Badge-2 Kaggle Competition

Out date: Jul 27, 2022 @ 18:00HRS
Due date: Jul 31, 2022 @ 23:59HRS

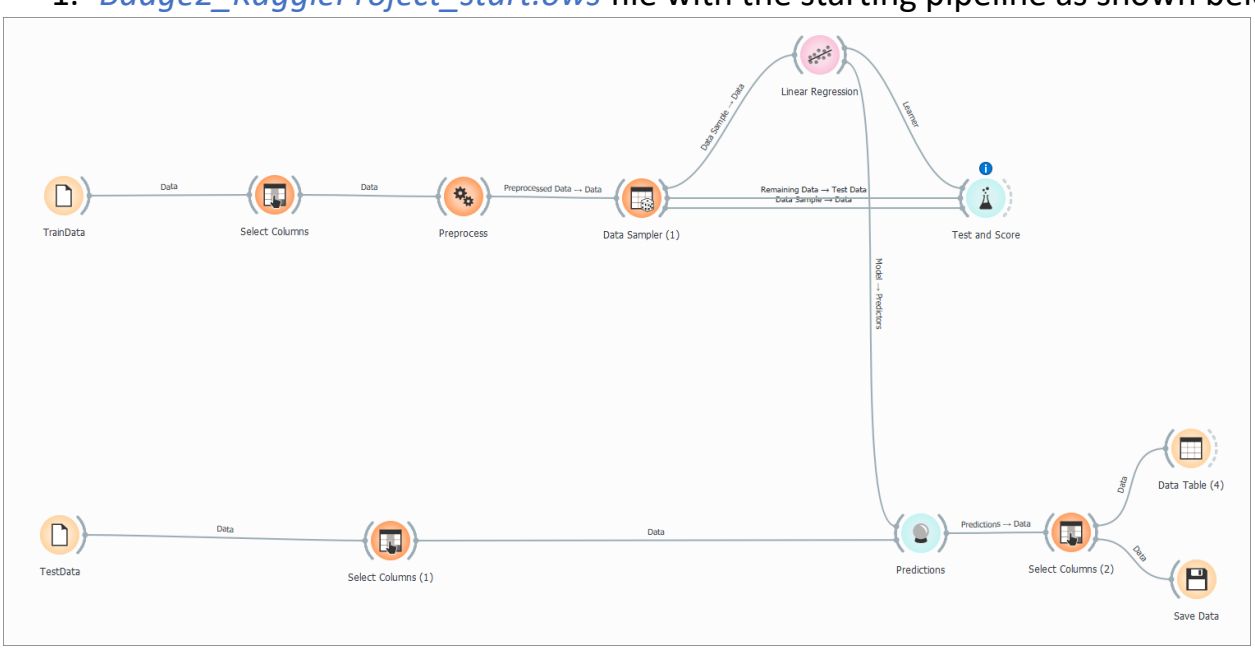
Submission

1. Prepare your solution in Orange and save the workspace (e.g., Badge2_Kaggle_LastName.ows).
2. Complete the tables provided in this document and save the document. (e.g., Badge2_Kaggle_LastName.docx)
3. Save your final predictions in a csv file (e.g., Badge2_Kaggle_Predictions_LastName.csv).
4. Upload the following files to Canvas :
    a. Badge2_Kaggle_LastName.ows
    b. Badge2_Kaggle_LastName.docx
    c. Badge2_Kaggle_Predictions_LastName.csv

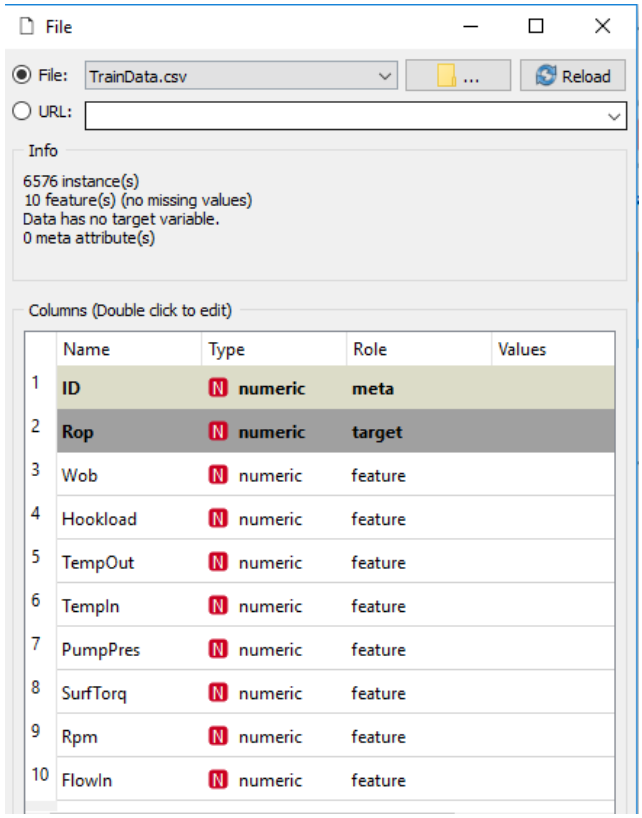
Objective: Your customer is interested in predicting **Rate of Penetration (ROP)** using drilling diagnostics data contained in the *TestData.csv* file. As a data scientist, your job is to train machine learning models using the *TrainData* file provided to you using **Orange,** deploy your best performing model to predict **ROP** and submit the predictions to your client.

