**Lithology Classification using X-Ray Fluorescence and a Random Forest**

Denise Robinson

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Jamia Mills

October 30, 2022

**Abstract**

Copper is an element essential to the manufacturing of greener technologies such as solar panels and wind farms, and the world is currently facing a copper deficit. To discover more copper, organizations drill holes into the subsurface. The holes are subsequently logged for lithology, which is a subjective process. One technique to add objectivity to lithology logging is through lithogeochemistry. Lithogeochemistry is the process by which rocks are described using multi elemental chemical analyses. A portable x-ray fluorescence device (pXRF) provides real-time multi elemental analysis of rocks. This study investigates if lithology units can be accurately classified using multiple element geochemistry collected using a portable x-ray fluorescence (pXRF) device using a random forest predictive algorithm. The pXRF data was collected from a drill hole originating at a copper porphyry project in Arizona every ten centimeters. The final dataset had 11 elements and seven units to be classified. A random forest was generated using python that had an accuracy of 96.3%. The most important elements were titanium, iron, strontium, and copper. We conclude that pXRF data can be used as a tool to assist geologists in adding a level of objectivity to lithology logging.

*Keywords:* Copper, lithology, portable x-ray fluorescence (pXRF), lithogeochemistry, random forest

**Lithology Classification using X-Ray Fluorescence and a Random Forest**

As the world pivots to more sustainable energy, so too must the technologies adapt to a cleaner footprint. Greener technologies must still be manufactured using resources found on earth. Copper, which has both high thermal and conductive properties, is one of the most widely used elements for the development of such tools (Visual Capitalist, 2021). For example, Rapier (2021) reports that each megawatt of energy produced in an offshore windfarm requires 9.6 metric tons of copper, while solar panels need 5 metric tons of copper per MW produced. Not only is copper critical, but the world is currently facing a copper deficit. In fact, Rapier (2021) states “Analysts have warned that the copper industry needs to invest more than $100 billion to avoid what could be an annual supply deficit of 4.7 million metric tons by 2030,” (para. 7). As such, the discovery and extraction of copper has become a high priority for the renewable energy sector.

Copper is an element that will aid in the production of more sustainable energies, but to do so, more copper must be discovered and extracted from the ground. One of the ways to assist geologist with mineral exploration and modelling is lithogeochemistry. Essentially, lithogeochemistry is the process by which lithological units are characterized by their geochemical signatures (Pantazis, 1988). Correctly logging lithological units can reduce the time and resources necessary to find and mine copper.

Traditionally, to obtain the geochemical composition of surface, reverse circulation (RC), or core samples each sample must be split, crushed, and pulverized before being analyzed. The sample preparation process often causes a long lead time between sample collection and actual data acquisition. These issues prevent many exploration companies from fully implementing lithogeochemistry into the logging process, instead relying solely on logging geologists to classify lithologic units. More recently, a tool called a portable x-ray fluorescence (pXRF) machine has become popular in mineral exploration.

A pXRF is a handheld device that provides near real-time multiple element geochemistry. “From the analysis of X-ray fluorescence data the energy and intensity of the characteristic photons emitted by the sample can be found,” (Mihut & Ianasi, 2021). In its simplest form, the pXRF bombards a substance, in this case a rock, with x-rays and then interprets the chemical composition based on the energy and intensity of the photons received by the machine.

The deployment of a pXRF has many benefits in the mineral exploration industry. First, the time frame to receive the data is nearly non-existent, while lab analysis takes weeks to months. By analyzing and receiving geochemical data immediately, it’s possible the results could be applied to the classification of the lithology units. For example, Gazley et al. (2014) successfully developed a schema to utilize pXRF data in rock classifications at a plutonic gold mine in Western Australia. Further, Trott et al. (2022) deployed a random forest to classify lithological units using pXRF data with a final model accuracy of 87%.

The implication of such successful deployments of pXRF data to classify lithological units is that pXRF integration into mineral exploration could add a layer of objectivity to geologic logging, which tends to be subjective based on the logger (Gazley et al., 2014). Consequently, better and faster geological logging could lead to more accurate and more successful exploration campaigns for necessary minerals such as copper.

