Predicting Critical Elements in Coal Mine Waste:

A Machine Learning Approach for a Low-Emission Future

Project Presentation

Evan Ginting
Yuhao Long

INTRODUCTION

Critical elements are **vital** for modern tech, economies, and security, but their **supply chains are vulnerable** to political, geographical, and environmental factors.

Our project is particularly interested in predicting **REE** (one of critical elements), and its subdivisions: **HREE** & **LREE**.

REE (Rare Earth Elements) consist of 17 lanthanide series. HREE is Heavier and less common REE. LREE is Lighter and more abundant REE.

In the last few years, **coal** has been identified as a **potential source** of critical elements.

Utilising **machine learning** to **predict the quantity of REE, HREE, and LREE** in coal mine waste.

Photo by Vardan Papikyan on Unsplash

Photo by Peter Pryharski on Unsplash

Photo by shraga kopstein on Unsplash

DATA INTRODUCTION...(1)

Data Source

4 data source

1 ACARP project C29030

3 Confidential Data

Sample Data

Coal Sample ID (Total: 252)

Elements (Total: 49)

Element	Symbol	CBLA-01	CBLN-02	CBLN-O1	CLBOW-1	CPOTTS-1	SCCDE-1
Lithium	Li	30	12	40	13	47	79
Beryllium	Be	1	0,7	1	1	2	2
Aluminium	Al	45000	19000	96000	13000	43000	57000
Strontium	Sr	16	32	310	97	35	240
Scandium	Sc	6,2	4,1	34	3,1	6,7	7,4
Vanadium	V	28	15	310	18	41	46
Chromium	Cr	30	17	120	15	22	26
Manganese	Mn	43	19	520	62	73	27
Iron	Fe	16000	1700	49000	3700	17000	6700
Cobalt	Co	11	81	49	7	130	63
Nickel	Ni	5	5	72	2	13	23

README | Sample coordinates | Collinsville | Newlands | Coppabella

Concentration value (Parts per Million - ppm)

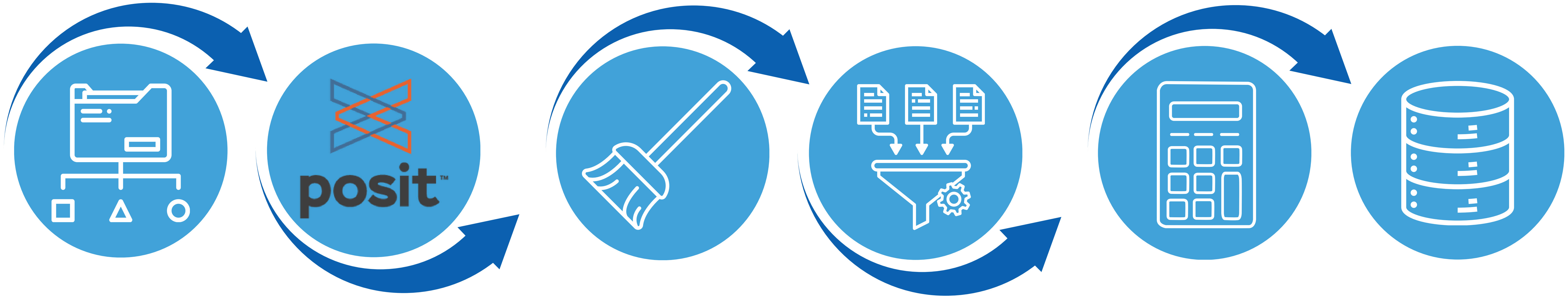
Project Area (Total: 17)

Final Data

Project_Name	Sample_ID	Ba	Ce	Co	Cr	Cs	Cu
Collingwood Park	CP-013	0.008	0.04	14.1	8.0	12	0.23
Collingwood Park	CP-014	0.004	0.22	9.6	5.0	23	0.70
Collinsville	CBLA-01	0.003	0.05	41.8	11.0	30	0.72
Collinsville	CBLN-02	0.004	0.05	26.6	56.0	17	0.36
Collinsville	CBLN-O1	0.005	0.13	25.7	56.0	120	2.10
Collinsville	CLBOW-1	88	30.9	7.0	15	*NA*	2.0
Collinsville	CPOTTS-1	130	31.2	56.0	22	1.50	17.0
Collinsville	SCCDE-1	190	84.2	56.0	26	1.40	21.0

DATA INTRODUCTION...(2)

Data Preparation



Data Sources	Data Upload	Data Cleaning & Transforming	Join Data	Calculate REE, HREE, LREE	Tidy Data
• All .csv format	• Cloud server • Workspaces	• Data type conversion • Long to wide format	• Join multiple datasets	• Perform addition for multiple elements that constructed REE, HREE, and LREE	• Ready to use

