Chemistry of Mars

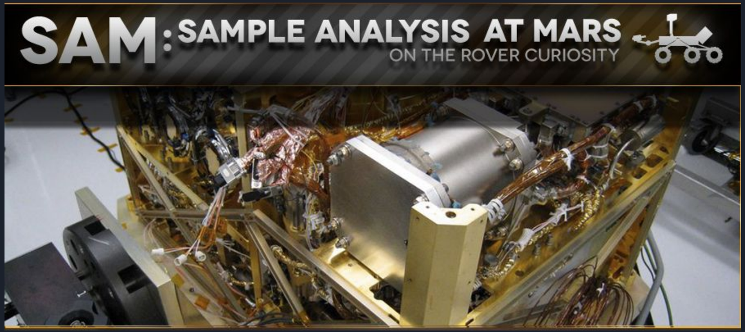
Erdős Data Science Bootcamp Fall 2024

Katherine Martin, Sara Mezuri, Michele Myong, and Nikolas Eptaminitakis

Project Description

NASA posted this challenge on Driven Data:

Goal: “detect the presence of certain families of chemical compounds in geological material samples using evolved gas analysis (EGA) mass spectrometry data collected for Mars exploration missions.”



The Data

Multi-label classification problem

- 754 rock samples with train features and labels
- Train features - EGA data (each csv contains 10,000 rows)

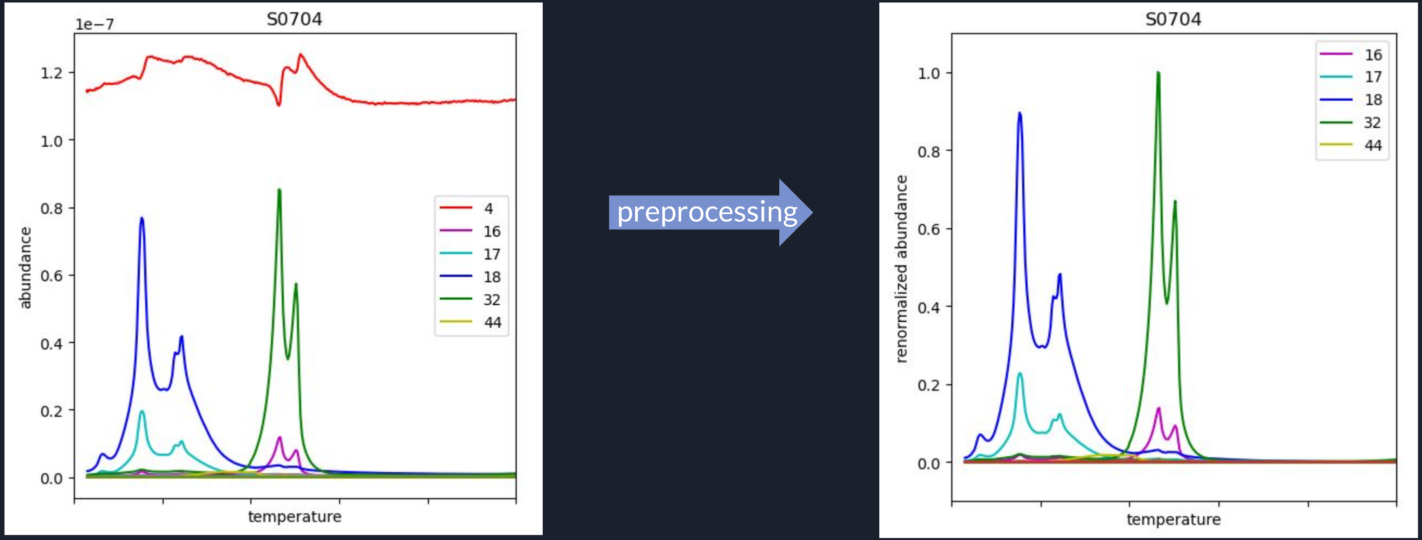
time	temp	m/z (mass-to-charge ratio)	abundance
0.0	-60.37	1.0	2.62e-10

