Final Report

1.0 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

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

	Alloy code	C	Si	Mn	P	S	Ni	Cr	Mo	Cu	V	Al	N	Ceq	Nb + Ta	Temperature (°C)	0.2% Proof Stress (MPa)	Tensile Strength (MPa)	Elongation (%)	Reduction in Area (%)
0	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	27	342	490	30	71
1	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	100	338	454	27	72
2	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	200	337	465	23	69
3	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	300	346	495	21	70
4	MBB	0.12	0.36	0.52	0.009	0.003	0.089	0.97	0.61	0.04	0.0	0.003	0.0066	0.0	0.0	400	316	489	26	79

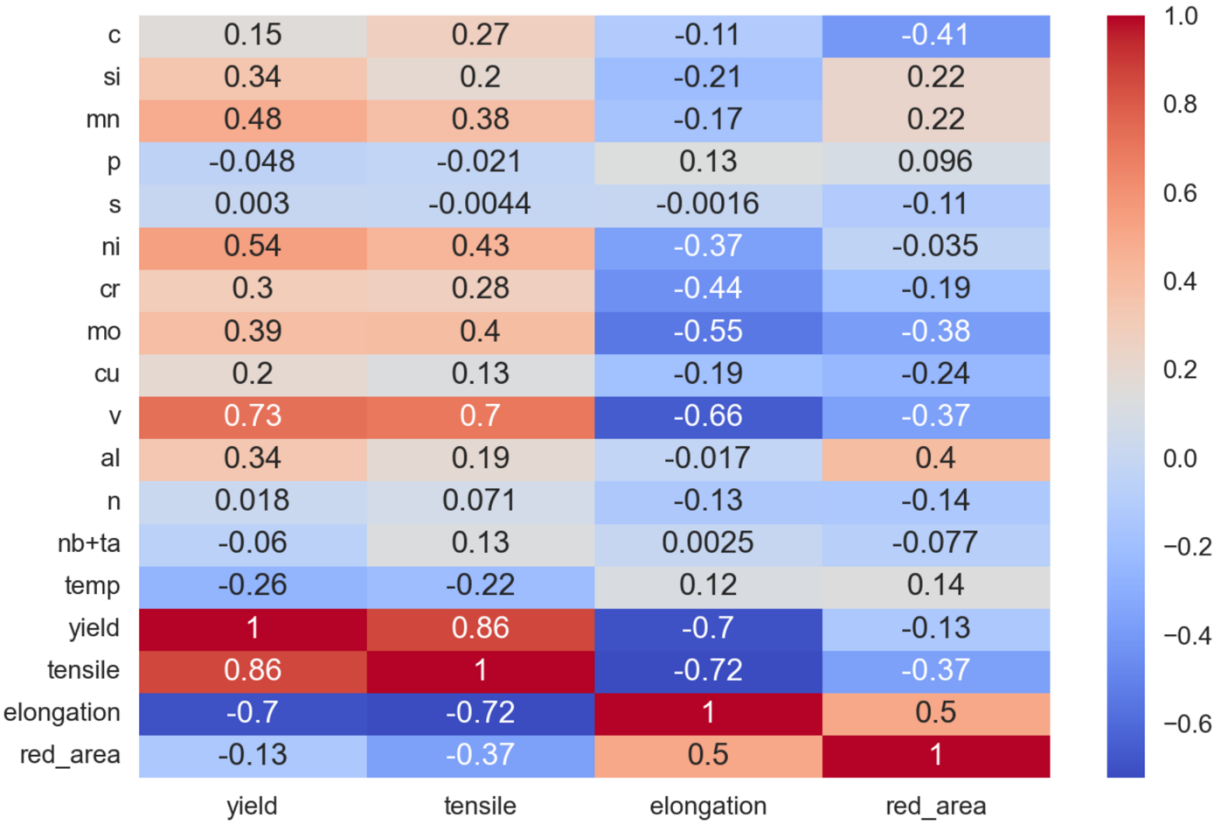
2.0 Data Cleaning

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

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



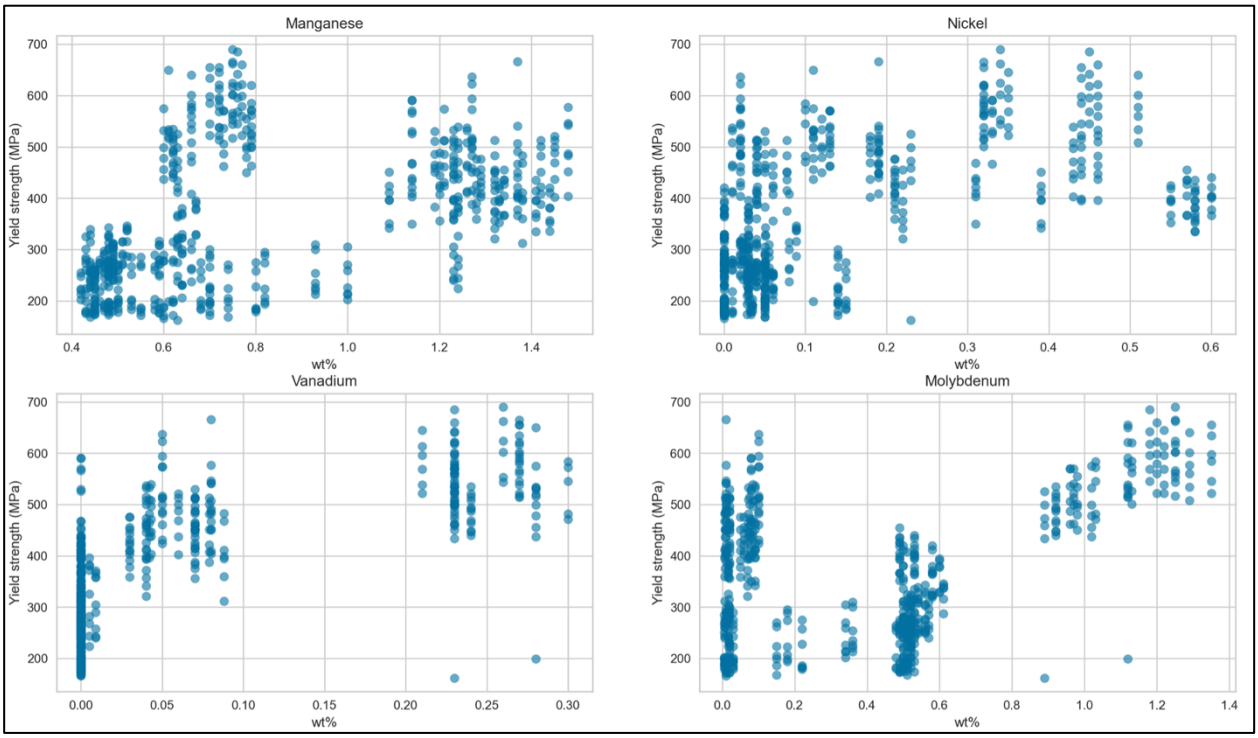
Temperature is negatively correlated with both Yield and Tensile strength which is expected. The higher the temperature, the weaker a metal gets. Correlations between the other strength variables are all expected as well; however, we want to find relationships between the elements and strength!

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.



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.

