Winter 2018 – PREDICT 454 Section 55

Northwestern University

Midterm: "Nomad2018 Predicting Transparent Semiconductors"

Predictions of Formation and Bandgap Energy

Pearson, Daniel

DanielPearson2018@u.northwestern.edu

February 11th, 2018

Problem Statement

Compounds that are both stable and transparent are called transparent conductors. These compounds conduct electricity and do not absorb much visible light. These compounds are essential components of modern technologies and will almost certainly be part of some pretty world-altering inventions down the road, particularly in the field of energy. Aluminum, gallium, and indium sesquoixides are some of the most effective semiconductors. Unfortunately, discovering new compounds based on these structures requires calculations built with density-functional theory[1]. These calculations are both time-consuming and expensive. As a result, Nomad asked participants in Kaggle competition to come up with models that can be used to predict bandgap energy and formation energy of different structures to facilitate the discovery of new transparent semiconductors. Models have built to predict bandgap energy[2] and formation energy[3] for other types of semiconductors.

I chose to participate in this competition for the midterm project. There were two sets of data: a training dataset containing 2,400 training records and a dataset containing 600 test records. I conducted predictive modeling in four phases: 1) Data exploration, 2) Data preparation, 3) Building models, and 4) Selecting the final model. My test predictions were uploaded to Kaggle under the username DanPNU.

Data Exploration

The dataset used to build the models contained 2,400 records and 14 numeric variables. The test data contained 600 records and 12 numeric variables. Table 1 lists the variables and a description of each. An abbreviation for each variable is listed next to the variable name in Table 1. These abbreviations will be used for the remainder of this report.

Before conducting exploratory data analysis, I merged the training and test datasets. Summary statistics for the variables are listed in Table 2. The values in Table 2 make sense. None of the variables have negative values. All of the composition numbers are between 0 and 1. All of the angular data are between 0 and 360 degrees. The data in Table 2 also show that the only missing variables are the target variables (Formation Energy and Bandgap Energy) for the 600 test observations. Therefore, I did not have to create any flag variables to note missing values.

1. Kaduk, B.; Kowalczyk, T.; Van Voorhis, T.; Constrained Density Functional Theory. Chemical Reviews. 2012, 112, 321−370.

2. Taoreed O. Owolabi, Kabiru O. Akande, Sunday O. Olatunji, Nahier Aldhafferi, & Abdullah Alqahtani. (2017). Modeling energy band gap of doped TiO2 semiconductor using homogeneously hybridized support vector regression with gravitational search algorithm hyper-parameter optimization. *AIP Advances, 7*(11), 115225-115225-13.

3. Zhao, C., Li, N., Wei, T., Wang, S., & Lu, K. (2013). Model for the formation energy of In–N clusters and their effect on the energy band gap of the Ga-rich and As-rich InxGa1−xNyAs1−y semiconductor alloys. *Physica B: Condensed Matter, 414*, 115-118.

Table 1: List of Variables in Dataset

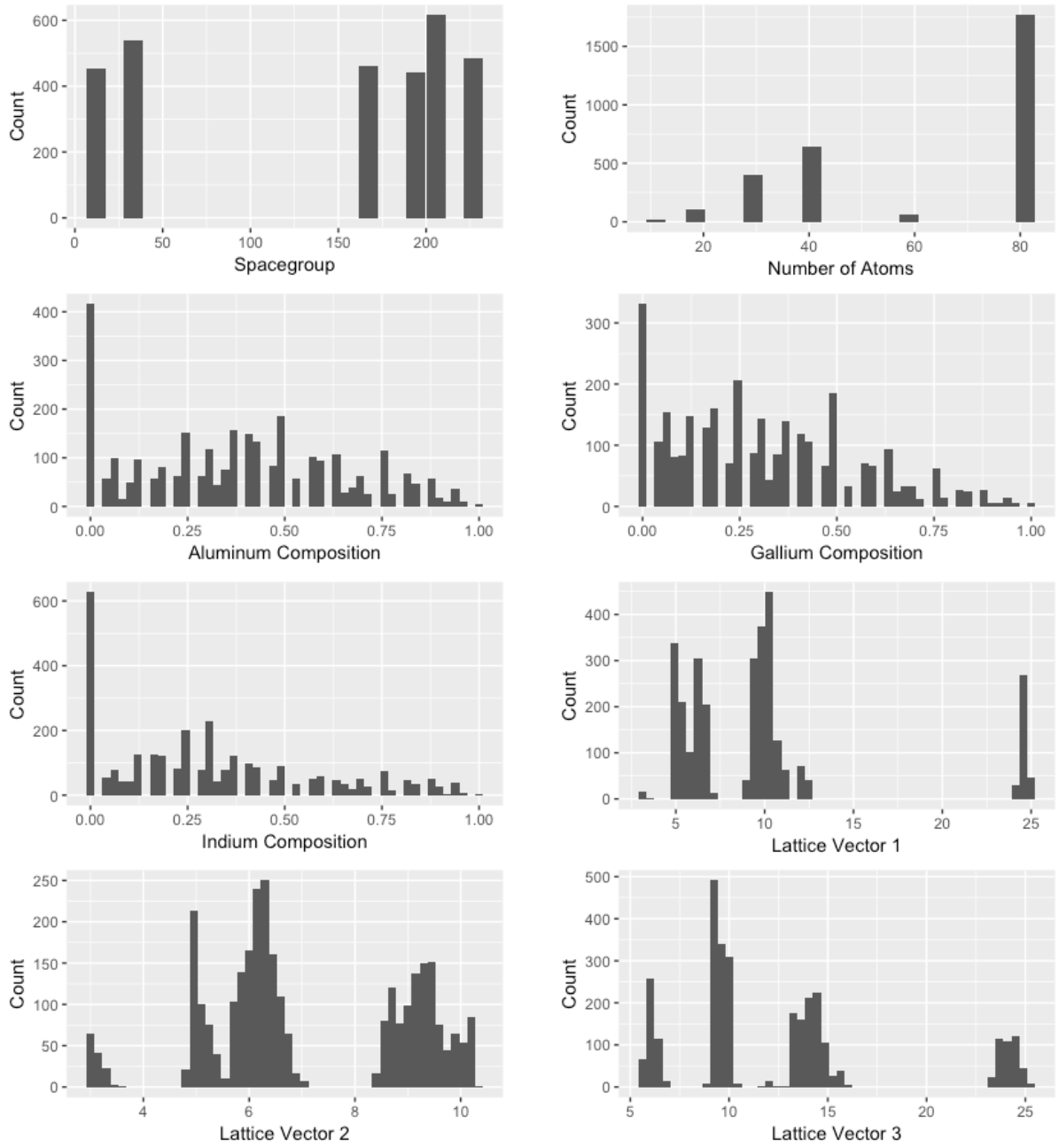
Variable	Description
id (ID)	Identification variable
spacegroup (SG)	Measure of symmetry
number_of_total_atoms (N)	Total number of atoms in the unit cell
percent_atom_al (Al)	Relative composition of Aluminum
percent_atom_ga (Ga)	Relative composition of Gallium
percent_atom_in (In)	Relative composition of Indium
lattice_vector_1_ang (LV1)	Length of lattice vector 1 in angstroms
lattice_vector_2_ang (LV2)	Length of lattice vector 2 in angstroms
lattice_vector_3_ang (LV3)	Length of lattice vector 3 in angstroms
lattice_angle_alpha_deg (Alpha)	Lattice angle alpha in degrees
lattice_angle_beta_deg (Beta)	Lattice angle beta in degrees
lattice_angle_gamma_deg (Gamma)	Lattice angle gamma in degrees
formation_energy_ev_natom (Formation Energy)	Formation energy in electron volts
bandgap_energy_ev (Bandgap Energy)	Bandgap energy in electron volts

