Energy Consumption

Nicholas Morris

Machine Learning

Wrangling → Feature Engineering → Modeling

Timestamp

Converting Year Made to Date Made.

Additional Data

Adding economic trends such as unemployment and CPI. Converting categorical features to binary data points. Attempting feature transformations.

Predictions

Training linear regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling

Dataset

Below is the first two housing units in the data. There are 11,000 total units and 939 columns. The target we are predicting is NWEIGHT. [Link to the dataset and code]

REGIONC	DIVISION	REPORTABLE_DOMAIN	TYPEHUQ	NWEIGHT	HDD65	CDD65	HDD30YR	CDD30YR	Climate_Region_Pub	...	SCALEKER	IECC_Climate_Pub	HDD50	CDD80	GND_HDD65	WSF	OA_LAT	GWT	DesignDBT99	DesignDBT1
2	4	12	2	2471.68	4742	1080	4953	1271	4	...	-2	4A	2117	56	4250	0.48	6	56	9	96
4	10	26	2	8599.17	2662	199	2688	143	5	...	-2	3C	62	26	2393	0.61	0	64	38	73

Date Made

The YEARMADE column was converted to a date so we can extract economic indicators.

Feature Engineering

Binary Data

Various categorical features, some words and others are integers, were converted to binary data points.

Economic Data

The NASDAQ closing price, Unemployment rate, CPI, PPI, GDP, GDI, and Federal Funds Rate were pulled from FRED (Federal Reserve Economic Data).

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

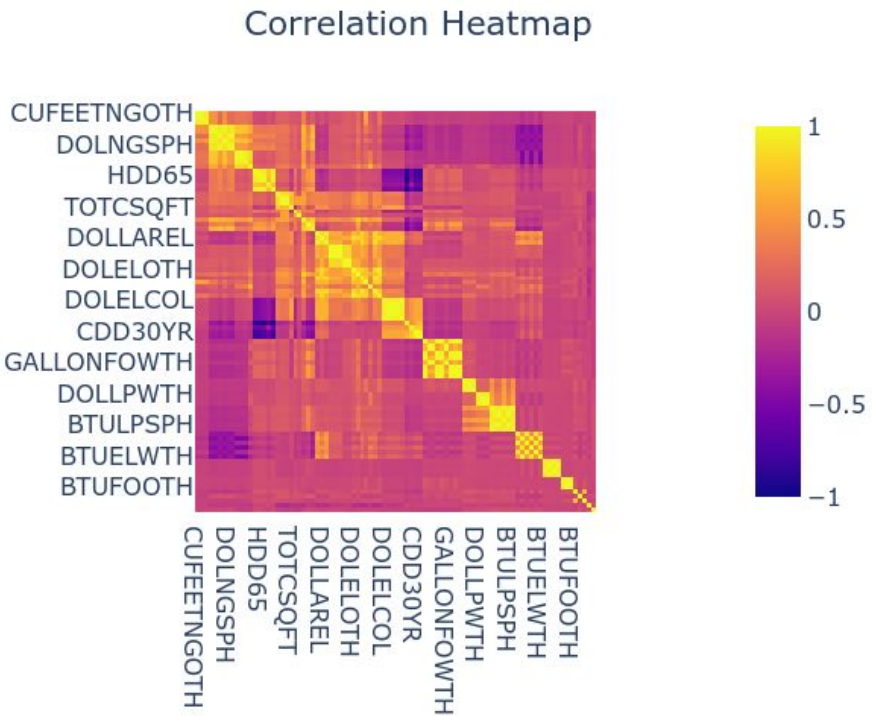
Reciprocals did not improve model performance, so, it was left out of the final model.

