Model Card

**Forecasting TTF Natural Gas Prices**

See the [example Google model cards](https://modelcards.withgoogle.com/model-reports) for inspiration.

## Model Description

#### Linear regression

       One of the most basic types of regression in machine learning, linear regression comprises a predictor variable and a dependent variable related to each other in a linear fashion. Linear regression involves the use of a best fit line. We should use linear regression when our variables are related linearly. For example, if we are forecasting the effect of increased advertising spend on sales. However, this analysis is susceptible to outliers, so it should not be used to analyze big data sets.

**Support Vector Machines**

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

#### Random Forest

A random forest is a machine learning technique that’s used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

#### Bayesian Ridge

       In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response, y, is not estimated as a single value, but is assumed to be drawn from a probability distribution. The model for Bayesian Linear Regression with the response sampled from a normal distribution is:

𝑦~𝑁(𝛽𝑇𝑋,𝜎2𝐼)

The aim of Bayesian Linear Regression is not to find the single “best” value of the model parameters, but rather to determine the posterior distribution for the model parameters. Not only is the response generated from a probability distribution, but the model parameters are assumed to come from a distribution as well.

\*\* Data Features \*\*

All the regression models were fitted with the following features as is

Open

High

Low

Close

Volume

Dividends

Stock Splits

\*\*Input:\*\*

Open

High

Low

\*\*Output:\*\*

Close

\*\*Model Architecture:\*\*

## Performance

Model Performance using K-Fold Crosss Validation

####################################################

Linear Regression

Chart

Description automatically generated

Regression Metrics Report

Mean Absolute Error: 0.194514

Mean Squared Error: 0.291012

Root Mean Squared Error: 0.539455

Mean Squared Log Error: 0.000075

Median Absolute Error: 0.030114

R² : 0.999939

Adjusted R² : 0.999938

Mean Poisson Deviance: 0.004673

Mean Gamma Deviance: 0.000078

Max Error: 3.730768

Explained Variance: 0.999939

Mean Absolute Percentage Error: 0.003174

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Percentiles

5: -0.485795

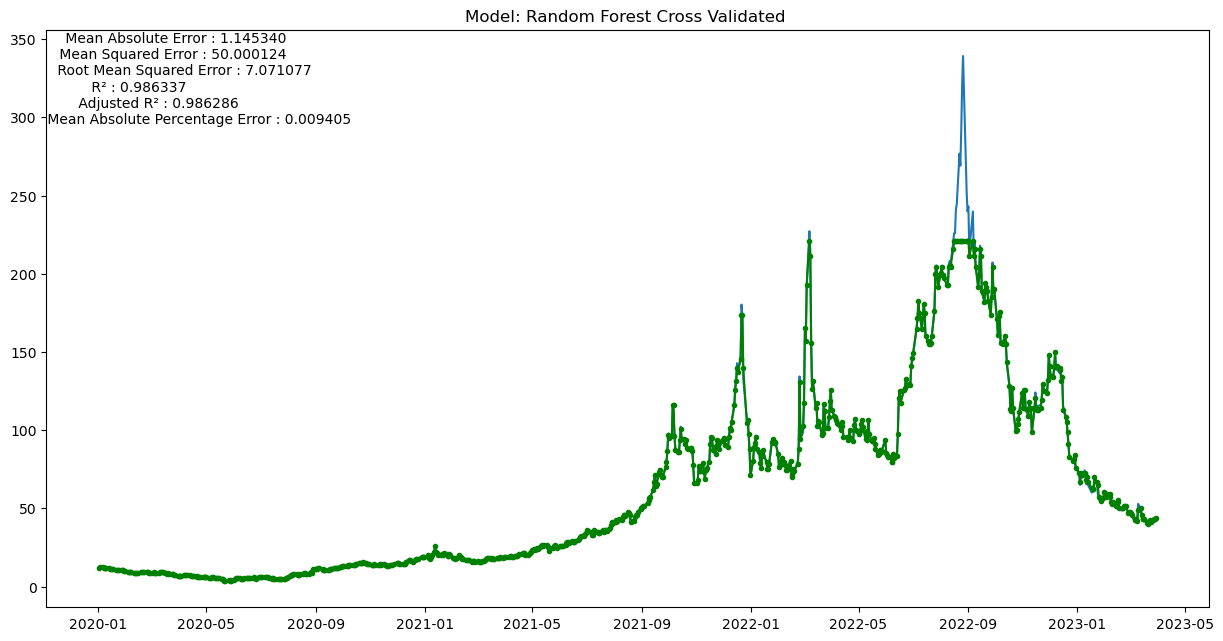
25: -0.042830

50: -0.013930

75: 0.000157

95: 0.375633

Random Forest Regression



Regression Metrics Report

Mean Absolute Error: 4.621943

Mean Squared Error: 248.589515

Root Mean Squared Error: 15.766722

Mean Squared Log Error: 0.003938

Median Absolute Error: 0.816886

R² : 0.947660

Adjusted R² : 0.946672

Mean Poisson Deviance: 1.017762

Mean Gamma Deviance: 0.004391

Max Error: 117.980851

Explained Variance: 0.950228

Mean Absolute Percentage Error: 0.024507

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Percentiles

