Gawler Unearthed

Mine-Now

JULY 2020

PROBLEM STATEMENT

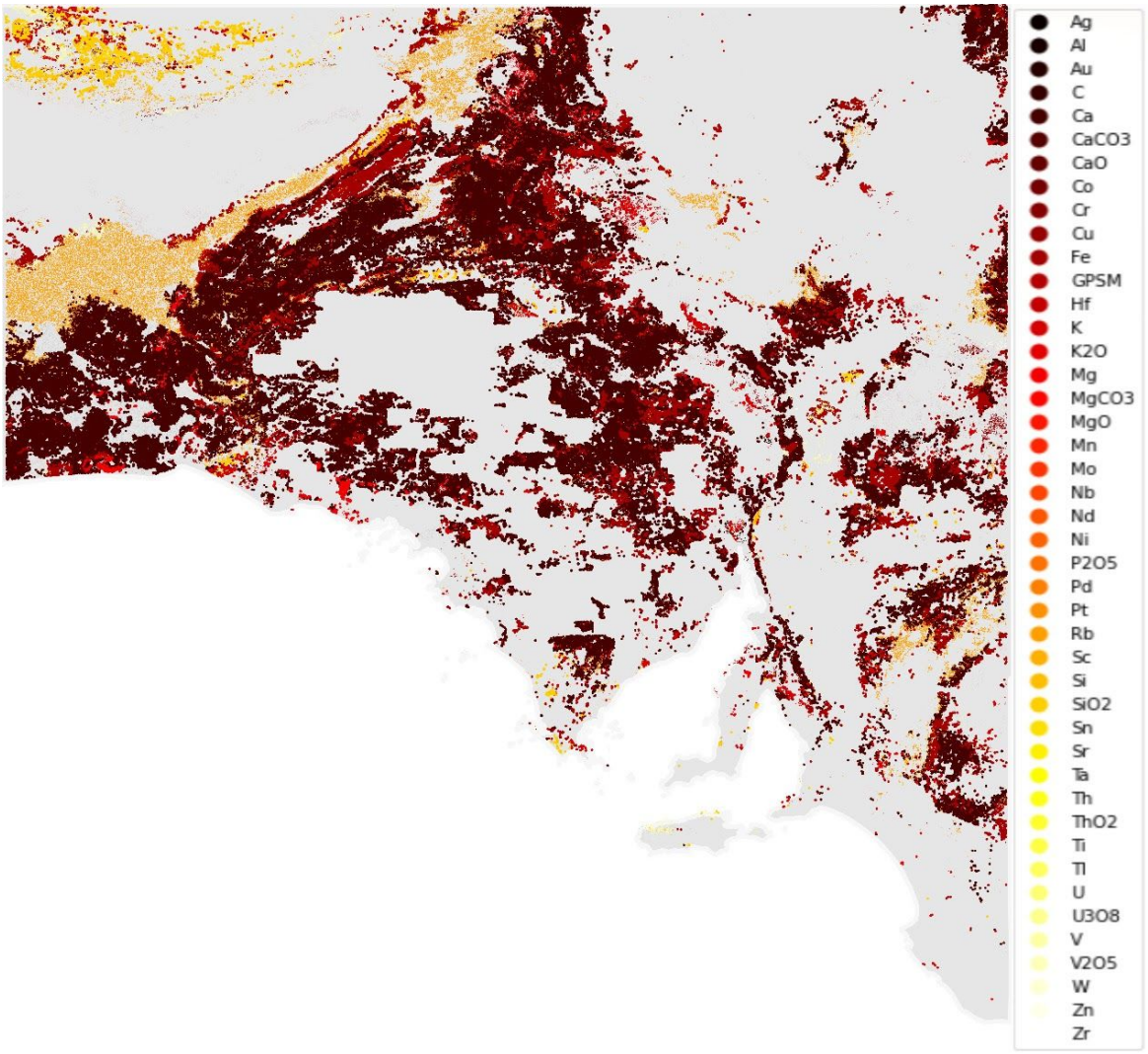
Since the discovery of the first coal mine near Newcastle in the late 1800s, mining has turned out to be an integral part of Australia's Culture & Development generating around 77 Billion $ of Annual Revenue.

Since most outcropping deposits have already been discovered and mature mining camps have started to dry out, the discovery of new mineral deposits has become important, increasingly expensive, and risky in the last 15 years. And there lacks an invertible model that can very appropriately exploit and learn from humongous amounts of data that has been generated by new geophysical & geochemical methods.

PROPOSED SOLUTION

We, at Mine Now, have created a three-step approach to predict Mineralization in the Gawler region including which Mineral will most dominantly be found and the size of the deposit.

Three different Machine Learning models were trained, all of which attained an AUC of more than 0.97 individually and overall accuracy of ~88%, ~90% & ~91.5% respectively on the testing set.



Using the trained model, 750,000 generated data points (excluding the already dug locations) were evaluated. Our models successfully predicted around 350,000 points out of the whole to contain mineralization along with the Mineral Class which would most dominantly be found there and the deposit type based on the size.

