Predicting Volatility in US Electricity Prices

Miles Franklin, David Harper,
Adam Goldstein, Himanshu Ghritalhre

Recap

- **Target**
    - The goal of this project is to produce a model predicting the price volatility of electricity for a given **year**. We measure the volatility using the Coefficient of Variation, to normalize the data and facilitate state to state comparisons.

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

- **Features**
    - Futures Market Volume
        - How well do market participants understand market direction?
    - Drought and Temperature Trends
        - Are the effects of climate patterns felt more by renewable sources?
    - Distribution of Fuel Sources in a State
        - What percent of production is from coal, solar, etc.?
    - Consumption measured in electricity sales
        - Do states that have more businesses and industry have a less stable price?

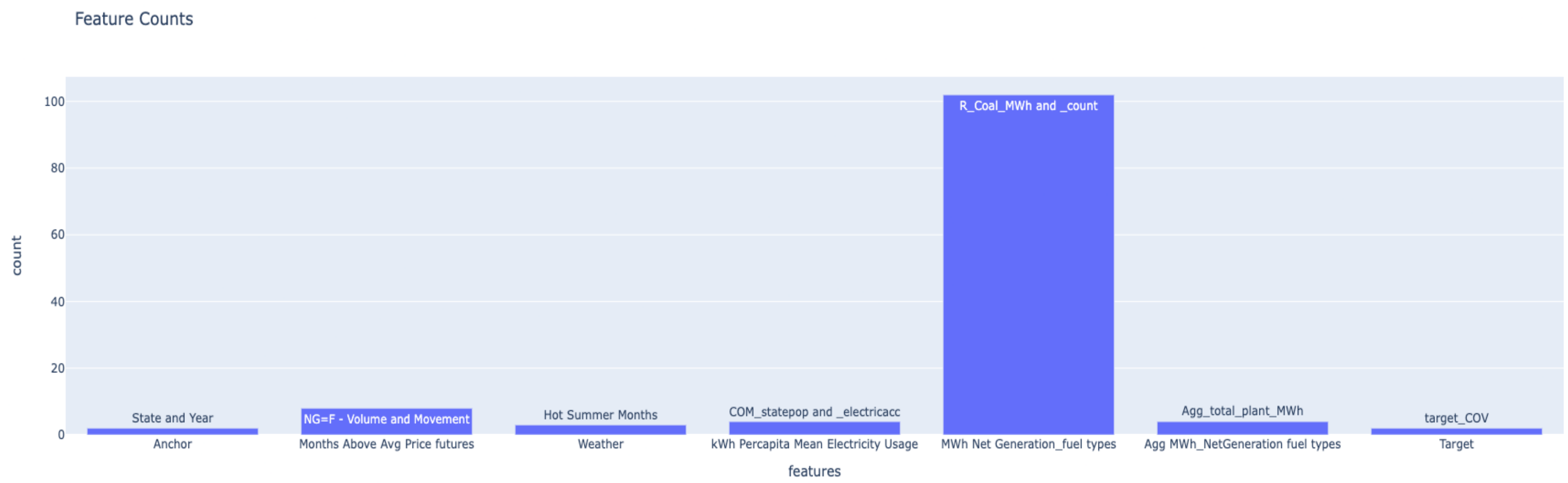
Phase 2 Project Overview

- Dataset
- Expanded dataset from 2008-2020 to 2001 - 2020
    - Backfilled electric accounts for RES, COM, and IND population
        - RES - US Census
        - COM and IND - Statistics of U.S. Businesses (SUSB, a program run by US Census)
- Started modeling with Decision Tree and Linear Regression
- Explored Research Questions

Dataset

Rows = 50_states * 20_years (2001-2020) = 1000

Features - 123 (Fuel Type - 37 unique)



Back Fill Model

Problem - Electric Accounts data was not recorded for **2001-2008**

Solution - Predict using other **yearly** data with **greater** than **0.75 R^2**

Approach 1 - US Census

Sector	R-Squared
RES	0.942734
COM	0.855026
IND	0.690939

≈ + 0.02

≈ + 0.09

Approach 2 - SUSB Employment Count

Sector	R-Squared
RES	0.942734
COM	0.878632
IND	0.787149

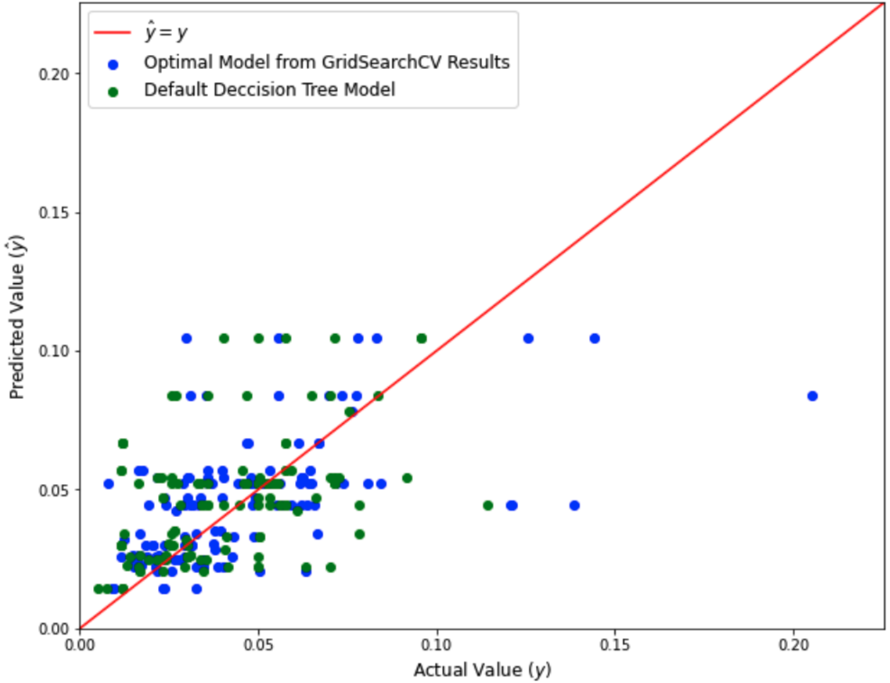
BackFill 2001-2008 - Approach 2

- Was given a description of types of business
- Made assumption on key words for IND and COM classification
- Used a string search so plurals, prefixes, and suffixes did not impact the classification

