Well log data prediction with classification algorithms

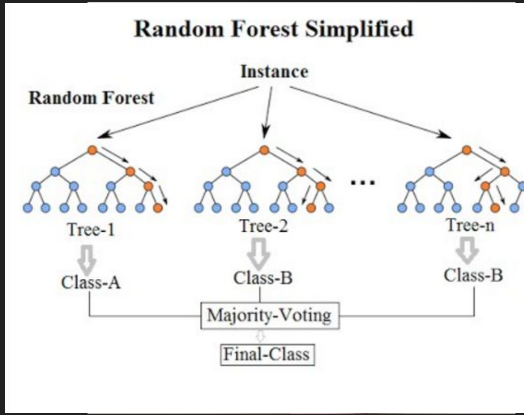
Haoxuan Yang, Rosie Zhu

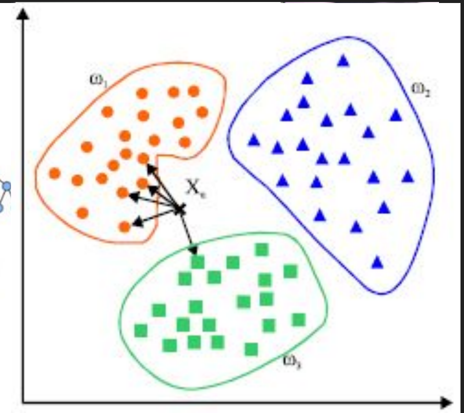
4/26/2018

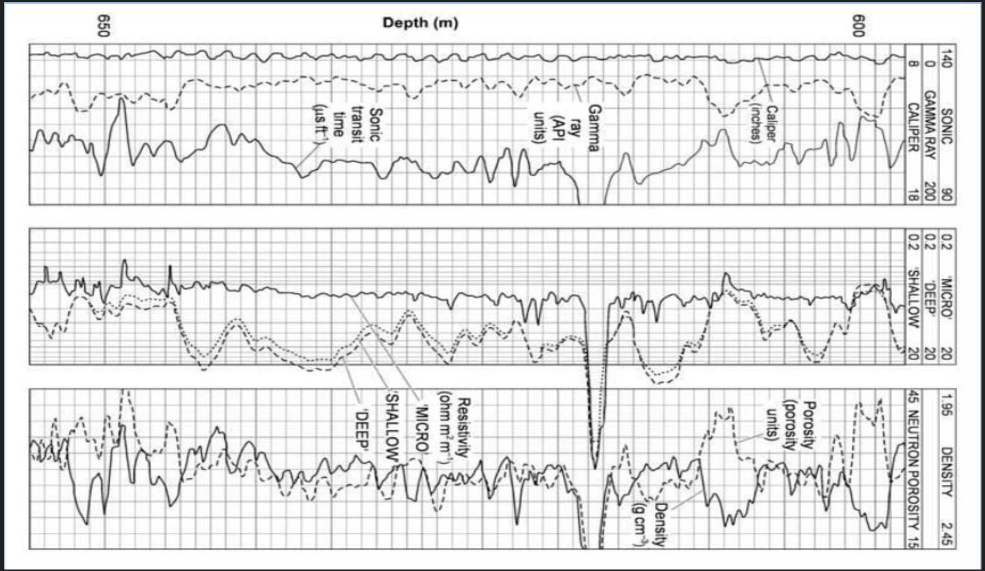
Summary

This project will use classification algorithms (Random Forest and Knn) to predict the depth of natural gas from the well log data in the Boonesville Field in Texas. It returns the most promising depth/layers the geophysicists should later focus on.

We also analyse which feature is more important for well log data using random forest.







