INFOSYS 722 - Assignment 4 Scientific Report

Yunke Sun

Github repository: <https://github.com/ysun948/INFOSYS722-assignment4.git>

October 10, 2023

# Contents

1. [**Business Understanding**](#_bookmark0) **2**
   1. [Situation Objectives](#_bookmark1) 3
   2. [Assessment of the situation](#_bookmark2) 4
      1. [Resource inventory](#_bookmark3) 4
      2. [Requirements, Assumptions and Constraints](#_bookmark4) 4
      3. [Risks and contingencies](#_bookmark5) 5
   3. [Data mining goals](#_bookmark6) 5
   4. [Project plan](#_bookmark7) 5
2. [**Data Understanding**](#_bookmark8) **5**
   1. [Data collection](#_bookmark9) 5
   2. [Initial Data Screening and Description](#_bookmark11) 7
   3. [Data Exploration](#_bookmark13) 9
   4. [Data quality inspection](#_bookmark17) 12
3. [**Data Preparation**](#_bookmark19) **13**
   1. [Data Selection](#_bookmark20) 13
   2. [Data formatting](#_bookmark21) 13
   3. [New features derivation](#_bookmark23) 14
   4. [Data integration](#_bookmark24) 14
4. [**Data transformation**](#_bookmark26) **15**
   1. [Data reduction](#_bookmark27) 15
   2. [Data projection](#_bookmark29) 16

### [Data-Ming Method Selection](#_bookmark30) 16

* 1. [Discussion of Data Mining Methods in Context of Data Min-](#_bookmark31)

[ing Objectives](#_bookmark31) 16

* 1. [Data types accessible for data mining](#_bookmark34) 18
  2. [Data mining objectives](#_bookmark35) 18
  3. [Modelling Requirements,Assumptions and Criteria](#_bookmark36) 18
     1. [Prediction Methods](#_bookmark37) 18
     2. [Description Methods](#_bookmark38) 19
  4. [Methods Selection](#_bookmark39) 20

1. [Data-Mining Algorithms Selection](#_bookmark40) 20
   1. [Possible data mining algorithms](#_bookmark41) 20
   2. [Selection of algorithms in context of data mining objectives](#_bookmark42) . 20
      1. [For predicting CO2 emissions](#_bookmark43) 20
      2. [For recognizing patterns and trends](#_bookmark44) 21
   3. [Build models with algorithms and parameters](#_bookmark45) 21
      1. [Prediction models](#_bookmark46) 21
      2. [Time-series models](#_bookmark48) 22
2. [Data Mining](#_bookmark49) 23
   1. [Logical tests](#_bookmark50) 23
   2. [Output of models](#_bookmark51) 23
   3. [Search for patterns](#_bookmark52) 23
3. [Interpretation](#_bookmark54) 24
   1. [Study and discuss mined patterns](#_bookmark55) 24
   2. [Data visualization](#_bookmark58) 26
   3. [Assessing and Evaluating Models](#_bookmark60) 27
   4. [Multiple iteration](#_bookmark61) 27
4. [Final discussion](#_bookmark62) 28

[References](#_bookmark64) 29

# Business Understanding

CO2 is a major greenhouse gas responsible for trapping heat in the Earth’s atmosphere. Elevated CO2 levels contribute to global warming and climate change. High levels of CO2 emissions have direct and indirect consequences

on the environment. The United Nations has established the 17 Sustain- able Development Goals (SDGs) as a comprehensive framework to address various pressing global issues but not limited to health, education, gender equality, clean water, and climate action(Della [Santa Navarrete, Borini, &](#_bookmark66) [Avrichir, 2020).](#_bookmark66) Among the 17 SDGs, healthcare improvement, promoting sustainable industry, mitigating climate change, and conserving marine re- sources are related to CO2 emission. Understanding current emission levels throughout the world is crucial for achieving international agreements and international agreements.

## Situation Objectives

CO2 emission remains a significant issue throughout the world. One of the striking issues related to CO2 emission is the global inequality in CO2 emission. According to Hannah Ritchie’s research, regarding per capita CO2 emissions, high-income countries emit more than 30 times as much as those in low-income countries. And geographically, North America shares 16.5 percent of global CO2 emissions but only 7.5 percent of global population, while in contrast, Africa shares 3.9 percent of global CO2 emissions with only

* 1. percent of global population(Ritc[hie & Roser, 2023).](#_bookmark73) By analyzing the comprehensive CO2 emission data from multiple countries with timestamp, researchers can analyze historical data to identify trends in CO2 emissions over time. By comparing emissions data across countries and regions, re- searchers can find variations and disparities, NGOs like Union Nation may make use of this information to collaborate internationally, share best prac- tices, and implement effective policies. Meanwhile, this is also useful for climate adaptation strategies planning. If we can build a robust CO2 emis- sion prediction model based on date, region and industry, policymakers can use it to plan for climate adaptation and mitigation strategies.

Hence, the study will be conducted with the following objectives:

* + - To to provide data-driven insights and visual representations that can effectively raise public awareness and encourage active participation in efforts to reduce CO2 emissions.
    - To explore prominent trends and patterns that could have implications for the achievement of the Sustainable Development Goals (SDGs).
    - To conduct comparative evaluations of carbon dioxide (CO2) emis- sions across different countries and areas, and benchmark top players

with global norms and goals. This comparative analysis will elucidate optimal methodologies and identify regions necessitating more measures.

Regarding the success criterion, we would consider this study as a success if the following criterion are met:

* + - Identify the regions and countries with the highest CO2 emissions over time, help to solve global CO2 emissions inequality.
    - Identify trends and patterns regarding CO2 emissions.
    - Increase public awareness of controlling CO2 emissions.

## Assessment of the situation

To assess the situation, an examination of the current state of CO2 emissions is conducted, cross-referencing with individual SDGs, considering how each country’s efforts contribute to or hinder the global aims.

### Resource inventory

And the tools we leveraged such as Python, SPSS Modeler, and Tableau are free of use through the university analytic platforms. The data set we will use

is from h[ttps://www.kaggle.com/datasets/saloni1712/co2-emissions?resource=do](http://www.kaggle.com/datasets/saloni1712/co2-emissions?resource=download)wnload.

### Requirements, Assumptions and Constraints

Regarding the requirements, for data completeness, according to Kaggle

, this data set should fully complete and has no missing values, for data accuracy, this data set should be accurate and reliable, for data consistency, different columns or sources should not contradict each other.

