Analysis and Forecasting of Coal Prices

Pratyush Singh
Department of Informatics
University of Zurich
pratyush.singh@uzh.ch
19-762-988

Rohit Kaushik
Department of Informatics
University of Zurich
rohit.kaushik@uzh.ch
17-747-6680

Mesut Ceylan
Department of Informatics
University of Zurich
mesut.ceylan@uzh.ch
18-748-012

Andris Prokofjevs
Department of Informatics
University of Zurich
andris.prokofjevs@uzh.ch
18-796-755

***Abstract*—Forecasting plays an important role in various fields of the concern. With the help of forecasting, the enterprises can find out whether they can succeed in the new business; whether they can face the existing competition. In this paper, we describe the analysis and importance of time series forecasting by taking the example of the coal prices of South Africa. We perform extensive data analysis of time series and feature selection before training the classical ARIMA model, machine learning based XGBoost model and deep learning based LSTM model. We prove that selecting better features help in increasing model's performance. We also perform large-scale hyperparameter search using TALOS and report the results and compare all the models. We found that our XGBoost model outperforms ARIMA and LSTM model with an R^2 score of 0.91 as compared to R^2 0.83 of LSTM model and R^2 0.71 of ARIMA model.**

***Index Terms*—forecasting, time series, LSTM, ARIMA, OECD**

I. Introduction

Although many countries have been transitioning from traditional carbon-emission heavy energy resources such as natural gas, coal and oil to renewable and green energy sources, recent economic developments in Sub-Saharan Africa countries result in increased energy demand that still satisfied with fossil fuels, mainly coal.

Specifically, country of South Africa highly differs from others in Africa as it possesses largest coal reserves in the continent and is the 6th largest coal producer in the world. [1] The country utilizes coal as a dominant energy resource for domestic and continent usage. As coal is crucial energy source for South Africa: "77% of South Africa's energy is directly derived from coal" [2] and coal mining industry established in South Africa, consists of state-owned and private local as well as global enterprises conducting coal business. Thus, domestic and global coal price plays significant role for their businesses and commercial operations.

Accurately predicted coal prices provide a strategic and operational advantage for the private and state-own enterprises, covering supply chain and production planning and its effected elements. Additionally, it assists policy makers and governmental bodies to make guided and data-driven decision on such an important energy source not only in South Africa, but also globally. Price of coal in South Africa affects environmental efforts of reducing CO2 impact and coal usage from EU and OECD countries, who are continuously decreasing their coal production and consumption. [1], [3] and also may be of interest for rising economies with high demand for energy i.e China, which are increasing their coal consumption over time [1].

With this end, this study introduces the time series analysis of coal prices in South Africa discussing seasonality, stationarity, and trend analysis. Additionally, we perform predictive modelling with varying statistical and machine learning approaches for coal price prediction in South Africa between the years of 2016 and 2020. Our contributions to this project are as follows:

- We present and discuss the extensive data analysis of time series and analyze for stationarity, seasonality and trend in time series qualitatively and quantitatively.
- We perform feature selection on the basis of exploratory data analysis and prove that our selected features help in increasing the predictive model performances.
- We train three models: a classical ARIMA model, a machine learning based XGBoost model and a deep learning based LSTM model
- We also perform in-depth hyperparameter tuning of the LSTM model using TALOS library

The remainder of this report is structured as following: In Section II, we describe the dataset and perform an exploratory data analysis to deal with outliers, analyze for the trend, stationarity and seasonality of time series and select the relevant features for training. The proposed Models and Experimental Setup is discussed in Section III. In Section IV, we provide an insight to the obtained results and key observations along with the comparison of all the models and lastly, Section V, concludes the report with discussion and future work.

II. Data Analysis and Feature Selection

A. Dataset

The utilized dataset in this study consists of 343 observations: monthly time series between January 1992 to July 2020, and the 17 following attributes:

- Coal price in Australia ($ / tonne)
- Coal price in South Africa ($ / tonne)
- Crude oil price ($ / barrel)
- Fuel Index
- Natural gas price ($ / Million BTUs[1])
- Coal Producer Index
- Gold price ($ / ounce)

[1]A British thermal unit (BTU) is a measure of the heat content of fuels or energy sources. [4]

- Silver price ($ / ounce)
- Copper price ($ / ounce)
- Iron Ore price ($ / Dry Metric Tonne)
- Crude Oil Index
- Metal Index
- Non-Fuel Index
- Commodity Index
- Composite Leading Indicator[2]
- Consumer Confidence Index[3]
- Business Confidence Index[4]

The example subset of described dataset is presented in Table I. The dataset is obtained from OECD website via scraping directly from the OECD public resources.

	Coal.Australia	Coal.sa	CrudeOil	Fuel.Index	NaturalGas
Date					
1992-01-01	39.5	31.0	17.38	23.71	1.28
1992-02-01	39.5	31.0	17.62	23.83	1.21
1992-03-01	39.5	31.0	17.45	23.79	1.28

	Coal.Producer.Index	Gold	Silver	Copper	IronOre
Date					
1992-01-01	93.6	354.45	4.11	2139.23	33.1
1992-02-01	93.8	353.91	4.15	2205.97	33.1
1992-03-01	93.6	344.34	4.11	2227.33	33.1

	CrudeOil.Index	Metal.Index	NonFuel.Index	Commodity.Index	CLI
Date					
1992-01-01	40.23	35.59	48.40	50.72	99.20515
1992-02-01	41.09	37.12	48.71	50.72	99.22767
1992-03-01	41.17	37.41	48.74	50.75	99.25794

	CCI	BCI
Date		
1992-01-01	99.36520	99.32731
1992-02-01	99.46713	99.51028
1992-03-01	99.62660	99.67402

TABLE I: Exemplary subset of dataset used in this study

B. Data Preprocessing

Standard preprocessing steps were taken in this study, firstly dataset was checked for missing values, resulting that the dataset carries only four null values in different attributes as shown in Table II. Since number of observations is very limited, missing values were imputed with mean values of respective attribute to maintain observations, instead of discarding.