5: -2.004225

25: -0.823430

50: -0.145828

75: 0.792962

95: 21.588835

Chart

Description automatically generated

Regression Metrics Report

Mean Absolute Error: 36.702407

Mean Squared Error: 4831.197075

Root Mean Squared Error: 69.506813

Mean Squared Log Error: 0.213018

Median Absolute Error: 2.688681

R² : -0.017199

Adjusted R² : -0.036392

Mean Poisson Deviance: 37.821451

Mean Gamma Deviance: 0.319752

Max Error: 267.675460

Explained Variance: 0.251335

Mean Absolute Percentage Error: 0.180284

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Percentiles

5: -2.301520

25: -0.884486

50: 2.423998

75: 59.236544

95: 163.328662

Bayesian Ridge Regression

Chart

Description automatically generated

Regression Metrics Report

Mean Absolute Error: 0.194656

Mean Squared Error: 0.291007

Root Mean Squared Error: 0.539451

Mean Squared Log Error: 0.000075

Median Absolute Error: 0.030129

R² : 0.999939

Adjusted R² : 0.999938

Mean Poisson Deviance: 0.004672

Mean Gamma Deviance: 0.000078

Max Error: 3.724794

Explained Variance: 0.999939

Mean Absolute Percentage Error: 0.003176

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Percentiles

5: -0.487415

25: -0.042849

50: -0.013941

75: 0.000151

95: 0.377215

## Improvements

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Linear Regression

Improved by 0.362%

Base Accuracy : 99.32271143537098

Cross Validated Accuracy : 99.68262119124071

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Linear Regression

Improved by nan%

Base Accuracy : 99.32271143537098

Test Improved Accuracy : nan

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Linear Regression

Improved by nan%

Base Accuracy : 99.68262119124071

Cross Validated Test Improved Accuracy : nan

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Random Forest

Improved by -1.151%

Base Accuracy : 98.68564760171864

Cross Validated Accuracy : 97.54931056968233

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Random Forest

Improved by -0.246%

Base Accuracy : 98.68564760171864

Test Improved Accuracy : 98.44293188581312

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Random Forest

Improved by -0.812%

Base Accuracy : 97.54931056968233

Cross Validated Test Improved Accuracy : 96.75760495605316

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Support Vector

Improved by -3.697%

Base Accuracy : 85.11850978384804

Cross Validated Accuracy : 81.97158866138835

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Support Vector

Improved by 16.703%

Base Accuracy : 85.11850978384804

Test Improved Accuracy : 99.33608388479102

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Support Vector

Improved by 21.486%

Base Accuracy : 81.97158866138835

Cross Validated Test Improved Accuracy : 99.58366845614128

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bayesian Ridge

Improved by 0.362%

Base Accuracy : 99.32259687374373

Cross Validated Accuracy : 99.6823887260662

