Mining Rock Chemistry Data





Overview

The chemistry of magma and the fluid it gives off as it cools are the most important source of metals in ore deposits.

Ore deposits have changed drastically over Earth's history. Certain deposit types are limited to the Archean or Proterozoic. Can we get age labels (eons) from bulk rock chemistry data for prospecting purposes?

PERIPHERAL CENTRAL Inner edge NACONDA of sphalerite INTERMEDIATE ZONE 1 km Outer edge of copper ore Map credit: Evans, A. M. An Introduction to Ore Geology, Elsevier, Zone of numerous Zone of overlap of copper quartz-molybdenite veinlets NY (1980) p. 148-150.



Archean (4-2.5 Ga) gold/uranium conglomerate (Witwatersrand, South Africa) Photo credit: INTERNATIONAL **GEOLOGICAL CONGRESS**

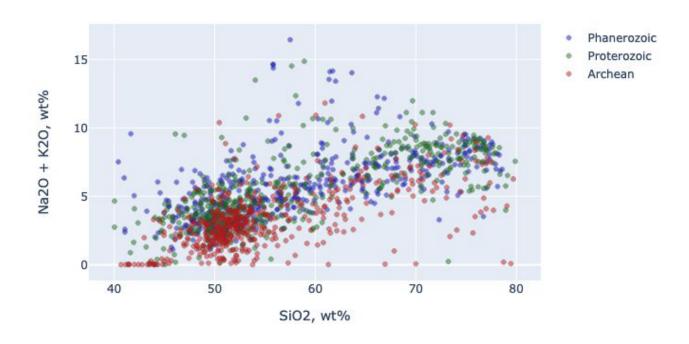
Phanerozoic (0.55 Ga to present) mineralized halos around igneous intrusion (Butte, MT)

=

The Data:

- EarthChem database
- ~160,000 usable rock compositions (major elements, lat & long, age)
- Feature engineering: manual.
 - It's all about the ratios.
 - Except when it's not.

TAS plot by geologic eon for EarthChem data



TAS diagram for rock classification and preliminary data visualization

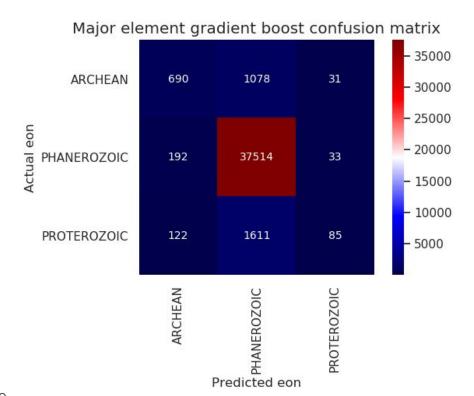
The Analysis 01

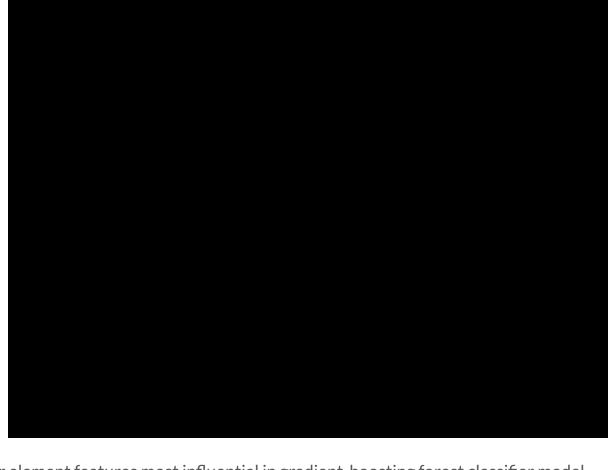
Tools used:

- Logistic regression + regularization
- Linear support vector classifiers
- Decision trees
- Random forest classifiers
- Gradient boosting forests

Techniques & metrics:

- Confusion matrices
- ROC curves (have to convert to binary problem)
- Accuracy, precision, and recall all have their place





Three major element features most influential in gradient-boosting forest classifier model

