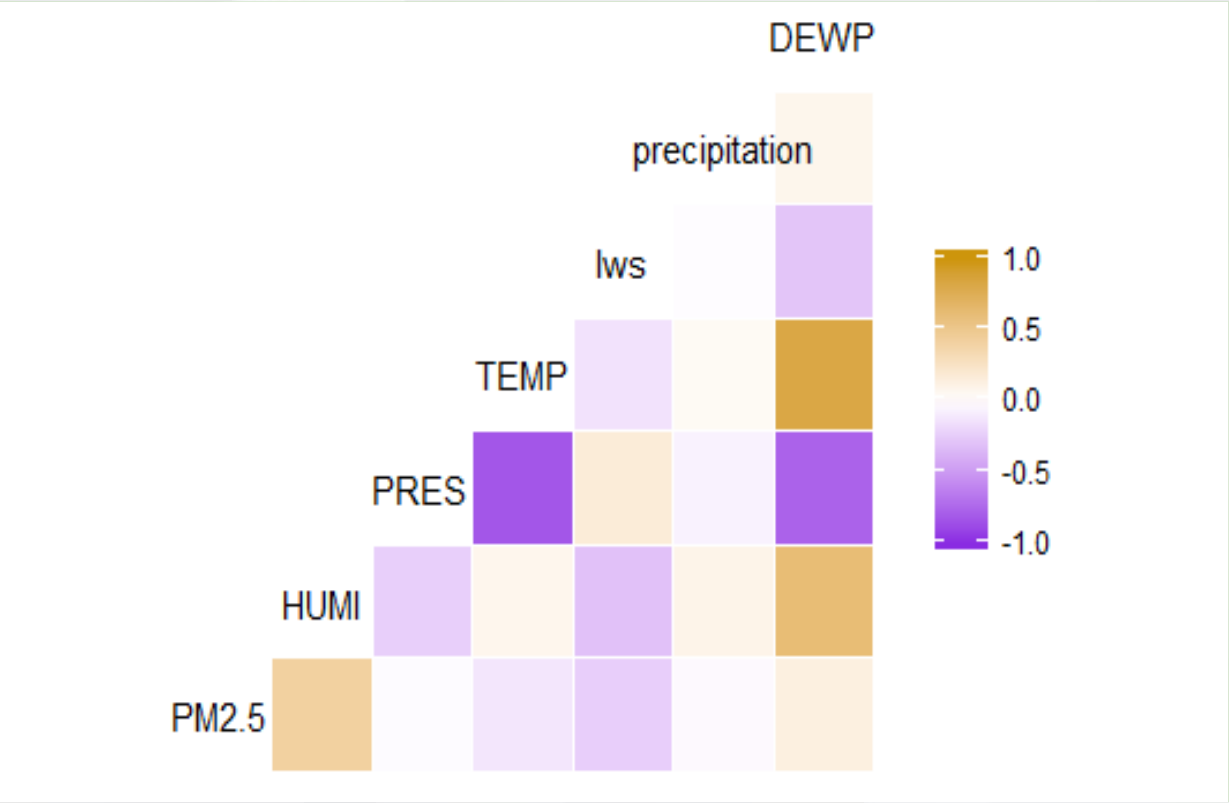
Trifacta Data Wrangling

49,579 valid observations and 13 variables; Time period 2010-2015

- Drop missing values (5.7%)
- Create 4 dummy variables for wind direction (NE, NW, SE, SW)

3 Data Visualization





Part Two

Multiple Regression

Multiple Regression-Modeling Process

Initial Model

- Use PM2.5 value as our target variable
- Training (80%) Validation (20%)

- R-Squared
- RMSE (training & validation)
- Interpretation

Final Model

Add Logarithmic form

- Log(PM_2.5)

VIF test

- Temperature > 5
- Dew Point > 5

Variables Selection

- forward selection, backward selection and exhaustive search
- 10 variables (delete temperature & dew point)

Try polynomial term in our model

- Air pressure^2
- Wind speed*wind direction

Outliers

- Delete 21 Obs (0.053%)