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bayesian Ridge

Improved by nan%

Base Accuracy : 99.32259687374373

Test Improved Accuracy : nan

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bayesian Ridge

Improved by nan%

Base Accuracy : 99.6823887260662

Cross Validated Test Improved Accuracy : nan

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Best Model based

Support VectorHyperparameter boosted Cross Validated model governing parameters

C : 10

cache\_size : 200

coef0 : 0.0

degree : 3

epsilon : 0.5

gamma : 1e-07

kernel : linear

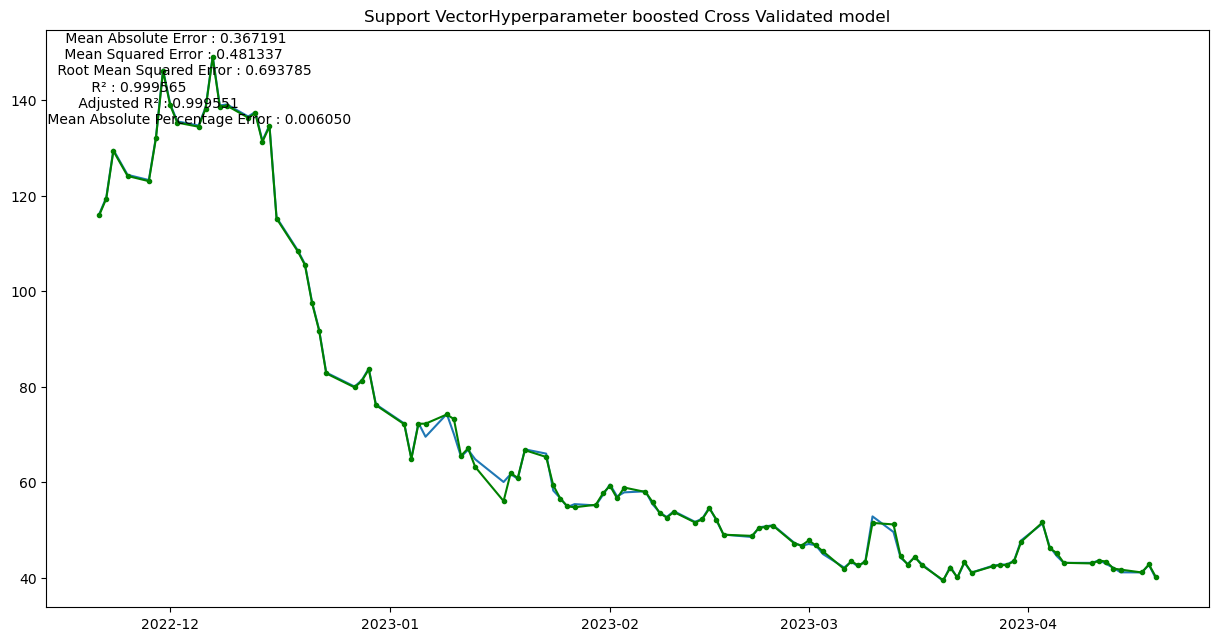
max\_iter : -1

shrinking : True

tol : 0.001

verbose : False

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



LSTM Model Forward Price Prediction

Chart, histogram

Description automatically generated

## Limitations

**Linear Regression:**

* The cross-validated accuracy improved by 0.362%, indicating that the model might be overfitting the data.
* The test accuracy improvement is not available, so it is unknown how the model performs on unseen data.
* The cross-validated test improvement is not available, so it is unknown how the model generalizes to new data.

**Random Forest:**

* The cross-validated accuracy decreased by 1.151%, indicating that the model might not be able to capture the complex relationships between the features and the target variable.
* The test accuracy improved by 0.246%, suggesting that the model might be better at generalizing to unseen data.
* The cross-validated test improvement is not available, so it is unknown how the model generalizes to new data.

**Support Vector:**

* The cross-validated accuracy decreased by 3.697%, indicating that the model might not be able to capture the complex relationships between the features and the target variable.
* The test accuracy improved by 16.703%, suggesting that the model might be better at generalizing to unseen data.
* The cross-validated test improvement increased by 21.486%, suggesting that the model might generalize well to new data.
* The best model has a high regularization parameter C, which might cause the model to underfit the data.

**Bayesian Ridge:**

* The cross-validated accuracy improved by 0.362%, indicating that the model might be overfitting the data.
* The test accuracy improvement is not available, so it is unknown how the model performs on unseen data.
* The cross-validated test improvement is not available, so it is unknown how the model generalizes to new data.

## Trade-offs

The trade-off between the models is the balance between accuracy, generalization, and overfitting. Support Vector seems to have the best performance among the models, with high test accuracy improvement and cross-validated test improvement, but it also has a high regularization parameter that might cause underfitting. The other models have varying degrees of accuracy and generalization performance but might suffer from overfitting or underfitting issues.