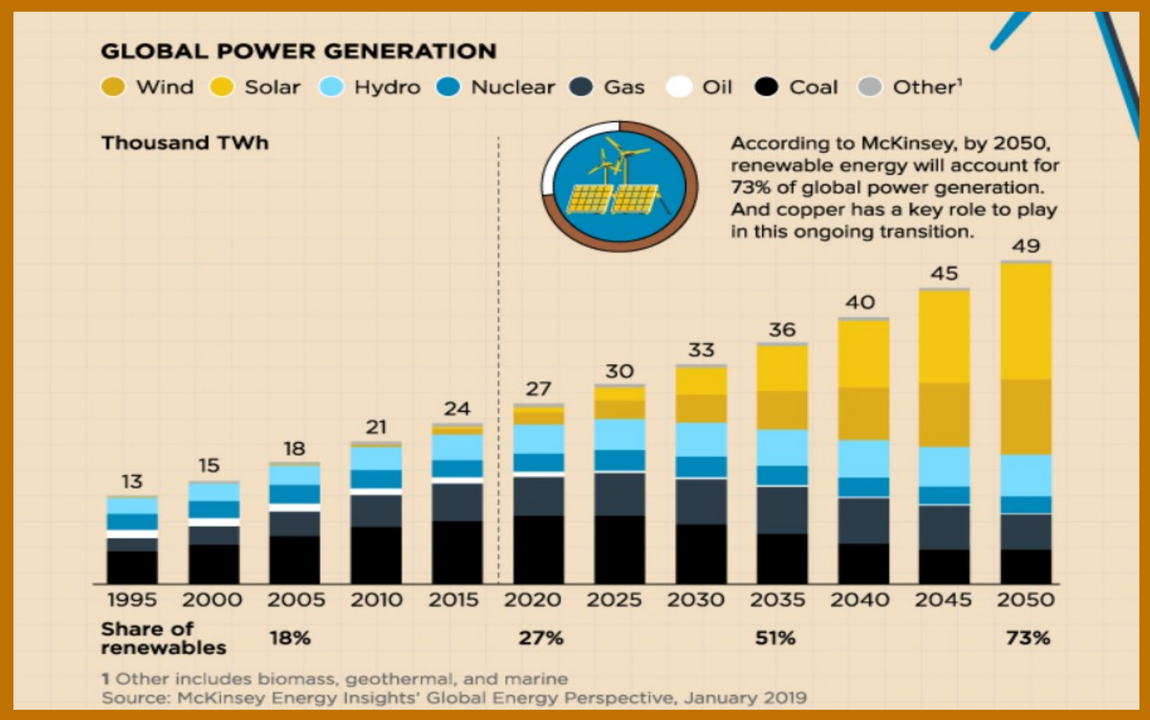
Predicting Magma Fertility for Porphyry Copper Exploration Using Machine Learning

By Precious Mungin, Nhi Hoang, Kamalesh Muthu and Johan John
UH Data Science for Energy Transition



Background/Purpose

Since copper is a great conductor of heat, renewable energies like wind and solar farms will require huge amounts of the metal. Subsequently, making copper a high demand commodity for renewable energy in the future. Currently the best location to collect copper from is porphyry deposits located near volcanoes. However, not all porphyry deposits are fertile or filled with valuable metals. In the paper "Application of machine learning to characterizing magma fertility in porphyry Cu deposits," S. Zou and his team showed that through the analysis of zircon geochemistry, machine learning models of Neural Network and gradient boosted decision trees were able to generate highly accurate models to separate fertile and infertile pophry copper locations.
As noted in the paper the most difficult part of creating these models was the collection of the data samples. One of the most basic rules of machine learning is that only "good" data will result in a "good" output from the model. In the manner of zircon geochemistry a "good" sample would be a sample that contains all of the following 19 elements (La, Pr, Ti, Nd, Er, Tm, Y, Hf, Eu, Sm, Gd, Ce, Dy, Lu, Tb, Th, Ho, Yb, U) along with containing no outliers in any of the elements mentions. This stringent set of requirements scoured our 3000 sample sized dataset until only 1300 usable samples were available. Zircon geochemistry is not an easy process, so losing half of the sample size in each batch is a terrible inefficiency, however it is better than the alternative, which is the cost to mine in an infertile porphyry copper deposit.The paper proves that using machine learning on zircon geochemistry samples can accurately predict whether porphyry deposits are fertile with copper (S. Zou). In our study we wanted to expand on this concept by trying to alleviate one of the primary inefficiencies of the previous model, the removal of zircon samples. Our research attempts to see if we could preserve samples by using the SciPy imputation library to find the missing values of the different elements and to see the effect of such a change on the end behavior on different machine learning models.

