MINE 432 FINAL PROJECT REPORT

SIDDHANT SONI – 92606482

JAMED AMBROSE – 18569905

JOSIAH TSANG – 74191248

Table of Contents

Table of Figures

Table of Tables

Executive Summary

Within the mining industry, novel easy-to-mine mineral discoveries are becoming less frequent, and cut-off mill feed grades at established projects are reducing (TÜV SÜD Global Risk Consultants, 2018). As such, the importance of testing and sampling of tailings material in minerals is ever-increasing. However, assaying is a notoriously time-intensive process, for which results may take between of 60 minutes to 5 business days per sample. Given this, aberrations in assay values may only be observed using chemical testing methods at low frequencies in comparison to other flotation parameters which may be observed in real time. This is where the need for improvements to be implemented, such as machine learning, to hasten time-consuming processes by providing supplementary insights arises.

The focus of the report involves the use of an open-source data set containing various operating metrics of an iron ore flotation plant (Oliveira, 2017) and follows the construction of a machine learning model based on the data set – which will aim to predict the percent silica in iron concentrate streams.

The report will include an introduction and background of the issues within the mining industry and how machine learning can be a solution, then will follow the construction of the model, consisting of data cleanup to data analysis, leading to the results and future recommendations.

Introduction

As the mining industry's reliance on legacy technologies persists alongside increasing difficulties associated with the discovery of novel orebodies, project economics have suffered (Durrant-Whyte, Geraghty, Pujol, & Sellschop, 2015). The industry also faces disruptions from volatile commodity prices, existing resource depletion, reducing grades, increasing haulage times, and an increasingly challenging macroeconomic environment (Deloitte, 2022). The Global Mining and Metals Industry is also at peak data-literacy and maintains ever-larger amounts of data which requires processing that often still done manually, many being data related operations such as equipment performance, plant operations, and orebody resources and reserves (Chaudhuri, n.d.). To tackle these issues and to enhance inexpensive data-driven mineral engineering applications, innovative technologies machine learning, are becoming increasingly attractive candidates.

What is machine learning? "Machine learning is a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy." (IBM Cloud Education, 2020). Machine learning is the science and art of programming computers so that they can learn from data. Machine learning is special in that there is no need to explicitly state what the decision process is. Instead, the rules to the algorithm are automatically deduced from the sample data given (Caté, 2019).

Machine learning differs from statistics. Statistics is the science of exploring data and gaining insight from the data. Machine learning, on the other hand, is the science of utilizing data to compute a prediction. Commercial-scale applications of such technology include enabling data-driven anticipation of corrective actions to control impurities in the concentrate stream. The use of machine learning in the mining industry will be reviewed by focusing on an open-source dataset containing various operating metrics of an iron ore flotation plant. (Oliveira, 2017).

Background

This study is based on an open-source dataset pertaining to various operating metrics at an iron ore flotation plant. The primary tailings material (impurity) in the feed is silica. The 24-column dataset is a report of metrics such as feed throughput, feed stream iron and silica grades, concentrate stream iron and silica grades, all with timestamps. Some columns were sampled every 20 seconds, others hourly. Some exploratory data analysis was conducted, and following information was collected about key fields in the dataset:

Rows: 737,453

Columns: 24

Column Name	Example	Units	Measurement Frequency (Hz)
Timestamp	2017-03-10 01:00:00	–	1.0
Ore Pulp Flow	395.713	t/h	0.2
Ore Pulp Density	1.74	kg/cm^3	0.2
Ore Pulp pH	10.0672	–	0.2
Feed Si (%)	16.98	%	0.000028
Feed Fe (%)	55.20	%	0.000028
Conc. Si (%)	1.31	%	0.000028
Conc. Fe (%)	66.91	%	0.000028

Table 1: Key Fields in Dataset

As is evident from the measurement frequency values in Table 1, assay measurement reporting frequencies at the subject plant are extremely low in comparison to the reporting frequencies of other operational metrics. As such, there exists an incentive to be able to predict certain assays to anticipate change ahead of confirmatory testing using a predictive model.