DATA PREPARATION

The data that was used can be broadly divided into two categories - Geophysical data in the form of visual/gridded rasters of the whole landmass of Gawler province and the Geochemical data comprising details such as Depth, Mineral Classes, etc. in the form of Delimited Text/Shape Files of the already existing mineral locations.

For the purpose to extrapolate novel unmined locations, geophysical inputs were taken as the features for the model to learn due to the limitation of Geochemical data to only the mineralized sites. The geochemical outcomes - Mineralization (Yes/No), the most dominant mineral (Mineral Class), and the type of Deposit (Small/Medium/Large) were used as the target labels to predict.

A total of 6 geophysical inputs were used as the models' features ->

- Gravity (GRAV) - .tif
- Gravity 1-Vertical Derivative (GRAV_1VD) - .tif
- Digital Elevation Modelling (DEM-9s) - .asc
- The 9 Second Flow Direction Grid (D8-9S) -.asc
- Total Magnetic Intensity (TMI) - .ers
- TMI Variable Reduction to Pole (TMI_VRTP) - .ers

The raw formats (with one channel) were obtained for the different rasters from different locations. (All the rasters are compiled and could now be obtained through this link.)

Other Geophysical inputs such as Radiometric & Spectroscopy were not included as features due to the smaller and/or incomplete size of their raster files.

The geochemical data were obtained from the Unearthed_5_SARIG_Data_Package.zip file. Two of the files were only required after extractions:

- sarig_dh_details_exp.csv
- sarig_rs_chem_exp.csv

The first file (**sarig_dh_details_exp.csv**) was used to obtain the data for training the first model. Three columns were selected and the null values were dropped to train the first model:

- LONGITUDE_GDA94
- LATITUDE_GDA94
- MINERAL_CLASS

A total of **147407** data points were obtained for Mineral Class - 'Y' and **174436** for Mineral Class - 'N'.

The second file (**sarig_rs_chem_exp.csv**) was used to prepare the data to train the second and the third model. This file contained 37 Million Data Points (~11 GB). Four columns were selected to prepare the data for the second and the third model:

- LONGITUDE_GDA94
- LATITUDE_GDA94
- CHEM_CODE
- UNIT_PPM_VALUE

All the duplicates were dropped and only one data point for a unique location was selected. To evaluate which datapoint to choose and which to drop, we calculated the value at every row. This value column was obtained by multiplying the price/unit of the respective mineral and the Amount in PPM.

	LONGITUDE_GDA94	LATITUDE_GDA94	CHEM_CODE	UNIT_PPM	VALUE
0	134.141459	-30.215218	Au	50000.0	2.240000e+09
1	134.775396	-30.303629	Au	50000.0	2.240000e+09
2	134.781993	-30.334396	Au	50000.0	2.240000e+09
3	136.051548	-27.543329	Sc	410000.0	1.418600e+09
4	136.034312	-27.604906	Sc	380000.0	1.314800e+09
...	...	...	...	...	...
401701	137.907468	-29.919152	Cu	0.0	0.000000e+00
401702	137.893009	-29.801222	Cu	0.0	0.000000e+00
401703	137.892969	-29.799842	Cu	0.0	0.000000e+00
401704	137.893393	-29.816697	Cu	0.0	0.000000e+00
401705	137.900277	-29.815406	Cu	0.0	0.000000e+00

401706 rows × 5 columns

Then, the Amount obtained in PPM (UNIT_PPM) was converted into either of the three classes based on the following range for the purpose of converting the regression problem into a classification one -

- **Low**: 0-100 PPM
- **Medium**: 100-50,000 PPM
- **High**: 50,000 + PPM

	LONGITUDE_GDA94	LATITUDE_GDA94	CHEM_CODE	UNIT_PPM	VALUE	UNIT_PPM_CLASS
0	134.141459	-30.215218	Au	50000.0	2.240000e+09	MED
1	134.775396	-30.303629	Au	50000.0	2.240000e+09	MED
2	134.781993	-30.334396	Au	50000.0	2.240000e+09	MED
3	136.051548	-27.543329	Sc	410000.0	1.418600e+09	HIGH
4	136.034312	-27.604906	Sc	380000.0	1.314800e+09	HIGH
...	...	...	...	...	...	...
401701	137.907468	-29.919152	Cu	0.0	0.000000e+00	LOW
401702	137.893009	-29.801222	Cu	0.0	0.000000e+00	LOW
401703	137.892969	-29.799842	Cu	0.0	0.000000e+00	LOW
401704	137.893393	-29.816697	Cu	0.0	0.000000e+00	LOW
401705	137.900277	-29.815406	Cu	0.0	0.000000e+00	LOW

401706 rows × 6 columns

Finally, the value column and the UNIT_PPM column were dropped and the file was exported to act as the training for Model 2 and model 3.