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

1. *Badge2_KaggleProject_start.ows* file with the starting pipeline as shown below:

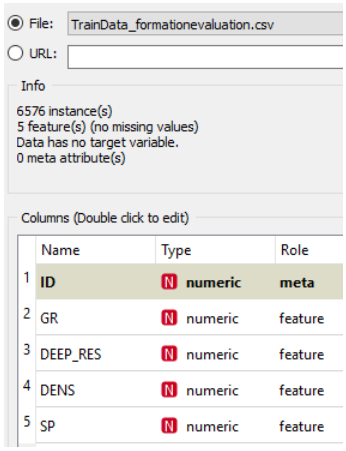


2. *TrainData.csv* containing the following features:



	Name	Type	Role	Values
1	**ID**	**numeric**	**meta**	
2	**Rop**	**numeric**	**target**	
3	Wob	numeric	feature	
4	Hookload	numeric	feature	
5	TempOut	numeric	feature	
6	TempIn	numeric	feature	
7	PumpPres	numeric	feature	
8	SurfTorq	numeric	feature	
9	Rpm	numeric	feature	
10	FlowIn	numeric	feature	

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.



	Name	Type	Role
1	**ID**	**numeric**	**meta**
2	GR	numeric	feature
3	DEEP_RES	numeric	feature
4	DENS	numeric	feature
5	SP	numeric	feature

4. *TestData.csv and TestData_formationevaluation.csv* contains the same features as shown in points 3 and 4 above except the ROP target variable

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)

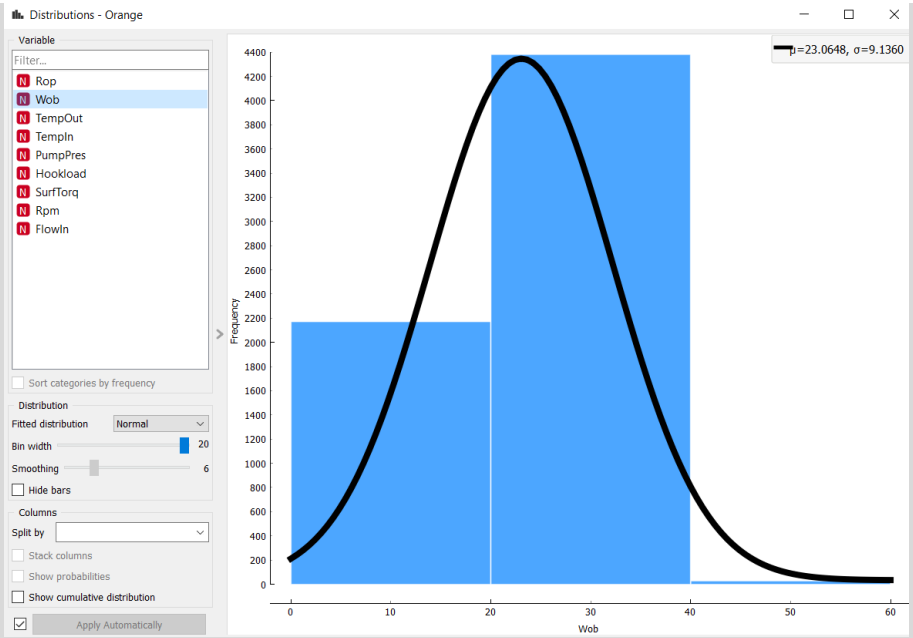
https://www.youtube.com/watch?v=guFiQ87tg_s (a video about the oil well drilling process)

