Industrial Internship Report on

” Quality Prediction in Mining Process”

Prepared by

Naincy Naiyar

Executive Summary
This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT). This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time. My project was Quality Prediction in a Mining Process. The main Objective of this project is to Explore real industrial data and help manufacturing plants to be more efficient. This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1 Preface

The mining sector holds significant importance in extracting and processing valuable resources. To enhance operational efficiency and ensure quality control in mining processes, the development of precise predictive models is crucial. This project endeavors to investigate and establish a quality prediction system specifically designed for mining operations, with a primary focus on accurately predicting the percentage of silica concentrate. By obtaining timely information regarding impurity levels, engineers can take proactive measures to mitigate impurity and minimize the volume of ore that is directed to tailings. The dataset utilized in this project originates from an authentic flotation plant, encompassing diverse process variables and quality metrics that are meticulously recorded.

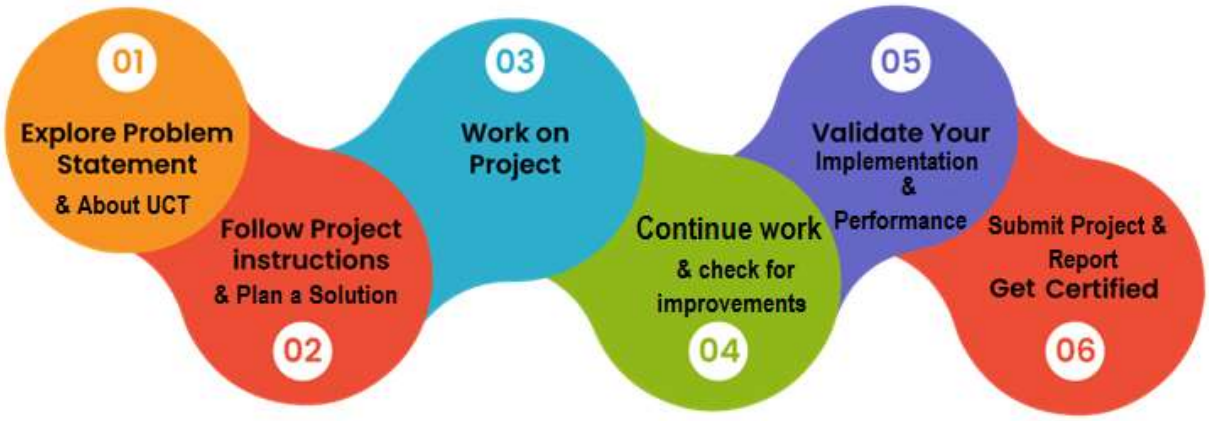
In this 6-week Internship, I learnt to apply Machine Learning Models in a real Database. This internship offers a transformative and immersive learning opportunity for individuals aspiring to pursue careers in these fields. Participants undergo an intensive program that equips them with crucial skills and knowledge essential for success in the industry. Starting with an orientation to introduce fundamental concepts and programming languages like Python, the internship then progresses to cover data exploration, visualization, and the foundations of machine learning, including career opportunities. Real-world projects enable interns to apply their knowledge to practical scenarios like I applied random forests, Regression using Lasso and ridge in my project. While mentors provide valuable guidance and feedback.

ML & DS Internship in career development plays a huge role. Unlocking a successful career in the rapidly evolving fields of Data Science and ML begins with essential internships. These invaluable opportunities offer hands-on experience, allowing interns to apply theoretical knowledge to real-world projects and sharpen their technical skills, including proficiency in Python, machine learning algorithms, and data visualization. By curating a portfolio of completed projects, aspiring professionals gain a competitive edge in the job market. Additionally, networking with industry experts opens doors to potential job opportunities and provides valuable guidance. Moreover, exposure to industry-leading tools and technologies enhances practical understanding, ensuring interns are job-ready and well-prepared for future success.

This Project aims to use the given data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions (empowering!). Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and help the environment (reducing the number of ore that goes to tailings as you reduce silica in the ore concentrate). It is not always easy to find databases from real world manufacturing plants, especially mining plants. Also, this database comes from one of the most important parts of a mining process: a flotation plant.

This Opportunity given by USC/UCT gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship. This experience concludes with project presentations and reflections, leaving participants with a strong foundation, a portfolio, and a valuable network to launch their careers as accomplished data scientists and machine learning practitioners.

