Badge-4 Lab-4 [SVM]

Out date: Aug 10, 2022
Due date: Aug 14 at 11:59PM

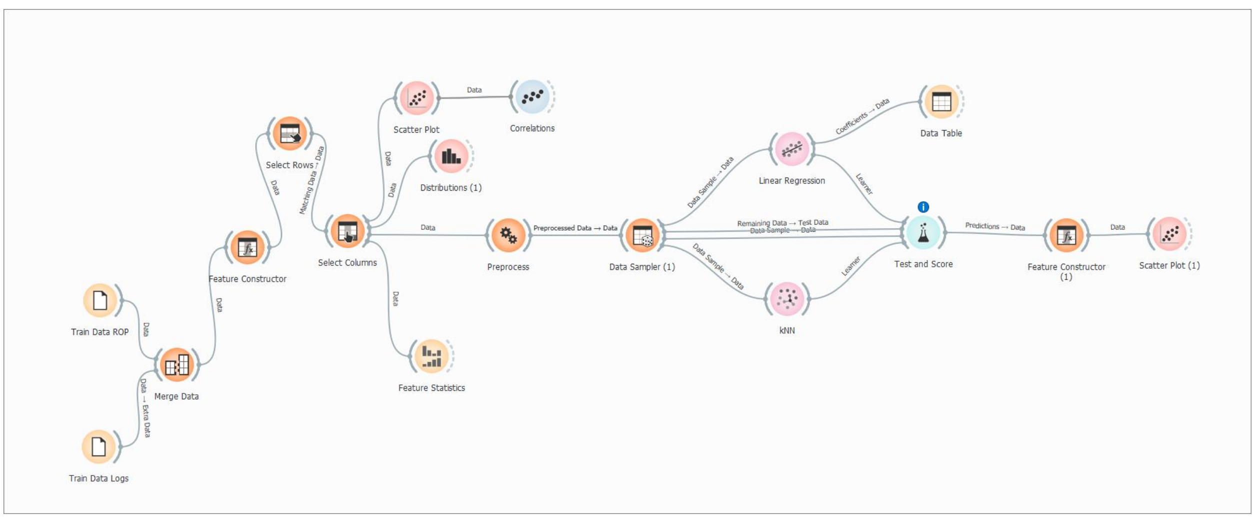
Submission

1. Prepare your solution in Orange and save the workspace for Problem 1 (e.g., Lab-4_SVM_LastName.ows) **[20 points]**
2. Complete the tables given below and save the file (e.g., Lab-4_SVM_LastName.docx). **[80 points]**
3. Upload the files to the Canvas.

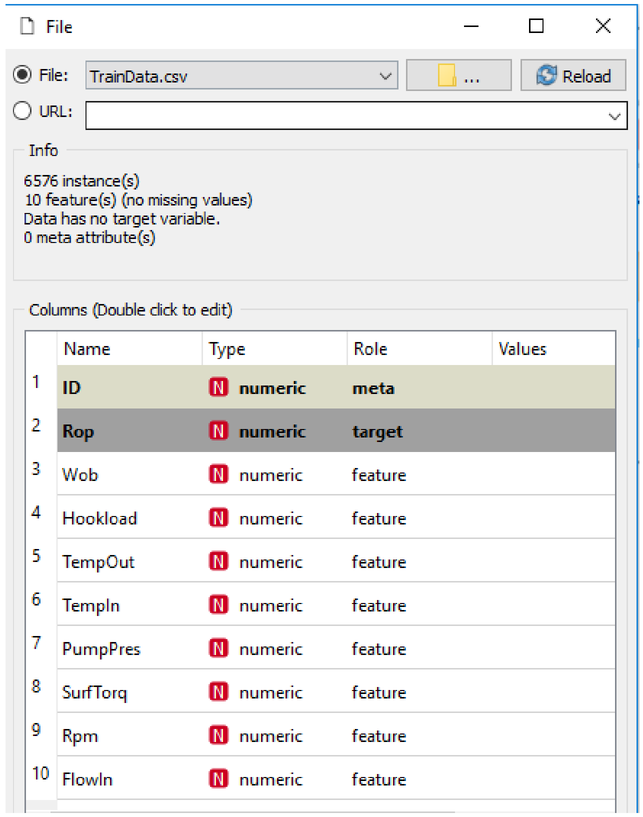
Objective(s):
To apply SVM algorithm for a regression problem and compare its performance with other machine learning algorithms.

Data: Please download the following files from Canvas to complete your assignment:

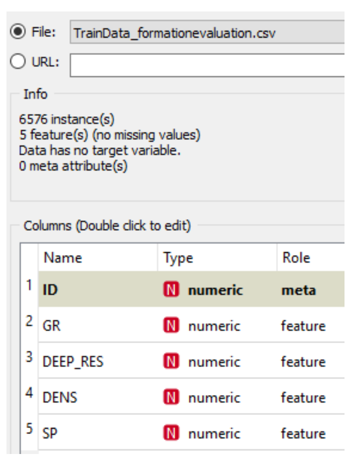
1. *Badge3_Lab4_start.ows* file with the starting pipeline as shown below:



2. *TrainData.csv* containing the following features:



3. *TrainData_formationevaluation.csv* is an optional dataset available to you for use and it contains the following features. Since drilling involves penetrating subsurface rocks, considering one or more of these features may help in improving your model performance.