Decision Tree Regression Model

- Grid Search Cross Validation for parameter optimization
  - Criterion (Loss Function)
  - Max Number of Features
  - Max Tree Depth
  - Minimum Number of Sample per Leaf

	R^2	Adjusted R^2	MSE
Optimal Model from GridSearchCV Results	0.2715	-9.4416	0.0006
Default Parameters	0.1139	-11.70	0.0007



Ridge Regression

Feature Filtering

Comparing different models

Dealing with multicollinearity

```
Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)
===================================================================
Training MAE: 1.8892762804663488
-------------------------------------------------------------------
Validation MAE: 1.739064984907136
-------------------------------------------------------------------
Validation R2 score: 0.16805893803490302
===================================================================
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
===================================================================
Training MAE: 1.6962987856290794
-------------------------------------------------------------------
Validation MAE: 1.7966512572361908
-------------------------------------------------------------------
Validation R2 score: 0.018216250226247285
===================================================================
MSE train: 8.927, test: 6.758
```

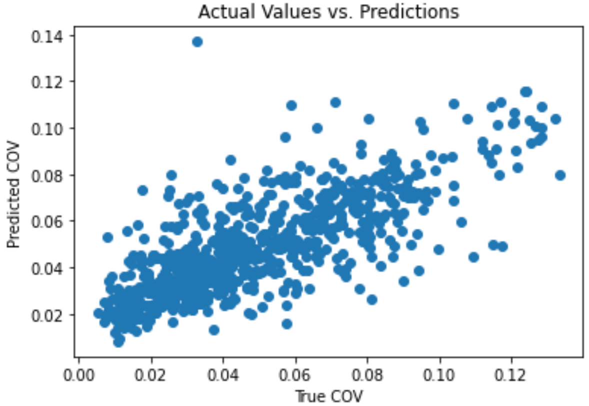
Plain Linear Regression Model

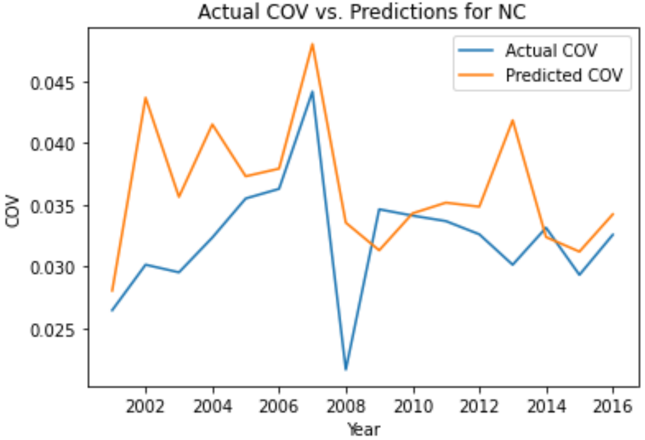
- **MSE ->** 0.0003 ; **Adjusted R-squared ->** 0.588

OLS Regression Results

Dep. Variable:	target_COV	**R-squared:**	0.591
Model:	OLS	**Adj. R-squared:**	0.588
Method:	Least Squares	**F-statistic:**	189.4
Date:	Wed, 03 Nov 2021	**Prob (F-statistic):**	5.55e-149
Time:	22:32:37	**Log-Likelihood:**	2087.1
No. Observations:	794	**AIC:**	-4160.
Df Residuals:	787	**BIC:**	-4128.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Months Above Average Price Movement BZ=F	-0.0010	0.000	-5.203	0.000	-0.001	-0.001
Months Above Average Price Movement NG=F	-0.0021	0.000	-5.558	0.000	-0.003	-0.001
Months Above Average Price Volume NG=F	-0.0020	0.001	-3.871	0.000	-0.003	-0.001
Months Above Average Price Volume HO=F	0.0032	0.001	5.608	0.000	0.002	0.004
COV	0.7107	0.022	32.112	0.000	0.667	0.754
Agg_total_plant_count_x	8.325e-07	3.94e-07	2.111	0.035	5.82e-08	1.61e-06
Constant	0.0189	0.004	4.435	0.000	0.011	0.027





Next Steps

- **Data Backfill**
  - Improve accuracy of word classification by using lemmatization before searching
- **Dataset Engineering to Improve Model Performance**
  - **Sampling**
    - Synthetic Minority Over-Sampling Technique (SMOTE) for Regression
  - **Feature Engineering**
    - Reduce Plant Fuel Type by aggregation categories. Ex. Coal, Biomass
    - Remove features based on thresholds of correlation. Ex. **corr > list(0.30,0.40,...)**
- **Modeling**
  - Reframe problem to classification where classes are ranges of COV (i.e [0.05 - 0.10])
  - Evaluate what features are the strongest for our model (i.e LR - Weights)
- **Logging -** Log performance and model metadata to understand experiments overtime