Now to obtain the geophysical features at these locations, using QGIS, all the 6 rasters were sampled using the Point Sampling Tool. The resulting sampled csvs were combined and the output looked similar to this:

```
In [3]: train.head()
```

C1	LONGITUDE_	LATITUDE_G	MINERAL_CL	gravity_ma	dem-9s	SA_TMI	gravity_1V	SA_TMI_VRT	d8-9s
142	129.11	-31.5842	N	-33.482	95.7766	-36.3184	-0.00034	-192.407	2
143	129.346	-31.5211	N	-25.5448	96.9994	-143.02	0.00044	-423.11	2
145	129.385	-31.6402	N	-28.2794	87.0945	1077.01	-0.00015	667.035	1
147	129.167	-31.1688	Y	-32.5929	129.967	-1565.33	0.00068	-1701.49	4
148	129.164	-31.1625	Y	-33.3853	131.218	-227.706	0.00042	640.236	4
149	129.11	-30.7499	N	-45.8616	154.939	172.563	-0.00022	43.9881	1
150	129.247	-30.5269	N	-52.8326	159.276	64.9359	-0.00034	52.2999	2
151	129.371	-30.5701	N	-53.8595	138.374	206.621	-0.00018	327.643	2
153	129.376	-26.6621	N	-4.67423	538.87	360.422	0.00034	68.9003	16
154	129.404	-26.6963	N	-6.10203	536.082	-157.93	0.0017	292.317	64

This is the file used for the training of the first model. The target label is the **MINERAL_CL** which is either ‘Y’ or ‘N’ and except for **C1, LONGITUDE_ & LATITUDE_G**, all the remaining columns are used as the training features.

OVERVIEW OF THE MODELS

We tried to create 3 different models stacked over each other to predict mineralization in different regions of Gawler. The model training and metrics are presented down below.

1. Model 1 - “Wh’re Art Thee Min’ral?”

The 1’st model “Wh’re Art Thee Min’ral” tries to perform a naive, though a very crucial task, of simply predicting if, given the Geophysical features (such as Gravity, Elevation) of a pair of Latitude, Longitude, is there a Mineral of Not?

An h2o’s AutoML Algorithm was trained for 20 Minutes over 80% of the prepared dataset (80% - Training & 20% Testing) to attain a whopping AUC of ~0.99 & accuracy of ~91 % on testing data.

```
aml.leaderboard
```

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20200727_130454	0.977349	0.195728	0.972297	0.0753031	0.237675	0.0564893
StackedEnsemble_BestOfFamily_AutoML_20200727_130454	0.977349	0.195728	0.972297	0.0753031	0.237675	0.0564893
DRF_1_AutoML_20200727_130454	0.976129	0.216466	0.970375	0.0769354	0.243678	0.0593789
XRT_1_AutoML_20200727_130454	0.972935	0.243863	0.965447	0.0804539	0.249788	0.0623942
XGBoost_2_AutoML_20200727_130454	0.961389	0.263156	0.952788	0.101885	0.277815	0.0771813
GBM_5_AutoML_20200727_130454	0.949435	0.351741	0.938337	0.120675	0.318762	0.101609
XGBoost_1_AutoML_20200727_130454	0.947571	0.300069	0.935833	0.12216	0.300862	0.0905181
XGBoost_grid__1_AutoML_20200727_130454_model_1	0.944831	0.30148	0.931614	0.124733	0.303365	0.0920301
GBM_4_AutoML_20200727_130454	0.942094	0.349772	0.929859	0.128643	0.32109	0.103099
GBM_grid__1_AutoML_20200727_130454_model_2	0.92922	0.35629	0.914609	0.146284	0.329473	0.108552

***Note - this shows the metrics of the models used in the training of the AutoML Model**

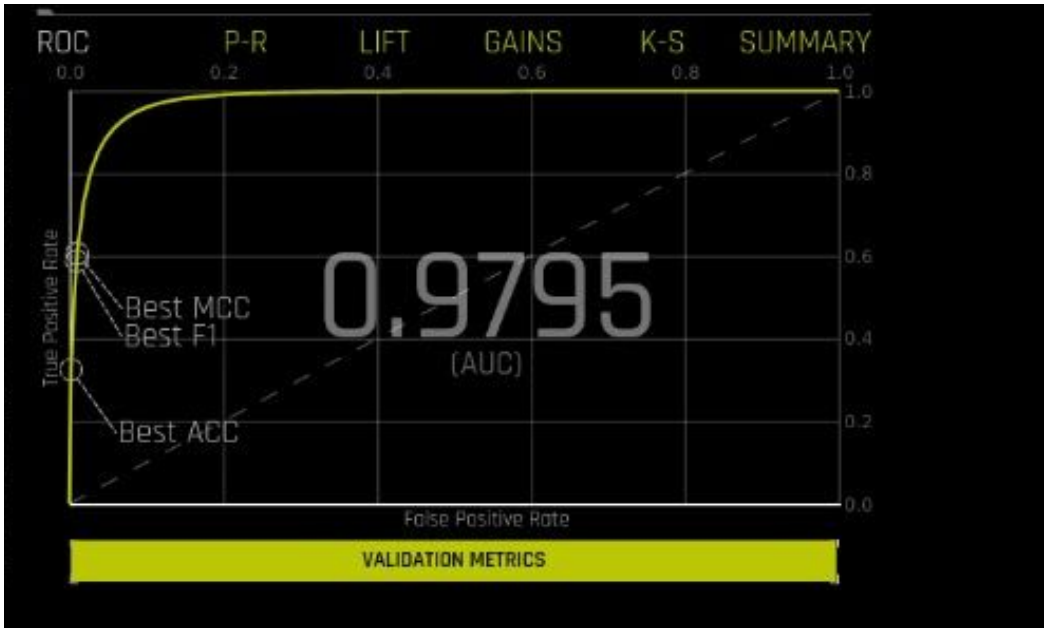
		N	Y	Error	Rate
0	N	125827.0	11894.0	0.0864	(11894.0/137721.0)
1	Y	7564.0	110176.0	0.0642	(7564.0/117740.0)
2	Total	133391.0	122070.0	0.0762	(19458.0/255461.0)

