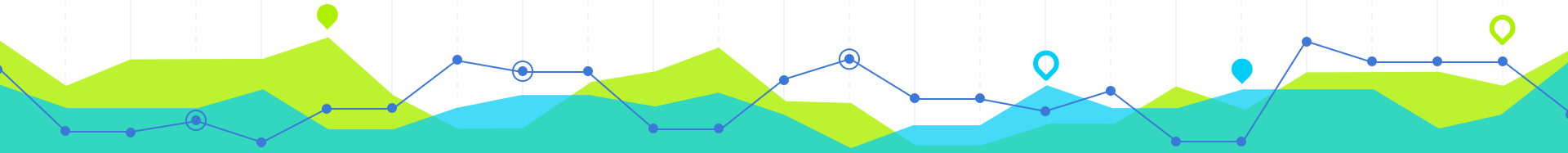


Team Mutants303

Anshuman Chakravarty
Arnab Manna
Sushovan Haldar

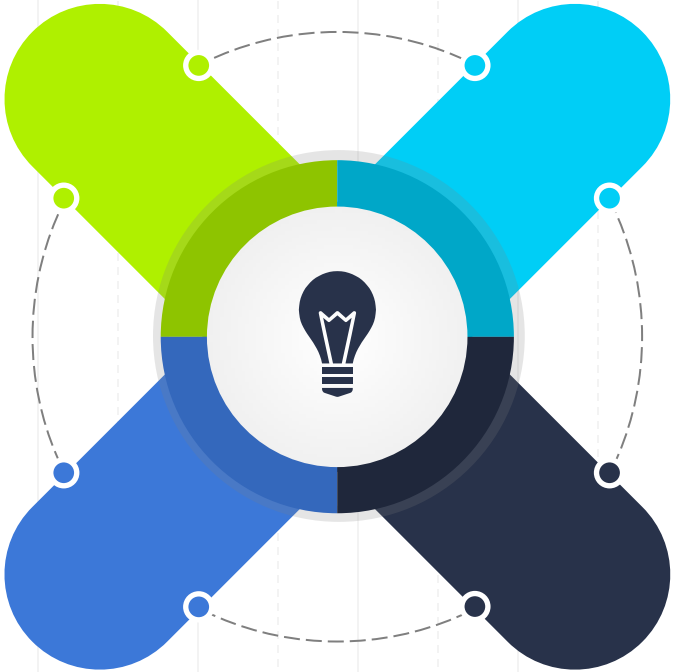
Data Description

The dataset consists of 11 different features of the alloy of conducting materials and the goal is to predict two properties i.e., the formation energy and band gap energy to facilitate the discovery of new transparent conductors.

Steps

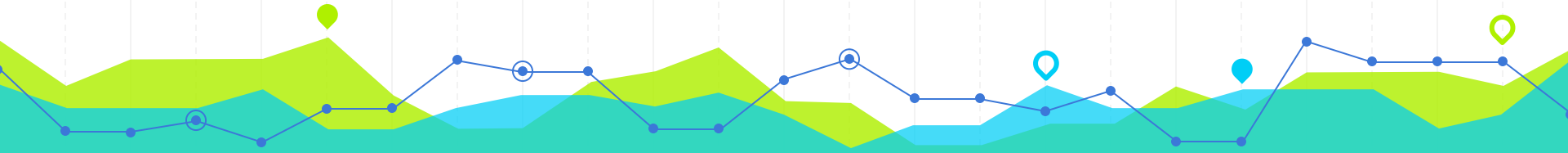
Exploratory Data Analysis

Data Preparation



Model Selection

Model Fitting



Exploratory Data Analysis

1

Data Exploration

The dataset contains 1900 rows and 14 columns, the last two columns being the target variables. It does not contain any null values and hence no data imputation was needed.

```
data.isnull().sum()
```

id	0
spacegroup	0
number_of_total_atoms	0
percent_atom_al	0
percent_atom_ga	0
percent_atom_in	0
lattice_vector_1_ang	0
lattice_vector_2_ang	0
lattice_vector_3_ang	0
lattice_angle_alpha_degree	0
lattice_angle_beta_degree	0
lattice_angle_gamma_degree	0
formation_energy_ev_natom	0
bandgap_energy_ev	0

dtype: int64

Data Exploration

There are two categorical variables “*spacegroup*” & “*number_of_total_atoms*” and the rest predictor variables are continuous variables.