[2]designed to provide early signals of turning points in business cycles showing fluctuation of the economic activity around its long term potential level. CLIs show short-term economic movements in qualitative rather than quantitative terms

[3]indication of future developments of households' consumption and saving, based upon answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment and capability of savings. An indicator above 100 signals a boost in the consumers' confidence towards the future economic situation, as a consequence of which they are less prone to save, and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude towards future developments in the economy, possibly resulting in a tendency to save more and consume less.

[4]provides information on future developments, based upon opinion surveys on developments in production, orders and stocks of finished goods in the industry sector. It can be used to monitor output growth and to anticipate turning points in economic activity. Numbers above 100 suggest an increased confidence in near future business performance, and numbers below 100 indicate pessimism towards future performance.

Feature	Number of NaN values
Coal.Australia	0
Coal.sa	0
CrudeOil	0
Fuel.Index	1
NaturalGas	0
Coal.Producer.Index	0
Gold	1
Silver	0
Copper	0
IronOre	1
CrudeOil.Index	0
Metal.Index	0
NonFuel.Index	0
Commodity.Index	0
CLI	1
CCI	0
BCI	0

TABLE II: Number of missing values by each attribute across whole dataset

Additionally, dataset was analyzed for possible outliers for every attribute with monthly and yearly perspective. Figure 1 depicts monthly boxplot of "Coal price in South Africa", clustering prices to respective months from all years included in the dataset. With similar fashion, Figure 2 shows boxplots for yearly clustered"Coal price in South Africa".

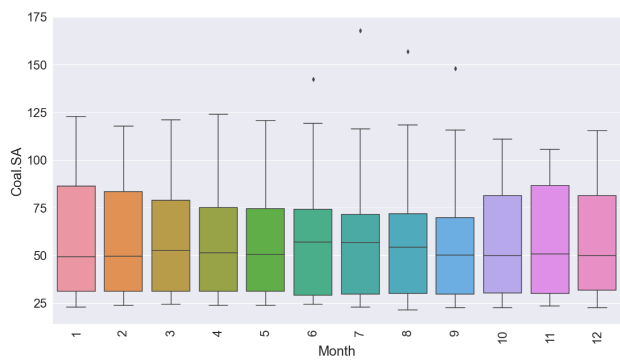


Fig. 1: Boxplots of clustered "Coal prices in South Africa" by months from all years.

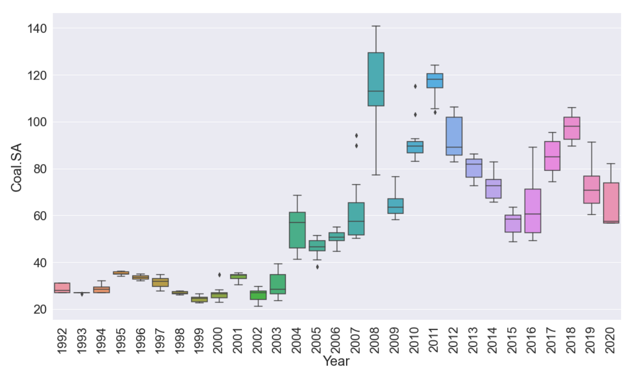


Fig. 2: Boxplots of clustered "Coal prices in South Africa" by years.

Many attributes had low number of outliers. In order to account for that, we benefited from IQR (interquartile range) scores to detect outliers. For the case of detected outliers were below the lower whisker, we assigned `lower_whisker + 1` to those values, whereas for the case of they were above the upper whisker, we assigned `upper_whisker - 1` to them. We believe the main reason for the observed outliers is the economic crisis in 2008. Figure 3 and Figure 4 represents the boxplots after removing the outliers for our example case using “Coal price in South Africa”.

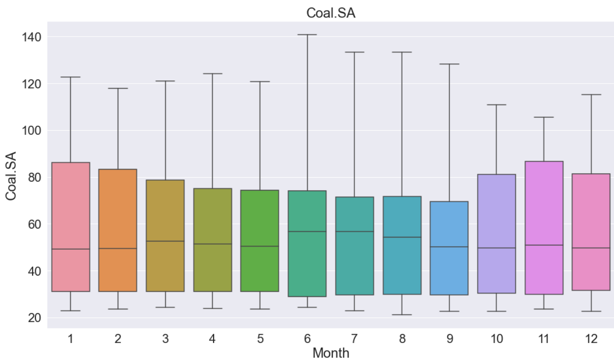


Fig. 3: Boxplots of clustered “Coal prices in South Africa” by months from all years after outlier removal.

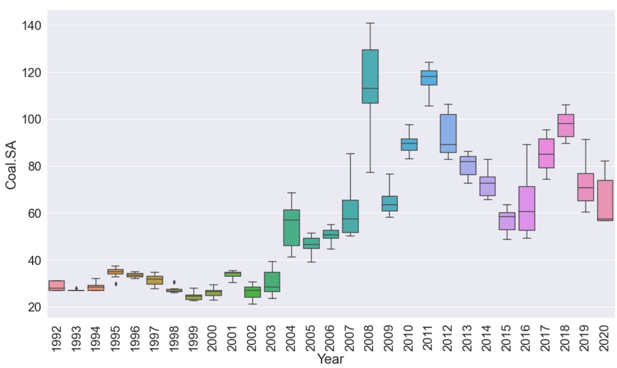


Fig. 4: Boxplots of clustered “Coal prices in South Africa” by years after outlier removal.

Above-explained methodology finalizes the data preprocessing steps and prepares data for exploratory data analysis.

C. Exploratory Data Analysis

Prior to formulating well-performing machine learning pipeline, this section sheds a light on main characteristics of the dataset and answers the main model design questions. Fundamental goal of this section is to analyze whether the time series of “Coal price in South Africa” has a trend, seasonality and stationarity.

1) Trend: It is defined as ”a pattern in data that shows the movement of a series to relatively higher or lower values over a long period of time” [5]. In order to observe if addressed time series has a trend, visual analysis is conducted.