DATA SCOPE

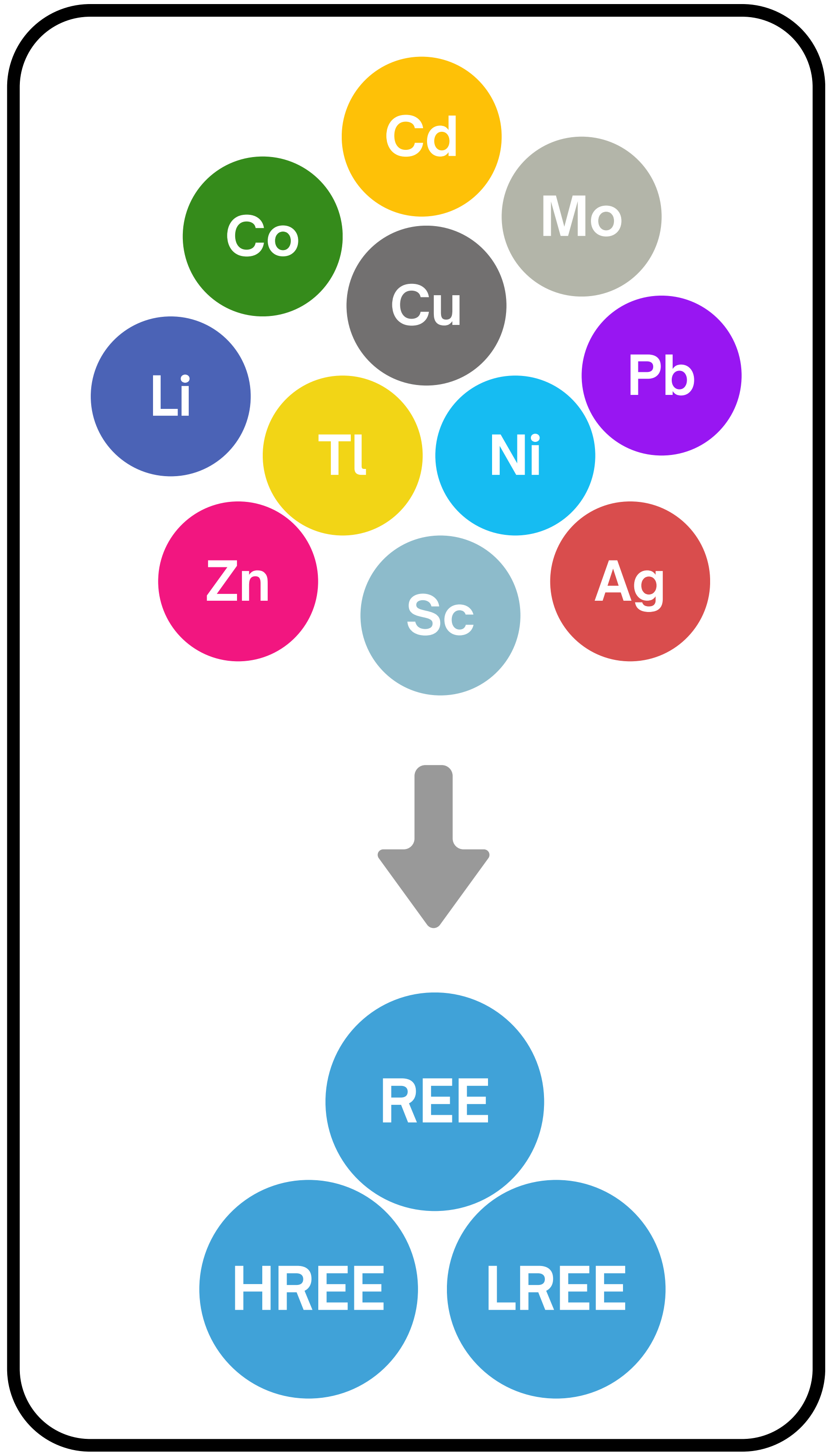
Coal waste samples undergo a test to determine concentration values. Our samples were tested between two test:

- **Test A** (ME-4ACD81): **$9.96 per sample**
- **Test B** (ME-MS81): **$35.84 per sample** (**3,6x** more expensive!)

Objective: Predict REE, HREE, and LREE concentrations using the results from Test A.

Test A measures the following 11 elements:

- Cadmium (Cd),
- Molybdenum (Mo),
- Lead (Pb),
- Silver (Ag),
- Scandium (Sc),
- Zinc (Zn)
- Lithium (Li),
- Cobalt (Co),
- Copper (Cu),
- Nickel (Ni),
- Thallium (Tl).