The Analysis 02

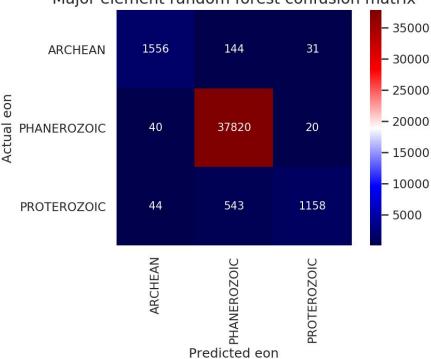
You can break this problem wide open*

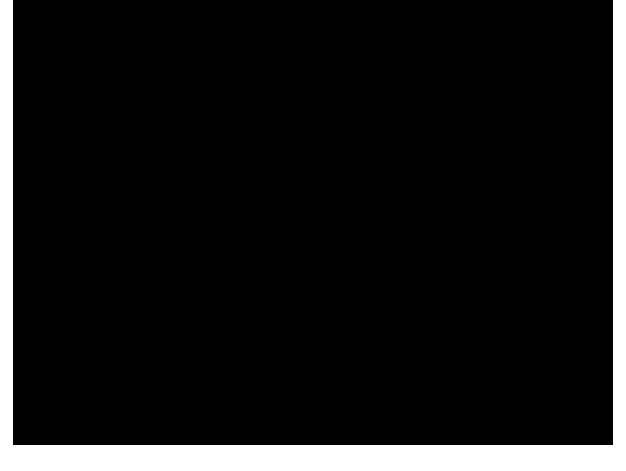
...if you make latitude and longitude features in your model.

Is that cheating?

* - and by break it wide open, I mean get the recalls all above 0.5

Major element random forest confusion matrix



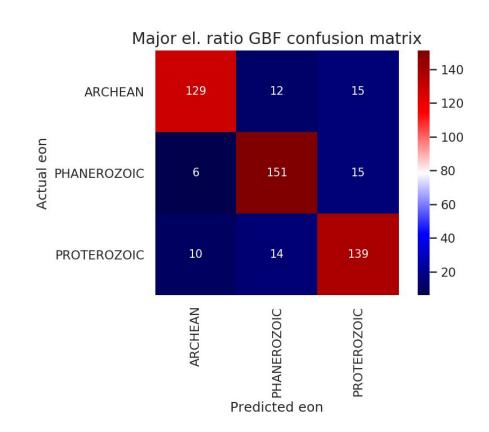


Region in focus: Southern Africa

The Analysis 03

The Southern Africa data sure do clean up nice, don't they?

- Balanced dataset (1,000 analyses per eon)
- Even recall > 0.92...
- Model shown: just five ratio features
- World domination may be too much to ask, but regional hegemony is within our grasp.



Prospecting - Use Cases & Model Optimization

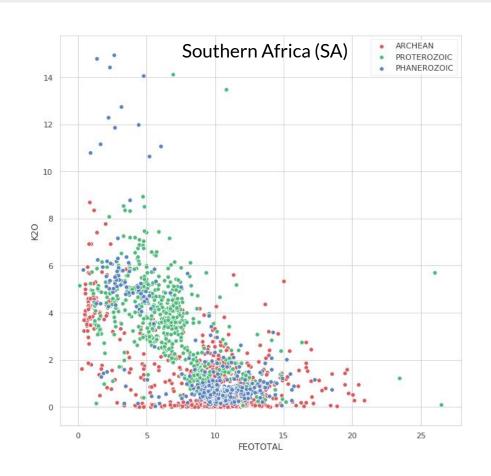
- Tune models on a regional basis.
- Regions that are still only partly explored (SE Asia, Africa, etc.)
- High recall make sure that no good sites escape notice.
- High precision make sure we're only investigating promising sites.
 Exploration is really expensive.





If Only There Were Time

- Refinement of probability thresholds to adjust precision & especially recall on models for specific use cases.
- Undersampling Phanerozoic compositions when constructing a new generation of models.
- More data visualization for 1 on 1 relationships between variables.
- Clustering analysis to take a new look at the VERY old, VERY divisive problem of rock classification.
- Use this data to construct a tutorial series in data science for geology students.

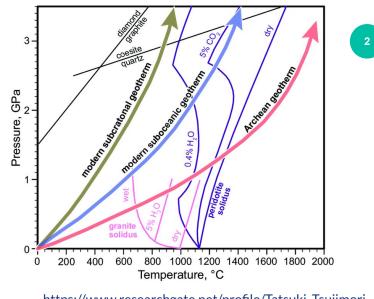


Thank you.

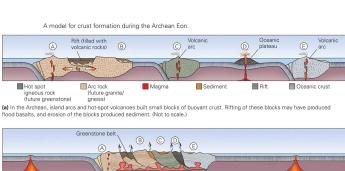


How has Earth changed over time?