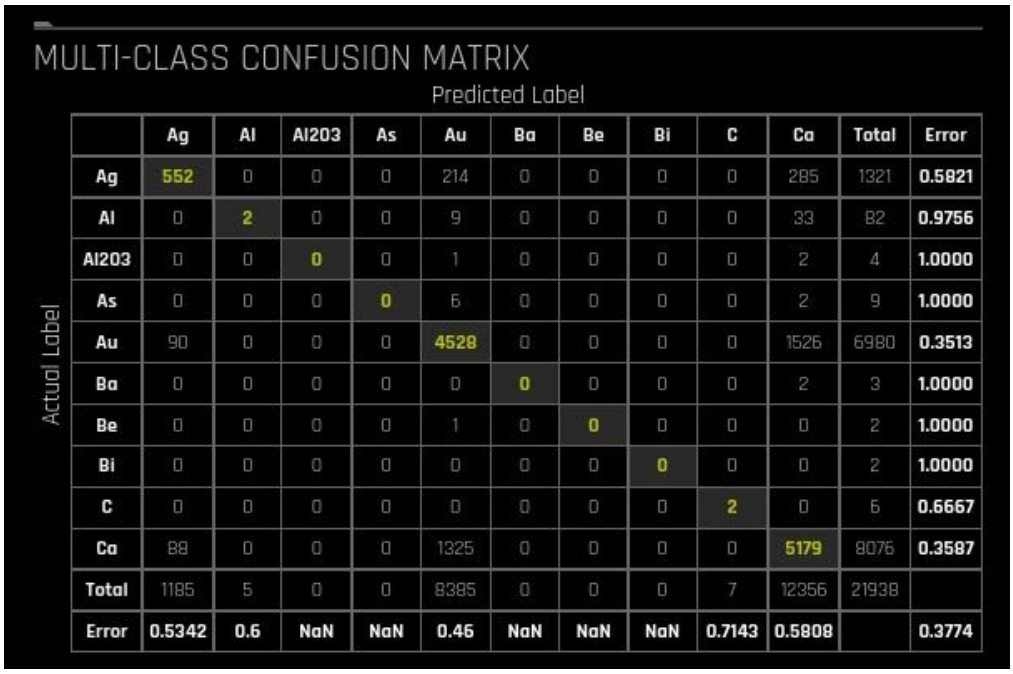
2. Model 2 - “Bid Me, Thy Nameth!”

The second model “Bid Me Thy Nameth” was trained to find out about which mineral/metal is present at the location for which the geophysical data is fed.

This is a multi-class problem and a more powerful tool was needed for the same. After trying various state-of-the-art solutions including Sklearn, Keras, Pytorch, Autokeras, MLBox, the best results were obtained using h2o’s Driverless AI.