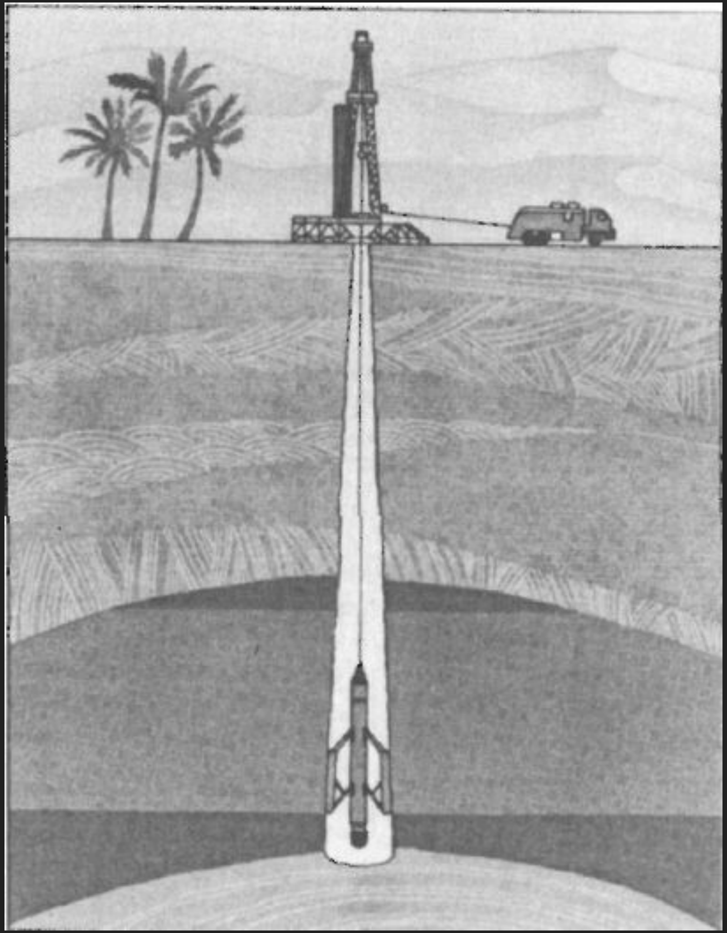
Contents

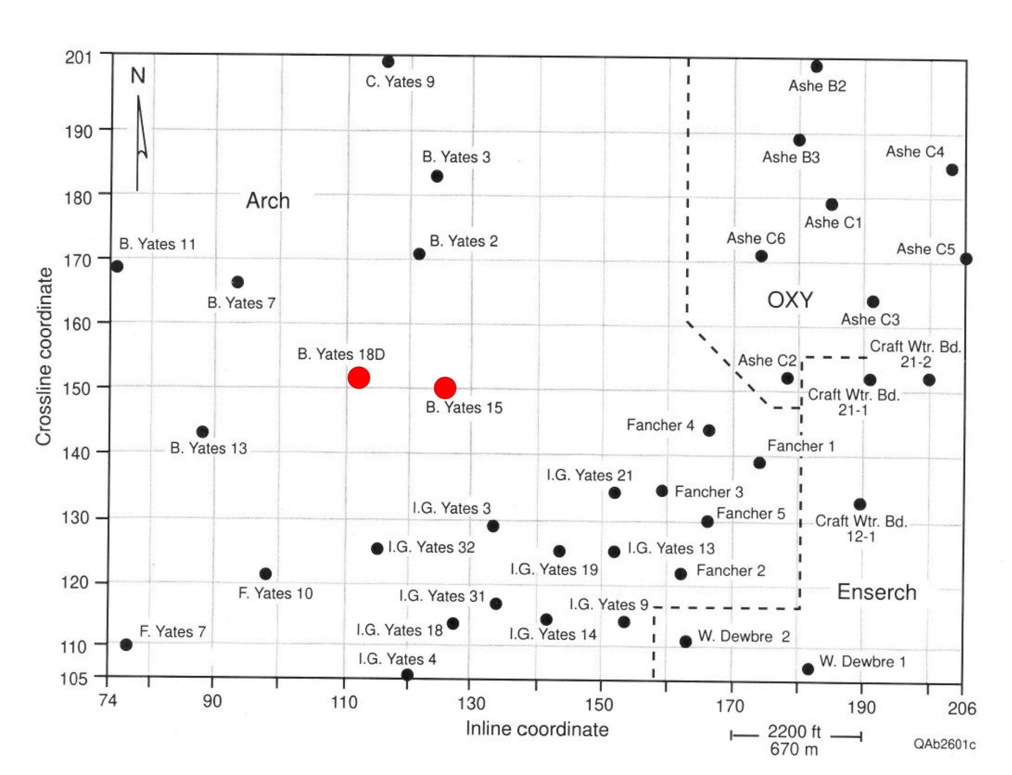
- Data Description
- Data Processing
- Prediction and Evaluation
- Conclusion

Data Description

Well logging operations is a geophysical method which provide measurements of borehole and formation properties at measured depth [1].

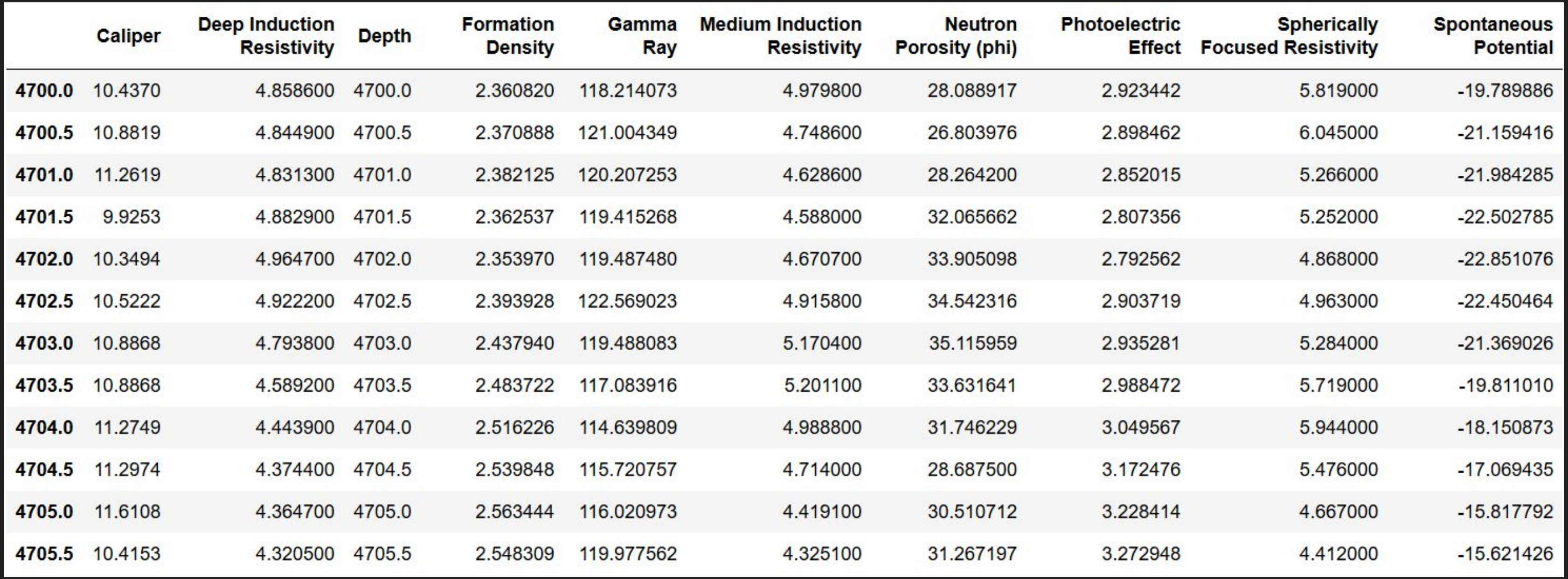
This dataset comes from the Midcontinent sandstone natural gas reservoir in Boonsville Field, from the Secondary natural gas recovery (SGR) program by U.S.department of Energy and the Gas Research Institute. (Latitude: 33-02'48"N, Longitude: 097-49'05"W)





Well locations

Two wells in this field: B.Yates 15 (BY15) and BY 18, are chosen to be assessed in this project. Locations of the wells are shown in the figure on the left.