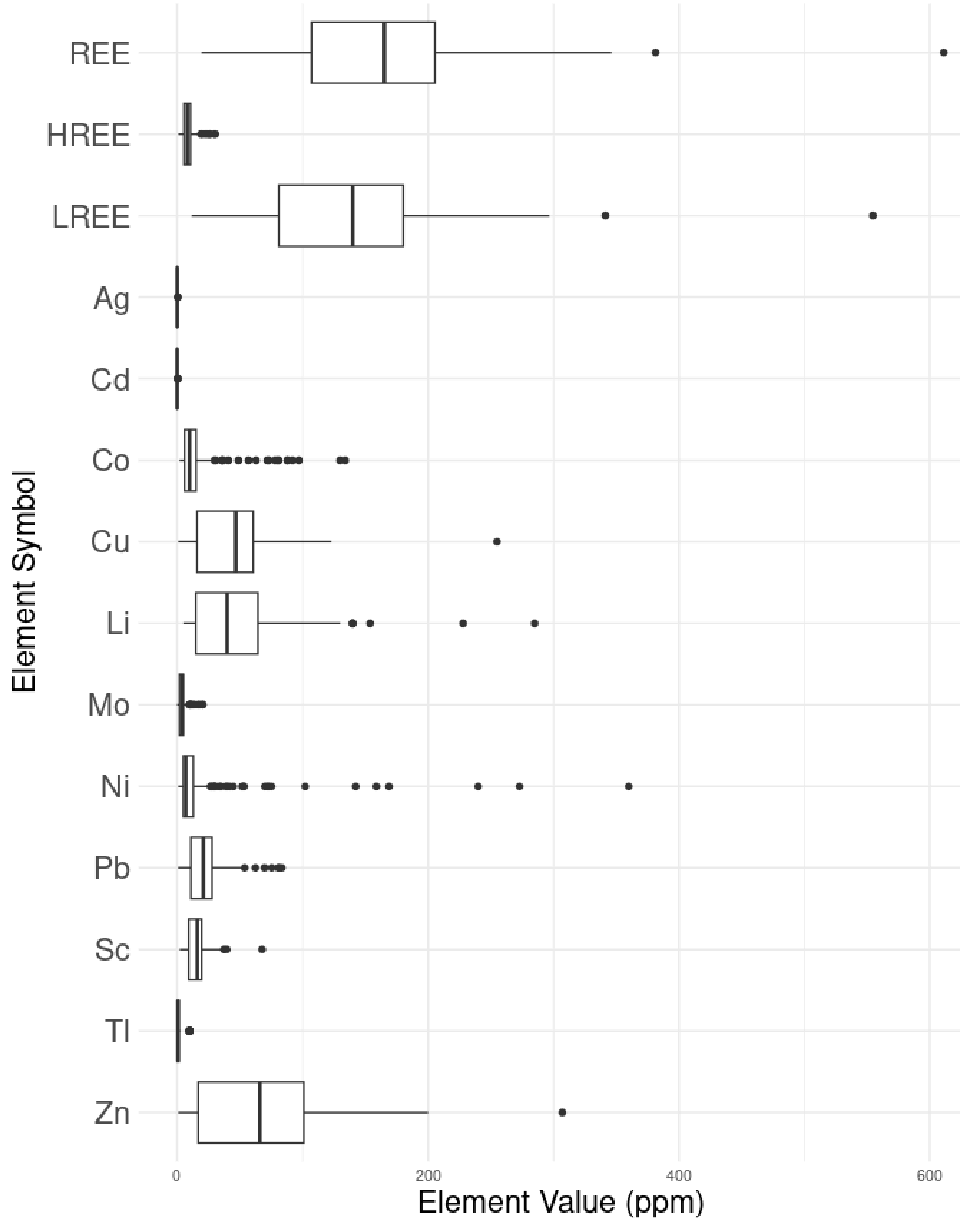


EXPLORATORY DATA ANALYSIS ...(1)