Drilling diagnostics predictor variables:

WOB- Weight applied to the drill bit in kilopounds (k-lbs)
Hookload – Total weight of the suspended drill string in kilopounds (k-lbs)

TempOut and TempIn- Temperature of the drilling fluid going in and coming out in degF

PumpPres: Pressure exerted by the surface pump when pumping drilling fluid (mud) downhole, in psi.

RPM- Rotations per minute – Speed at which the drill string is rotated at surface

SurfTorq – Torque as a result of the drill string rotation, in psi

FlowIn – Flow rate at which the drilling fluid is pumped downhole, in gallons per minute

Optional formation evaluation predictor variables:

Gamma Ray(GR), Density of the formation (DENS), Resistivity of the formation (DEEP_RES) and Spontaneous Potential (SP)

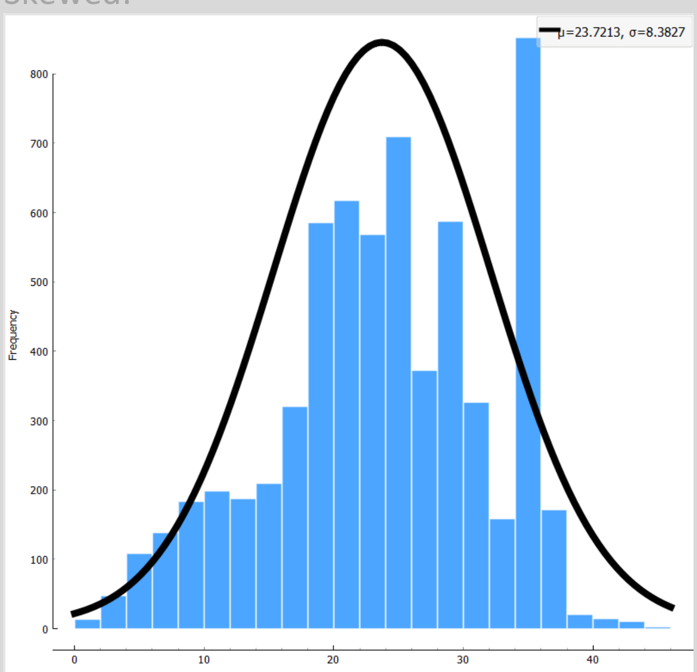
Target variable: Rate of Penetration (ROP) in feet/hour, a measure of how fast the drilling has progressed.

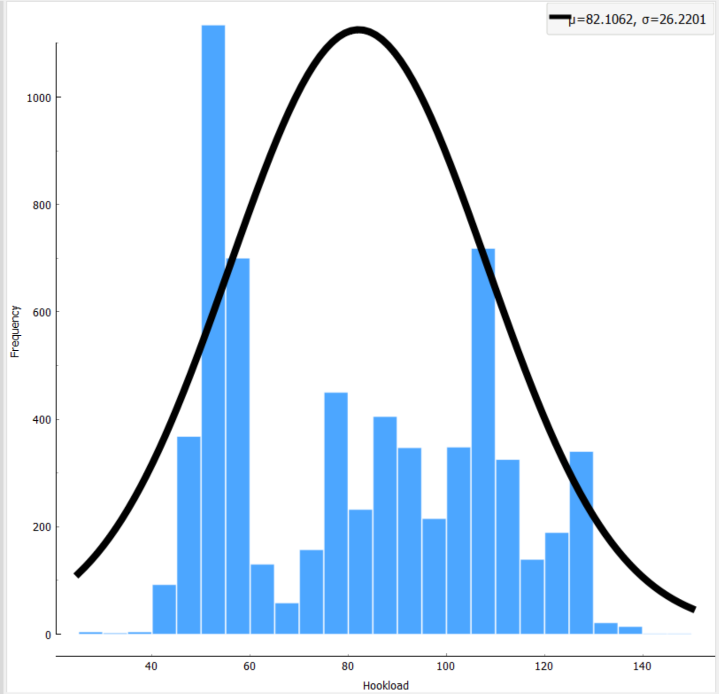
Reference:

https://gdr.openei.org/submissions/1113 (data)

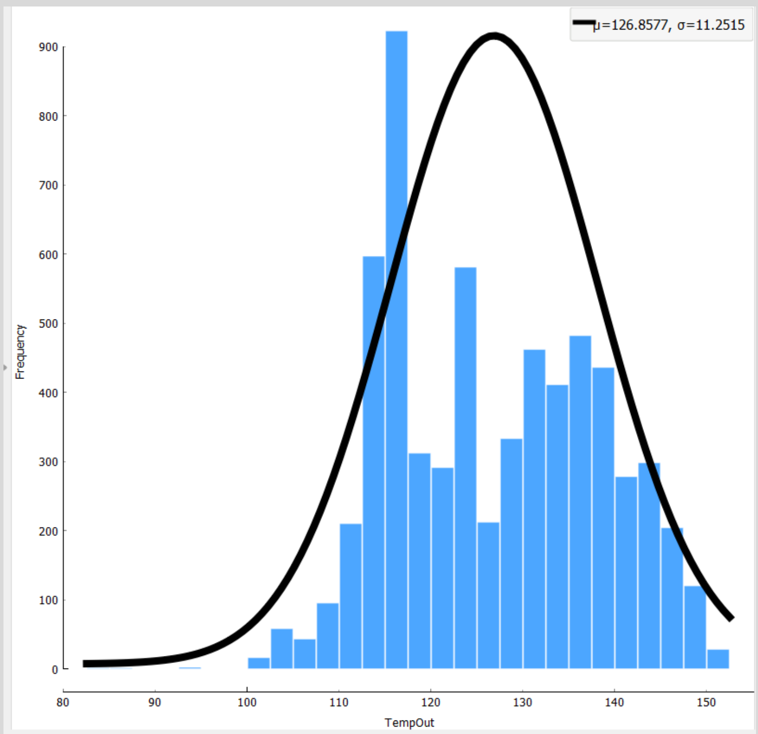
https://www.youtube.com/watch?v=guFiQ87tg_s (a video about the oil well drilling process)

