Time-Series Forecasting for Predicting Sustainable Performance of Companies

Georgia Institute of Technology

CS6748 Practicum, Jun 2024

Midterm Review: Sprint 1

Team 2: R. ElGendi (relgendi3@gatech.edu), S. Lee (slee3510@gatech.edu), M. Palfenier (mpalfenier3@gatech.edu)

Agenda

1 **Problem Statement (Goals & Objectives)**
Hypothesis and assumptions

2 **Exploratory Data Analysis**
Contextual Analysis - Sustainability

3 **Forecasting Methods Exploration**
Method selections, pros & cons

4 **SPRINT 1 & Prelim-Findings**
From Lasso, LSTM,

5 **Q&A**
Feedback

6 **Reference & Appendix**

Predicting Sustainable Performance of Companies

1. Problem Statement (Goals & Objective): Hypothesis and assumptions

G&O: ClarityAI is looking to *enhance* the ***predictive accuracy*** *of sustainability indicators*, focusing on ***Scope 1 Emission value*** for mid-term forecasting (2022~2030), using ***time-series*** methods.

Hypothesis and Assumptions: Relationship between *Company's REVENUE* and SCOPE 1 EMISSION can be understood by 1) operational efficiency, 2) industry type and 3) regulatory environment. In this forecasting exercise we will apply 'industry type' to segment the data. Detailed reasons are as follows:

1) **Operational Efficiency**: Companies that optimize for energy efficiency tend to have lower Scope 1 emissions per unit of revenue. The efficiency gain leads to cost savings and higher profitability. However, in this scenario we *assume* that all companies and industries are embarking on the GHG accounting journey, ***process not mature*** *yet* to dissect the data into various population groups for time-series forecasting.

2) **Industry Type**: Industries such as manufacturing, mining, and oil & gas will produce higher Scope 1 emission due to the *nature their business operations.* As a result, close ties between revenue and emission levels can be expected (increased production leading to higher emission and higher revenue). Whereas *Service-oriented industries* like finance or technology is less likely to have directly correlated with emissions. As a result, **industry** *population is categorized into* ***two types, directly correlated and weakly correlated industries.***

3) **Regulatory environment**: companies operating in strict environmental regulatory countries may incur higher cost to reduce GHG emissions effecting revenue. While those in less stringent countries companies might have higher emission value with higher short-term revenue. In this forecasting exercise, we assume regulatory environment is not significant for categorizing companies as '*Carbon Tax' is* ***not penalizing*** *at a significant level to* ***enforce*** *business's* ***operational*** *and* ***structural change.*** Currently (in 2024) Sweden, Finland and Singapore implemented more progressive compliance mechanism and higher fines, however given encoded company names via ID and assumed global practice, ***regulatory environment tied to the HQ country is assumed not meaningful*** hence excluded.

Data Overview (1)

2. Exploratory Data Analysis

Data Split:

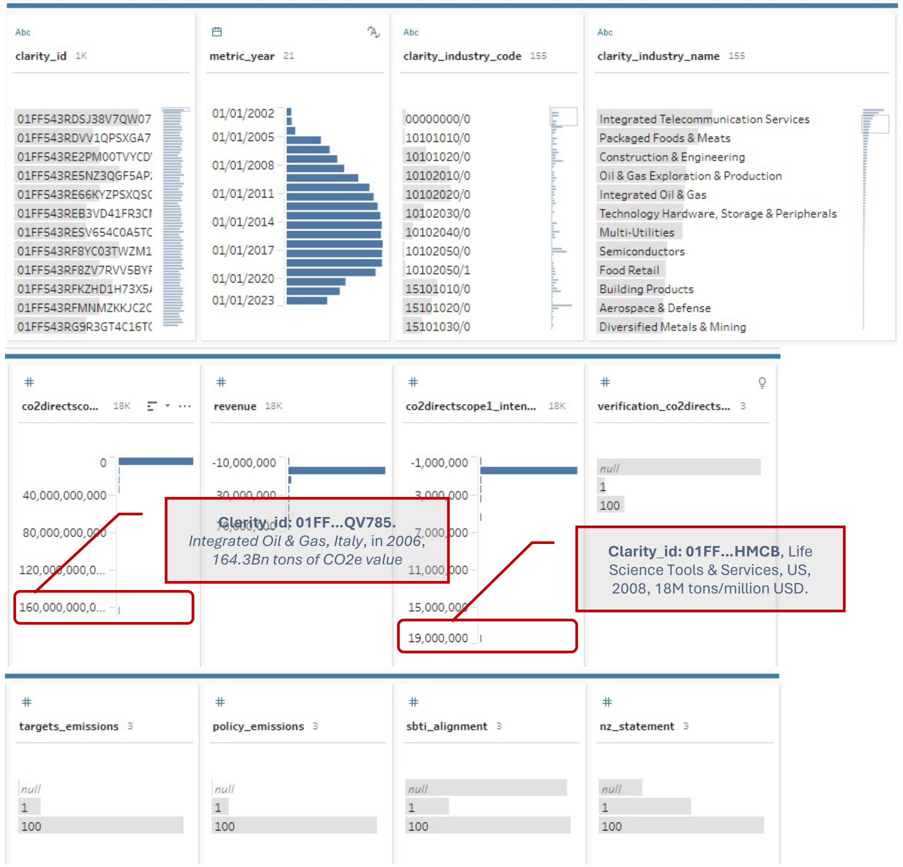
- Total Records (Training 18283 rows, Validation 274 rows), 98.5 vs 1.5 split
- Subset training data to validate algorithm efficiency

Attributes: (example of noticeable attributes)

- *clarity_id:* 1372 unique records = companies
- *metric_year:* majority data centered around 2009-2019, good sample size to build theory, noticeable year – a) global financial crises: 2007-2008, b)US presidential election: G.W. Bush – Obama transition in 2008 to 2019, b) COVID-19 in 2020-2021
- *clarity_industry_name*: is there a pattern around industry vs carbon intensity, cyclic industry unlike consumer business adding complexity to the prediction
- *country_code*: DGP vs sum of revenue considered however country at HQ level, insignificant
- *revenue*: revenue pattern against economic cycle as an adjustment factors
- verification_co2directscope1: 14% of population has 3rd party verification, relatively premium quality dataset, high potential for higher accuracy
- Heatmaps and country specific data evaluation can be found in the appendix