This data is assumed to be not normally distributed, considering the global CO2 emissions inequalities, we expect a large extent of variation among CO2 emissions of different industries and regions. Moreover, the time series data is assumed to be not independent.

Regarding the constraints, for one thing, the time slots in this data set is limited, which means it is difficult for us to build a reliable CO2 emission prediction model with time series analysis. For another, the feature columns contain only sectors, countries, and timestamp, thus our options for feature engineering and feature selection are limited.

### Risks and contingencies

For potential risks, due to the fact of CO2 emission inequality, we may have outliers that could impact model performance. The contingency plan is to carefully pre-process the data and perform data imputation or cleansing if necessary. Another issue is that the lack of features may cause over-fitting, for contingency plan we should use cross-validation to assess model general- ization and optimize the parameters or consider ensemble multiple models. Last but not least, for large scale data mining, computational resource may not be sufficient, we could use cloud computing to solve this issue.

## Data mining goals

The data mining goal of this project are listed as below.

* Identify the trend and pattern of the distribution of CO2 emissions around the world, find out countries and sectors contributing the most to the emissions.
* Build CO2 emissions prediction models to predict the CO2 emissions for certain regions, sectors and time.
* In the future, dig deeper into this topic with multiple iteration, trying different data mining methods.

## Project plan

We plan to carry out this project in 20 days. The planning is listed as below. The Gantt chart can be seen in figure [1.](#_bookmark10)

# Data Understanding

## Data collection

As aforementioned, the data set was downloaded from Kaggle [(JHALANI,](#_bookmark68) [2023).](#_bookmark68) And the URL is h[ttps://www.kaggle.com](http://www.kaggle.com/datasets/saloni1712/co2-)/datas[ets/saloni1712/co2-](http://www.kaggle.com/datasets/saloni1712/co2-) emissions?resource=download. This data set provides a good example of carbon emissions exploration for combating climate change. Regarding the data collection, we can figure out from the web page that the data set is from https://carbonmonitor.org/, which provides official data release with countries and sectors CO2 emissions changes up to August 31 2023. We do

|  |  |  |
| --- | --- | --- |
| Phase | Time | Risks |
| Kick off | 2  days | Fail to load the data set  ans set up pySpark |
| Data Un-  derstanding | 2  days | Overwhelming size of  data set, no clear expla- nation on columns |
| Collection | 2  days | Handling large volumes  of data can be challeng- ing in terms of storage, processing |
| Initial Data  Screen | 2  days | Some of the data vari-  ables or attributes are irrelevant to the analy- sis |
| Data Ex-  ploration | 2  days | Outliers and null values  might have negative im- pact |
| Data Min-  ing | 2  days | Difficult to find out the  suitable metrics to mea- sure the performance of algorithms |
| Multiple  iterations and Report | 5  days | Repeatedly refining  models on the same dataset can lead to overfitting. |

Table 1: Project plan

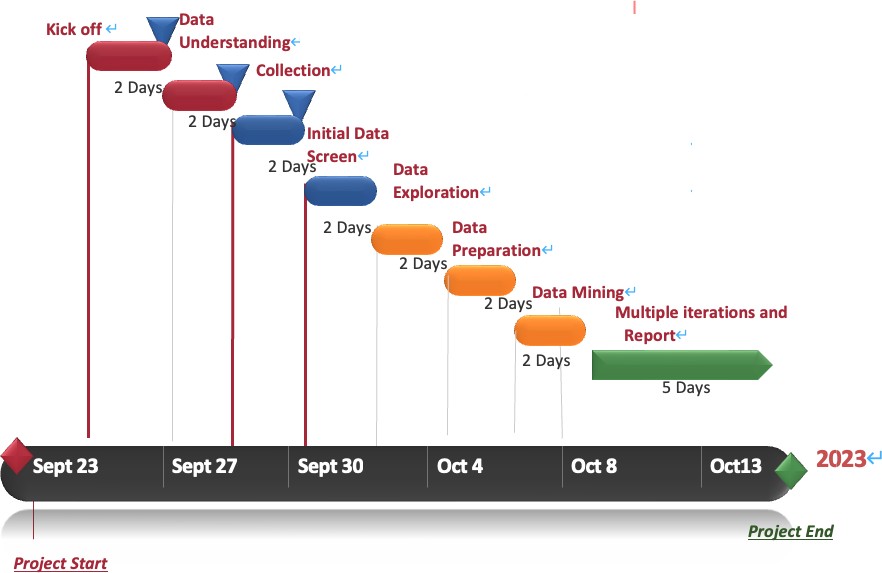


Figure 1: Project plan

encounter an issue since we do not know how the data set was collected and which methodology was applied, that is not transparent.

## Initial Data Screening and Description

Anaconda platform was employed to apply Python packages and pyspark. The data set was imported into Jupyter and read through spark.read(). The result is shown below in [Figure2.](#_bookmark12) It reveals that the data set has five columns such as country, date, sector, value and timestamp. It also shows that each line records the power-related record at different date and time points in various countries or regions.

According to printSchema() function, there are 135408 rows, and the data types of each column are summarized in [Figure3.](#_bookmark14) And the contents in Figure 2 and Figure 3 unveils the data set as a preview are summarized in the table below.

|  |  |
| --- | --- |
| Column | Description |
| Country | Text, the geographical location or  territory that is being referred to in the given record. This en- compasses specific nations such as Brazil, China, France, as well as col- lective areas such as EU27 and UK, and a worldwide representation re- ferred to as WORLD. Also, the rest of the world shows as ROW in the data set |
| Date | Date, the date of the data collected |
| Sector | Text, the preview shows the data of  “Power” sector |
| Value | Float, a numeric value of CO2 con-  sumption of each country at specific date and time points |
| Timestamp | Integer, the original date, and time  of the data collected in UNIX times- tamp format. |