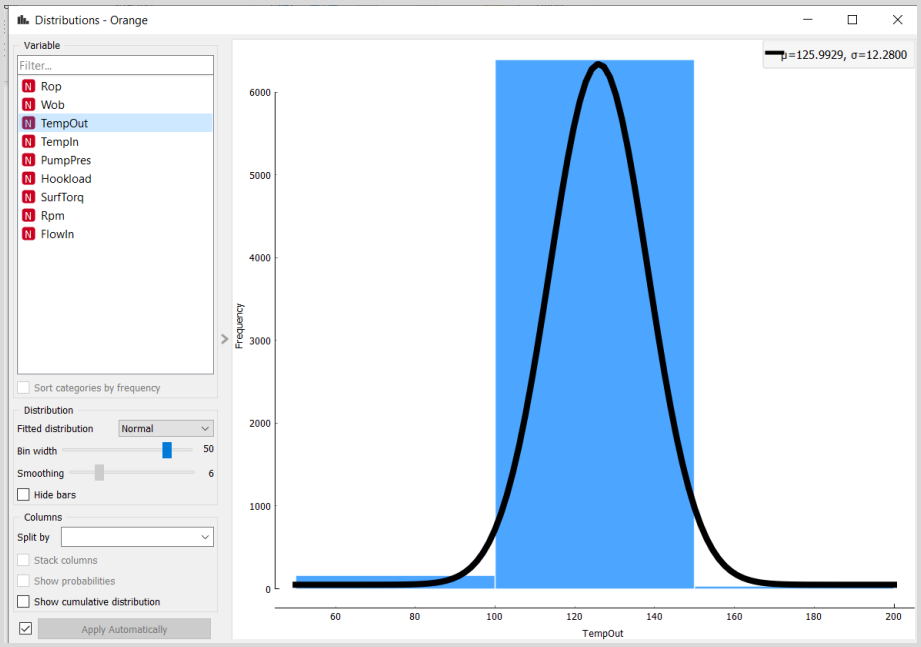
Project Instructions

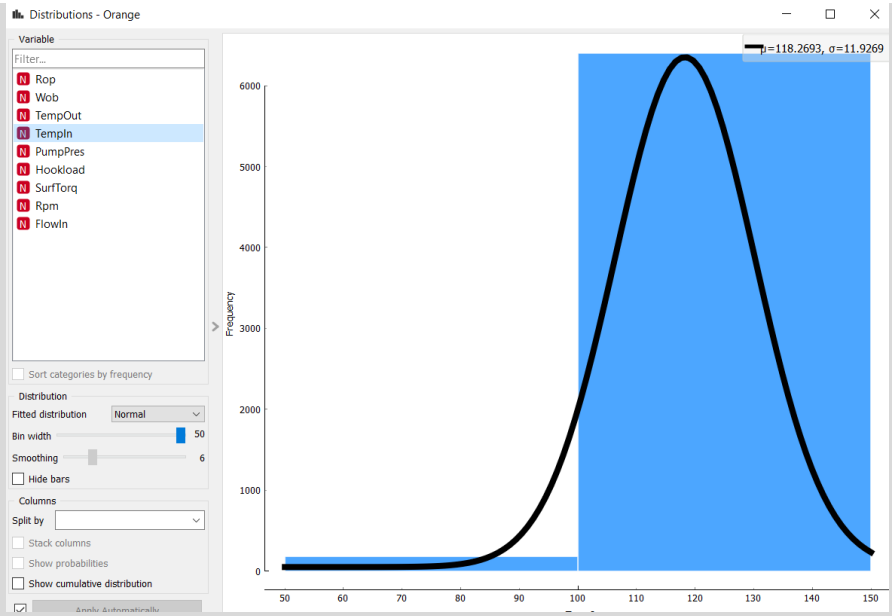
1. Launch Orange and open the *Badge2_KaggleProject_start.ows file*.
2. Add required widgets to the pipeline and answer the questions below: (15 points)

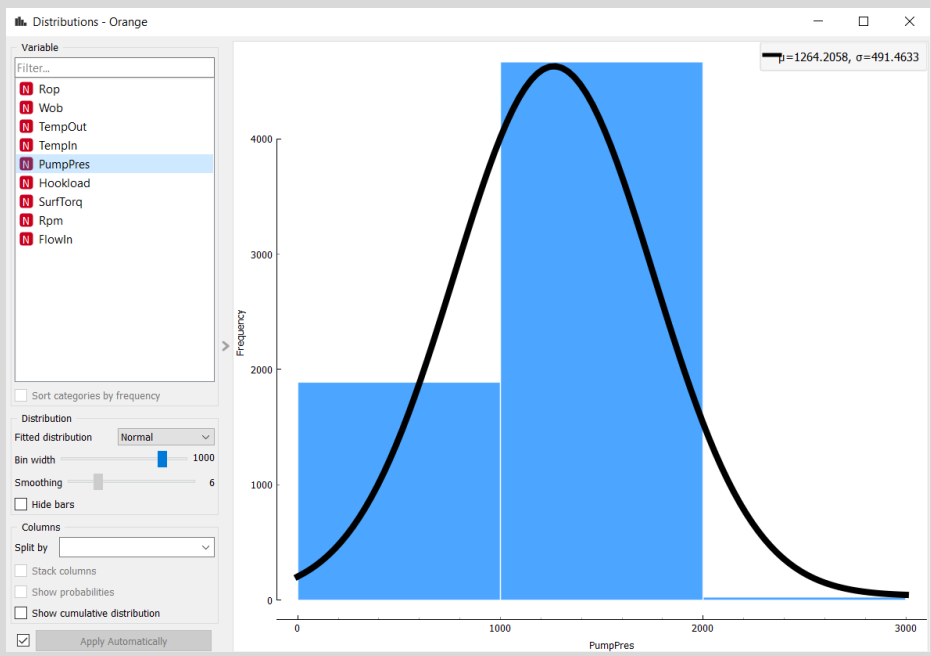
How many samples?	Data Set Size Rows: 6576 Columns: 10
How many features?	Features Categorical: - Numeric: 8

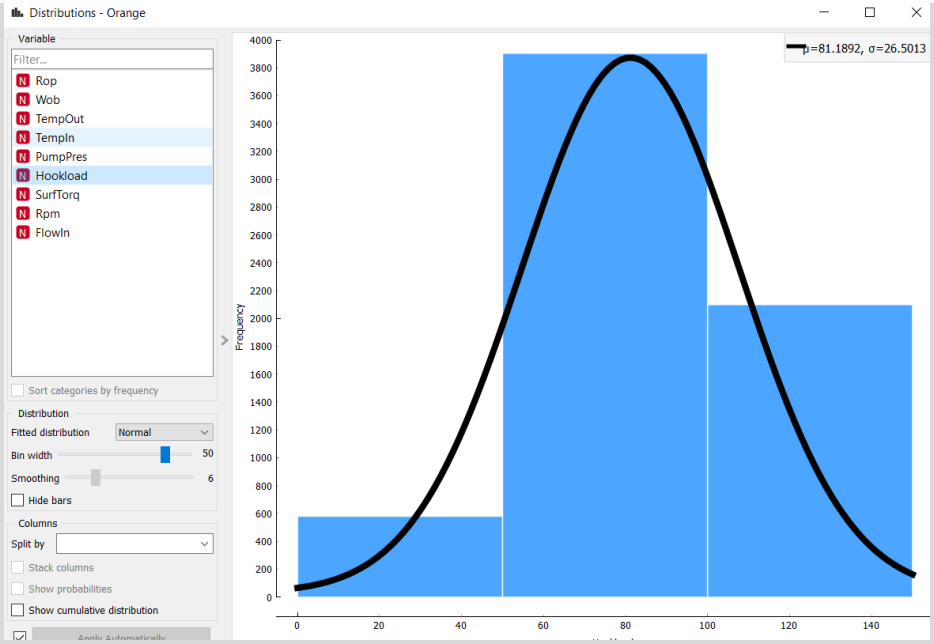
What are the distribution types and ranges for your predictor features?

