Badge-1 (Kaggle Competition)

Total Points: 100

Out date: Jun 29, 2022
Due date: Jun 29, 2022 by 9:00 PM

Submission

1. Prepare your solution in Orange and save the workspace (e.g., Lab5_LastName.ows).
2. Complete the table given below and save the file (e.g., Lab5_LastName.docx)
3. Upload the csv file to Kaggle.
4. Upload the ows file to the Canvas.

Background information: Oil and gas reservoirs lie deep beneath the Earth's surface. Geologists and engineers cannot examine the rock formations in situ, so tools called sondes go there for them. Specialists lower these tools into a wellbore and obtain measurements of subsurface properties. The data are displayed as a series of measurements covering a depth range in a display called a well log. Often, several tools are run simultaneously as a logging string, and the combination of results is more informative than each individual measurement (https://www.slb.com/resource-library/oilfield-review/defining-series/defining-logging).

Objective: To train a Machine Learning model to predict lithology using log measurements from two different geothermal wells. Use kNN algorithm in Orange to train your model. You will participate in a Kaggle competition which will run from 18:30HRS to 20:00HRS on 19Nov as part of the class.

Data: Excel files (*Train.xlsx and Test.xlsx*) and *Badge1_Kagglestart.ows* which can be downloaded from Canvas.
Data Source: https://gdr.openei.org/submissions/1111

Lab Instructions- Problem-1

1. Launch Orange. Open *Badge1_Kagglestart.ows* and verify that you can see the two pipelines as shown below:

2. Load the *Train.xlsx* file using the **Training Data (File)** widget. Complete the following table for the Concatenated dataset (12 points):

How many features?	5 if Lithology is the target
How many samples?	30410
Which attributes have missing values and what is the % of missing values?	DENS (6%), GR (0%)
What are the distribution types and ranges for GR, SP, DENS and DEEP_RES features?	Numeric

3. Inspect the rest of the top pipeline and complete the table below: (3 points)

Data Set	Model Parameters	AUC	CA	F1	Specificity
(Cross Validation) CV-5	Lithology	0.988	0.986	0.986	0.946
Test Set (test on train data)	Lithology	1.00	0.994	0.994	0.978

