eda_v2_notebook_version

October 20, 2020

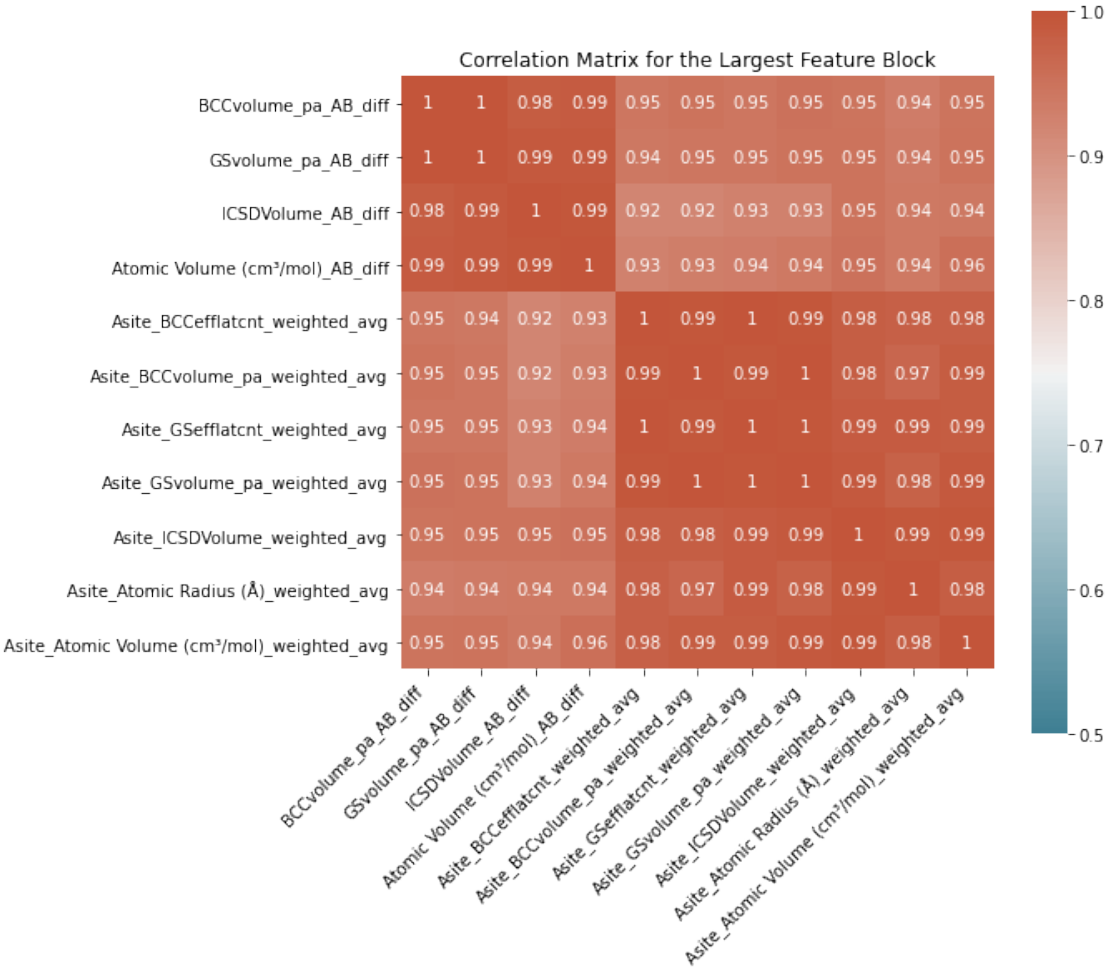
The author has provided a dataset of 1929 Perovskite compounds with 962 features. Since this is a large count of features, through the EDA we aim to remove uninformative features, and also highlight the most informative features in predicting the stability.

0.0.1 1. Variable Scoping

(a) Separate Continuous vs. Discrete Variables Since some of the algorithms used for modeling will need distinguishment of continuous vs. discrete variables, we separate the variable types here. We try to go with the author's intention as much as possible in deciding this, through the continuous and discrete variables mentioned in the code repository (https://github.com/uw-cmg/perovskite-oxide-stability-prediction) prepared by the author.