Interactions

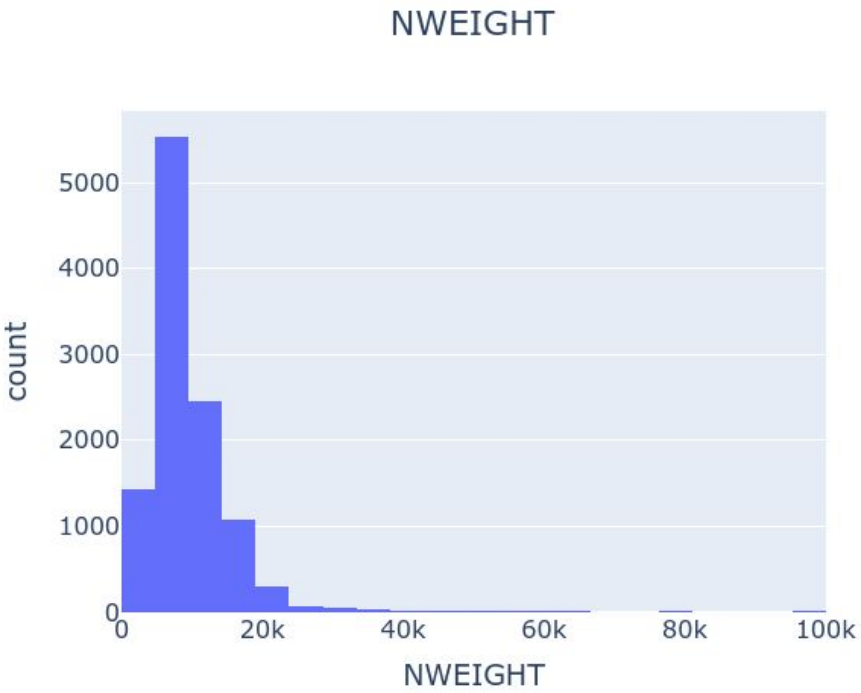
An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration



There are various zones of correlations in the data. The yellow regions show positive correlation meaning that as one indicator increases in value another indicator follows. There's a few dark purple regions of negative correlation showing that as one indicator increases another indicator decreases in value.



The distribution of NWEIGHT is heavily skewed to the right which increases the difficulty of modeling. Non-linear relationships will need to be found to account for the skew. Most of the values for NWEIGHT are between 4766 and 14.2 thousand (the two highest bars).

Significant Patterns

There were no correlations with NWEIGHT above 0.7 or below -0.7. There were no differences of at least 10% in average NWEIGHT between any two labels of a categorical feature. This will increase the difficulty of modeling NWEIGHT.

Modeling

Model Parameters

Linear Regression

Library: scikit-learn
Length Of Path: 1e-9
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 500

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Linear Regression

R2: 0.64
RMSE: 3061
In Control: 97%

Model Indicators:
1. REPORTABLE_DOMAIN_20
2. REPORTABLE_DOMAIN_6
3. REPORTABLE_DOMAIN_14
4. DIVISION_3
5. DIVISION_4

XGBoost

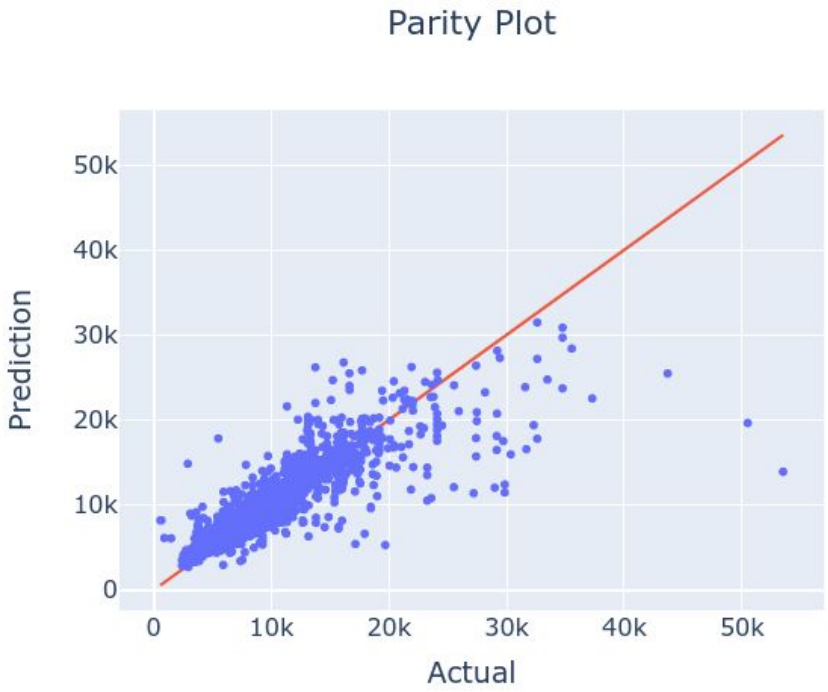
R2: 0.75
RMSE: 2589
In Control: 96.5%

Model Indicators:
1. REGIONC_2
2. DIVISION_3
3. Climate_Region_Pub_4
4. REPORTABLE_DOMAIN_9
5. DIVISION_4

