Final Project Proposal
CSCI303 Intro to Data Science

Xiaoyu Zhu, Haoxuan Yang

April 2018

1 Summary

This project will use traditional machine learning algorithms to predict the depth of hydrocarbons from the well log data in the Billie Yates portion of the Boonesville Field in Texas. (Latitude: 33-02'48"N, Longitude: 097-49'05"W)

2 Dataset Description

Well logging operations is a geophysical method which provide measurements of borehole and formation properaties at measured depth (Atlas ((2002)). This dataset comes from the **GPGN268** (Geophysical Data Analysis) class, which is used for practising predicting hydrocarbon.

There are five wells in this field: **B.Yates 7 (BY7),BY11, BY13, BY15, BY18D**. Each well log data would include the following features: gamma ray, neutron porosity, formation density, spontaneous potential and resistivity. The primary characteristics that indicate the location of hydrocarbon reservoirs include abnormally high resistivity, a decrease in spontaneous potential, low gamma ray readings, and low formation density.

2.1 Attribute Description

Listing of attributes:

- **Depth**: continuous value
- **Gamma Ray**: continuous value
  Reflect the concentration of radioactive elements in the rock, radioactive material is concentrated in in shaley rocks while most sandstones (could reserve oil) are very weakly radioactive.
- **Neutron Porosity**: continuous value
  Measure the hydrogen ion concentration in a formation, a lower neutron log reading indicates abundant formation hydrogen.
- **Formation Density**: continuous value
  Measure the density (grams/cm3) of the formation based on the density of electrons in the formation.
- **Spontaneous Potential**: continuous value
  Reflecting higher permeability in sandstone (could reserve oil) relative to lower permeability in shale.

- **Resistivity**: continuous value
  Hydrocarbons, together with rock, and fresh water are all very highly resistive to electric. current flow.

3 Proposed Work

1. Label the well log **BY18D** based on information from *Intro to Well Log.* Add an output of "yes" or "no" to the end of each row, indicating if there is oil at that depth.
2. Training data with the supervised learning model. Our initial options are random forest, GBDT, Xgboost and k-Nearest Neighbors. We will finally pick the best model based on their performance.
3. The cross validation will be used on our log data of the well to evaluate the model.
4. Since this is a classification problem, we will be using the following evaluation methods: AUC, accuracy and F1-score.
5. After choosing the best model, we will make the prediction and try to output the result into probability space.

4 Challenges

There are several challenges in this classification task:

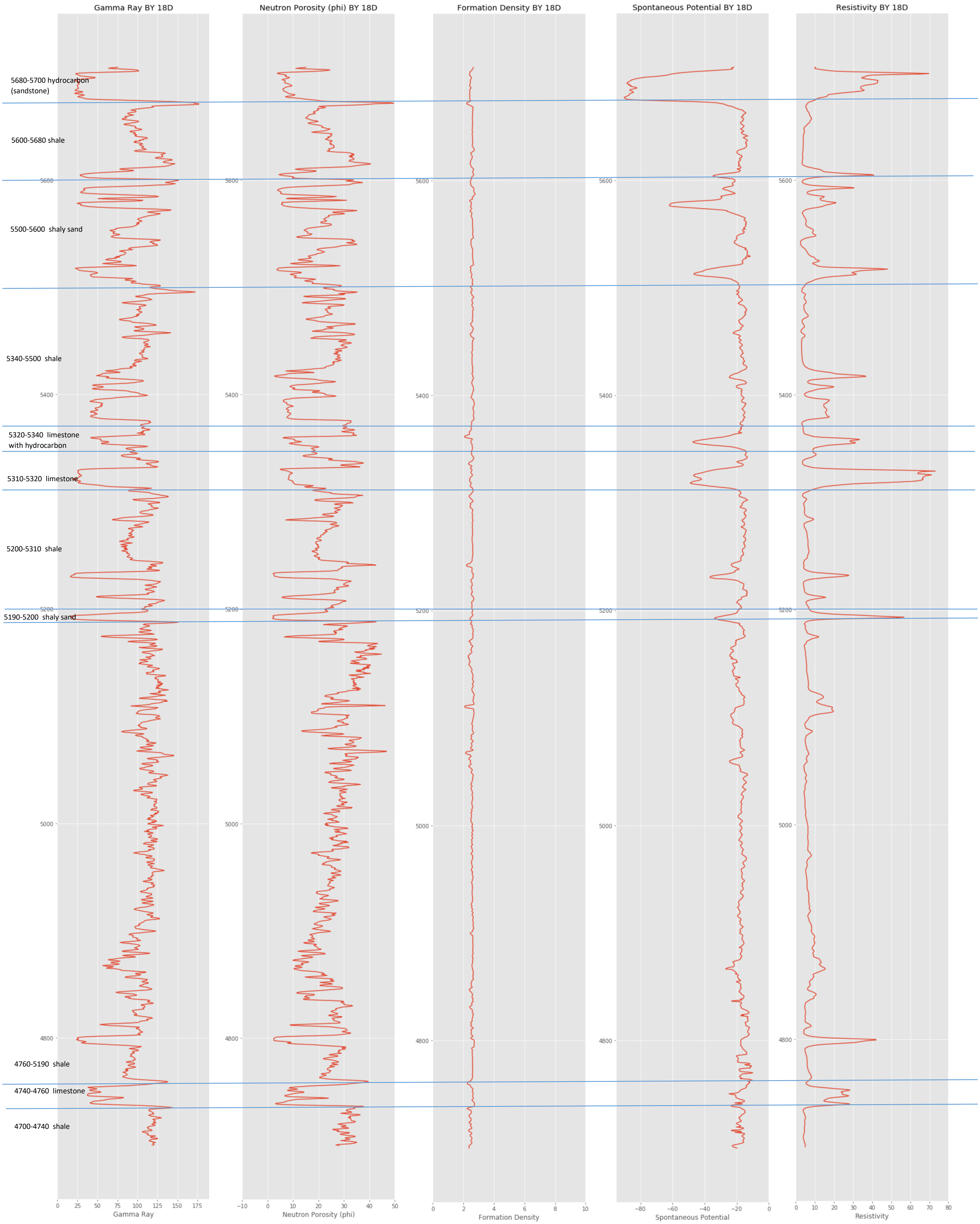
Since labeling would have a heavy influence on training result, it needs prior geophysics knowledge to manually label the data based on the tool book. If the labeling task is not good enough, the model will not learn enough from the data.

In order to avoid overfitting problem, we need to fine tune the parameters in the model and use cross validation to test our model.

Moreover, our attributes are all continuous variables which will need the regression CART tree when using a decision tree based model. In this case, the output "Yes" or "No" will be encoded using a binary variable.

5 Initial Results

We have cleaned the data: removed the outliers and plot the features at different depth in Well BY18D (Page 3).



Gamma Ray BY 18D
Neutron Porosity (phi) BY 18D
Formation Density BY 18D
Spontaneous Potential BY 18D
Resistivity BY 18D
5680-5700 hydrocarbon
(sandstone)
5600-5680 shale
5500-5600 shaly sand
5340-5500 shale
5320-5340 limestone
with hydrocarbon
5310-5320 limestone
5200-5310 shale
5190-5200 shaly sand
4760-5190 shale
4740-4760 limestone
4700-4740 shale
5600
5400
5200
5000
4800
0
25
50
75
100
125
150
175
Gamma Ray
−10
0
10
20
30
40
50
Neutron Porosity (phi)
0
2
4
6
8
10
Formation Density
−80
−60
−40
−20
0
Spontaneous Potential
0
10
20
30
40
50
60
70
80
Resistivity

References

B. Atlas. *Introduction to Wire Log Analysis.* Baker Hughes Inc., 2002.