Visual Inspection & Outliers:

abnormality, outside of 96 percentile



Clarity_id: 01FF…QV785.
Integrated Oil & Gas, Italy, in 2006, 164.3Bn tons of CO2e value

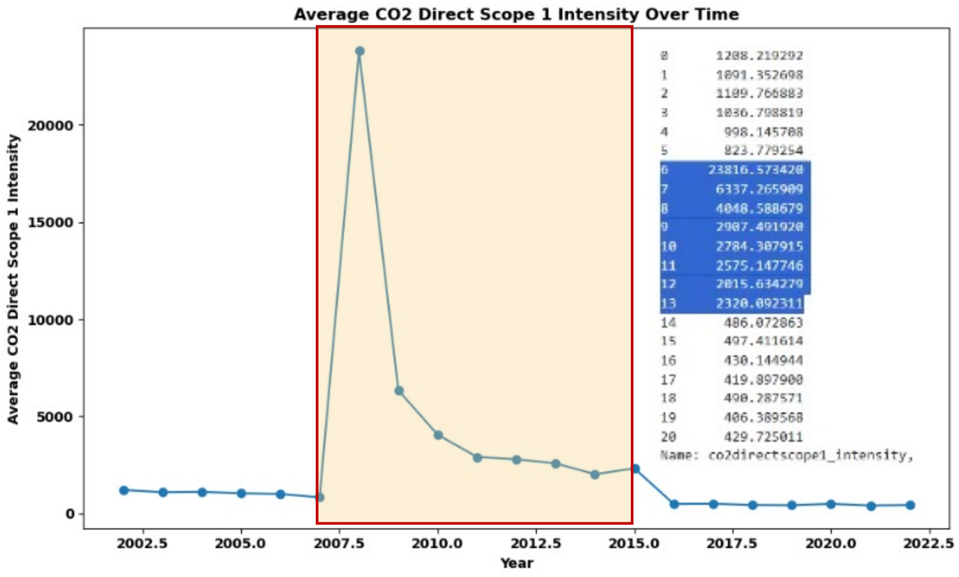
Clarity_id: 01FF…HMCB, Life Science Tools & Services, US, 2008, 18M tons/million USD.

Data Overview (2)

2. Exploratory Data Analysis … Outlier exclusion reasoning

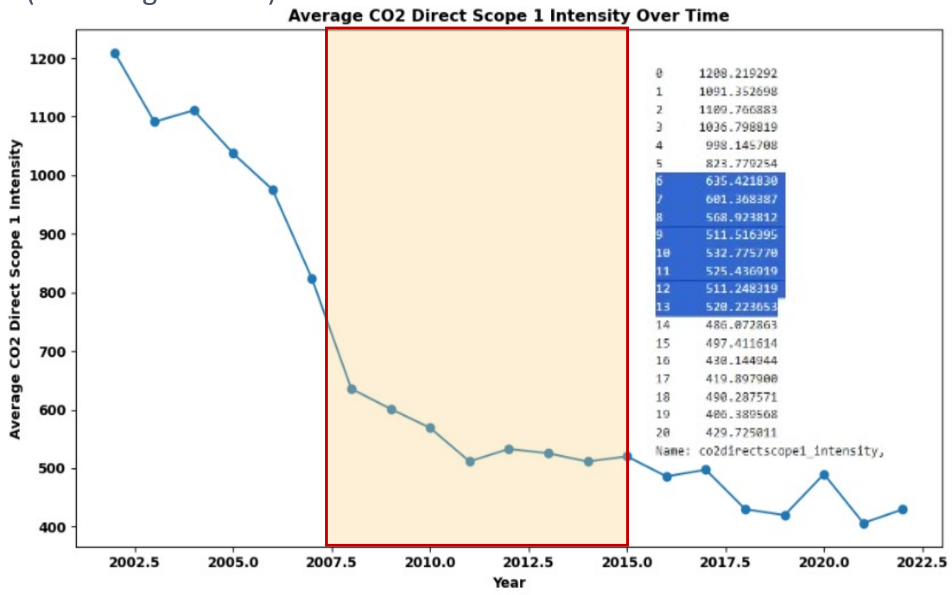
Average CO2 Direct Scope 1 Intensity Over Time

(As-is)



```
0       1208.219292
1       1091.352698
2       1109.766883
3       1036.798819
4        998.145708
5        823.779254
6      23816.573420
7       6337.265909
8       4048.588679
9       2907.491920
10      2784.307915
11      2575.147746
12      2015.634279
13      2320.092311
14       486.072863
15       497.411614
16       430.144944
17       419.897900
18       490.287571
19       406.389568
20       429.725011
Name: co2directscope1_intensity,
```

(Excluding Outliers)

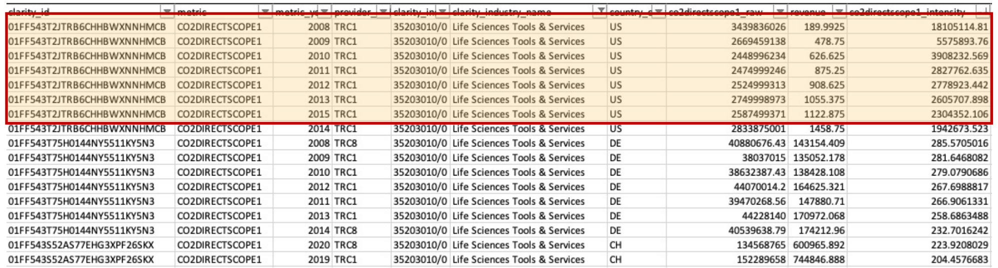


```
0       1208.219292
1       1091.352698
2       1109.766883
3       1036.798819
4        998.145708
5        823.779254
6        635.421830
7        601.368387
8        568.923812
9        511.516395
10       532.775770
11       525.436919
12       511.248319
13       520.223653
14       486.072863
15       497.411614
16       430.144944
17       419.897900
18       490.287571
19       406.389568
20       429.725011
Name: co2directscope1_intensity,
```