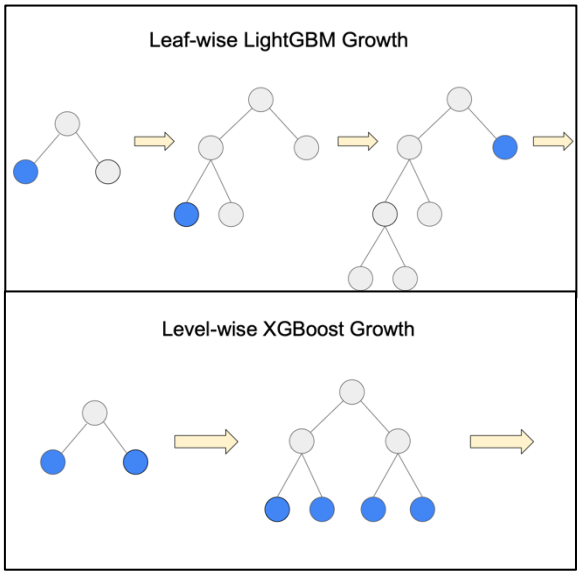
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	14.9141	735.7980	24.5215	0.9557	0.0699	0.0454	0.6460
lightgbm	Light Gradient Boosting Machine	16.5407	824.1776	26.0469	0.9508	0.0731	0.0493	0.2310
et	Extra Trees Regressor	17.2057	812.8257	26.4365	0.9508	0.0759	0.0511	0.1250
gbr	Gradient Boosting Regressor	17.9938	875.5949	27.4723	0.9475	0.0789	0.0545	0.0810
xgboost	Extreme Gradient Boosting	16.9166	894.2253	27.1273	0.9468	0.0756	0.0500	0.1150
rf	Random Forest Regressor	17.9213	882.6691	27.7950	0.9468	0.0789	0.0529	0.1540
dt	Decision Tree Regressor	22.7973	1534.3729	35.7140	0.9083	0.1026	0.0646	0.0150
ada	AdaBoost Regressor	35.2418	2148.1269	45.6406	0.8711	0.1412	0.1130	0.0820
knn	K Neighbors Regressor	35.2107	2448.9948	48.7628	0.8511	0.1421	0.1054	0.0180
lar	Least Angle Regression	35.6690	2459.3196	48.0963	0.8506	0.1394	0.1070	0.0150

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

5.1 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.

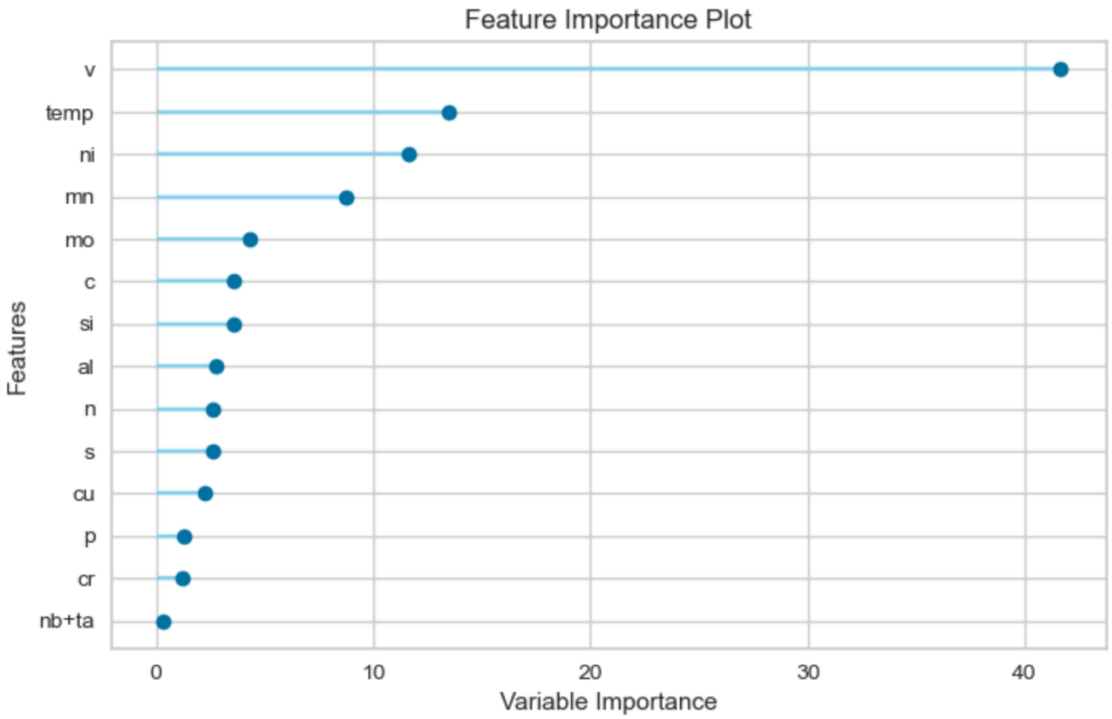


Extra Trees models are also an ensemble decision tree model like Random Forest models. The main difference between them is that they do not bootstrap the data to train on each tree. Instead, they train each tree on the entire dataset while splitting randomly, not to reduce loss like Random Forest Regressors.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