2 Multiple Regression-Variable Selection

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.984e+01  6.140e-01  32.311   <2e-16 ***
HUMI            2.022e-02  1.739e-04 116.279   <2e-16 ***
seasonadjSummer -7.310e-01  1.285e-02 -56.869   <2e-16 ***
seasonadjFall   -2.820e-01  1.209e-02 -23.318   <2e-16 ***
seasonadjWinter  3.547e-01  1.345e-02  26.375   <2e-16 ***
Iws             -2.964e-03  8.908e-05 -33.272   <2e-16 ***
cbwdNW          -1.262e-01  1.384e-02  -9.119   <2e-16 ***
cbwdSE           6.476e-01  1.346e-02  48.127   <2e-16 ***
cbwdSW           4.733e-01  1.428e-02  33.155   <2e-16 ***
precipitation   -5.723e-02  4.481e-03 -12.773   <2e-16 ***
PRES            -1.663e-02  6.038e-04 -27.536   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7797 on 39652 degrees of freedom
Multiple R-squared:  0.4481,    Adjusted R-squared:  0.448
F-statistic:  3219 on 10 and 39652 DF,  p-value: < 2.2e-16
```

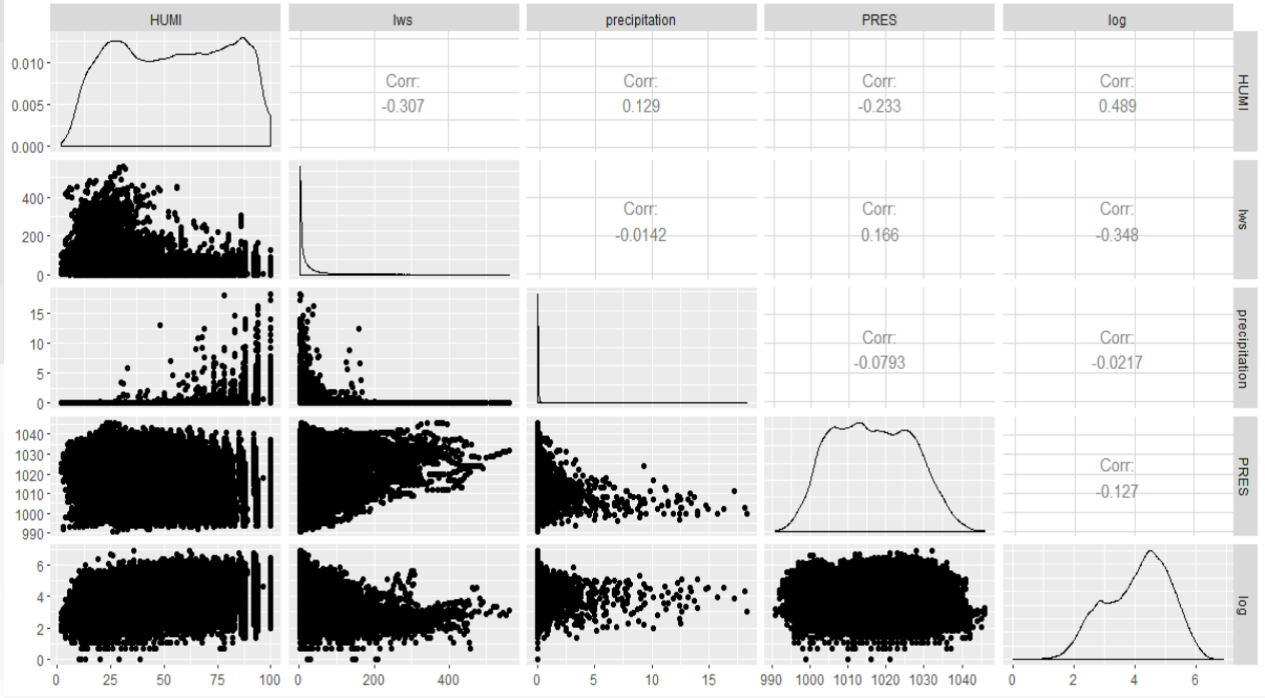
Forward selection, backward selection:
10 variables: humidity, wind speed, precipitation, air pressure, three season dummies, three wind direction dummies.

Exhaustive search:
set maximum variables number: 10, (still add 12 variables into the model.)
Delete temperature and dew point

3 Multiple Regression- Polynomial Terms

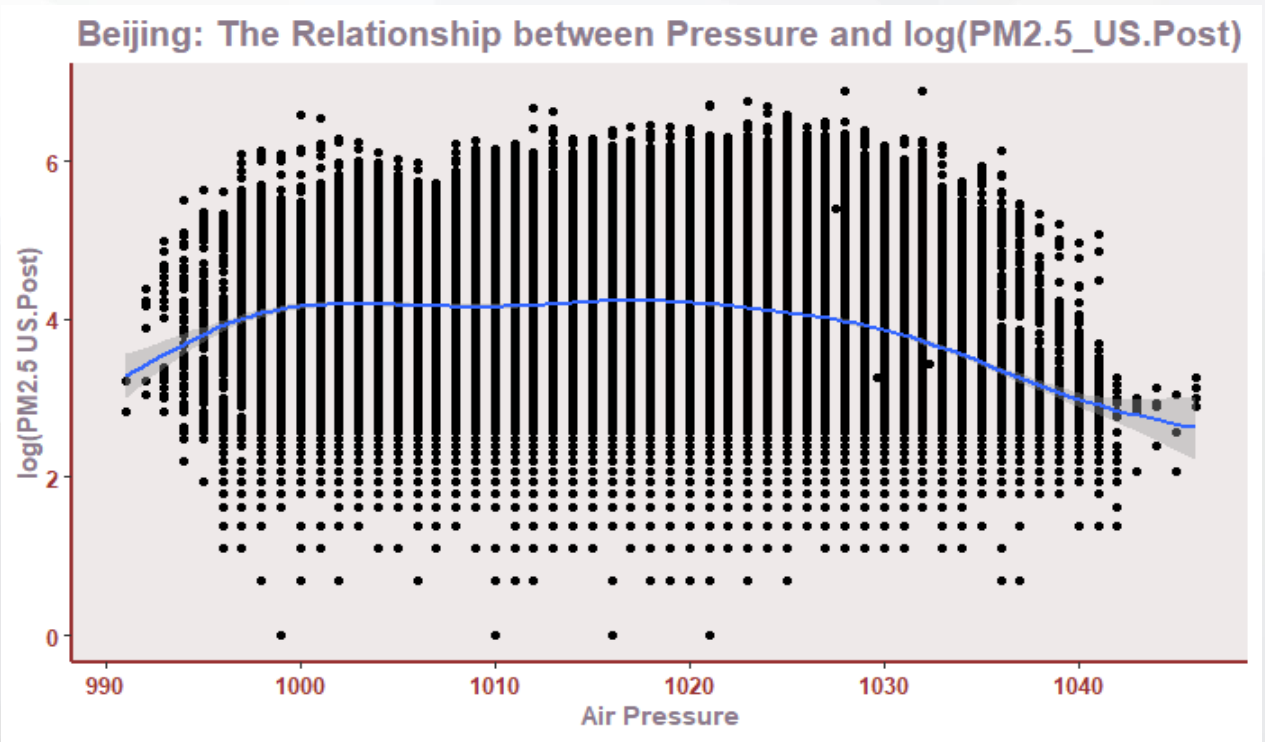
- Wind speed* SW, Wind speed* SE, Wind speed* NW

Hourly Wind Speed*combined wind direction (categorical variable).
Wind from some specific directions may have more influence on PM2.5 in Beijing

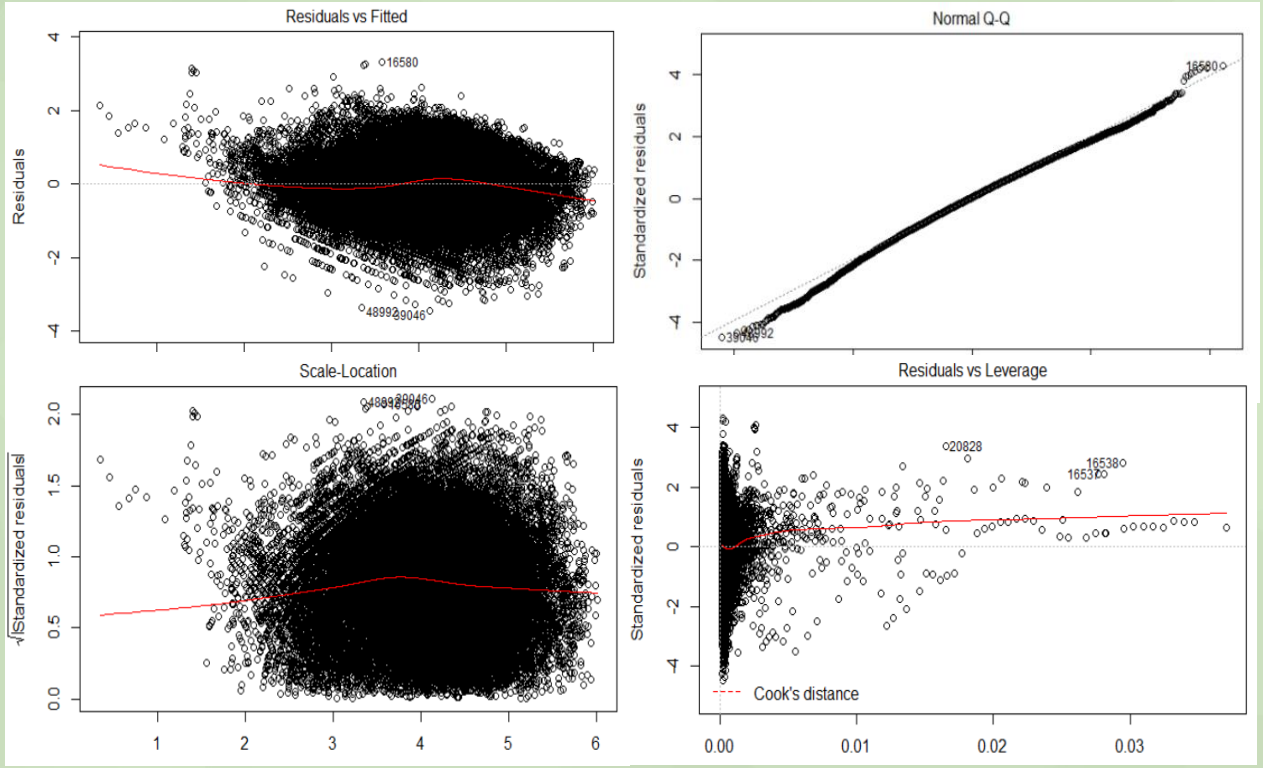


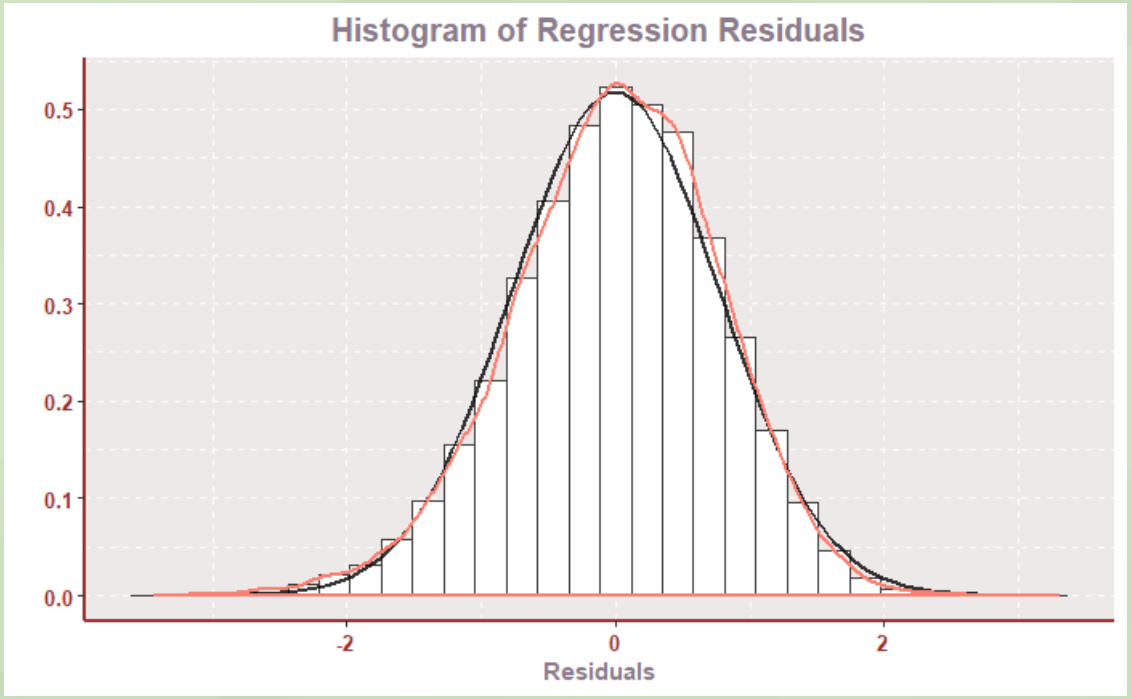
- Pressure ^2

Plots between numeric variables and log(PM 2.5)
log(PM 2.5) - Air Pressure - quadratic function
log(PM 2.5) - Air Pressure^2 – linear relationship



4 Multiple Regression-Outliers





5 Multiple Regression-Interpretation

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.573e+02  3.975e+01 -19.053  < 2e-16 ***
HUMI            1.978e-02  1.744e-04 113.439  < 2e-16 ***
seasonadjSummer -6.368e-01  1.341e-02 -47.481  < 2e-16 ***
seasonadjFall   -2.972e-01  1.212e-02 -24.529  < 2e-16 ***
seasonadjWinter  3.930e-01  1.345e-02  29.224  < 2e-16 ***
Iws            -6.065e-02  3.640e-03 -16.662  < 2e-16 ***
cbwdNW         -2.532e-01  1.519e-02 -16.667  < 2e-16 ***
cbwdSE          4.961e-01  1.555e-02  31.901  < 2e-16 ***
cbwdSW          3.479e-01  1.574e-02  22.101  < 2e-16 ***
precipitation  -1.176e-01  7.111e-03 -16.535  < 2e-16 ***
I(PRES^2)      -7.498e-04  3.838e-05 -19.537  < 2e-16 ***
PRES            1.510e+00  7.812e-02  19.334  < 2e-16 ***
I(Iws * SW)     1.313e-02  1.944e-03   6.751 1.49e-11 ***
I(Iws * SE)     1.581e-02  8.529e-04  18.536  < 2e-16 ***
I(Iws * NW)     1.445e-02  7.929e-04  18.221  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7697 on 39627 degrees of freedom
Multiple R-squared:  0.4616,    Adjusted R-squared:  0.4614
F-statistic:  2427 on 14 and 39627 DF,  p-value: < 2.2e-16

                      ME      RMSE       MAE  MPE MAPE
Test set -6.565243e-17 0.7695204 0.6098349 -Inf  Inf
                      ME      RMSE       MAE  MPE MAPE
Test set -0.0006729632 0.7770129 0.616476 -Inf  Inf
```

