Badge-2 Lab-4 [Regression Trees & Logistic Regression]

Out date: July 20, 2020
Due date: July 24, 2020 at 11:59PM

Submission

1. Prepare your solution in Orange and save the workspace for Problem 1 (e.g., Lab-4_1_LastName.ows) **[10 points]**
2. Prepare your solutions in Orange and save the workspace for Problem 2 (e.g., Lab-4_2_LastName.ows) **[10 points]**
3. Complete the tables given below and save the file (e.g., Lab-2_LastName.docx). **[80 points]**
4. Upload the files to the Canvas.

Objective(s):

To compare performance of multi-linear regression with regression tree and other machine learning algorithms.

To apply Logistic regression classifier for a classification problem and compare its performance with other machine learning algorithms.

Data:

Horizontal drilling which provides additional exposure of the reservoir to the wellbore (lateral length) is one of the primary drivers that made economic production of oil and gas from the tight shale reservoirs in the US a success, in conjunction with stimulation using hydraulic fracturing.

Objective of the exercise is to understand effect of lateral length of a horizontal well on shale gas production in some of the US shale plays using multiple linear regression analysis. While lateral length would be one of the primary predictors, other variables considered are vertical section length of the horizontal well, proppant volume and fluid volume pumped for stimulating the horizontal section of the well using hydraulic fracturing and type of shale plays.

Data source: National Oil and Gas Gateway (http://www.noggateway.org/explore). From participating states that contribute to this data source, Arkansas and Colorado were identified as states of interest for the project owing to the level of shale gas drilling and production activity in these states over the period 2008-2018 and data availability on Fluids and Proppants.

Variables of interest:

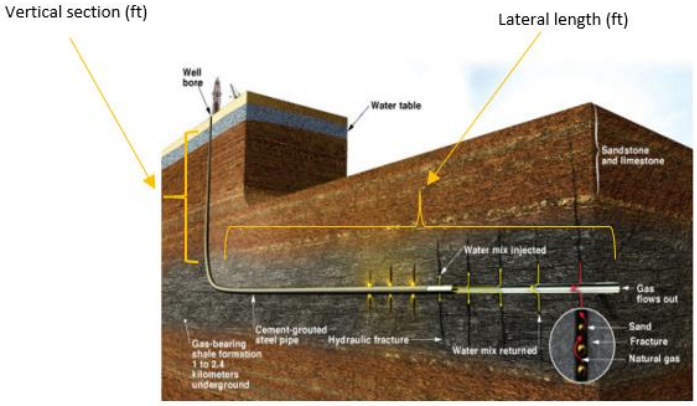


Figure 1: Horizontal well schematic (from The Academy of Medicine, 2017)

Referring to the Figure-1 above, the following were identified as variables of interest after reviewing the dataset:

Response (Target) variable: Max_Gas, Maximum Gas Production (Million Standard Cubic Feet, MSCF)
Since shale gas wells have their peak production in the first 2-3 years after the well is drilled, stimulated and put on production, maximum annual gas production is selected as the best response variable to predict using the predictor variables identified below

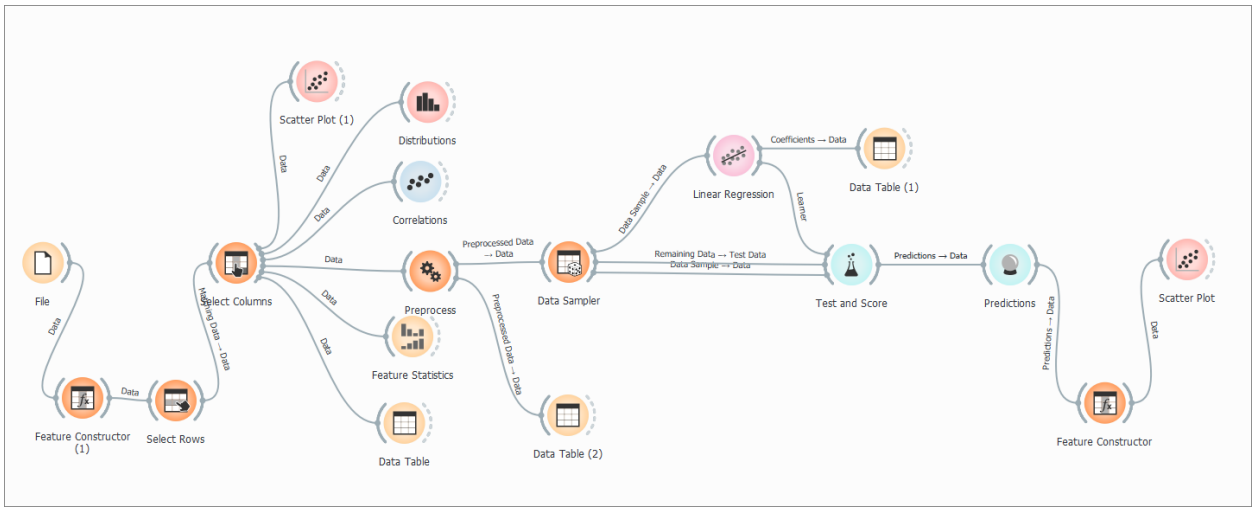
Predictor variables:
Wellhead Latitude and Longitude (degrees)
Lat_Len, Lateral length of the horizontal section (feet)
Vert_Len, Length of the vertical section (feet)
Fluid, Total amount of fracturing fluid (typically water) pumped to create the hydraulic fractures (Barrels)
Proppant, Total amount of proppant (typically sand) pumped along with the fracturing fluid as a slurry to keep the hydraulically created fractures open (Pounds)
ShalePlay (a dummy variable, 0- Fayetteville, Arkansas Shale gas play and 1- Mancos, Colorado Shale gas play)
Drilling Orientation (dummy variable)