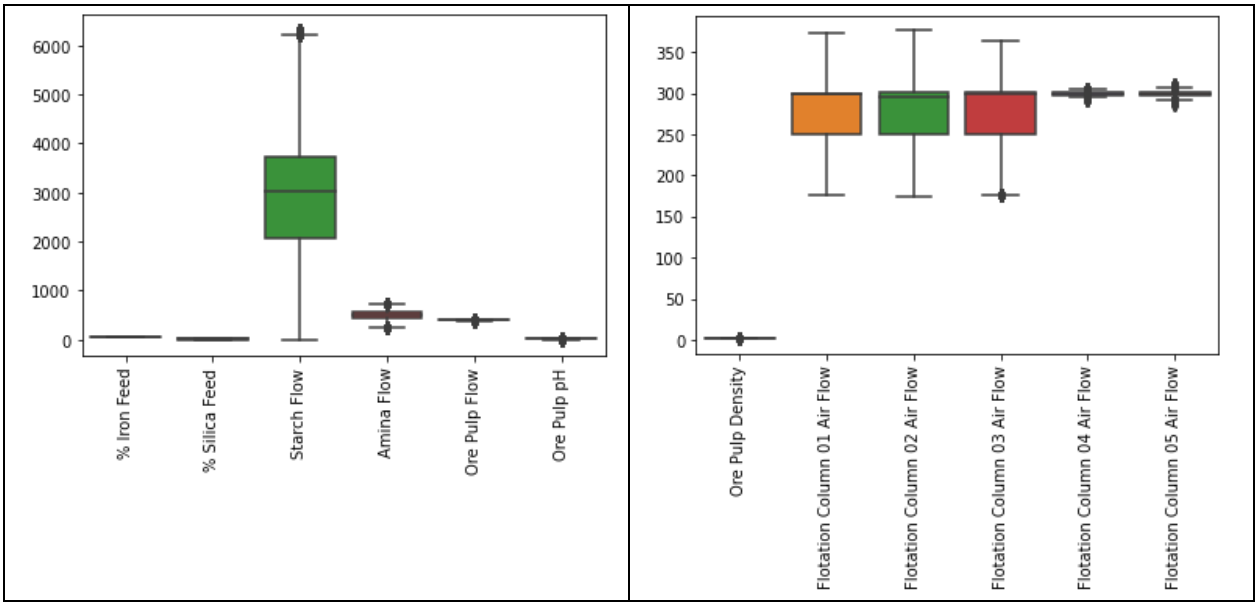
Using existing Python libraries, it was possible to build one such predictive model through data cleaning, analysis, and regression. The process followed, insights drawn, and lessons learned are detailed in this technical report.

Data Cleaning

Exploratory Data Analysis

To analyze and manipulate the dataset, our analysis relies on various pre-existing Python libraries including Seaborn, NumPy, and PyPlot. Seaborn provides a high-level interface for creating attractive and informative statistical visualizations.

First, we looked at outliers in the dataset. An outlier is a data point that is significantly different from the other data points in the dataset which can be caused by errors in the data collection process, or they can be legitimate observations that are simply unusual. To visualize this, box plots of each feature were created as shown in Figure 1.