Adjusted R^2 :0.4614
RMSE(Training): 0.7695
RMSE(Validation): 0.7770

Humidity:
coefficient = 0.01978 > 0.
When humidity increases by 1%, PM2.5 increases by 1.978%.

Wind direction:
(compared with wind from northeast)

PRES and PRES^2:
Coefficient(PRES) is positive, Coefficient(PRES^2) is negative.
(under extreme situations, PM2.5 tends to have a negative relationship with pressure.)

Part Three
Classification

Target Variable Justification

Target variable: Air Condition

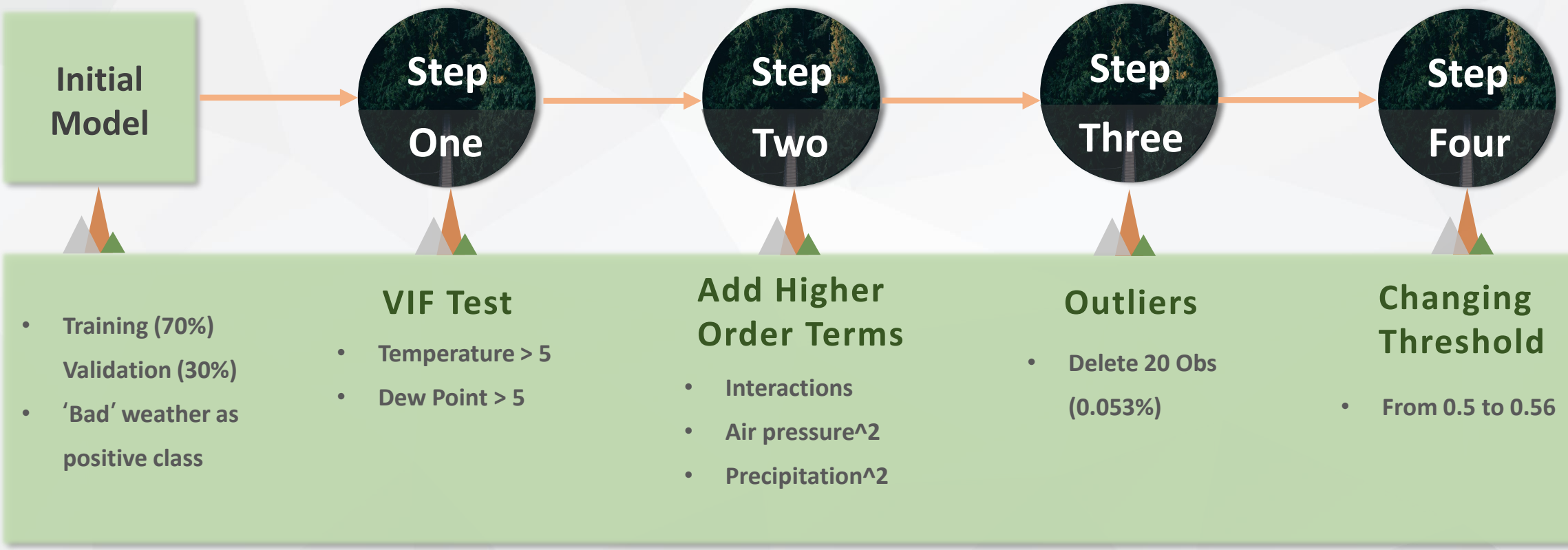
If PM2.5 is less than or equal to 50 µg/m³, we consider it to be a 'Good' air condition, which is only unhealthy for sensitive groups.