Problem 1/2. [50 points]

Data: For this lab, we will continue from the lab-3 on linear regression.

Lab Instructions

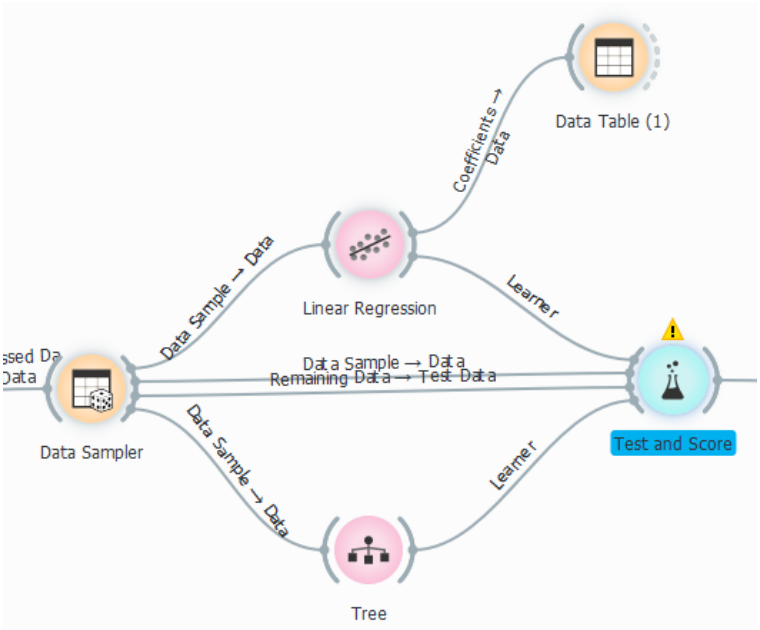
1. Open the *gas_prod_pred_lab3.ows* file in Orange. Your pipeline should look as below. Confirm the *data.xlsx* dataset is loaded by opening the **File** widget.



2. Inspect the pipeline. Open **Test and Score** widget and complete the table below using **Cross validation, 5 folds** for sampling: (10 points)

Model	RMSE	MAE	R2
Linear Regression (Cross Validation, 10 folds)	0.622	0.431	0.635

3. Add **Tree** widget as shown below:



4. Inspect the rest of the pipeline. Open **Test and Score** widget and observe the model performances for the following model scenarios using **cross validation (5 folds)** as **Sampling** method. (30 points)

Model	RMSE	MAE	R2
Linear Regression	0.622	0.431	0.635
Tree (5 Min,5 subset)	0.619	0.418	0.639
Tree (5,10)	0.619	0.418	0.639
Tree (5,20)	0.597	0.404	**0.664**
Tree (10, 10)	0.606	0.409	0.654
Tree (10, 20)	0.606	0.409	0.654
Tree (10, 50)	0.598	0.405	0.663
Tree (30,50)	0.592	0.401	**0.670**

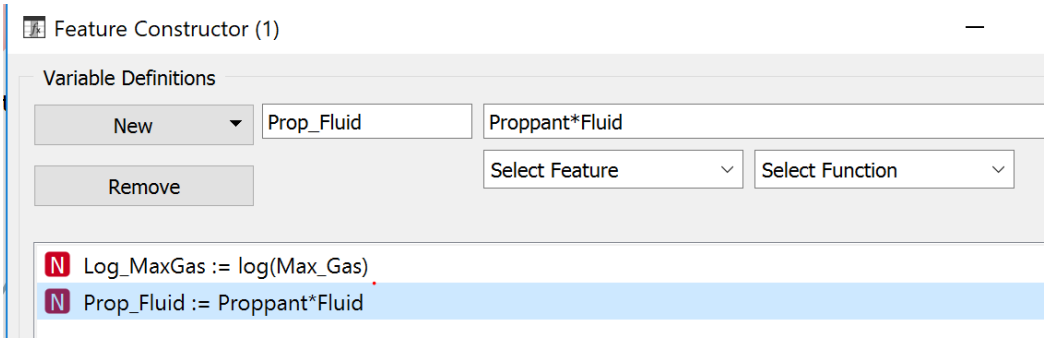
What are your observations about the various model performances?	One case where increasing the do not split improves the model Other is when increasing both values improve the model