Table 2: Data types of each column

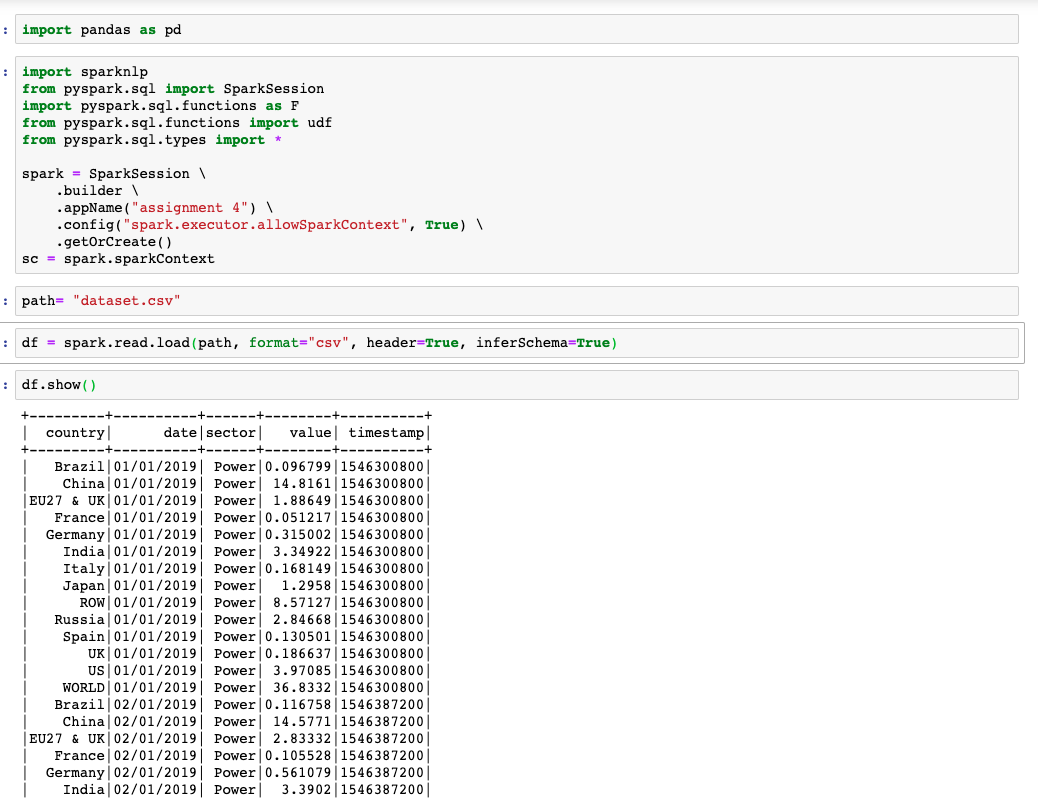


Figure 2: Initial Data Screening

## Data Exploration

After downloading the data set and initial data screening through Python, a preliminary analysis was conducted to understand the data. The initial timestamp was captured in the UNIX timestamp format, and subsequently, a datetime column was generated by converting the UNIX timestamp. This process resulted in the extraction of time, which is depicted as a new column in Figure [4.](#_bookmark16) The results also indicates that the data set records 135,408 records with 7 columns. The process of conversion enhances the legibility of the timestamp data.

Further, we convert the pyspark dataframe to pandas dataframe, the non-numeric columns are summarized in Python Pandas as shown in Figure

[5.](#_bookmark18) The dataset comprises a total of 14 countries, spanning a duration of 1,612 days, and encompassing 6 distinct sectors. However, it is important to note that all time points within the dataset are identical. The data contains 127,135 unique values in Column Value with 2.36 average, and 5.91

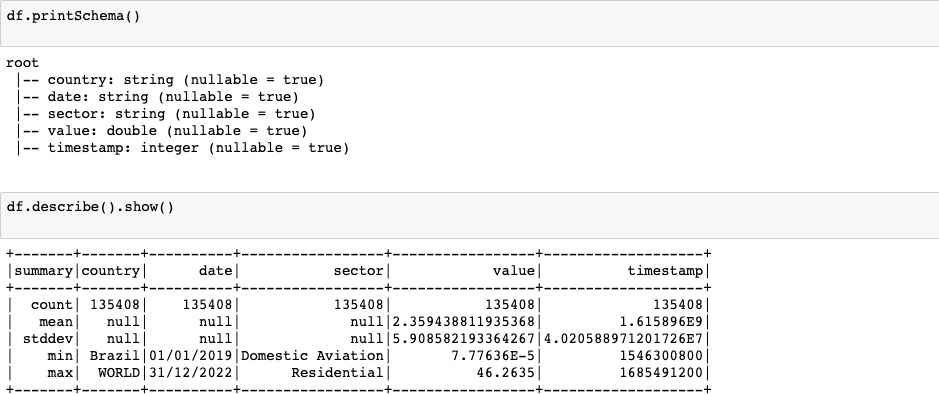


Figure 3: Data types of each column

standard deviation. No missing value is contained in the data set.

The summary of data exploration can be consolidated as follows in Ta- [ble3.](#_bookmark15)

Similar as previous analysis in SPSS, [Figure6](#_bookmark22) shows that the latest status of each country in Tableau. The line of the world indicates that the world trend has higher but more fluctuate CO2 consumption value compared to other individual countries, as it is a lump-sum value. Apparently, the line of China keeps the same pace of the line of the World. It has higher value of power consumption compared to other countries, followed by the rest of the world (ROW). Below the green line of ROW, it is the trend line of the United States. Other countries are below the lines of China, ROW and the United States, crowded in the bottom, we can analyze further in the later steps.

Based on the heatmap of Tableau, China has highest CO2 consumption value on average shown in [Figure7](#_bookmark25) .

|  |  |  |
| --- | --- | --- |
| Column | Data type | Exploration Summary |
| Country | Categorical | It consists of 14 unique data  values with 135,408 valid records. |
| Date | Continuous | The date ranges from January  1, 2019 to May 31, 2023 with  135,408 valid records. |
| Sector | Categorical | 6 unique values with 135,408  valid records. |
| Value | Continuous | It ranges from 0 to 46.264,  with 135,408 valid records. The mean of the column Value is 2.359. Standard Deviation is 5.909, and Skewness is 4.195. |
| Timestamp | Continuous | It ranges from 1546300800  to 1685491200, with mean and standard deviation are 1615896000, and  40205889.712 respectively.  Skewness is zero. |