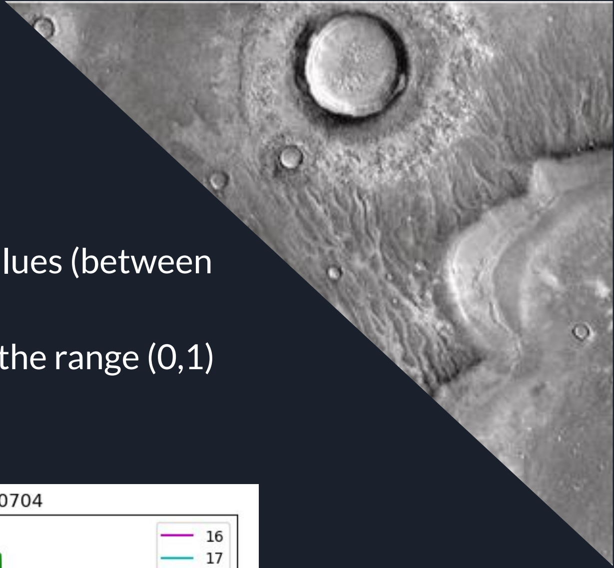
- Train labels (10) - minerals with binary label to indicate presence

sample_id	basalt	carbonate	chloride	iron_oxide	oxalate	oxychlorine	phyllosilicate	silicate	sulfate	sulfide
S0001	0	1	0	0	0	0	0	0	0	0

Data Cleaning

- Drop non-integer m/z values and selecting range of relevant values (between 0-100, ignore the value 4)
- Subtract background abundance and normalize the result into the range (0,1)

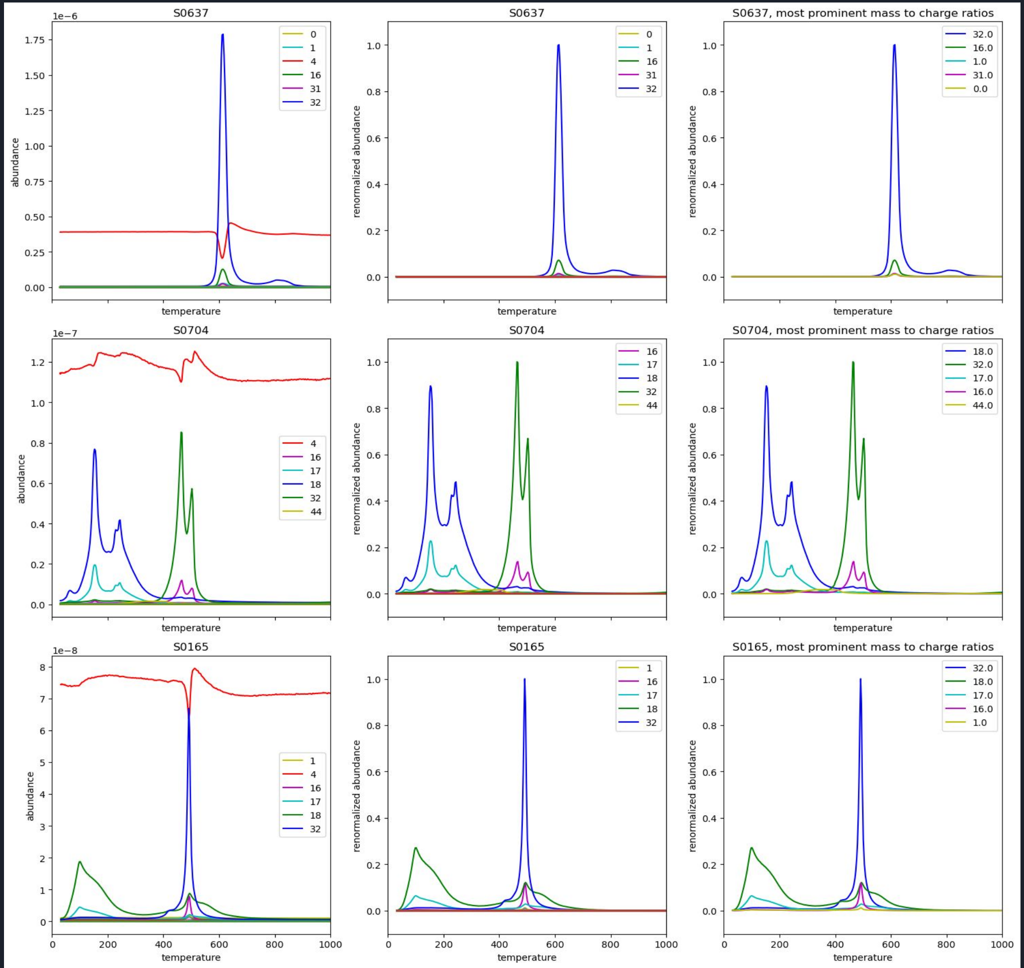




Features Engineering

Choose the five most prominent mass-to-charge ratios, and for them record the ratios themselves, the peak abundance, the temperature at which it occurs, and the sum of the mean and standard deviation for the abundance.

