Final Report

# Data

Steel is one of the most important materials in existence. From bridges to buildings to cars, steel is considered the material of choice due to its cost to strength ratio. In the field of steelmaking, it could be useful to metallurgists to have an estimate for the strength of a grade of steel prior to it being manufactured. In this project, I create a regression model that estimates the strength of a grade of steel solely based on its constituent elements.

Steel chemistry data was collected from the machine learning data repository

[Kaggle](https://www.kaggle.com/datasets/rohannemade/mechanical-properties-of-low-alloy-steels?resource=download)

Graphical user interface, text, application, table

Description automatically generated It consists of 915 samples of steel each with its respective steel chemistry and strength parameters. An example of the dataset is shown below:

# 2.0 Data Cleaning

Various features needed to be dropped, Alloy code wasn't useful in this context, neither was Carbon equivalent (Ceq) since the chemistry was known. Columns were then renamed. 0.2% Proof Stress is another name for Yield strength and was renamed accordingly.

There were no null values, however there was one unusually high strength property observation which was dropped. Additionally the temperatures the samples were pulled ranged from 27degC to 650degC. A cut-off of 450degC was chosen since most steel applications don't reach temperatures that high. 450degC is still unusual, however I couldn't risk removing too much data.

# 3.0 EDA

Table

Description automatically generatedThe first step of EDA was to look for general patterns in the data therefore a heatmap was created as shown below:

Temperature is negatively correlated with both Yield and Tensile strength which is expected. The higher the temperature, the weaker a metal gets. Correlations between the other strength variables are all expected as well; however, we want to find relationships between the elements and strength!

The elements that stick out are Vanadium (v), Molybdenum (mo), Nickel (ni) and Manganese (mn). Surprisingly Carbon doesn't have a huge role to play in determining strength. There are no elements that contribute significantly negatively to steel strength.

The following scatterplots show the relationship between the Yield strength and the weight percent of each element in that sample of steel.

The target variable 0.2% Proof Strength, otherwise known as Yield strength was chosen to be the target variable in this project since it is the most important strength parameter. It determines when a material will permanently deform under stress which usually one would try to avoid.

Chart, scatter chart

Description automatically generated

# 4.0 Preprocessing

The remaining data was split into training and test sets, and the X datasets were transformed using a Standard Scaler.

# 5.0 Modelling

PyCaret is a low-code machine learning library that automates the model selection process. It scores various models using k-fold cross-validation and returns a hierarchy of the best models. Using this library, the top 3 models were chosen. The models were put into an ensemble Voting Regressor which returns the average of the weighted predictions of each model. The top 10 models are shown below.

Table

Description automatically generated

The CatBoost Regressor, Light Gradient Boosting Machine and Extra Trees Regressor were chosen to be input into the Voting Regressor

## Explaining Models

The CatBoost Regressor is a relatively new machine learning model. This model is an evolution of decision trees and gradient boosting and is best at working with categorical data. In this instance it works well with numeric values as well!

The LightGBM is like XGBoost however, the main difference is in how the trees grows. In LightGBM trees are grown vertically or leaf-wise, whereas in XGBoost, leaves are grown level-wise. This distinction results in LightGBM being faster, but it does tend to overfit.

A picture containing diagram

Description automatically generated

Diagram

Description automatically generated

Extra Trees models are also an ensemble decision tree model like Random Forest models. The main difference between them is that they do not bootstrap the data to train on each tree. Instead, they train each tree on the entire dataset while splitting randomly, not to reduce loss like Random Forest Regressors. In addition, splitting is random in the Extra Trees, whereas the split is based on the applied criterion in Random Forests.

## 5.2 Feature Importance

Here are the feature importance graphs for each of the regressors.