Program was planned nicely in a 6-week schedule.



This DS & ML internship provides a well-rounded learning journey that includes developing technical expertise, problem-solving capabilities, fostering teamwork and effective communication, gaining insights into industry practices, and creating networking possibilities guided by experienced professionals, interns gain industry exposure and build a portfolio showcasing their achievements. Networking opportunities may lead to future job prospects. Overall, the internship serves as a crucial stepping stone, equipping individuals with the necessary knowledge and experience to excel in the dynamic fields of data science and machine learning.

Thank to UniConverge Technologies Pvt Ltd , upskill Campus and everyone related to this team for such an amazing opportunity.

For a newbie, this is a perfect platform to learn and apply on real world Problems. I must suggest my juniors and peers to go through this once. It is an amazing opportunity, just grab it!

2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

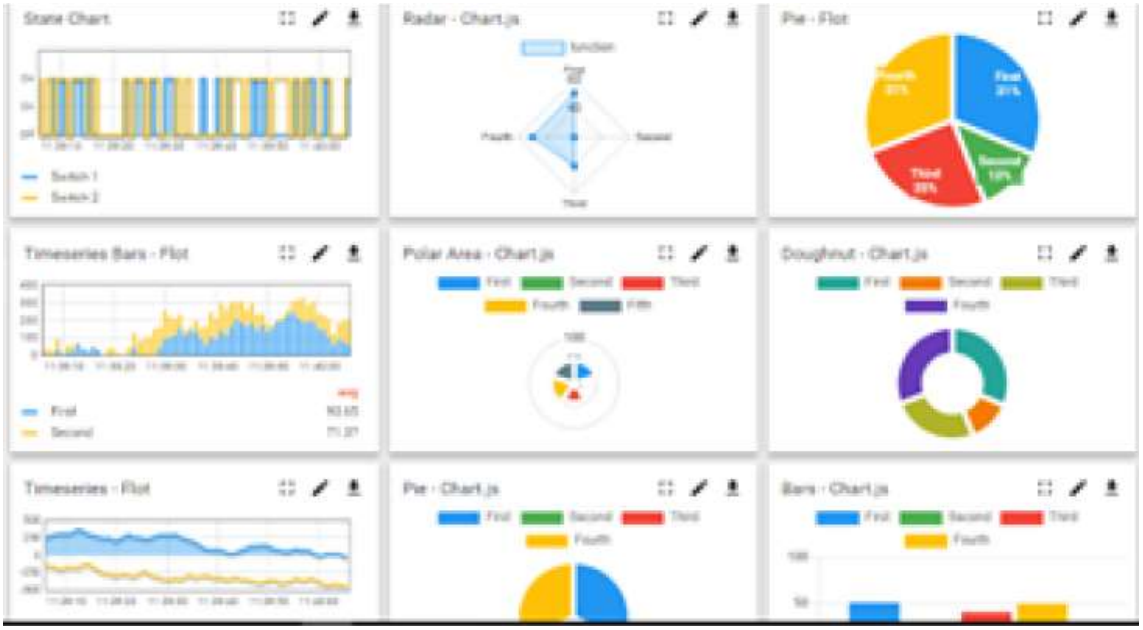
i. UCT IoT Platform (UCT Insight)