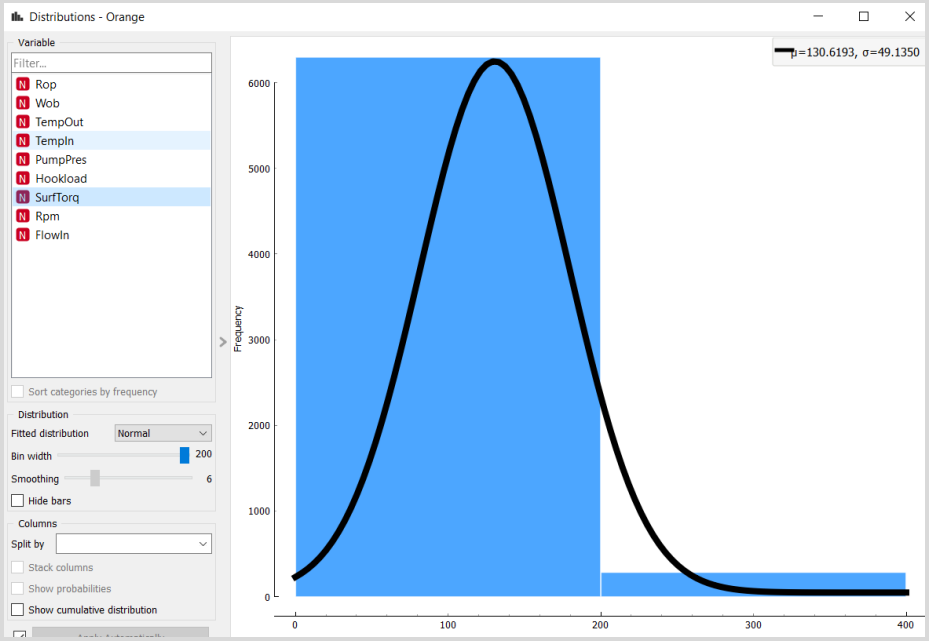


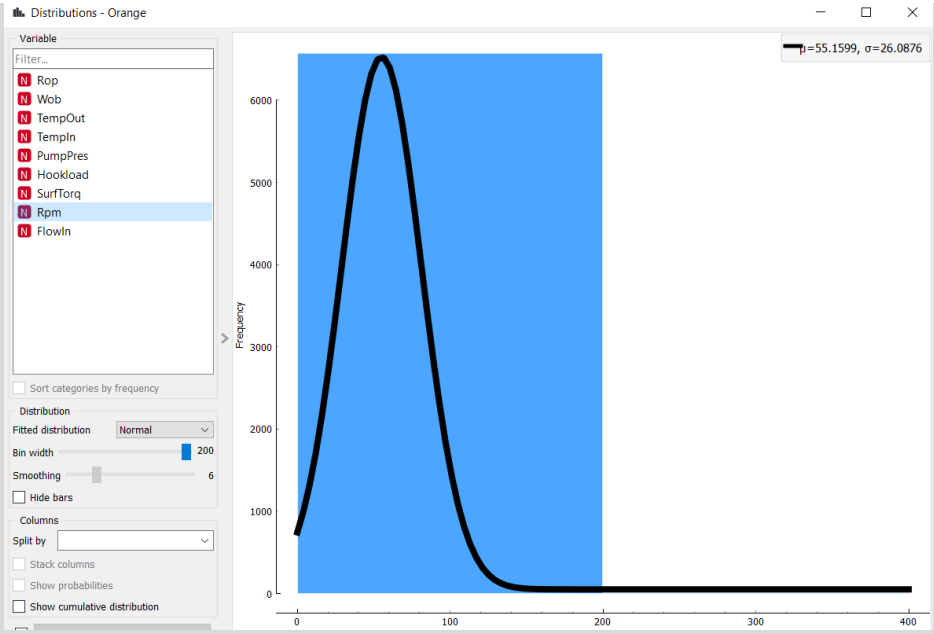


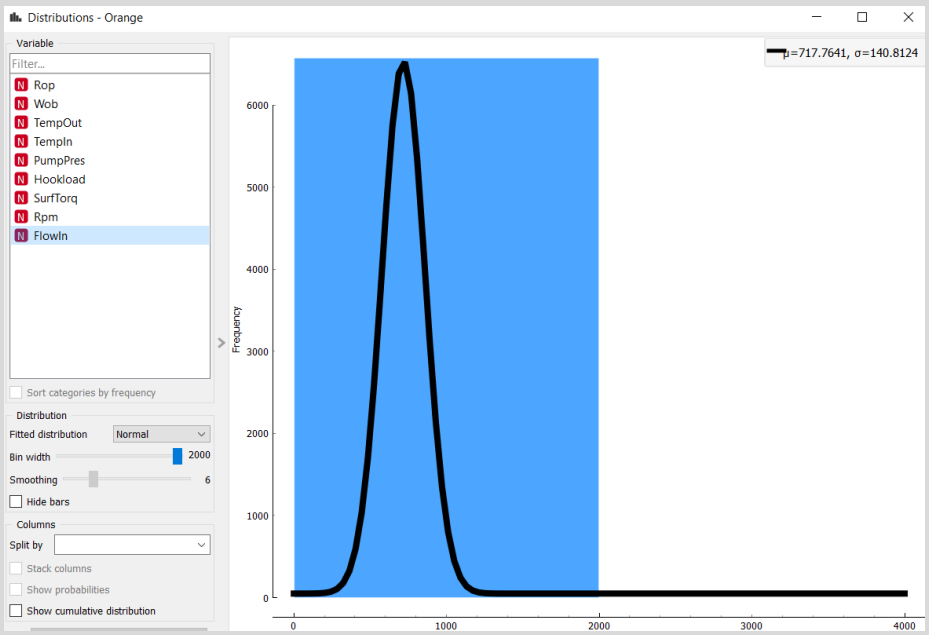

Distributions - Orange
Variable
Filter...
Rop
Wob
TempOut
TempIn
PumpPres
Hookload
SurfTorq
Rpm
FlowIn
Sort categories by frequency
Distribution
Fitted distribution
Normal
Bin width
50
Smoothing
6
Hide bars
Columns
Split by
Stack columns
Show probabilities
Show cumulative distribution
μ=118.2693, σ=11.9269
Frequency
6000
5000
4000
3000
2000
1000
0
50
60
70
80
90
100
110
120
130
140
150


Distributions - Orange
Variable
Filter...
Rop
Wob
TempOut
TempIn
PumpPres
Hookload
SurfTorq
Rpm
FlowIn
Sort categories by frequency
Distribution
Fitted distribution
Normal
Bin width
1000
Smoothing
6
Hide bars
Columns
Split by
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
μ=1264.2058, σ=491.4633
Frequency
4000
3000
2000
1000
0
0
1000
2000
3000
PumpPres