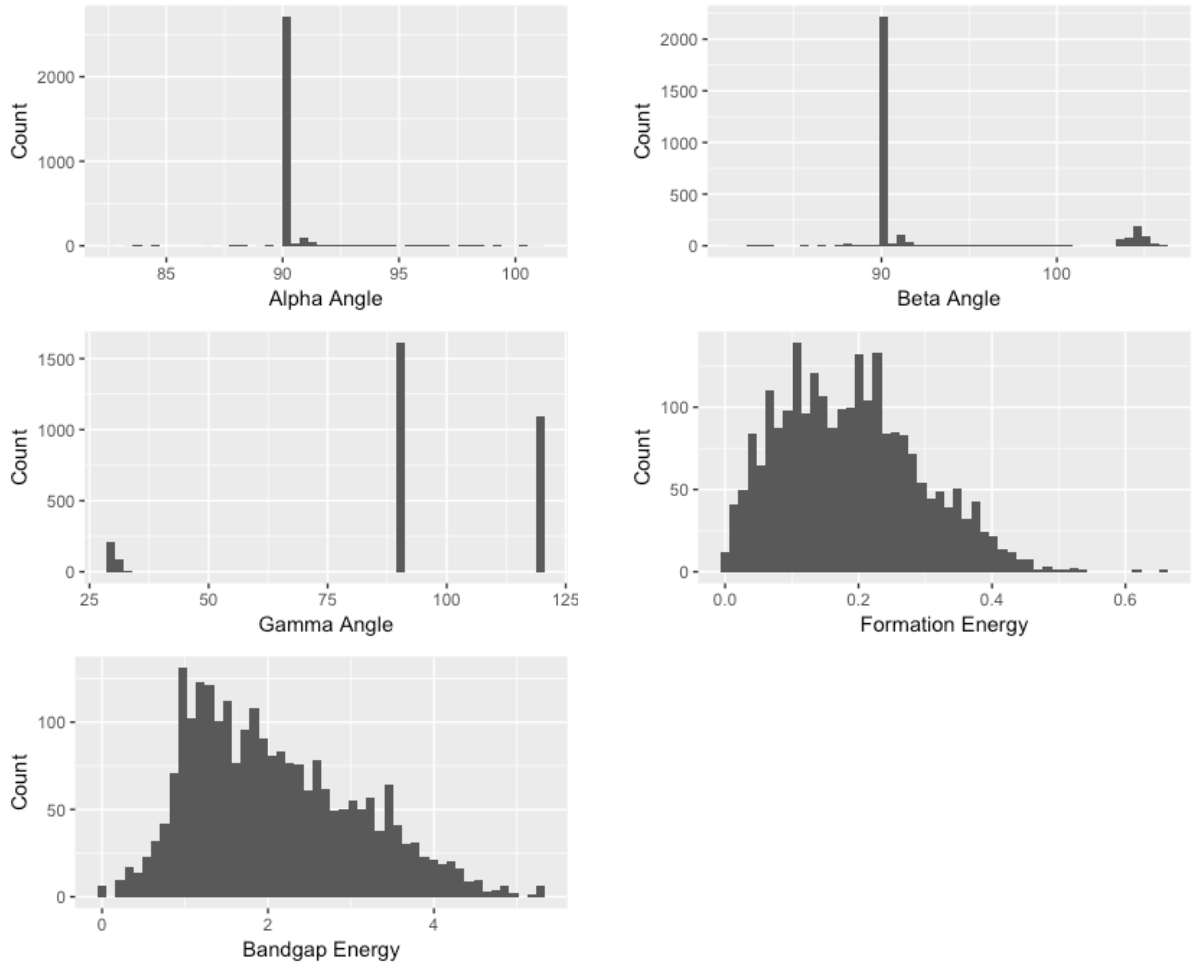
Table 2: Summary Statistics for Dataset

```
      id          spacegroup    number_of_total_atoms percent_atom_al  percent_atom_ga  percent_atom_in  lattice_vector_1_ang
Min.   :   1.0   Min.   : 12.0   Min.   :10.00         Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 3.037
1st Qu.: 375.8   1st Qu.: 33.0   1st Qu.:40.00         1st Qu.:0.1562   1st Qu.:0.0938   1st Qu.:0.0625   1st Qu.: 6.138
Median : 900.5   Median :194.0   Median :80.00         Median :0.3750   Median :0.2812   Median :0.2500   Median : 9.532
Mean   :1020.5   Mean   :141.1   Mean   :61.69         Mean   :0.3826   Mean   :0.3095   Mean   :0.3079   Mean   :10.044
3rd Qu.:1650.2   3rd Qu.:206.0   3rd Qu.:80.00         3rd Qu.:0.5625   3rd Qu.:0.4688   3rd Qu.:0.4688   3rd Qu.:10.309
Max.   :2400.0   Max.   :227.0   Max.   :80.00         Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :24.913

lattice_vector_2_ang lattice_vector_3_ang lattice_angle_alpha_degree lattice_angle_beta_degree lattice_angle_gamma_degree
Min.   : 2.942       Min.   : 5.673       Min.   : 82.74            Min.   : 81.64           Min.   : 29.72
1st Qu.: 5.832       1st Qu.: 9.301       1st Qu.: 90.00            1st Qu.: 90.00           1st Qu.: 90.00
Median : 6.386       Median :10.121       Median : 90.00            Median : 90.00           Median : 90.00
Mean   : 7.086       Mean   :12.563       Mean   : 90.23            Mean   : 92.42           Mean   : 95.10
3rd Qu.: 9.098       3rd Qu.:14.370       3rd Qu.: 90.01            3rd Qu.: 90.01           3rd Qu.:120.00
Max.   :10.290       Max.   :25.346       Max.   :101.23            Max.   :106.17           Max.   :120.05

formation_energy_ev_natom bandgap_energy_ev
Min.   :0.0000            Min.   :0.0001
1st Qu.:0.1056            1st Qu.:1.2785
Median :0.1818            Median :1.9078
Mean   :0.1876            Mean   :2.0772
3rd Qu.:0.2563            3rd Qu.:2.7620
Max.   :0.6572            Max.   :5.2861
NA's   :600               NA's   :600
```