1. Open the *Badge3_Lab4_start.ows file* using Orange.
2. Inspect the **entire** pipeline: (15 points)

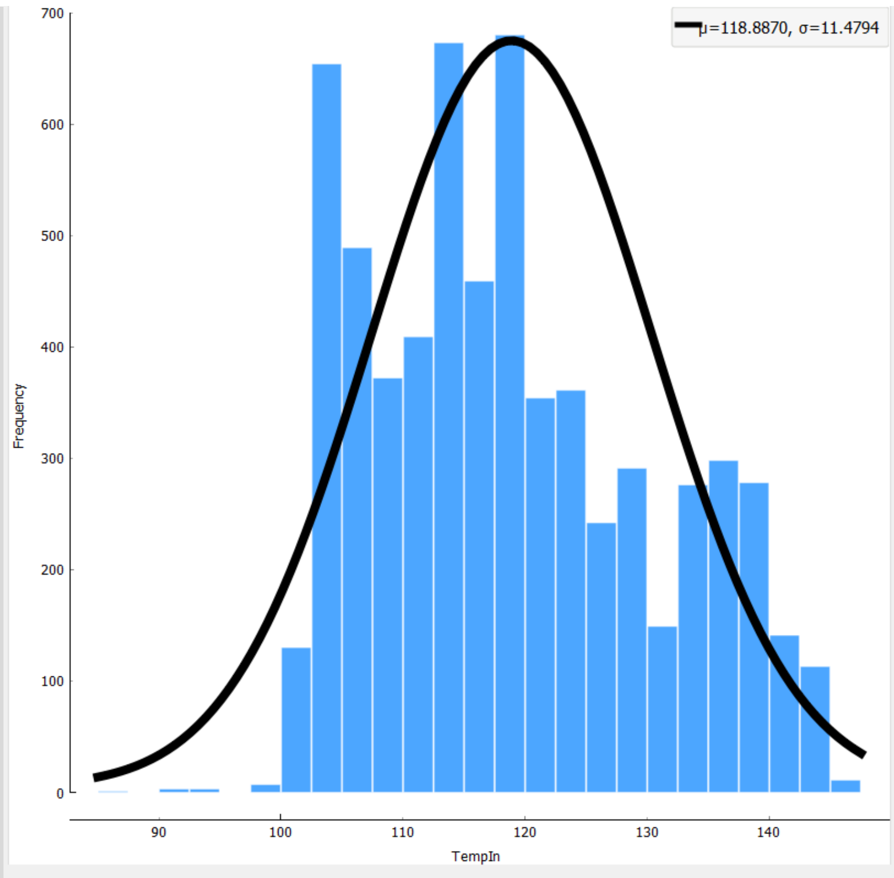
<table>
<tr><td>How many samples?</td><td>6576</td></tr>
<tr><td>How many features?</td><td>15 (combined)</td></tr>
<tr><td>What are the distribution types and ranges for your predictor features?</td><td>Skewed:

</td></tr>
</table>



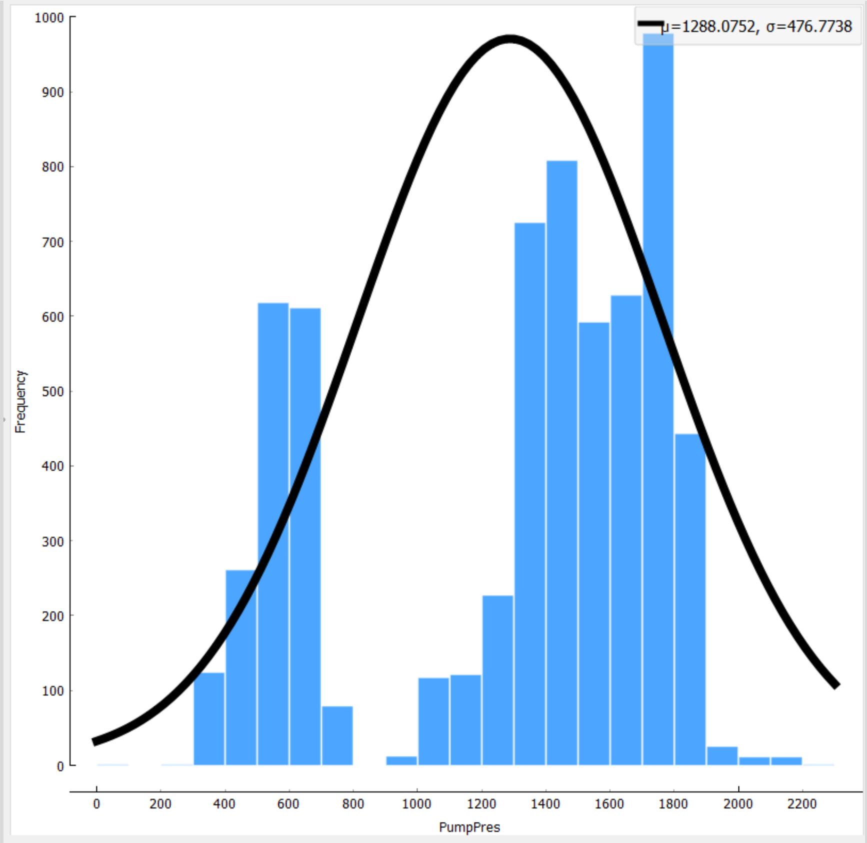
μ=82.1062, σ=26.2201
Frequency
Hookload

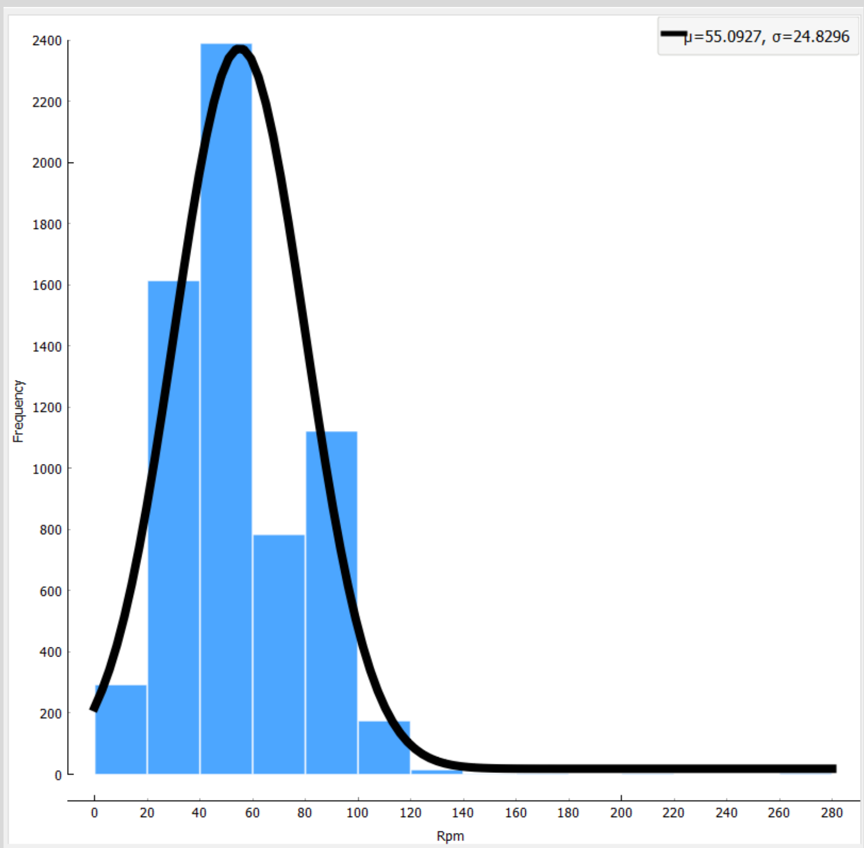


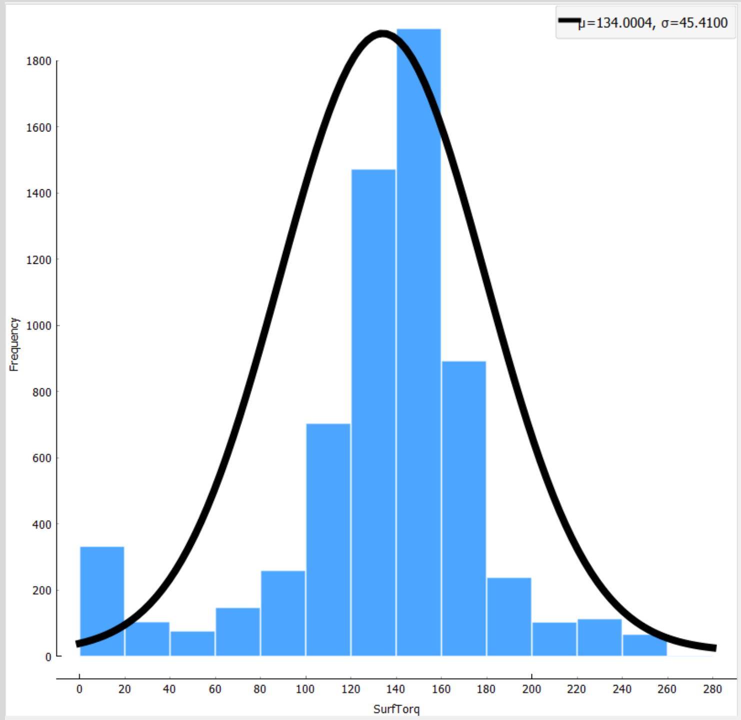
μ=126.8577, σ=11.2515
Frequency
TempOut

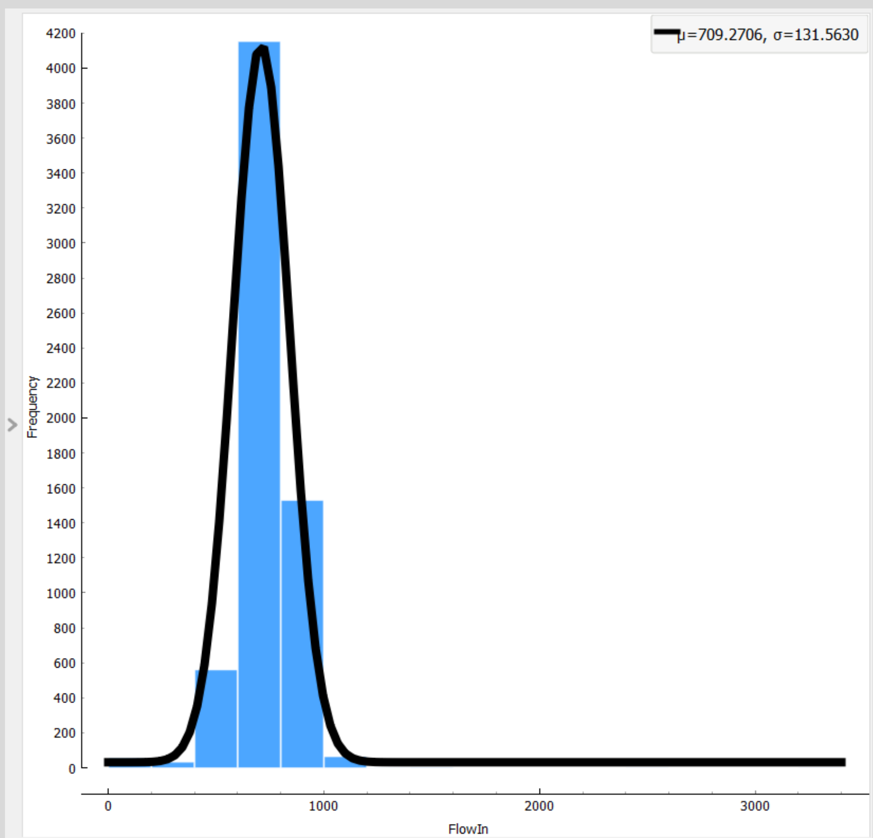


μ=118.8870, σ=11.4794
Frequency
TempIn
0
100
200
300
400
500
600
700
90
100
110
120
130
140


μ=1288.0752, σ=476.7738
Frequency
PumpPres
1000
900
800
700
600
500
400
300
200
100
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200

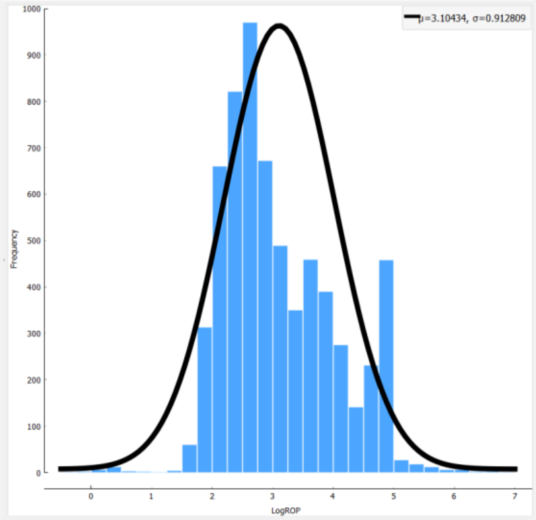

μ=55.0927, σ=24.8296
Frequency
Rpm
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280


μ=134.0004, σ=45.4100
Frequency
SurfTorq