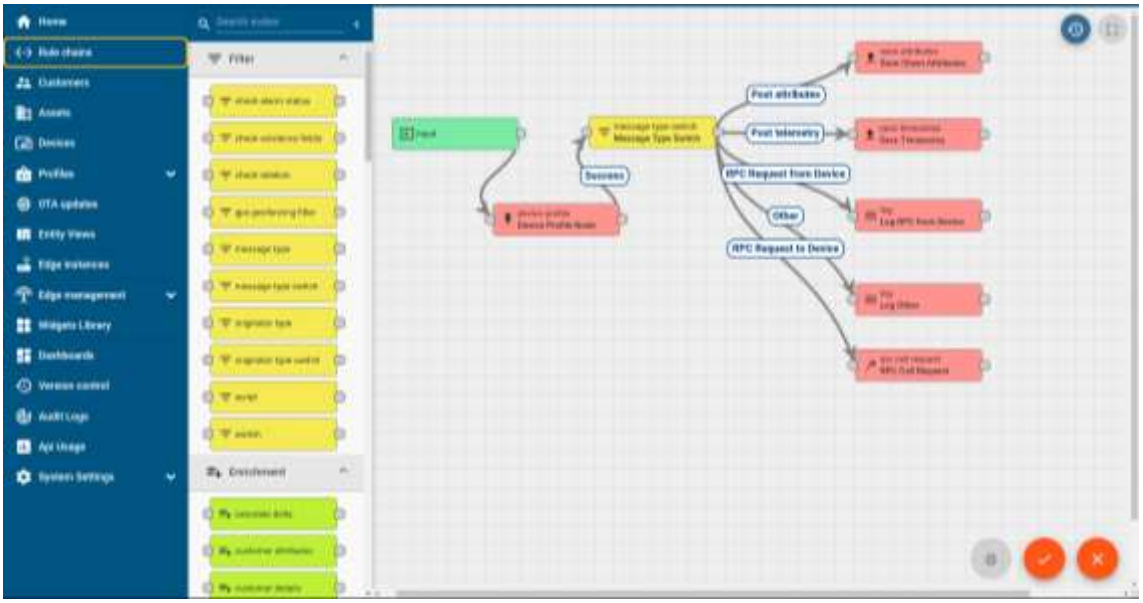
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application(Power BI, SAP, ERP)
- Rule Engine





ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

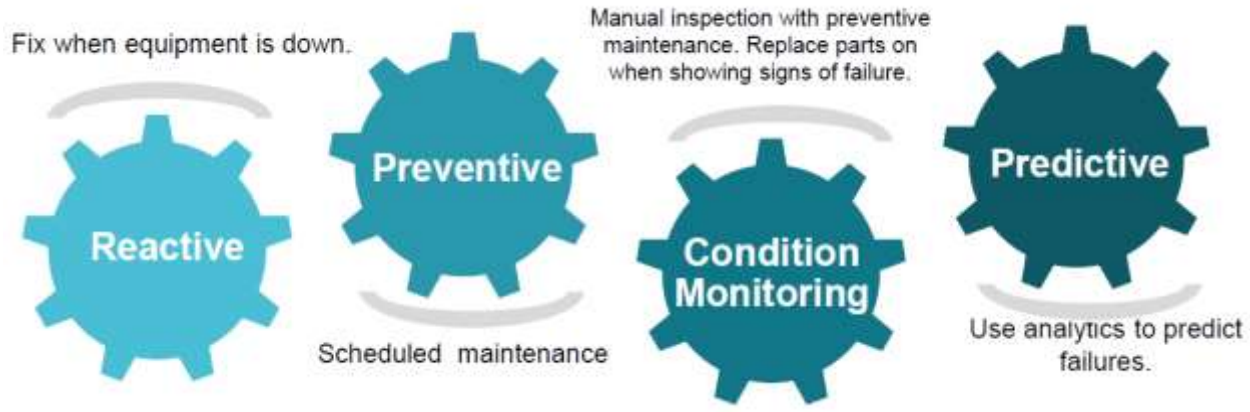
Its unique SaaS model helps users to save time, cost and money.

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

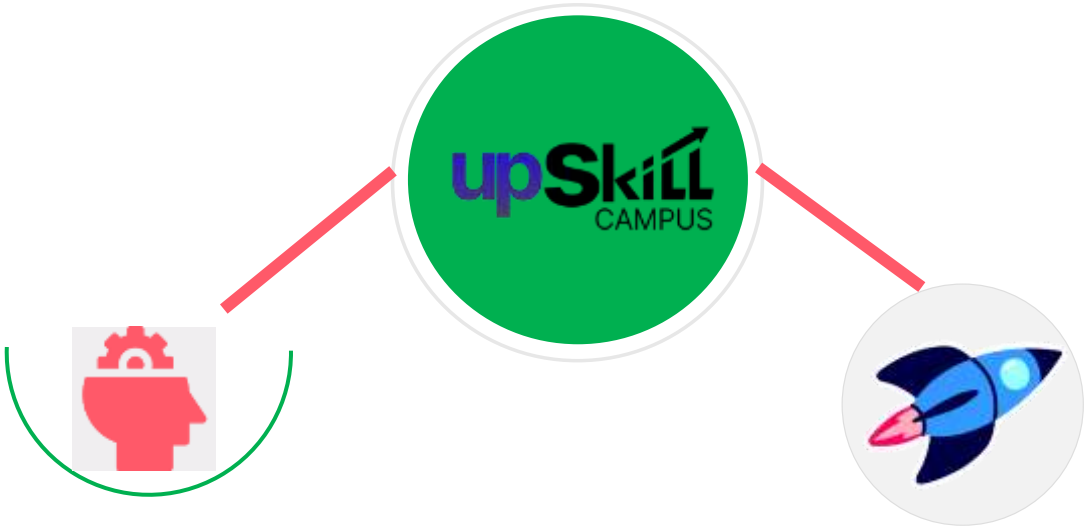
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

