Augmented Random Over-sampling with Classification Algorithms on Well Log Datasets

Final Project Report
CSCI303 Intro to Data Science

Xiaoyu Zhu, Haoxuan Yang

April 2018

1 Abstract

This project will use two well-known classification algorithms (Random Forest and K-Nearest Neighbors) 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. In this project, we also find an approach called **Augmented Random Over-sampling** strategy involving K-means clustering to balance the numbers of instances of two class to improve the performance of the prediction result. We also analyze the feature contribution during the training task for the well log data using Random Forest. Empirical studies shows that our method can take advantage of the current information we have and further generate the synthetic data based on the patterns from the original one.

2 Challenges

There are ***so many*** challenges in this classification task:

- **Realistic Simulation of the well log Scenario**: Datasets are complex and in high dimensional space. In order to simulate the real world scenario, we use data from one well log to predict another one, which is commonly used strategy in this domain, instead of having more training data than testing data (because well log data is kind of hard to get from the real world).

- **Missing and highly Imbalanced data**: There are hundreds of NaNs in the datasets and due to the amount of NaNs in some observations are large and it is difficult to reconstruct their information, we just simply remove those observations. After removing the observations containing NaNs, **we only have 25 data points** labeled with "yes" ("yes" means having natural gas in this observation) and we need to predict thousands of data points in the testing dataset.

- **Data labeling**: 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.

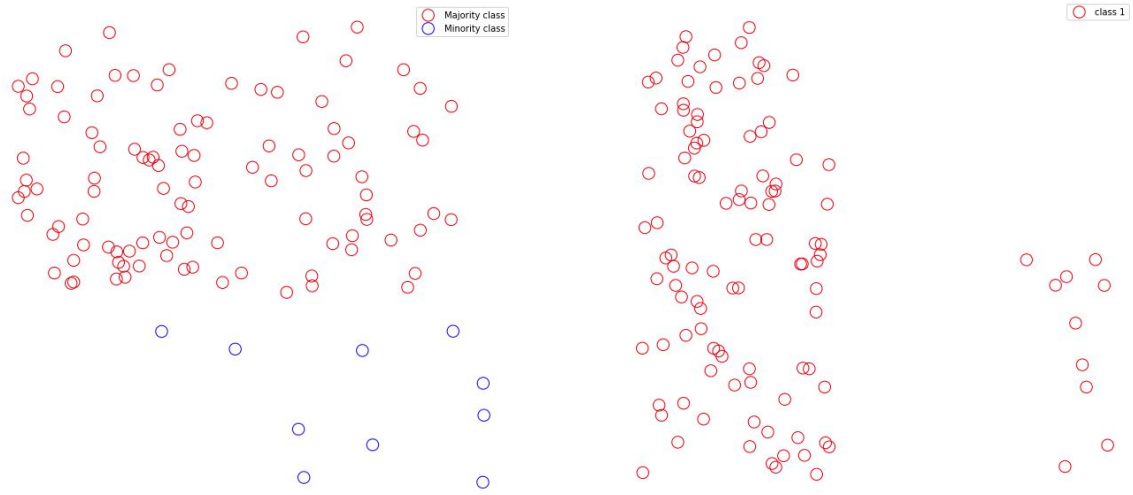
- **Overfitting Problem**: Since we are using the decision tree based Random Forest and K-Nearest Neighbors, overfitting should be considered to have a more generalized model. In order to avoid overfitting problem, we need to fine tune the parameters in the model and use cross validation to test the effectiveness our model.

3 Problem Background and Related Work

3.1 Problem of Imbalanced Data in Real World

In recent years, imbalanced data problem receives more and more attentions in many data related domains such as the detection of unknown and known network intrusions [1], and the prediction of oil spills in satellite radar images [2] or well logs not because its impact on the results but also has its research values. In these domains, the performance of prediction on the minority classes are very important.

Generally, there are two kinds of imbalanced dataset, between-class imbalance [3] and within-class imbalance [4]. Between-class imbalance usually occurs when some classes have much more instances than others as is shown in Figure 1a while within-class imbalance means some subsets in one class have much more instances than other subsets as is shown in Figure 1b. By convention, we define classes having more instances the majority classes in an imbalanced dataset and classes having fewer instances the minority classes.



(a) Between-class imbalanced data. (b) Within-class imbalanced data.

Figure 1: Examples showing between-class imbalanced data and within-class imbalanced data.