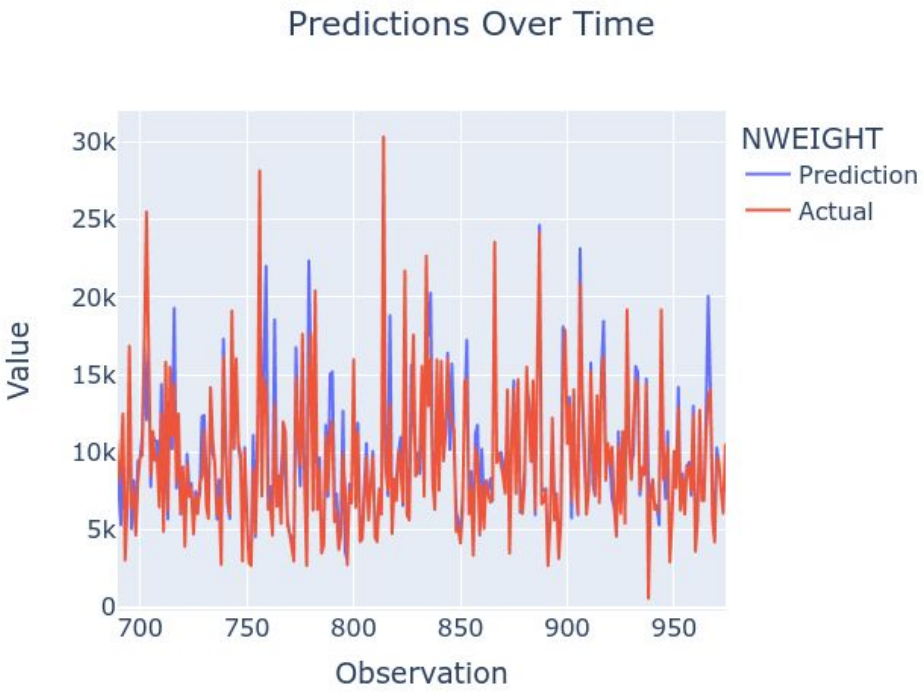
Neural Network

R2: 0.65
RMSE: 2962
In Control: 97.36%

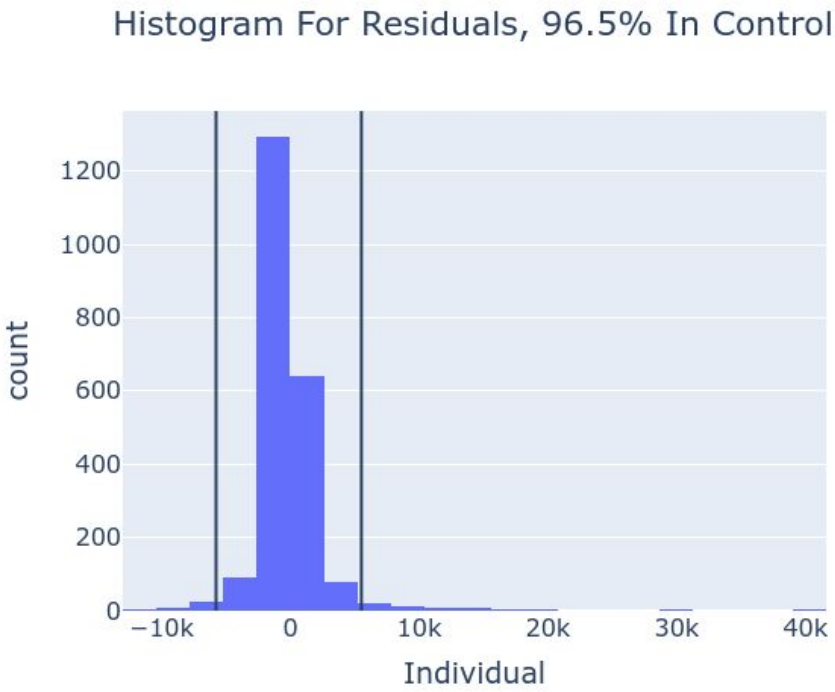
Model Indicators:
DNF



These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. The predictions are centered on the red line (perfect predictions). The model under-predicts the larger values of NWEIGHT, which we saw would be a problem area.