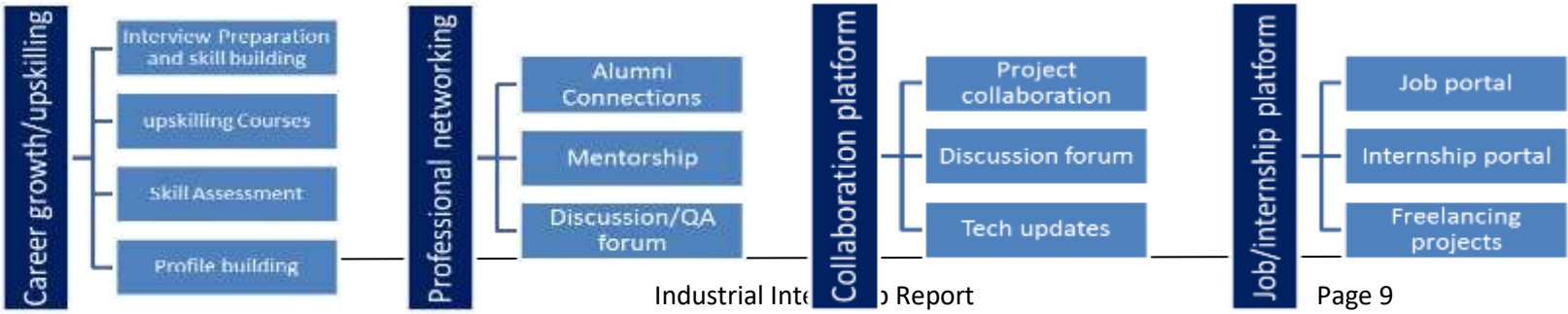
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

- ☛ get practical experience of working in the industry.
- ☛ to solve real world problems.
- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.5 Reference

[1] https://www.geeksforgeeks.org/

[2] https://scholar.google.com/

2.6 Glossary

Terms	Acronym
Data Science	DS
Machine learning	ML
Residual Sum of Squares	RSS
Root Mean Square Error	RMSE

3 Problem Statement

The "Quality Prediction in a Mining Process" problem about Industrial Manufacturing and Production. It revolves around predicting the quality of final products in mining operations. It aims to Explore real industrial data and help manufacturing plants to be more efficient Variations in geological formations, mining techniques, and processing conditions can cause differences in raw material composition and quality. The objective is to develop a predictive model that analyzes various parameters and data points collected during mining and processing to accurately forecast the quality attributes of the final products. Implementing this predictive model can optimize production, reduce waste, and ensure consistent product quality, leading to improved efficiency, cost reduction, and increased customer satisfaction in the mining industry.

It is not always easy to find databases from real world manufacturing plants, especially mining plants. This database comes from one of the most important parts of a mining process: a flotation plant.

The main goal is to use this data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions (empowering!). Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and also help the environment (reducing the amount of ore that goes to tailings as you reduce silica in the ore concentrate).

4 Existing and Proposed solution

The mining sector is crucial in the extraction and processing of valuable materials required by numerous economic sectors. Accurate parameter prediction is crucial in mining processes in order to maximise operating effectiveness and guarantee product quality. Mining engineers and operators may make educated judgements and act promptly to maintain and improve product quality by using quality prediction as a proactive strategy.