**Objectives**

Lithological logging is one of the most important aspects of mineral exploration. Logging of exploration materials such as core or rock chips directly impacts the project. Understanding of the subsurface geology and structures paint the picture of the geologic history responsible for the area. Subsequently, this knowledge can lead geologists to refine exploration targets and gain more understanding of where the minerals in question may exist underground.

That said, there are usually multiple geologists responsible for logging at any given project. For a process so critical to the understanding of a project and geologic occurrences, inconsistencies and subjectivism are often displayed in logging between geologists (Gazley et al., 2014). Additionally, from a data perspective, it’s difficult to determine and track such inconsistencies. It is therefore important to incorporate objective classification methods to the lithology logging, such as utilizing the portable x-ray fluorescence (pXRF) device.

The aim of this project will be to determine if lithology units can be accurately classified using multiple element geochemistry collected using a portable x-ray fluorescence (pXRF) device. To achieve this aim, the following objectives will be met:

* To identify if data provided by a pXRF scan can be used to classify lithological data;
* To determine which chemical elements provided by a pXRF are the most useful in lithology classification; and
* To assess the success of the classification model.

**Overview of Study**

Mineral exploration companies require a way to increase objectivity to lithologic logging. By adding quantitative elemental data, geologists can have more evidence to determine lithology units. Further, the portable x-ray fluorescence (pXRF) can easily be deployed in field settings and provides near real time geochemical data.

We will exploration the relationships between the elemental results of a pXRF device and the lithology unit using a random forest classification method. We will determine if pXRF data can be used to assist geologists with lithology logging. We will also confirm which predictors demonstrate the strongest relationship to the target variables.

**Research Questions and Hypotheses**

As lithology logging is subject, especially when completed by multiple geologists, implementation of a chemical analyses in conjunction with lithological logging would provide a level of objectivity to the process. For many organizations, it’s not time nor cost effective to send entire lengths of core to the lab to be processed prior to logging. A handheld pXRF unit would address both the cost and time issues of submitting samples to an analytical lab. The pXRF provides near real time data. This would allow geologists to make more confident lithology calls, as well as selectively send samples to the lab for finalized analyses.

Can a combination of heavy and light elements provided by pXRF readings every ten centimeters be used to accurately categorize the lithology unit of a copper porphyry deposit located in Arizona, United States? The heavy elements for this project consist of copper (Cu), rubidium (Rb), and strontium (Sr). The light elements include aluminum (Al), silicon (Si), sulfur (S), potassium (K), calcium (Ca), titanium (Ti), manganese (Mn), and iron (Fe).

**Ho:** All elements (Cu, Rb, Sr, Al, Si, S, K, Ca, Ti, Mn, and Fe) have an impact on the lithology unit.

**Ha:** All elements (Cu, Rb, Sr, Al, Si, S, K, Ca, Ti, Mn, and Fe) do not have an impact on the lithology unit.

**Literature Review**

**Mineral Exploration and Lithologic Logging**

In mineral exploration projects, surficial geologic mapping and sampling is often the first step, but geologists will ultimately need an understanding of the subsurface. One popular method to investigate the subsurface is by drilling diamond core holes into it. Diamond core holes are referred to as such, because they deploy a diamond drill bit. The diamond coated, hollow drill bit rotates in a circular motion and produces a solid cylinder of the rock that it drilled through (Trenchlesspedia, 2021). This is called core.

Logging of the core is a critical step to understanding the rocks below ground. By logging in this scenario, we refer to lithology logging. Lithology is “the mineralogical composition of a rock unit,” (Thompson, 2012, p. 544). Lithology logging produces a continuous downhole report of the rock units contained within the core. Lithology logging is one of the most important and subjective activities of mineral exploration.