μ=709.2706, σ=131.5630
Frequency
FlowIn

What is your target variable? What is the distribution type and range for the target variable?
Do you see the need for a variable transformation?

Rate of Penetration (ROP) in feet/hour, a measure of how fast the drilling has progressed.



No, seems already has log and looks normal

Use the Scatter Plot and Correlations widget to understand the relationship between the target variable and predictors.

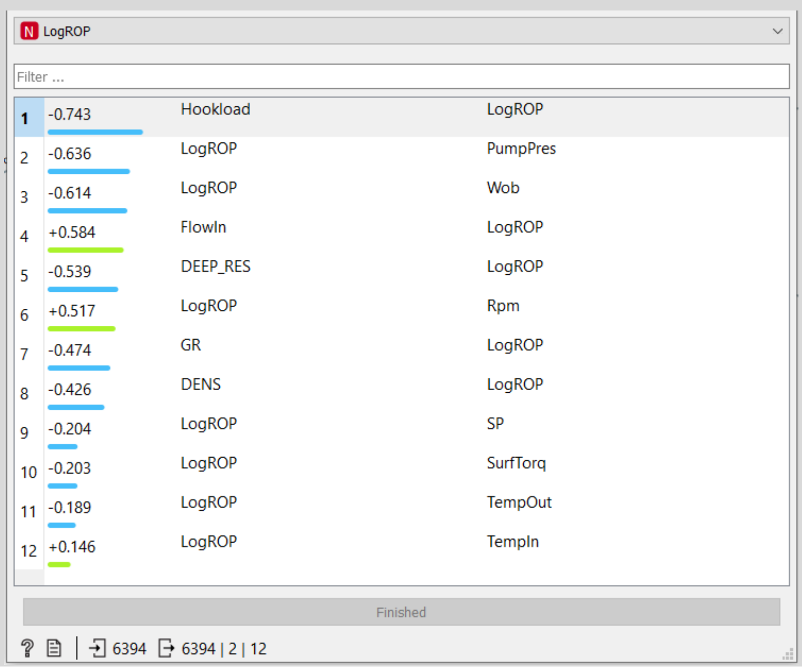
Comment on the relationships, predictors you think will have the biggest impact on predicting the target variable.

Not all of the predictors follow the r line, but very concentrated on one area.