The goal of this research is to create a quality prediction system that is especially suited for a mining procedure. Predicting the percentage of silica concentrate, a crucial quality indicator in the mining business, is the main goal of our work. Engineers may improve the quality of their products and cut down on waste by precisely estimating the silica content and identifying potential problems early on.

We use a data-driven strategy and machine learning tools to accomplish our goals. The study makes use of a sizable dataset gathered from a flotation facility in the real world, which captures a variety of process variables and quality measures. We investigate the data to learn more about the connections between various variables and how they affect silica concentration. The goal of our research is to create a predictive model that can accurately predict the silica concentration during the mining process through data preprocessing, feature selection, and model training.

In-depth data exploration is used in the process to comprehend the dataset's features and spot any patterns or anomalies. To deal with missing values, outliers, and make sure the data is appropriate for modelling, preprocessing techniques are used. To determine the most pertinent features, feature selection techniques are used.

I compared Linear Regression Model with Regularization Term (Lassso Model), Linear Regression Model with Regularization Term (Ridge Model) & Random Forest to best find the perfect model depending upon its R square and error values.

4.1 Code submission (Github link)

https://github.com/naincy-n/Quality_Prediction_in_a_Mining_Process

5 Proposed Design/ Model

Here are the steps for my Model –

5.1 Data Acquisition and Exploration

The dataset is about a flotation plant which is a process used to concentrate the iron ore. This process is very common in a mining plant. In content, the first column shows time and date range (from march of 2017 until September of 2017). Some columns were sampled every 20 second. Others were sampled on an hourly base. The second and third columns are quality measures of the iron ore pulp right before it is fed into the flotation plant. Column 4 until column 8 are the most important variables that impact in the ore quality in the end of the process. From column 9 until column 22, we can see process data (level and air flow inside the flotation columns, which also impact in ore quality. The last two columns are the final iron ore pulp quality measurement from the lab. Target is to predict the last column, which is the % of silica in the iron ore concentrate.

I am finding answer ton these questions through this project-

- Is it possible to predict % Silica Concentrate every minute?
- How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.
- Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?

Here, after receiving the dataset, a preliminary investigation was carried out to learn more about its characteristics and structure. The quantity, format, and overall dispersion of the dataset were evaluated during this exploration. We counted the samples (number of instances) and variables (number of features) in the dataset.
If any missing values existed in the dataset, they were found during the initial exploration. Missing values can happen for a number of reasons, including incorrect data capture or damaged sensors. It was important to address these missing values before further analysis and modeling to ensure the integrity and accuracy of the results.

In addition to addressing missing values, exploratory data analysis techniques were employed to visualize the data and identify any patterns or relationships. Visualization methods such as histograms, scatter plots, and correlation matrices were utilized to understand the distributions of variables, detect outliers, and assess the interdependencies between variables.

5.2 Data Preprocessing

The preparation of the dataset for modelling is greatly aided by data preprocessing. The data was preprocessed in the ways listed below to ensure its quality and compatibility for the mining process' quality prediction task. It includes-

- Handling Missing Values.
- Outlier Detection.
- Data Transformation.

5.3 Feature Selection

Correlation analysis and domain expertise were combined in the feature selection procedure for the quality prediction project for a mining process. The most pertinent and instructive properties for the prediction job were found using both methods.

Based on their association with the target variable and their significance in the mining process, the features for the quality prediction in a mining process project were chosen. These characteristics comprised the iron ore pulp quality measurements, process factors associated with the flotation plant's operational circumstances, and specific final product measurements derived from laboratory testing. Chemical composition parameters, impurity levels, flotation column parameters (level, air flow), and other pertinent process control variables are a few examples of selected features.

5.4 Model Selection and Training

Features- (737453, 8)

Working with following models -

- **Linear Regression Model with Regularization Term (Lasso Model)-**

The Lasso model is a type of linear regression model that incorporates a regularization term to prevent overfitting and improve the model's performance. It introduces a penalty term based on the absolute values of the coefficients, which encourages the model to select a subset of features and produce sparse solutions. By adding this regularization term, the Lasso model helps control the complexity of the model and avoids excessive reliance on irrelevant or noisy features. The strength of the regularization is controlled by a parameter called lambda (λ), which determines the trade-off between fitting the training data and the magnitude of the coefficients. The Lasso model is particularly useful when dealing with high-dimensional datasets where feature selection is important. It automatically performs feature selection by shrinking the coefficients of less relevant features towards zero, effectively eliminating them from the model. To fit the Lasso model, various optimization algorithms can be used, such as coordinate descent or least angle regression. These algorithms iteratively update the coefficients to minimize the sum of squared errors along with the regularization term.
In summary, the Lasso model extends linear regression by incorporating a regularization term to prevent overfitting and promote feature selection. It is a valuable tool for building parsimonious models that can handle high-dimensional datasets and improve generalization performance.