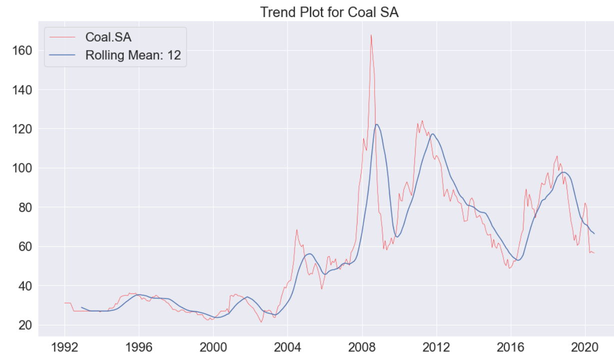


Fig. 5: Coal price in South Africa over time (red) and mean value of this attribute (from 12 Months) (blue)

Time series in Figure 5 reveals that the coal prices have generally increased over time. The variance is low between 1992-2002 and is generally higher from 2004 onwards. To account for the annual effect, the simple moving average, as the unweighted mean of the previous M=12 data points, was computed and compared. We can see that Coal Price in South Africa series has a trend. The price raises until around 2008 and then the trend appears to change direction or at least stay stationary. However, observing the fact that price values never do fall significantly under the values at the beginning of 2008 peak, we consider the Coal Price in South Africa series to have generally positive trend. Due to the presence of the trend, we can state that the data is not stationary based on qualitative analysis, yet such a hypothesis needs to be confirmed quantitatively, thus we perform the statistical tests to check for the stationarity.

2) Stationarity: To check whether concerned time series is stationary, thus “the statistical properties of a process generating a time series do not change over time” [6], we initially analyzed the autocorrelation plot depicted in Figure 6, comparing relation of autocorrelation values and confidence intervals. Theoretically, if most of the (95%) autocorrelation values are within the confidence interval, we can conclude that the series is stationary based on this observation.

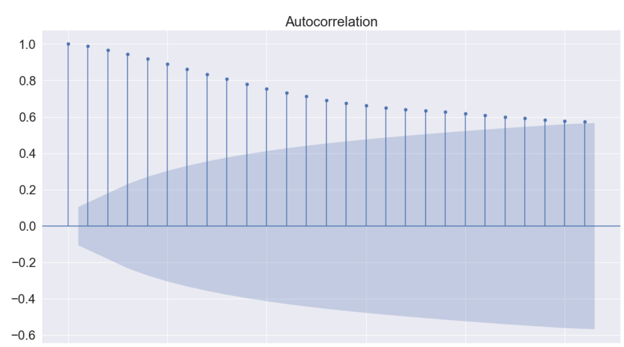


Fig. 6: Autocorrelation plot for Coal price in South Africa, indicates lags on x-axis while correlations on y-axis.

Analyzing the ACF plot proved that none of the lags are within the confidence interval, thus we conclude that the Coal price in South Africa series is non-stationary. However, if the differenced series analyzed, where each value is a difference of current and previous value, one can interpret that differenced series highlight stronger stationarity than the original series as we can observe in Figure 7.

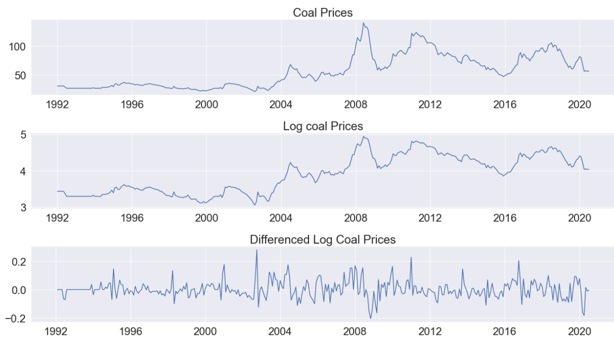


Fig. 7: Given Coal price in South Africa, along with log coal and differenced price series.

Qualitative analysis of the time plots lead to further inspection of stationarity using Augmented Dickey-Fuller Test (ADF), the extended version of simple Dickey Fuller test by including extra lagged in terms of the dependent variables to eliminate the autocorrelation problem. Dickey Fuller test in turn examines the null hypothesis of an autoregressive integrated moving average (ARIMA) against stationary [7]. The test results comprise of a Test Statistic and critical values for different confidence levels. Theoretically, if the test statistic value is less than the critical value (0.05), one can conclude that time series is stationary.

ADF test statistic	-1.820749
p-value	0.370210
# lags used	13.000000
# observations	329.000000
critical value (1%)	-3.450384
critical value (5%)	-2.870365
critical value (10%)	-2.571472

TABLE III: ADF results

According to the results shown in Table III, we observe weak evidence against the null hypothesis, thus rejection of the null hypothesis fails and we conclude that data has a unit root and is non-stationary. Aligned with qualitative observations, we observe lower test statistic value for the same test conducted on differenced data, shown in Table IV.

ADF test statistic	-8.133588e+00
p-value	1.073929e-12
# lags used	2.000000e+00
# observations	3.390000e+02
critical value (1%)	-3.449788e+00
critical value (5%)	-2.870104e+00
critical value (10%)	-2.571332e+00

TABLE IV: ADF results for differenced data

Here, we observed strong evidence against the null hypothesis, thus rejected it and concluded that data has no unit root and is stationary. Thus, in order to make the series stationary, we are using differencing to trend effects from the original series.

3) Seasonality: Another important time series property we investigate is seasonality: “the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly” [8]. If present, seasonality can be easily observed in seasonal plot such as Figure 8.