The features “*spacegroup*” & “*number_of_total_atoms*” were mapped as follows :

```
data['number_of_total_atoms'].unique()
```

```
array([80., 40., 30., 20., 60., 10.])
```

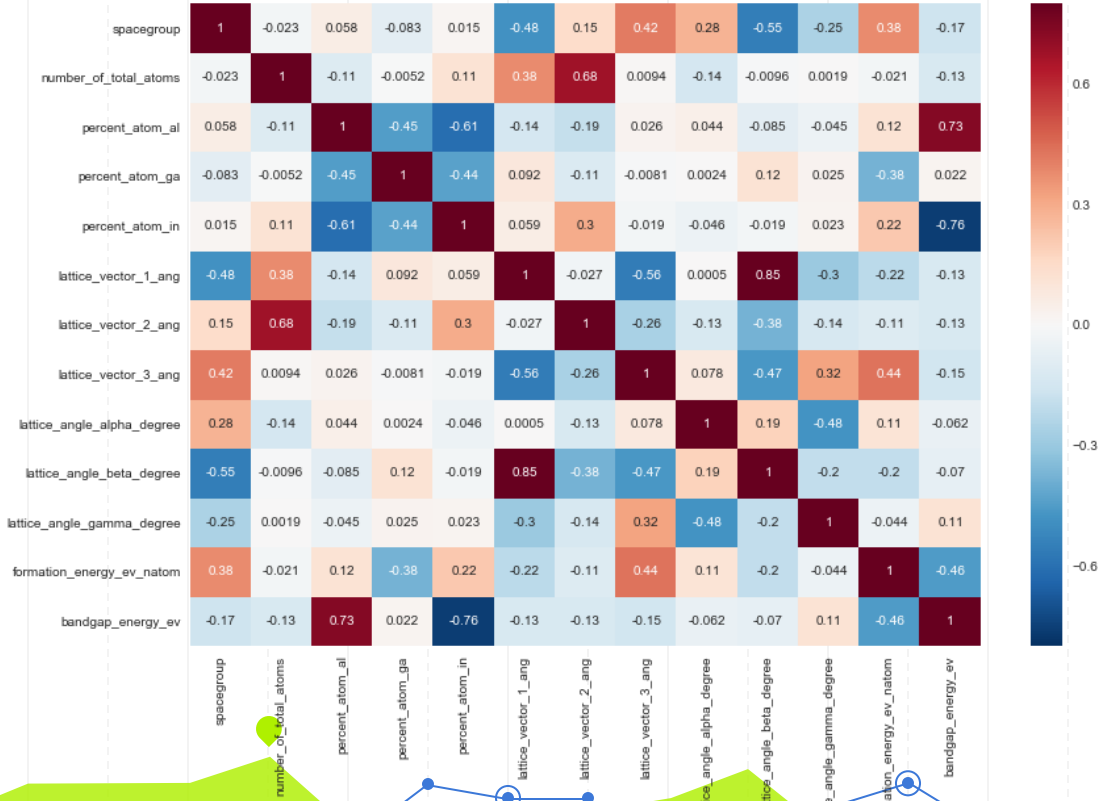
```
data['number_of_total_atoms'] = data['number_of_total_atoms'].map({10: 0,20:1,30:2,40:3,60:4,80:5})
```

```
data['spacegroup'].unique()
```

```
array([ 33, 194, 227, 167, 206,  12], dtype=int64)
```

```
data['spacegroup'] = data['spacegroup'].map({12: 0,33:1,167:2,194:3,206:4,227:5})
```