(b) Remove Variables with Minimal Variance There are some features that have minimal variance, which would be uninformative. We use a threshold of 1e-5 in standard deviation to remove these features.

(c) Remove Redundant Variables Since the author has created derived features combining the element properties of A site / B site with their min / max / average, etc., it is highly likely that these features contain redundant information. To determine the clusters of features that contain redundant information, we run a hierarchical clustering on the feature correlation matrix, with threshold set to 90% (only pairwise correlations above 90% can be considered as a candidate to belong to same cluster). Once we have determined these cluster blocks, there are many ways to represent the feature block for modeling purpose - such as extracting components through PCA, using average of the features, etc. For EDA, we just pick one feature from each cluster block and remove all other redundant features from the block. After removing redundant variables, we are left with 314 continuous variables and 123 discrete variables.



0.0.2 2. Determine Informative Features

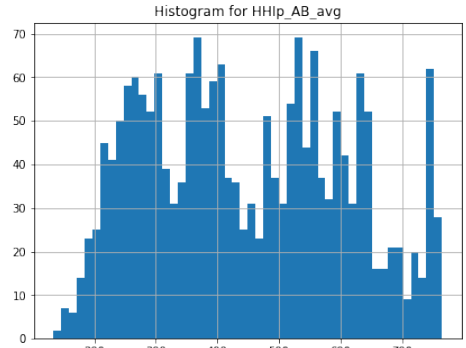
Below, we perform a univariate analysis on feature's relationship to target, to gain insight for the most informative features. This would be beneficial in sanity checking the model result, but since it is a univariate analysis it may not coincide with the model result which could have interaction effects from multiple features.

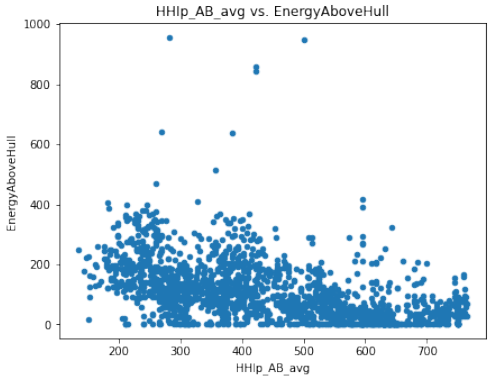
(a) Continuous Variables Here we look at the feature correlation vs. the Energy above convex hull to get insights of the most informative features. Through this analysis, it is suggested that the features **'HHIp_AB_avg', 'Asite_HHIr_min', 'Coeficient of Thermal Expansion (10^-6 K-1)_AB_avg', 'Asite_Coeficient of Thermal Expansion (10^-6 K-1)*max', 'Asite* BP (K)_min'** would be the most informative. Although continuous variables, we can see that the min / max features below show very discrete values of features. Hence, it is likely that tree based models that can handle non-linear splits / groupings of these discrete intervals, would work better than linear models.

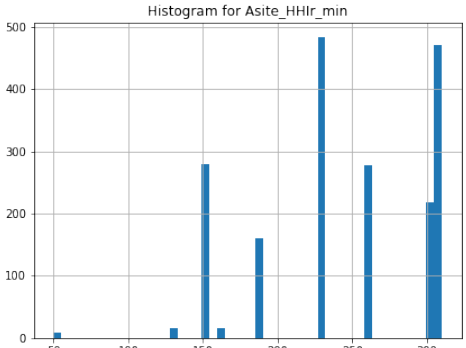
`Top continuous features determined by correlation`

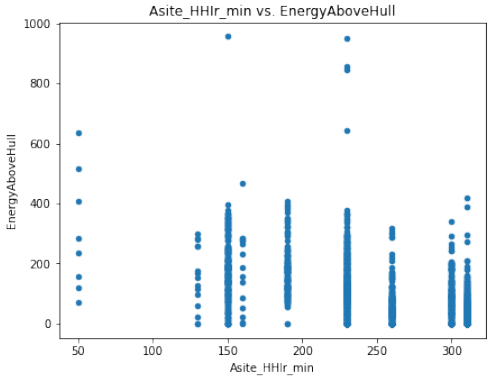
[26]:

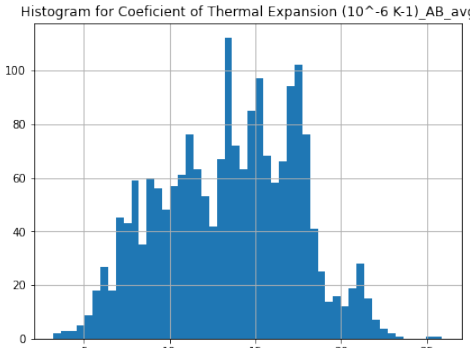
	corr	abs_corr
HHIp_AB_avg	-0.509767	0.509767
Asite_HHIr_min	-0.503562	0.503562
Coeficient of Thermal Expansion (10^-6 K-1)_AB_avg	0.502667	0.502667
Asite_Coeficient of Thermal Expansion (10^-6 K-…	0.491303	0.491303
Asite_ BP (K)_min	-0.477988	0.477988

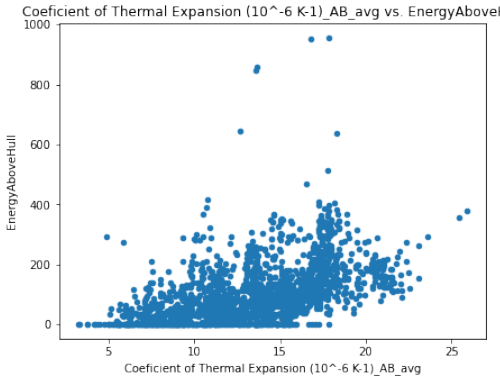

Histogram for HHIp_AB_avg

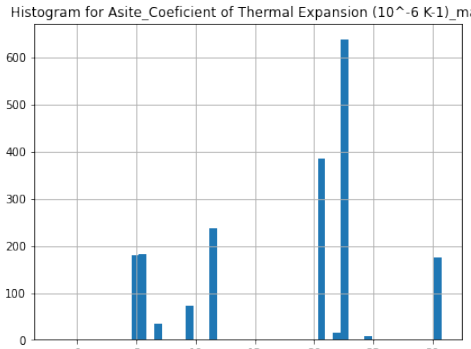

HHIp_AB_avg vs. EnergyAboveHull
EnergyAboveHull
HHIp_AB_avg

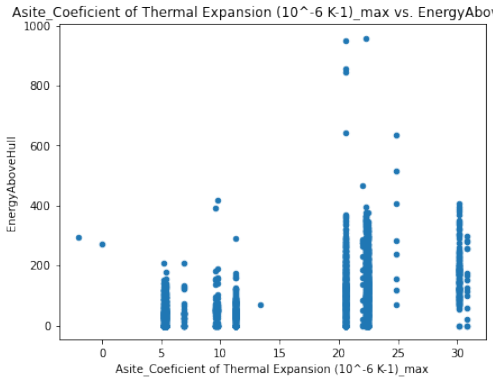

Histogram for Asite_HHIr_min

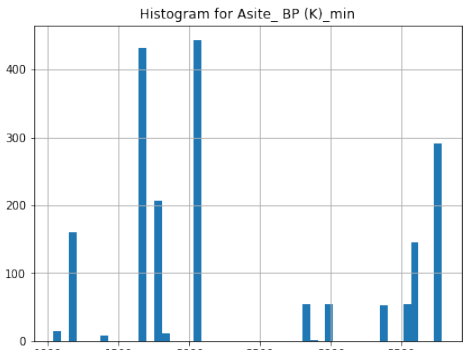

Asite_HHIr_min vs. EnergyAboveHull
EnergyAboveHull
Asite_HHIr_min

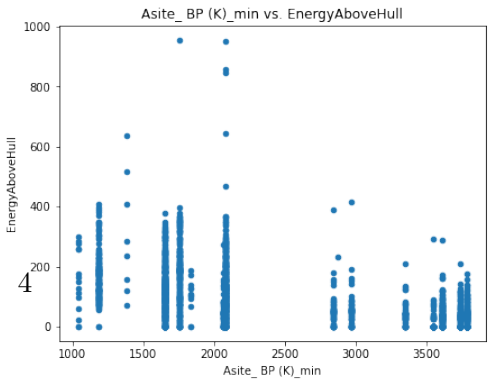

Histogram for Coeficient of Thermal Expansion (10^-6 K-1)_AB_avg


Coeficient of Thermal Expansion (10^-6 K-1)_AB_avg vs. EnergyAboveHull
EnergyAboveHull
Coeficient of Thermal Expansion (10^-6 K-1)_AB_avg