Excluded 2 companies from the data set:

- **Clarity_id: 01FF…QV785.** *Integrated Oil & Gas, Italy*, in *2006, 164.3Bn tons of CO2e value*
- **Clarity_id: 01FF…HMCB**, Life Science Tools & Services, US, 2008, 18M tons/million USD *(Across 2008 – 2015):*
  - *highest intensity value of any company*
  - *highest CO2 emission within the 'Life Sciences Tools & Services' peer group, and*
  - *lowest revenue of their peer group.*

(Not populating/filing the data gap with Gradient boosting)



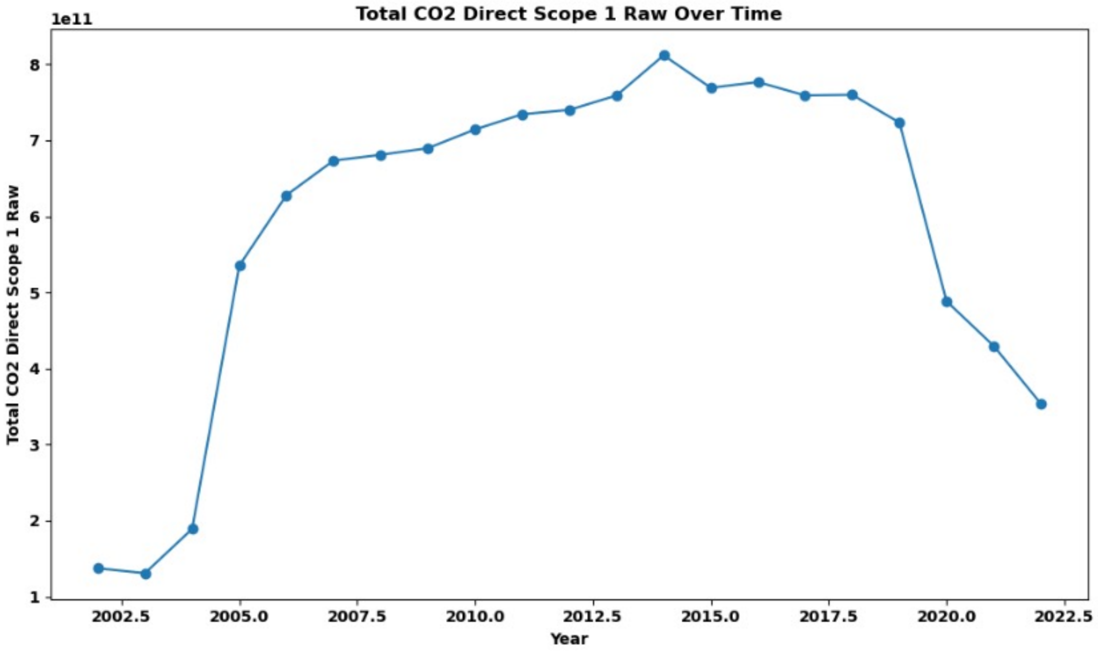
clarity_id	metric	metric_ye	provider_	clarity_in	clarity_industry_name	country_c	co2directscope1_raw	revenue	co2directscope1_intensity
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2008	TRC1	35203010/0	Life Sciences Tools & Services	US	3439836026	189.9925	18105114.81
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2009	TRC1	35203010/0	Life Sciences Tools & Services	US	2669459138	478.75	5575893.76
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2010	TRC1	35203010/0	Life Sciences Tools & Services	US	2448996234	626.625	3908232.569
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2011	TRC1	35203010/0	Life Sciences Tools & Services	US	2474999246	875.25	2827762.635
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2012	TRC1	35203010/0	Life Sciences Tools & Services	US	2524999313	908.625	2778923.442
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2013	TRC1	35203010/0	Life Sciences Tools & Services	US	2749998973	1055.375	2605707.898
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2015	TRC1	35203010/0	Life Sciences Tools & Services	US	2587499371	1122.875	2304352.106
01FF543T2JTRB6CHHBWXNNHMCB	CO2DIRECTSCOPE1	2014	TRC1	35203010/0	Life Sciences Tools & Services	US	2833875001	1458.75	1942673.523
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2008	TRC8	35203010/0	Life Sciences Tools & Services	DE	40880676.43	143154.409	285.5705016
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2009	TRC1	35203010/0	Life Sciences Tools & Services	DE	38037015	135052.178	281.6468082
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2010	TRC1	35203010/0	Life Sciences Tools & Services	DE	38632387.43	138428.108	279.0790686
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2012	TRC1	35203010/0	Life Sciences Tools & Services	DE	44070014.2	164625.321	267.6988817
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2011	TRC1	35203010/0	Life Sciences Tools & Services	DE	39470268.56	147880.71	266.9061331
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2013	TRC1	35203010/0	Life Sciences Tools & Services	DE	44228140	170972.068	258.6863488
01FF543T75H0144NY5511KY5N3	CO2DIRECTSCOPE1	2014	TRC8	35203010/0	Life Sciences Tools & Services	DE	40539638.79	174212.96	232.7016242
01FF543S52AS77EHG3XPF26SKX	CO2DIRECTSCOPE1	2020	TRC8	35203010/0	Life Sciences Tools & Services	CH	134568765	600965.892	223.9208029
01FF543S52AS77EHG3XPF26SKX	CO2DIRECTSCOPE1	2019	TRC1	35203010/0	Life Sciences Tools & Services	CH	152289658	744846.888	204.4576683

Data Overview (3)

2. Exploratory Data Analysis … Political Events

Addressing Political Events (unprecedented or force majeure) :

- **2007-2008 Financial Crisis:** do not see large impact from average intensity
- **2009 US Presidential Election (G.W. Bush ➔ Obama):** majority data centered around 2009-2019
- **2015 Paris Agreement:** Post Paris Agreement average intensity is kept between 500-600MMT/$
- **2020-2021 COVID19:** No significant impact from average intensity, from absolute CO2 emission perspective about 32.3% reduction has been observed



Methods Selections

3. Forecasting Methods Exploration

General objective is to select models given theoretical understanding of its' strengths and weaknesses, ***understand the limitation of analysis*** and ***explain what the model can support with degree of confidence*** for ClarityAI. Developing time-series model will enhance understanding of the data and the *underlying causes* associated with Scope 1 emission and *forecast with intention in dynamic (at times unprecedented) economy*.