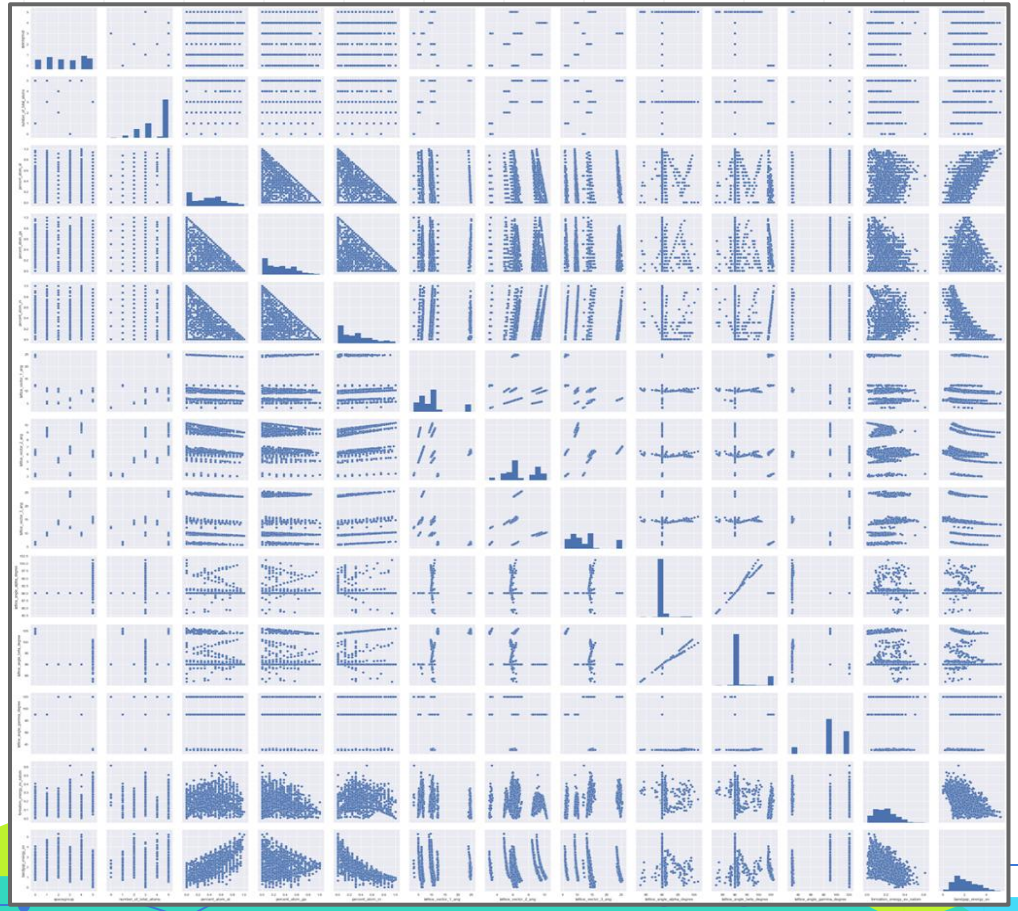
Correlation Matrix



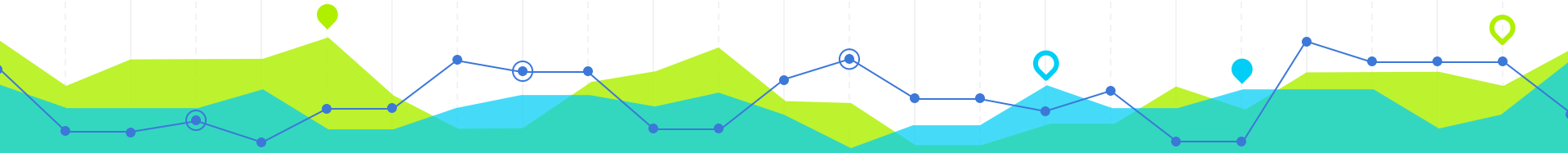
The correlation between different variables were plotted on a heatmap.

Since no variables were found out to be highly correlated, so no variables were dropped.