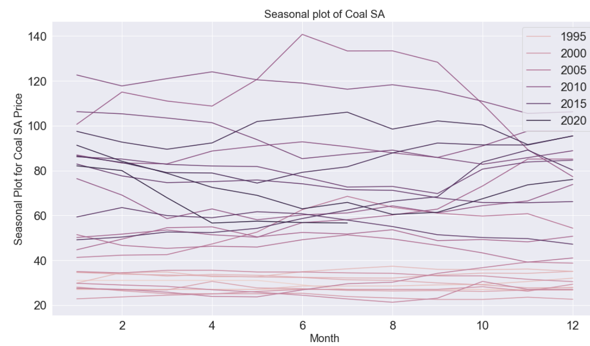


Fig. 8: Seasonal plot for Coal price in South Africa

Based on qualitative analysis from the seasonal plot, we did not observe significant increase or decrease among months over the years, although in the obvious outlier, price increase from April to July is evident. However, the overall plots reflects no ”obvious” seasonality. To be absolutely confident in our conclusion, we build further seasonal boxplot in Figure 9 and subseries plot in Figure 10.

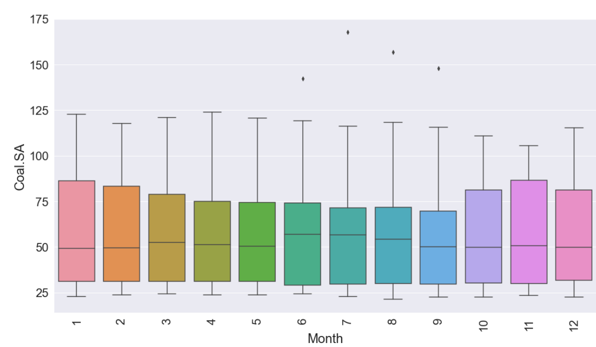


Fig. 9: Seasonal boxplot for Coal price in South Africa

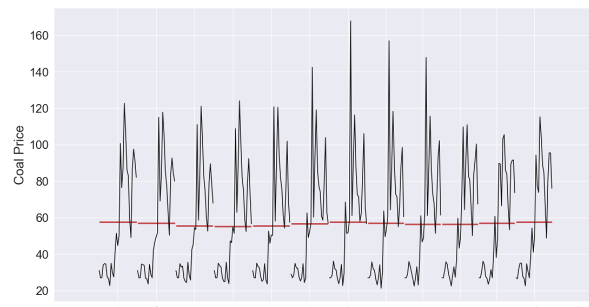


Fig. 10: Seasonal subseries plot for Coal price in South Africa

As it can be seen from the seasonal boxplot, the overall level of all months are similar to each other, therefore we can again conclude that there is no obvious seasonality. In attempt to explain repeatedly occurring outliers, we investigate the data for those levels and discover that main root of outliers are the years of 2008-2009 which is attributed to the economic crisis. Also, as it can be seen from the seasonal subseries plot, the overall level of all months are around the same, therefore indicating absence of seasonality.

Finally, we concluded seasonal decomposition using additive model and obtained following results presented in Figure 11. From the classical decomposition, we can conclude that there appears to be no significant seasonality in the data since the seasonal component ranges from -1 to 1 (in the additive case it does not have significant influence). We can observe seasonality plays a negligible role by plotting the seasonally adjusted and the original observations together. In the Figure 12, we can see that seasonality does not have a significant effect whereas the trend does.

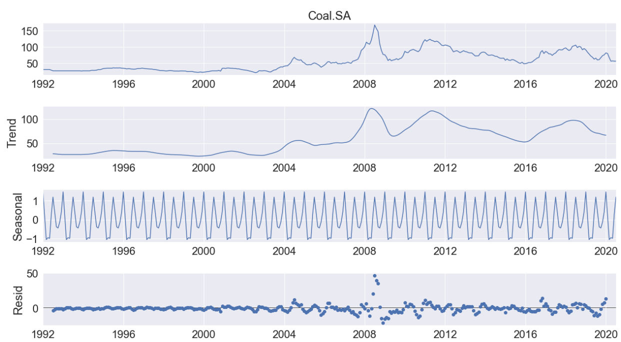


Fig. 11: Seasonal decomposition Coal price in South Africa

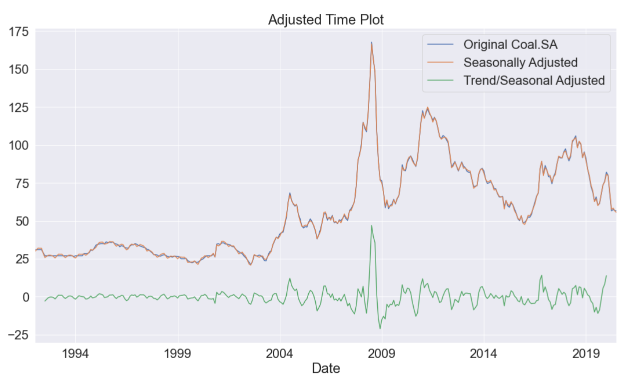


Fig. 12: Seasonality vs Trend Coal price in South Africa

D. Feature Selection

In order to select most useful feature subset from given 17 attributes, we performed a feature selection based on attribute correlations. Feature selection is crucial step for training the machine learning and deep learning models as redundant features increases the number of parameters and can also degrade the performance of the model. The choice was made as follows. First, we created two correlation heatmaps using Pearson and Spearman correlation in Figure 13 and Figure 14. The difference between them as described in [9] while both Pearson and Spearman are used for measuring the correlation, the difference between them lies in the type of analysis we expect: Pearson correlation evaluates the linear relationship between two continuous variables, while Spearman correlation evaluates the monotonic relationship. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data. Since we observe no substantial difference that should be considered between two different correlation coefficients depicted in Figure 13 and Figure 14, we will continue with our feature selection approach using Pearson Correlation Coefficient.