4. Add a **Confusion Matrix** widget to the **Predictions** widget. Examine this widget and complete the table below. Use **test data** (remaining data) and consider **Granitiod** as the positive class:
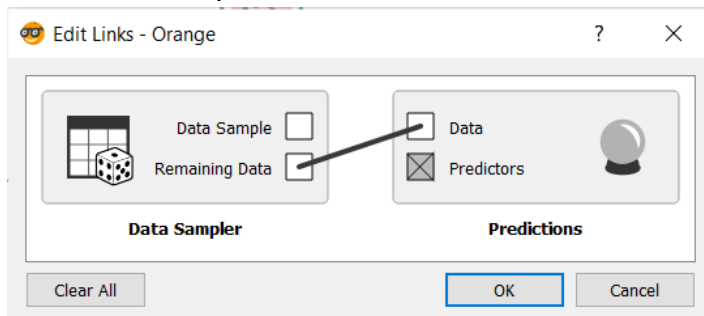
(5 points)
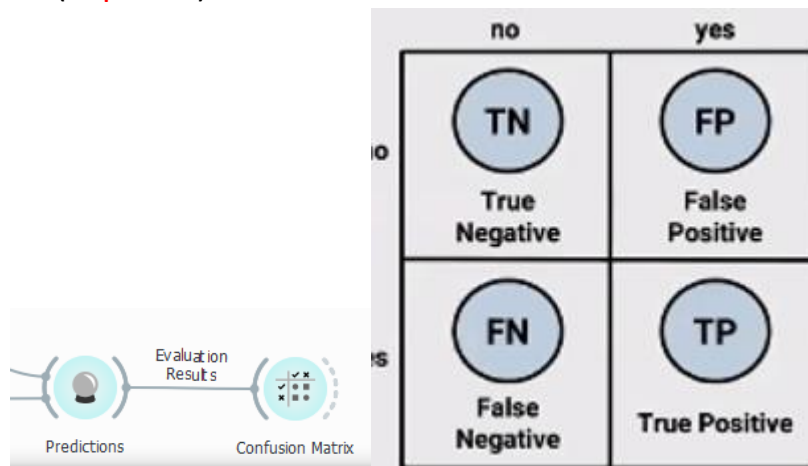
TP	TN	FP	FN
21272	2836	95	125
5337	673	35	37
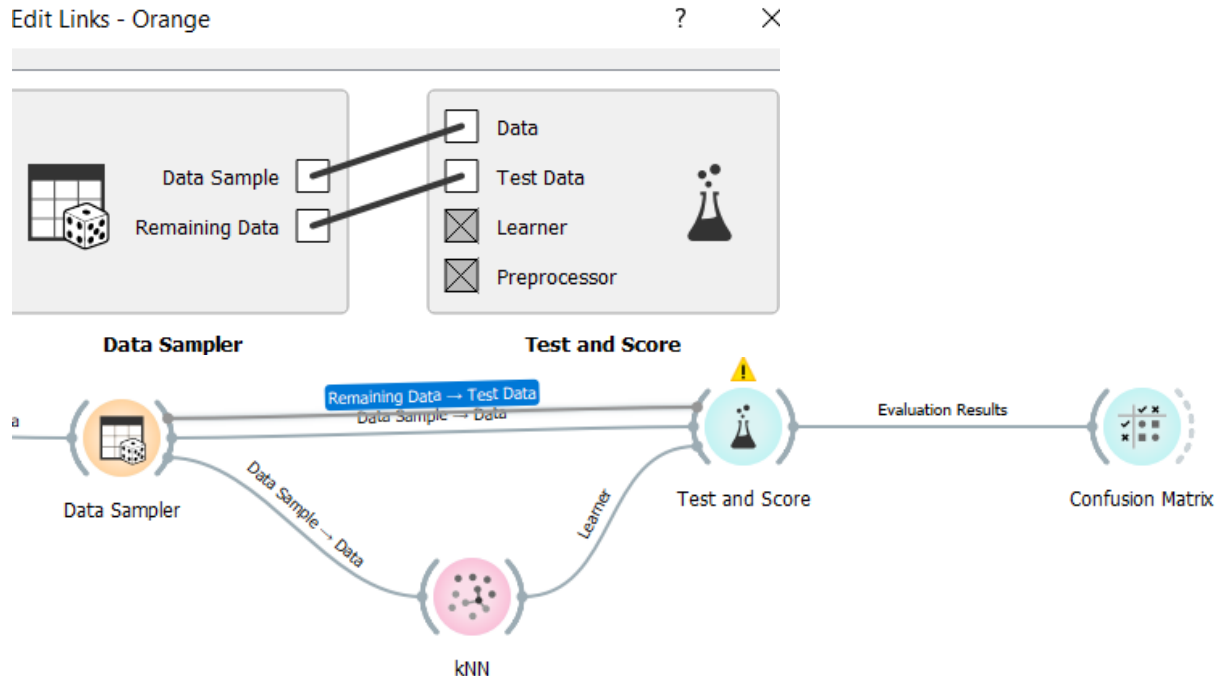
Lab Instructions - Problem-2 (Kaggle Competition)

Your customer is asking you to use Machine Learning to predict Lithology in a new data set that contains some log measurements and is in the _Test.xlsx_ file. As a data scientist, your job is to train machine learning models using the _Train.xlsx_ file provided to you using **Orange,** deploy your **best performing model** to predict **Lithlogy** for the samples in the _Test.xlsx file_ and submit the predictions to your client via Kaggle.

Please use the following Kaggle link for downloading the datasets (Train.xlsx and Test.xlsx) and uploading your predictions:
https://www.kaggle.com/t/595395dfbdec480682aba2351494cd8f

Specifically,

1. You will be using kNN algorithm to train your Machine Learning models on the _Train.xlsx_ dataset.
   You are encouraged to consider the following aspects:
   a. Preprocessing features (Imputing missing values, Standardizing or Normalizing features)

b. Selection of features
c. Dimensionality Reduction using PCA
d. Parameters to consider for tuning:
    i. Sample size (e.g. 70%, 80%, 90%)
    ii. Value of k
    iii. Euclidean or Manhattan distance
    iv. Uniform or Distance weighting
e. Use Predictions and Confusion Matrix widget to evaluate the performance of your various models on the training and test set.