Most traditional classifier such as Random Forest, K-Nearest Neighbors, Logistic Regression, etc. are based on the assumption that the classes in the classification tasks are well-balanced. However, the classification using class-imbalanced data will cause a biased result in favor of the majority class. In this case, the minority classes are often ignored or treated as noises. That is why there is always a misclassification on the minority classes than the minority classes. What's even worse, This kind of bias in classification will be larger especially for high-dimensional data, where the number of features are large.

3.2 Over-sampling such as SMOTE Provides a Solution

To fix this problem, There are two main types of strategies. One is to use cost sensitive learning in which distinct costs have been assigned to improve the performance of the classification tasks [5, 6, 7]. Another one is to re-sample an original dataset [8, 9] in which under-sampling and over-sampling strategies has been intensively investigated in many studies to produce class-balanced data. Among these, over-sampling is very popular and has its advantage of creating synthetic data based on the prior knowledge from the data we currently have without losing the original data information. Synthetic minority over-sampling technique (SMOTE) [10] is one of the famous over-sampling algorithms which has been widely used in many application domains. This has led to two other methods such as borderline-SMOTE1 and borderline-SMOTE2 [11] which were proposed to achieve better performance (TP rate and F-value) based on the traditional method. All of these algorithms are using K-Nearest Neighbors to iteratively generate synthetic data.

However, there are some weaknesses in these algorithm when dealing with imbalanced data with some noise. One of the weaknesses is due to the fact that SMOTE uses uniform probability when choosing instances in a minority to perform over-sample as is shown in Figure 2. This issue raises when there is a between-class imbalance and the noise will be over-sampled which is not what we desired. What's worse, it will further amplify the noise from the minority class but located among majority class in the data.

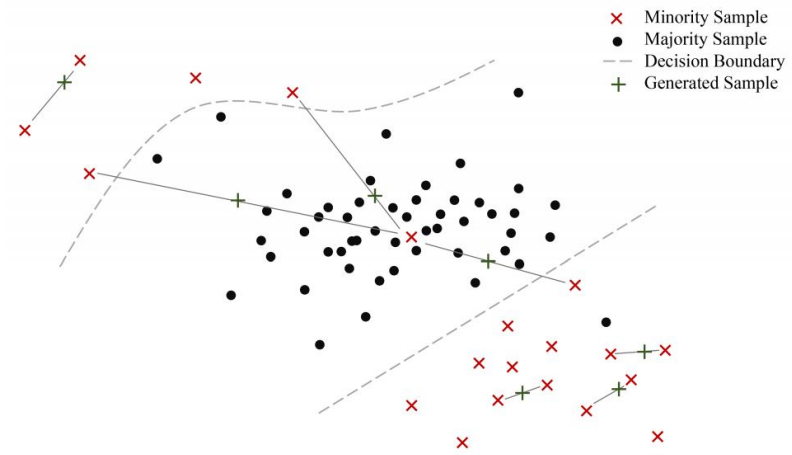


Figure 2: The performance of SMOTE algorithm on data with some noise.

3.3 Augmented Random Over-sampling Algorithm

Over-sampling involving K-means [12] algorithms can solve this problem by finding the several centroids of the minority class. Thus, the noise in the minority data will be ignored during the K-means process. We simply call this approach "Augmented Random Over-sampling (AROS) Algorithm". To implement this, we first extract the minority class data (data points labeled with "yes") from well log training dataset, then we perform K-means clustering with K = 1, 3, 5, 10. Based on the clusters computed from K-means, we generate synthetic data points around those centroids of the clusters under the assumption that the synthetic data should satisfy Gaussian distribution. To simply illustrate our approach, an example of this algorithm using K-means with K = 10 is shown in Figure 3.