Time-series models focus on historical data.

- PROS: Efficient for forecasts handling large number of output with stable historical input data, e.g. prediction of demand/patterns. Efficient in smoothing out small random fluctuation. Good for short-term forecasting
- CONS: Requires large amount of historical data. It does not cope with extreme variations (jumps and drips) and it does not use 'feedback loops' to dynamically upgrade model adjustment features. Inefficient for long forecast horizon.

Eight (8) Time Series Algorithms were reviewed for preliminary assessment. Recent finding highlight 'advanced time series forecasting models' which can be found in Appendix.

SPRINT 1 Methods:

Team 2's approach, to bring out creativity and benefit from diverse industry background -

1) Range of Time-Series with 2019~ data only (Ramy)
2) Exponential Smoothing with 'verified' data only (Marshall)
3) ARIMA by added Industry Categorization (production vs service industry)

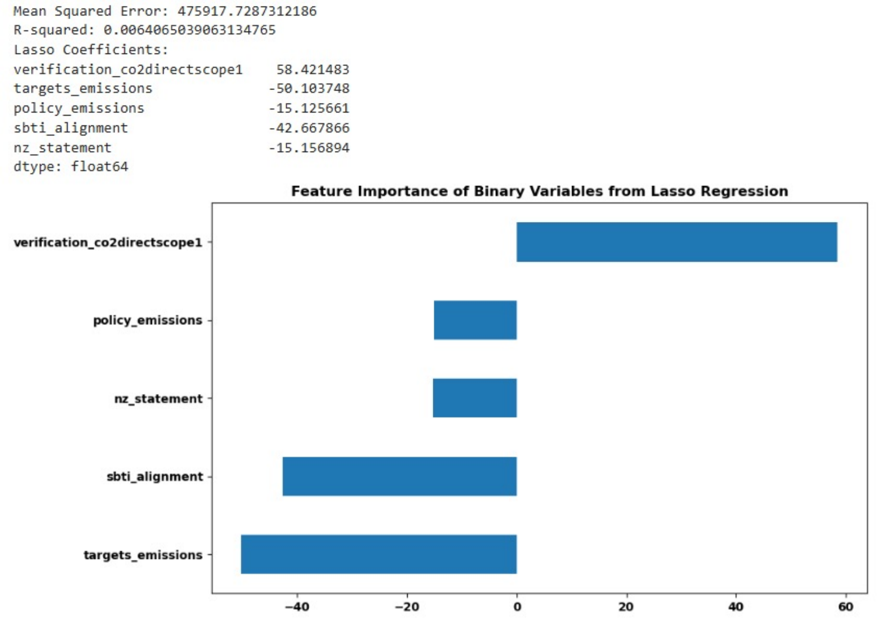
Moving forward for SPRINT 2:

Select a single method to improve overall correlation value, align on the population selection and range of forecasting, clear documentation on the tried methods, single ipynb file for submission

SPRINT 1: Lasso & Exponential Smoothing (1)

Feature Selection - LASSO

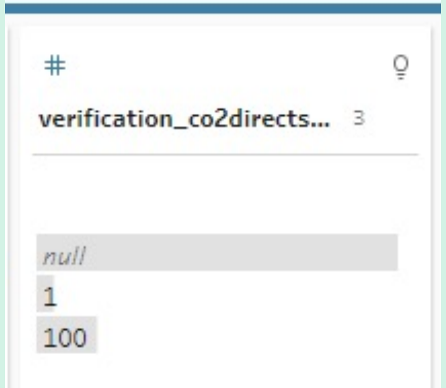
Companies that goes through rigorous verification process ‘verification_co2directscope1 = 100’ have better quality data, useful for the forecasting model.



```
Mean Squared Error: 475917.7287312186
R-squared: 0.0064065039063134765
Lasso Coefficients:
verification_co2directscope1    58.421483
targets_emissions              -50.103748
policy_emissions               -15.125661
sbti_alignment                 -42.667866
nz_statement                   -15.156894
dtype: float64
```

Forecasting – EXPONENTIAL SMOOTHING (limitations)*

Due to small (13.8%, 2500 rows) subset of data that are ‘verified’ by 3rd party the time-series forecasting algorithm, simple exponential smoothing model, does not converge.

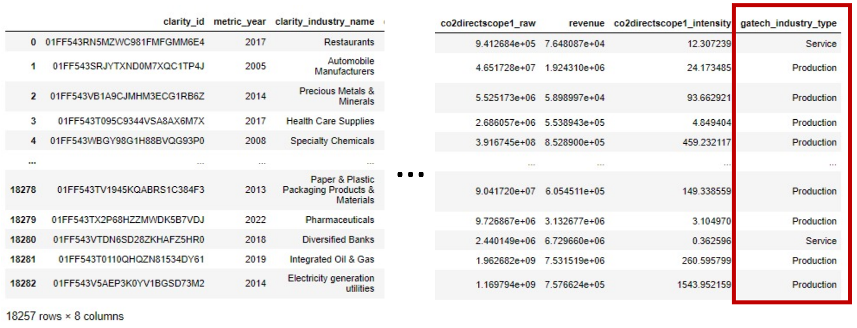


# verification_co2directs... 3

null
1
100

SPRINT 1: ARIMA by Industry Type (2)