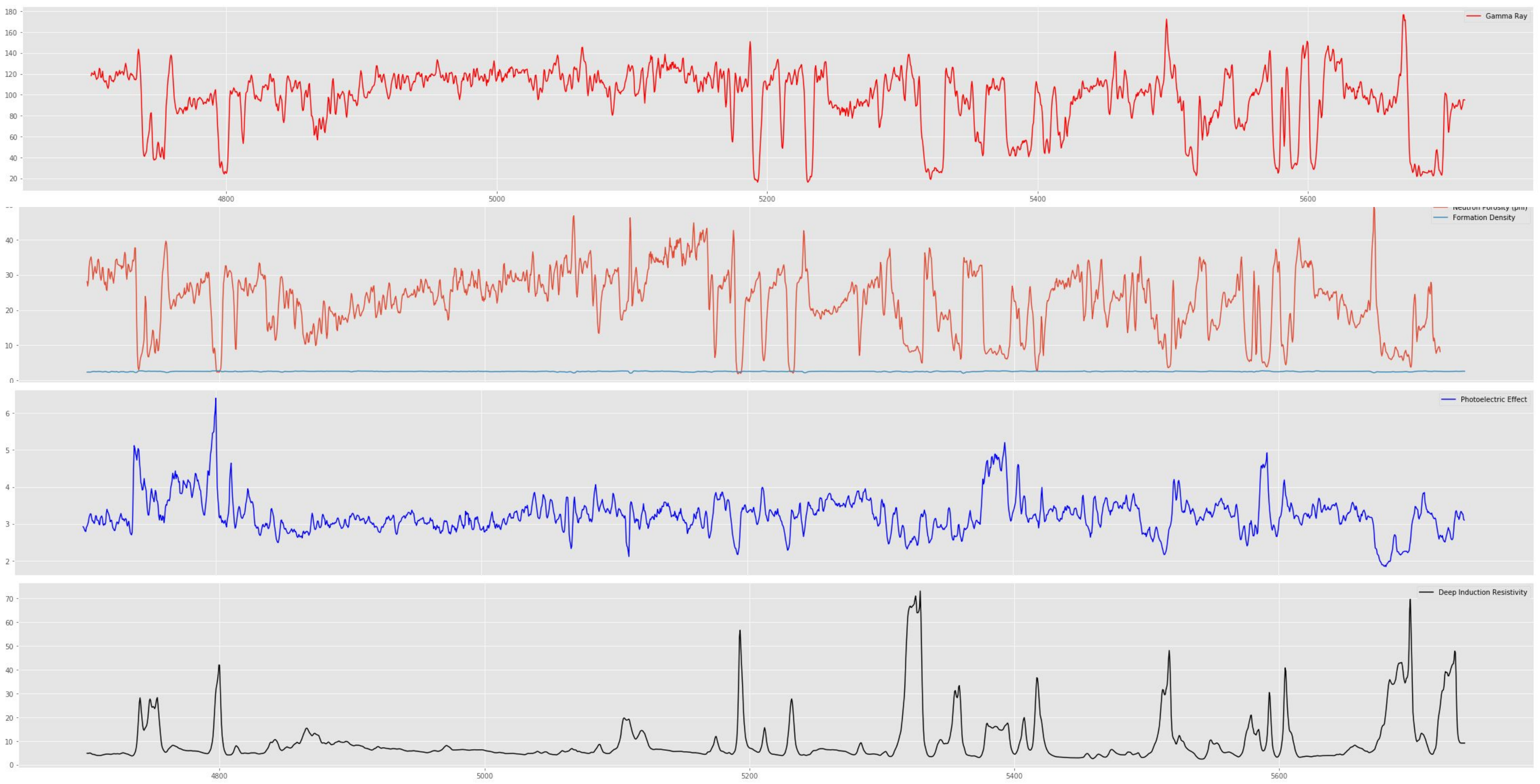
	Caliper	Deep Induction Resistivity	Depth	Formation Density	Gamma Ray	Medium Induction Resistivity	Neutron Porosity (phi)	Photoelectric Effect	Spherically Focused Resistivity	Spontaneous Potential
4700.0	10.4370	4.858600	4700.0	2.360820	118.214073	4.979800	28.088917	2.923442	5.819000	-19.789886
4700.5	10.8819	4.844900	4700.5	2.370888	121.004349	4.748600	26.803976	2.898462	6.045000	-21.159416
4701.0	11.2619	4.831300	4701.0	2.382125	120.207253	4.628600	28.264200	2.852015	5.266000	-21.984285
4701.5	9.9253	4.882900	4701.5	2.362537	119.415268	4.588000	32.065662	2.807356	5.252000	-22.502785
4702.0	10.3494	4.964700	4702.0	2.353970	119.487480	4.670700	33.905098	2.792562	4.868000	-22.851076
4702.5	10.5222	4.922200	4702.5	2.393928	122.569023	4.915800	34.542316	2.903719	4.963000	-22.450464
4703.0	10.8868	4.793800	4703.0	2.437940	119.488083	5.170400	35.115959	2.935281	5.284000	-21.369026
4703.5	10.8868	4.589200	4703.5	2.483722	117.083916	5.201100	33.631641	2.988472	5.719000	-19.811010
4704.0	11.2749	4.443900	4704.0	2.516226	114.639809	4.988800	31.746229	3.049567	5.944000	-18.150873
4704.5	11.2974	4.374400	4704.5	2.539848	115.720757	4.714000	28.687500	3.172476	5.476000	-17.069435
4705.0	11.6108	4.364700	4705.0	2.563444	116.020973	4.419100	30.510712	3.228414	4.667000	-15.817792
4705.5	10.4153	4.320500	4705.5	2.548309	119.977562	4.325100	31.267197	3.272948	4.412000	-15.621426

The number of observations is about 2000: from about 4700 ft to 6000 ft underground, at depth increments of 0.5 ft.