Models

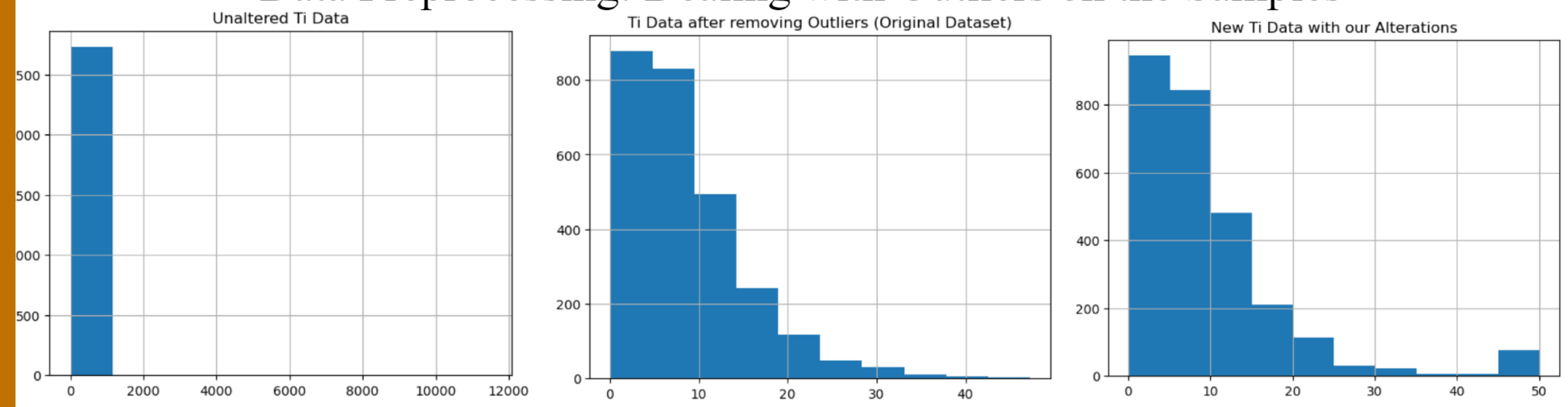
Neural Networks

NN models are good at processing large datasets to find relationships or patterns in the data. This was one of the models used by the article to evaluate if the zircon sample was taken from a porphyry deposit that was fertile with valuable metals. In general Neural Networks typically do better with more samples provided however this is only true if "good' samples or samples that accurately contain the relationship between different samples are provided. In the paper to preserve the integrity of the data the authors removed the "bad" samples which contained outliers or was missing element values. In our study we wanted to see if there was a way to preserve the integrity of the data without throwing away the samples. We predict that some of our imputations would improve the accuracy and predictability of this model.

Logistic Regression

This model is considered the baseline model for classification problems. Although it is sensitive to outliers, the size of a dataset is not a factor when trying to classify with Logistic regressions. This makes it a perfect the model to judge how much noise or meaningless data is added to the dataset when we impute values.

Data Preprocessing: Dealing with Outliers on the Samples



As you can see in the unaltered Ti histogram, majority of the sample exist in the first bin (0-1000 ppm) with outliers existing in the rest of the bins. These outliers can mess with the machine learning models so in the the paper (Zou, S.) the authors just removed the outliers from the dataset, which in the case of Ti is any sample over 50 ppm. Since our team was trying to preserve samples instead of removing outliers samples we just replaced the outlier value with the max value the sample can be which in this case is 50 ppm.