Lithology logging is a crucial part of mineral exploration, because it allows drill holes to be spatially correlated in 3-D, which illustrates the rock units and their orientations in the subsurface (Blackbourn, 2009). For such a critical task, most organizations rely solely on geologists to correctly identify the lithologies. In some instances, differentiating between rock units can be easily accomplished, but in others, rocks that are very different chemically can look the same, complicating identification. Logging tends to be more subjective than objective, which can cause data inconsistencies (Gazley et al., 2014). Incorrect logging is an issue that most exploration projects will experience, thus it’s critical to apply some objective methods to assist with logging.

**Portable X-Ray Fluorescence (pXRF) and its uses in Geology**

The chemical composition of rocks, commonly referred to as geochemistry, can be extremely useful to understand depositional environments, and thus ore deposits (Gazley et al., 2011). Wet lab analyses are well known and used throughout the mineral exploration industry to determine concentrations of elements in rock samples. While lab analyses are considered industry standard for result reporting, the process is both expensive and time consuming.

To combat the expense and time involved in wet lab analyses, some organizations have started incorporating x-ray fluorescence. “X-ray fluorescence (XRF) spectrometry is a non-destructive, rapid, simultaneous multi-element analytical methodology,” (Shand & Wendler, 2014, p. 31). X-ray fluorescence is the process by which x-rays bombard a surface and the device interprets the chemical composition by analyzing the energy and intensity of the photons received (Mihut & Ianasi, 2021). In portable x-ray fluorescence (pXRF), the device is handheld and can easily be deployed in remote areas. Not only is pXRF easy to deploy, but it provides near real-time results.

Essentially, with a pXRF, it’s possible for remote mineral exploration projects to have immediate geochemistry to assist in lithological logging. Utilizing the real-time geochemistry information from a pXRF device, geologists can apply a level of objectivity to an otherwise subjective practice of lithology logging (Gazley et al., 2014).

**Random Forest Applications in Lithogeochemistry**

As previously discussed, lithology logging is often subjective and not repeatable by multiple geologists. By deploying a pXRF in field settings, we can add a layer of objectivity and quantification to the logging practice. That said, it’s essential to understand which elemental concentrations are the most important when classifying the rock units. To accomplish this, many organizations are relying on machine learning methods (Trott et al, 2022).

A random forest is a supervised machine learning technique that involves the ensemble of many decisions trees (Yiu, 2019). Because random forests are ensemble machine learning methods, they tend to have fewer issues with overfitting the data than standard decision trees. One of the most useful aspects of a random forest is that it clearly defines which input variables are the most important (Yiu, 2019). A clear understanding of the most important elements for lithologies could provide geologists with enough knowledge to implement a pXRF and interpret the likely lithology unit without using any further equipment. Further, Trott et al. (2022) clearly demonstrated the successful application of a random forest using pXRF data to classify lithology units at a porphyry-epithermal project located in Chile.

**Research Design**

**Methodology**

The project is comprised of entirely quantitative methodologies as the independent variables are numeric. The quantitative technique will be experimental, as we will be evaluating the cause and effect (University of Newcastle Library Guides, n.d.). In this case, the potential cause evaluated is the chemical composition and the effect on the lithological unit. It will be conducted using hypothesis testing. The null hypothesis states that all elements (Cu, Rb, Sr, Al, Si, S, K, Ca, Ti, Mn, and Fe) have an impact on the lithology unit. In this case, impact will be defined as the presence of all elements within the final predictive model. The pXRF data will be cleaned and combined with the lithology data, and the final dataset will be loaded into python, and then classified using a random forest.

**Dataset**

The pXRF data was collected and processed by the Center for Advanced Subsurface Earth Resource Models (CASERM), which is a research consortium comprised of the Colorado School of Mines and Virginia Tech (CASERM, 2022a). Separate datasets (heavy and light elements) were provided. These data were delivered separately, as they were scanned using different pXRF device settings. Specifically, the light elements were collected using 15 kV and 24 mA, while the heavy elements were gathered using 40 kV and 8 mA using a silver (Ag) anode x-ray tube in a helium atmosphere (CASERM, 2022b). As part of the transformation process the heavy and light elements were combined into a single data set, the join id was created by concatenating the hole number with the ‘from’ and ‘to’ columns. Only intervals that occurred in both datasets were considered. Further, some of the elements contained a significant number of ND or not detectable values. Again, as part of data transformation, these elements were removed from the data. Of the remaining elements, any samples containing any ND values were also removed. This process was completed to ensure only readily detectable elements would be considered for the predictive model. The final dataset contains the following elements: Al, Si, S, K, Ca, Ti, Mn, Fe, Cu, Rb, and Sr.