Figure 1 shows the distribution of every variable in the original data set.

Figure 1: Variable Distributions



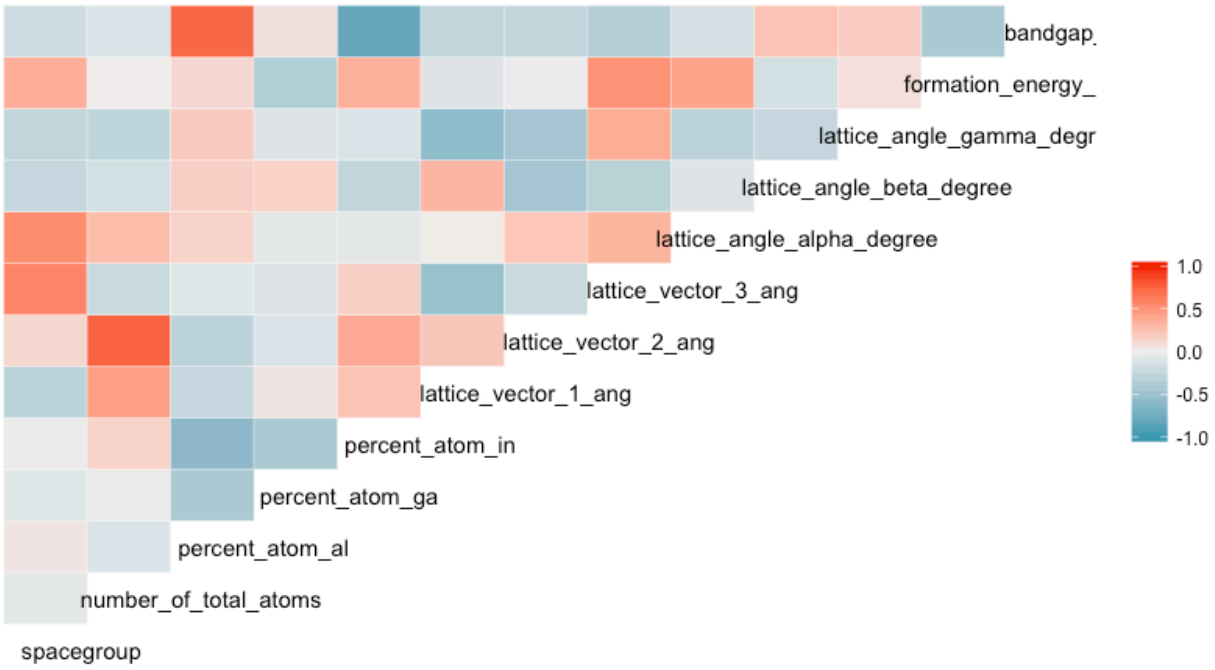


The histograms show that there are only 6 unique values of SG and 6 unique values of N. The composition data (Al, Ga, In) are essentially zero-inflated poisson distributions. This is because aluminum, gallium, and indium were each absent from a sizeable number of the structures. The lattice vector lengths appear to have 3 or 4 separate distributions per vector. Alpha and Beta distributions show basically two separate distributions with a small number of other values that don't really fit with the majority of the data. The Gamma histogram shows three separate distributions. The Bandgap and Formation Energy distributions have a positive skew. The Formation Energy histogram shows that that the distribution may be bimodal.

I elected not to remove what look like outliers for this exercise. I decided not to for two reasons. 1) The values that look like outliers do not look they were recorded incorrectly and 2) I don't know if this dataset is random sample or if these values were targeted specifically.