Data Preprocessing: Dealing with Missing Values in the Samples

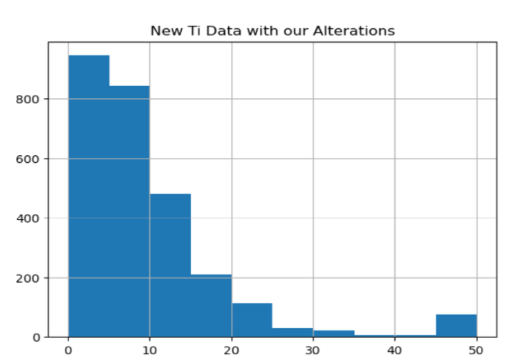
Original Dataset

The dataset we were provided is full of empty values since not all of the 14 elements analyzed are present in every zircon sample. Since the model looks at the ratios between the elements to make decisions an element without a value would lead the program to see a ration of either a zero, a non-number or infinity. It is very clear this is a problem we need to solve. In the paper (Zou, S.) the authors solved this problem by removing samples with missing values. This is a good solution since it preserves the original relationships between the samples without introducing noise into the sample.

Ti	Ce	Sm	Feature
x		x	
	x		
x			
	x		
		x	

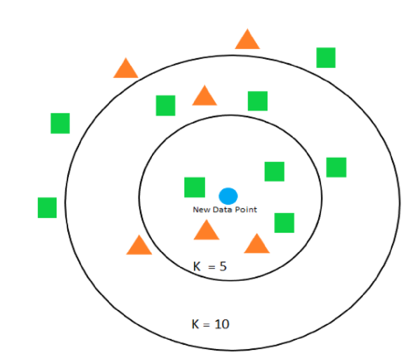
Univariate Imputation (Median)

Our goal in this project was to preserve the samples collected instead of removing the undesired ones. To do this we need to impute or replace the missing values with data that preserves the integrity of the dataset. Since the missing values are all numeric we decided to use the scipy imputation function to find the missing values. The first method we used was univariate imputation which takes in the known values of a particular feature to fill in the missing values of that feature. In this case we filled in the missing values with the median of the feature since as you can see in the Ti graph below the data is skewed to the left, so the median value is a better representation of the average of the data than the mean value.



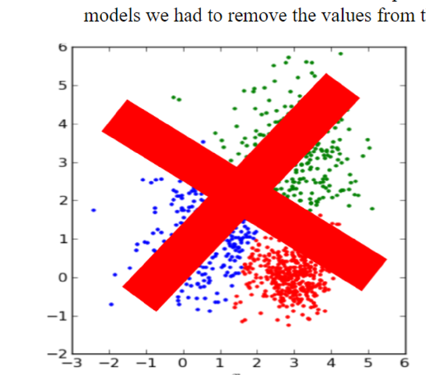
K-Nearest Neighbor

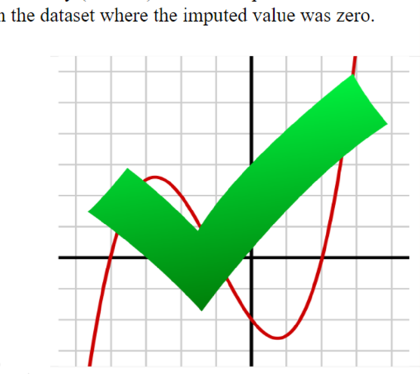
In the K-Nearest Neighbor (KNN) imputation the method takes in all of the features of the dataset except the missing feature and "plots" them and finds the nearest "neighbors" (in our case 2) and averages there values and gives it to the missing values. This type of imputation makes clusters very distinct and easy to see for the machine, however this comes at a cost of accuracy. We predict that this method would gives us the least loss while also being one of the more inaccurate methods.