Table 3: Summary of data exploration

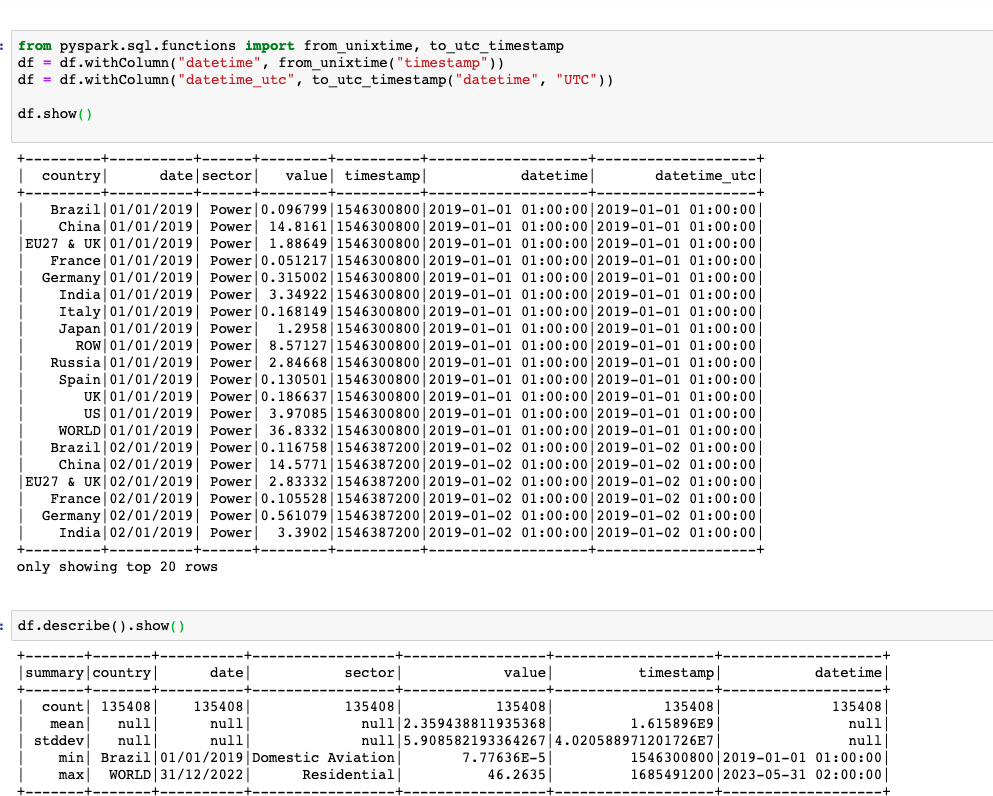


Figure 4: Data Exploration

We can also plot the sum of CO2 emissions by region and sector as can be seen in figure [8](#_bookmark28) and figure [9.](#_bookmark32)

## Data quality inspection

As can be seen in figure [10,](#_bookmark33) the completeness results is 100 percent for each column, which indicates that there is no missing data in each column. This is aligned with the introduction of the data set on Kaggle website.

In this section, we focus on outlier inspection. We use the z-value method to detect outlier s from the data, and in the actual data cleaning process,. The z-score is a well-known statistical measure that quantifies how far a particular data point is from the mean of a distribution in terms of standard deviations. It is calculated using:

*z* = *x − µ*

*σ*

In the formula, x is the data point for which you want to calculate the z-score, *µ* is the mean of the distribution and *σ* is the standard deviation of the distribution. For this part we group the data based on country and sector, for each combination of country and sector we calculate the mean and

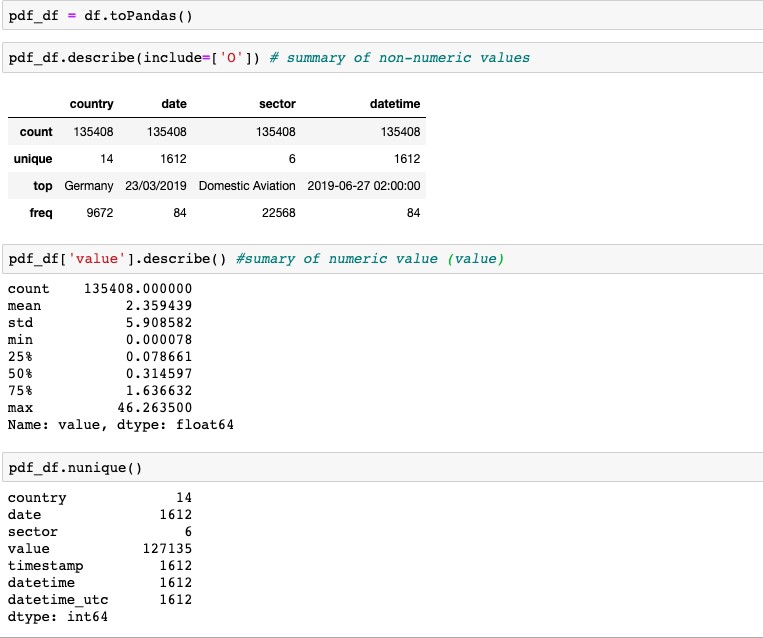


Figure 5: Data Exploration

standard deviation, if z-score is higher than 3 then we consider the current data point is an outlier. Finally we get 725 outliers for 135408 rows, this indicates the issue of outliers is not serious.

# Data Preparation

## Data Selection

Based on the discussion above, we found no missing values and 725 outliers in the data set. We decide to remove the outliers from the data set and use the remaining part of the data. We then find out that the data type of column date is string, which has no use for further data mining, so we decide to format the timestamp column and remove the date column.

## Data formatting

One of the initial challenges encountered was the management of Unix times- tamps. The timestamps mentioned in this context are represented as integer values, denoting the number of seconds that have passed since the initial date of the UNIX system, which is January 1, 1970. To facilitate the compre-

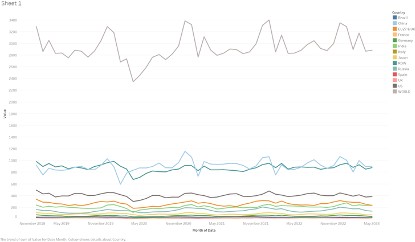


Figure 6: CO2 emissions world trend

hension of these Unix timestamps, a derivation procedure was utilized. The procedure involved the inclusion of a time offset of 2,208,988,800 seconds to every timestamp. The transformation was of paramount importance in facil- itating a more comprehensive and temporally sensitive analysis of our data set, enabling us to align our data manipulation with the goals of our study.