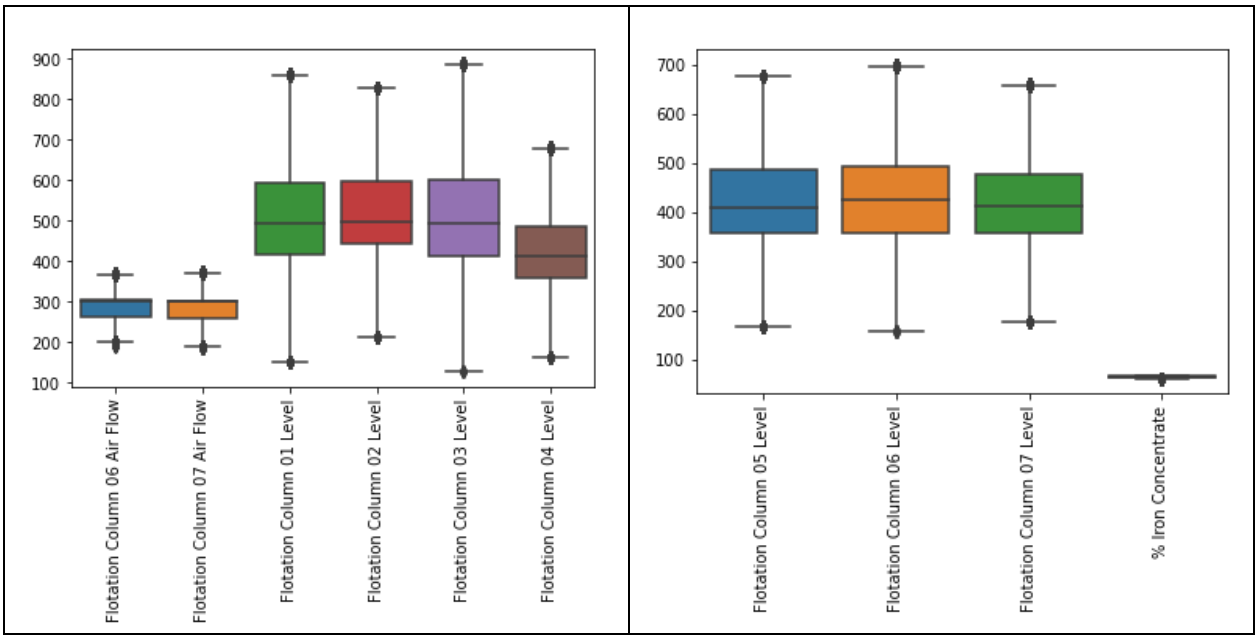


Figure 1: Boxplots Features before Outlier Removal

Based on the boxplots in Figure 1, it was apparent that there are a few outliers in the dataset for each feature, leading removing them from the dataset.

The next step was to check for multicollinearity between features. Collinearity refers to the situation in which two or more predictor variables in a regression model are highly correlated. This can cause problems with the interpretation of the model because it can be difficult to determine the unique effect of each predictor on the outcome variable.

A correlation matrix was plotted to identify features with high collinearity.