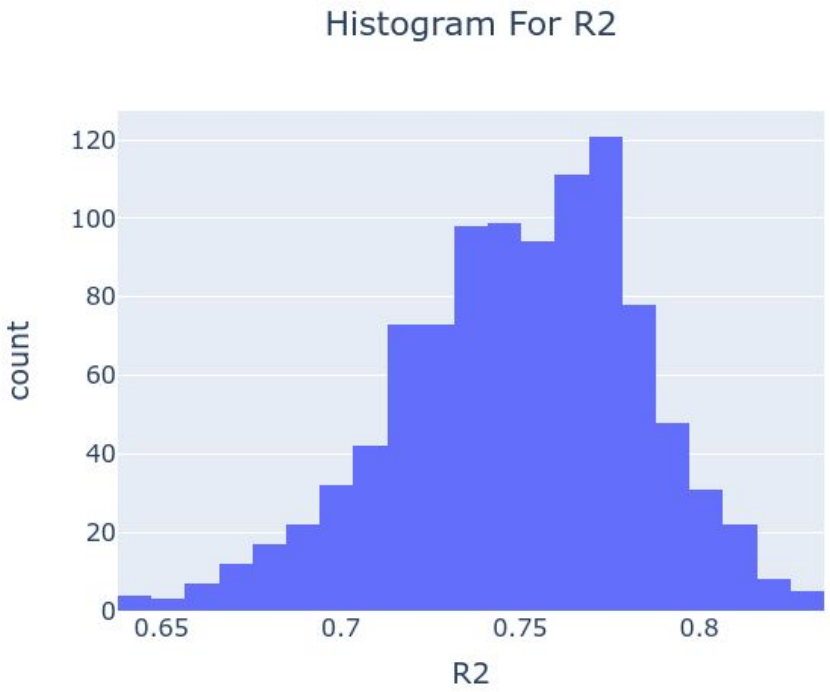


Here’s a snapshot of the predictions over time. We can see the the blue predictions follow the actual values well. There is some extreme values where the red line drops down or jumps up well past the predictions. We can see there is no seasonality in the data.



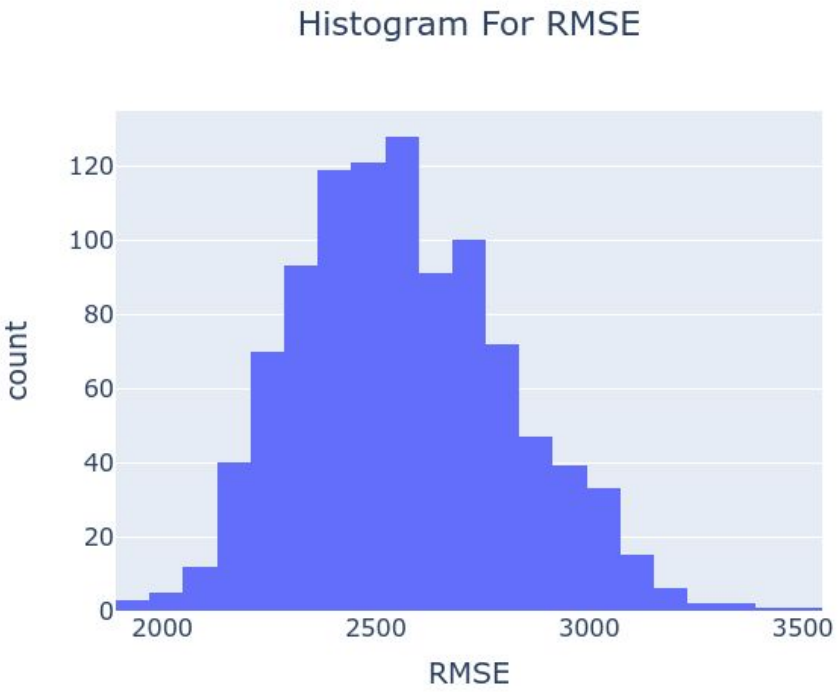
The residuals are prediction error = actual - predicted.

The residuals have a tight bell shape, which is good, and they are centered on zero. Control limits were computed on the residuals and we can see that the prediction error is mostly under control. We can see a skew to the right, which shows the model has a tendency to under-predict NWEIGHT.