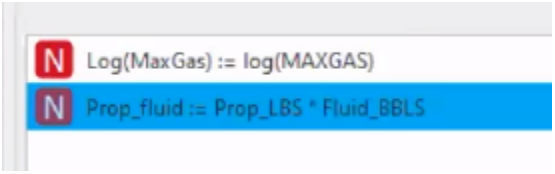
Complete the following table for **Test on test data**

Model	RMSE	MAE	R2
Linear Regression	0.590	0.428	0.655
Tree (30,50)	0.562	0.401	0.687

5. Let's try interaction between the features Proppant and Fluid. Create a construct a new feature as shown below.

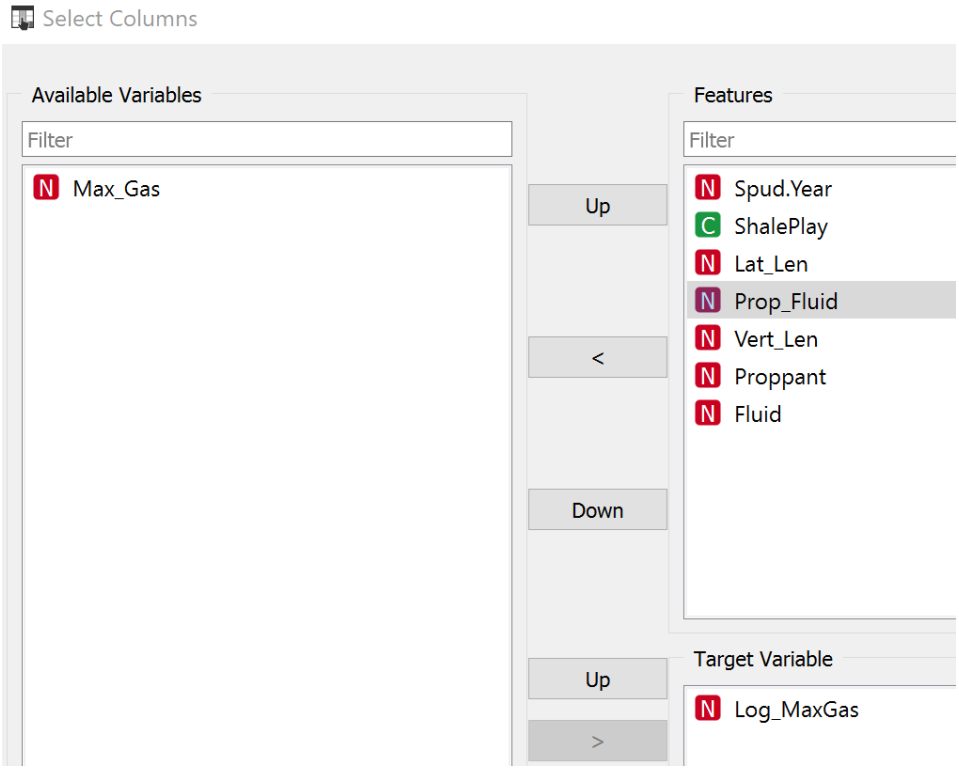
(4 points)





6. Make sure the new feature is added to the Select Columns' Features list.

(6 points)



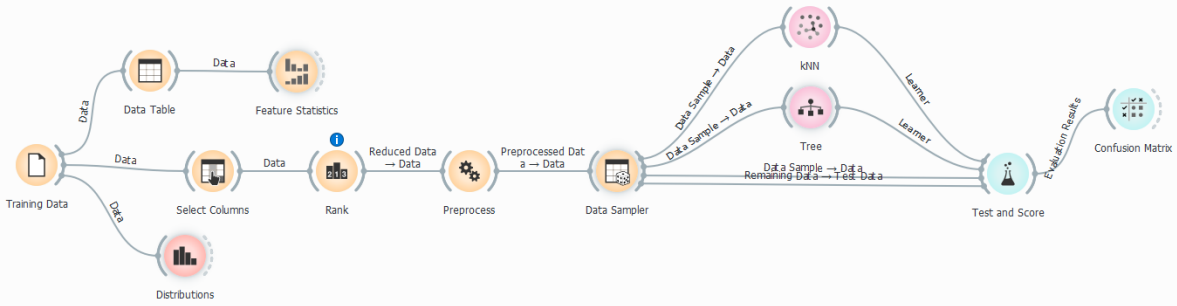
What do you think about the impact of the interaction?	Didn't significantly improved the model performance, but indeed made a big impact that improved

Problem 2/2. [50 points]

Data: For this lab, please download *Train.xlsx* and *Lab4_LogisticReg_Start.ows* files from Canvas to your folder.