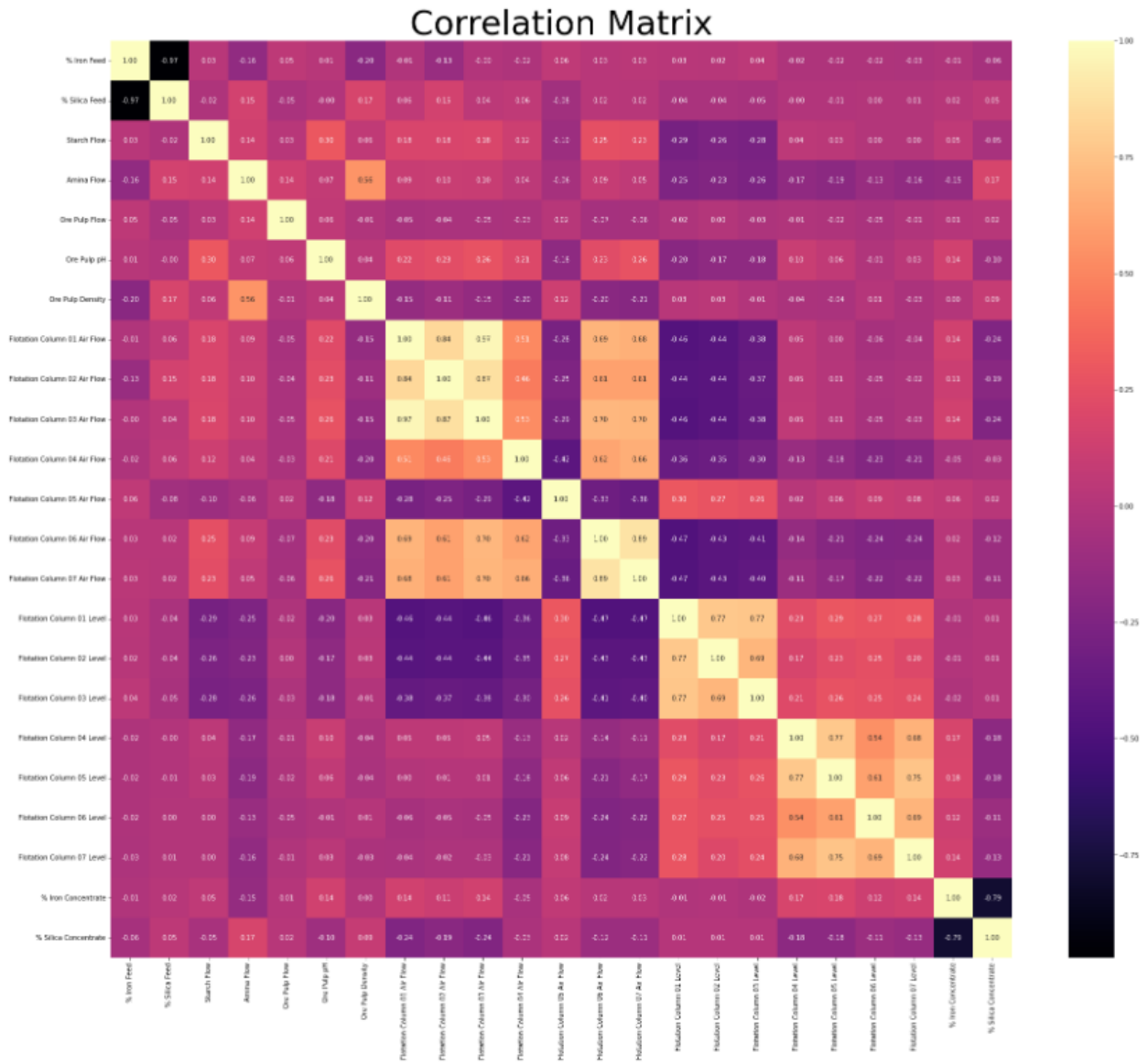


Figure 2: Correlation Matrix before Removing Highly Correlated Features

Due to the high number of features, predicting features that correlate with another predicting features were removed to aid interpretability. In doing so, it was not clear if it influenced model performance.

Preprocessing

The date column was dropped due to its lack of use for the machine learning model.

Outliers

Outliers were removed for each of the input features using the interquartile range (IQR) method. The interquartile range (IQR) is a measure of central tendency that is used to describe a dataset. It is calculated as the difference between the 75th percentile and the 25th percentile of the data. Two parameters were established, Q1 and Q3, denoting the 25th percentile and the 75th percentile of the data respectively. The IQR is defined as IQR = Q3 - Q1.

The IQR is used to identify outliers in the dataset. Any data point that is more than 1.5 times the IQR below the 25th percentile or above the 75th percentile is considered an outlier.

Collinear Features

To remove collinearity one or more of the correlated predictor variables are removed. This is done by identifying one of the correlated predictor variables using the correlation matrix and setting a threshold value. In the event where two pairs of variables share a correlation of 0.7 or greater, one is removed from the dataset.

Data Analysis

The critical output identified for this plant is the Silica Grade (%) in the concentrate stream. This is an output metallurgists may be interested in predictively estimating to anticipate near-term reagent consumptions. This output would be estimated by the Machine Learning model using the inputs that remained following removal of outliers and corrections for multicollinearity.

Following data cleaning, 13 features fed into training the linear regression model of the one output, Silica Grade (%) in the concentrate stream:

% Iron Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 04 Air Flow	Flotation Column 05 Air Flow	Flotation Column 01 Level	Flotation Column 04 Level	Flotation Column 06 Level	% Iron Concentrate	% Silica Concentrate
55.2	3019.53	557.434	395.713	10.0664	1.74	249.214	295.096	306.4	457.396	443.558	446.370	66.91	1.31
55.2	3024.41	563.965	397.383	10.0672	1.74	249.719	295.096	306.4	451.891	448.086	445.922	66.91	1.31
55.2	3043.46	568.054	399.668	10.0680	1.74	249.741	295.096	306.4	451.240	449.688	447.826	66.91	1.31
55.2	3047.36	568.665	397.939	10.0689	1.74	249.917	295.096	306.4	452.441	446.210	437.690	66.91	1.31
55.2	3033.69	558.167	400.254	10.0697	1.74	250.203	295.096	306.4	452.441	453.670	443.682	66.91	1.31

Table 2: Features feeding into linear regression model

Upon removal of the outliers, the following box plots in Figure 3 are plotted, illustrating the removal of all or most of the outliers.