The lithology logging was completed on site in Arizona by an experienced geologist. To combine the lithology log with the pXRF data, the interval values (from and to) were used. Only pXRF readings that occurred completely with in a lithology unit interval were included. In other words, if a pXRF value crossed a lithology break, it was excluded from the final data. The final dataset has 3917 entries.

**Methods**

To analyze the data two main programs were deployed: PyCharm, a python graphical user interface (GUI) and Orange, an open source data visualization tool. Descriptive statistics were calculated for each of the independent variables, in this case, the separate elements using the Orange. This process provided insights into the distributions and possible outliers of each element. For actual data analysis and modeling, Pycharm and python libraries were used. First, the data was partitioned into 70% for training data and 30% into a test set. The model used for this project was a random forest. Random forests are supervised machine learning ensemble classifiers (Provost et al., 2017) (Yiu, 2019). Random forests are considered ensemble classifiers as they combine many decision trees into a single model, which leads to less overfitting. In this case, our random forest combined the effects of 500 decision trees. Random forests have also been previously proven as successful models to use for lithology classifications based on pXRF data (Trott et al, 2022). To assess the results of the model, the final accuracy percentage and a confusion matrix were used. Finally, to evaluate the null hypothesis, we will determine if all independent variables exist with in the final classification model.

**Limitations**

The limitations of this project are mostly encompassed with the dataset. While there were several thousand entries of data, there were only pXRF results for a single drill hole available. As such, we must infer results based on a single drill hole for an entire exploration program. The use of hundreds or thousands of drill holes in the same area would provide much larger dataset for analysis. Another limitation of the data was the number of elements that had to be removed entirely because they contained few actual quantitative values. X-ray fluorescence detects some elements easily, while others are more difficult or impossible for the machine to pick up. The limitations of the pXRF device readings restrict the project to only those elements that are readily detectable. Another limitation is that of the original logging data. While a very experienced geologist logged the hole used, it’s still possible that some units were mislabeled. Because we are only using the data from one drill hole, any misidentifications in the drill log could greatly impact the model. Finally, the logging data is very skewed regarding the distribution of the different lithology units. A more equally distribution of target variables would be preferential, as the model would have more data to train on, especially for the less common lithology picks.

**Ethical Considerations**

Ethical considerations are extremely important to research projects. While the data in this project does not involve human participation, there are still confidentiality concerns that must be addressed (Byrne, 2017). The organization involved in this project is a publicly traded exploration company. For public companies, material data disclosure is critical to staying within the trading laws. Material information can be considered any data which would affect the price of stock and thus the company. We do not expect this project to produce any material information, but as a precaution, all identifying information from the company as well as the drill hole name have been redacted. Further, we obtained consent from the organization prior to conducting the experiment. This research project will be managed with the same confidentiality and considerations as one involving humans.

