*University of Utah*
UNDERGRADUATE RESEARCH JOURNAL

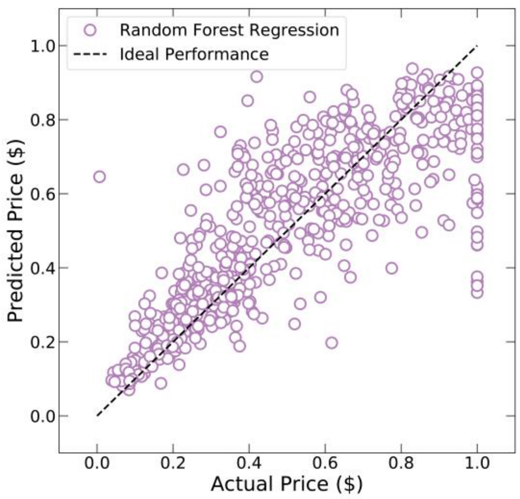
# UTILIZATION OF MACHINE LEARNING TO BUILD A PRICE PREDICTION MODEL FOR RAW MINERALS SOURCING

**Brennan Theler, Kaai Kauwe, Taylor Sparks**
**Department of Materials Science and Engineering**

In the modern world, most machinery and devices require a wide range of raw minerals, many of which are rare or hard to refine. The world minerals market faces a high level of concentration of supply due to geographical and political constraints on mineral deposits. While materials scientists are primarily concerned with the performance of any given material, businesses must take into account the supply and price volatility of any given mineral component of their products. Machine learning is a powerful tool commonly used to forecast outcomes, but the materials science field has not yet utilized this tool to analyze materials flow from an economic perspective on a large scale to create a comprehensive picture of both performance and economic aspects of a material.

Using data gathered primarily from the United States Geological Survey (USGS), the researchers collated over 55,000 datapoints on a large set of materials for the years 1998-2015. Measures collected included mineral identity, mineral price, market concentration using the Hirfindahl-Hirschmann Index (HHI), inflation indexes, transportation prices through oil price, and total production amounts. These data were normalized and placed into a set of machine learning algorithms (random forest, linear regression, and support vector regression) as descriptive datasets, in order to predict the mineral price of the following year.

Of the three algorithms, similar results were produced with r-squared scores just below .7. The best model by a slight amount in repeated tests with random starting seeds was the random forest regression. One example, with an r-squared score of .6846, is showcased here.



The r-squared score given shows predictive ability, but one hampered by a significant amount of error. This stems from two primary sources. First, the normalization method used normalized the maximum price to a value of 1, with all other datapoints being less than 1. This creates the line of data points on the far-right side of the graph. This normalization method distorts the data by creating a set of guaranteed extreme values and is an analyzation artifact rather than an inherent property of the dataset. Work is ongoing to identify and use better normalization methods, such as mean normalization, to improve the model. Second, all three models used are strongly interpolative—near the extremes of price, 0 and 1, the model ceases to predict accurately, consistently overpredicting near 0 and underpredicting near 1. The random forest model minimizes this aspect, out of the three models tested, but work is continuing to attempt to reduce this error via parameter refining to improve extreme value prediction of the models.

In addition, efforts to expand the dataset both in terms of years covered and in number of additional variables explored, in order to eliminate confounding variables and improve the model further, is ongoing.

For access to the dataset used in this work, please contact Assistant Professor Taylor Sparks.

This work was supported by funding from the Summer Program for Undergraduate Research at the University of Utah awarded to Brennan Theler. The authors also wish to thank the USGS for the use of their data and for their interest in a cooperative venture to extend this research into the future.