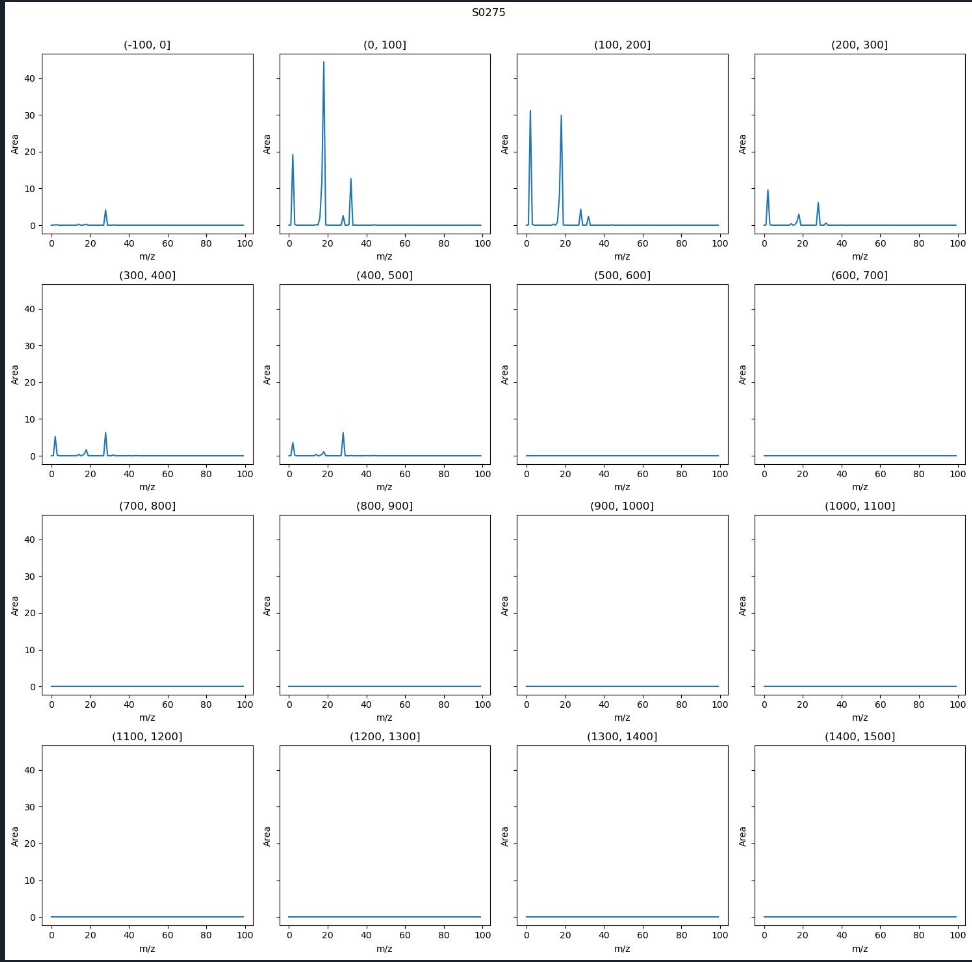
This gives 5·4 = 20 features.



Binning by Temperature

Create 16 temperature intervals and compute the area under the abundance curve over each interval, for each value of the mass to charge ratio. This gives 16·100=1600 features.