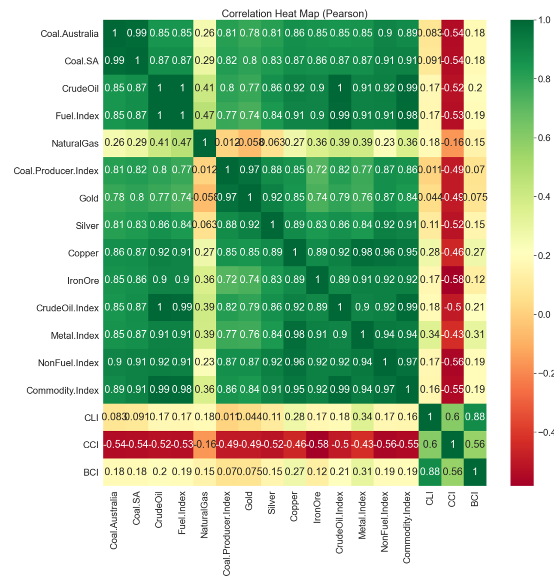


Fig. 13: Pearson correlation heatmap for all attributes

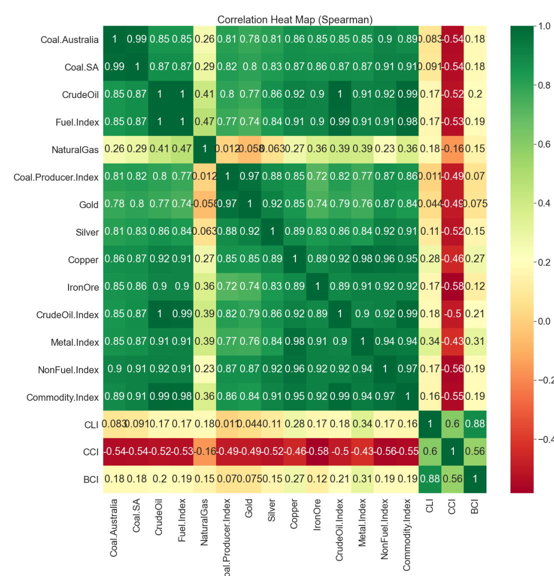


Fig. 14: Spearman correlation heatmap for all attributes

We sorted out the features that have correlation with our target variable lower than 0.5 resulting following list in Table V.

Coal.Australia	0.987597
Coal.sa	1.000000
CrudeOil	0.871322
Fuel.Index	0.873123
Coal.Producer.Index	0.823938
Gold	0.801117
Silver	0.829874
Copper	0.874786
IronOre	0.862946
CrudeOil.Index	0.870686
Metal.Index	0.865686
NonFuel.Index	0.910450
Commodity.Index	0.907782
CCI	0.539615

TABLE V: Filtered attributes and their pearson correlations with target attribute

Next, we calculated the correlation between these variables. The reason why we need to examine the correlation between the attributes besides our target attribute, is to find the unnecessary ones, which will possibly cause multi correlation. Thus, we have to detect them and remove them from the feature list.

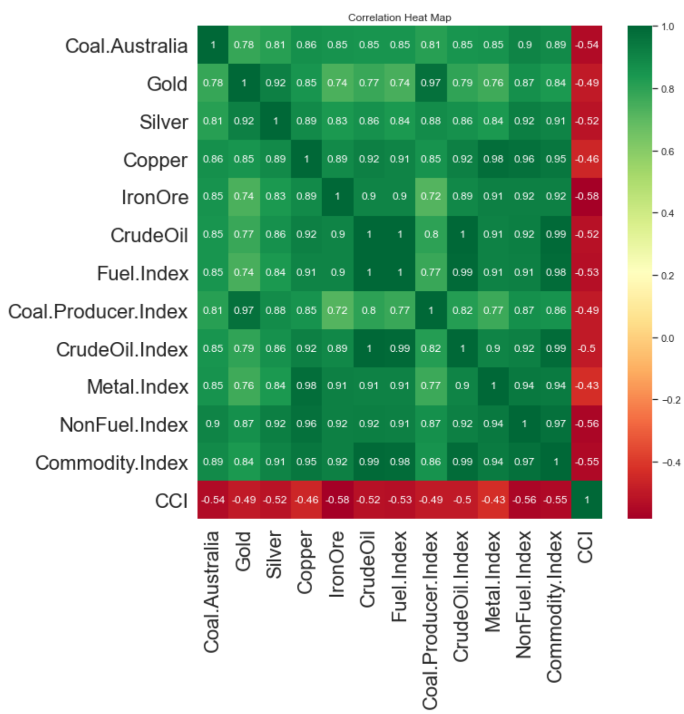


Fig. 15: Feature correlation heatmap

From correlation matrix in Figure 15, we got pairs of attributes who have high correlations (e.g. gold and silver) and investigate correlation of these variables with our target variable in Figure 16. Then variable from the pair, which has lower correlation with target variable is discarded.

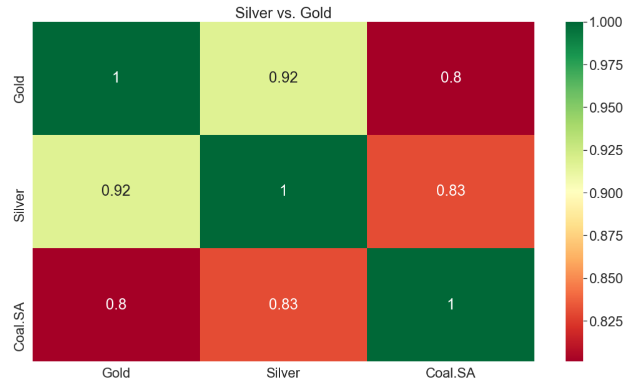


Fig. 16: Silver, Gold and Coal.SA correlation heatmap

In our example of determining between variables, silver and gold, we decide to drop variable gold, since it has slightly lower correlation with our target variable Coal.SA. This approach is repeated for all attributes, which were identified in previous step. In the end following list of features were selected:

- Coal price in Australia ($ / tonne)
- Coal price in South Africa ($ / tonne)
- Fuel Index
- Coal Producer Index
- Iron Ore price ($ / Dry Metric Tonne)
- Crude Oil Index
- Metal Index
- Non-Fuel Index
- Commodity Index
- Consumer Confidence Index

These features are utilized for further evaluations.