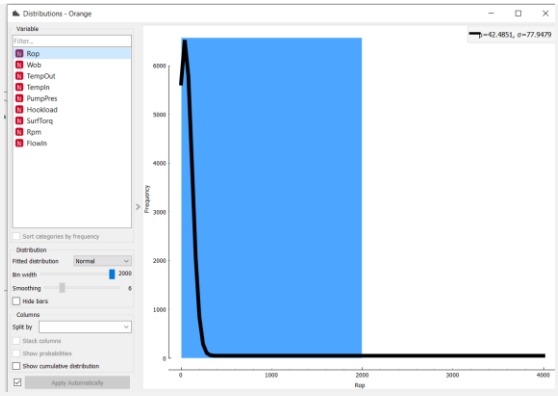

Distributions - Orange
Variable
Filter...
Rop
Wob
TempOut
TempIn
PumpPres
Hookload
SurfTorq
Rpm
FlowIn
Sort categories by frequency
Distribution
Fitted distribution
Normal
Bin width
50
Smoothing
6
Hide bars
Columns
Split by
Stack columns
Show probabilities
Show cumulative distribution
μ=81.1892, σ=26.5013
Frequency


Distributions - Orange
Variable
Filter...
Rop
Wob
TempOut
TempIn
PumpPres
Hookload
SurfTorq
Rpm
FlowIn
Sort categories by frequency
Distribution
Fitted distribution
Normal
Bin width
200
Smoothing
6
Hide bars
Columns
Split by
Stack columns
Show probabilities
Show cumulative distribution
μ=130.6193, σ=49.1350
Frequency





What is your target variable? What is the distribution type and range for the target variable?	2 **Rop** N **numeric** **target**

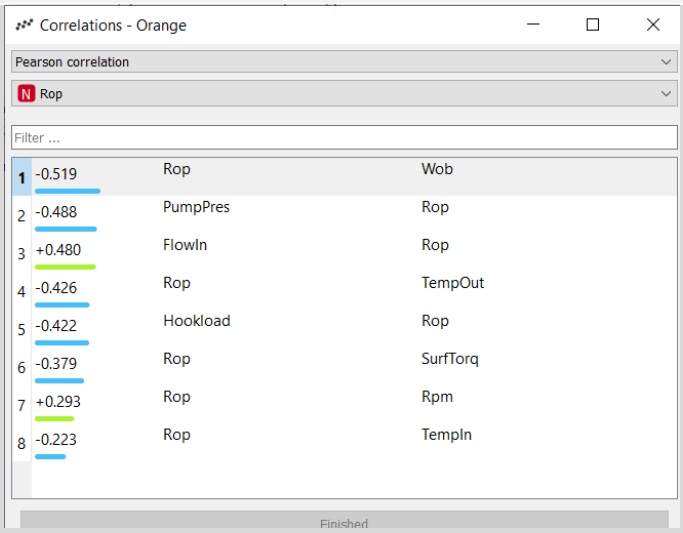


Do you see the need for a variable transformation?