1 Descriptive Statistics

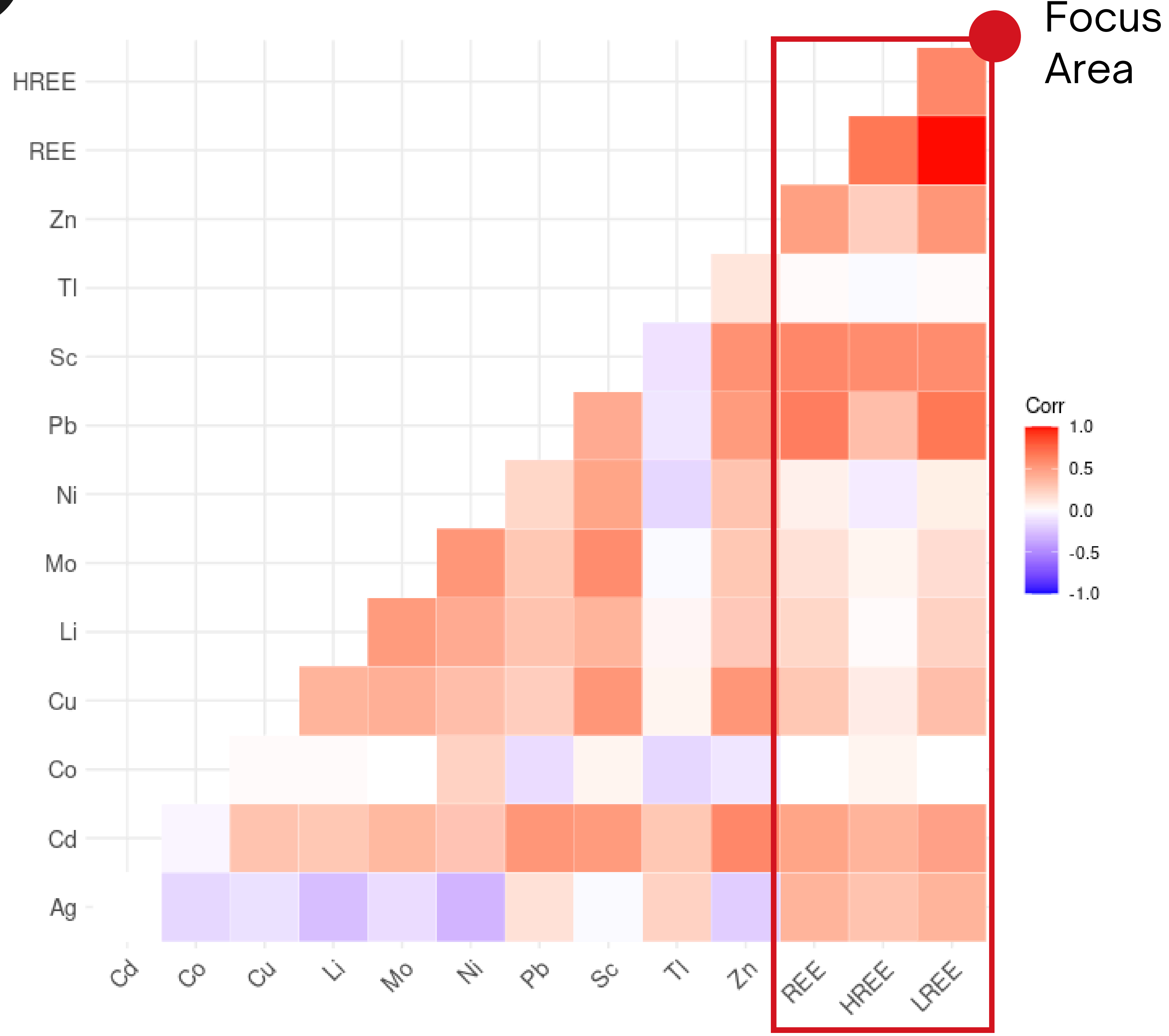
Element	Min	Max	Mean	Median	Range	Q1	Q3	IQR	SD	kurtosis	Missing Obsv.	Total Obsv.
REE	19,60	611,00	159,33	165,43	591,40	107,10	205,51	98,41	78,89	6,57	15	237
HREE	1,10	30,60	8,67	8,20	29,50	5,46	10,75	5,29	4,81	6,71	6	246
LREE	11,70	554,50	134,28	140,06	542,80	81,10	180,30	99,20	71,05	6,99	11	241
Ag	0,10	0,66	0,21	0,18	0,56	0,11	0,27	0,16	0,12	5,78	190	62
Cd	0,01	0,72	0,13	0,09	0,71	0,05	0,19	0,14	0,11	9,84	133	119
Cu	1,00	255,00	42,03	47,00	254,00	16,00	60,75	44,75	28,99	13,14	2	250
Li	5,00	285,00	47,72	40,00	280,00	15,00	64,50	49,50	39,49	9,53	30	222
Mo	0,10	20,60	4,46	4,00	20,50	2,00	5,00	3,00	3,29	10,24	67	185
Ni	1,00	360,00	16,97	7,00	359,00	5,00	13,00	8,00	38,25	44,10	18	234
Pb	0,89	83,45	21,69	21,00	82,56	11,06	28,00	16,94	14,54	6,14	-	252
Sc	2,20	67,80	15,47	16,20	65,60	9,33	19,68	10,35	8,30	8,48	6	246
Tl	0,03	10,00	1,94	0,72	9,97	0,36	1,51	1,15	3,12	5,78	196	56
Zn	1,00	307,00	65,38	66,00	306,00	17,00	101,00	84,00	49,65	4,13	3	249
Co	2,00	134,00	15,40	10	132,00	6,00	15,25	9,25	19,27	17,66	8	244

2 Data Distribution



EXPLORATORY DATA ANALYSIS ...(2)

3 Correlation Matrix



4 Correlation Coefficient

Elements	REE	HREE	LREE
Pb	0,64	0,32	0,68
Sc	0,59	0,57	0,57
Zn	0,49	0,24	0,52
Cd	0,46	0,38	0,47
Ag	0,39	0,31	0,38
Cu	0,28	0,09	0,32
Li	0,21	0,02	0,23
Mo	0,15	0,05	0,16
Ni	0,06	-0,07	0,08
Tl	0,01	-0,02	0,02
Co	0	0,05	0

Above > 0,5 Corr. Coeff.