**Chart, scatter chart

Description automatically generatedCatBoost Regressor:**

**Extra Trees:**

Chart, scatter chart

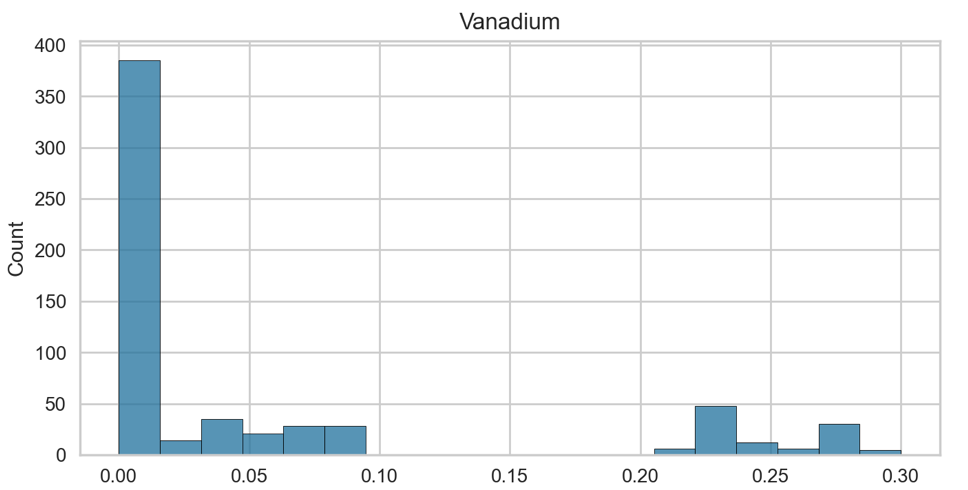
Description automatically generated

Chart, scatter chart

Description automatically generated**Light Gradient Boosting Machine:**

There are a few elements that are commonly major contributors in these 3 models, Vanadium (v), Manganese (mn) and Nickel (ni). Temperature is also a major contributor. Another interesting feature is that the LGBM model seems to take more input from the other elements in its prediction, whereas the other two models rely heavily on the first 3 or 4 features.

As it can be seen, Vanadium is the element which contributes most to strength. Interestingly enough, most of the samples didn't contain this element as it can be seen in the histogram below:



Vanadium forms secondary carbide phases when added to steel. It reduces the grain sizes of the steel thereby reducing the spread of dislocations. Basically, this reduction in grain size prevents the physical movement of the atoms making the steel more resistant to stress. This reduction is evident in the increase in Yield Strength. This process is commonly referred to as grain boundary strengthening [1].

Nickel and Manganese are also elements that strengthens steel via grain-boundary strengthening [2] [3]. However, Manganese also contributes to the formation of another phase in the steel known as austenite which interestingly also makes it more ductile.

Temperature plays a crucial role in reducing Yield strength. An increase in temperature makes the movement of dislocations in most metals since atoms are jostling much more. This makes them less resilient to stress.

## 5.3 Hyperparameter Tuning

Here are the metrics of the untuned and tuned models trained on the training set, and tested on the training set, validation set, test set and cross-validated on the entire dataset.

Table

Description automatically generated

The untuned Cat Boost Regressor was chosen to be included in the final Voting Regressor model since it performed better than the untuned regressor. Both the tuned Light Gradient Boosting Machine and Extra Trees Regressor performed better than their untuned counterparts. They all tended to overfit on the training sets, but still performed admirably on the other sets.

# 6.0 Final Model

As mentioned above, a Voting Regressor was chosen to combine all models. In this meta-model, a weighted average of each model's predictions is used to form a final prediction. The algorithm is shown below:

Text

Description automatically generated

The most accurate weights for the Cat Boost Regressor, Light Gradient Boosting Machine and Extra Trees Regressor had optimum weights of 0.7, 0.1 and 0.2 respectively. The final metrics table is shows below:

Table

Description automatically generated

# 7.0 Conclusion

This model does do quite a good job in predicting steel strength. It appears the Voting Regressor is less prone to overfitting and thus resulted in higher scores on the test, validation and entire datasets. Surprisingly, data on the samples' microstructure resulting from its heat treatment was not needed in this analysis. A limitation to this model is that the data is probably representative of a certain set of steel samples and may not be generalizable to other steel with different chemistries and heat treatments. Additionally, the inclusion of temperature in this analysis might not be useful in most cases.

# 8.0 Sources

[1] Applications of vanadium in the steel industry. (2021). Vanadium, 267–332. https://doi.org/10.1016/b978-0-12-818898-9.00011-5

[2] Applications of vanadium in the steel industry. (2021). Vanadium, 267–332. https://doi.org/10.1016/b978-0-12-818898-9.00011-5

[3] Kaar, S., Krizan, D., Schneider, R., Béal, C., &amp; Sommitsch, C. (2019). Effect of manganese on the structure-properties relationship of cold rolled AHSS treated by a quenching and partitioning process. Metals, 9(10), 1122. https://doi.org/10.3390/met9101122