Characteristics between a company’s revenue and Scope 1 emission by type of business operation were incorporated. 155 unique industry name, divided into 2 industry type, ‘Production’ vs ‘Service’ (column ‘gatech_industry_type’).



	clarity_id	metric_year	clarity_industry_name	…	co2directscope1_raw	revenue	co2directscope1_intensity	gatech_industry_type
0	01FF543RN5MZWC981FMFGMM6E4	2017	Restaurants		9.412684e+05	7.648087e+04	12.307239	Service
1	01FF543SRJYTXND0M7XQC1TP4J	2005	Automobile Manufacturers		4.651728e+07	1.924310e+06	24.173485	Production
2	01FF543VB1A9CJMHM3ECG1RB6Z	2014	Precious Metals & Minerals		5.525173e+06	5.898997e+04	93.662921	Production
3	01FF543T095C9344VSA8AX6M7X	2017	Health Care Supplies		2.686057e+06	5.538943e+05	4.849404	Production
4	01FF543WBGY98G1H88BVQG93P0	2008	Specialty Chemicals		3.916745e+08	8.528900e+05	459.232117	Production
...	...	...	...	•••	...	...	...	...
18278	01FF543TV1945KQABRS1C384F3	2013	Paper & Plastic Packaging Products & Materials		9.041720e+07	6.054511e+05	149.338559	Production
18279	01FF543TX2P68HZZMWDK5B7VDJ	2022	Pharmaceuticals		9.726867e+06	3.132677e+06	3.104970	Production
18280	01FF543VTDN6SD28ZKHAFZ5HR0	2018	Diversified Banks		2.440149e+06	6.729660e+06	0.362596	Service
18281	01FF543T0110QHQZN81534DY61	2019	Integrated Oil & Gas		1.962682e+09	7.531519e+06	260.595799	Production
18282	01FF543V5AEP3K0YV1BGSD73M2	2014	Electricity generation utilities		1.169794e+09	7.576624e+05	1543.952159	Production

18257 rows × 8 columns

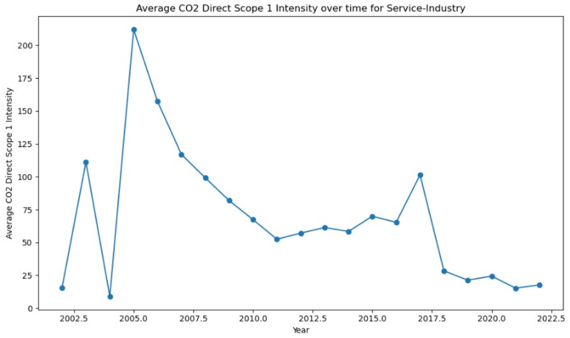
Findings:

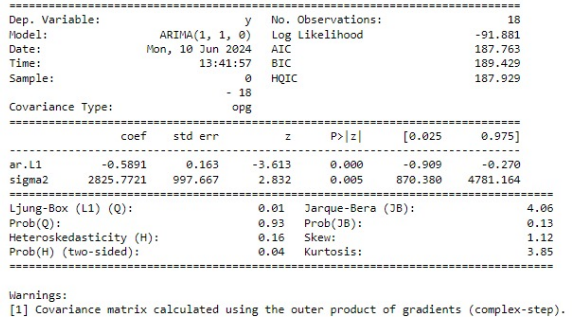
ARIMA model with energy intensives industry types tends to perform better, while the service-industry tends to have lower correlation when Moving Average regression was applied.

Check points:

Industry segmentation adequacy needs to be cross validated with S&P500’s standard definition, need to make a decision on whether we want ‘general trend’ or ‘industry specific’ prediction model for SPRINT 2. Check to see if LSTM (Long-Short Term Memory Recurrent Network will be valuable as we are hoping to forecast til 2030 (mid-range).