If PM2.5 is more than 50µg/m³, we consider it to be a 'Bad' air condition, which is unhealthy for everyone.

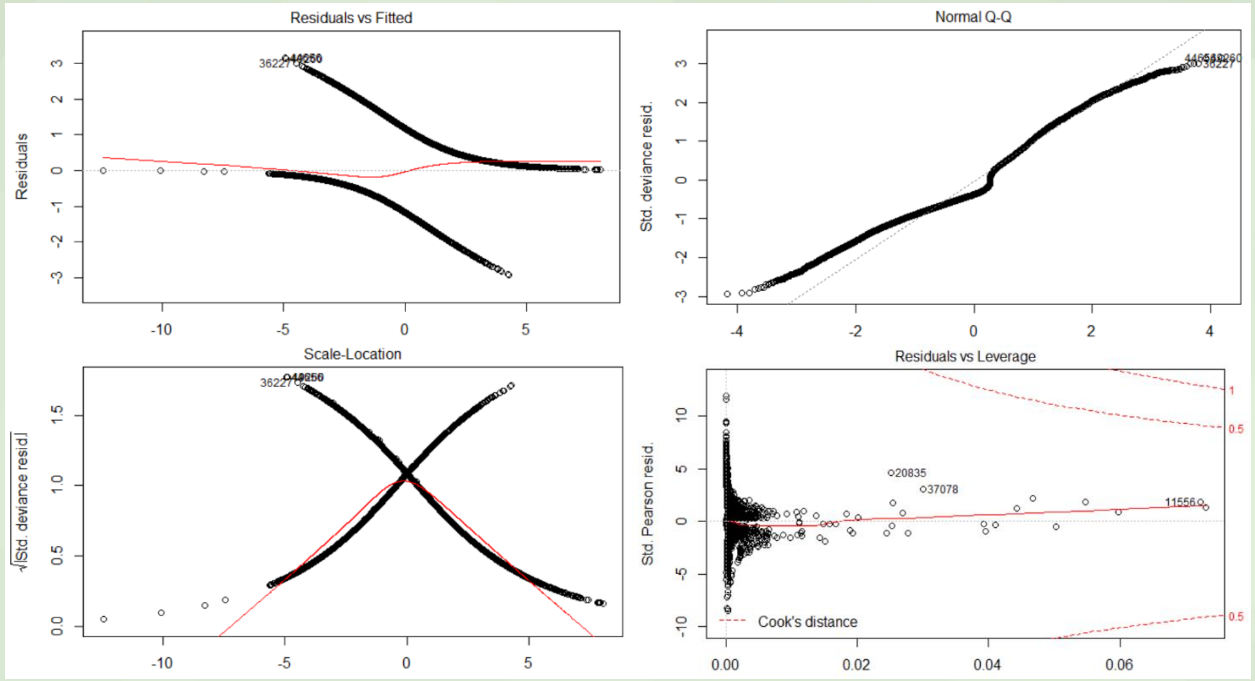
AQI Category	Index Values	Revised Breakpoints (µg/m³, 24-hour average)
Good	0 - 50	0.0 – **12.0**
Moderate	51 - 100	12.1 – **35.4**
Unhealthy for Sensitive Groups	101 – 150	35.5 – **55.4**
Unhealthy	151 – 200	55.5 – 150.4
Very Unhealthy	201 – 300	150.5 – 250.4
Hazardous	301 – 400	250.5 – 350.4
Hazardous	401 – 500	350.5 – 500

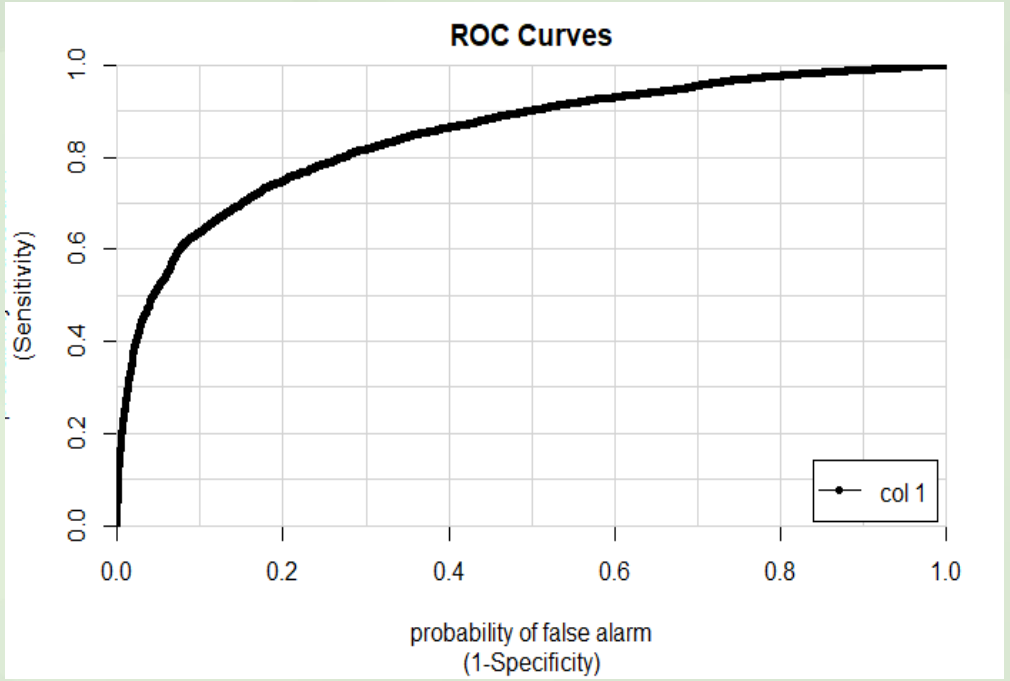
Image source: https://medium.com/mongolian-data-stories/air-pollution-part-2-f9f4da33a1bd

2 Logistic Regression-Modeling Process



3 Logistic Regression-Final Model





Test Set		Reference	
		Bad	Good
Prediction	Bad	8375	2235
	Good	737	3527

Accuracy: **0.8002** (0.7969)
Sensitivity: **0.9191** (0.9156)
F1: **0.8493** (0.8450)
AUC: **0.8499**

4 Logistic Regression-Coefficient Interpretation

Odds Ratios of Logistic regression

Covariates	Logit coef.	Odds Ratio
Humidity	-0.050	0.951
Air pressure	-4.227	0.015
Air pressure^2	0.002	1.002
North west	0.280	1.323
South east	-1.830	0.160
South west	-1.162	0.313
Wind speed	0.010	1.010
Precipitation	0.953	2.594
Precipitation^2	-0.077	0.926
Summer	1.641	5.160
Fall	1.014	2.756
Winter	-0.775	0.461

Positive Correlation: Humidity, South wind, Winter.

Negative Correlation: North wind, Wind speed, Summer, Fall.