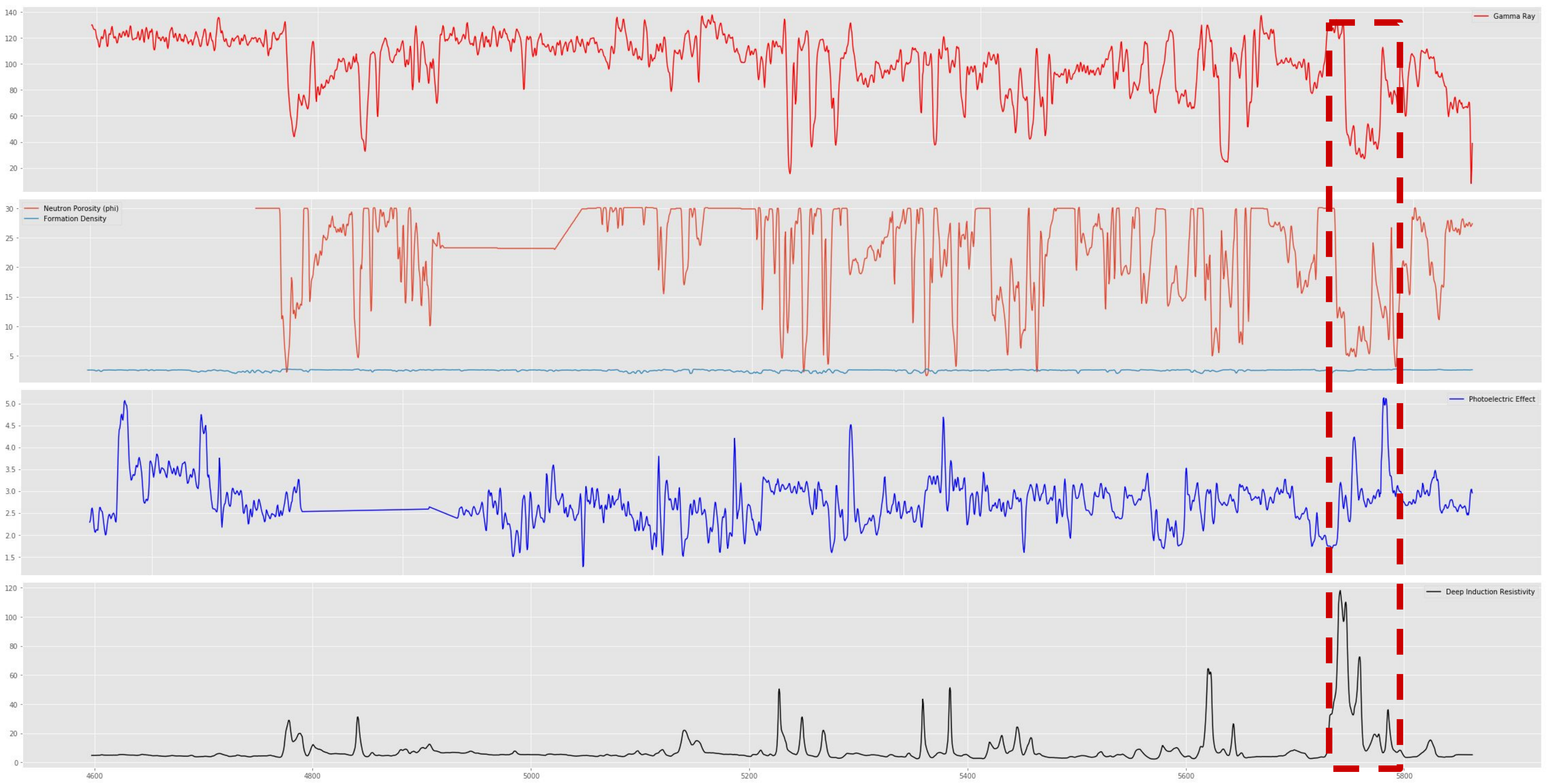
The 9 features of log data are: Caliper data, Gamma ray data, Deep Induction Resistivity, Medium Induction Resistivity, Neutron Porosity, Formation Density, Photoelectric Effect, Spherically Focused Resistivity and Spontaneous Potential.

Characteristics of hydrocarbon may include abnormally high resistivity,etc.

Data Processing



Well log data at B Yates 18D (5 of the 9 features)



Well log data at B Yates 15 (5 of the 9 features)

Label targets:

According to a document for the dataset: *Secondary natural gas recovery : targeted applications for infield reserve growth in midcontinent reservoirs, Boonsville Field, Fort Worth Basin, Texas*, layers may reserve gas are defined as a result of the 3-D seismic survey, geologic evaluation and well log interpretation.

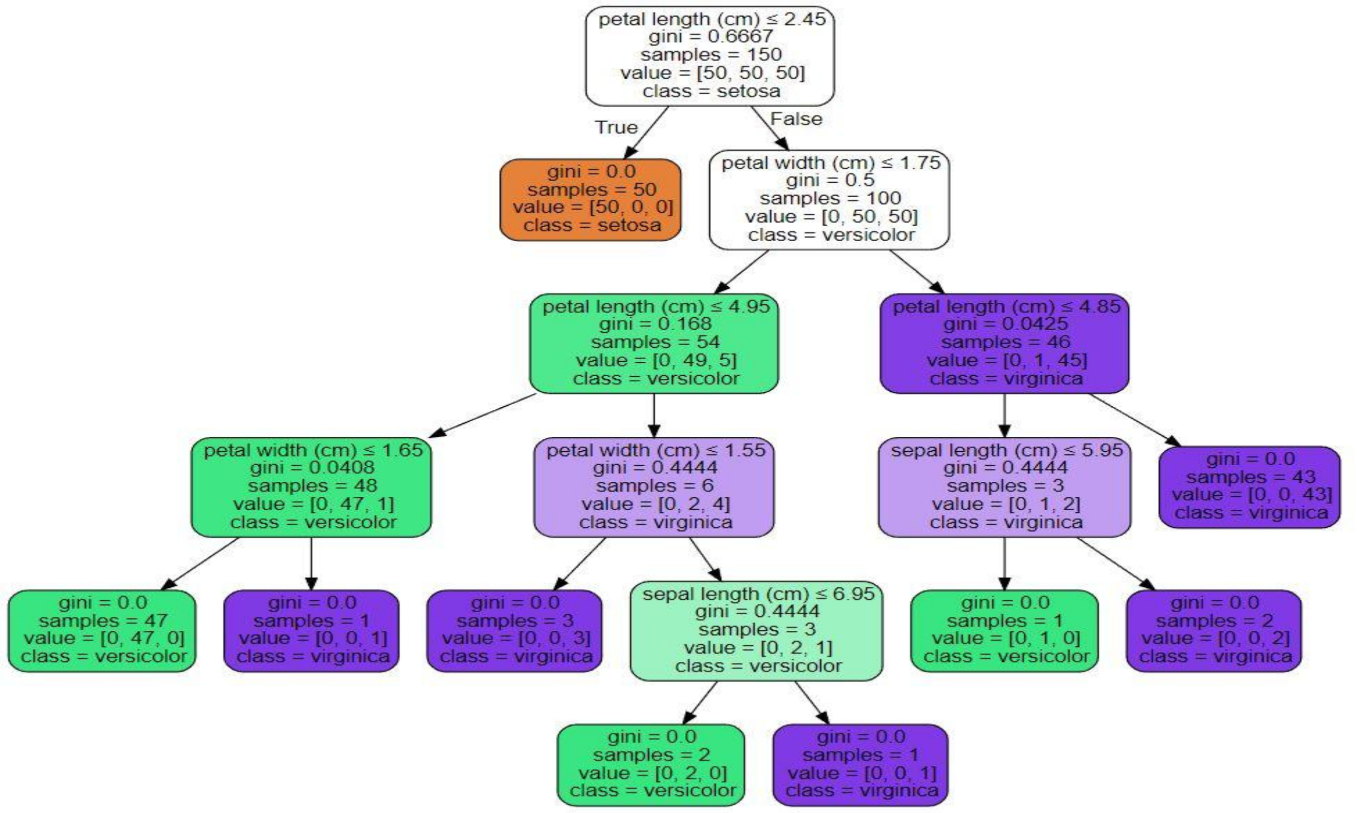
We created a new attribute contains the targets in the dataset, after each row (each depth) of data, manually label ‘Yes’ (indicating there is gas as predicted in the document) and ‘No’ (there is not gas).