Examining relationships between variables is a critical component of exploratory data analysis. Figure 2 is a correlation plot. The opaquer a color is (either red or blue), the stronger the linear correlation between the variables. Most of the variables don't appear to be heavily correlated with either Bandgap Energy or Formation Energy. However, Bandgap Energy does look to have a positive correlation with Al and a negative correlation with In. N appears to have a positive correlation with LV2. In and Al appear to be negatively correlated.

Figure 2: Correlation Plot



It is valuable to view scatterplots of these relationships. Figure 3 shows Bandgap Energy vs composition. Figure 4 shows correlations that were not identified in Figure 2 as strong correlations. The lattice vectors (LV1, LV2, and LV3) appear to be correlated strongly with Bandgap Energy and somewhat correlated with Formation Energy. The reason the lattice vectors were not recognized as highly correlated with the target variables is because there are several separate linear relationships evident. The correlation plot fits a correlation to the entire dataset. Had the correlation function looked at "clusters" of data separately, a very strong correlation would have been recorded.

Figure 3: Bandgap Energy vs. Composition

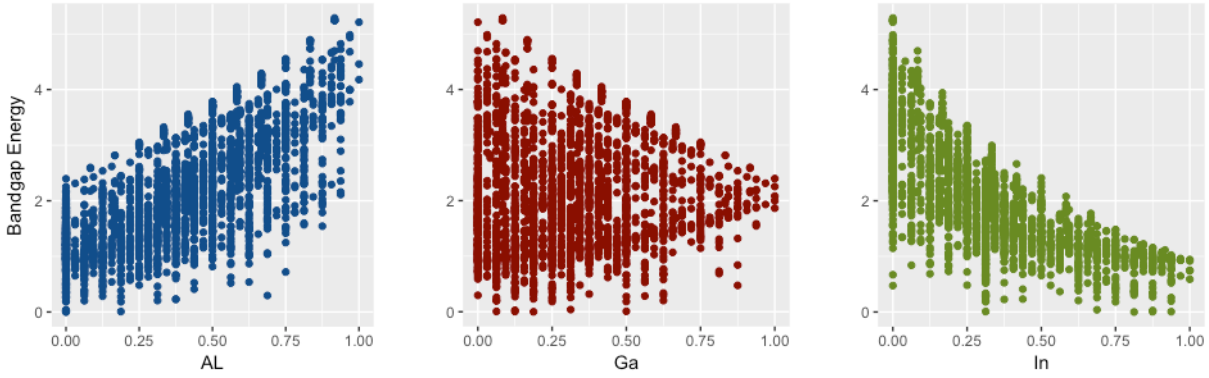
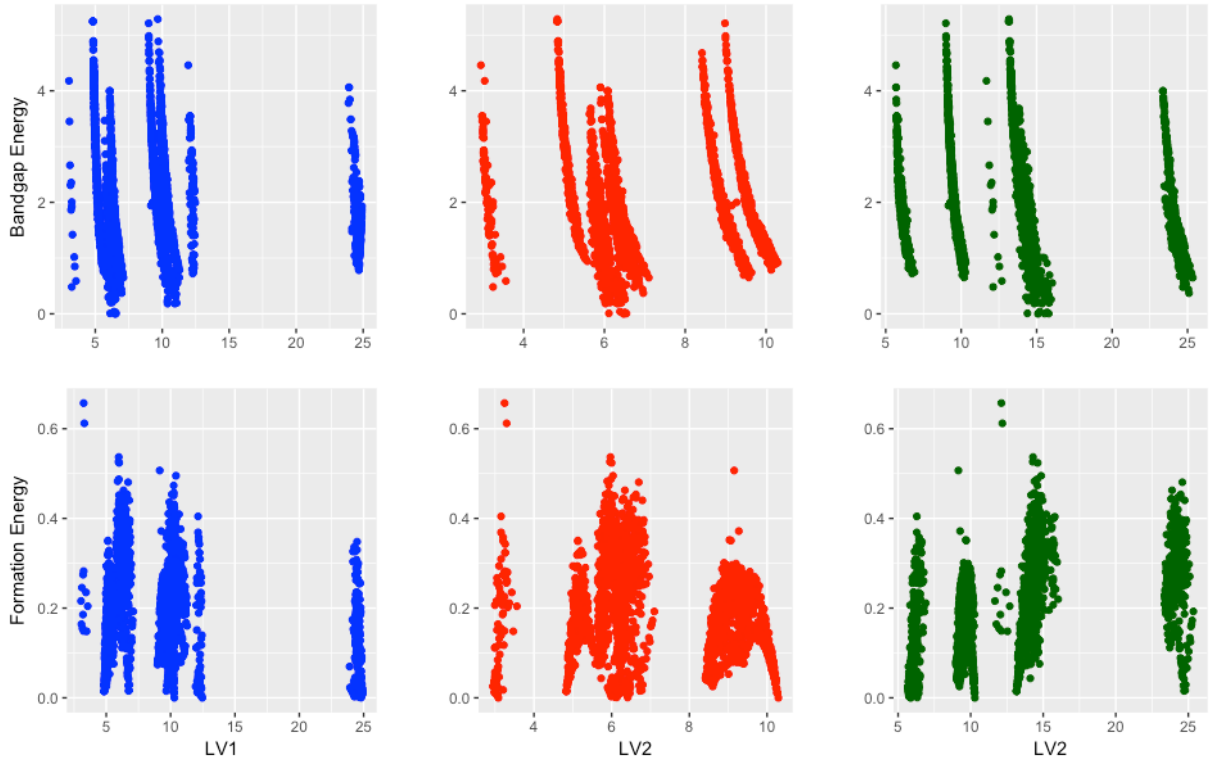
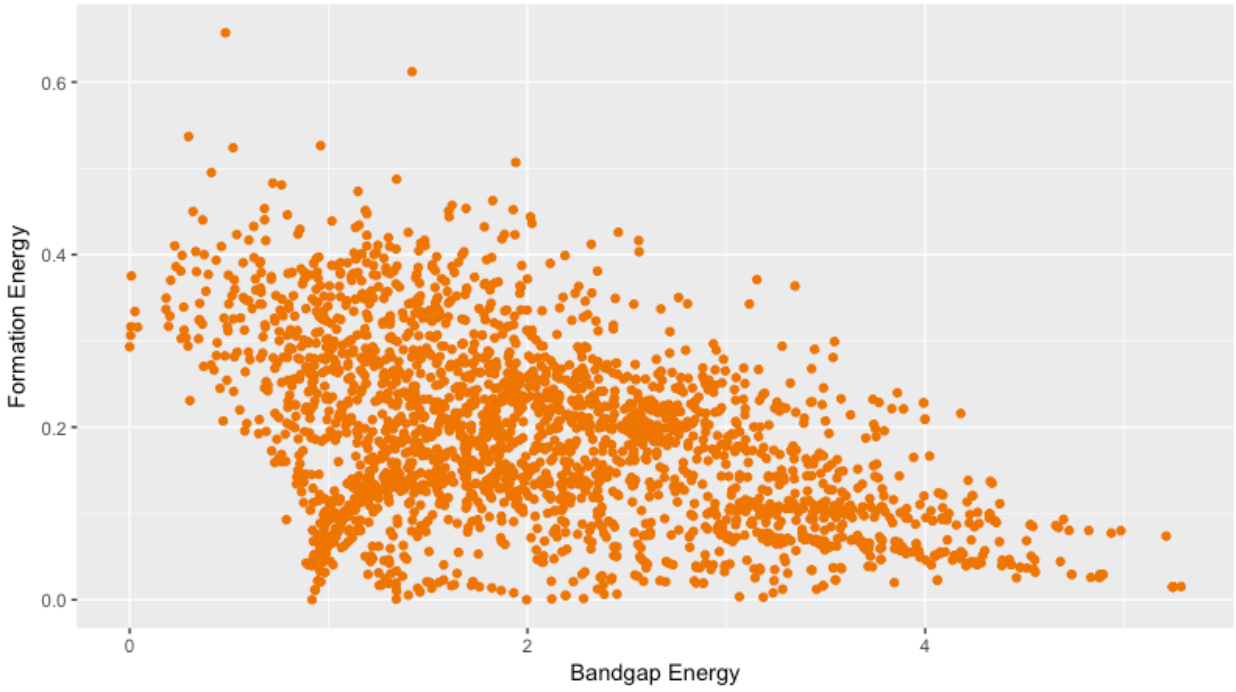