## New features derivation

As of now, there are only two categorical variables, country and sector in the data set. For further analysis, we create dummy values for country and sector. Based on the feature date, we can further extract Year, Quarter, and Month as new features. Furthermore, the date information was disassem- bled into discrete properties for year, quarter, and month, and subsequently encoded as integer values by creating dummies.

## Data integration

In the previous step we conducted one-hot coding to both categorical vari- ables and datetime variables, we merge all the derived features into a dataframe containing all the original and generated features except datetime.



Figure 7: CO2 emissions heatmap

# Data transformation

## Data reduction

Following the implementation of the newly introduced features, there has been a significant augmentation in the number of columns within the data set. We build a correlation matrix, and based on the result we can remove redundant or insignificant characteristics. To illustrate, we measure the cor- relation coefficient of each feature with the dependent variable and set a threshold of 0.1, which means we drop features whose correlation coefficient with value are lower than 0.1. The result is shown in Figure**??**. Finally, we keep the following features: ’country WORLD’, ’sector Power’, ’sec- tor Domestic Aviation’, ’sector International Aviation’, ’sector Industry’, ’country China’, ’country ROW’, ’country Spain’, ’country France’, ’coun- try Italy’, ’country UK’, and ’country Brazil’.

By employing feature selection techniques, we have not only enhanced the effectiveness of our modeling procedure, but also reduced the risk of over-fitting.

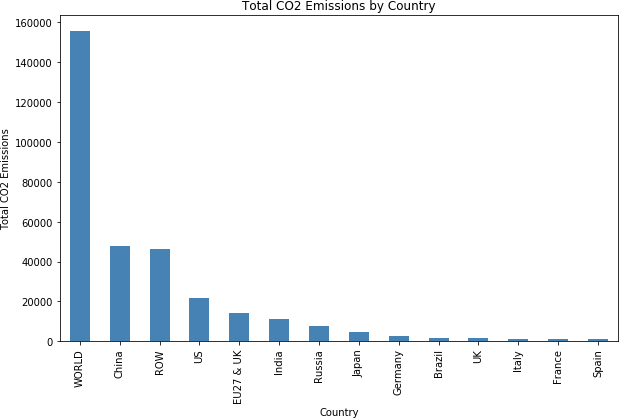


Figure 8: CO2 emissions by country

## Data projection

Since we are dealing with one-hot encoded features, It is possible to multiply different one-hot encoded features or perform other mathematical operations to create new interactive features. This can help the model capture relation- ships between features. We figured out that top two contributors of CO2 emissions are China and ROW, and by sectors the top two are power and industry, we can thus separately multiply the two country-related features with the two sector-related features to create four new features.

# Data-Ming Method Selection

## Discussion of Data Mining Methods in Context of Data Mining Objectives

For possible data mining methods, we may choose among classification, clus- tering , description and prediction.

Classification is a data mining technique that entails the allocation of items within a specified collection to specific target groups or classes.It is process of generalizing the data according to different instances(Gama [&](#_bookmark67)

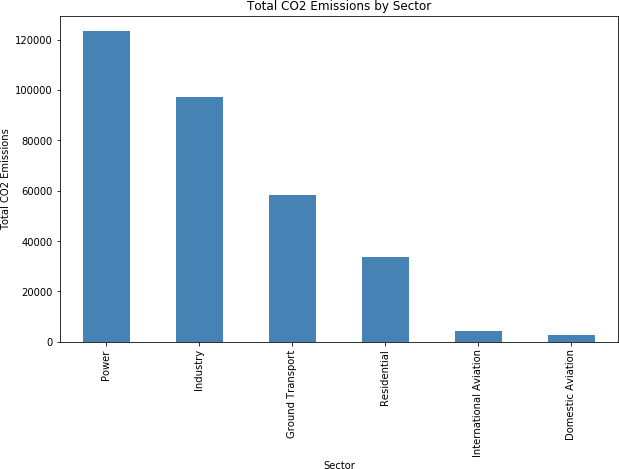


Figure 9: CO2 emissions by sector

[Brazdil, 1995)](#_bookmark67) .The primary goal is to accurately predict the specific target class for each instance within the data set. This diverges from our primary objective, which entails forecasting the future power consumption value.

Clustering is a data mining method used for grouping similar data points or objects into clusters or categories based on their inherent similarities(Verma, [Srivastava, Chack, Diswar, & Gupta, 2012).](#_bookmark76) The primary goal of clustering is to discover hidden patterns, structures, or natural groupings within a data set without any prior knowledge of the groups.

Description refers to the process of summarizing and understanding the characteristics and patterns within a data set without necessarily making predictions or building models. Descriptive analysis is a fundamental step in data mining that focuses on providing insights and a comprehensive view of the data.

Prediction involves using historical data to make informed forecasts or predictions about future events or outcomes. This system possesses the abil- ity to integrate both categorization and estimation (Kaur, [Singh, & Josan,](#_bookmark69) [2015).](#_bookmark69)

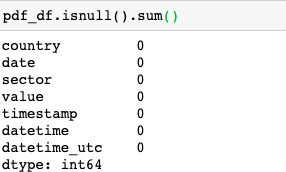


Figure 10: Missing values

## Data types accessible for data mining

For this project, our features are mostly one-hot coded categorical features, thus we can not apply estimation which refers to to the examination of outcomes characterized by continuous values. For categorical variables we are not able to apply clustering algorithms. Since the dependent variable, the CO2 emissions is a continuous variable, we prefer to apply prediction rather than classification. For exploratory analysis, we can apply description to figure out the pattern within the data set.

## Data mining objectives

The most important data mining objective of our project is to build CO2 emissions prediction models to predict the CO2 emissions for certain regions, sectors and time, for this purpose we straightforwardly apply prediction. An- other goal is to figure out the trends and patterns by time-analysis, for this purpose when can apply description for understanding the characteristics of data set.

## Modelling Requirements,Assumptions and Criteria

### Prediction Methods

Prediction methods requires that the input data is accurate, complete, and representative of the problem domain, which is met in our project. The



Figure 11: Outliers

assumption is that the data meets quality standards and the criteria is that data processing steps should be applied a needed. For features, prediction models require reasonable feature engineering and assume that feature engi- neering methods are properly chosen, the criteria are that features with high significance should be kept. Based on the discussion above we meet this requirement.