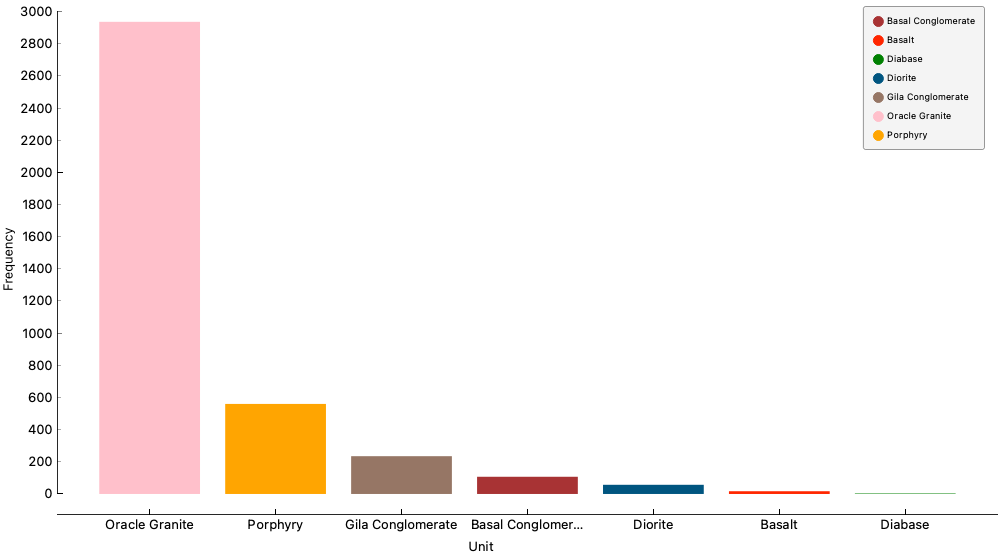
**Findings**

**Data Exploration**

The dataset contains 3915 entries and 17 variables. Our target variable, Unit, is a categorical variable that is the lithological unit pick made by the geologists during the logging process. There are seven distinct unit categories: basal conglomerate, basalt, diabase, diorite, gila conglomerate, oracle granite, and porphyry. The frequency of each category was plotted using a bar chart (see Figure 1).

**Figure 1**

*Unit Frequency in Bar Chart*



*Note*. Frequency bar chart of lithology units

The bar chart clearly illustrates the skewness of the data. Oracle Granite comprises the vast majority of the dataset with 75% of the total data set having been logged as Oracle Granite. With some model types, such disproportionality among the target variables could affect the performance of a machine learning model, but random forests are considered appropriate algorithms in this case as they are ensemble models and use numerous individual classification trees (Zhang, 2015).

We then examined the distributions of the 11 numerical variables that will act as our predictor variables: Al\_pct, Si\_pct, S\_pct, K\_pct, Ca\_pct, Ti\_ppm, Mn\_ppm, Fe\_pct, Cu\_ppm, Rb\_ppm, and Sr\_ppm (see Figure 2). Several of the predictor variables show fairly normal distribution such as Al\_pct, Si\_pct, and K\_pct. All other predictor variables display heavily right skewed distributions.

**Figure 2**

*Distributions and Descriptive Statistics for Predictor Variables*

**

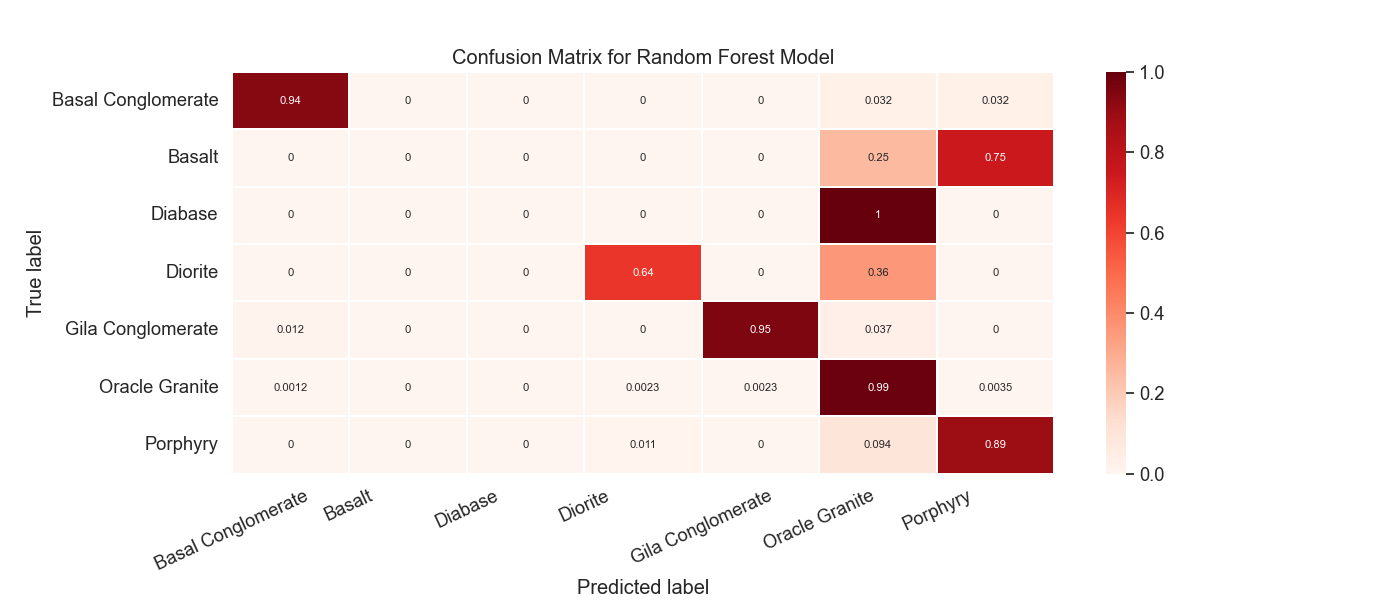
*Note.* Histograms and descriptive statistics for the predictor variables created with Orange.