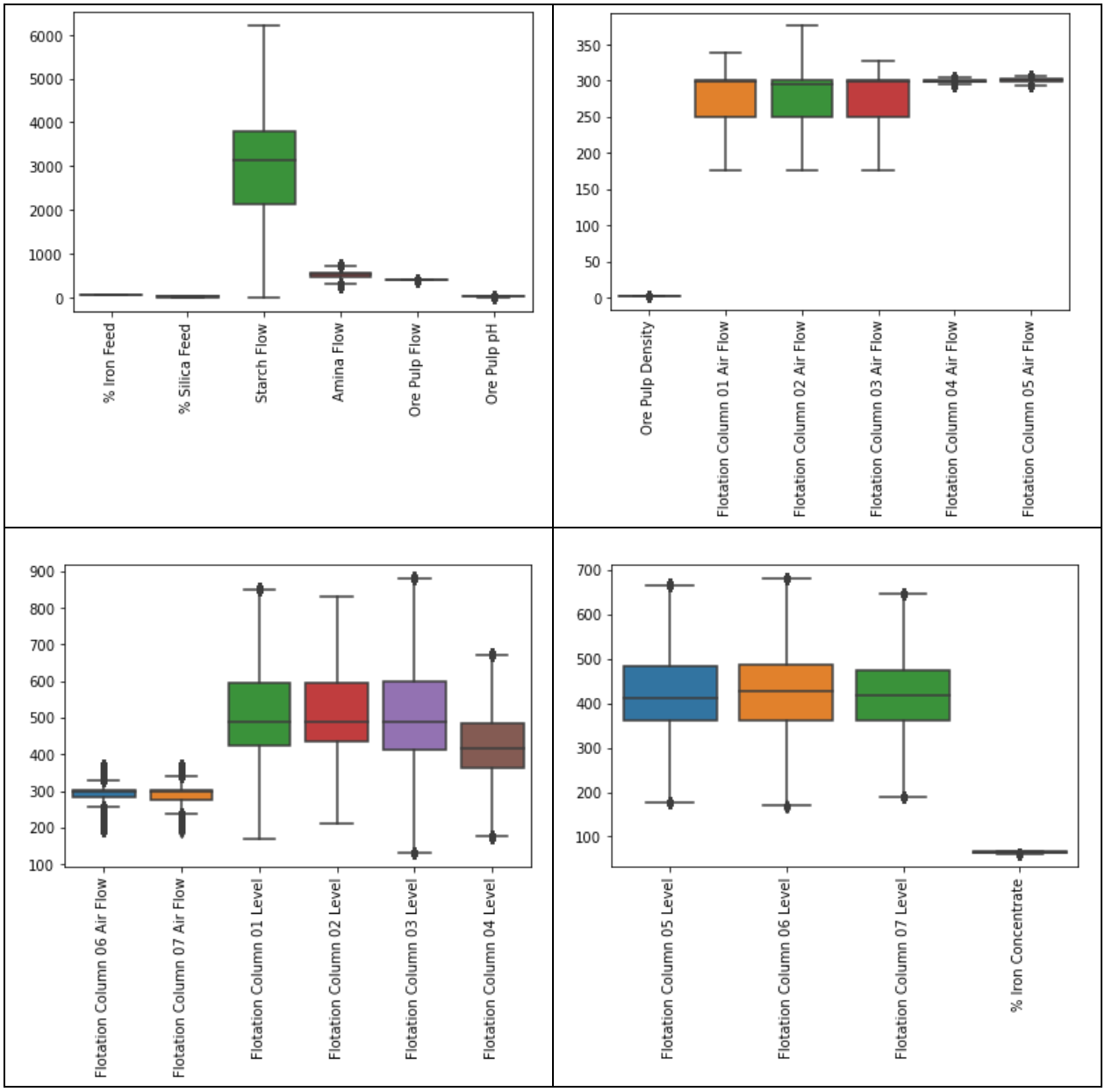


Figure 3: Boxplots Features after Outlier Removal

After removing collinear features, the identified variables were excluded from the dataset to avoid multicollinearity (the correlation of independent variables). The reduced correlation matrix is shown below in Figure 4.