### Description Methods

For data exploration, we are required to explore and summarize the data to provide a comprehensive description, we assume meaningful patterns can be recognized. The criteria is that the patterns should be interpretable. For data visualization, we require and assume it clearly convey information which could be implicit in the data set, the criteria is that visualizations should be clear and informative. Based on the figures below we are able to make suitable visualizations.

## Methods Selection

Based on the discussions above, we will choose prediction methods and de- scription methods as data mining methods. For concrete algorithms we will discuss later.

# Data-Mining Algorithms Selection

## Possible data mining algorithms

In this section, we will be discussing possible data mining algorithms for the project. As we have mentioned, we have one-coded categorical features which are not linearly correlated, and considering the dimension of features is 12, a relatively high dimension, we should choose algorithms good with han- dling high-dimensional non-linearly correlated data. The possible solutions are Ridge Regression, Decision Tree Regression, Random Forest Regression and Support Vector Regression.

For recognition of CO2 emissions patterns, besides data visualizations and statistical analysis, we can carry out time-series analysis. The com- monly used algorithms include Autoregressive Integrated Moving Average (ARIMA) and Seasonal Decomposition of Time Series (STL).

## Selection of algorithms in context of data mining objec- tives

### For predicting CO2 emissions

Ridge regression is a regularized version of linear regression, meaning it’s designed to handle multicollinearity and prevent over-fitting. Ridge regres- sion adds a penalty term to the linear regression equation, known as L2 [regularization(McDonald, 2009).](#_bookmark71) By adding the L2 penalty term, it prompts the coefficients of correlated features to be spread out, preventing any single feature from dominating the model.

Decision Tree Regression builds a model in a tree structure to make pre- dictions on continuous variables. In the trees each internal node represents a decision or split based on one of the input features, and each leaf node contains the prediction value(Rathore [& Kumar, 2016).](#_bookmark72) It can capture non- linear relationships between features and the target variable. In our project decision trees can model complex interactions among these binary features. Random Forest is the aggregation of decision trees. To illustrate, it is an ensemble algorithm that combines the predictions of many decision trees for

more accurate results. Each tree in the forest is built on a random subset of the data and features, adding randomness and diversity to the mo[del(Liu,](#_bookmark70) [Wang, & Zhang, 2012).](#_bookmark70)

Support Vector Regression is used to find a decision boundary that best fits the given data points while minimizing the marginal error. The advan- tage of SVR in our case is that it can handle high-dimensional data efficiently and effectively (Smola [& Sch¨olkopf](#_bookmark75), [2004).](#_bookmark75)

As we have discussed earlier, the multicollinearity is not a main concern in our project, thus Ridge Regression has little effect on improving the per- formance. For the time being, for prediction models we choose Decision Tree Regression, Random Forest Regression and Support Vector Regression.

### For recognizing patterns and trends

ARIMA models are employed to capture and understand the underlying patterns and trends within time-series data(Sh[umway, Stoffer, Shumway, &](#_bookmark74) [Stoffer,](#_bookmark74) [2017).](#_bookmark74) ARIMA contains three main components: AutoRegressive (AR) Component which accounts for the relationship between the current observation and previous observations in the time series; Integrated (I) Com- ponent represents the differences of the time series data to make it station- ary, which means removing trends and making the series mean and variance constant over time; Moving Average (MA) Component which stands for re- lationship between the current observation and past white noise (random) error terms.

STL stands for ”Seasonal-Trend decomposition using LOESS,” it is used to decompose a time series into three components: Seasonal, Trend, and Residual(Clev[eland, Cleveland, McRae, & Terpenning, 1990).](#_bookmark65) The seasonal trend decomposition is useful for acquiring seasonal fluctuations in our case. For this part, we apply both ARIMA model and STL model.

## Build models with algorithms and parameters

### Prediction models

From pyspark.ml package we can import needed components including Deci- sionTreeRegressor, RandomForestRegressor, and LinearRegression. Before building models we apply VectorAssembler to create a feature vector by as- sembling all the feature columns into a single vector column. The column” value” is our dependent variable, and all the other variables are dependent variables. We build the model with train data set, and we can print the parameters. As we can see, for decision tree the initial depth is 5, number

of nodes is 25, and the number of used features is 15; for random forest regression we apply 20 trees and all 15 features; for SVR model we apply 15 features as well. The parameters should be tuned later. This can be seen in [Figure12.](#_bookmark47)



Figure 12: Build prediction models

### Time-series models

For time-series models we combine the above-mentioned two algorithms. For each combination of country and sector we build a pipeline, firstly conduct ARIMA then STL and save the results. Finally we get 84 combinations of country,sector, ARIMA model and STL model. We can check the details of the models using summary() for ARIMA and trend,seasonal and resid for STL.

# Data Mining

## Logical tests

We apply a 70/30 training/testing split in the data mining process. The reasons contains over-fitting prevention, data balance and statistical signif- icance. For one thing, large size of training data set reduces the risk of over-fitting because the algorithm would have enough data to learn from, and the remaining 30 percentage of data provides a reasonable amount of data to evaluate. For another, the 70/30 ratio provides statistically signif- icant results for model evaluation. With a sizable test set, one can have more confidence in the estimated model performance metrics. Also, we are handling large scale data set, so we have to take computing efficiency into consideration. Splitting the data into smaller test sets can be computation- ally intensive. 70/30 splits strike a balance between computational efficiency and model evaluation.

## Output of models

We run the models and measure the performance on test set. For metric, we choose Root Mean Squared Error(RMSE) for evaluation. The RMSE of Decision Tree Model is 1.412, for Random Forest is 2.068, for SVR 3.68. The resulsts can be seen in figure [13.](#_bookmark53) Regarding the time-series models, for each combination of country and sector we can plot the diagnosis of ARIMA including the coefficient estimates and residuals, for STL the three seasonal, trend and residual component. For instance the combination of Brazil and Power, we can plot figure [14.](#_bookmark56)

## Search for patterns

Firstly, we search for the distribution of global CO2 emissions among years and months. We plot five line graphs each representing global CO2 emis- sions for one year, we can figure out a pattern that the CO2 emissions in winters(for Northern hemisphere) is higher than other seasons, and in sum- mers we observe the lowest emissions. For instance, The fluctuation in 2021 can be seen in figure [15.](#_bookmark57)

Secondly, for each country we can try to plot the sum of CO2 emissions for all the sectors with date as x axis. For example, the main contribu-