Air Pressure: Always negative.

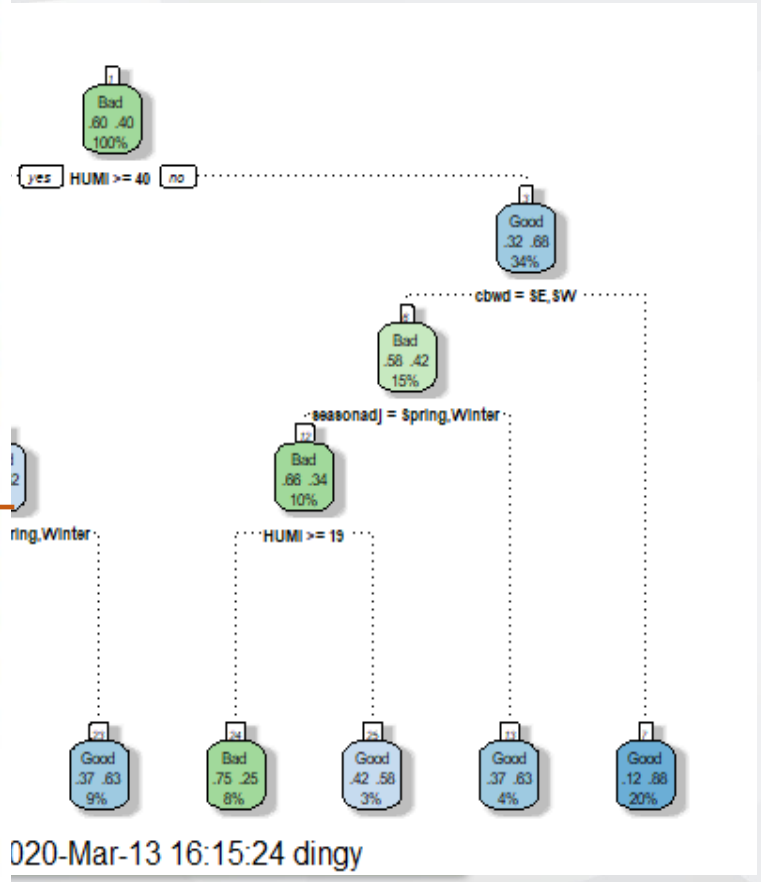
Precipitation: First positive and then becomes negative.

5 Decision Tree-Pruning Process

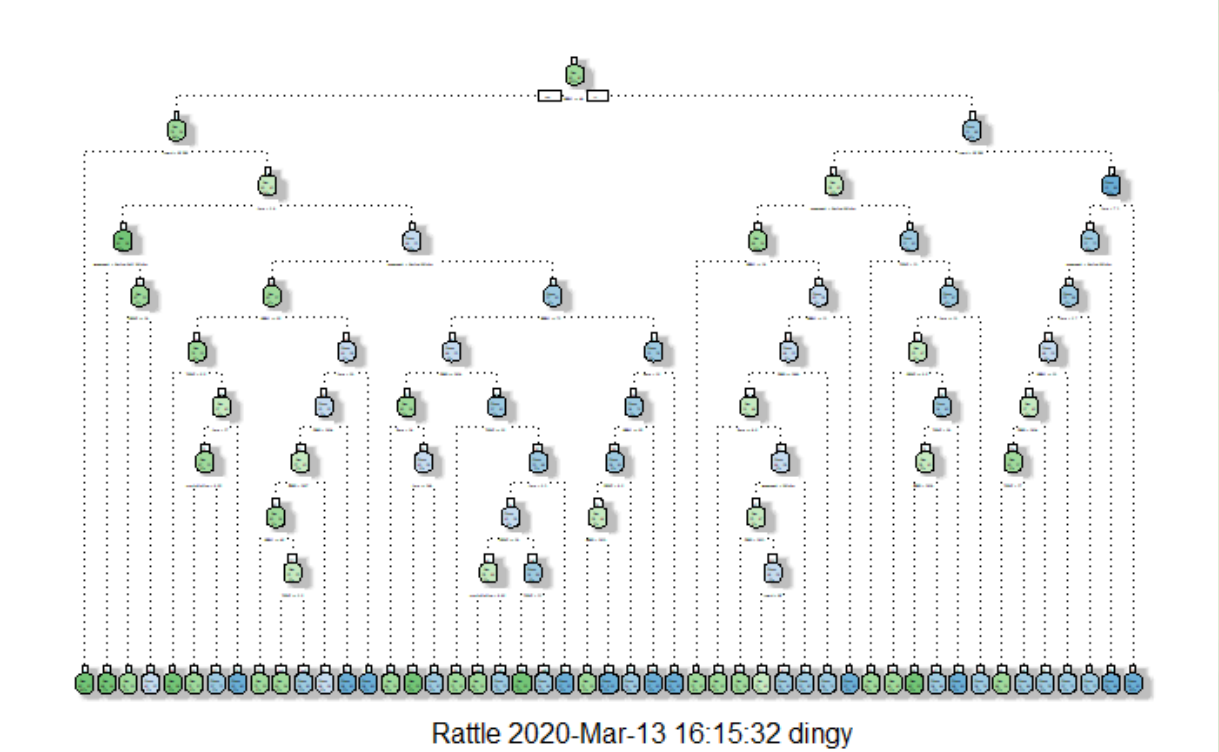
Initial Model

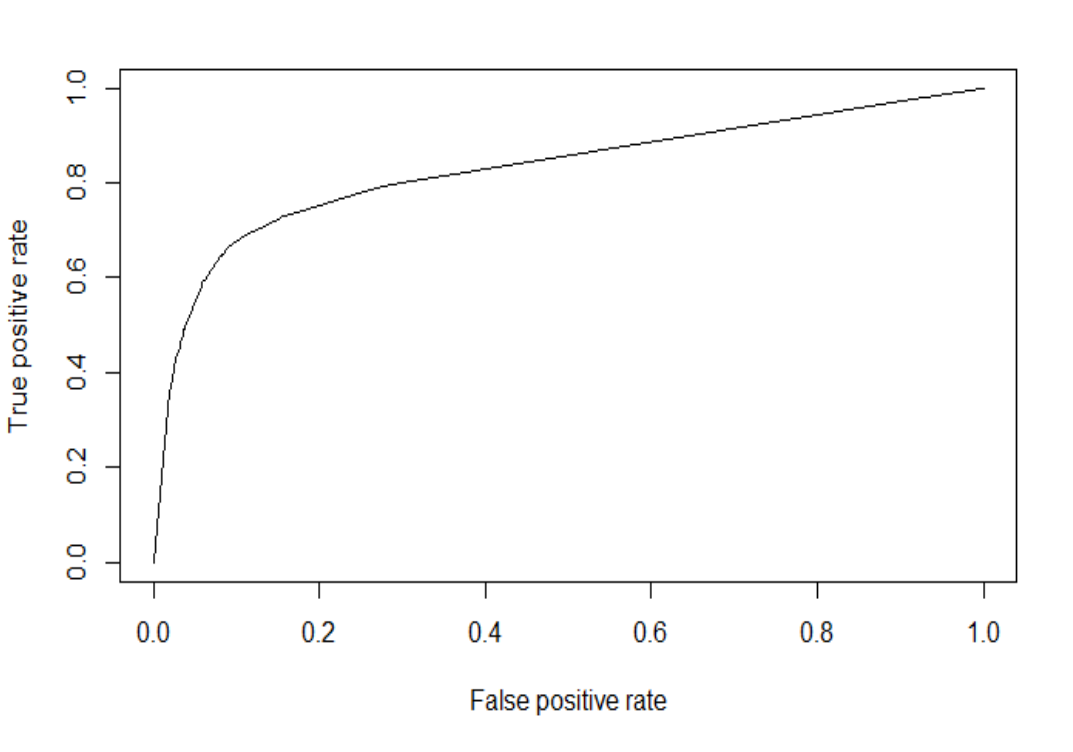
- Training (70%) Validation (30%)
- 'Bad' weather as positive class