Multivariate Imputation

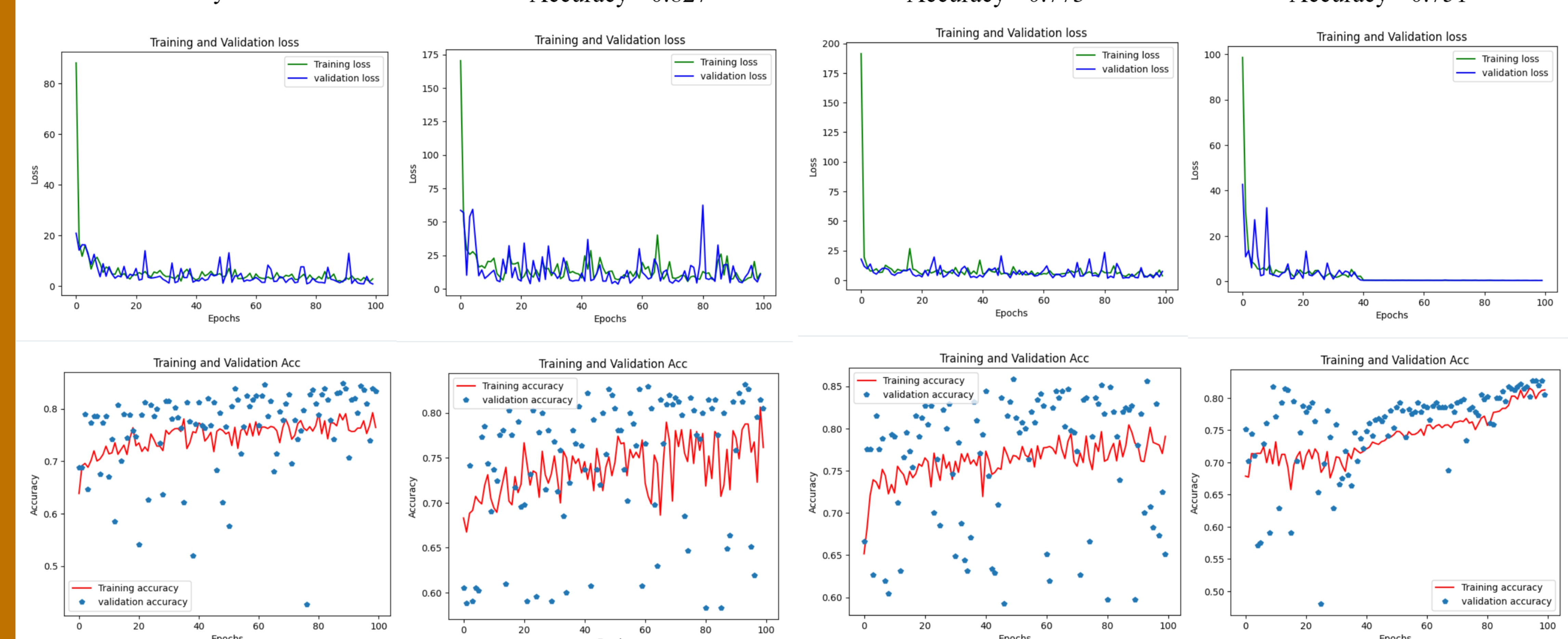
Multivariate imputation solves the missing value problem in a similar manner to the KNN imputation method. Like KNN the multivariate imputation takes in all of the features to calculate the imputed value. The difference between the two is that the multivariate imputation creates a function from the input that will output the missing value. Of the three imputation methods we tried the multivariate method would be the one that most closely preserves the integrity of the original dataset and the most accurate. However it is also the method that is the most prone to noise. The method assumes that there is a function that can be generated from the data that explains the relationship between the values. While this could be true real data is typically organized in clusters, leading to the actual value of the imputation to be different that the one recorded by the multivariate imputation. Furthermore this imputation might give values of zero into the dataset, which would mess up the models since ratios where zero is the divisor equal infinity ($x/0 = \infty$). In order to preserve the models we had to remove the values from the dataset where the imputed value was zero.