Top 3 below will have the most negative correlation

FlowIn would have the most positive correlation

Up to 8 rank have strong correlations



What do you think about the correlation between predictors?
Identify predictors with moderate to strong correlations.
Explain how would you consider this when selecting variables for your model training.

Wob

Filter ...

1	-0.614	LogROP	Wob
2	+0.576	Hookload	Wob
3	+0.546	PumpPres	Wob
4	-0.542	Rpm	Wob
5	+0.537	GR	Wob
6	-0.507	FlowIn	Wob
7	+0.443	DEEP_RES	Wob
8	+0.325	TempOut	Wob
9	+0.291	DENS	Wob
10	+0.255	SP	Wob
11	+0.182	SurfTorq	Wob
12	+0.025	TempIn	Wob

Finished

TempOut

Filter ...

1	+0.897	TempIn	TempOut
2	+0.528	PumpPres	TempOut
3	-0.352	SP	TempOut
4	+0.325	TempOut	Wob
5	+0.313	DENS	TempOut
6	+0.242	SurfTorq	TempOut
7	-0.230	FlowIn	TempOut
8	-0.216	Rpm	TempOut
9	-0.189	LogROP	TempOut
10	+0.174	Hookload	TempOut
11	-0.140	DEEP_RES	TempOut
12	+0.049	GR	TempOut

TempIn

Filter ...

1	+0.897	TempIn	TempOut
2	-0.593	SP	TempIn
3	-0.405	DEEP_RES	TempIn
4	+0.248	PumpPres	TempIn
5	-0.241	Hookload	TempIn
6	+0.232	DENS	TempIn
7	-0.200	GR	TempIn
8	+0.164	SurfTorq	TempIn
9	+0.146	LogROP	TempIn
10	+0.060	Rpm	TempIn
11	+0.055	FlowIn	TempIn
12	+0.025	TempIn	Wob

Hookload

Filter ...

1	-0.743	Hookload	LogROP
2	+0.628	DEEP_RES	Hookload
3	+0.594	Hookload	PumpPres
4	-0.591	FlowIn	Hookload
5	+0.576	Hookload	Wob
6	-0.566	Hookload	Rpm
7	+0.537	GR	Hookload
8	+0.520	Hookload	SP
9	-0.241	Hookload	TempIn
10	+0.174	Hookload	TempOut
11	+0.138	DENS	Hookload
12	+0.133	Hookload	SurfTorq

SurfTorq

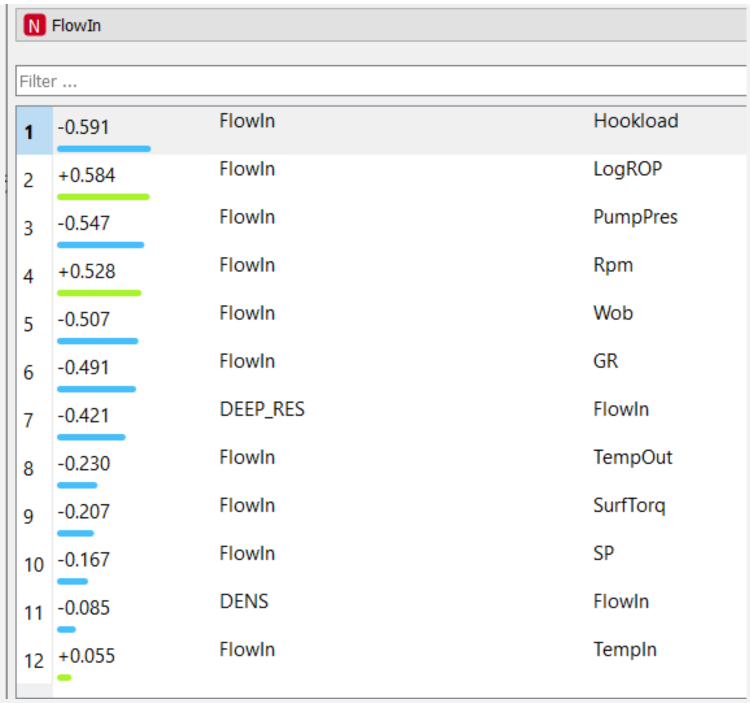
Filter ...

1	+0.262	PumpPres	SurfTorq
2	+0.242	SurfTorq	TempOut
3	-0.207	FlowIn	SurfTorq
4	-0.203	LogROP	SurfTorq
5	+0.182	SurfTorq	Wob
6	+0.179	GR	SurfTorq
7	+0.164	SurfTorq	TempIn
8	+0.133	Hookload	SurfTorq
9	+0.093	DENS	SurfTorq
10	+0.072	DEEP_RES	SurfTorq
11	+0.050	Rpm	SurfTorq
12	-0.045	SP	SurfTorq

Rpm

Filter ...