Figure 4: Bandgap Energy and Formation Energy vs. Lattice Vector Lengths



One difficulty was figuring out what correlated with Formation Energy. Most variables had stronger relationships with Bandgap Energy. However, it appears that Bandgap Energy could be predictive of Formation Energy.

Figure 5: Formation Energy vs. Bandgap Energy



Data Preparation

Three new variables were added to the data set before the models were built. The first was molar weight. Molar weight (MW) was calculated with the molar weights of the individual compounds as:

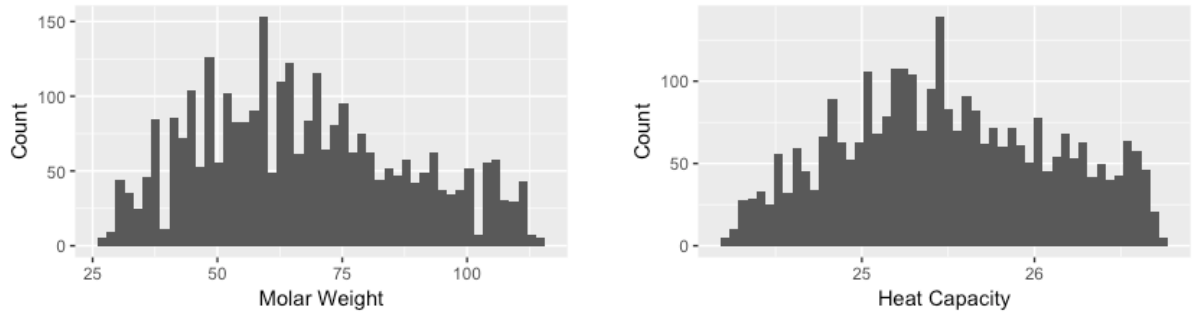
$$MW = 26.98 \cdot Al + 69.72 \cdot Ga + 114.82 \cdot In$$

The next variable added to the dataset was heat capacity (Cp). This calculation was similar to the calculation of molar weight. That is, it was calculated using the heat capacities of individual compounds.

$$Cp = 24.20 \cdot Al + 25.86 \cdot Ga + 26.74 \cdot In$$

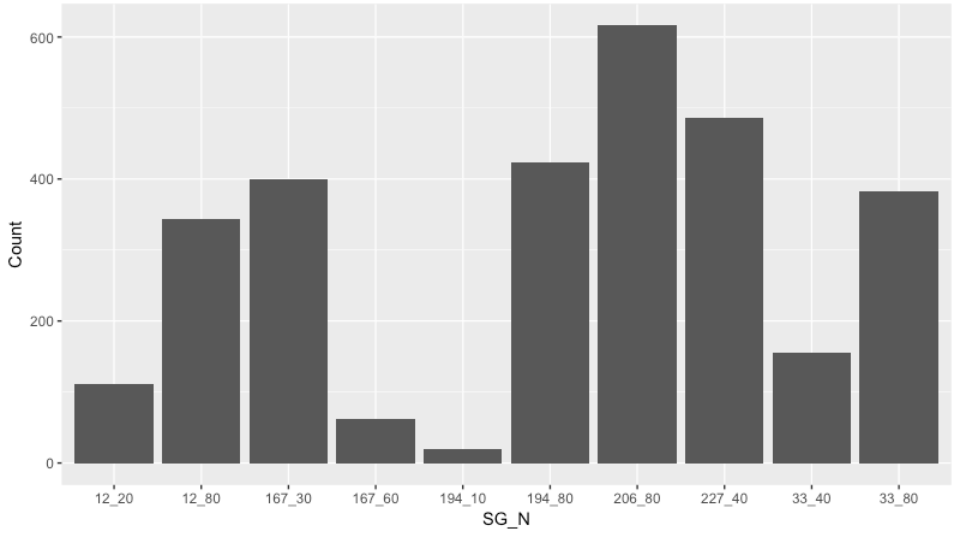
Histograms of MW and Cp are shown in Figure 6. Nothing here appears out of the ordinary.