We added six period of ‘Yes’ to each well, for example, at BY 18D:

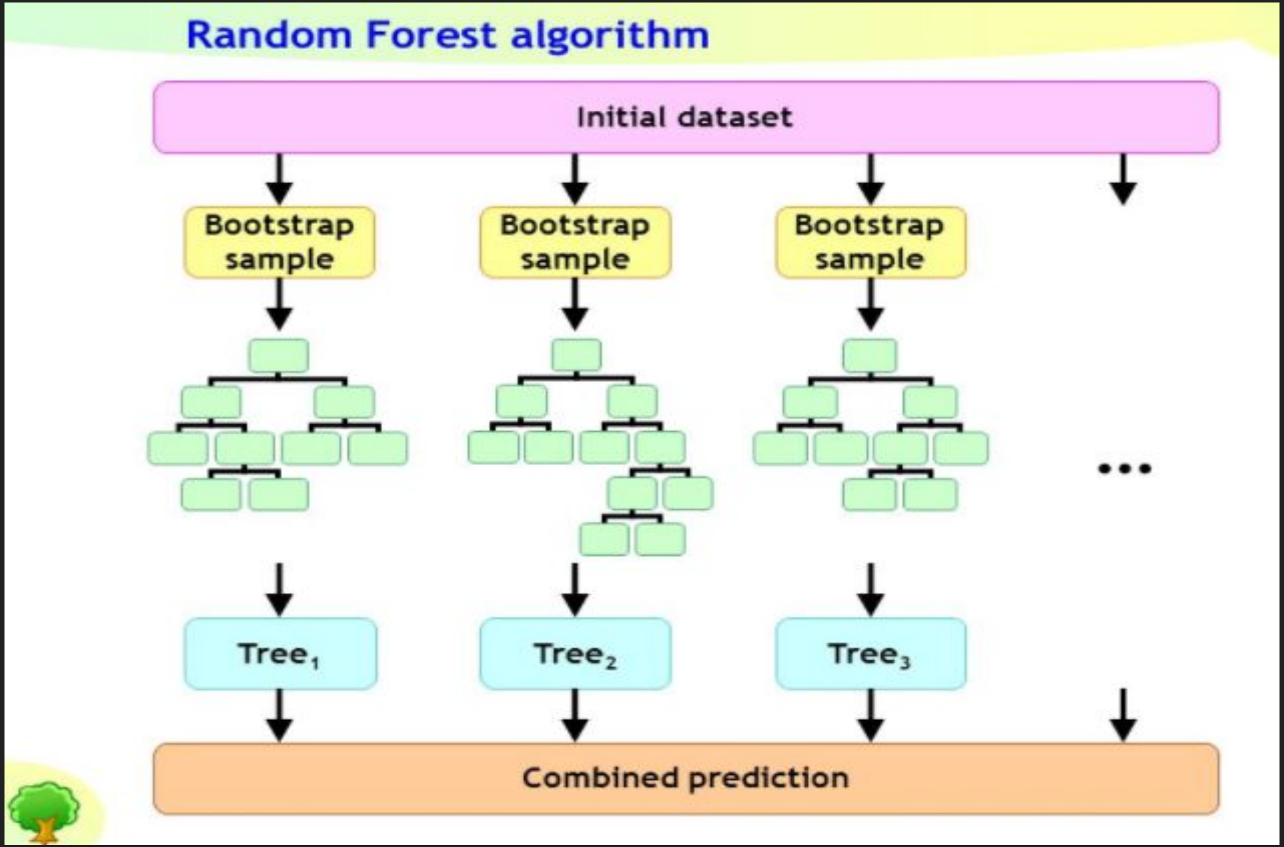
We labeled ‘Yes’ to depth: 5670 to 5740 ft, 5508 to 5520 ft, 5571 to 5582 ft, 5600 to 5612 ft, 5310 to 5332 ft and 5189 to 5198 ft.