METHODOLOGY

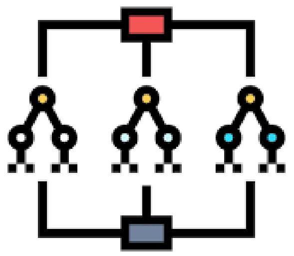
Data Pre-processing:

- **Replacing outliers with median:** Outliers for each element are replaced with the median value of that element within their respective project.

- **Replacing missing values with median:** Missing values for each element are replaced by the median value within their respective project; if no values exist in the project, the global median for that element is used.

Model Building Scenarios

- All elements from Test A
- Elements with high coefficient to target variable (>0.5)
- Top variable importance in model contribution.
- Strata Grouping



Random forest

Gradient boosting

Hyper-parameter Tuning

Random forest: mtry, min_n, trees

Gradient boosting: tree_depth, learn_rate, sample_size, min_n, mtry, loss_reduction

with Cross-Validation resamples

Model Performance Comparison

Used the trained model to predict REE, HREE, and LREE. Utilise these model evaluation performance to choose the best model:

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- R-Square
- MAPE (Mean Absolute Percentage Error)

PREDICTIVE MODELLING...(4)

Target	Method	Scenarios	Elements	RMSE	R-Sq	MAE	MAPE
REE	Random Forest	Default tuned	All Elements	47,57	0,62	33,85	0,26
REE	Random Forest	Elements with high variable importance	Pb, Tl, Sc, Zn, Cu, Cd	48,82	0,60	35,76	0,28
REE	XGBoost	Default tuned	All Elements	50,37	0,57	37,26	0,29
REE	Random Forest	Elements with correlation coefficient above 0.5	Pb, Sc	55,65	0,48	38,04	0,29

PREDICTIVE MODELLING...(5)

Target	Method	Scenarios	Elements	RMSE	R-Sq	MAE	MAPE
HREE	Random Forest	Default tuned	All Elements	3,33	0,46	2,13	0,29
HREE	Random Forest	Models with high variable importance	Pb, Sc	3,60	0,37	2,48	0,37
HREE	XGBoost	Models with high variable importance	Pb, Sc	3,54	0,39	2,36	0,34

PREDICTIVE MODELLING...(6)

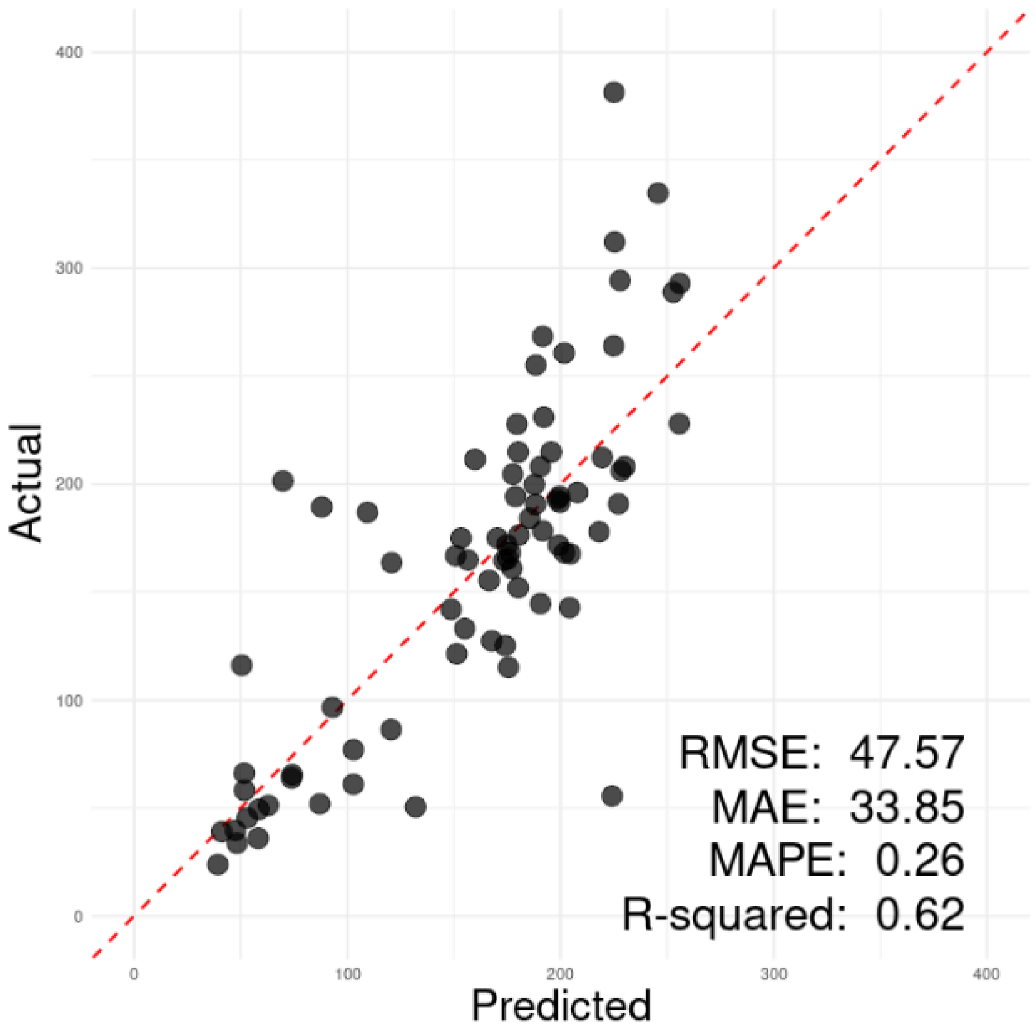
Target	Method	Scenarios	Elements	RMSE	R-Sq	MAE	MAPE
LREE	Random Forest	Default tuned	All Elements	39,61	0,61	26,62	0,25
LREE	Random Forest	Elements with high variable importance	Tl, Pb, Sc, Cd, Zn, Cu	40,48	0,59	27,7	0,27
LREE	XGBoost	Elements with correlation coefficient above 0.5	Pb, Sc, Zn	41,7	0,57	31,08	0,27
LREE	Random Forest	Elements with correlation coefficient above 0.5	Pb, Sc, Zn	38,44	0,63	29,6	0,31