Figure 6: MW and CP Distributions



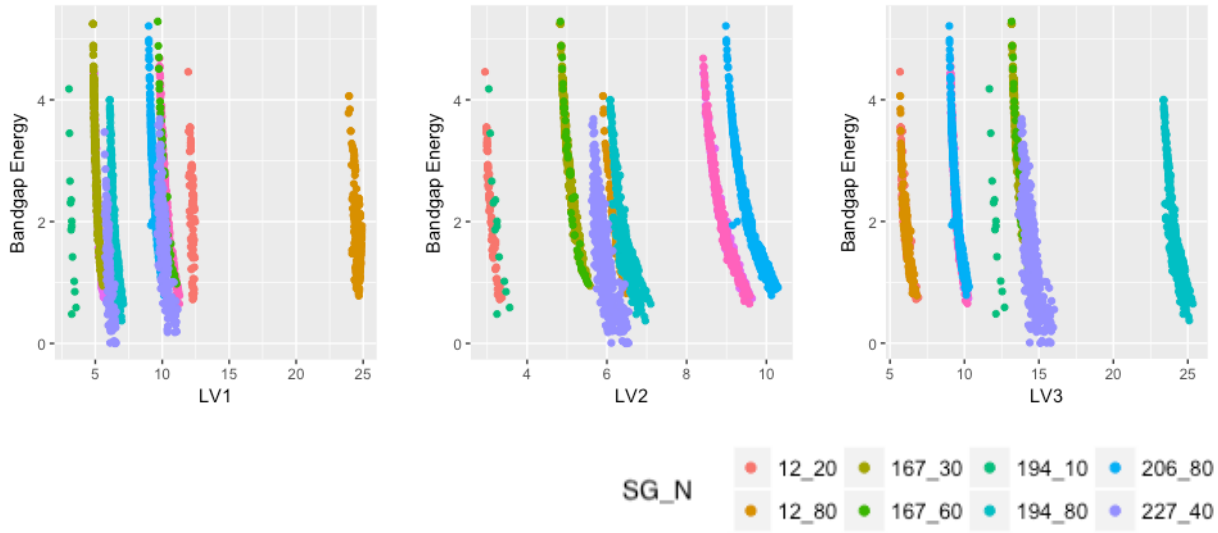
The last variable that was created was a combination of SG and N. Recall, there were only 6 unique values of SG and 6 unique values of N. There are actually only 10 unique combinations of SG and N. SG_N is a variable that represents one of those 10 unique combinations. Figure 7 shows a bar chart of SG_N.

Figure 7: SG_N Bar Chart



SG_N can be used to segregate the data into clusters. Using SG_N, the relationships between lattice vectors and Bandgap Energy become much clearer.

Figure 8: Bandgap Energy vs. Lattice Vectors by SG_N



Building and Selecting Predictive Models

Two models were required for this project. One model was needed to predict Bandgap Energy and one model was needed to predict Formation Energy. Before building any models, 30% of the training data was separated a labeled as a validation set. Additionally, all numerical data was standardized. The validation data and test data were standardized using the means and standard deviations of the training data.

Bandgap Energy Model

Table 3 lists the main models constructed to predict Bandgap Energy.

Table 3: Model Construction Summary

Model	LM1	LM3	Lasso1
Model Type	Linear Regression	1 Linear Regression Model for Each SG_N	1 Lasso Model for Each SG_N
Variable Selection	Manual	Manual	Minimum Lambda
Variables Modeled	All	LV + LV^2, MW, Cp	Varied Based on SG_N
Mean Squared Error	0.1073	0.0577	0.0620
Standard Error	0.00749	0.00534	0.00561

Table 3 Continued: Model Construction Summary

	Hybrid1	XGB1
Model Type	Combination of LM3 and Lasso1	Extreme Gradient Boosting
Variable Selection	Manual	Manual
Variables Modeled	Varied Based on SG_N	LV + LV^2, MW, CP, Alpha, Beta, Gamma, SG_N
Mean Squared Error	0.0565	0.0508
Standard Error	0.00530	0.00465

The first model that is reported here was a basic linear model using all of the original variables (including 2nd degree polynomial terms for each lattice vector) and the created variables (MW, Cp, SG_N). This was just to provide a crude baseline. This model performed poorly.

The second model that is being reported is actually a series of 10 linear models. For this model, a separate linear model was generated for each value of SG_N. The models took the form:

$$Bandgap\ Energy = \beta_0 + \beta_1 \cdot LV + \beta_2 \cdot LV^2 + \beta_3 \cdot X$$

Where LV is one of the three lattice vectors (L1, L2, or L3) and X is either MW or Cp.

Different combinations of LV and X were modeled manually. The combination which gave the lowest validation error was selected for the final model.