Complete the table below: (20 points)

What preprocessing options did you consider?	Normalize (to interval [0,1], Impute Missing Values (Average/Most Frequent)
Did you explore PCA for dimensionality reduction and evaluate its impact on model performance?	Yes, seems 5 in Preprocess is the sweet spot but did not make the model perform any better based on below
If you used dimensionality reduction using PCA for your final model, provide details on how many components you considered and the % variance explained:	I messed with 2 components and 65% variance when testing, but not much further due to time in competition
What is your best performing model and model parameters?	kNN 8, Manhattan, Uniform

Complete the table below based on model performance of your training data and CV-5. You should provide at least 5 different models.
(25 points)

Model	AUC	CA	F1	Specificity
M1: kNN (3, Euc, Uniform), Example model	0.988	.986	.986	0.946
M2: kNN (8, Man, Uni)	0.999	0.988	0.988	0.956

M3: kNN (8, Euc, Uni)	0.993	0.983	0.983	0.941
M4: kNN (8, Euc, Dist)	0.993	0.987	0.987	0.939
M5: kNN (8, Man, Dist)	0.999	0.990	0.990	0.951

2. Open the **Unlabelled Data(File)** widget and load <mark>the</mark> *topredict_facies.xlsx* file.
(5 points)

How many samples are there in this data set? What is the % of missing value?	3378 1.1%

3. Connect the **kNN** model widget to **Predictions(1)** widget as shown below:
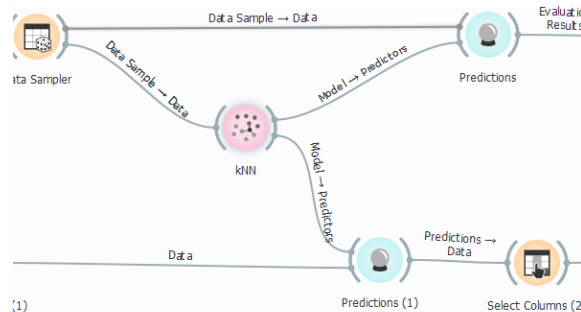
4. Verify the rest of this pipeline and ensure **Predictions** widget contains the predicted **Lithology** values for all the samples.

5. Use **Select Columns(2)** widget shown in the above below to select only the predicted values and verify using the **Data Table** widget that you have only one column of predicted **Lithology** values.
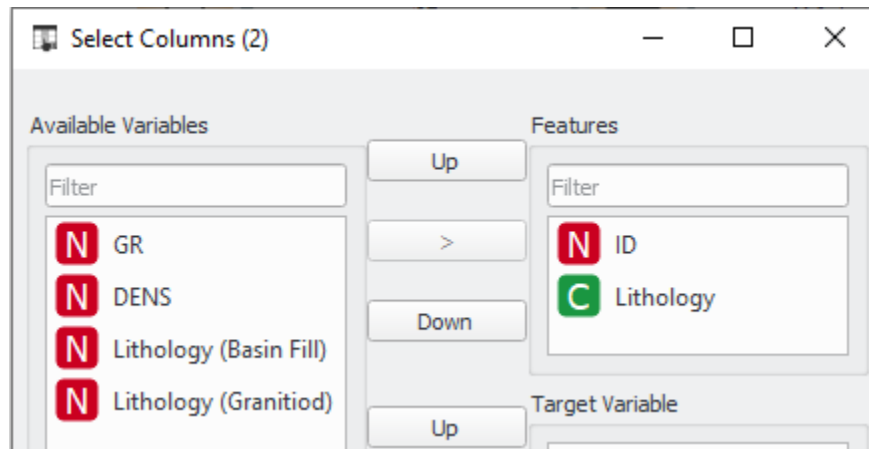
6. Use **Save Data** widget to save your predictions as *Badge1_Kaggle_Predictions_LastName.csv.* Uncheck **Add type annotations to header** and save the .csv file to your local folder.

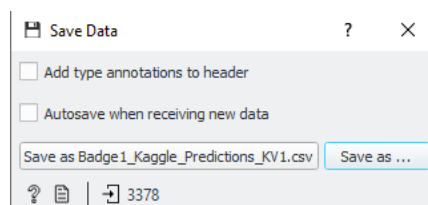7. Predictions csv file should output results in two columns: 1) ID (in text format only), and 2) the name of the target variable (i.e Lithology).

ID in the numeric format　　　　　　　ID in the text format

8. Upload the csv file to Kaggle. Make at least 5 submissions for full credit. Maximum allowable submission is **10**.　　　　　　　　(15 points)

https://www.kaggle.com/t/595395dfbdec480682aba2351494cd8f

9. Upload your final .ows to Canvas. (15 points)

Look at 1:43:00

Select A column (ID), right click to format, change to text

YOUR RECENT SUBMISSION

Badge1_Kaggle_Predictions_Rattan - best.csv
Submitted by Riyan Rattan · Submitted just now

FAILED
Submission Error

Evaluation Exception: Submission must have 3378 rows