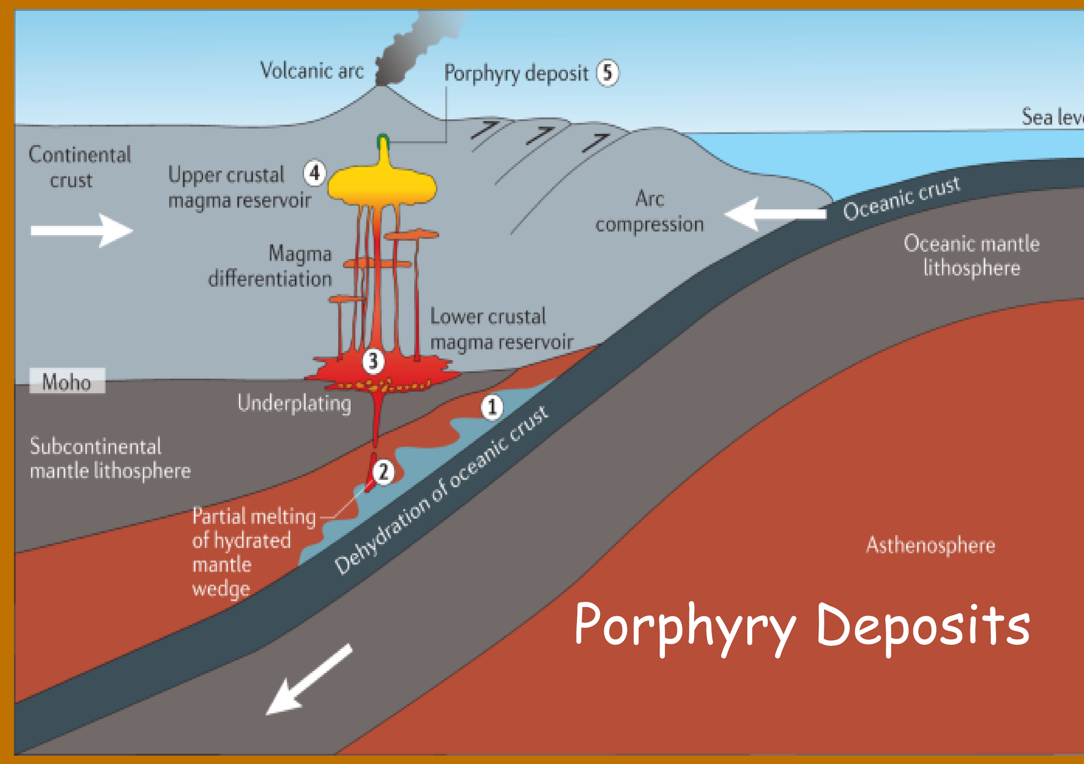




Results

KNN Imputation	Original	Multivariate Imputation	Univariate Imputation
Neural Network	Neural Network	Neural Network	Neural Network
Loss-0.466, Accuracy-0.788	Loss-1.846, Accuracy-0.812	Loss-1.426, Accuracy-0.854	Loss-0.466, Accuracy-0.788
Logistic Model	Logistic Model	Logistic Model	Logistic Model
Accuracy - 0.795	Accuracy - 0.827	Accuracy - 0.773	Accuracy - 0.751





Conclusion

In our study, we used three imputation techniques to compare the techniques with how closely we could preserve the integrity of the original dataset. We used KNN, univariate, and multivariate imputation to observe the differences it creates among the end behavior of the models. Univariate imputation implies that there is no relationship between the missing values and the other features in the dataset. By choosing the median to replace the missing values, it is able to preserve the distribution of the data. Multivariate imputation is useful there is a relationship between the missing values and the other features in the dataset. However, it can be prone to noise when the relationship between the variables is clustered. KNN imputation, like, multivariate imputation is useful when missing values are not random. It predicts missing values based on the values of the k-nearest data points in the dataset. This technique can preserve relationships between clustered data points. Looking at the Logistic Regression model we notice the most accurate model is the original dataset. The next accurate logistic model is the KNN dataset. This tells us that among the three imputations, KNN is the closest to the original dataset implying that the relationship between fertile deposits and infertile deposits are clusters. The most accurate model for the NN was the multivariate model with the accuracy of 85 percent followed by the other three methods which have an around 80 percent. However we can not say the multivariate model is the best model since it has a high value for its loss function. In general we found that imputation missing values decreases the loss function in the NN model which can be seen in the fact that the highest loss value is recorded by the original dataset which is the smallest. The two functions with the lowest loss function was the univariate imputation and the KNN imputation. Looking at the two models KNN seems to be the best general imputation technique for this dataset since it closely models the original dataset and has the least loss value. However it is clear that there are better imputation alternatives since both the original dataset and multivariate dataset outperformed the KNN imputation. Therefore it is important to know what your models need before choosing an imputation technique.

Acknowledgement

National Science Foundation (NSF IIS-2123247)
Data Science for Energy Transition
Dr. Jiajia Sun
Dr. Jonny Wu
Dr. Jeremy Tsung-jui Wu

Zou, S., Chen, X., Brzozowski, M. J., Leng, C.-B., & Xu, D. (2022). Application of machine learning to characterizing magma fertility in porphyry Cu deposits. Journal of Geophysical Research: Solid Earth, 127, e2022JB024584. https://doi. org/10.1029/2022JB024584
"Copper: Critical Today, Tomorrow, and Forever." *Visual Capitalist, 22 Oct. 2020, www.visualcapitalist.com/sp/copper-critical-today-tomorrow-and-forever.*

Link to Notebook