Prediction and Evaluation

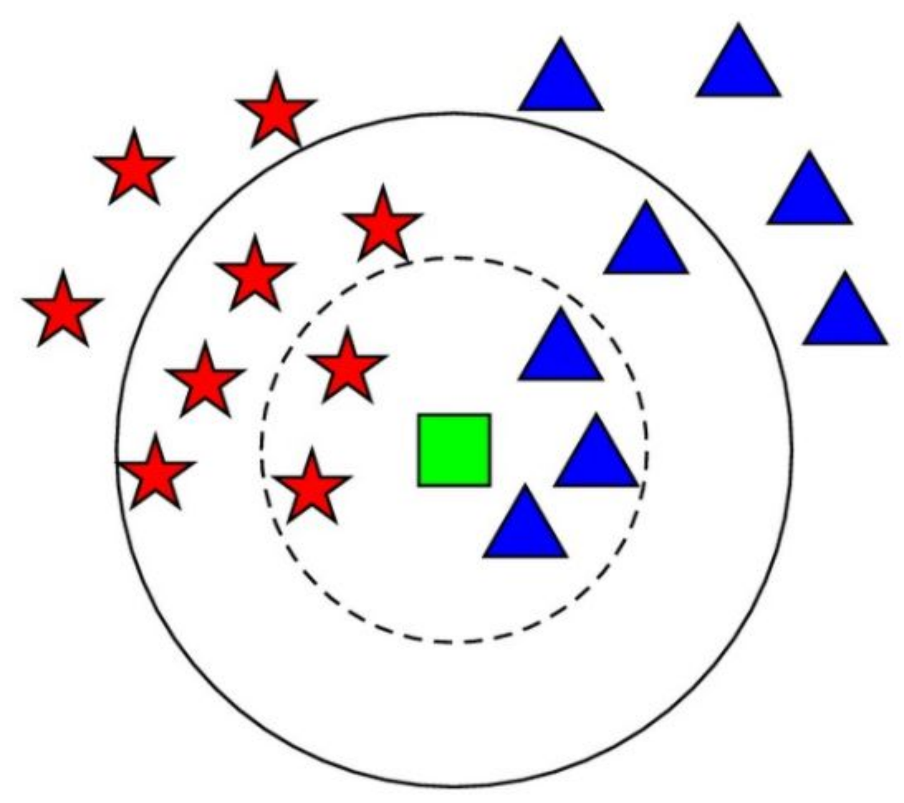
Decision Tree



Random Forest



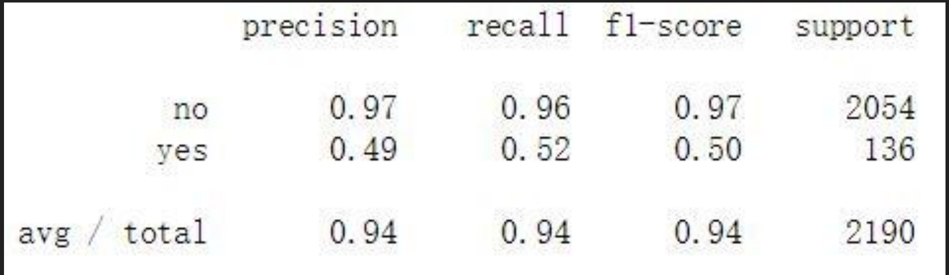
KNN



Data We Have

	no	yes
BY18D (training)	1808	25
BY15 (testing)	2044	146

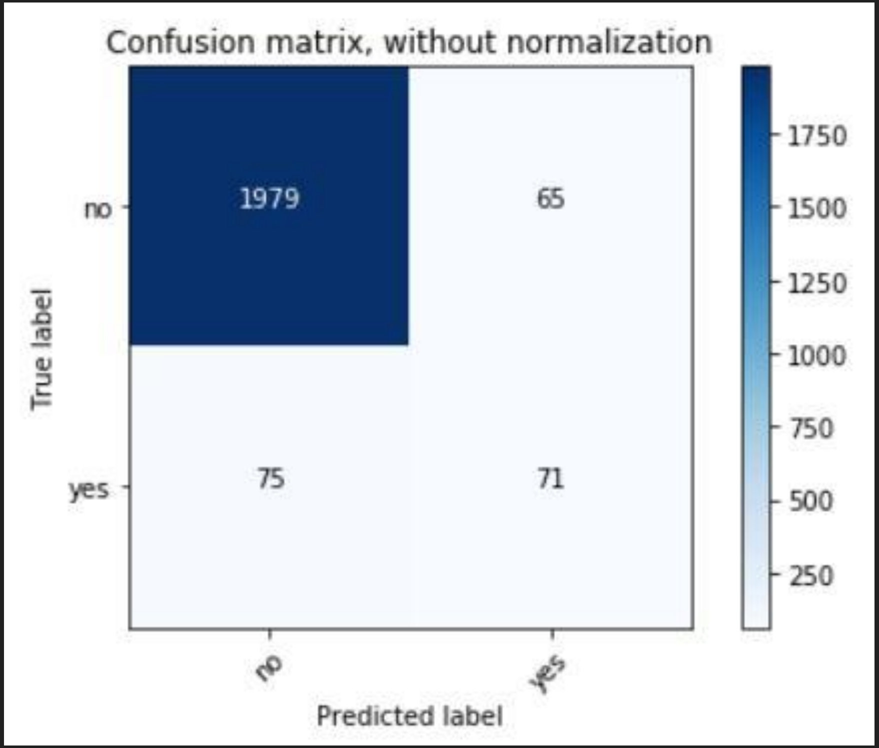
Random Forest Report



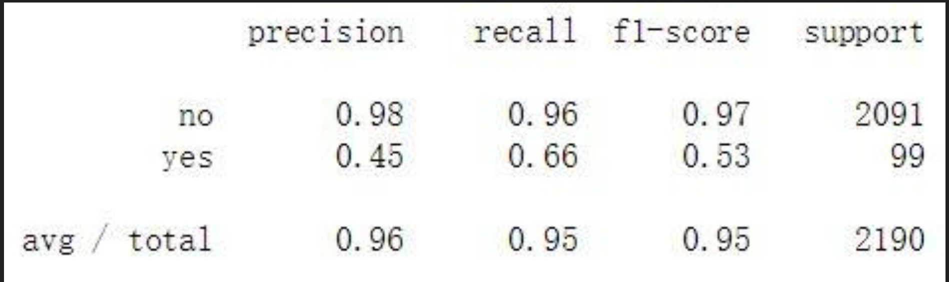
```
             precision    recall  f1-score   support

         no       0.97      0.96      0.97      2054
        yes       0.49      0.52      0.50       136

avg / total       0.94      0.94      0.94      2190
```



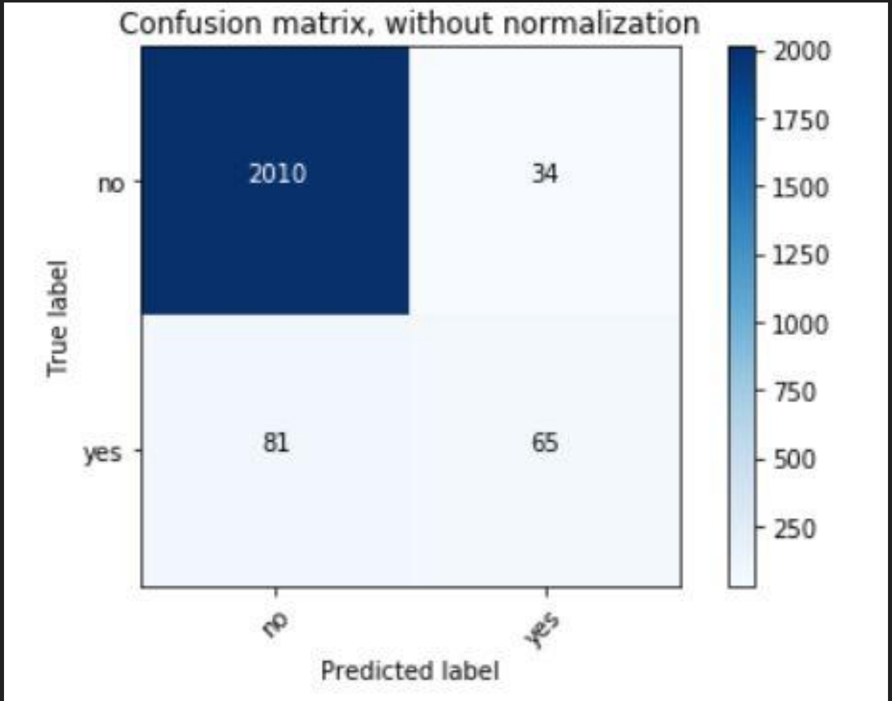
KNN Report



```
             precision    recall  f1-score   support

         no       0.98      0.96      0.97      2091
        yes       0.45      0.66      0.53        99

avg / total       0.96      0.95      0.95      2190
```



Why so BAD?