The model was trained for about an hour. These were the final metrics->





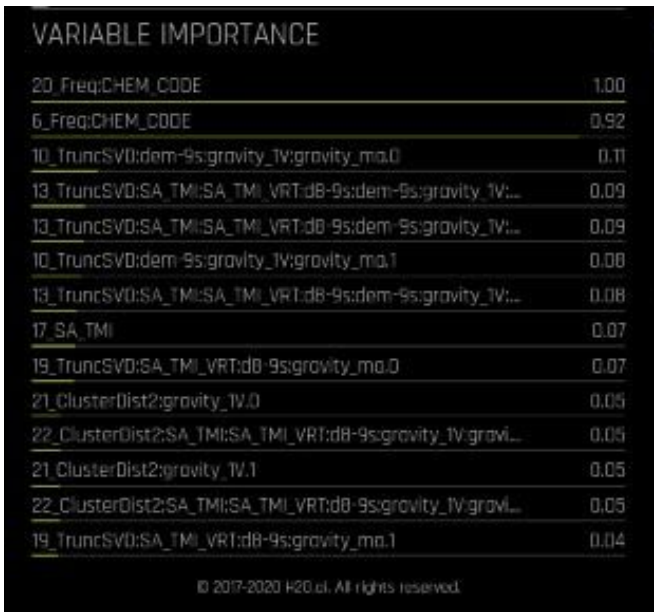
	Ag	Al	Al2O3	As	Au	Ba	Be	Bi	C	Ca	Total	Error
Ag	552	0	0	0	214	0	0	0	0	285	1321	0.5821
Al	0	2	0	0	9	0	0	0	0	33	82	0.9756
Al2O3	0	0	0	0	1	0	0	0	0	2	4	1.0000
As	0	0	0	0	6	0	0	0	0	2	9	1.0000
Au	90	0	0	0	4528	0	0	0	0	1526	6980	0.3513
Ba	0	0	0	0	0	0	0	0	0	2	3	1.0000
Be	0	0	0	0	1	0	0	0	0	0	2	1.0000
Bi	0	0	0	0	0	0	0	0	0	0	2	1.0000
C	0	0	0	0	0	0	0	0	2	0	6	0.6667
Ca	88	0	0	0	1325	0	0	0	0	5179	8076	0.3587
Total	1185	5	0	0	8385	0	0	0	7	12356	21938	
Error	0.5342	0.6	NaN	NaN	0.46	NaN	NaN	NaN	0.7143	0.5808		0.3774

The accuracy was calculated by a different approach for this model. As for every location the model with the highest value was obtained; though other minerals were also found and so if our model predicts those minerals, it's not wrong, right. When the accuracy was calculated that way, there was a significant boost than the actual. The final obtained accuracy was ~88%.

3. Model 3 - "How Big Ar't Thee?"

The third and the final model "How Big Art Thee?" was trained to find out about how big the deposit is. It's a multiclass problem as it classifies mineral deposits into either low, medium or high on the basis of an estimate of its amount in ppm.

The same dataset that was used for training the first model was used to train this model as well. An estimated 1 hour of training was required which resulted in the model to obtain these outstanding metrics ->



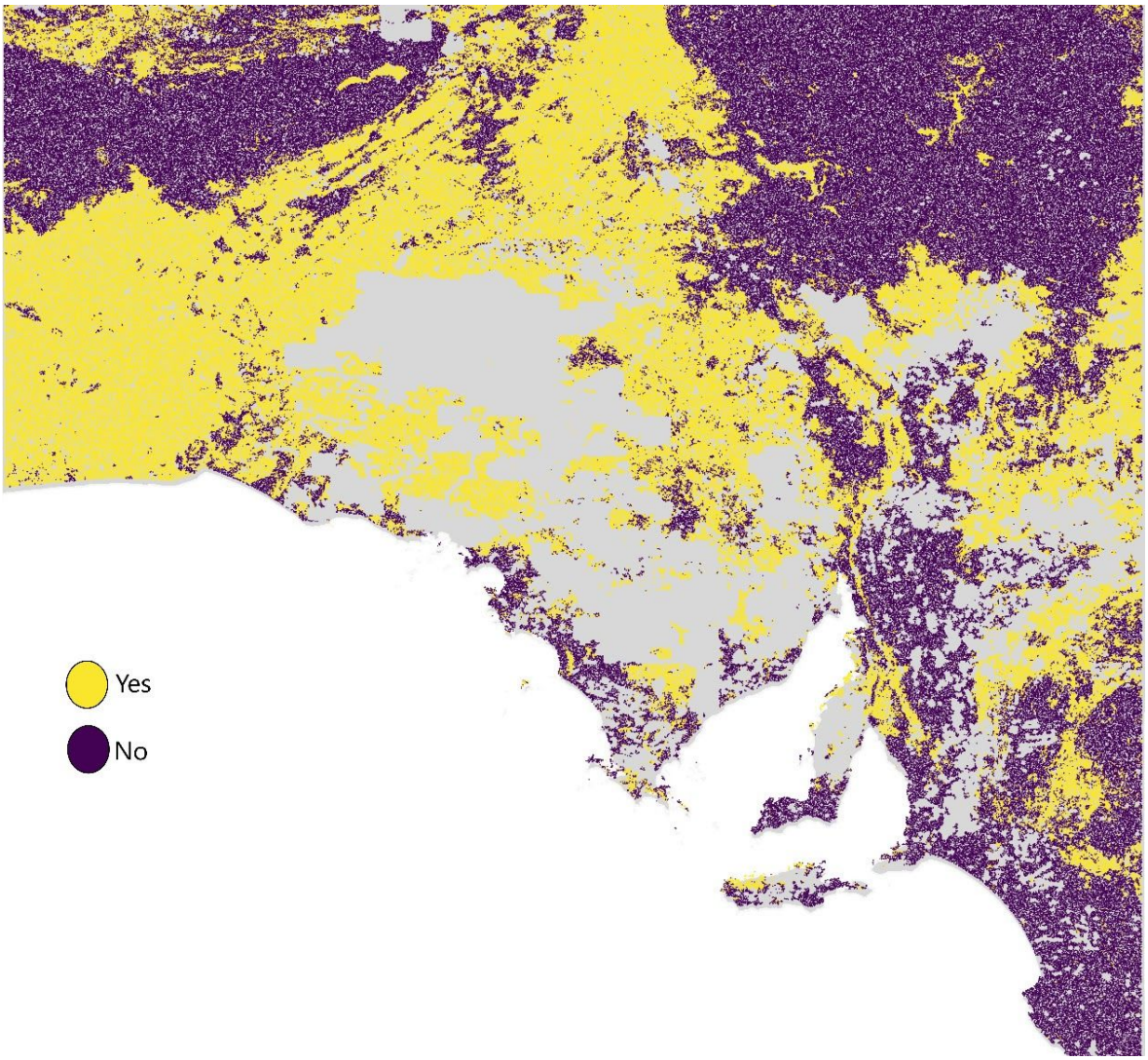
GENERATION OF NOVEL MINERALIZED DATA POINTS