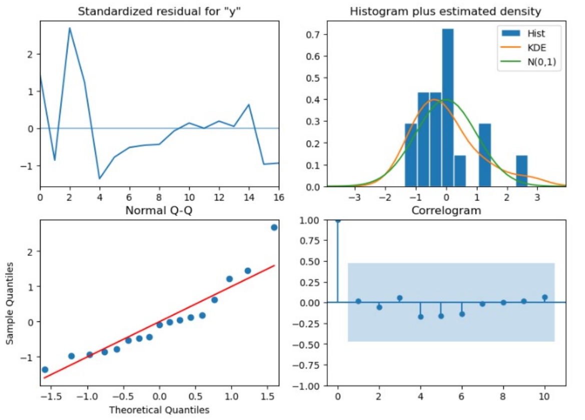
Service Industry



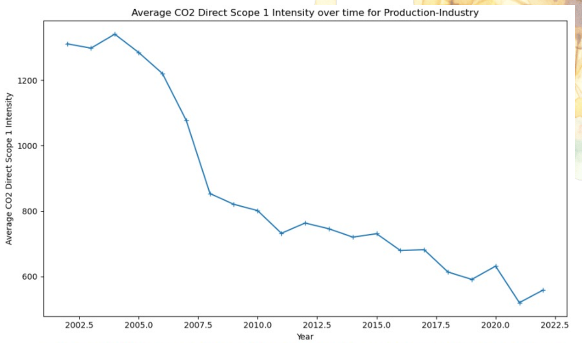


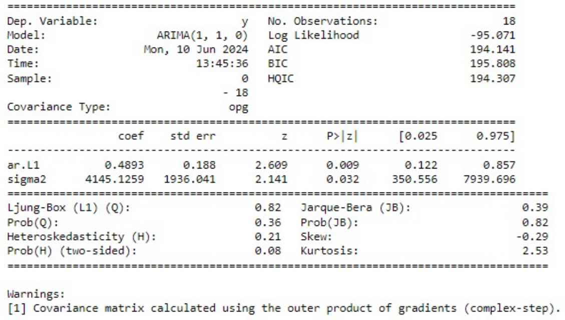
```
==============================================================================
Dep. Variable:                      y   No. Observations:                   18
Model:                 ARIMA(1, 1, 0)   Log Likelihood                 -91.881
Date:                Mon, 10 Jun 2024   AIC                            187.763
Time:                        13:41:57   BIC                            189.429
Sample:                             0   HQIC                           187.929
                                 - 18
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5891      0.163     -3.613      0.000      -0.909      -0.270
sigma2      2825.7721    997.667      2.832      0.005     870.380    4781.164
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 4.06
Prob(Q):                              0.93   Prob(JB):                         0.13
Heteroskedasticity (H):               0.16   Skew:                             1.12
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.85
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



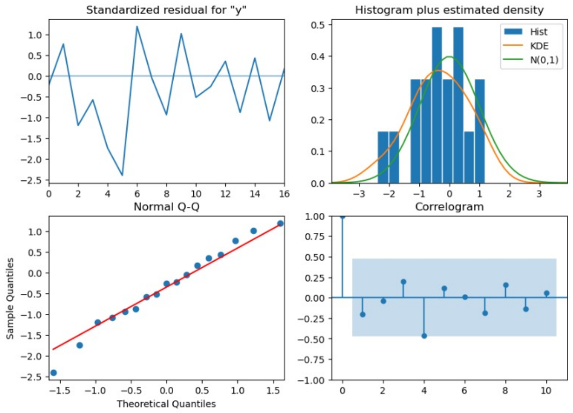
Production Industry





```
==============================================================================
Dep. Variable:                      y   No. Observations:                   18
Model:                 ARIMA(1, 1, 0)   Log Likelihood                 -95.071
Date:                Mon, 10 Jun 2024   AIC                            194.141
Time:                        13:45:36   BIC                            195.808
Sample:                             0   HQIC                           194.307
                                 - 18
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4893      0.188      2.609      0.009       0.122       0.857
sigma2      4145.1259   1936.041      2.141      0.032     350.556    7939.696
===================================================================================
Ljung-Box (L1) (Q):                   0.82   Jarque-Bera (JB):                 0.39
Prob(Q):                              0.36   Prob(JB):                         0.82
Heteroskedasticity (H):               0.21   Skew:                            -0.29
Prob(H) (two-sided):                  0.08   Kurtosis:                         2.53
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



NEXT STEP

- *Sprint 1 was dedicated session to bring* out creativity of each team member, leveraging their diverse industry experience and professional experience in approaching scope 1 emission data.
- *Moving forward for SPRINT 2:*
  - *Set SPRINT 1's best approach as a baseline for evaluation*
  - *Select a single method to improve overall correlation value*
  - *Align on the population selection and range of forecasting*
  - *Clear documentation on the tried methods, to share learnings from Practicum to ClarityAI*
  - *Single ipynb file for submission and structured repository*

Q&A

Constructive Feedback from ClarityAI

Appendix

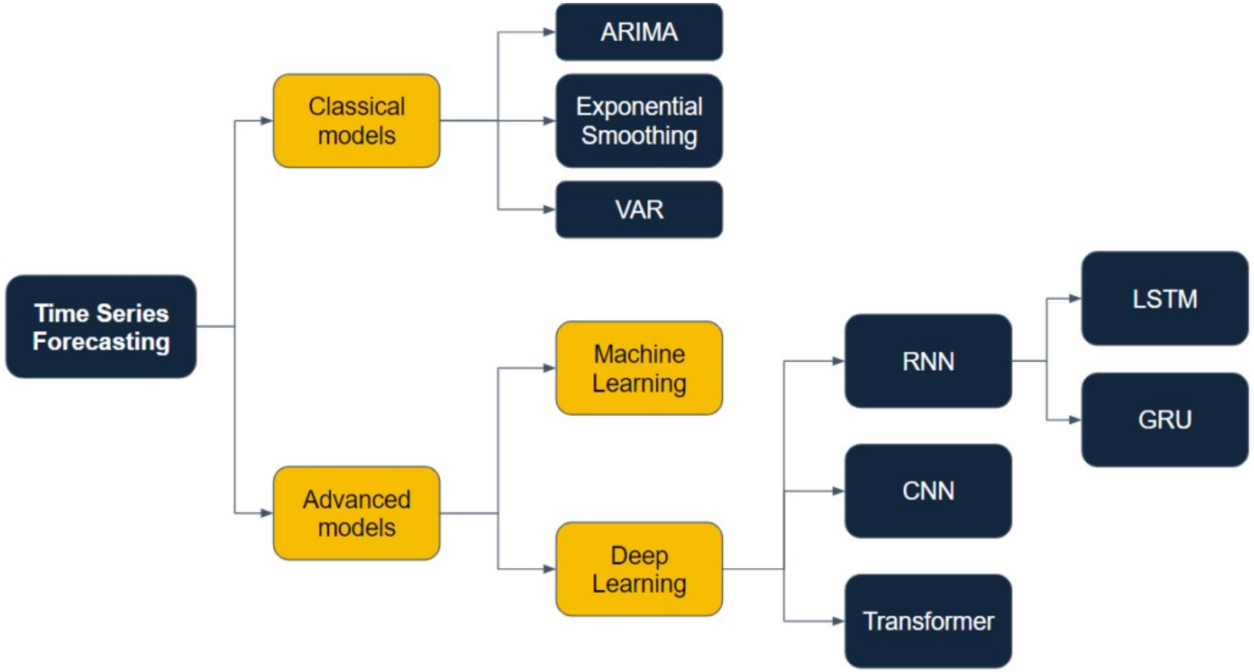
Advanced Timed Series Forecasting Method
Procs and Cons of Time Series algorithm considered for Sprint 1
Feature Selection