- **Linear Regression Model with Regularization Term (Ridge Model)-**

The Ridge model, also known as Ridge regression, is a linear regression model that includes a regularization term to improve its performance and address the issue of multicollinearity in the data. It adds a penalty term based on the squared magnitude of the coefficients to the ordinary least squares objective function. The regularization term in the Ridge model helps control the complexity of the model by shrinking the coefficients towards zero. This helps reduce the impact of highly correlated features and prevents overfitting. The strength of the regularization is controlled by a parameter called alpha (α), which determines the trade-off between fitting the training data and the magnitude of the coefficients. By adding the regularization term, the Ridge model provides a more stable solution, especially when dealing with datasets that have multicollinearity, where predictor variables are highly correlated. It helps prevent the model from assigning excessive importance to any one feature and balances the influence of multiple correlated features. To fit the Ridge model, various optimization algorithms can be used, such as gradient descent or closed-form solutions. These algorithms adjust the coefficients iteratively to minimize the sum of squared errors and the regularization term.

In summary, the Ridge model extends linear regression by incorporating a regularization term that addresses multicollinearity and improves model stability. It is effective in situations where there are highly correlated features and provides a balanced and robust solution for regression problems.

- **Random Forest Model-**

The Random Forest model is a versatile and powerful machine learning technique that may be utilised for regression as well as classification applications. It is an ensemble learning method that makes predictions by combining numerous decision trees. The Random Forest approach generates a set of decision trees, each of which is trained on a different sample of the data and employs a random subset of the characteristics. During training, each decision tree is constructed individually and makes predictions based on the majority vote or the average prediction of all trees in the forest. The Random Forest model's key advantages include its ability to handle high-dimensional data, numerical and categorical features, and missing values without requiring costly data preprocessing.

6 Performance Test

While testing, here are the main constraints needed for this project-

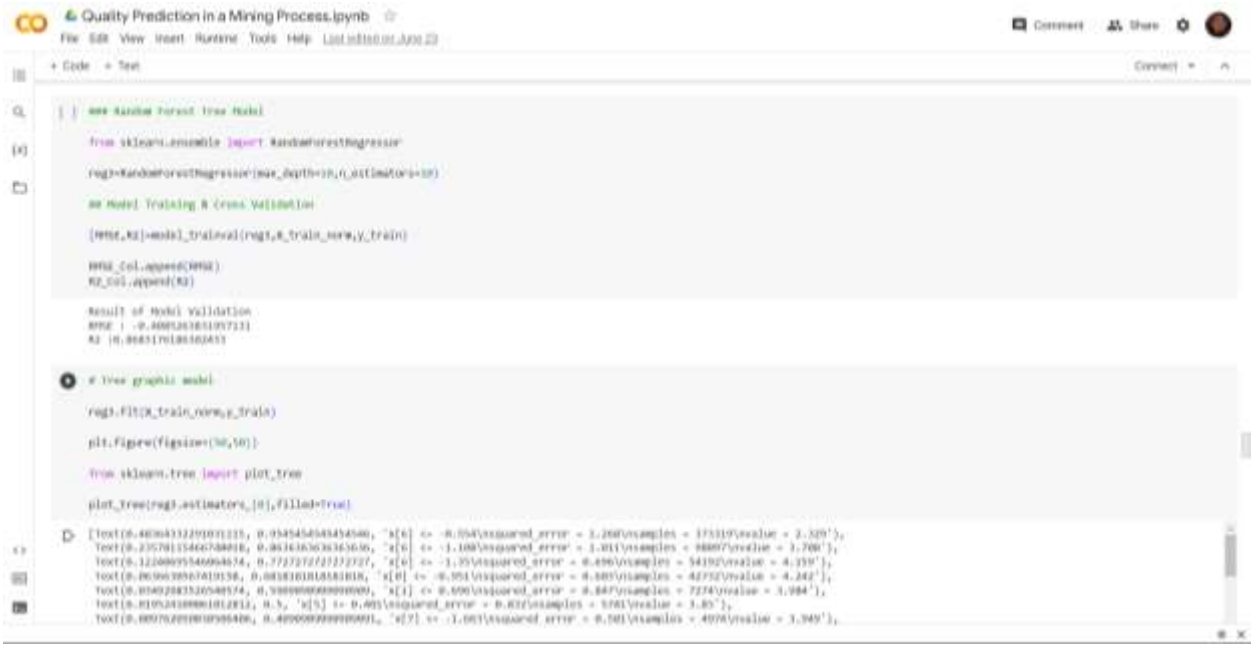
- Memory Constraint: Implement data compression and efficient model architectures to handle large datasets with limited memory.
- MIPS and Speed Constraint: Optimize algorithms and use hardware accelerators to enhance prediction speed.
- Accuracy Constraint: Thorough model evaluation, hyperparameter tuning, and ensemble techniques to improve prediction accuracy.
- Durability Constraint: Deploy the model on durable hardware and consider ruggedized components for harsh mining environments.
- Power Consumption Constraint: Optimize the model to reduce power-intensive operations and explore energy-efficient hardware solutions.
- Data Collection Constraints: Develop robust data collection protocols and use data augmentation techniques to address data scarcity.
- Environmental Constraints: Preprocess data to remove noise and use sensors with appropriate filtering to mitigate environmental interference.