Pairplotting features



We get a grid of plots for each variable in our dataset. Hence, we can quickly see how all the variables are related. This can help to infer which variables are useful, and which have skewed distribution.



Data Preparation 2

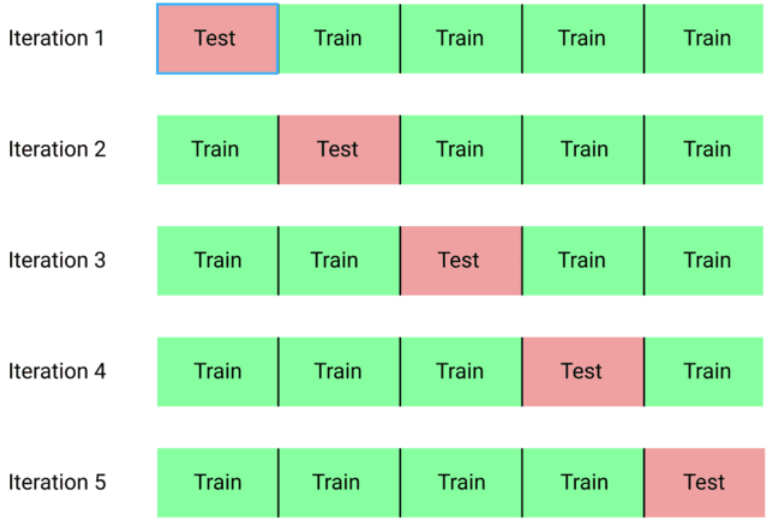
Train-Test Split

The Dataset was split (70:30) into

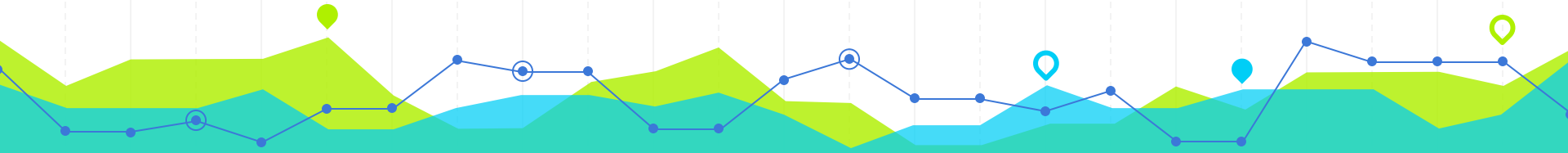
- Train Data: Used to train the different models.
- Test Data: Separate Dataset strata on which the model was tested upon to find the various parameters. This dataset can be thought of as an unknown Dataset on which the model can be applied.

The total Dataset of 1900 was split into 1330:570 for Train and Test data respectively.