Histogram for Asite_Coeficient of Thermal Expansion (10^-6 K-1)_max


Asite_Coeficient of Thermal Expansion (10^-6 K-1)_max vs. EnergyAboveHull
EnergyAboveHull
Asite_Coeficient of Thermal Expansion (10^-6 K-1)_max


Histogram for Asite_ BP (K)_min

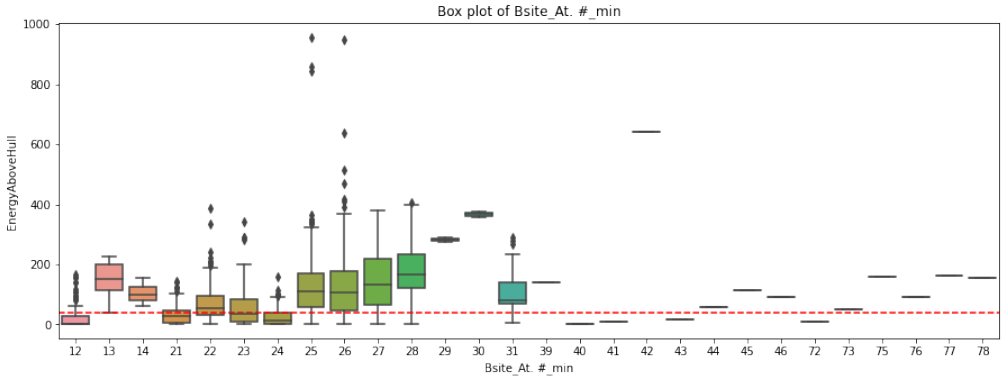

Asite_ BP (K)_min vs. EnergyAboveHull
EnergyAboveHull
Asite_ BP (K)_min

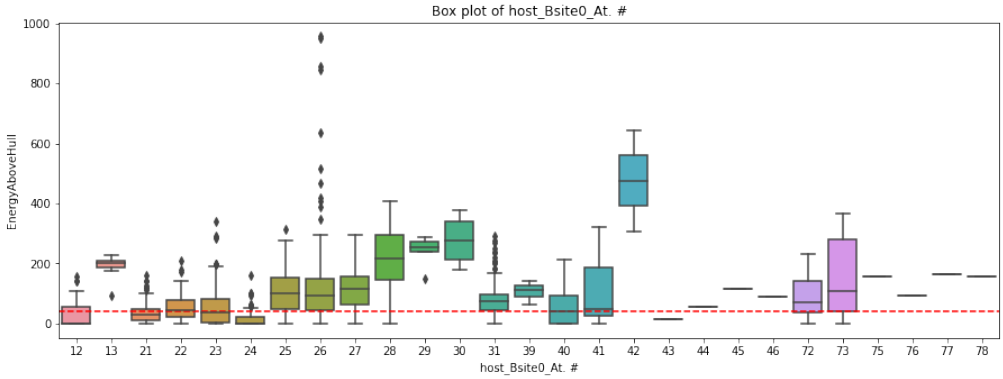
(b) Discrete Variables Here we look at the mutual information gain from the feature and the stability binary variable to get insights of the most informative features. We use the same criteria as author of Energy above convex hull being under 40meV/atom to determine stability of the compound. Through this analysis, it is suggested that the features **'Bsite_At. #_min', 'host_Bsite0_At. #', 'host_Bsite0_Group', 'host_Bsite0_NUnfilled', 'Asite_At. #_min'** would be the most informative discrete variables. From the box plots we can see that these features do seem quite useful, as depending on the value the range of energy above convex hull is quite different.