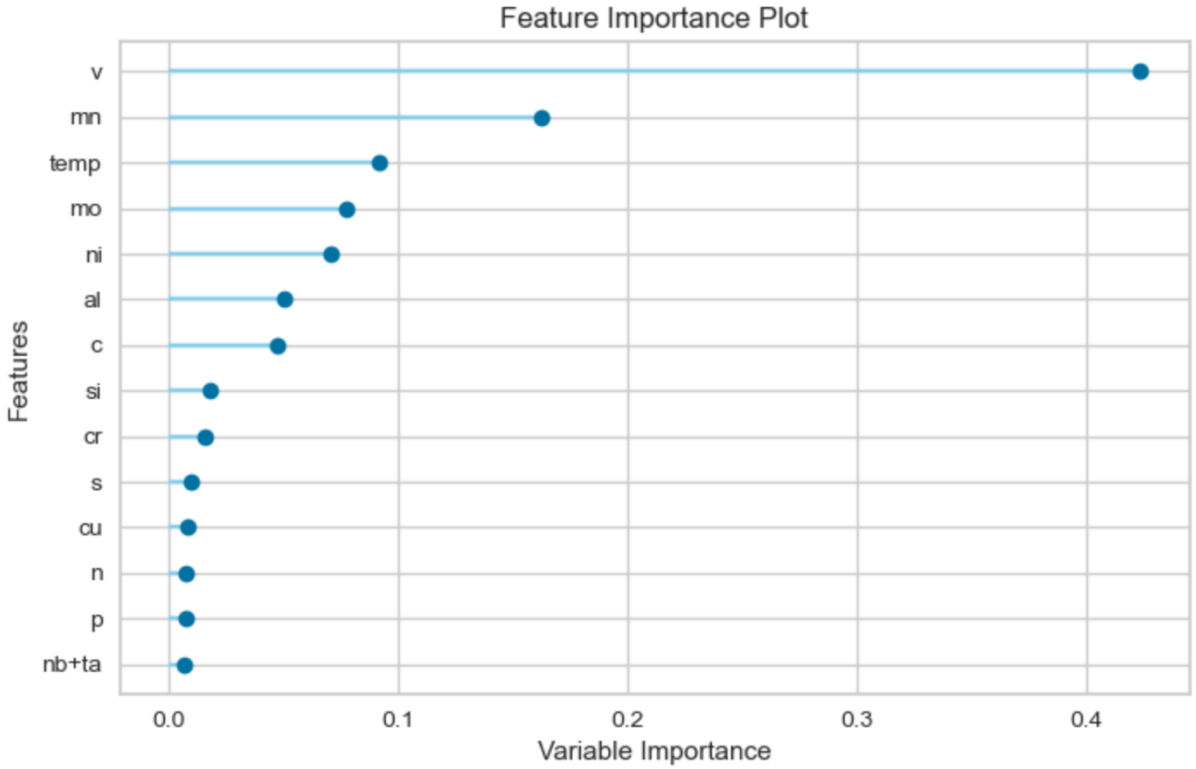
5.2 Feature Importance

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