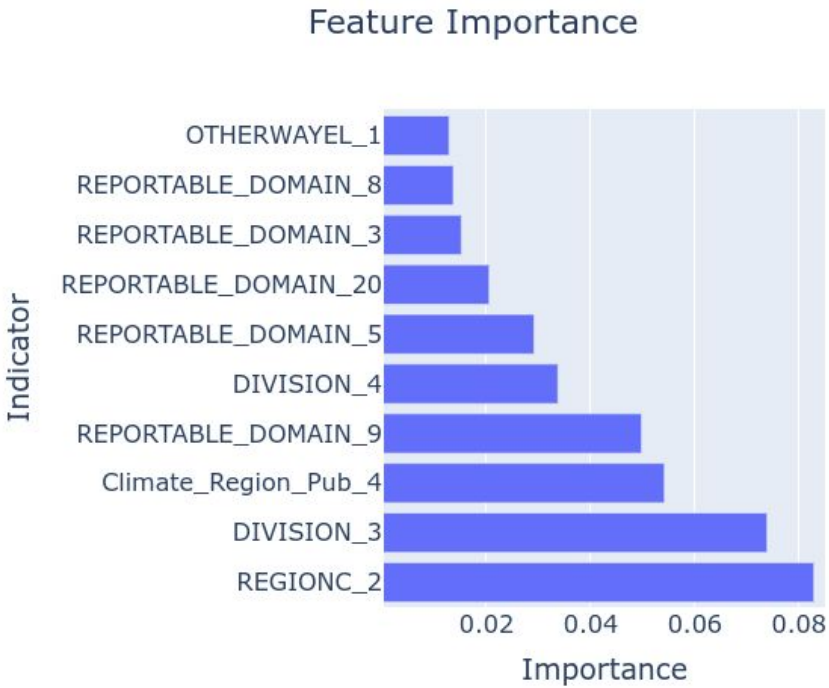
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then R2 was computed on each sample to get a distribution.

R2 has a wide range between 0.65 and 0.8, which isn't good. R2 has a bell shape, which is good, and a skew to the left.

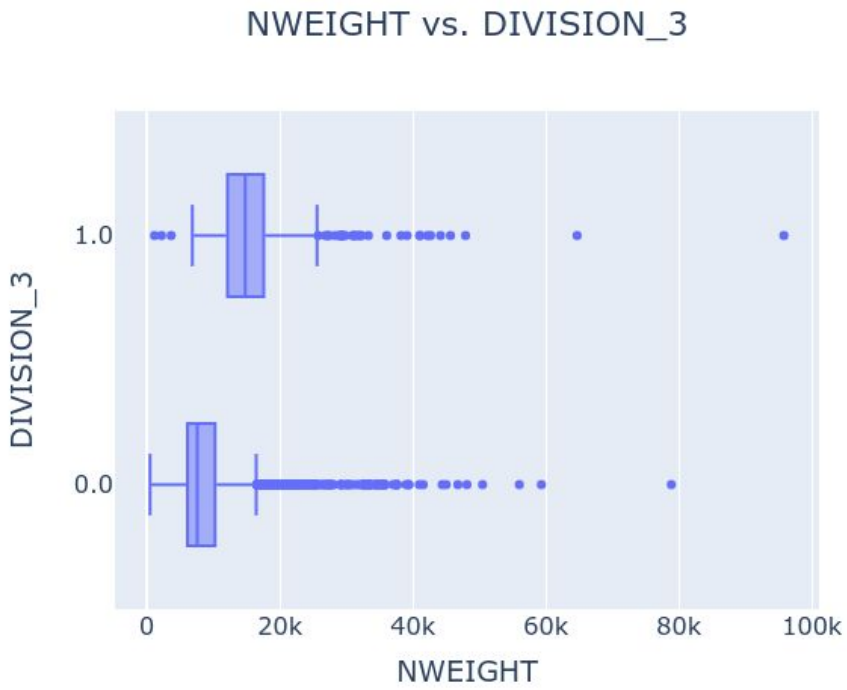


The prediction error was resampled as previously mentioned to get a distribution for RMSE.