k-Fold Cross Validation



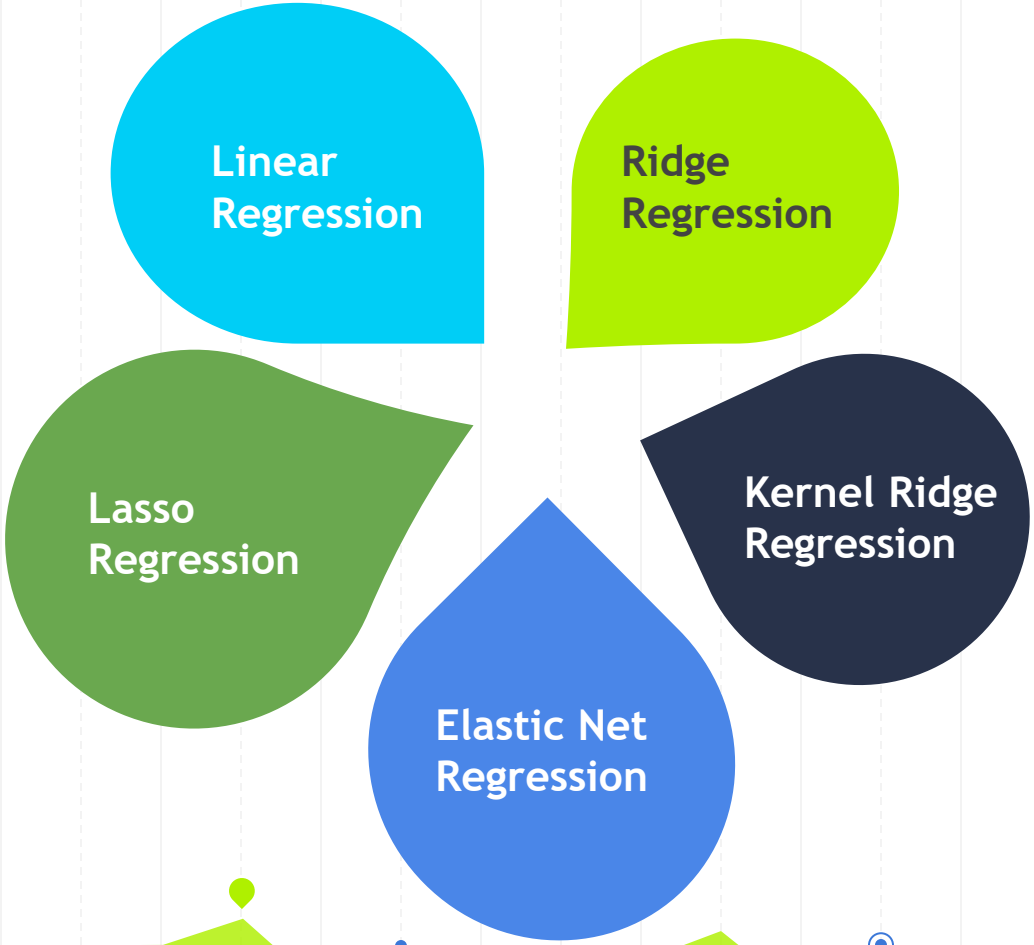
The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.



Fitting Models

3

Regression techniques



Linear Regression

Linear regression is a linear approach for modelling the relationship between a scalar dependent variable *y* and one or more independent variable denoted by *X* .

The **RMSE** error found on the cross validation set for the attribute **“Formation energy of an atom”** is around **0.0801**.

The **RMSE** error found on the cross validation set for the attribute **“Band gap energy”** is around **0.437**.

Ridge Regression

Ridge Regression adds a small bias factor to the predictor variables because in most of the cases the variables turn out to be highly collinear in nature.

The **RMSE** error found on the cross validation set for the attribute **"Formation energy of an atom"** is around **0.08083**.

The **RMSE** error found on the cross validation set for the attribute **"Band Gap Energy"** is around **0.43714**.

Lasso Regression

Lasso Regression (Least absolute shrinkage and selection operator) is a type of regression technique that performs both variable selection and regularization in order to improve the accuracy of the model.

The **RMSE** error found on the cross validation set for the attribute **"Formation energy of an atom"** is around **0.08013**.

The **RMSE** error found on the cross validation set for the attribute **"Band Gap Energy"** is around **0.43691**.

Elastic Net Regression

Elastic Net regression uses linear combination of both L1 and L2 regularization penalties of the lasso and ridge regression.

The **RMSE** error found on the test set for the attribute **“Formation energy of an atom”** is **0.05505**.

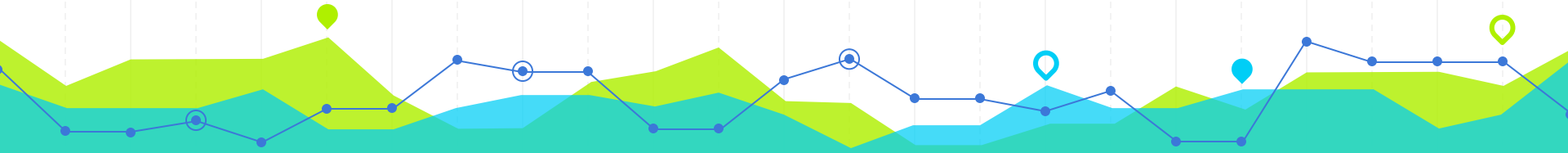
The **RMSE** error found on the cross validation set for the attribute **“Band Gap Energy”** is around **0.26913**.