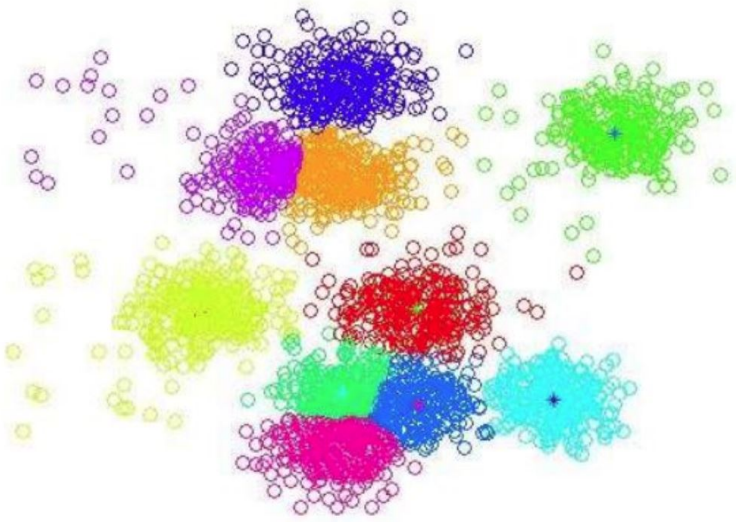


Figure 3: An example of Augmented Random Over-sampling Algorithm using K = 10. Each cluster in the figure has its different color.

4 Dataset Description

Well logging operations is a geophysical method which provide measurements of borehole and formation properties at measured depth [13]. 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)

Two wells in this field: **B.Yates 15 (BY15) and B.Yates 18D (BY18D)** are chosen to be assessed in this project. The number of observations is about 2000: from about 4700 ft to 6000 ft underground, at depth increments of 0.5 ft. Each well log data would include the following 9 features: Caliper data, Gamma ray data, Deep Induction Resistivity, Medium Induction Resistivity, Neutron Porosity, Formation Density, Photoelectric Effect, Spherically Focused Resistivity and Spontaneous Potential. 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.

4.1 Attribute Description

Listing of attributes:

- **Depth**: continuous value
  from about 4700 ft to 6000 ft
- **Caliper**: 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.
- **Deep Induction Resistivity**: continuous value
  Hydrocarbons, together with rock, and fresh water are all very highly resistive to electric.

- **Medium Induction Resistivity**: continuous value
  Hydrocarbons, together with rock, and fresh water are all very highly resistive to electric.
- **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.
- **Photoelectric Effect**: continuous value
- **Spherically Focused Resistivity**: continuous value
- **Spontaneous Potential**: continuous value
  Reflecting higher permeability in sandstone (could reserve oil) relative to lower permeability in shale. current flow.

5 Data Processing

5.1 Data Visualization

We plotted the well log data before processing to see the correlation between oil and well log features (Figure: 4, 5). Gamma Ray data, Neutron Porosity, Formation Density, Photoelectric Effect, Deep Induction Resistivity and Spontaneous Potential at different depth are plotted in curves. There are peaks of the curves occur at the same depth, indicating the abnormal layers under the ground. However, the peaks can be gas, limestone or salt water, so the gas is sometimes hard to recognize by this graph.