Advanced Timed Series Forecasting Method

D. Andres, 2023, MLPills.dev

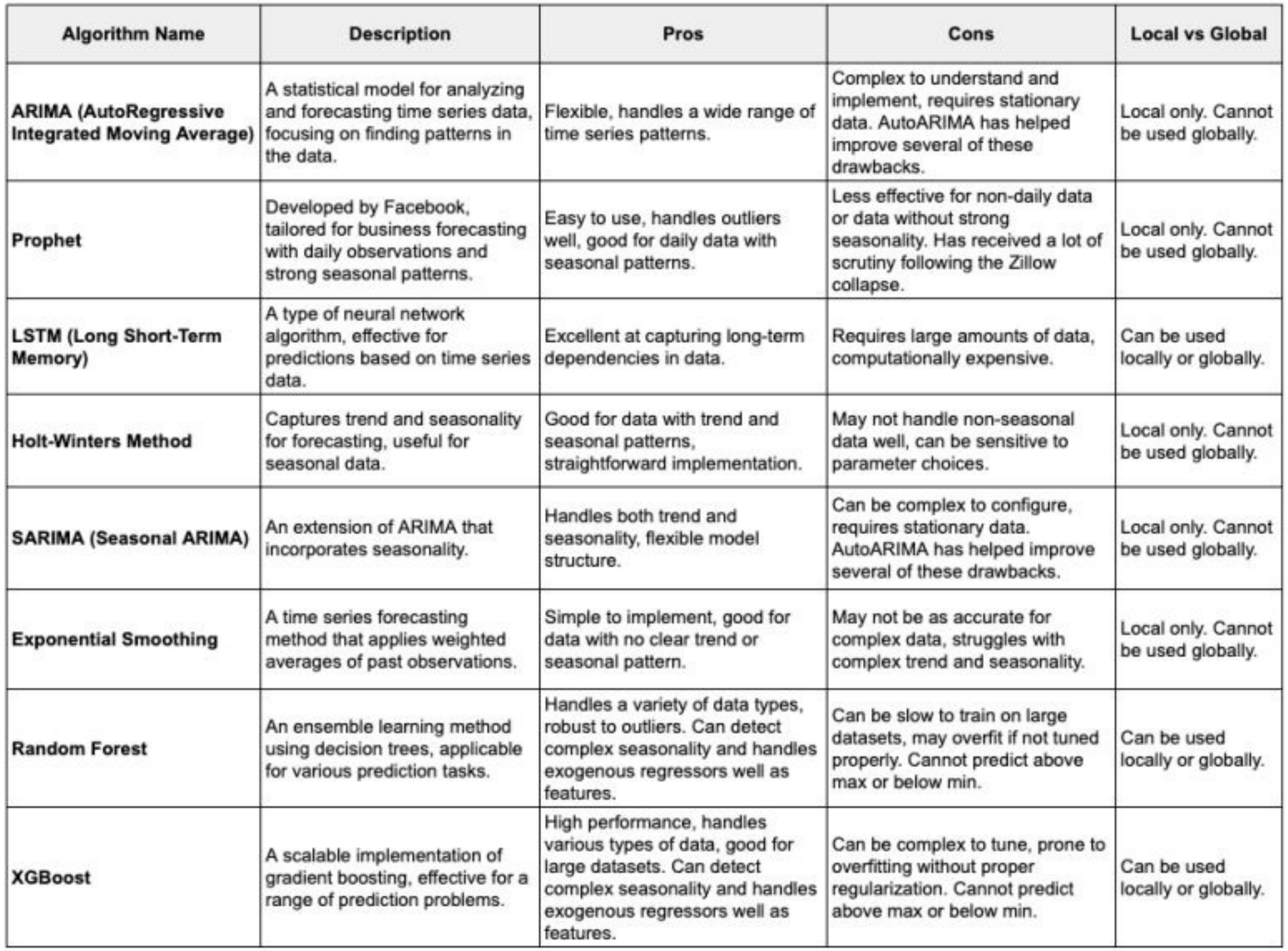
There are mixed view regarding the accuracy of the approaches that some say these ‘advanced techniques underperform classical ones’.

However, a useful *guideline* is to *consider classical techniques when dealing with data over a short time period* and *opt for machine learning or deep learning approaches when dealing with data over a long period*.