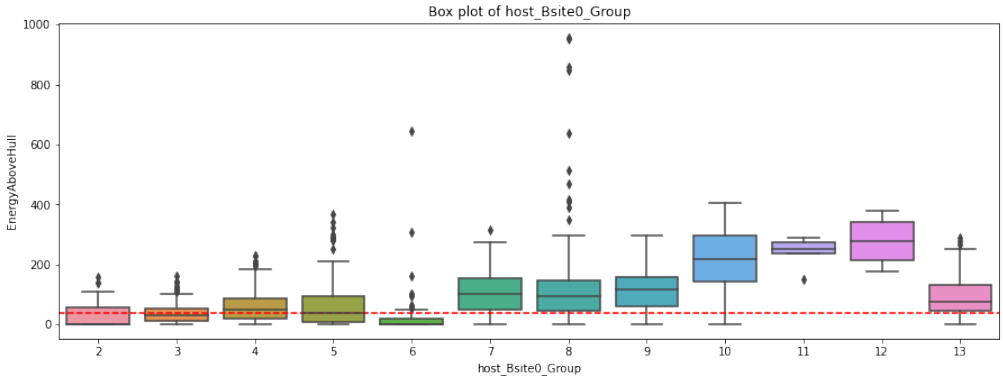
```
Top discrete features determined by mutual information gain
```

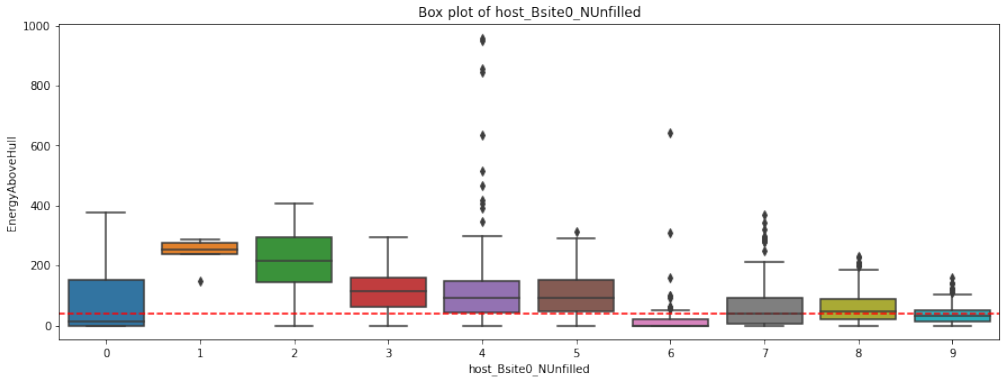
[31]:

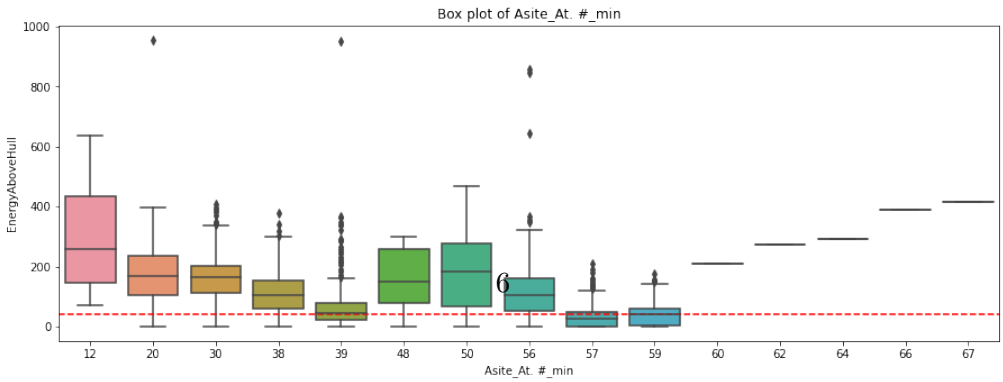
	feature	MI
81	Bsite_At. #_min	0.130
18	host_Bsite0_At. #	0.121
19	host_Bsite0_Group	0.113
31	host_Bsite0_NUnfilled	0.109
69	Asite_At. #_min	0.106


Box plot of Bsite_At. #_min
EnergyAboveHull
Bsite_At. #_min


Box plot of host_Bsite0_At. #
EnergyAboveHull
host_Bsite0_At. #


Box plot of host_Bsite0_Group
EnergyAboveHull
host_Bsite0_Group


Box plot of host_Bsite0_NUnfilled
EnergyAboveHull
host_Bsite0_NUnfilled


Box plot of Asite_At. #_min
EnergyAboveHull
Asite_At. #_min