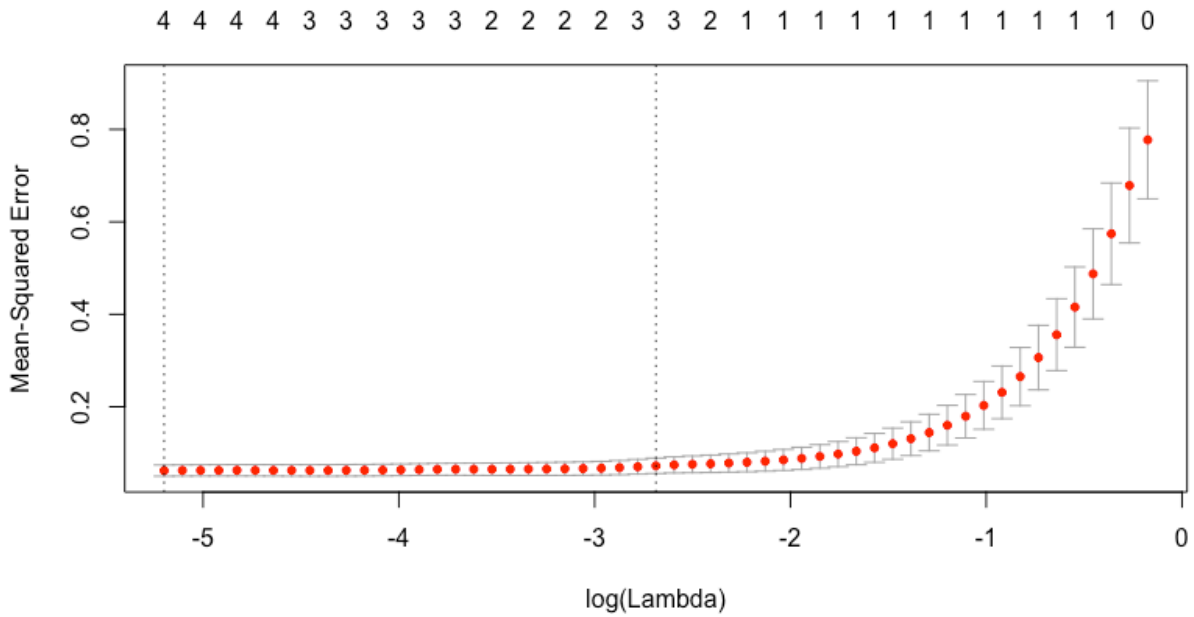
A similar process was followed for the lasso models. Meaning, a different model was created for each SG_N value. Recall that this equation is minimized for lasso regression[4]:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

The cv.glmnet function in R was used to conduct cross validation on a grid of λ values. The goal of the cross validation was to minimize the mean squared error. A plot of this format was generated for every lasso model. The hybrid model gave lower mean squared errors than the simple linear models and the lasso models alone.

4. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; An Introduction to Statistical Learning with Applications in R.

Figure 9: MSE vs Log (λ) Example for Lasso Cross-Validation



The value of λ that gave the lowest mean squared error was used when making predictions on the validation dataset. The λ term in the lasso equation has the effect of selecting variables because as λ increases, the model coefficients decrease. A sufficiently large λ can even result in model coefficients equal to zero.

The fourth model that was tried was a hybrid. Basically, this model was also a combination of 10 models (one for each SG_N value). However, no new models were fit for the hybrid. Instead, either the simple linear model from the second model or the lasso model was chosen. The selection criteria was mean squared error on the validation data. The hybrid was comprised of 8 simple linear models and 2 lasso models.

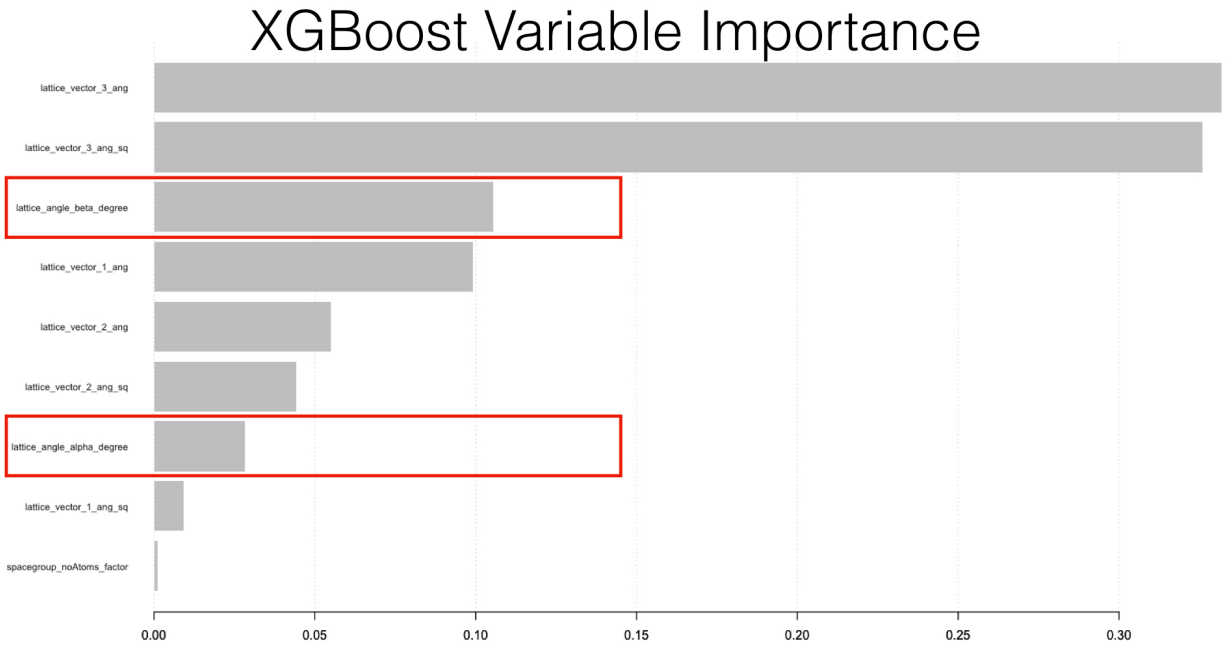
The last technique used to predict Bandgap Energy was extreme gradient boosting. Extreme gradient boosting was done with the XGBoost package in R. While using XGBoost to predict bandgap energy for transparent semiconductors is a relatively new concept, there are examples of XGBoost being used to predict quantitative structure – activity relationships (QSAR) in the pharmaceutical industry[5]. These XGBoost predictions can be used to prioritize experiments during drug discovery research. Conceptually, this is pretty similar to what was done for this project.