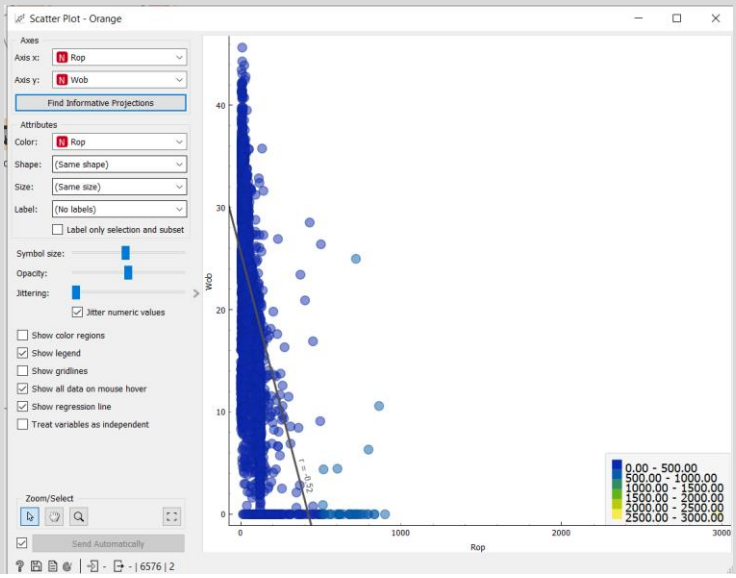
It might make it look and to distinguish meaning better

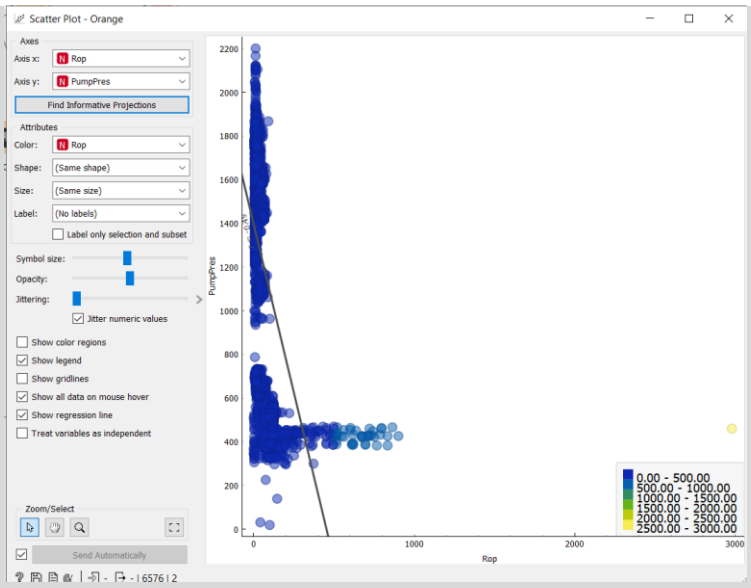
Use the Scatter Plot and Correlations widget to understand the relationship between the target variable and predictors.

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



Seems Wob is the strongest relationship; Flowin the most positive with the target



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.	 Not too perfect since not all the plots are lined to the r and likely not deploy the model, but still wouldn't discard this to see what can be improved Each predictors have their strong correlations, expected to be independent. A change in the target can make an impact on the predictors.
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? Yes, like for example, there are large amounts of zero values in WOB I would try to preprocess the data by adding an impute widget, since it didn't register anything; some of them could have been a fixed value depending on what it is like average. The missing values can be replaced with a zero.

Feature Statistics - Orange

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
PumpPres		1264.2058	1430.4050	0.3888	19.94	2200.43	0 (0%)
FlowIn		717.7641	700.52	0.1962	0.00	3317.51	0 (0%)
SurfTorq		130.6193	140.18	0.3762	0.00	271.54	0 (0%)
TempOut		125.9929	124.83	0.0975	84.07	151.70	0 (0%)
TempIn		118.2693	117.18	0.1008	85.00	146.31	0 (0%)
Hookload		81.1892	80.07	0.3264	27.27	148.93	0 (0%)
Rpm		55.1599	50.5750	0.4729	0.00	271.58	0 (0%)
Wob		23.0648	23.82	0.3961	0.00	45.64	0 (0%)
Rop		42.4851	18.08	1.8347	0.71	2977.91	0 (0%)

Consider the Scatter plot between your strongest predictor and target variable. Do you see any outliers? How would you deal with the outliers?	Yes, I would look into the Feature Constructor as we did in Lab 3. We could average out to the mean or apply transformation (look at log or square root of variable) to the target variable (changes distribution of target to make it valid more likely)

