Thanks for taking the time to apply!

Contained here is an example problem related to machine learning in mineral exploration. Currently the work that Caldera Analytics is focused on is targeting IOCG (Iron Ore Copper Gold) deposits in the Gawler Craton in South Australia. One of the key issues in this area is the depth to the basement rocks, which is where the IOCG deposits are contained. In these areas, the rock types that contain Copper deposits are under cover, that is, there are layers of sediment from different geological eras above the target rocks. This makes exploration hard, as it reduces the available data sources available. A lot of the already discovered mines had their target rocks at surface, so it's easier for geologists to inspect these rocks and determine if its worth drilling. This is not the case in the Gawler Craton.

In the Gawler craton, depth to basement can be anywhere from 50m up to 1200 meters (and probably more). The depth to the basement affects the economics of mining (deeper means more cost) and also affects how you interpret some data sources such as magnetic intensity, which means its critical for machine learning.

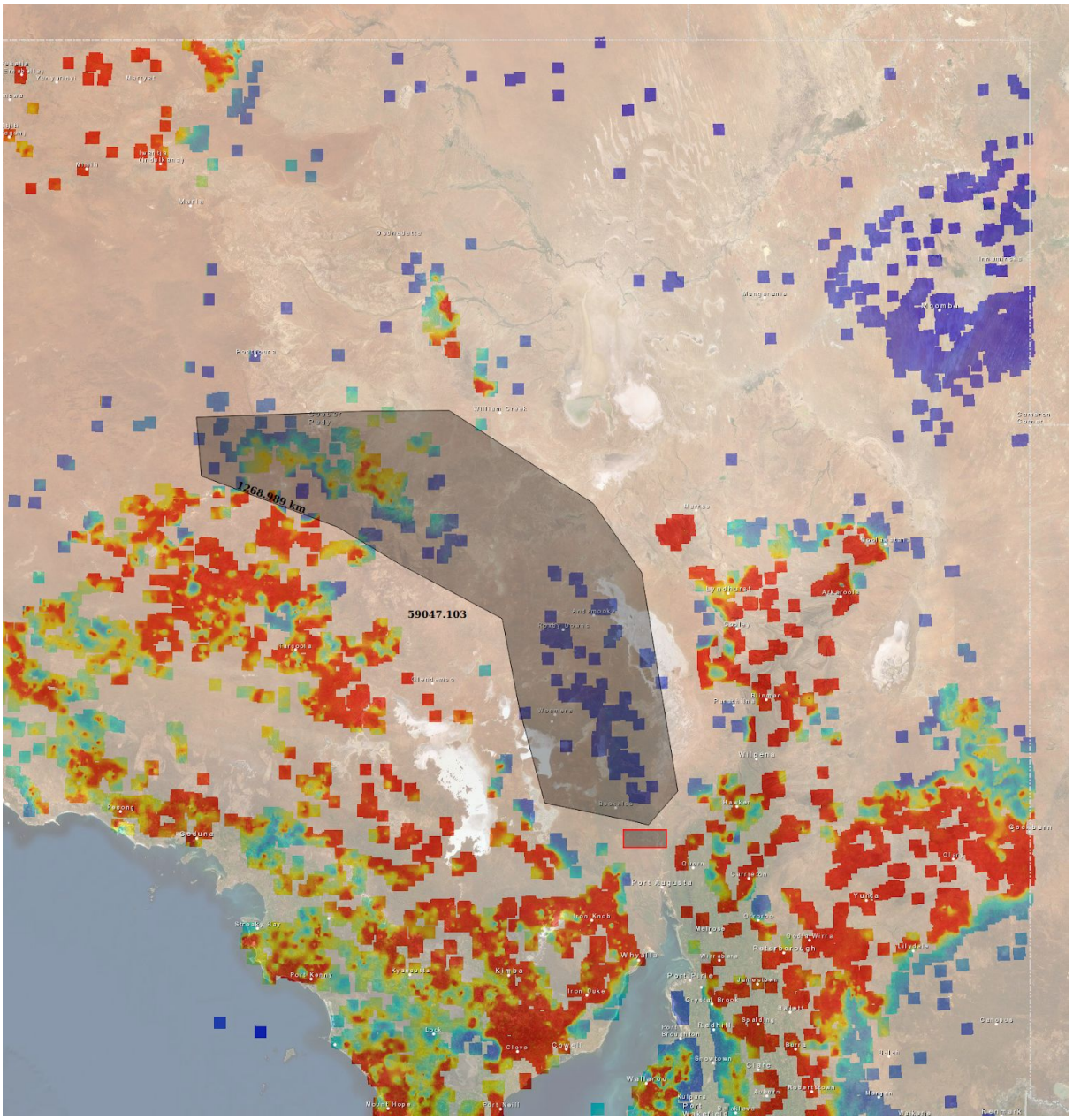
The Challenge
The challenge is to build a machine learning regression model to predict the depth to basement at a set location. Attached is a dataset containing historical drillholes, their location and the depth to the basement rocks in metres. Also contained are the potential geophysics predictors.

Area
The drill holes are located in the IOCG corridor of the Gawler Craton in South Australia (http://www.portergeo.com.au/database/mineinfo.asp?mineid=mn668). The link goes into pretty technical detail in terms of geology, so best just to look at the image to get a feel of the location.

Dataset in Detail
The predictors (along with the depth to basement targets) are sourced from SARIG, the SA government website for geology data (https://map.sarig.sa.gov.au/). All of the predictors are sourced from geophysics. When exploration is done under cover - geophysics is critical to understand what's underneath the ground. Different rocks have different gravity and magnetic signatures, so these data sources are used to guide exploration.

If you go to the All Map Layers tab, and then go to the Regional Geophysical Images category, you can see the source layers for the predictors, such as TMI (Total Magnetic Intensity) and Gravity 1VD (1st Vertical Derivative of the Gravity readings).

You can also see the Depth to Basement dataset as a patchy regional image in that category - it's a pretty crude dataset as they've taken a point where a drillhole has intersected the basement and extrapolated that to a set distance. I've drawn a rough polygon in the below image of where the dataset is located. Red means shallow depth to basement, blue means deep depth to basement.



Why is this a challenge?

One of the issues in mineral exploration machine learning is the small amount of samples and the high amount of potential predictors. Therefore feature selection is very important, but often feature selection can lead to overfitting if the validation is not done correctly.

For each of the geophysics data sources (eg 1st Vertical Derivative of Total Magnetic Intensity), I've taken the actual value of the data source at the historical drill hole, and then a heap of statistical aggregations such as the mean in a 2km radius, the mean in a 4km radius, the mean in a 8km radius, medians, mins etc. So some form of feature selection will be required to reduce the predictors of the machine learning model.

Another challenge is validation - random cross validation, cross validation or a holdout set? What is appropriate here?

Output

Please do this work in a jupyter notebook, showing your reasoning/method of work. Python or R is fine. This isn't a kaggle style problem, so there isn't a hidden holdout set. How you choose your validation samples is up to you.

Things I am interested in are:

- Feature Selection
- Method of Validation
- Model used (and why?)
- Performance Metrics using for tuning of model

Please send me the jupyter notebook by **15th of Sunday.**

As the timeframe of 2 weeks is quite short for what is a complex problem, I'm not expecting an amazing in depth approach that covers every angle - what is more important is your reasoning and creativity.