**Random Forest Application**

Our random forest model seeks to predict the lithology unit using the numerical pXRF values as the input values. Our dataset was imported into a Pandas data frame within the PyCharm GUI, and the column headers were defined. Two separate data frames were created. The ‘X” or horizontal data frame was comprised of all 11 predictor variables and ‘y’ or our vertical data frame which contained only the target variable. We used the Scikit library to split the dataset into 70% training and 30% validation. Essentially, our model was trained using a random selection of 70% of the data, and the model was validated using the remaining 30%. We further employed the Scikit python library to train our random forest. As random forests are ensemble methods, we used a combination of 500 individual decisions trees to complete the final model. Once the model was trained, we deployed the model on the test dataset. Our final model had an accuracy of 96.3%. The Scikit and Matplot libraries were used to produce a confusion matrix to investigate the success of the model regarding each unit rather than overall (see Figure 3).

**Figure 3**

*Random Forest Confusion Matrix*

*Note.* Confusion matrix produced using Scikit Learn and Matplot.

The model performed well with some lithology units and poorly with others. Based on the confusion matrix, the random forest was affective at correctly identifying gila and basal conglomerates, but did not perform well at all regarding the basalt and diabase. If we view the classification report, we can see that basalt and diabase both had very few instances tested by the random forest with four and two, respectively (see Figure 4). This is a direct consequence of our small dataset and highly skewed classification variables. Further, based on the f-1 score, which takes both precision and recall into consideration, the model performed well for the basal conglomerate, gila conglomerate, oracle granite, and porphyry units. It performed moderately at classifying diorite, and it failed to classify basalt or diabase. Basalt and diabase did not have enough test instances to be considered adequately tested. Thus, we do not feel we can comment as to the accuracy of the model regarding basalt and diabase given the current data set.