3. You are free to apply all the techniques you have learnt in Energy Data Analytics Bronze Belt (Badges 1 & 2) program to train your Machine Learning models on the input dataset that you loaded in the above steps.
   You are encouraged to consider the following aspects:
   a. Preprocessing features (Imputing any missing values, Standardizing or Normalizing features)

   b. Dimensionality reduction using PCA

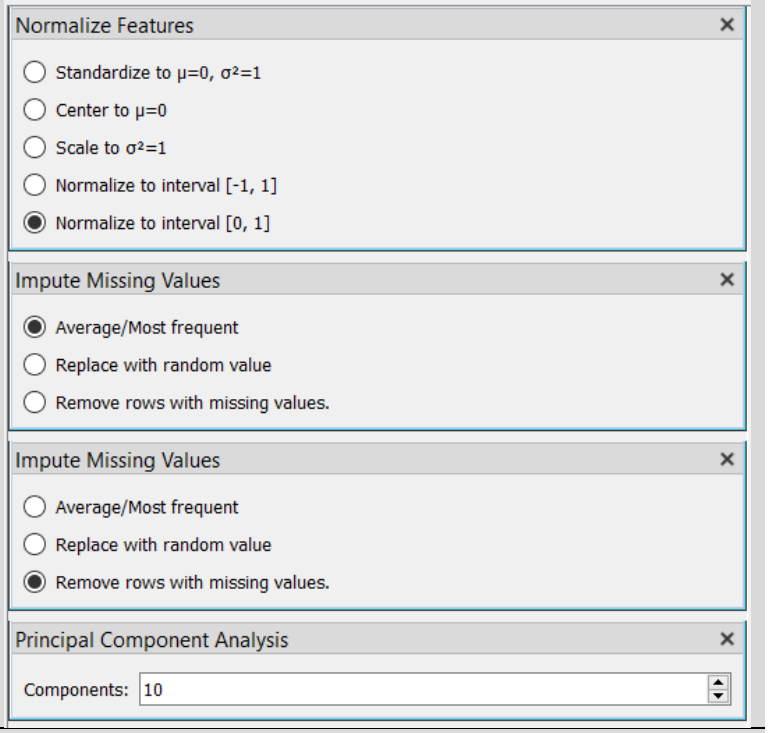
   c. Splitting your training dataset to train and test to test and compare the performance of your model(s). Consider selecting Replicable sampling for repeatable results.

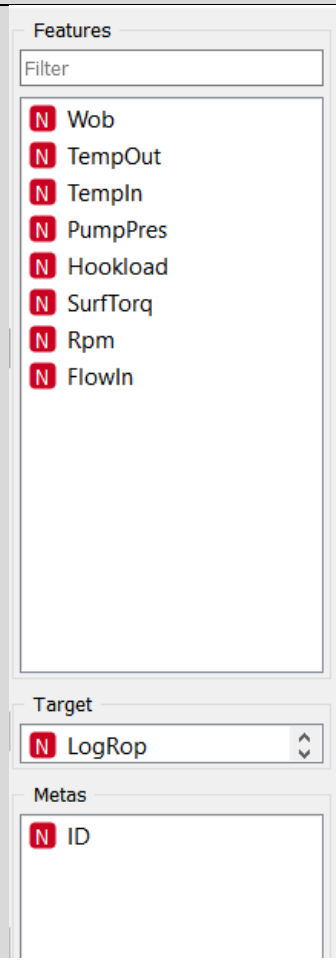
   d. Using Test and Score widget and metrics such as R^2, RMSE and MAE evaluate the performance of your various models on your training / test dataset and when you tune various parameters.

e. Consider diagnostic procedures such as plotting your predicted vs actual target variable and residual vs predicted output from your model to understand the validity of the base linear regression model and decide if your target variable and / or predictors would need variable transformation such as log or square root.

f. Consider the effect formation evaluation parameters provided as a separate dataset in your model performance.

Complete the table below: (15 points)

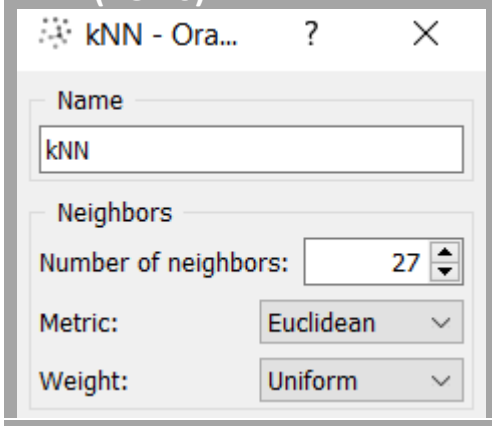
What data preprocessing options did you consider?	
Did you consider any transformation for your target variable and predictors? List the variables and transformations used for your final model.	Yes, using the feature constructor widget as log(Rop)
Did you explore PCA for dimensionality reduction and evaluate its impact on model performance?	Yes, added some PCA widgets and explores increasing and decreasing
If you used dimensionality reduction using PCA for your final model, provide details on how many components you considered and the % variance explained:	3 components to get 80% variance, 1 if not normalized

What is your best performing model and model parameters?	Modified kNN with different PCA.
What is the final list of predictors you considered for your best performing model?	

Complete the table below based on **cross validation (10-fold)** on your training data using metrics from Evaluation Results in the **Test and Score** widget. You are required to try at least <u>five different</u> models. (30 points)

Model	RMSE	MAE	R^2
Linear Regression (example entry)	64.73	28.13	0.37
ROP (Feature constructor and Impute)	RMSE 0.548	MAE 0.423	R2 0.689

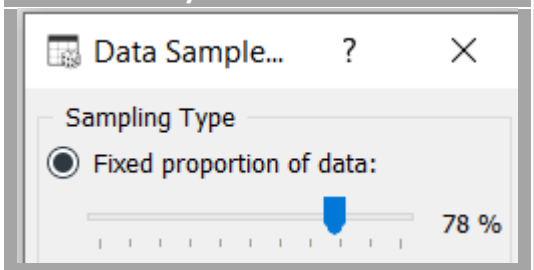
kNN (PCA 6)