After the models were trained, to predict new mineralized locations, 1 million points were generated using QGIS (spaced 2 Km Apart) inside the polygon comprising the landmass of the smallest raster (TMI) and excluding the data points that have already been dug.

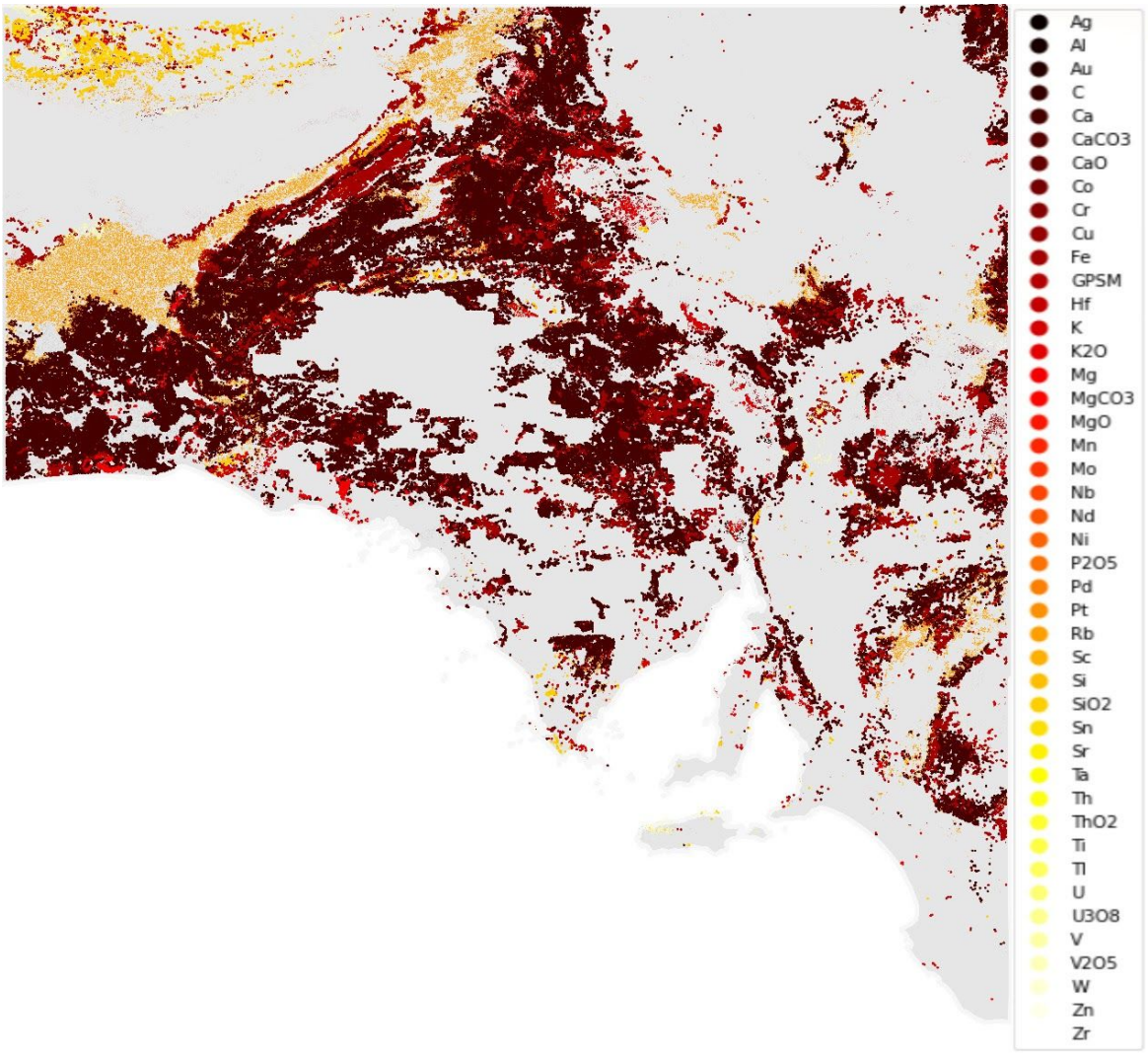
These data points were then sampled to obtain their respective geophysical features at those points. After sampling and excluding the rows with null values, the length of the file shrunk to 724,000. The final combined CSV was then fed to the 3 models serially to obtain predictions.