6.1 Test Plan/ Test Cases/Procedure

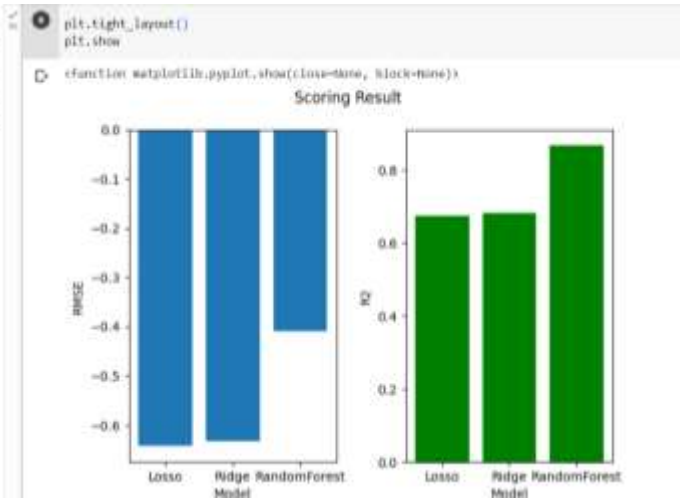
After Model Evaluation, I have taken RMSE and RSS using Linear Regression Model with Regularization Term (Lassso Model) , Linear Regression Model with Regularization Term (Ridge Model) and Random Forest Model.

Based on valuation, I selected Random Forest Model and plotted its Graphic model –

- I try to reduce n_estimators to reduce time computation



- Created Graph based on its Scoring Result and selected model.



- The Random Forest Model is selected as the prediction model with Root Mean Square error at 0.181 with R Square at 0.857

- The % iron concentrate is still be used as feature because other features are not related with the label significantly, so we are still had to sample to check lab result.

6.2 Performance Outcome

- It possible to predict % Silica Concentrate every minute.
- Compared the predicted % silica concentration with test set as below table-

index	% silica concentrate	predicted % silica concentrate
0	198870	3.540000	4.123465
1	18768	1.300000	1.686601
2	19259	2.080000	2.068832
3	616058	1.450000	1.434063
4	5905	2.970000	2.885636
5	292500	1.180000	1.118109
6	278272	1.570000	1.930445
7	605221	5.520000	4.010870
8	559308	2.920000	2.177462
9	132395	4.927199	4.719341
10	133474	4.501191	4.749818
11	32119	4.280000	3.776531
12	350034	2.180000	2.498830
13	229427	5.050000	4.491923
14	682341	1.770000	2.163193
15	161206	3.330000	2.477250
16	658923	1.680000	2.045549
17	580514	4.913343	4.817483