DISCUSSION...(1)

REE

Random Forest - Default tuned - All Elements

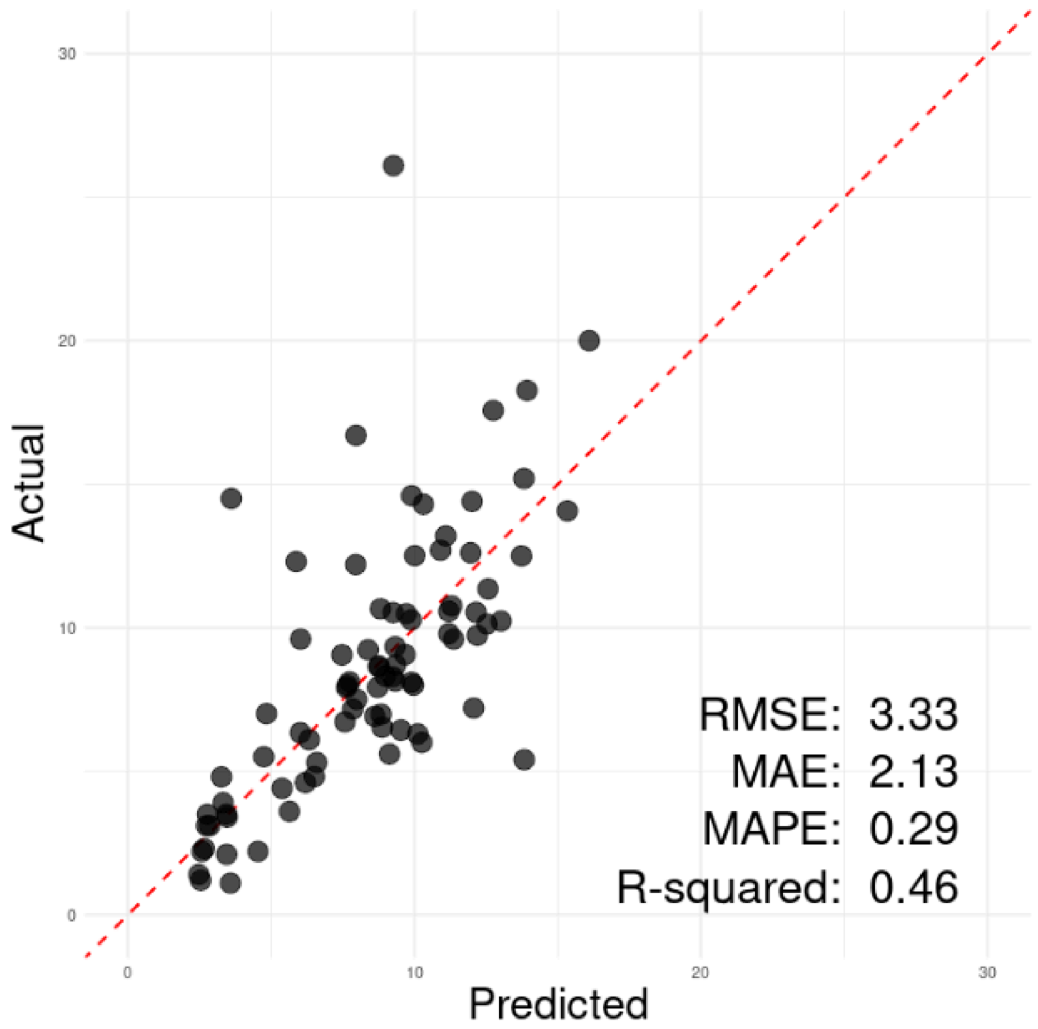
Predicted vs Actual Chart



HREE

Random forest - Models with high variable importance

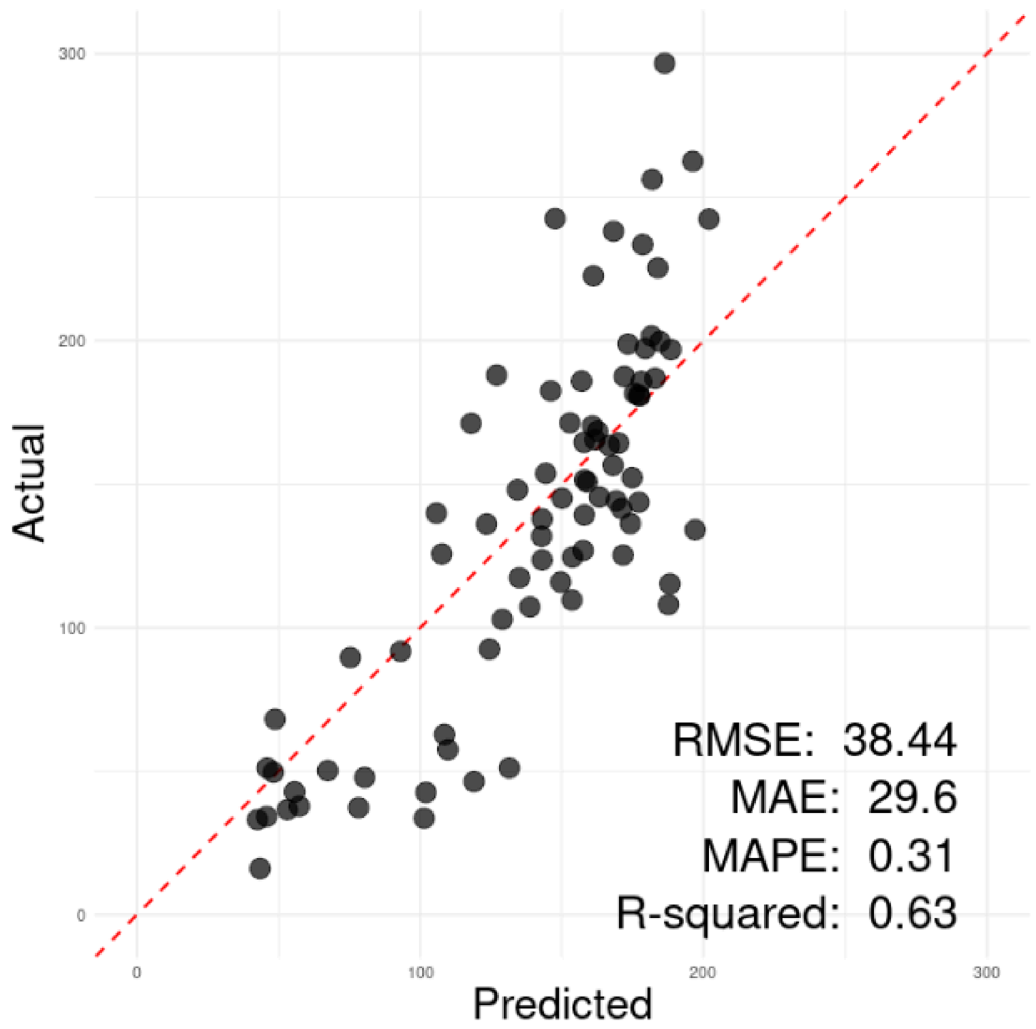
Predicted vs Actual Chart



LREE

Random forest - Elements with correlation coefficient above 0.5

Predicted vs Actual Chart



DISCUSSION...(2)

Pros

- **RMSE and MAE** of the best models are **below Standard Deviation** of REE, HREE, and LREE accordingly.
- These results indicates the models can be deemed as **good models**.

Cons

- R-Square of the best models are ranging from 0,60 - 0,65, indicating the models are **moderately strong** in capturing the variability of the data.
- MAPE of the best models are **relatively high**, ranging from 17% - 30%, indicating the models are, on average, off by these numbers.

