Citrine Challenge

Role: Data Science and Engineering Internship

Suzie Petryk

Overview

In this challenge, the correlation between composition and yield strengths of steel is analyzed. The dataset used is from Citrination, entitled "Mechanical properties of some steels" [1], which includes 842 steel samples. Each entry contains the weight percent composition of the sample's 13 constituent elements, as well as mechanical properties including yield strength, ultimate tensile strength, and fracture toughness. Here, we focus only on the yield strengths.

Much of the discussion on the methodology is present in the accompanying Jupyter notebooks. This report summarizes the methodology, includes a discussion on the approach's limitations and provides physical insight to the results.

Accompanying files

The files accompanying in this report are described as follows:

SteelYieldStrengthDemo.ipynb This Jupyter notebook provides a demo of the analysis described. It walks the reader through the process of extracting the data to a .csv file, creating t-SNE plots and a random forest model, and analyzing the feature importances.

SteelsTesting.ipynb This notebook contains the work behind tuning the hyperparameters for the t-SNE plots and random forest model used in the demo. It also includes an initial attempt at data visualization using PCA.

steel_data.json This file is the complete downloaded dataset in JSON format directly from Citrination.

steel_data.csv This csv file contains the 842 entries from the dataset, with the weight percent composition of the 13 elements for each steel and its corresponding yield strength in MPa.

yield_strength_predictions.csv This csv file contains the yield strength predictions from a test set, subsampled from the total dataset. It includes the weight percent composition of the steel, the actual yield strength given in the data, and the predicted yield strength from the random forest model.

Methods

First, a visualization of the data using t-SNE is created (Figure 1). This exposes a tight cluster of samples that all have high yield strengths, indicating the presence of some correlation.

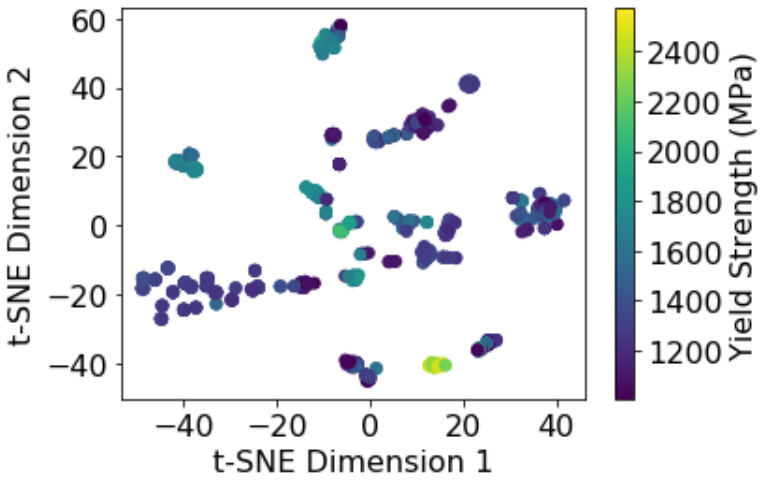

Figure 1. t-SNE embedding of weight percent composition matrix for steels.

Next, we use a random forest model to predict the yield strengths, yielding an R^2 value of approximately 80% on a test set. While this correlation is relatively high, the limitation of using only the composition to predict yield strength is discussed in the next section. The feature importances from the random forest model are then used to identify the elements that contributed the most to the yield strength predictions. From the plot of feature importances (Figure 2), we find that titanium had a significantly higher impact on the predictions than did the other elements.

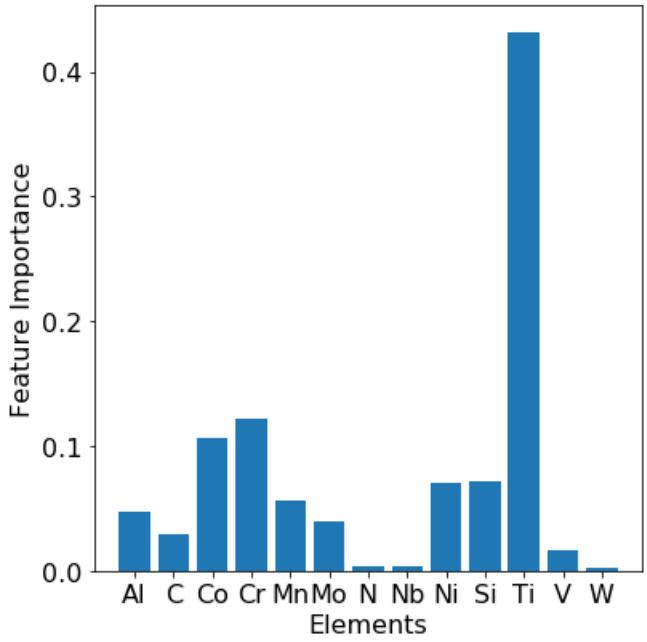

Figure 2. Contributions of various elemental compositions to yield strength predictions.