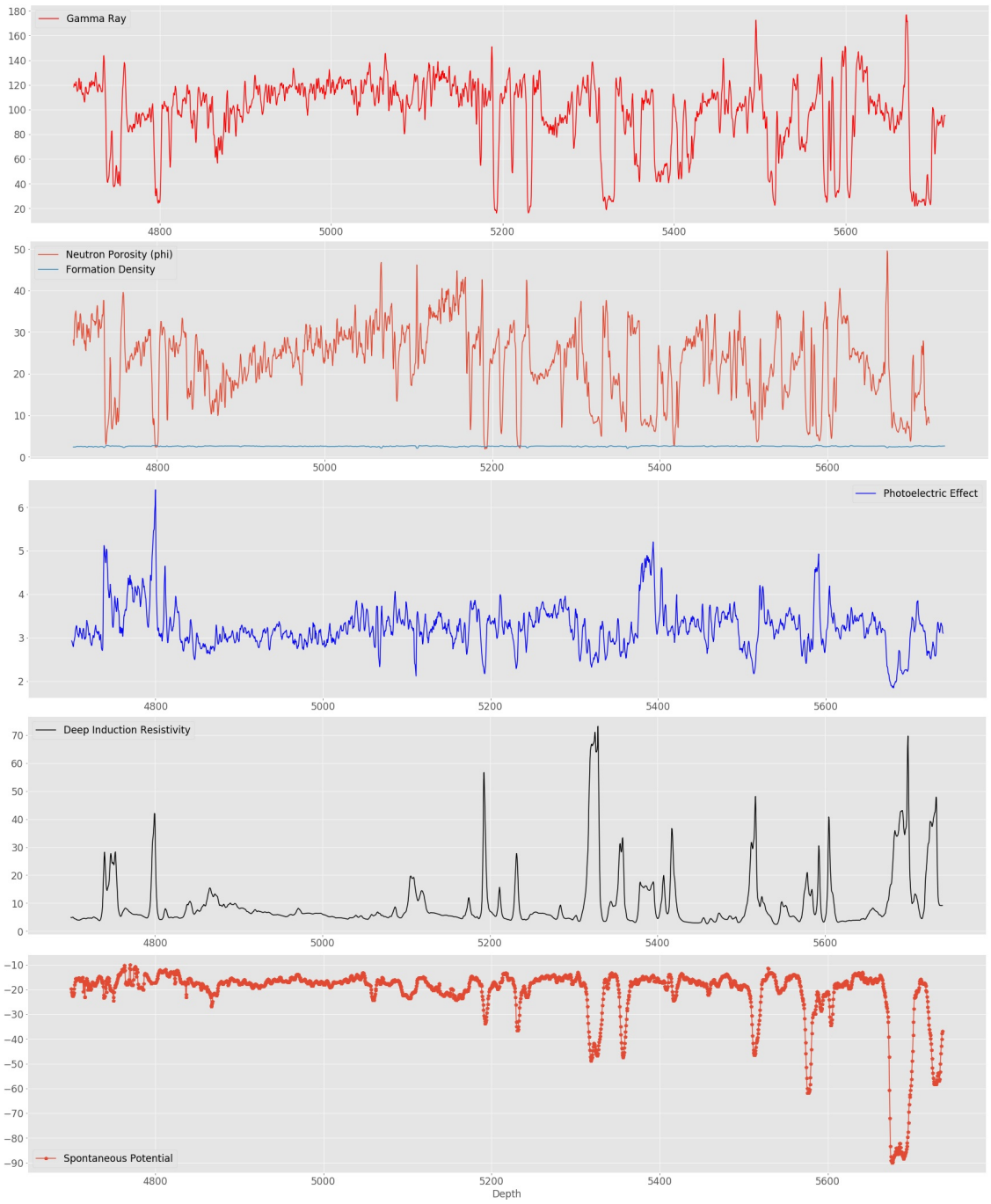


Figure 4: Gamma Ray data, Neutron Porosity, Formation Density, Photoelectric Effect, Deep Induction Resistivity and Spontaneous Potential at different depth in well BY 18D.

Figure 5: Gamma Ray data, Neutron Porosity, Formation Density, Photoelectric Effect, Deep Induction Resistivity and Spontaneous Potential at different depth in well BY 15.

5.2 Label targets

According to the SGR topical report of Boonsville Field [14], 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.

5.3 Data Visualization in Low-dimensional Space

After performing PCA on the BY18D data, we reduce 9 features to 3 principal components and plot the observations in 3-D dimension (Figure: 6). As we can see from the figure, there is a rough boundary between the 'Yes' samples and 'No' samples with a small intersected part. Since PCA is just a linear embedding method, sometimes it is hard to embed non-linear data such as one with a "Swiss Roll" shape. We also found that the performance before PCA is always better than the one after PCA which explains that our well log data is complex and might contain some non-linear parts.