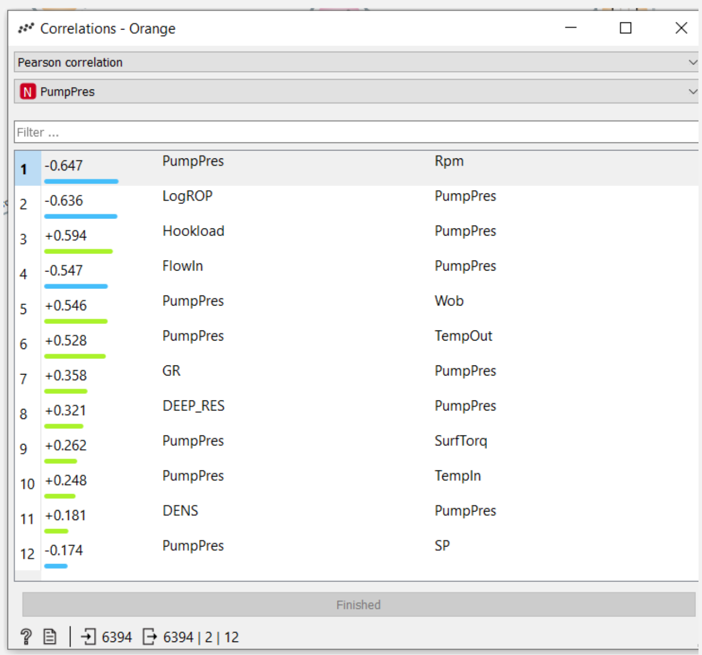
1	-0.647	PumpPres	Rpm
2	-0.566	Hookload	Rpm
3	-0.542	Rpm	Wob
4	+0.528	FlowIn	Rpm
5	+0.517	LogROP	Rpm
6	-0.379	GR	Rpm
7	-0.358	DEEP_RES	Rpm
8	-0.216	Rpm	TempOut
9	-0.125	Rpm	SP
10	+0.060	Rpm	TempIn
11	+0.050	Rpm	SurfTorq
12	+0.013	DENS	Rpm



FlowIn

Filter ...

1	-0.591	FlowIn	Hookload
2	+0.584	FlowIn	LogROP
3	-0.547	FlowIn	PumpPres
4	+0.528	FlowIn	Rpm
5	-0.507	FlowIn	Wob
6	-0.491	FlowIn	GR
7	-0.421	DEEP_RES	FlowIn
8	-0.230	FlowIn	TempOut
9	-0.207	FlowIn	SurfTorq
10	-0.167	FlowIn	SP
11	-0.085	DENS	FlowIn
12	+0.055	FlowIn	TempIn



Correlations - Orange

Pearson correlation

PumpPres

Filter ...

1	-0.647	PumpPres	Rpm
2	-0.636	LogROP	PumpPres
3	+0.594	Hookload	PumpPres
4	-0.547	FlowIn	PumpPres
5	+0.546	PumpPres	Wob
6	+0.528	PumpPres	TempOut
7	+0.358	GR	PumpPres
8	+0.321	DEEP_RES	PumpPres
9	+0.262	PumpPres	SurfTorq
10	+0.248	PumpPres	TempIn
11	+0.181	DENS	PumpPres
12	-0.174	PumpPres	SP

Finished

6394 6394 | 2 | 12

GR

Filter ...

1	+0.537	GR	Hookload
2	+0.537	GR	Wob
3	+0.504	DEEP_RES	GR
4	-0.491	FlowIn	GR
5	-0.474	GR	LogROP
6	-0.379	GR	Rpm
7	+0.358	GR	PumpPres
8	+0.309	GR	SP
9	-0.200	GR	TempIn
10	+0.179	GR	SurfTorq
11	-0.055	DENS	GR
12	+0.049	GR	TempOut

DEEP_RES

Filter ...

1	+0.628	DEEP_RES	Hookload
2	-0.539	DEEP_RES	LogROP
3	+0.504	DEEP_RES	GR
4	+0.443	DEEP_RES	Wob
5	+0.441	DEEP_RES	SP
6	-0.421	DEEP_RES	FlowIn
7	-0.405	DEEP_RES	TempIn
8	-0.358	DEEP_RES	Rpm
9	+0.321	DEEP_RES	PumpPres
10	-0.140	DEEP_RES	TempOut
11	+0.072	DEEP_RES	SurfTorq
12	+0.047	DEEP_RES	DENS

DENS

Filter ...

1	-0.426	DENS	LogROP
2	+0.313	DENS	TempOut
3	+0.291	DENS	Wob
4	+0.232	DENS	TempIn
5	+0.181	DENS	PumpPres
6	+0.138	DENS	Hookload
7	+0.093	DENS	SurfTorq
8	-0.085	DENS	FlowIn
9	-0.055	DENS	GR
10	+0.047	DEEP_RES	DENS
11	-0.045	DENS	SP
12	+0.013	DENS	Rpm

So based on above of predictors, each will have their own correlation that they are strong to and can be unique. The stronger correlations will definitely have impacts when we are selecting the variables.

Do you see any anomalous data instances? If yes, what % of the dataset is affected and how would you deal with this anomaly?

Hint: Can you drill with zero WOB?