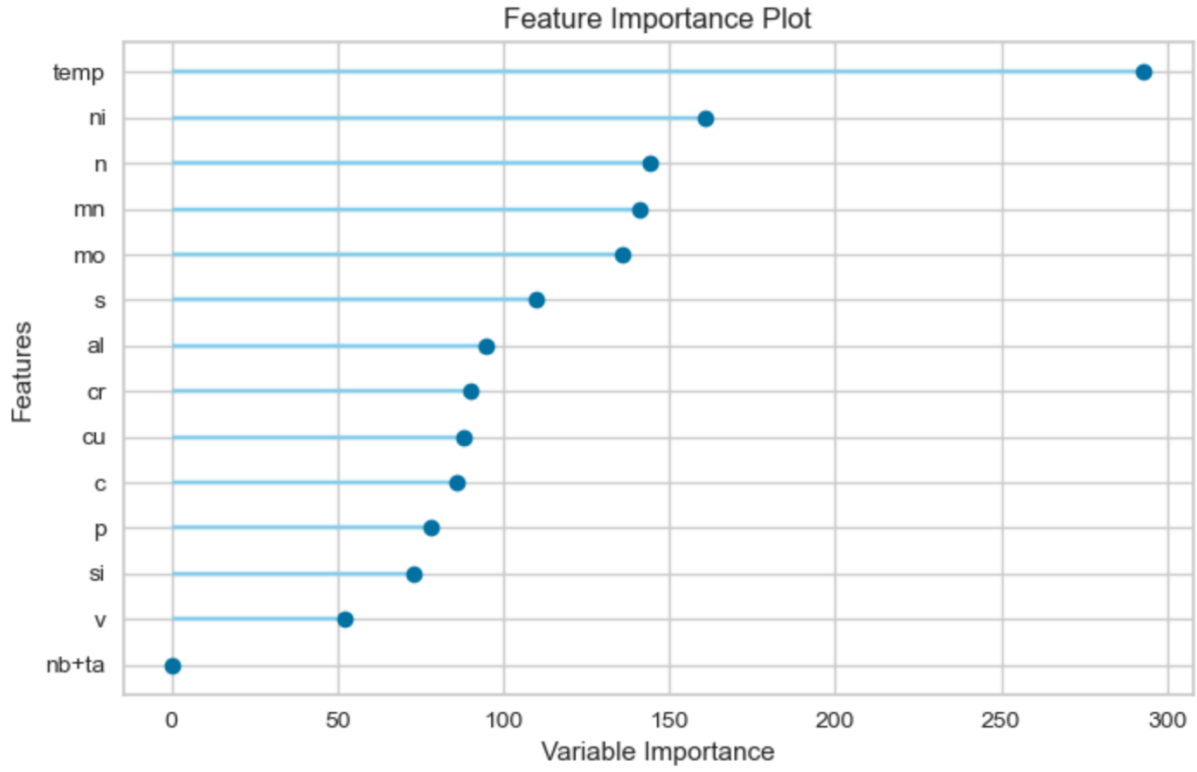
CatBoost Regressor:



Extra Trees:

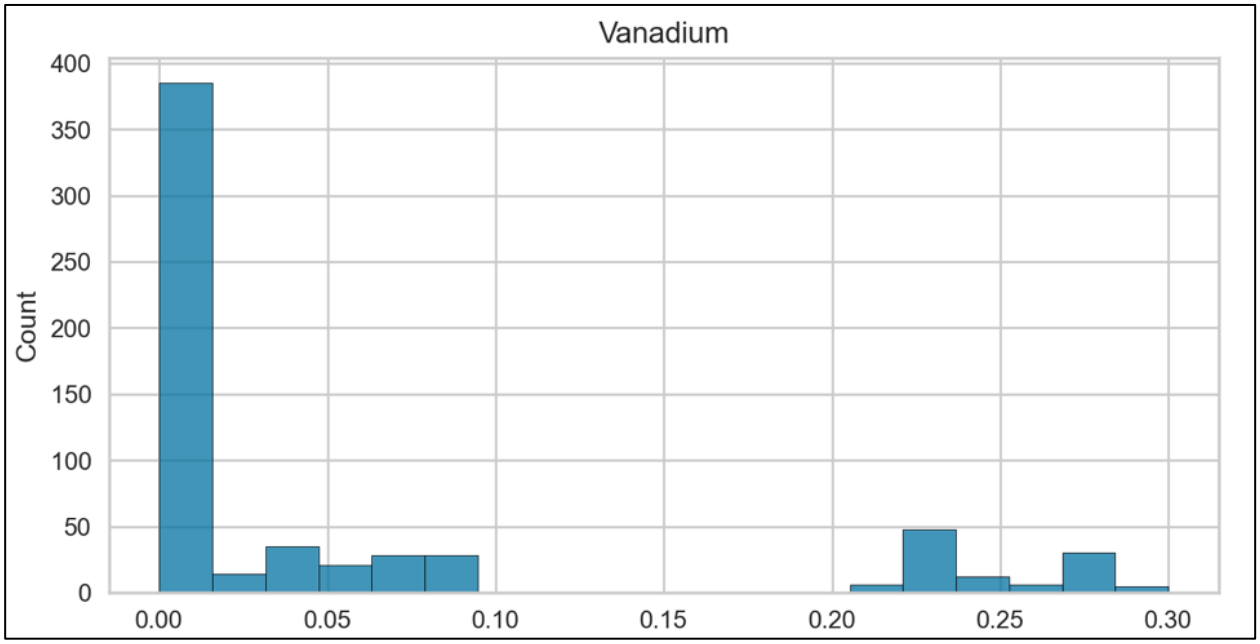


Light Gradient Boosting Machine:



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.

		Untuned CatBoost	CatBoost	Untuned LGBM	LGBM	Untuned ExtraTrees	ExtraTrees
Train	**R2**	0.998044	0.999429	0.984709	0.988166	1.000000	0.997733
	MAE	4.381501	2.509424	8.633846	8.457923	0.000000	3.277293
	MSE	34.055890	9.948996	266.223392	206.045577	0.000000	39.463553
Valid	**R2**	0.983944	0.982376	0.982793	0.984886	0.982518	0.982405
	MAE	12.452790	13.496551	13.783181	13.106688	15.066129	15.224032
	MSE	311.425003	341.836147	333.740443	293.156436	339.085687	341.274511
Test	**R2**	0.927087	0.923388	0.919589	0.920844	0.916665	0.922419
	MAE	16.969241	17.255153	16.959305	18.451850	19.181290	18.308929
	MSE	1303.243063	1369.360729	1437.257607	1414.834575	1489.526024	1386.680810
CV entire	**R2**	0.956715	0.955564	0.956984	0.953459	0.947689	0.950650
	MAE	14.185969	14.636312	15.737631	15.970620	17.693424	16.727021
	MSE	762.429021	784.740332	843.110852	823.500995	1105.556508	875.942224

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:

```
# Weights will be assigned iteratively to each model in a Voting Regressor to discover the most accurate model
weights1 = []
weights2 = []
weights3 = []
scores = []

for i in np.arange(0.1,1,0.1):
    for j in np.arange(0.1,1,0.1):
        for k in np.arange(0.1,1,0.1):
            vote_reg = VotingRegressor([('cat', cat), ('lgbm', best_lgbm), ('xt', best_xt)], weights = [i,j,k])
            vote_reg.fit(X_train, y_train)
            y_pred = vote_reg.predict(X_test)
            score = r2_score(y_pred, y_test)
            scores.append(score)
            weights1.append(i)
            weights2.append(j)
            weights3.append(k)
```

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:

		Untuned CatBoost	LGBM	ExtraTrees	VotingRegressor
Train	**R2**	0.998044	0.988166	0.997733	0.997801
	MAE	4.381501	8.457923	3.277293	4.254489
	MSE	34.055890	206.045577	39.463553	38.284181
Valid	**R2**	0.983944	0.984886	0.982405	0.985509
	MAE	12.452790	13.106688	15.224032	12.410774
	MSE	311.425003	293.156436	341.274511	281.061241
Test	**R2**	0.927087	0.920844	0.922419	0.927101
	MAE	16.969241	18.451850	18.308929	16.846251
	MSE	1303.243063	1414.834575	1386.680810	1302.991689
CV entire	**R2**	0.956715	0.953459	0.950650	0.957005
	MAE	14.185969	15.970620	16.727021	14.218077
	MSE	762.429021	823.500995	875.942224	758.996468

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., & 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