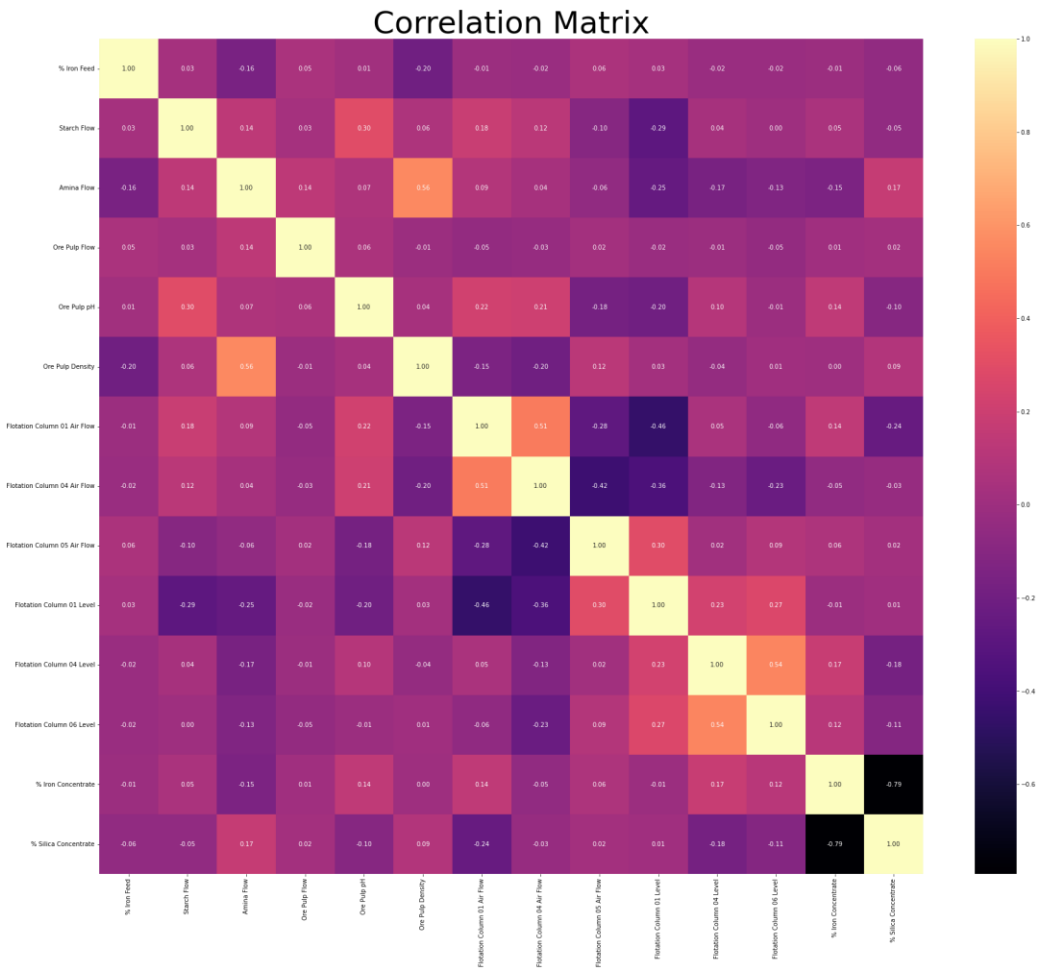


Figure 4: Correlation Matrix after Removing Highly Correlated Features

As shown in Figure 4, the matrix was reduced after removing multicollinearity, making the data, and hence the model more interpretable.

Building the Model

To construct the model, the dataset was split into X and Y arrays, with feature scaling performed on the X arrays. Feature scaling is the process of standardizing or normalizing the range of independent variables or features of a data set. It is a necessary step as the range of values of different features can vary greatly, potentially affecting the performance of the machine learning model. By subtracting the mean value of each feature from each data point, then dividing it by the standard deviation of the feature, it results in a new feature with a mean of 0 and a standard deviation of 1.

Next, the dataset was split into inputs and outputs for a training set and a test set. A training set is a subset to train a model; a test set is a subset to test the trained model. As is standard practice, 80% of the dataset was used for training and the remaining 20% was used to test the model's accuracy to quantify the robustness of our findings.