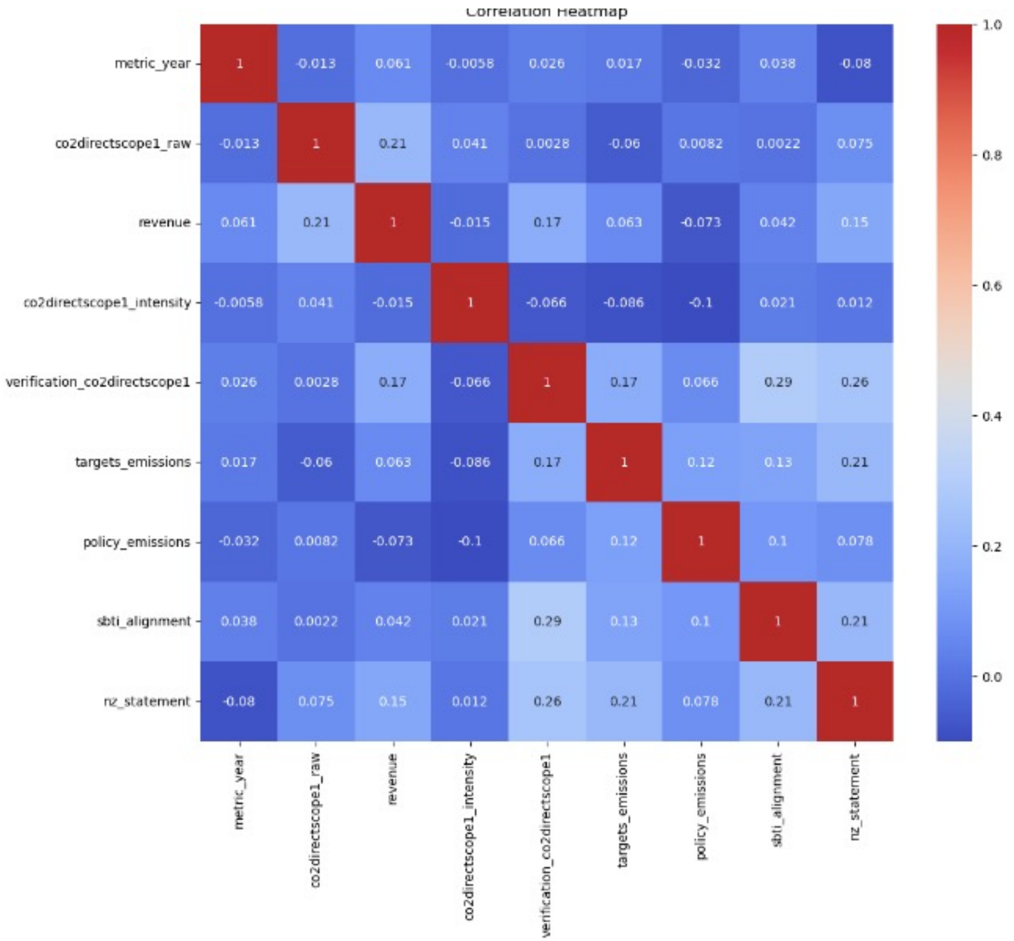
Eight (8) Time-Series Algorithms

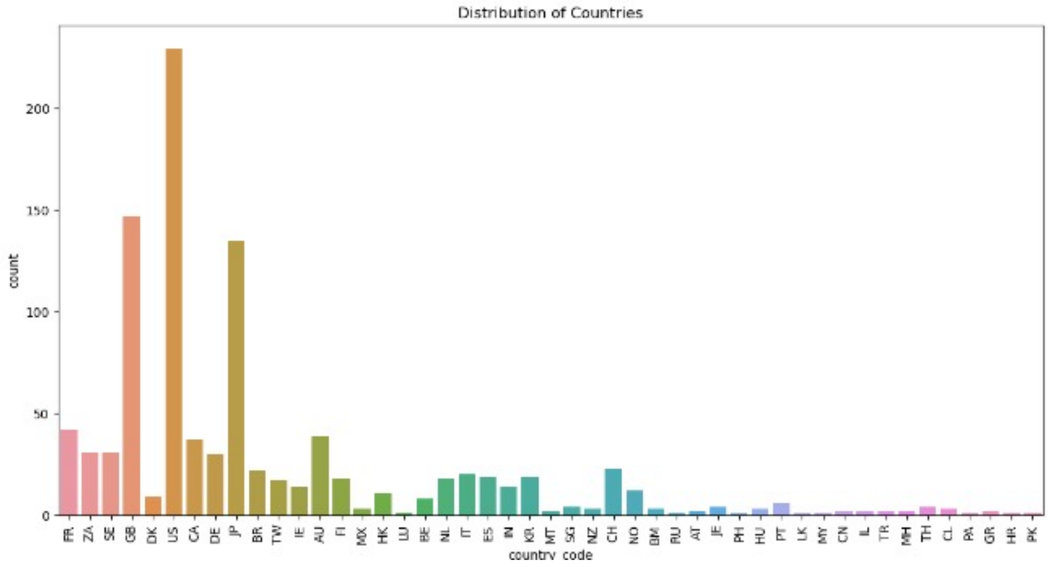
M.Dancho, 2024, LinkedinPost



Algorithm Name	Description	Pros	Cons	Local vs Global
ARIMA (AutoRegressive Integrated Moving Average)	A statistical model for analyzing and forecasting time series data, focusing on finding patterns in the data.	Flexible, handles a wide range of time series patterns.	Complex to understand and implement, requires stationary data. AutoARIMA has helped improve several of these drawbacks.	Local only. Cannot be used globally.
Prophet	Developed by Facebook, tailored for business forecasting with daily observations and strong seasonal patterns.	Easy to use, handles outliers well, good for daily data with seasonal patterns.	Less effective for non-daily data or data without strong seasonality. Has received a lot of scrutiny following the Zillow collapse.	Local only. Cannot be used globally.
LSTM (Long Short-Term Memory)	A type of neural network algorithm, effective for predictions based on time series data.	Excellent at capturing long-term dependencies in data.	Requires large amounts of data, computationally expensive.	Can be used locally or globally.
Holt-Winters Method	Captures trend and seasonality for forecasting, useful for seasonal data.	Good for data with trend and seasonal patterns, straightforward implementation.	May not handle non-seasonal data well, can be sensitive to parameter choices.	Local only. Cannot be used globally.
SARIMA (Seasonal ARIMA)	An extension of ARIMA that incorporates seasonality.	Handles both trend and seasonality, flexible model structure.	Can be complex to configure, requires stationary data. AutoARIMA has helped improve several of these drawbacks.	Local only. Cannot be used globally.
Exponential Smoothing	A time series forecasting method that applies weighted averages of past observations.	Simple to implement, good for data with no clear trend or seasonal pattern.	May not be as accurate for complex data, struggles with complex trend and seasonality.	Local only. Cannot be used globally.
Random Forest	An ensemble learning method using decision trees, applicable for various prediction tasks.	Handles a variety of data types, robust to outliers. Can detect complex seasonality and handles exogenous regressors well as features.	Can be slow to train on large datasets, may overfit if not tuned properly. Cannot predict above max or below min.	Can be used locally or globally.
XGBoost	A scalable implementation of gradient boosting, effective for a range of prediction problems.	High performance, handles various types of data, good for large datasets. Can detect complex seasonality and handles exogenous regressors well as features.	Can be complex to tune, prone to overfitting without proper regularization. Cannot predict above max or below min.	Can be used locally or globally.

Feature Assessment


Correlation Heatmap
metric_year
co2directscope1_raw
revenue
co2directscope1_intensity
verification_co2directscope1
targets_emissions
policy_emissions
sbti_alignment
nz_statement


Distribution of Countries
count
FR ZA SE GB DK US CA DE JP BR TW IE AU FI MX HK LU BE NL IT ES IN KR MT SG NZ CH NO BM RU AT JE PH HU PT LK MY CN IL TR MH TH CL PA GR HR PK
country code