We then return to our t-SNE plot and compare the clusters colored by yield strength and titanium content (Figure 3). The clustered points of high yield strength correspond closely to the clustered points with the highest titanium content.

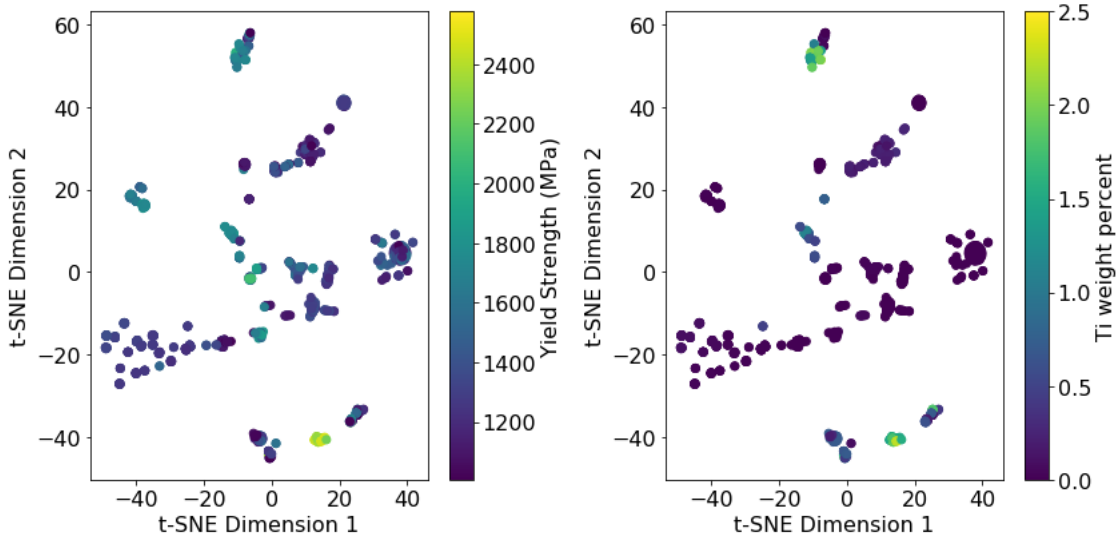

Figure 3. t-SNE plots of the weight percent composition matrix, colored by (left) yield strength, and (right) titanium weight percent.

Discussion

The R^2 value of approximately 80% from the random forest model supports a strong correlation between steel composition and yield strength. However, the remaining variability is not surprising: the yield strength of a steel is also heavily influenced by the processing steps involved in the steel's creation. The yield strength defines the stress at which a steel begins to plastically deform. If a steel sample is subjected to stresses below this point, then it always returns to its original shape when the stress is released. However, stresses beyond the yield strength lead to permanent plastic deformation. This is closely related to the steel's microstructure. For example, steels with large grains are more vulnerable to plastic deformation, and have lower yield strengths. One way to control the grain size is through specific heat treatments of the metal [2]. These heat treatments hold the metal at certain temperatures for a certain length of time to promote the growth of finer grains. The heat treatments were not expressed in this dataset, and so any variability in the yield strength due to these processes was not captured by only the steel's composition.

The feature importances provided by the random forest model led us to conclude that the weight percent of titanium had the most significant impact on the yield strengths of steel. Given a mixture of 13 alloying elements in the steel, it is not immediately obvious that titanium would have the most impact. This demonstrates the model's ability to uncover information through the dataset. However, this does not indicate the type of correlation: is high titanium content favorable or unfavorable for high yield strengths? From our knowledge of steel processing, we know that alloying elements can increase the yield strengths through solid solution strengthening [3]. The impurity atoms introduce strains in the lattice of the metal, which can trap dislocations and prevent them from propagating through the metal. The t-SNE plots in Figure 3 which relate high titanium content to high yield strengths validate the model's ability to uncover physical meaning in the dataset.

It may be clear from current knowledge on steel processing that alloying elements such as titanium leads to higher yield strengths. However, similar correlations may not be as clear for other material properties. The process of using t-SNE to uncover structure, using a random forest model to find important features, and relating the trends between those features and properties of interest back to the structure of the data can be applied to other cases to discover hidden correlations.

References

[1] Citrine Informatics. *Mechanical properties of some steels*. Dataset ID: 153092.

[2] *Yield Strength and Heat Treatment*. Technology, Products and Processes, TPP Information Centre, 2006

[3] Brush Wellman Alloy Products. *Solid Solution Hardening and Strength*. Technical Tidbits, no. 16, Apr. 2010.