Justification

- The correlation coefficient of all elements from Test A are weak to moderate (< 0,68).
- Cost and benefit trade-off.
- Still related to above, to improve the model in the future, more data are needed, it is reasonable to use elements from Test A.
- Error in the lab test when inputting concentration values, resulted in outliers value.

CONCLUSION

- All the best result models can be considered as **"reasonably good"** model. Several influencing factors are:
  - RMSE and MAE records lower value than target variables' standard deviation;
  - Lab test cost-benefit;
  - Weak-Moderate relationship of predictors to the target variables; and
  - Data inputation error from lab test.

- Limitation:
  Time constraints prevented us from getting corrected lab data, which could have reduced outliers and improved model accuracy.

- Future Improvement:
  The model can improve with more data, as Test A's affordability makes new samples practical. This could greatly boost prediction accuracy and performance.

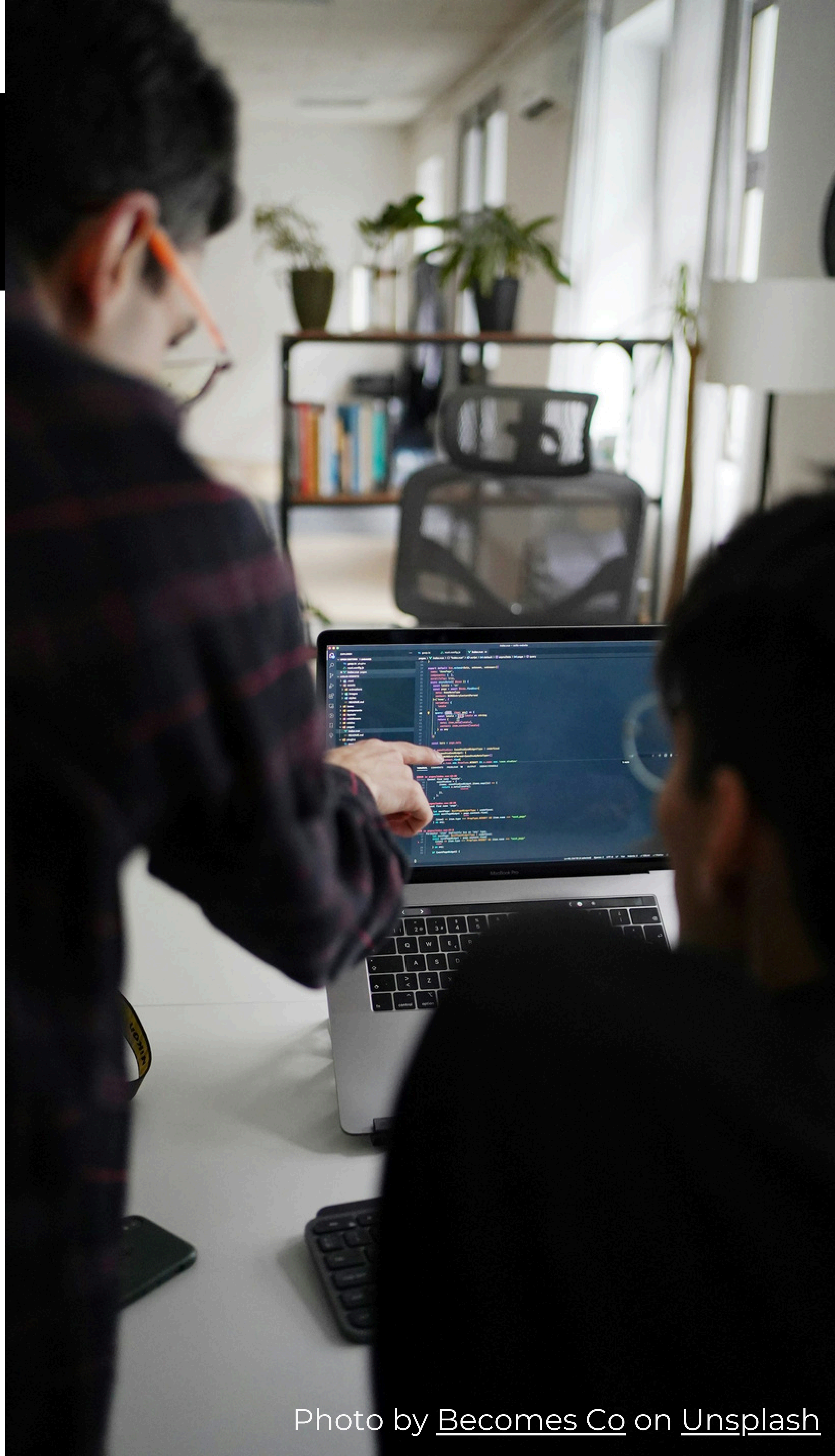


Photo by Becomes Co on Unsplash

THANK YOU

For your time and attention