- This is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.

- The predictive model developed using conventional machine learning algorithms could be biased and inaccurate

Random Over Sampling Gives a Solution!

- Random Over Sampling approach increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

- **BUT**, it has a disadvantage:
    - It increases the chance of overfitting since it simply replicates the minority class data

Augmented Random Over Sampling

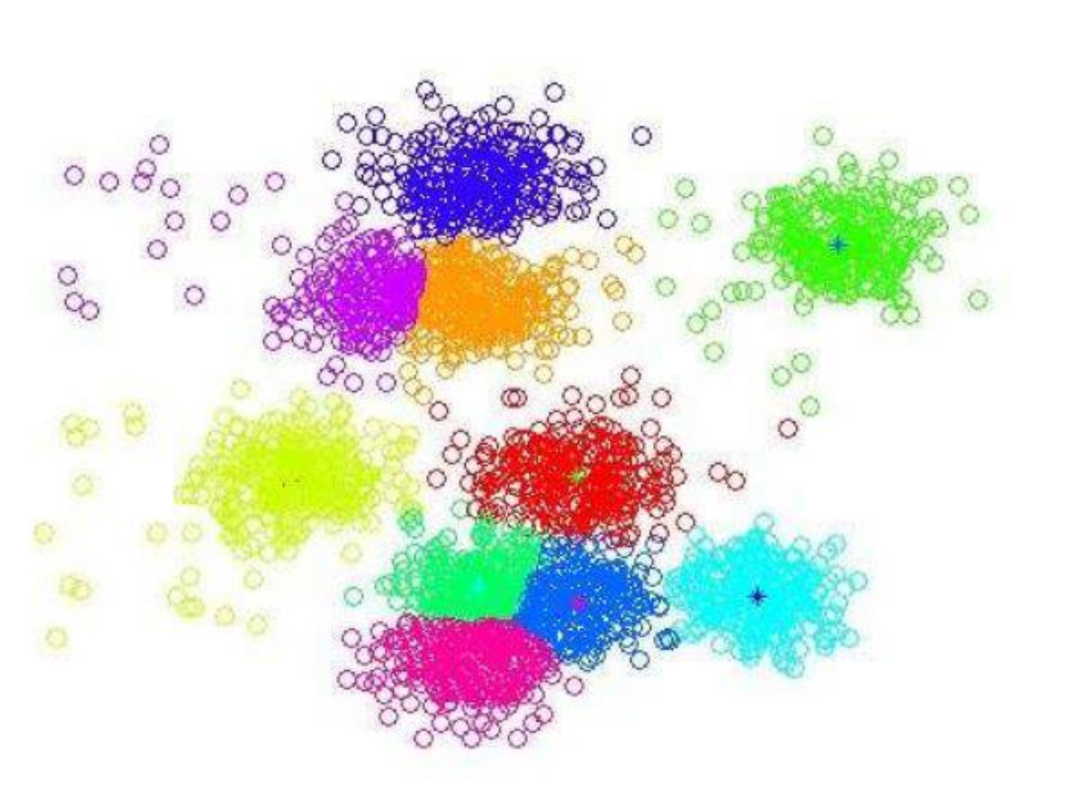
- To avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models. We call this “Augmented Random Over Sampling”.

K-means Clustering

Group the data with label “yes” into subsets which is the minority class data.

Each centroid of a cluster represent a subset of the data.

Then create some fake data around the centroid based on **Gaussian distribution** (I assume...).



Fake Data?

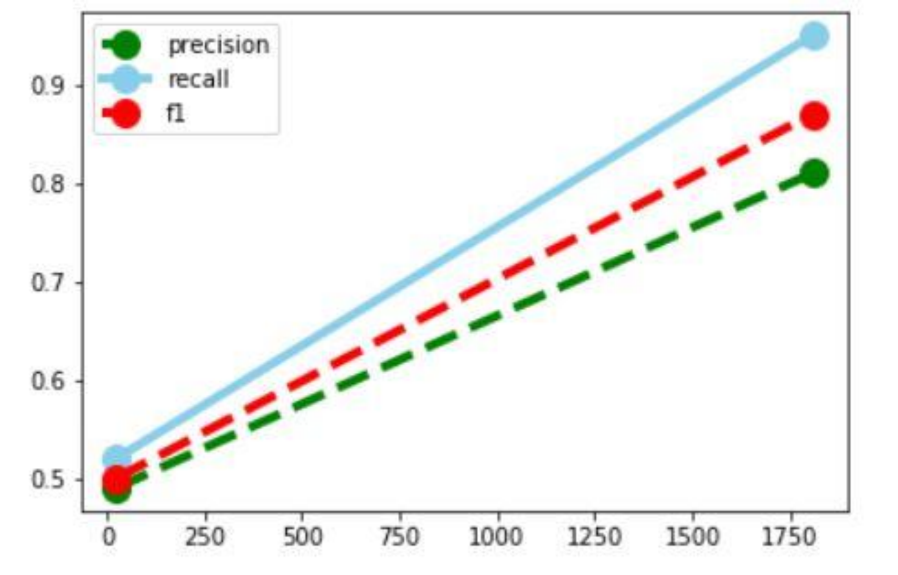
	no	yes
BY18D (training)	1808	25
BY15 (testing)	2044	146