PREDICTIONS

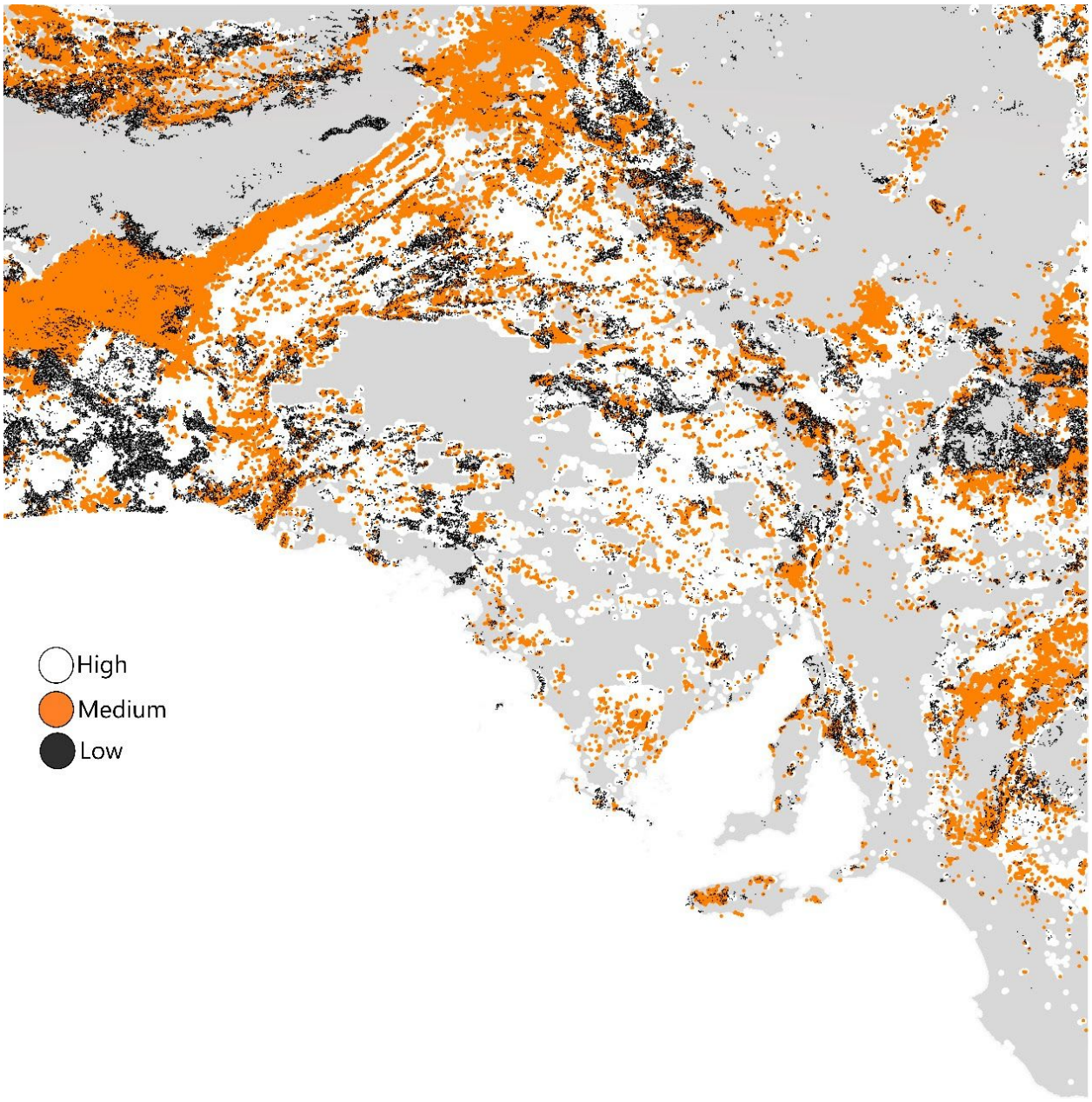
The MoJo pipeline was used to import all the saved files into Jupyter-Notebook. After loading the models, the predictions were obtained. Out of the 724K total data points, the first model predicted 342,000 points to contain mineralization. These sets of points were then selected and passed on to the second and the third model to find the predicted mineral class and sizes.