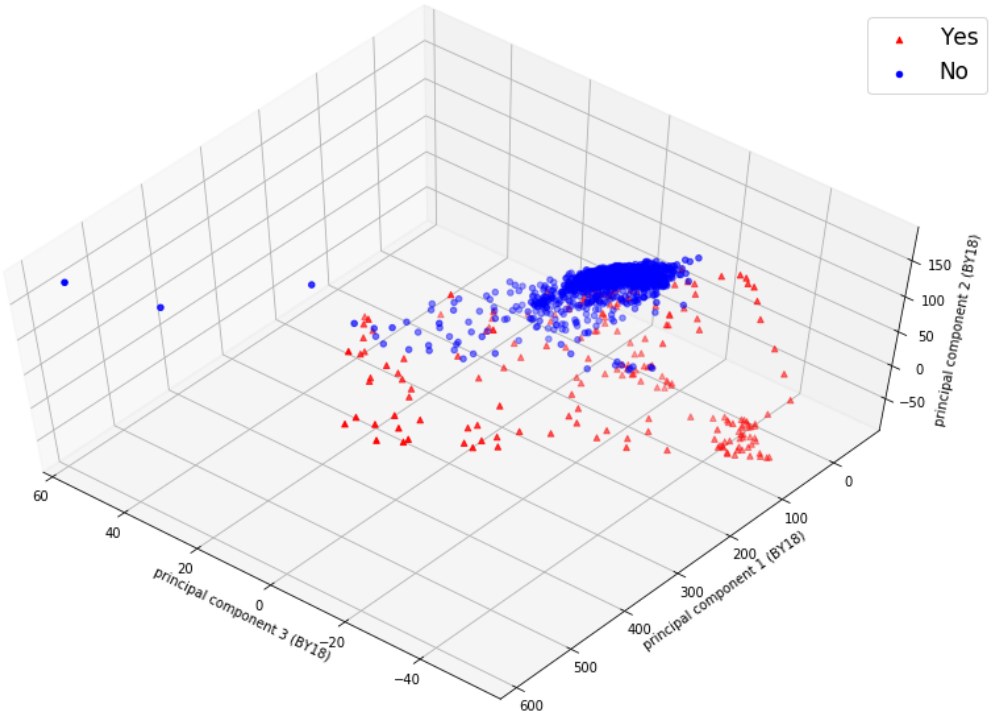


Figure 6: BY 18D Data Visualization in Low-dimensional Space

6 Prediction and Evaluation

6.1 Main Steps

1. We label the well log **BY18D** based on information from *Intro to Well Log* by adding an output of "yes" or "no" to the end of each row, indicating if there is oil in the layer of that depth.

2. We remove the observations with missing values in the data. Table 1 shows the data we have after the data cleaning process.

	No	Yes
BY18D (training)	1808	25
BY15 (testing)	2044	146

Table 1: Dataset Size before Augmented Random Over Sampling

3. Augmented Random Over-sampling algorithm with K = 10 has been performed in our minority class data points in the training dataset in order to get a balanced data. Table 2 shows the data we have after the over-sampling process.

	No	Yes
BY18D (training)	1808	1808
BY15 (testing)	2044	146

Table 2: Dataset Size after Augmented Random Over Sampling

4. We simply use PCA to project the data into a low-dimensional space just for visualization.

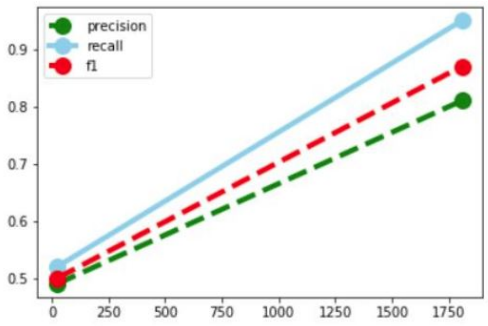
5. Supervised learning model by two traditional algorithms: Random Forest and K-Nearest Neighbors have been used to train and predict the data.

6. The grid search method combined with the 5-fold cross validation has been used on our log data of the well to evaluate the performance of the model.

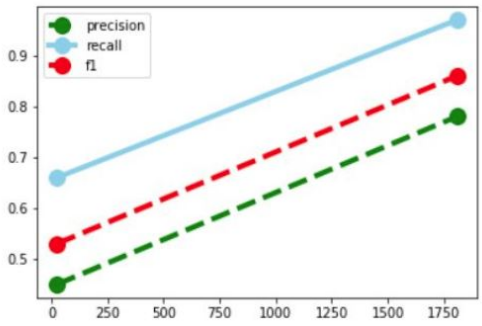
7. Since this is a classification problem, we just use the following evaluation methods: precision, recall and F1-score.

6.2 Experiment Results

Precision, recall and F1-score of the prediction tasks before and after Augmented Random Over-sampling on minority class (data labeled with "yes") of the testing dataset are shown in Figure 7a and 7b. Confusion matrix results are shown in Figure 8a and 8b. 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.

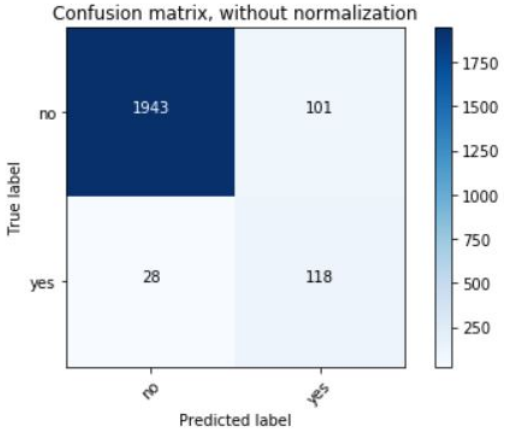


(a) Precision, recall and F1-score of the prediction task before and after Augmented Random Over-sampling with Random Forest model.