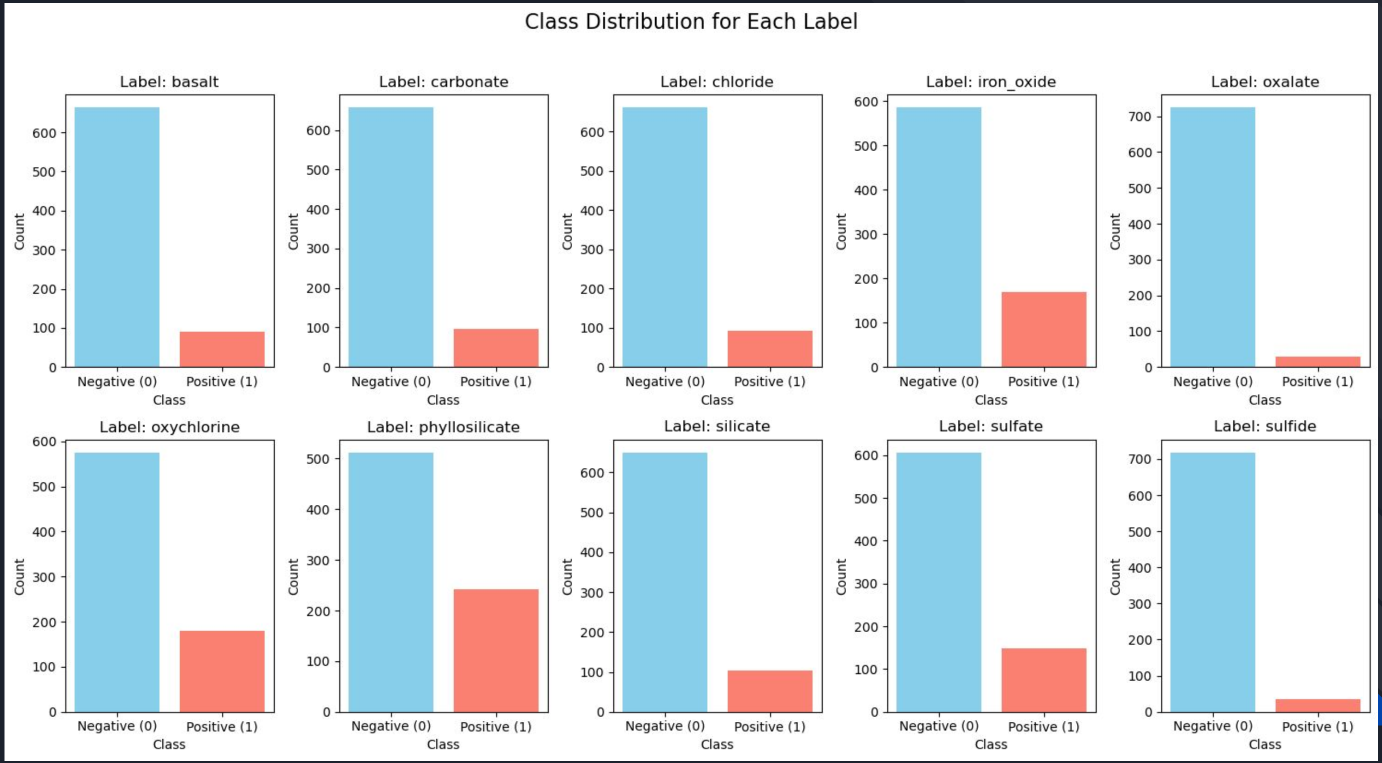
Then for each interval, perform principal component analysis to reduce the total number of features to 16·3=48.



Basic Classification Models

Model	**Training Set Accuracy**	**Per Label Accuracy**	**F1-Score**
Linear Regression	19.9%	84.4%	NA
Logistic Regression	23.8%	87.1%	0.14
Naive Bayes	12.6%	28.6%	0.27
KNN (with n=1)	68.2%	92.3%	0.20
KNN (with n=2)	52.3%	90.9%	0.16
KNN (with n=3)	57.0%	90.8%	0.19
KNN (with n=4)	47.7%	90.9%	0.19
KNN (with n=5)	49.7%	90.1%	0.19

Imbalanced Data



Best Model: Random Forest Classifier

This was done by binary relevance. Binary relevance converts a multi-label classification problem with L labels into L separate binary classification problems, each using the base classifier (we used four different classifiers). The final prediction is the combination of the results from all individual label classifiers. (micro-avg).

Classification Report for the Best Model

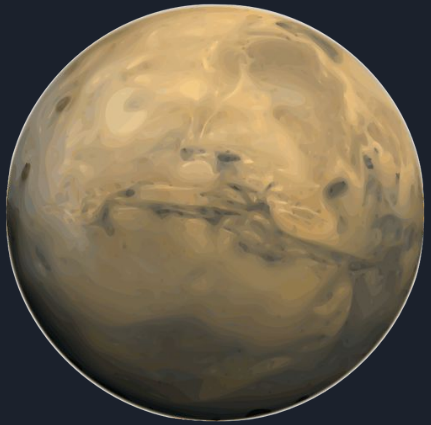
Accuracy for the Random Forest Classifier model: 0.48344370860927155
Hamming Loss for the Random Forest model: 0.08211920529801324

Classification Report for the Random Forest Classifier model:

	precision	recall	f1-score	support
0	0.86	0.32	0.46	19
1	0.80	0.57	0.67	14
2	1.00	0.40	0.57	20
3	0.77	0.52	0.62	33
4	1.00	0.78	0.88	9
5	0.86	0.74	0.79	34
6	0.91	0.67	0.77	48
7	0.71	0.26	0.38	19
8	0.95	0.62	0.75	32
9	1.00	0.50	0.67	4
micro avg	0.88	0.56	0.68	232
macro avg	0.89	0.54	0.66	232
weighted avg	0.88	0.56	0.67	232
samples avg	0.51	0.46	0.47	232

Next Steps

- Talk with NASA scientists about getting recently collected data from Mars rock in a nice format that we can try our highest performing model on.
- Engineer new features using wavelet decompositions of the abundance curves for the various mass-to-charge ratios and use them to train models.



Thank You

The Erdős Institute and the organizers of the Fall Data Science Bootcamp
Especially Stephen Gubkin and our project mentor Soumen Deb