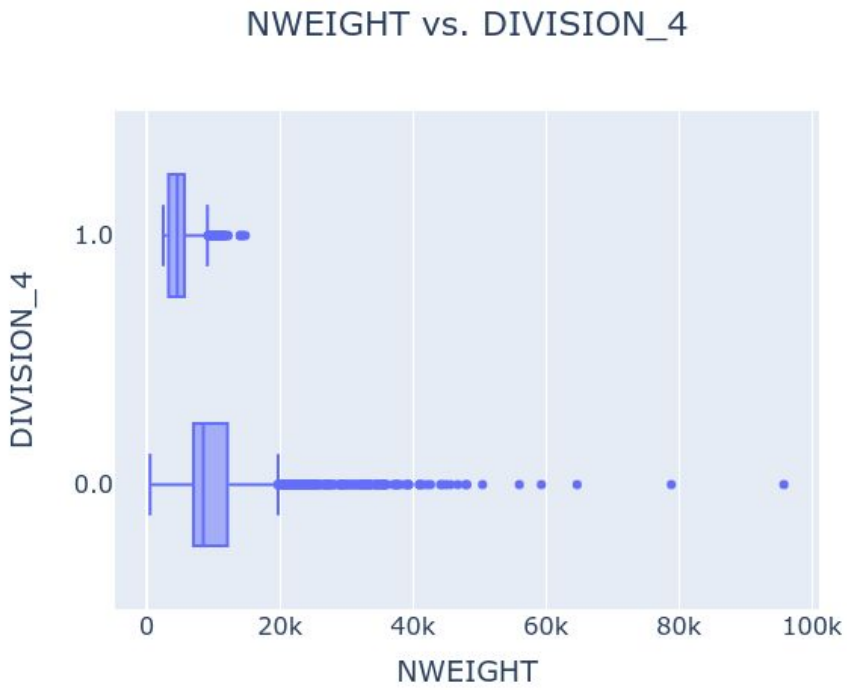
On average, the predictions are off by 2000 to 3000 units, which isn't a tight range. RMSE has a bell shape, which is good. There's a skew to the right.



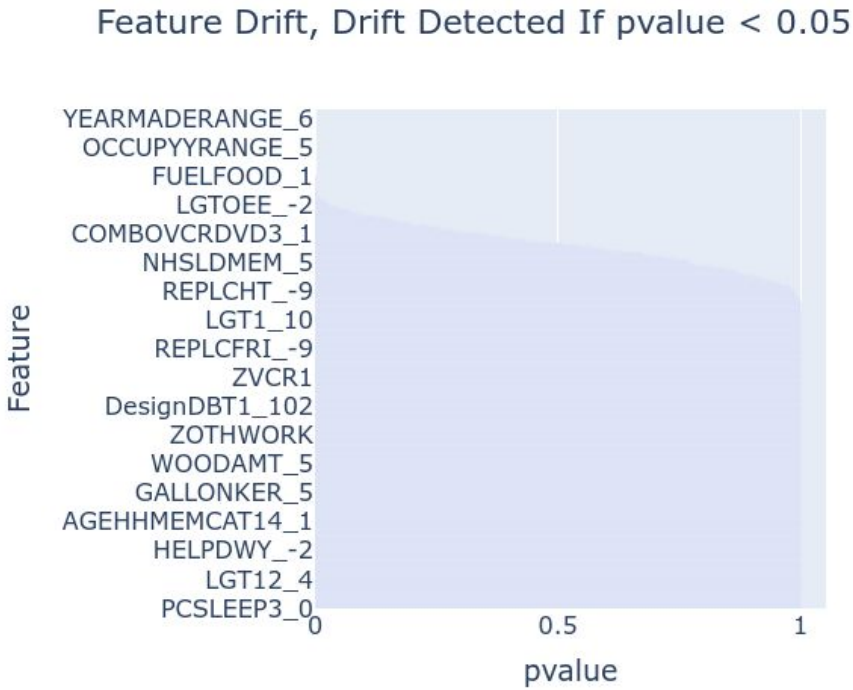
The top ten most important indicators of NWEIGHT are shown to the left. The top four features contribute most to the model, and then importance starts to taper off.



DIVISION_3 shows a higher NWEIGHT than other divisions with a median of 14.8 thousand, and the other divisions have a median of 7577.