**Figure 4**

*Classification Report*

*Calendar

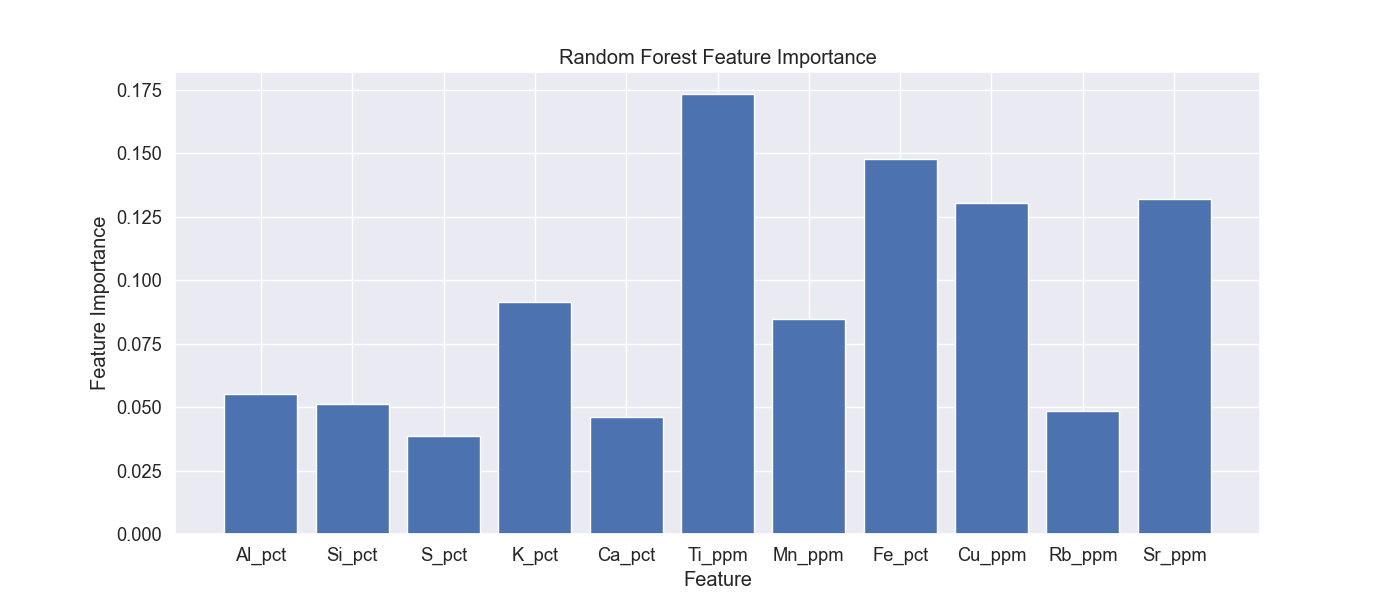
Description automatically generated*

*Note.* Screen shot of classification report using the Scikit library.

To test our null hypothesis, which states all predictor variables will have an impact on the final model, we produced a feature importance chart using the Scikit library (see Figure 5). The feature importance indicates the relative importance of each variable (Brownlee, 2020). For our purposes, a feature with an importance greater than zero will be considered as part of the final model.

**Figure 5**

*Feature Importance Bar Chart*



*Note.* Bar chart depicted the feature importance for predictor variables in random forest model created with Scikit and Matplot.

Because each of the 11 predictor variables has a feature importance greater than zero, we fail to reject the null hypothesis. Titanium (Ti\_pct) was the most important predictor variable in our final model, followed by iron (Fe\_pct), strontium (Sr\_ppm), and copper (Cu\_ppm). This means that titanium, iron, strontium, and copper were the most important elements when classifying the lithologic unit.

**Conclusion**

Lithologic logging is by nature a subjective endeavor, especially with multiple geologists are responsible for the logging. One of the major obstacles faced by mineral exploration companies is to produce true and accurate lithologic logs to assist in the exploration and location of minerals necessary for moving toward a greener future. The use of rapid geochemical results produced using a pXRF unit in conjunction with a random forest machine learning algorithm has been demonstrated as an affective technique to add objectivity to the lithologic logging of a copper porphyry deposit in Arizona. We also found that all 11 predictor variables (Cu, Rb, Sr, Al, Si, S, K, Ca, Ti, Mn, and Fe) impacted the final model, thus we failed to reject our null hypothesis.

**Recommendations**

Our model proved affected at adding objectivity to lithology logging. Other organizations could benefit by implementing a program where pXRF data is used in tandem with a random forest classifier to objectively classify lithologies. This tool could be used by geologists in the field to increase the accuracy and consistency of lithology logging. The use of pXRF data to supplement logging has real-time applications while waiting for analytical results from the lab can take weeks or even months.

Copper is an element critical to the development of a renewable energy. As such, the continued mining and thus exploration of copper is necessary to facilitate the manufacturing of greener technologies and infrastructure such as electric vehicle batteries, solar panels, and wind turbines. By providing explorationists with further tools to correctly and objectively classify lithological units using pXRF data will add efficiencies and streamline the process to discover and extract such crucial metals from the earth.

**References**

Blackbourn, G. (2009, March 19). Cores and Core Logging for Geoscientists (1st ed.). Whittles Publishing.