A linear regression model was then built from scikit-learn using the **LinearRegression** class. Linear regression is a technique where the independent variable has a linear relationship with the dependent variable. The goal is to produce a best fit linear line to fit the model.

Results

The results of our linear regression model are summarized below in Table 3:

R^2 Score	Root Mean Square Error	Average Accuracy
66.23%	69.86%	75.77%

Table 3: Summarized Results

The R^2 score method using the sklearn.metrics library is the most common method of measuring a regression model. It is calculated as the square of the Pearson correlation coefficient between the observed and predicted values of the target variable. The R^2 score ranges from 0 to 1, where a value of 0 indicates a poor fit, and a value of 1 indicates a perfect fit. Thus, a higher R^2 score indicates a better-fitting model.

We also compared each of the model predicted percent silica concentrate to the accepted % silica concentrate and found the average accuracy to be 75.77%.

Discussion

An average accuracy of 75.77% is an acceptable value for the model. While it has shown a clear relationship, there is still a small lack of accuracy that is needed to truly consider it as a proper solution. There are benefits in using a linear regression-based model: it is easy to implement, performs well for linearly separable data, and it is easy to interpret and train. (Waseem, 2022). However, as seen with the average accuracy, it is not perfect, especially within this scenario. The main issue with linear regression is that not all data can be captured linearly. Other issues include that it is sensitive to outliers and assumes that data is independent. (Waseem, 2022). These

disadvantages, however, were taken into consideration, as both outliers and any multicollinearity were removed prior to testing.

Conclusion and Recommendations

As the mining industry faces growing challenges such as volatile commodity prices along with the usage of legacy technology and methods, machine learning was tested as a potential solution using an example dataset of an iron ore flotation plant. Using python, a machine learning model was created to estimate the Silica grade in Iron concentrate at various datapoints, allowing early detection to reduce impurities or aid the environment.

To achieve this, unnecessary data and outliers were removed, along with correlated columns to avoid multicollinearity. Using the linear regression model, the results showed an average accuracy of 75.77%. As there are some shortcomings of the linear regression model, the model has room to improve and grow.

In our study of the given dataset, other regression models were fit and tested, but with varying degrees of success.

Overall machine learning is a highly useful tool that can be the solution for many of the issues that the mining industry is currently facing. Creating a more in-depth model would likely produce more reliable results. Machine learning models such as the one made in this report, can be used in a multitude of other fields withing the industry, evolving it to compete with the growing global demand of materials.

References

Caté, A. (2019, December). Retrieved from Machine Learning And Artificial Intelligence For Mining Geoscience: https://www.srk.com/en/publications/machine-learning-and-artificial-intelligence-for-mining-geoscience

Chaudhuri, S. (n.d.). Retrieved from Driving Insight from Data in Mining Industry: https://www.wipro.com/natural-resources/driving-insight-from-data-in-mining-industry/#:~:text=It%20is%20well%20known%20that,to%20name%20a%20few%20elements.

Deloitte. (2022). Retrieved from Future of mining with AI: Building the first steps towards an insight-driven organization: https://www.deloitte.com/content/dam/assets-shared/legacy/docs/deloitte-norcat-future-mining-with-ai-web.pdf

Durrant-Whyte, H., Geraghty, R., Pujol, F., & Sellschop, R. (2015, November 1). Retrieved from How digital innovation can improve mining productivity: https://www.mckinsey.com/industries/metals-and-mining/our-insights/how-digital-innovation-can-improve-mining-productivity

IBM Cloud Education. (2020, July 15). Retrieved from Machine Learning: https://www.ibm.com/cloud/learn/machine-learning

Oliveira, E. M. (2017). Retrieved from Quality Prediction in a Mining Process: https://www.kaggle.com/datasets/edumagalhaes/quality-prediction-in-a-mining-process

TÜV SÜD Global Risk Consultants. (2018, March 13). Retrieved from Meeting the Fast-Changing Challenges of the Mining Industry: https://riskandinsurance.com/meeting-fast-changing-challenges-mining-industry/

Waseem, M. (2022, January 1). Retrieved from How To Implement Linear Regression for Machine Learning?: https://www.edureka.co/blog/linear-regression-for-machine-learning/

Appendix (Jupiter Notebook Code)