III. Models and Experimental Setup

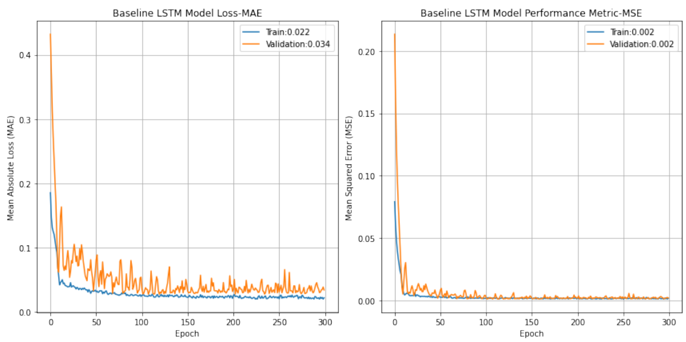
This study proposes three different models to tackle time series forecast problem, ARIMA as a statistical model, XG-Boost as a machine learning model and finally LSTM network as a black-box deep learning model.

In the following subsections, above-mentioned models are introduced.

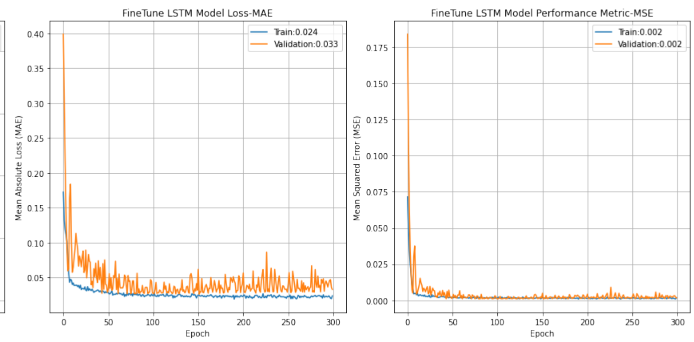
A. ARIMA

A time series is a sequence where a metric is recorded over regular time intervals. If only the previous values of the time series is used to predict its future values, it is called Univariate Time Series Forecasting. And if predictors other than the series (a.k.a exogenous variables) are used to forecast, it is then named Multi Variate Time Series Forecasting. In our project, we train both the models and demonstrate the differences.

ARIMA, short for 'Auto Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values [10]. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.



(a) Loss of Baseline LSTM



(b) Loss of Fine-tuned LSTM

Fig. 17: Loss and performance metric curves of Baseline and Fine-tuned LSTM models w.r.t number of epochs and measure.

An ARIMA model is characterized by 3 terms: p, d, q where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. To decide the p,q,d parameters we look at the ACF and PCF plots.

The data may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced data show the following patterns: the ACF is exponentially decaying or sinusoidal; there is a significant spike at lag p in the PACF, but none beyond lag p.

The data may follow an ARIMA(0, d, q) model if the ACF and PACF plots of the differenced data show the following patterns: the PACF is exponentially decaying or sinusoidal; there is a significant spike at lag q in the ACF, but none beyond lag q. More information in section III-E, training details.

B. XGBoost Method (tree based)

We also train a machine learning method using the XGBoost method which is a traditional tree based learning method. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

Salient features of XGBoost which makes it different from other gradient boosting algorithms include [11]:

- Clever penalization of trees
- A proportional shrinking of leaf nodes
- Newton Boosting
- Extra randomization parameter
- Implementation on single, distributed systems and out-of-core computation
- Automatic Feature selection

Before running XGBoost, we must set three types of parameters: general parameters, booster parameters and task parameters.

- General parameters relate to which booster we are using to do boosting, commonly tree or linear model
- Booster parameters depend on which booster you have chosen
- Learning task parameters decide on the learning scenario. For example, regression tasks may use different parameters with ranking tasks

C. LSTM

Apart from models introduced above, we introduce black-box deep learning model, namely Long-Short Term Memory architecture, and discuss its specification in this subsection.

Earlier efforts [12] and [13] in research extend capabilities of feedforward neural network structure by introducing feedback mechanism between neural units, resulting Recurrent Neural Network (RNN). Despite their capabilities, RNN networks suffer from vanishing/exploding gradient problem proven by [14], [15] and [16]. To over come these difficulties, first [17] introduced Long-Short Term Memory (LSTM) units that capable of capturing long term dependencies between observations and tackling vanishing/exploding gradient issues. Within the problem scope of this study, LSTM units are proven to be highly successful on time series sequences.

D. Performance Metrics

There are two performance metrics used in this study to evaluate and compare model performances. First performance metric is R^2 (*R-Squared*), the metric measures goodness of fit of regression model and indicates model performance on unseen observation by computing percentage of explained variance over total variance.

Second performance metric is *Root Mean Squared Error*, standard and widely used error metric for regression problems. This metric computes deviation between predicted and ground truth values for each sample and averages their squares.

R^2 may take values between 0 and 1 as well as negative values, closer to one indicated best model fit, as error metric