5. Sheridan, R.P.; Wang, W.M.; Liaw, A.; Ma, J.; Gifford, E.M.; Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J. Chem. Inf. Model. 2016, 56, 2353−2360.

Extreme Gradient Boosting is a type of modeling that involves building a sequence of models to reduce errors. These are iterative procedures, so the different models are not independent of one another. There are many inputs that go into an XGBoost model. The most important parameters are the input variables that are used, the loss function, and the criteria for model selection. For the XGBoost model that is reported here, the loss function was the regular linear function and the evaluation criteria was the RMSE. The best XGBoost model also did not use all of the input variables. Better results were achieved by removing the composition variables (Al, Ga, In) and replacing SG and N with SG_N (as a factor 1-10).

Because linear combinations of trees are difficult to visualize, the results are best interpreted with a variable importance plot. The XGBoost function was able to use Beta and Alpha to get better predictions. These were not used in the best performing linear models.

Figure 10: Variable Importance for XGBoost Bandgap Energy Predictions



The XGBoost model was selected as the final model for this project because it gave the smallest mean squared error on the validation dataset.

Formation Energy Model

Table 4 lists the main models constructed to predict Bandgap Energy.

Table 4: Model Construction Summary

Model	Lasso2	XGB3
Model Type	Lasso	Extreme Gradient Boosting
Variable Selection	Minimum Lambda	Manual
Variables Modeled	All	All
Mean Squared Error	0.0046	0.0018
Standard Error	0.00033	0.00019

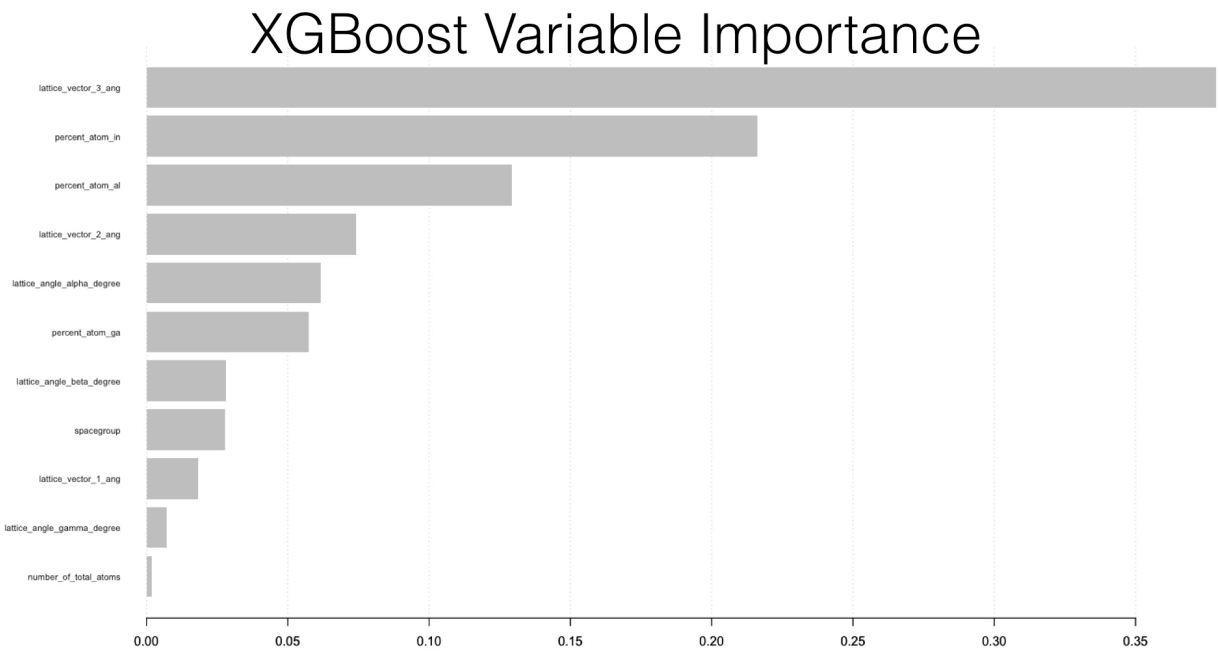
The first model that was tried was a lasso model. All of the input variables and the bandgap energy predictions were included in the Lasso model. However, the lasso removed N, In, LV2, Alpha, Beta, Gamma, Mw, and Cp. The final lasso model was:

$$Formation\ Energy = 0.187 + 0.003 \cdot SG + 0.047 \cdot Al - 0.003 \cdot In - 0.002 \cdot LV1 + 0.020 \cdot LV3 - 0.068 \cdot Bandgap\ Energy$$

While removing a number of parameters was impressive, the XGBoost model outperformed the Lasso model again. For the XGBoost model that is reported here, the loss function was the regular linear function and the evaluation criteria was the RMSE. All of the original variables were included in the XGBoost model. Figure 11 is the variable importance plot for the formation energy predictions. There are some similarities between this and the lasso model. Both Al and LV3 were significant contributors to both models. However, the bandgap energy predictions were not that important to the XGBoost model while they the most significant contributor to the lasso model.

The XGBoost model was selected as the final model for this project because it gave the smallest mean squared error on the validation dataset.

Figure 11: Variable Importance for XGBoost Formation Energy Predictions



Conclusion

As of 2/12 at 12:00 AM, I am in place 601 out of 867 on the Kaggle leaderboard for the 2018 Predicting Transparent Semiconductors challenge. While I wish I was ranked higher, I think I learned a few important lessons. First, I should have spent more time on advanced techniques such as extreme gradient boosting and neural nets. The Lasso models were definitely useful, but they aren't as powerful as other machine learning techniques. Second, while I was interested in this topic, I didn't quite budget my time as well enough for a project that essentially required me to build two separate models. I spent much more time on modeling Bandgap Energy than Formation Energy. Balancing my efforts might have resulted in a better score.