Lab Instructions

1. Launch Orange. Open *Lab4_LogisticReg_Start.ows, Train.csv* and verify that you can see the pipeline as shown below: (4 points)

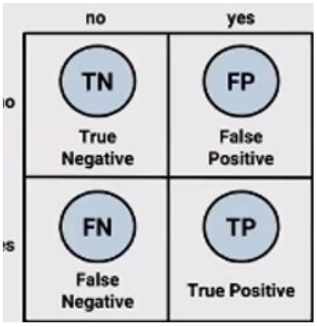


2. Inspect the pipeline and complete the table below: (8 points)

Data Set	Model Parameters	Features Used	CA	F1	Specificity
Training Set **CV-5folds**	kNN Tree	Training Set	0.990 0.990	0.990 0.990	0.959 0.958
Training Set **CV-5folds**	kNN Tree	CV-5folds	0.984 0.986	0.984 0.986	0.937 0.944

3. Open **Confusion Matrix** widget. Examine this widget and complete the table below. Consider **Granitiod** as the positive class: (8 points)

Model Parameters	TP	TN	FP	FN
KNN, 7 Neighbors, Euclidean, Uniform	**21283**	2795	136	114
Tree (Bin Tree- Yes Min Instances: 25 Subsets smaller than: 10 Majority %: Unchecked)	21295	2794	137	102



4. Add a **Logistic Regression** model to the above pipeline.

Complete the table below by considering different C values (Ridge L2) for Logistic Regression (30 points).

Data Set	Model Parameters	Features Used	CA	F1	Specificity (how well negative classes are dealt with)
CV-5folds, Training Set	Logistic Reg Tree kNN	C=1	0.908,0.908 0.986,0.990 0.984,0.990	0.893,0.894 0.986,0.990 0.984,0.990	0.434,0.443 0.944,0.958 0.937,0.959
CV-5folds, Training Set	Logistic Reg Tree kNN	C=100	0.904,0.904 0.986,0.990 0.984,0.990	0.895,0.895 0.986,0.990 0.984,0.990	0.486,0.486 0.944,0.958 0.937,0.959
CV-5folds, Training Set	Logistic Reg Tree kNN	C=10	0.905,0.905 0.986,0.990 0.984,0.990	0.895,0.895 0.986,0.990 0.984,0.990	0.479,0.481 0.944,0.958 0.937,0.959
CV-5folds, Training Set	Logistic Reg Tree kNN	C=50	0.905,0.904 0.986,0.990 0.984,0.990	0.895,0.895 0.986,0.990 0.984,0.990	0.485,0.485 0.944,0.958 0.937,0.959
CV-5folds, Training Set	Logistic Reg Tree kNN	C=0.001	0.880,0.880 0.986,0.990 0.984,0.990	0.823,0.823 0.986,0.990 0.984,0.990	0.120,0.120 0.944,0.958 0.937,0.959

Complete the table below for the best performing model on test data using **Test on test data**:

Data Set	Model Parameters	Features Used	CA	F1	Specificity
Test Data	Logistic Reg	C=50	0.909	0.901	0.502

C=0.001

Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.059	0.359	0.993	0.985	0.985	0.933
Tree	0.105	0.001	0.993	0.987	0.987	0.943
Logistic Regression	0.037	0.002	0.727	0.884	0.829	0.116

C=1

Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.059	0.359	0.993	0.985	0.985	0.933
Tree	0.105	0.001	0.993	0.987	0.987	0.943
Logistic Regression	0.097	0.004	0.931	0.916	0.904	0.465

C=10

Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.059	0.359	0.993	0.985	0.985	0.933
Tree	0.105	0.001	0.993	0.987	0.987	0.943
Logistic Regression	0.094	0.003	0.932	0.909	0.900	0.494

C=50

Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.059	0.359	0.993	0.985	0.985	0.933
Tree	0.105	0.001	0.993	0.987	0.987	0.943
Logistic Regression	0.100	0.003	0.932	0.909	0.901	0.502

C=100

Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.059	0.359	0.993	0.985	0.985	0.933
Tree	0.105	0.001	0.993	0.987	0.987	0.943
Logistic Regression	0.097	0.003	0.932	0.909	0.901	0.502

Confusion Matrix results for the best model on test data, **Granitiod** is the positive class:

TP	TN	FP	FN
5217	312	396	157