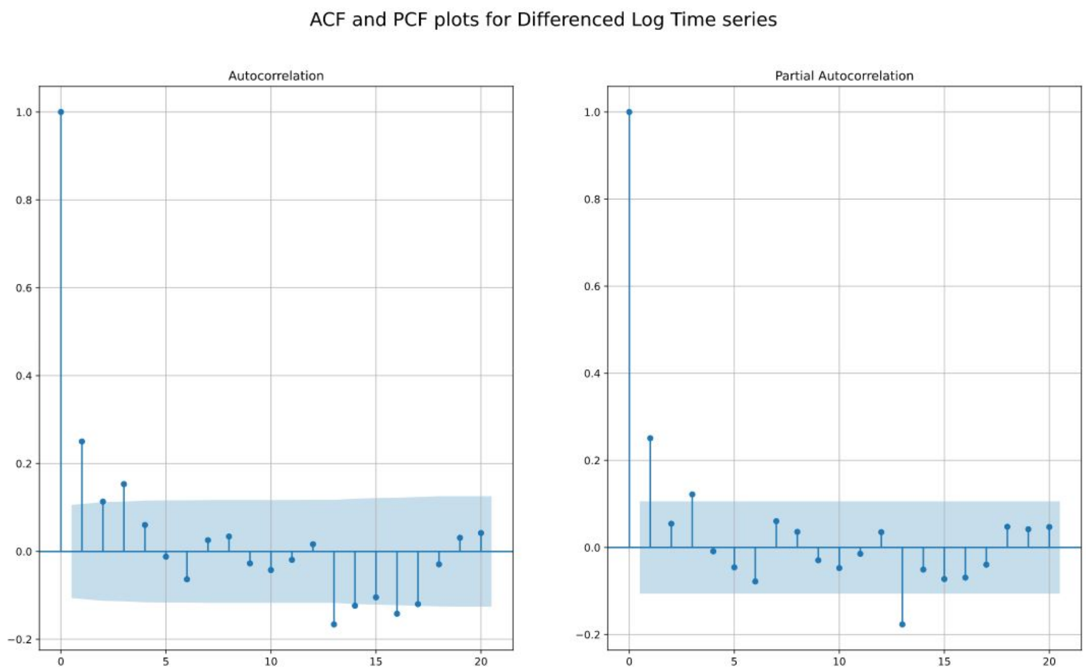


Fig. 18: ACF and PCF plot for the differenced time series

$RMSE$ takes positives values and lower $RMSE$ provides better model performance.

Performance metrics introduced above are computed as following:

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$RMSE(y_i, \hat{y}_i) = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where $n, y_i, \hat{y}_i \bar{y}$ corresponds to number of observations, groundtruth values, predicted values and mean value of groundtruth values respectively.

E. Training Details

1) ARIMA method: After investigating for stationarity and seasonality through several methods described before, we introduce the parameters chosen for ARIMA model in this section.

In order to choose p,d,q parameters we will use two different plots.

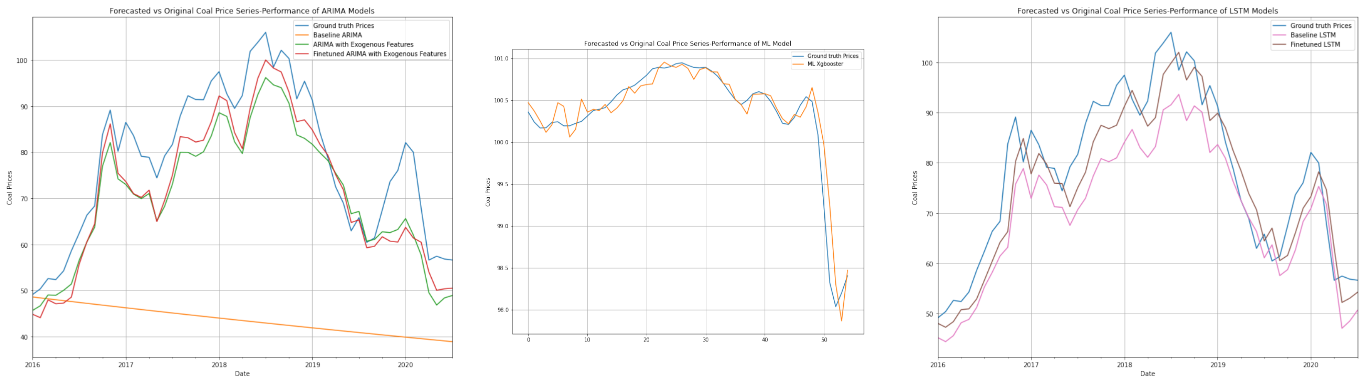
1) Autocorrelation Function (ACF): Measurement of the correlation between time series and lagged version of time series.
2) Partial Autocorrelation Function (PACF): This measures the correlation between the time series and lagged version of time series but after eliminating the variations already explained by the intervening comparisons

As seen from the Fig. 18, two dotted lines are the confidence intervals. We use these lines to determine the 'p' and 'q' values. Looking at the PACF plot, we can say that there is a significant spike at lag 1 and none beyond lag 1, therefore ARIMA(1,0,0) is a strong candidate. When we use the original series (not differenced), we need to use ARIMA(1,1,0) to take in account the differencing.

We train two models for ARIMA with and without the exogenous features. We use the standard `statsmodel` library to import `SARIMAX` model and set the seasonality to zero (since we train an ARIMA model). We also train a fine-tuned ARIMA model with exogenous features and search for the best 6 parameters in the model. We try all the combinations of the 6 parameters and use each combination individually to train our SARIMAX model. For each SARIMAX model, we look at the BIC/AIC [18], [19] scores and select the combination of parameters which gives us the least BIC/AIC score. We test the models for different values of p, q, d and seasonal_p, seasonal_q and seasonal_d in the range [0,2].

2) XGBoost method: We use the tree based booster model `gbtree` with a learning rate of 0.3 and regression with squared loss. We use the L1 regularization term on weights as 0.1 and L2 regularization term as 0.1. The maximum depth of the tree is chosen as 25 and the number of trees (or rounds) in an XGBoost model is specified as 1000.

3) LSTM method: Our baseline LSTM architecture is a small 2-layered LSTM model with 1 dropout layer. The hyper-parameters of this model are set manually and later we perform finetuning using TALOS. The baseline has 2 vertically stacked LSTM layers with 50 and 20 neurons. We use dropout of 0.15



(a) Performances of ARIMA models (b) Performance of XGBooster models (c) Performances of LSTM models

Fig. 19: Model performances compared w.r.t groundtruth coal prices of South Africa between 1992 and 2020

and Adam optimizer with a learning rate of 0.001. The model has total 18,289 trainable parameters.

The fine-tuned LSTM architecture using TALOS proposed in this study involves 52,289 trainable parameters. This simple yet effective network consists of vertically stacked 2-LSTM Layers, with a dropout layer in between. Both the layers consists of 64 units. Last LSTM layer is followed by one Fully Connected Layer with 1 neuron. Dropout is set at 0.15. This network is trained for 300 epochs by optimizing Mean Absolute Error loss function with Adam optimizer in which learning rate is set to 0.001, batch size to 64.