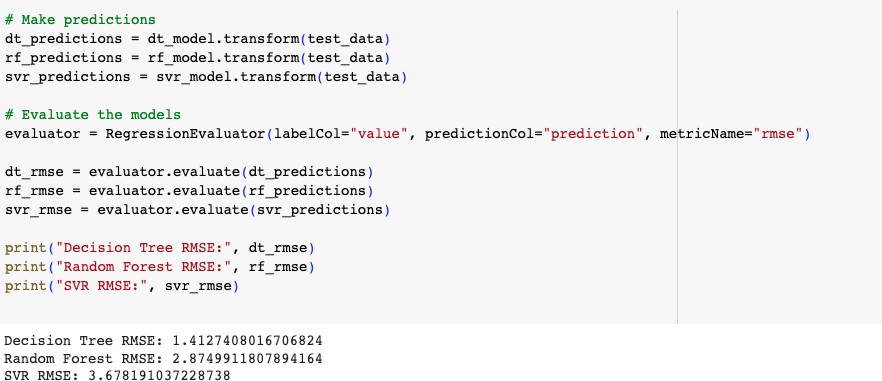


Figure 13: Result of models

tors are China,ROW, US and India. The patterns still follow our previous observation for most countries. We take US as an example in [figure16.](#_bookmark59)

If we combine the sector and country , and create a bar plot , we may find that China stands as the largest contributor to CO2 emissions globally, and compared with China, India exhibits significantly lower CO2 emissions per capita compared to other nations. For sectors, apparently power and industry are the top contributors.

# Interpretation

## Study and discuss mined patterns

Based on the RMSE of the three algorithms, this observation suggests that the collective structure of Random Forests did not yield a substantial en- hancement in performance compared to a solitary Decision Tree for the given dataset. In alternative terms, the inclusion of additional complexity in Ran- dom Forests may not be deemed essential in this particular context. The Decision Tree model is performing relatively well in terms of accuracy.

The interpretability of decision trees is relatively high, particularly when the depth of the tree is limited. In our case, Decision Trees possess the

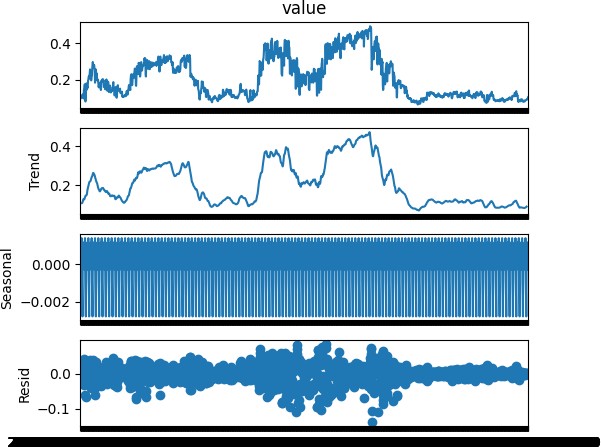


Figure 14: Result of time-series analysis models

ability to capture non-linear associations between features and the target variable. Additionally, feature importance allows for the identification and ranking of the most significant features, which can be advantageous when seeking to comprehend the underlying factors influence. Feature selection is recommended to assess the feature importance, particularly when utilizing tree-based models, in order to identify characteristics that have significant influence or insignificance. This has the potential to offer valuable insights for future feature engineering endeavors. Ensemble methods include the integration of several models, which might potentially result in a more re- silient and precise model. Considering the fact that several models have strong performance, employing an ensemble methodology could yield ad- vantageous outcomes.

For the descriptive analysis, we can find out that

* During winters , global CO2 emissions tend to rise due to an increase in the demand for heating.

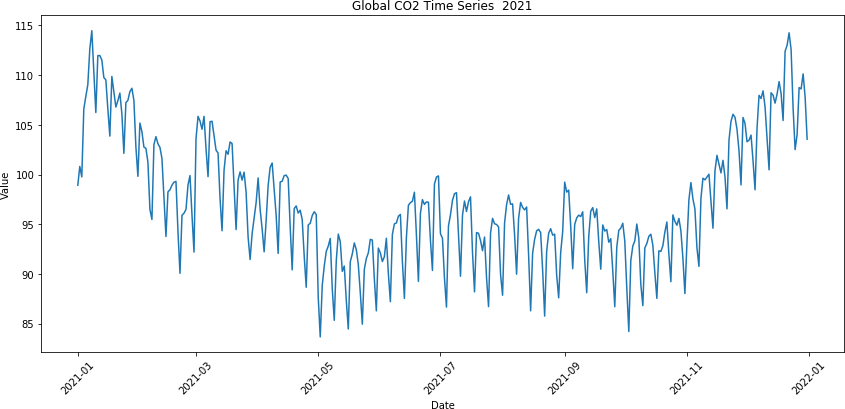


Figure 15: Global CO2 emissions 2021

* The top contributors of CO2 emissions are China, ROW, US, EU UK, India.
* The power sector leads with 60k MtCO2/day CO2 emissions, closely followed by industries and group transportation.
* Generally speaking the global CO2 emissions are increasing year by year .

## Data visualization

For time series , we visualize the result as Figure [18,](#_bookmark63) to illustrate we can find that for the specific combination of Brazil and power , the residual is almost normally distributed and the seasonal component has little pattern and the trend component is the most important.

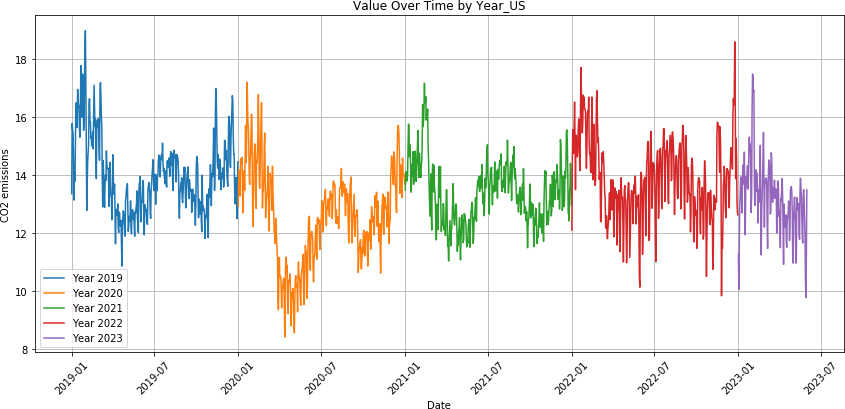


Figure 16: US CO2 emissions

## Assessing and Evaluating Models

We get the RMSE of 1.41 for prediction mode. Considering the data scale of our data set, the number of features and target variable’s values, this is acceptable. Due to the limit of data availability, we do not have enough useful features, and we manually one-hot encoded features based on two original features, although the risk of multicollinearity is minimal, we do lack the interpretability. And in context of our task, the prediction of global CO2 emissions is a very ambitious task and inherently challenging, so we can conclude that our model is a good attept.