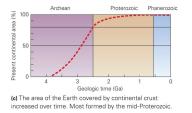
The Farth was hotter in the past and melted more easily. Perhaps (perhaps) melt fractions were larger, with a smaller proportion of "flux" elements (Na, K, P).



https://www.researchgate.net/profile/Tatsuki Tsuiimori



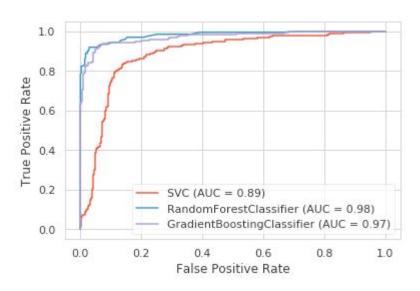


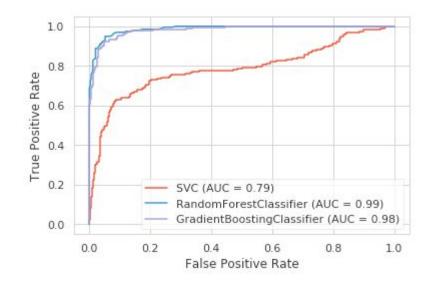


Continental cover & magma contamination less in the Archean.

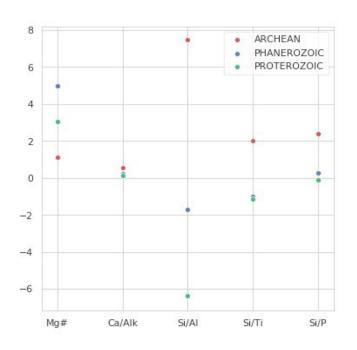
http://geologylearn.blogspot.com

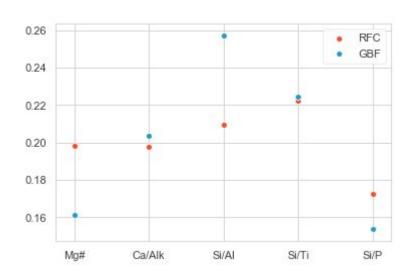
ROC Curves (SA)-Phan / PreC, Arch / ColdEarth





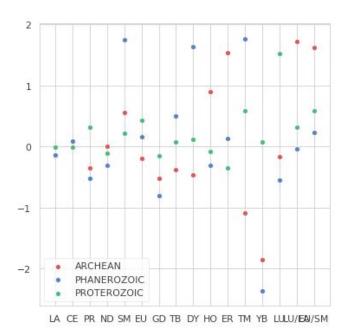
Ratios (SA): SVC coefficients; RFC, GBF importances

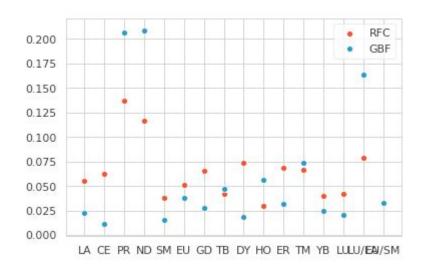




Gradient boosting forests have an OPINION.

Rare Earths (SA): SVC coefficients vs. RFC, GBF importances





Si/Al <= 0.105 gini = 0.666 samples = 1503 value = [514, 524, 465] class = PHANEROZOIC

Horticulture (SA)

Si/P <= 0.046 gini = 0.41 samples = 552 value = [105, 409, 38]

Mg# <= 0.165 gini = 0.599 samples = 951 value = [409, 115, 427] class = PROTEROZOIC

Ca/Alk <= 0.002 gini = 0.263 samples = 458 value = [43, 390, 25] lass = PHANEROZOI

Mg# <= 0.189 gini = 0.505 samples = 94 value = [62, 19, 13] class = ARCHEAN

Ca/Alk <= 0.002 gini = 0.567 samples = 692 value = [398, 93, 201] class = ARCHEAN

Si/Al <= 0.091 gini = 0.644 samples = 84 value = [28, 37, 19 value = [15, 353, class = PHANEROZC class = PHANEROZ

Mg# <= 0.35 gini = 0.107 samples = 374

Si/Ti <= 0.103 gini = 0.49 samples = 21

Si/Ti <= 0.03 gini = 0.266 samples = 73 value = [0, 12, 9] value = [62, 7, 4] value = [3, 12, 2] class = PHANEROZOIC class = ARCHEAN class = PROTERO

Ca/Alk <= 0.0 Si/Ti <= 0.036 gini = 0.653 gini = 0.591 samples = 31 samples = 465 value = [8, 10, 13 value = [223, 52, 19 value = [175, 41, 11] class = PROTEROZC class = ARCHEAN class = ARCHEAN

Si/Ti <= 0.019 gini = 0.371 samples = 227