Brownlee, J. (2020, March 29). *How to Calculate Feature Importance With Python*. Machine Learning Mastery. <https://machinelearningmastery.com/calculate-feature-importance-with-python/>

Byrne, D. (2017). Why Is Confidentiality an Important Aspect of Research?. Project Planner. 10.4135/9781526408556.n3.

CASERM. (2022a, August 25). Center for Advanced Subsurface Earth Resource Models. Center for Advanced Subsurface Earth Resource Models. Retrieved September 24, 2022, from https://caserm.mines.edu/

CASERM. (2022b). *Calibration Data Generation Laboratory assays graphed against Minalyze results for same intervals.* Center for Advanced Subsurface Earth Resource Models. Center for Advanced Subsurface Earth Resource Models.

Gazley, M. F., Vry, J. K., du Plessis, E., & Handler, M. R. (2011). Application of portable X-ray fluorescence analyses to metabasalt stratigraphy, Plutonic Gold Mine, Western Australia. Journal of Geochemical Exploration, 110(2), 74–80. https://doi.org/10.1016/j.gexplo.2011.03.002

Gazley, M. F., Tutt, C. M., Fisher, L. A., Latham, A. R., Duclaux, G., Taylor, M. D., & de Beer, S. J. (2014). Objective geological logging using portable XRF geochemical multi-element data at Plutonic Gold Mine, Marymia Inlier, Western Australia. *Journal of Geochemical Exploration*, *143*, 74–83. <https://doi.org/10.1016/j.gexplo.2014.03.019>

Mihut, N.M., & Ianasi, C. (2021). Overview of x-ray fluorescence spectrometry. *Fiability & Durability / Fiabilitate Si Durabilitate*, *27*(1), 71–74.

Pantazis, T.M. (1988). Lithogeochemical prospecting. General Geology. *Encyclopedia of Earth Science*. Springer, Boston, MA. <https://doi.org/10.1007/0-387-30844-X_64>

Provost, F., Hibert, C., Malet, J.-P. (2017). Automatic classification of endogenous landslide seismicity using the random forest supervised classifier. *Geophysical Research Letters, American Geophysical Union.* 44 (1), 113-120. <https://doi.org/10.1002/2016GL070709>.

Rapier, R. (2021, September 17). *Will the copper deficit derail the Renewable Revolution?* Forbes. Retrieved September 15, 2022, from https://www.forbes.com/sites/rrapier/2021/09/17/will-the-copper-deficit-derail-the-renewable-revolution/?sh=76f80afc775f

Shand, C. A., & Wendler, R. (2014). Portable X-ray fluorescence analysis of mineral and organic soils and the influence of organic matter. *Journal of Geochemical Exploration,* 143, 31–42.

Trenchlesspedia. (2021, March 15). Core Drilling. Retrieved October 9, 2022, from https://www.trenchlesspedia.com/definition/4203/core-drilling

Trott, M., Leybourne, M., Hall, L., & Layton-Matthews, D. (2022). Random forest rock type classification with integration of geochemical and photographic data. *Applied Computing and Geosciences*, *15*. <https://doi.org/10.1016/j.acags.2022.100090>

Thompson, D. J. (2012). Well Logging. *Earth Science: Earth Materials & Resources*, 544.

University of Newcastle Library Guides. Research methods: What are research methods? *LibGuides*. Retrieved October 1, 2022, from <https://libguides.newcastle.edu.au/researchmethods>

Visual Capitalist. (2021, May 20). *Visualizing the copper intensity of renewable energy*. Visual Capitalist. Retrieved September 15, 2022, from https://www.visualcapitalist.com/sp/copper-intensity-of-renewable-energy/

Yiu, T. (2019, June 12). *Understanding Random Forest - Towards Data Science*. Medium. Retrieved October 1, 2022, from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Zhang, Q. (2015). *Modern Models for Learning Large-Scale Highly Skewed Online Advertising Data*. [Master's thesis, University of California, Los Angeles]. <https://escholarship.org/content/qt7mc0k1v8/qt7mc0k1v8_noSplash_68fe9a60ebefb849e7eb26b341d072b0.pdf>

‌