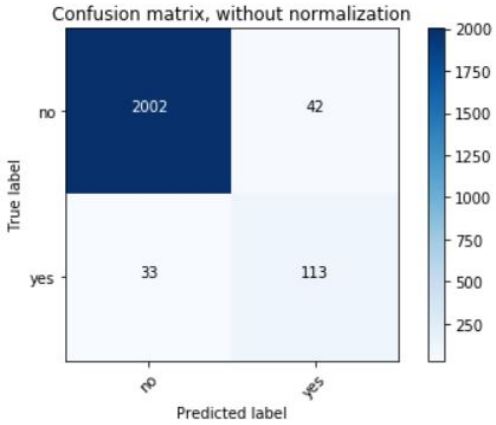


(b) Precision, recall and F1-score of the prediction task before and after Augmented Random Over-sampling with K-Nearest Neighbors model.

Figure 7: Performance of two models before and after AROS algorithm on minority class (data labeled with "yes") of the testing dataset



(a) Confusion matrix result using random forest.



(b) Confusion matrix result using K-Nearest Neighbors.

Figure 8: Performance of two models in terms of the confusion matrices of the testing dataset

Feature importance of each feature computed from Random Forest model is shown in Figure 3. The first feature Caliper has the biggest contribution in the classification task while Neutron Porosity, Medium Induction Resistivity and Spherically Focused Resistivity have comparatively very small contribution.

	Caliper	Gamma Ray	Deep Induction Resistivity	Medium Induction Resistivity	Neutron Porosity	Formation Density
importance	0.32	0.15	0.13	0.03	0.02	0.04

	Photoelectric Effect	Spherically Focused Resistivity	Spontaneous Potential
importance	0.07	0.04	0.21

Table 3: Feature Importance of Well log Data

7 Conclusion

This project has many challenges. Even though this well log datasets are highly imbalanced and in high dimension space, we still managed to get a satisfying classification performance with Random Forest and K-Nearest Neighbors models in terms of precision, recall and F1-score. Interestingly, we also noticed that Random Forest model seems to benefit more from the over-sampling method than K-Nearest Neighbors model. Augmented Random Over Sampling approach has no information loss. it can tackle the problem in traditional over-sampling method such as SMOTE and prevent overfitting at the same time.

8 Future Work

As the targets are not precise enough, the 'Yes' samples are not well distributed, so it becomes harder to train and predict the 'Yes' value. We may change the dataset to a more precise one and use several different datasets to test the effectiveness of the Augmented Random Over-sampling algorithm.

There is still some improvements can be done in the AROS algorithm. Since the K-means algorithms has many disadvantages, for example, it highly depends on the initialization result, K-mean++ can be used to have a better performance in the initialization process. Also, K-means clustering is also sensitive to the outliers in the dataset which means the locations of the centers computed and converged may be influenced by the noise far from the original data. To fix this problem, we can calculate the median value of the cluster in each iteration of the K-means clustering and this will lead to a more robust clustering result.

However, the datasets we have is complex and in high dimensions. We noticed that the performance of the two models did not benefit much from the PCA algorithm. This may be because there are some non-linear parts in our datasets due to the fact that PCA is a linear embedding method. Some non-linear based embedding approaches can be tested such as Laplacian Embedding, Locality Preserving Projection, etc to further improve the classification performance.

References

[1] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions," *Knowledge and Information Systems*, vol. 6, no. 5, pp. 507–527, 2004.

[2] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[3] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets. sigkdd explor newsl 6: 1–6," 2004.

[4] G. M. Weiss, "Mining with rarity: a unifying framework," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004.

[5] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 155–164.

[6] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: misclassification cost-sensitive boosting," in *Icml*, vol. 99, 1999, pp. 97–105.

[7] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, "Reducing misclassification costs," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 217–225.

[8] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000.

[9] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

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

[14] B. A. Hardage, Ed., *Secondary natural gas recovery : targeted applications for infield reserve growth in midcontinent reservoirs, Boonsville Field, Fort Worth Basin, Texas, volume I - topical report (May 1993 - June 1995)*. Chicago, Ill. : Gas Research Institute, 1995, no. Alma Record ID: 994353433502341.