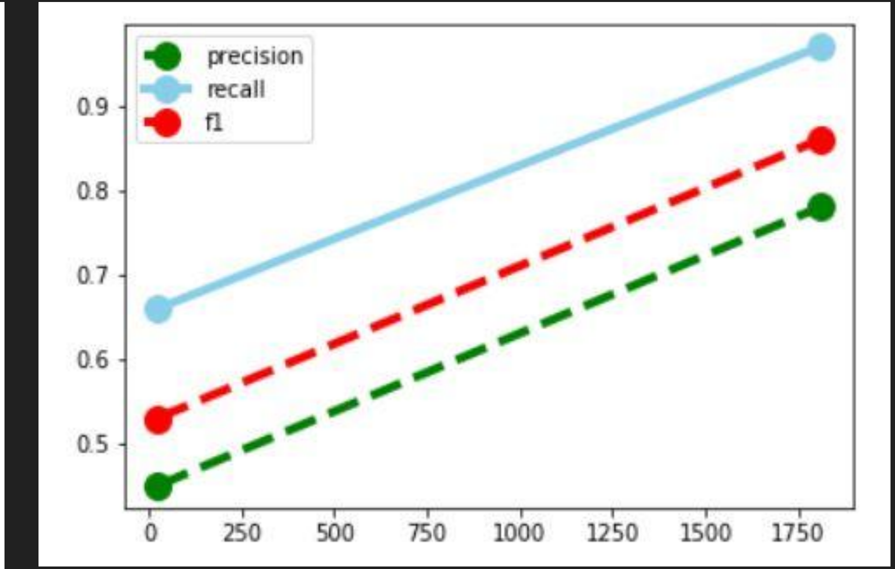
	no	yes
BY18D (training)	1808	1808
BY15 (testing)	2044	2044

Let’s see how they perform!

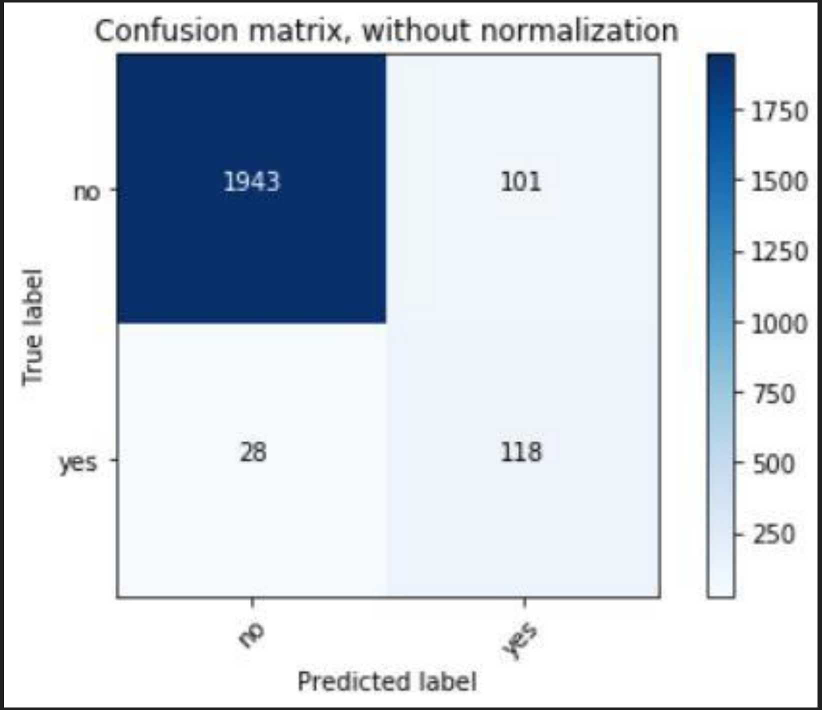
Random Forest

KNN

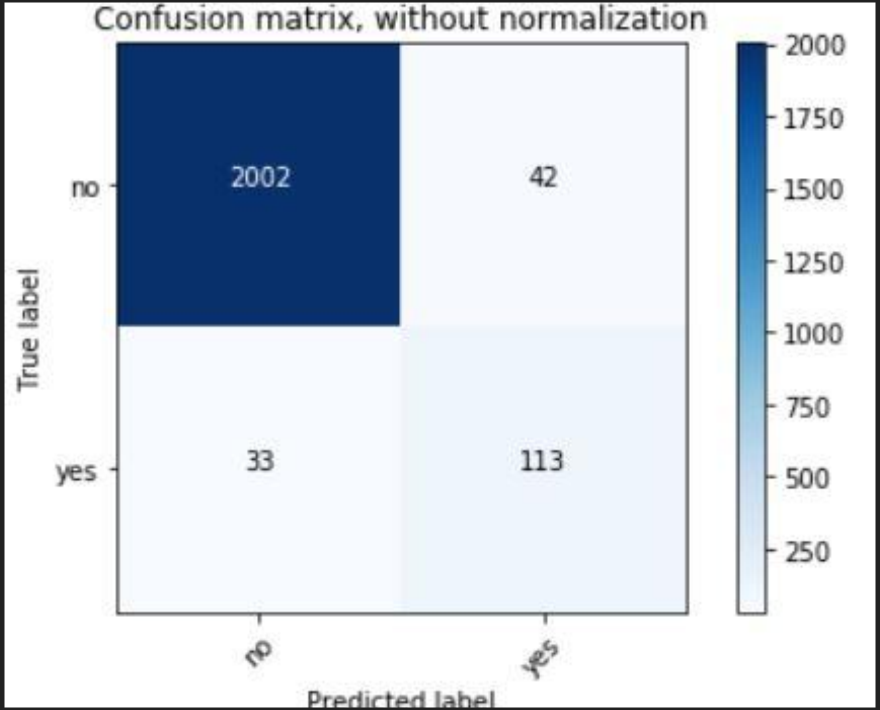




Random Forest



KNN



Feature Importance

	Caliper	Deep Induction Resistivity	Formation Density	Gamma Ray	Medium Induction Resistivity	Neutron Porosity	Photoelectric Effect	Spherically Focused Resistivity	Spontaneous Potential
Importance	0.32	0.13	0.04	0.15	0.03	0.02	0.07	0.04	0.21

Conclusion

- Datasets are complex and in high dimension space. In order to simulate the real world scenario, we use data from one well log to predict another one instead of having more training data and less testing data. (Because the data is kind of hard to get from the real world)
- Random forest has a better performance than KNN algorithm in this case.
- Creating some fake data can improve the performance dramatically in both two models and Random Forest seems to benefit more from it in terms of the performance.
- Augmented Random Over Sampling approach has no information loss and can prevent overfitting at the same time.

Thanks!