21	0.000619398	38	0.458938	0.47679	0.0053098
22	0.000582963	46	0.452889	0.47613	0.0053070
23	0.000558673	48	0.451723	0.47621	0.0053073
24	0.000546528	51	0.450047	0.47570	0.0053051
25	0.000534383	53	0.448954	0.47373	0.0052966
26	0.000473657	56	0.447351	0.47227	0.0052904
27	0.000437222	58	0.446404	0.47125	0.0052860
28	0.000412932	63	0.444072	0.47176	0.0052882
29	0.000400787	66	0.442833	0.47125	0.0052860
30	0.000388642	86	0.434453	0.47133	0.0052863
31	0.000364352	92	0.431684	0.471557	0.0052850
32	0.000327917	103	0.427676	0.46885	0.0052755
33	0.000315772	112	0.424543	0.46812	0.0052724
34	0.000291481	115	0.423595	0.46659	0.0052657
35	0.000267191	125	0.420681	0.46703	0.0052676
36	0.000255046	129	0.419588	0.46899	0.0052762
37	0.000247759	141	0.416527	0.46899	0.0052762
38	0.000242901	146	0.415288	0.46929	0.0052774
39	0.000218611	149	0.414559	0.46958	0.0052787
40	0.000200393	185	0.406398	0.46673	0.0052663
41	0.000189463	197	0.403410	0.46812	0.0052724
42	0.000182176	208	0.401297	0.46812	0.0052724
43	0.000170031	261	0.390950	0.46754	0.0052698
44	0.000163958	270	0.389346	0.46841	0.0052736
45	0.000145741	274	0.388691	0.46841	0.0052736
46	0.000133596	440	0.361801	0.47191	0.0052888



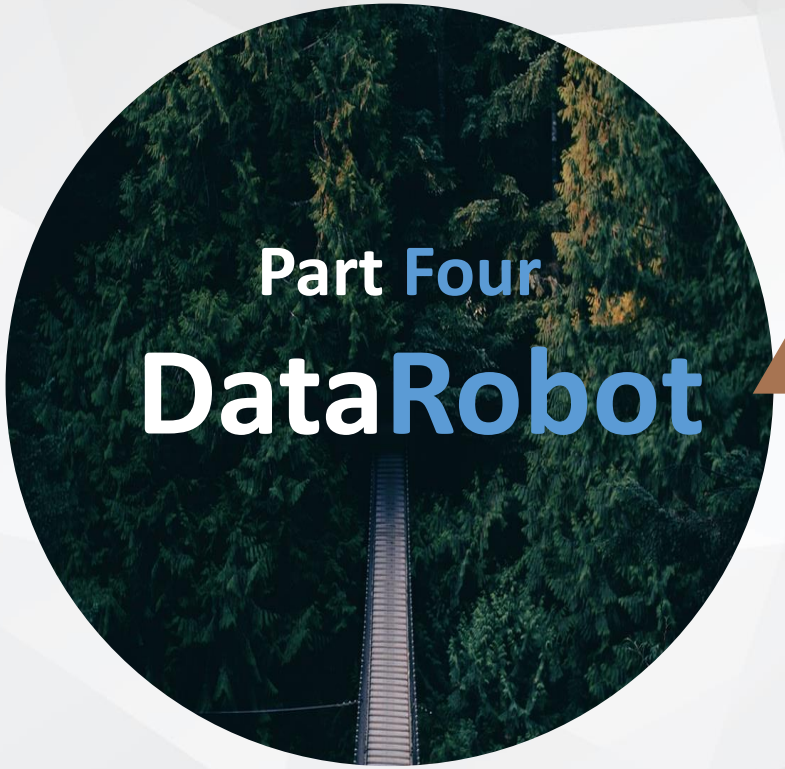
Decision Tree-Final Model





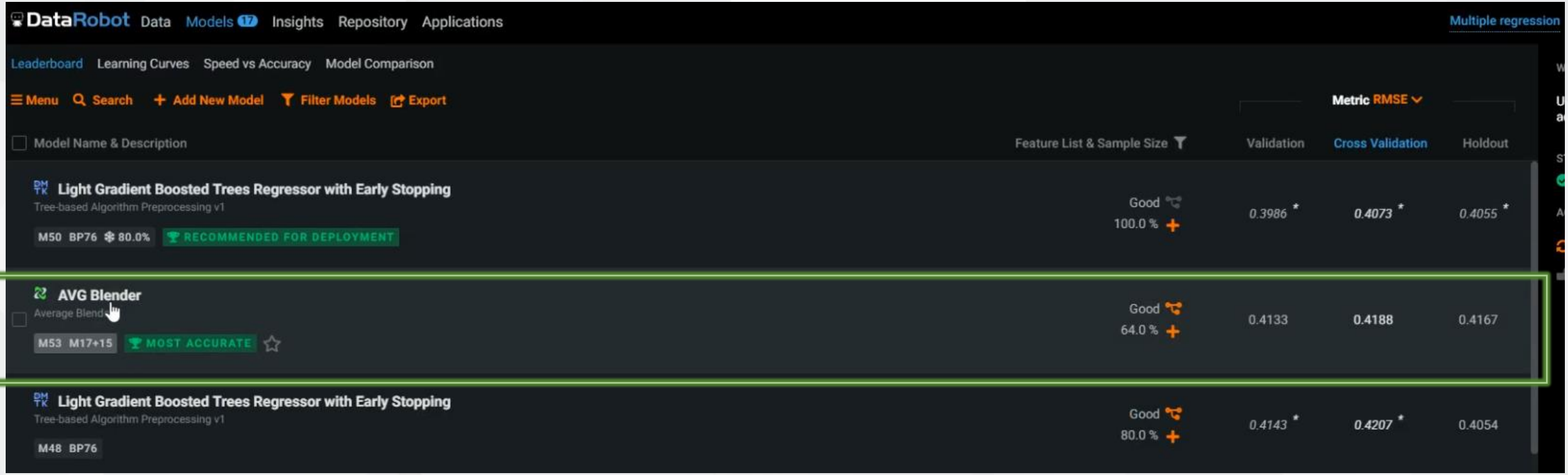
Test Set		Reference	
		Bad	Good
Prediction	Bad	8246	1890
	Good	866	3872

Accuracy:	0.8147 (0.8214)
Sensitivity:	0.9050 (0.9093)
F1:	0.8570 (0.8602)
AUC:	0.8290