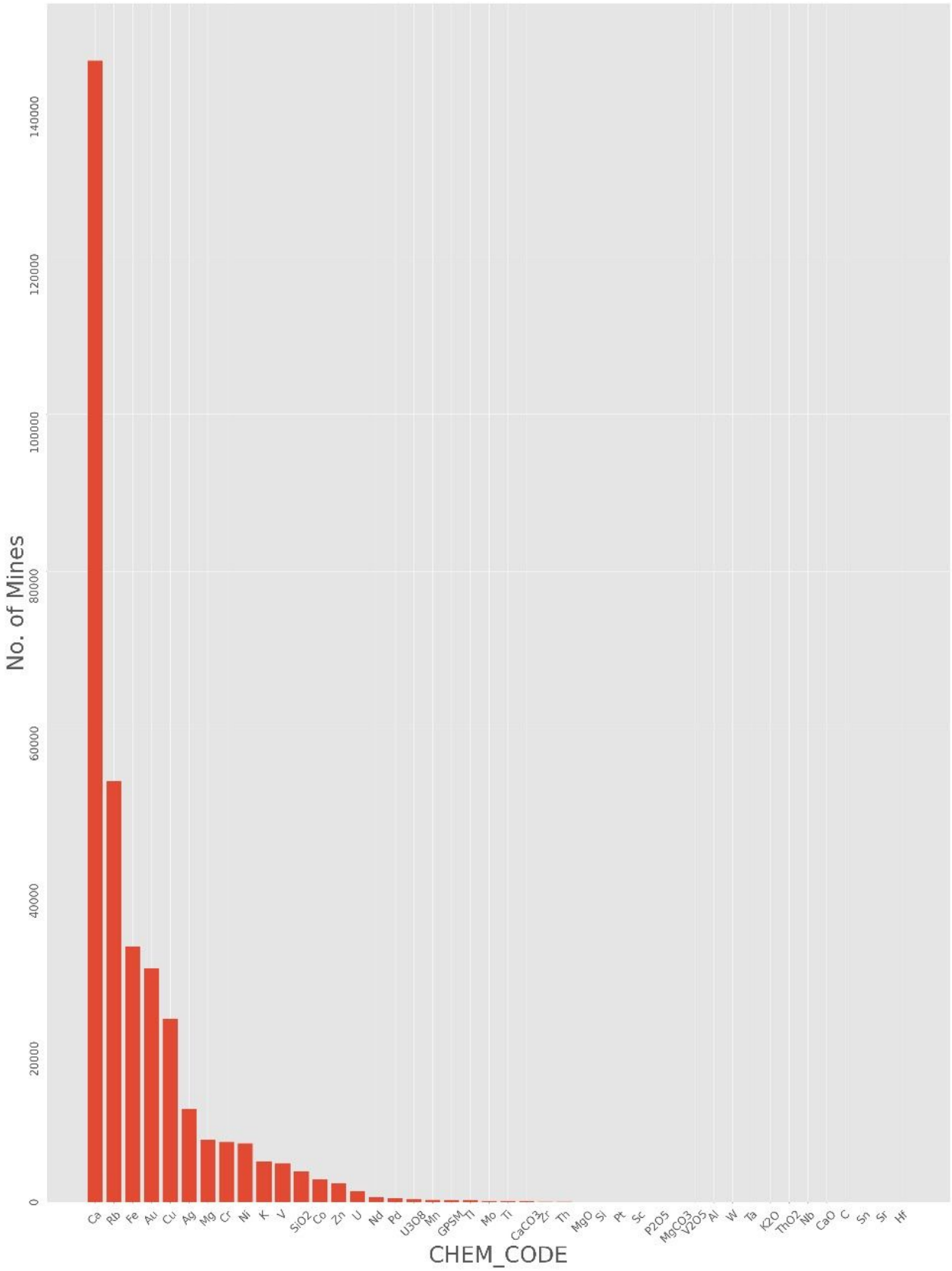
This plot represents the distribution of mineralized & unmineralized datapoints (predictions of the first model) across the Gawler region.



This plot represents the distribution of different minerals across the Gawler region as predicted by the model 2.



This plot represents the distribution of different sizes of minerals across the Gawler region as predicted by model 3.


No. of Mines
0
20000
40000
60000
80000
100000
120000
140000
Ca
Rb
Fe
Au
Cu
Ag
Mg
Cr
Ni
K
V
SiO2
Co
Zn
U
Nd
Pd
U3O8
Mn
GPSM
Ti
Mo
Ti
CaCO3
Zr
Th
MgO
Si
Pt
Sc
P2O5
MgCO3
V2O5
Al
W
Ta
K2O
ThO2
Nb
CaO
C
Sn
Sr
Hf
CHEM_CODE

CONCLUSION

Our findings are one of a kind and novel. Our model, even after using comparatively fewer features makes predictions with excellent accuracy. Just imagine, if more geophysical features such as Magentotellurics, Seismic included with lithology of rocks, the accuracy could be further improved. Also, novel features such as depth and rock types could be predicted using the same approach.

And the best part about it is - our approach is completely beginner-friendly & open-sourced. Anyone, with or without any background of Geology & Machine Learning could follow along with our specially curated documentation and find novel results. Our solution is just the beginning but it will set the pavement for the future of Mineral Exploration with Machine Learning.