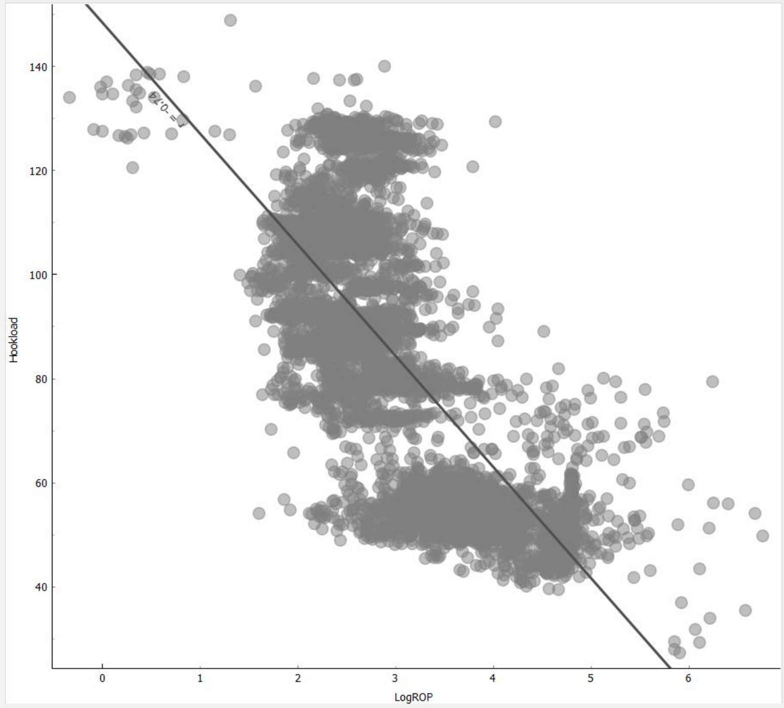
The original merge data, the an anomalous of ROP (target) is that:

Mean	Median	Dispersion	Min.	Max.	Missing
42.4851	18.08	1.8347	0.71	2977.91	0 (0%)

You cannot drill 2977 ft/hr

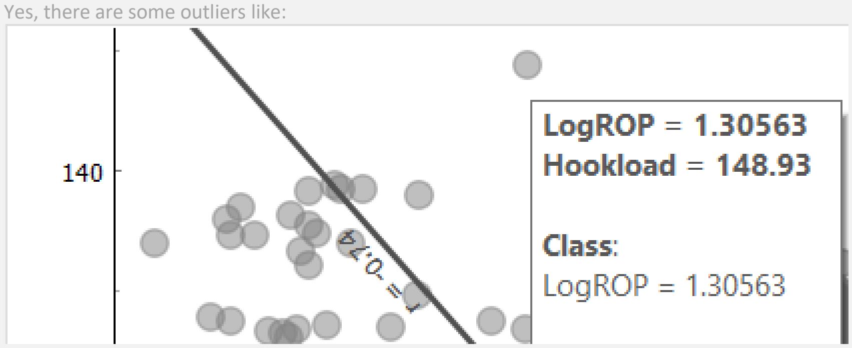
You cannot have 0.00 WOB as min

Consider the Scatter plot between your strongest predictor and target variable.



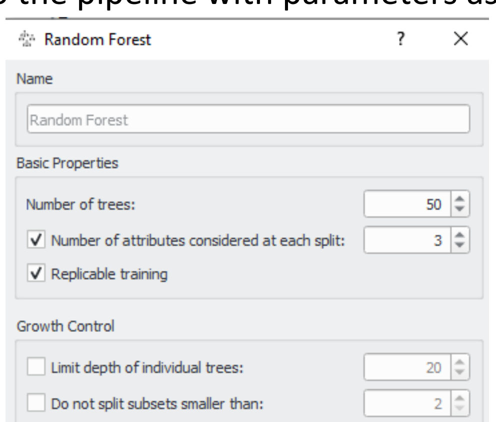
Do you see any outliers? How would you deal with the outliers?

Yes, there are some outliers like:



Remove one obvious outlier by filtering out ROP > 2900, transforming the variable like log…

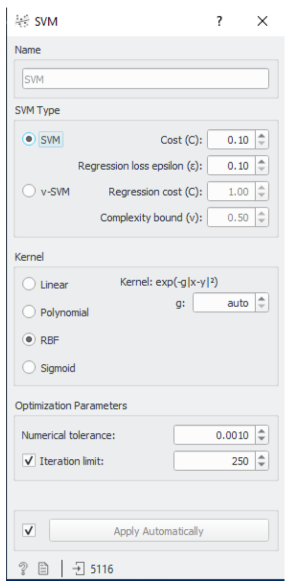
3. Add **Random Forest** trainer to the pipeline with parameters as shown below:



4. Complete the table below based on **cross validation (5-fold)** on your training data using metrics from Evaluation Results in the **Test and Score** widget. (30 points)

Model	RMSE	MAE	R^2
kNN	0.269	0.180	0.914
Linear Regression	0.433	0.326	0.778
Random Forest	0.251	0.168	0.926

5. Add **SVM (RBF)** trainer to the pipeline and tune model using following parameters.



SVM Parameters	R^2
C=0.1, Loss=0.1	0.646
C=0.1, Loss=0.5	0.828
C=0.1, Loss=1	0.694
C=0.5, Loss=0.1	0.731
C=0.5, Loss=0.5	0.866
C=0.5, Loss=1	0.751
C=1, Loss=0.1	0.675
C=1, Loss=0.5	0.870
C=1, Loss=1	0.763

C=10, Loss=0.1	0.643
C=10, Loss=0.5	0.834
C=10, Loss=1	0.750

6. Set **SVM** parameters to the best performing model from the above table and complete the table below with **Sampling** as **Cross Validation, 10 folds:**

Model	RMSE	MAE	R^2
kNN	0.269	0.180	0.914
Linear Regression	0.433	0.326	0.778
Random Forest	0.251	0.168	0.926
SVM	0.332	0.257	0.870

Complete the table for **Test on test data:**

Model	RMSE	MAE	R^2
kNN	0.271	0.184	0.907
Linear Regression	0.432	0.327	0.763
Random Forest	0.239	0.167	0.927
SVM	0.325	0.256	0.866