DataRobot

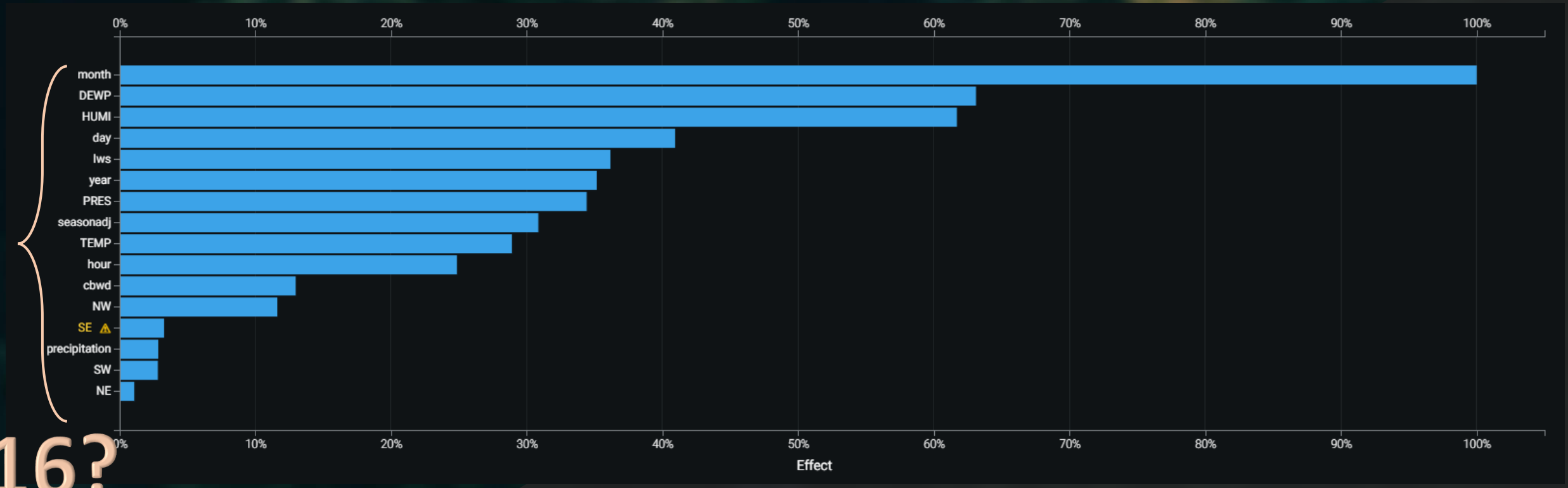
DataRobot-Comparison of Regression Models



The AVG Blender has a RMSE of 0.4167
Our best handmake model has a RMSE of 0.78

Something Important!

2 DataRobot-Feature Importance

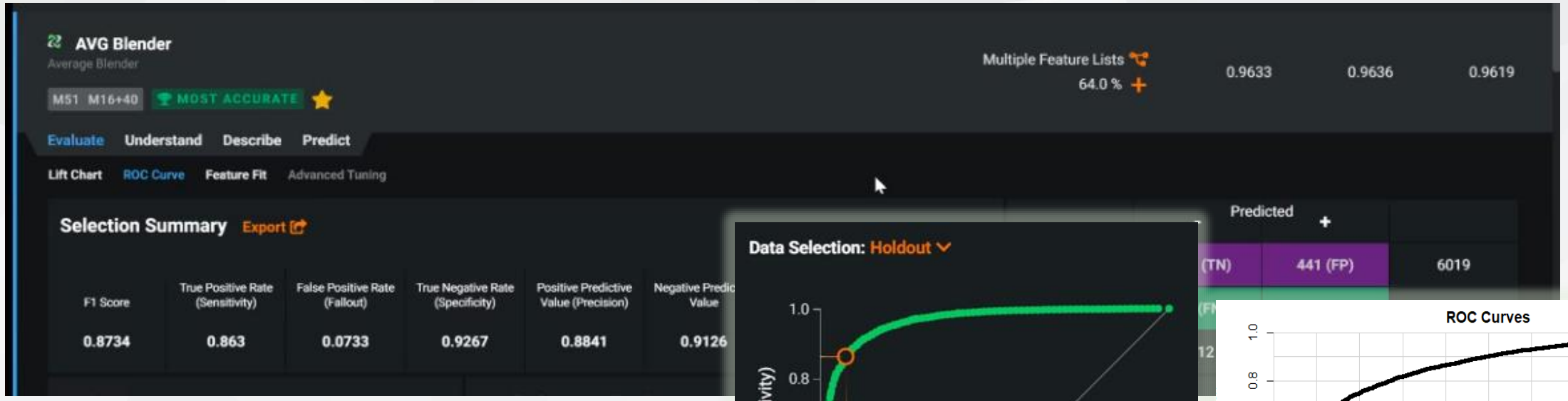


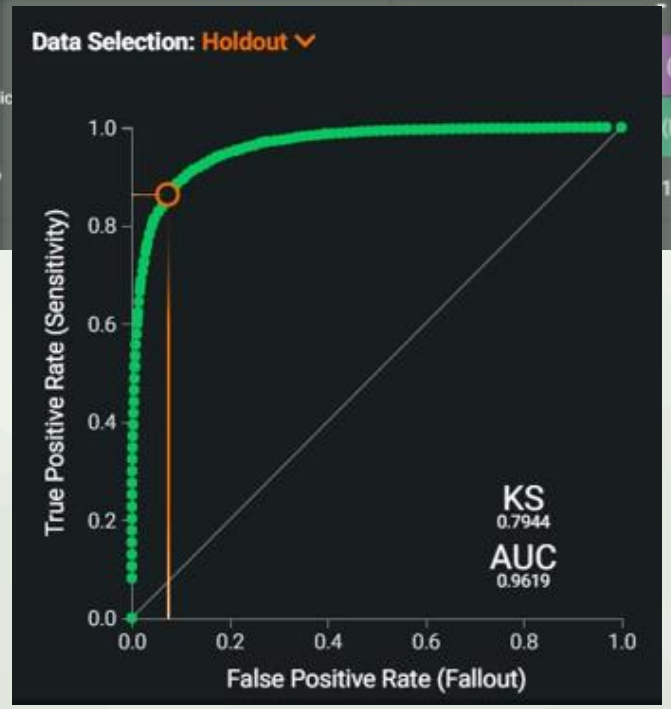
16?

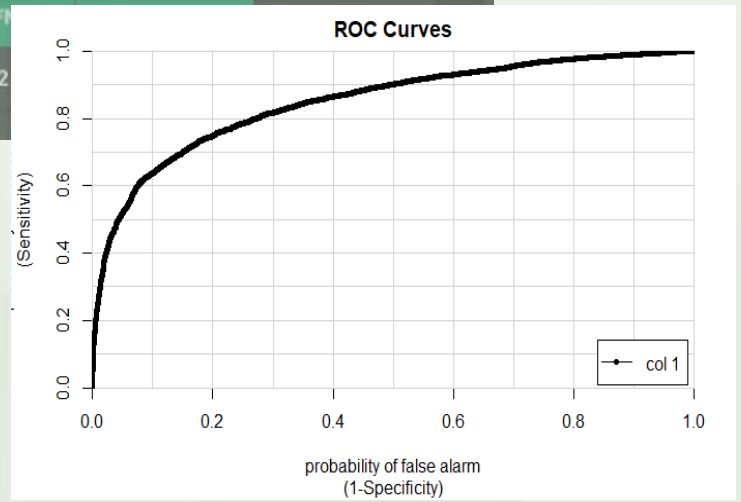
- Deleting Outliers
- Creating Polynomial Variables
- VIF Test

The AVG Blender has 15 variables.
Our best model has 14 variables.