We also perform the extensive hyperparameter search for the LSTM model using `TALOS` and report the results for the best hyperparameter combination.

IV. Results

Table VI depicts performances of models on evaluation metrics. Among the models categorized into three different groups, black-box LSTM Deep Learning model outperforms all the statistical and machine learning models during our experiments.

When ARIMA model is evaluated, one can clearly observe that including exogenous features introduced in subsection II-D assists the model to perform better compared to vanilla ARIMA model trained on only target variable, as depicted in Fig. 19, South Africa's Coal price. Hence, we prove that our feature selection analysis is valid. Furthermore, final ARIMA model is fine-tuned by optimizing *order* and *seasonal order* parameters and achieved the best results among other ARIMA models as shown in Fig. 19 as expected with **0.71** R^2 and **8.49 RMSE** scores.

The best results are obtained using the XGBoost model which is a tree based model. We handcraft the model parameters in order to search for the best performing model. We obtain **0.91** R^2 and **0.197 RMSE** score. This is because the gradient boosting methods are powerful classifiers/regressors that typically perform very well on structured data. It is an ensemble learning algorithm, which combines the predictions of multiple base learners (usually, each one being a fairly weak performer on its own) to generate one overall prediction for each input. This allows it to learn more complex relationships between the features and labels in the training set.

Deep learning models equipped with LSTM units also perform well on this dataset however, it is not the best performing model. Baseline LSTM network outperforms ARIMA models by far and achieved the high R^2 score, **0.72**, and low **RMSE**, **8.40** as shown in the loss curve in Fig. 17 and forecasting plot in Fig. 19. The fine-tuned LSTM model using TALOS implementation performs better than baseline LSTM. It achieves R^2 score, **0.83**, and **RMSE**, **6.59**.

Models	R^2	RMSE
ARIMA-Baseline	-4.89	38.53
ARIMA w/ Exogenous Features	0.45	11.74
Fine-Tuned ARIMA w/ Exogenous Features	0.71	8.49
XGBoost	**0.91**	**0.19**
Baseline LSTM	0.72	8.40
Fine-tuned (TALOS) LSTM	0.83	6.59

TABLE VI: Summary of model performances on test set based on performance metrics introduced in subsection III-D.

V. Conclusion and Future Work

In this project, we analysed the timeseries for the forecasting of Coal prices in South Africa. We performed extensive 'Exploratory Data Analysis' with the help of which we performed data cleaning, data preprocessing along with feature selection and extraction. By selecting better features, we proved improvement in the model performance compared to the model with non-exogenous features. We investigate the time series for stationarity and seasonality using both qualitatively and quantitatively. We handled outliers in certain years. We believe the reason for the outliers is the economic crisis in 2008 as there was a sharp slowdown in demand, and with mining output remaining stubbornly high. Therefore, coal benchmarks fell down.

We trained a classical time series forecasting model - ARIMA (Auto Regressive Integrated Moving Average), a machine learning gradient boosting method - XGBoost and, a deep Learning method using LSTM along with extensive hyperparameter search using TALOS library.

We calculated the metric R^2 score (coefficient of determination) regression score function. The R^2 score obtained using XGBoost is 0.91 in comparison to 0.83 and 0.71 which is achieved by LSTM and ARIMA model respectively.

In the future, we would like to experiment different gradient boosting methods such as LightGBM (Light Gradient Boosting Machine) [20] or Facebook's Prophet [21] to analyse the performance of these models on this time-series data.

References

[1] International Energy Agency, "World coal consumption, 1978-2019," 2020, [Online; accessed 19-July-2021]. [Online]. Available: https://www.iea.org/data-and-statistics/charts/world-coal-consumption-1978-2019

[2] Department of Mineral resources and Energy of South Africa, "Coal resources overview," 2021, [Online; accessed 19-July-2021]. [Online]. Available: http://www.energy.gov.za/files/coal_frame.html

[3] Eurostat, "Coal production and consumption statistics," 2021, [Online; accessed 19-July-2021]. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php

[4] The U.S. Energy Information Administration, "Units and calculators explained," 2021, [Online; accessed 19-July-2021]. [Online]. Available: https://www.eia.gov/energyexplained/units-and-calculators/british-thermal-units.php

[5] tejalkadam18m, "What is a trend in time series?" 2021, [Online; accessed 19-July-2021]. [Online]. Available: https://www.geeksforgeeks.org/what-is-a-trend-in-time-series/

[6] Wikipedia contributors, "Seasonality — Wikipedia, the free encyclopedia," 2021, [Online; accessed 19-July-2021]. [Online]. Available: https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322

[7] R. Mushtaq, "Augmented dickey fuller test," *SSRN*, 08 2011.

[8] Palachy Shay, "Stationarity in time series analysis," 2021, [Online; accessed 19-July-2021]. [Online]. Available: https://en.wikipedia.org/wiki/Seasonality

[9] SERENO, "Comparison of pearson and spearman correlation coefficients," 2020, [Online; accessed 19-July-2021]. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/comparison-of-pearson-and-spearman-correlation-coefficients/

[10] S. L. Ho and M. Xie, "The use of arima models for reliability forecasting and analysis," *Comput. Ind. Eng.*, vol. 35, no. 1–2, p. 213–216, Oct. 1998. [Online]. Available: https://doi.org/10.1016/S0360-8352(98)00066-7

[11] Wikipedia contributors, "Xgboost — Wikipedia, the free encyclopedia," 2021, [Online; accessed 17-July-2021]. [Online]. Available: https://en.wikipedia.org/w/index.php

[12] P. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, pp. 1550 – 1560, 11 1990.

[13] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[15] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," 2013.

[16] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. B. Schön, "Beyond exploding and vanishing gradients: analysing rnn training using attractors and smoothness," 2020.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. [Online]. Available: http://www.jstor.org/stable/2958889

[19] Wikipedia contributors, "Akaike information criterion — Wikipedia, the free encyclopedia," 2021, [Online; accessed 17-July-2021].

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157.

[21] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. [Online]. Available: https://doi.org/10.1080/00031305.2017.1380080