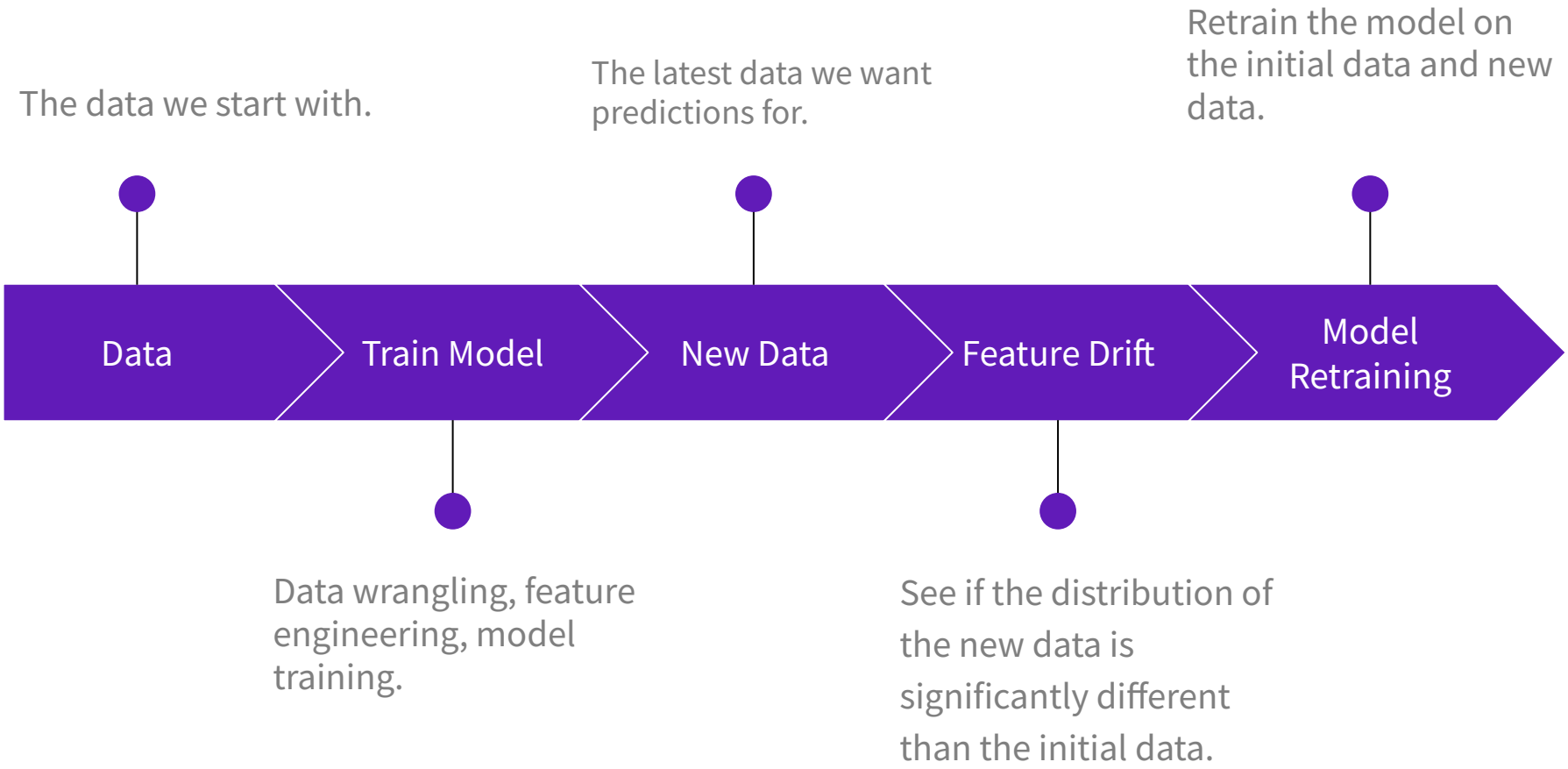


DIVISION_4 shows a lower NWEIGHT than other divisions with a median of 4569, and the other divisions have a median of 8465.



A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. Most of the columns do not experience a drift, which is good. The model was restrained to include the test data.

Deployment


The data we start with.
The latest data we want predictions for.
Retrain the model on the initial data and new data.
Data
Train Model
New Data
Feature Drift
Model Retraining
Data wrangling, feature engineering, model training.
See if the distribution of the new data is significantly different than the initial data.

Thank You