Kernel Ridge Regression

Kernel ridge Regression combines Ridge Regression with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a nonlinear function in the original space.

The **RMSE** error found on the cross validation set for the attribute "**Formation energy of an atom**" is around **0.080123**.

The **RMSE** error found on the cross validation set for the attribute "**Band Gap Energy**" is around **0.43684**.



Boosting Techniques

Gradient Boosting Regressor

Gradient boosting regressor produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The **RMSE** error found on the cross validation set for the attribute **"Formation energy of an atom"** is around **0.04389**.

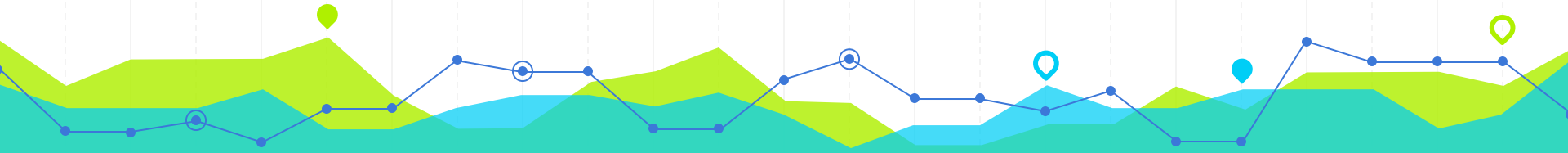
The **RMSE** error found on the cross validation set for the attribute **"Band Gap Energy"** is around **0.24247**.

XgBoost Regressor

Xgboost regressor produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees, similar to the Gradient Boosting Regressor. But uses a more regularized model formalization to control over-fitting, which gives it better performance.

The **RMSE** error found on the cross validation set for the attribute **"Formation energy of an atom"** is around **0.04887**.

The **RMSE** error found on the cross validation set for the attribute **"Band Gap Energy"** is around **0.23624**.



Model Selection 4

Model Selection

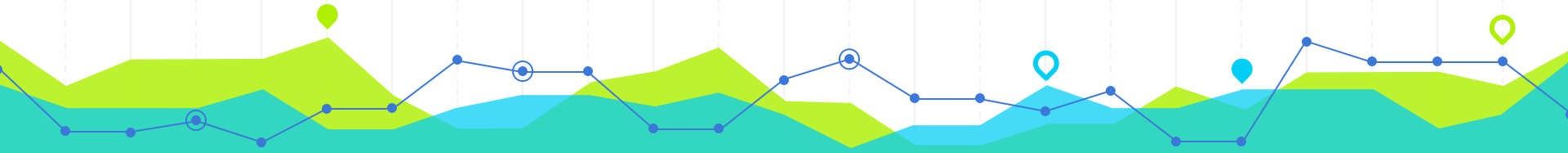
The criterion for choosing the best model is Root Mean Square Error (**RMSE**) . Going through the performances of different models,

formation_energy_ev_natom

	CV Mean	Std
Linear Regression	0.080125	0.006789
Ridge Regression	0.080383	0.007312
Lasso Regression	0.080127	0.001054
Kernel Ridge Regression	0.080123	0.001050
Elastic Net Regression	0.055049	0.002059
Gradient Boosting	0.043897	0.004217
XGBoost	0.048877	0.002607
Averaging Models	0.054580	0.001658

bandgap_energy_ev

	CV Mean	Std
Linear Regression	0.437077	0.010724
Ridge Regression	0.437144	0.010982
Lasso Regression	0.436812	0.011172
Kernel Ridge Regression	0.436846	0.001050
Elastic Net Regression	0.269135	0.016662
Gradient Boosting	0.242477	0.009324
XGBoost	0.236241	0.012923
Averaging Models	0.275858	0.011823



Hence, based on the calculated mean and standard deviations of the RMSE values we get the best accuracy with

Gradient Boosting Classifier

THANKS!

Any questions?