RMSE	MAE	R2
0.318	0.211	0.895

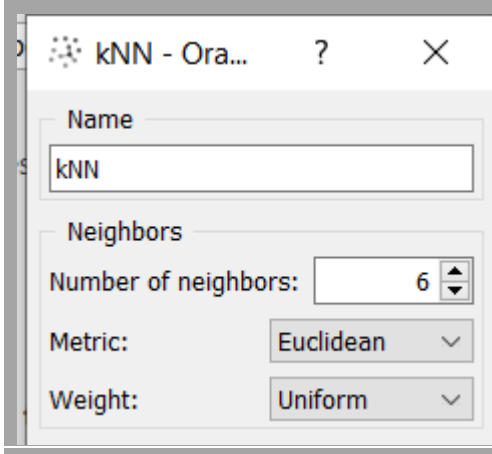
Tree (min-15, do not split-5, limit max and stop majority unchecked)

0.334	0.225	0.885



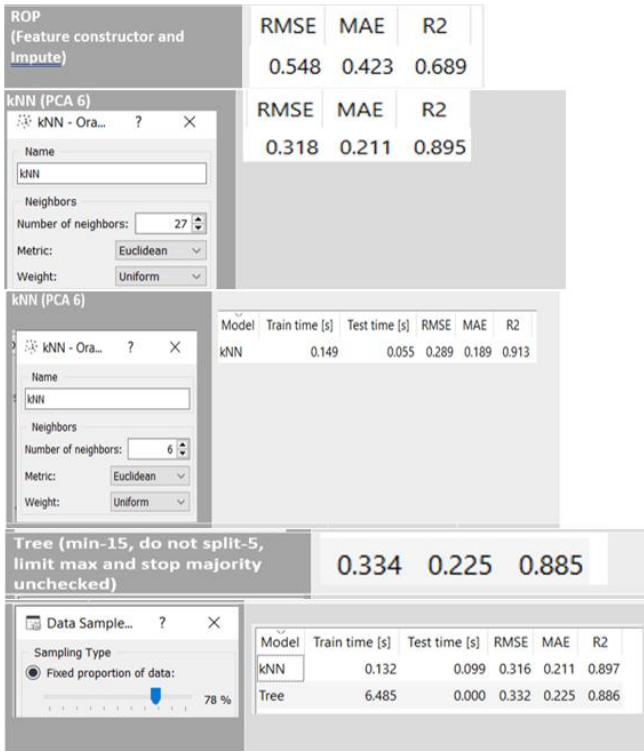
Model	Train time [s]	Test time [s]	RMSE	MAE	R2
kNN	0.132	0.099	0.316	0.211	0.897
Tree	6.485	0.000	0.332	0.225	0.886

kNN (PCA 6)



Model	Train time [s]	Test time [s]	RMSE	MAE	R2
kNN	0.149	0.055	0.289	0.189	0.913

BEST: kNN-6 (PCA-6)



Once you have trained your models and decided on the best model, select **Test on test data** in the **Test and Score** widget and use the Evaluation Results to complete the table below (10 points):

Test Data:

Model	RMSE	MAE	R^2

Model	Train time [s]	Test time [s]	RMSE	MAE	R2
kNN	0.027	0.012	0.291	0.192	0.910

CV-10:

Evaluation Results

Model	Train time [s]	Test time [s]	RMSE	MAE	R2
kNN	0.138	0.051	0.289	0.189	0.913

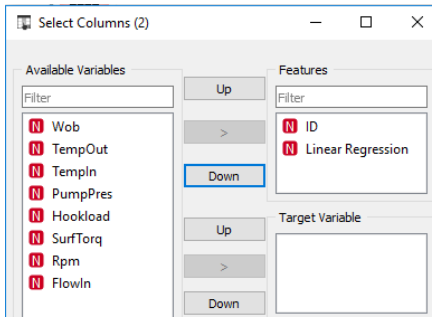
How do you interpret your final model based on its performance on the test data?	It took less train and test time, improved the RMSE and MAE, but the R2 went down slightly.

4. Open the **File** widget and load the *TestData.csv* file. (5 points)

What is the sample size of the TestData file?	Data Set Size Rows: 730 Columns: 9

5. Verify the rest of this pipeline and ensure **Predictions** widget contains the predicted **ROP** values for all the samples, from your best model. Please remember to transform your predicted output to **ROP**, if applicable.

6. Use **Select Columns(2)** widget shown in the above below to select only the predicted ROP values & ID meta data and verify using the **Data Table** widget.



(Move all except ID and the best model (like kNN for mine) to the left to be ignored and not show up in the excel columns)

7. Use **Save Data** widget to save your predictions as *Badge2_Kaggle_Predictions_LastName.csv.* Uncheck **Add type annotations to header** and save the .csv file to your local folder.

8. Ensure your Predictions csv file should output results in two columns: 1) ID, and 2) the predicted ROP values as shown below. Please ensure column B is renamed as ROP.

	A	B
1	ID	ROP
2	1	200.3691
3	2	231.421
4	3	212.5713
5	4	249.324
6	5	191.4097

9. Upload the csv file to Kaggle using the link below. You can upload up to **10** csv files during the competition. (15 points)

   https://www.kaggle.com/t/7095af5567fb46e5a95647e295b786bf

Make sure to select column A (ID), right click to format, change to text, if get errors

10. Upload your final '.ows' and final '.csv' predictions to Canvas. (10 points)