## Multiple iteration

For this section, we can look back and tune the hyper parameter for the decision tree regression model. To illustrate, we focus on the max depth , the max bins and the minumum instances per node. We apply grid search in pyspark to search the optimal parameters, again , we use RMSE to evaluate the performance of the algorithm.

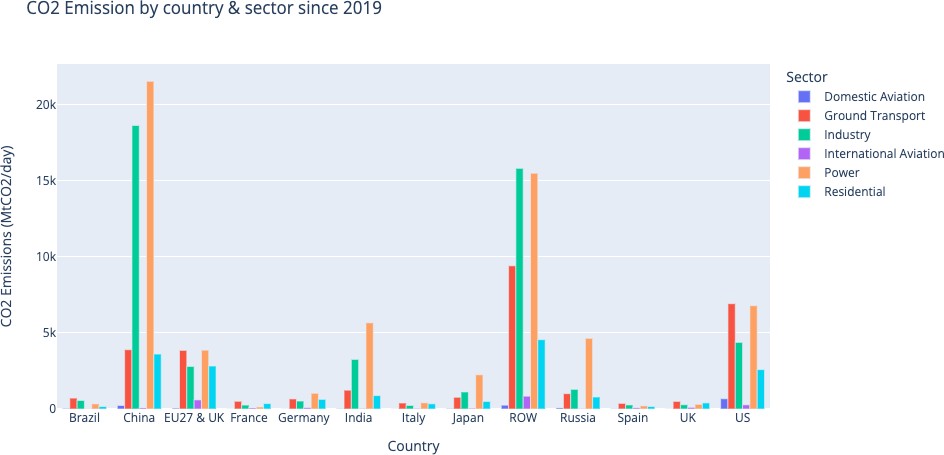


Figure 17: CO2 emissions by sector and country

Further, we create a cross validator to evaluate the performance of dif- ferent combinations of parameters. Finally we find out the best model is the model we have already built so we do not have to do more iterations.

# Final discussion

The incorporation of CRISP-DM (Cross-Industry Standard Process for Data Mining) was found to be an essential asset in this project, as it provided a systematic and structured framework for conducting data mining tasks. The continuous and iterative nature of the process allowed for ongoing re- finement and alignment with the project’s objectives, thereby showcasing the adaptable and resilient qualities of modern data analysis techniques.

In conclusion, this research emphasizes the importance of utilizing a purposeful and organized approach in the stages of data preparation and analysis. By employing the CRISP-DM framework and employing a variety of data mining techniques, we effectively transformed a raw dataset into meaningful and valuable insights, thereby laying the groundwork for further

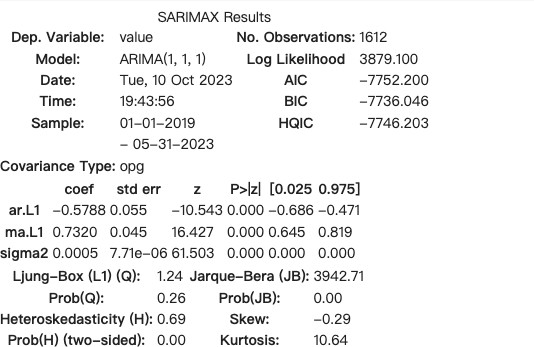


Figure 18: Result of ARIMA

exploration and potential real-world applications. Additional examination, contextual data, and enhanced modeling methodologies possess the capabil- ity to unveil deeper correlations and opportunities for progress. It is also a good practice to leverage business analytics tools and open stack tools.

图表

描述已自动生成

Figure 19: Result of STL

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 20: Parameters tuning

# References

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990).

Stl: A seasonal-trend decomposition. *J. Off. Stat*, *6* (1), 3–73.

Della Santa Navarrete, S., Borini, F. M., & Avrichir, I. (2020). Environmen- tal upgrading and the united nations sustainable development goals. *Journal of Cleaner Production*, *264* , 121563.

Gama, J., & Brazdil, P. (1995). Characterization of classification algorithms.

In *Portuguese conference on artificial intelligence* (pp. 189–200).

JHALANI. (2023, July). *Co2 emissions by sectors.* [https://www.kaggle](https://www.kaggle.com/datasets/saloni1712/co2-emissions?resource=download)

[.com/datasets/saloni1712/co2-emissions?resource=download](https://www.kaggle.com/datasets/saloni1712/co2-emissions?resource=download). (Retrieved August 20, 2023)

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education

sector. *Procedia computer science*, *57* , 500–508.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *Information computing and applications: Third international conference, icica 2012, chengde, china, september 14-16,* *2012. proceedings 3* (pp. 246–252).

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews:* *Computational Statistics*, *1* (1), 93–100.

Rathore, S. S., & Kumar, S. (2016). A decision tree regression based ap- proach for the number of software faults prediction. *ACM SIGSOFT* *Software Engineering Notes*, *41* (1), 1–6.

Ritchie, H., & Roser, M. (2023). Global inequalities in co2 emissions. *Our* *World in Data*.

Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). Arima models. *Time series analysis and its applications: with R ex-* *amples*, 75–163.

Smola, A. J., & Sch¨olkopf, B. (2004). A tutorial on support vector regression.

*Statistics and computing* , *14* , 199–222.

Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A comparative study of various clustering algorithms in data min-

ing. *International Journal of Engineering Research and Applications (IJERA)*, *2* (3), 1379–1384.

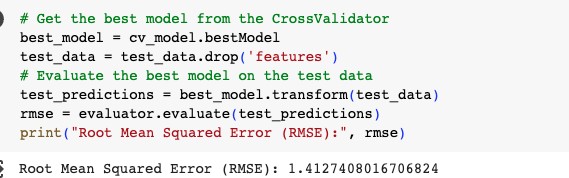


Figure 21: Result of Cross Validation

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: [https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.](https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html)).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data.