18	603707	1.410000	1.988853
19	73638	1.180000	1.678876
20	396995	2.190000	1.794006
21	239528	1.710000	1.858234
22	692316	5.060000	3.975011
23	22372	1.010000	1.206761
24	536822	4.690971	4.667929
25	649612	1.030000	1.441185
26	563449	2.180000	1.870806
27	583805	3.654036	3.718688
28	293832	1.170000	1.215713
29	654185	1.290000	2.155489
30	721533	2.590000	1.835059
31	162382	1.260000	1.286851
32	495275	1.460000	1.787588
33	254121	1.330000	1.489799
34	429096	2.500000	2.195220
35	707339	1.090000	1.872669
36	155091	4.919427	4.575431
37	708849	3.710000	4.231214
38	657251	1.260000	1.296690
39	698829	4.906676	3.601351
40	285506	1.020000	1.215713
41	90220	2.590000	2.658878

42	414611	1.430000	1.536191
43	210548	4.150000	4.192169
44	267720	1.900000	2.050101
45	520913	2.140000	2.137742
46	349953	2.180000	2.257643
47	717014	4.453802	4.160822
48	702694	1.240000	1.124935
49	99872	2.800000	2.523322

The table provided shows a comparison between the predicted % silica concentrate and the actual test set values. Each row represents a specific data point, and the three columns display the index, the actual % silica concentrate values, and the corresponding predicted % silica concentrate values.

From the table, we can observe that the predicted values are generally close to the actual values. However, there are some instances where the predicted values deviate from the actual values. This difference may be attributed to the inherent complexity of the mining process and various factors affecting the quality of the final product.

To assess the performance of the predictive model comprehensively, additional evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) could be calculated. These metrics provide a quantitative measure of how well the model's predictions match the actual values. A lower MAE and MSE and a higher R2 value indicate better model performance.

Also, To determine the How many steps (hours) ahead can we predict % Silica in Concentrate, a time series analysis needs to be performed on the data. This involves examining the time intervals between data points and identifying the appropriate lag (time steps) at which the % Silica in Concentrate shows significant correlation with its past values.

Regarding predicting % Silica in Concentrate without using the % Iron Concentrate column, it is possible to explore other features that might be correlated with % Silica. Feature selection techniques, such as correlation analysis, mutual information, or backward/forward selection,

can help identify relevant features that contribute to the prediction. By using other relevant variables, the model may still be able to make accurate predictions for % Silica in Concentrate without relying on % Iron Concentrate, provided those features capture meaningful information related to % Silica in Concentrate variations.

7 My learnings

In this Data Science and Machine Learning project on Quality Prediction of Mining Process, I have learned valuable insights into optimizing the efficiency of industrial mining operations. Through meticulous data collection and preprocessing, we tackled challenges related to missing values and outliers, ensuring the dataset's reliability. Feature engineering allowed us to identify key variables impacting ore concentrate quality. Model selection and evaluation enabled us to identify the best-suited predictive algorithms, while handling correlated features required careful consideration. Leveraging time-series analysis, we accounted for temporal dependencies in the data for better future predictions. Data visualization played a crucial role in gaining valuable insights from the dataset. By accurately predicting silica concentrate levels, engineers can now take proactive measures to minimize impurities and reduce waste, positively impacting both the environment and operational efficiency. Continuous improvement and collaboration among stakeholders were key components of this successful project, enabling better decision-making and optimization in the mining process.

8 Future work scope

The Quality Prediction of Mining Process project holds great promise for future advancements and practical applications in the mining sector. The project's potential lies in exploring advanced machine learning techniques, including deep learning algorithms like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to improve predictive accuracy for silica concentrate levels. Real-time predictions through continuous data monitoring can empower engineers to make timely decisions and optimize mining processes, while predictive maintenance can reduce downtime and enhance operational efficiency. Additionally, anomaly detection and ensemble techniques offer robustness to the model's predictions. Integrating data from various sources and collaborating with multiple mining plants can improve the model's generalizability and industry-wide adoption. Furthermore, the project can extend its scope to analyze the environmental impact of mining operations and promote sustainable practices. Through continuous innovation and data-driven decision-making, the Quality Prediction of Mining Process project can revolutionize the mining industry and contribute to more efficient and eco-friendly mining practices.