3 DataRobot-Comparison of Classification Models







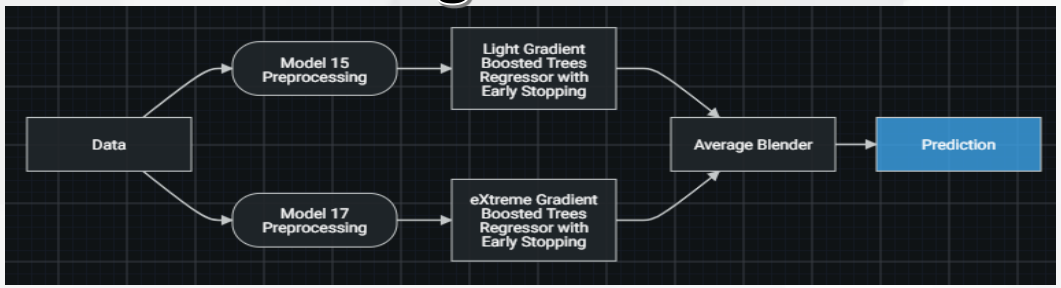
AVG Blender Again

The AVG Blender has an AUC of 0.9619
Our best handmake model has an AUC of 0.8498

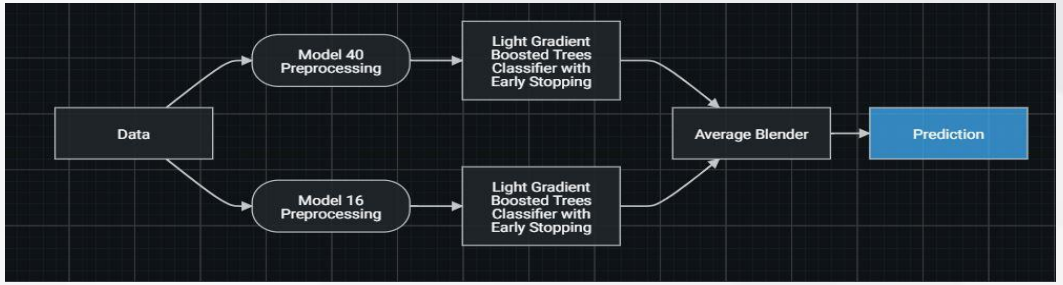
4 DataRobot-Overfitting issue

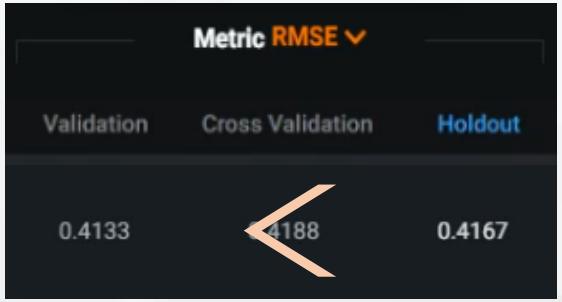
The AVG Blender

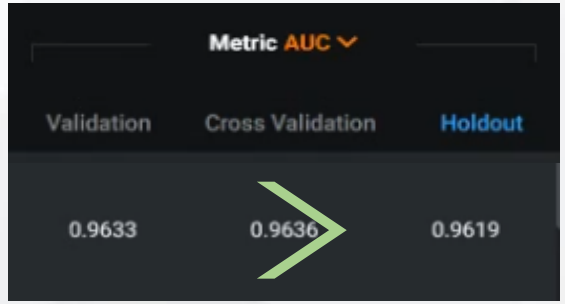
The Regression Model



The Classification Model

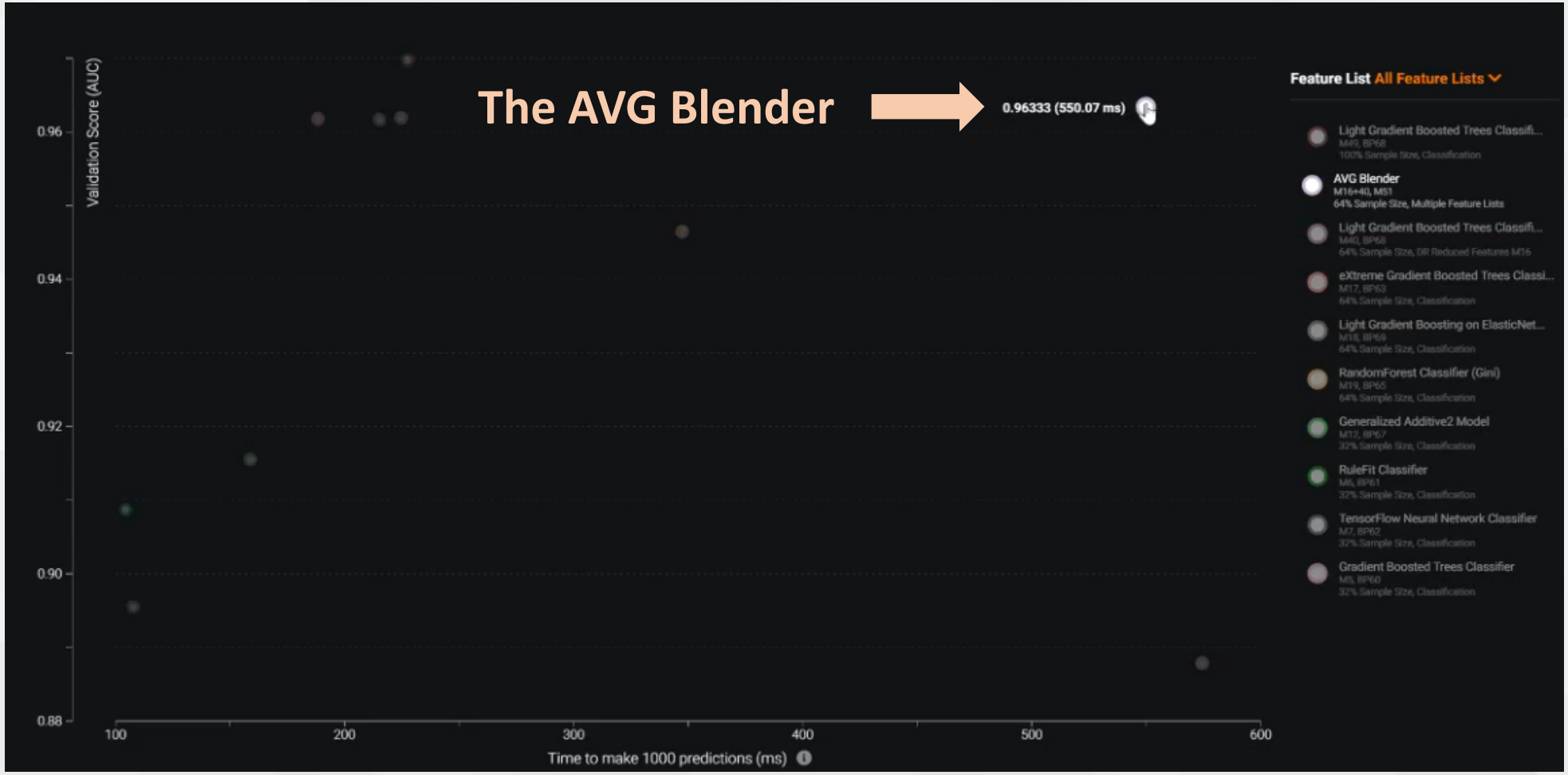






Help Prevent Overfitting Issue

5 DataRobot-Time Consumed



Time Consuming

6 Conclusion & Reflection

THANKS

For Watching