**REPORT**

PART 1:

* Method used for Data pre-processing –
  + Understood the overview of the dataset.
  + Dropped columns which were irrelevant for prediction.
  + Check for columns with Nan (empty) values or mismatching data types.
  + Replace them with mean and mode and refactored invalid data
  + Checked for outliers and removed them.
  + Converted Objects-data types to categorical values using one-hot encoding, label encoding and frequency encoding.
* Data Set –
  + **Breeding Bird Atlas** 
    - Domain - Breeding bird observations based on geographical locations
    - Type of Data - The dataset consists of categorical and numerical data
    - Features – 15 features
    - Number of Samples – 361582
    - Mean for each numerical column:
      * Fed. Region – 5.85
      * Month – 49.77
      * Day – 49.5
      * Year – 1964.18
      * Temperature – 49.46
      * Average UB Student – 2.85
    - Standard Deviation for each numerical column:
      * Fed. Region – 5.83
      * Month – 28.65
      * Day – 28.79
      * Year – 190.06
      * Temperature – 17.32
      * Average UB Student – 0.49
    - Missing Values for each column:
      * Fed. Region – 5795
      * Block ID – 2718
      * Map Link – 4717
      * County – 10602
      * Common Name - 10530
      * Scientific Name – 7485
      * NYS Protection Status – 8470
      * Family Name – 2456
      * Family Description -4733
      * Breeding Behavior -5183
      * Month – 358156
      * Day – 352244
      * Year – 10480
    - Graph-
      * Box Plot are used for detecting outliers, for column ‘Year’, we could see there are few values which lies beyond the first and third quartiles

A graph with numbers and a number of objects

Description automatically generated with medium confidence

* + - * For Fed. Region we do not have any outliers

A blue rectangular object with black lines

Description automatically generated

* + - * Bar Plot for County, here the data is linearly reducing and the most frequent value being Parulidae

A graph of a bar plot

Description automatically generated

* + - * Comparing the feature ‘NYC protection status’ with ‘Breeding status’ column we see the most repeating Status is ‘Protected’

A graph of breeding status

Description automatically generated

* + - * When comparing ‘year’ with ‘Breeding Status’, we could see at year 1984 there were lot of breeding with ‘confirmed’ status occurred

A graph of breeding status

Description automatically generated

* + - * Correlation matrix here gives us the relation of each feature with the target column ‘Breeding Status’, the breeding behavior column had the most relation with the target column and the least being Temperature which provides just 0.1% relation

A chart with numbers and a number of numbers

Description automatically generated with medium confidence

* + Dataset 2 –
    - Domain: Biological and ecological observations of penguins in Palmer Archipelago
    - Type of Data : Structured, tabular with numerical and categorical columns
    - Number of Samples: 344
    - Key Statistics:

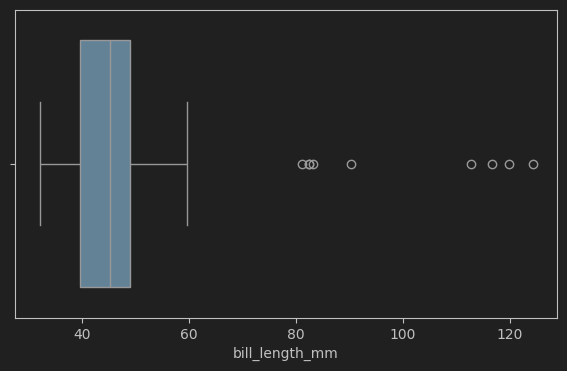
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | calorie requirement | average sleep duration | bill\_length\_mm | bill\_depth\_mm | flipper\_length\_mm | body\_mass\_g | year |
| count | 344 | 344 | 337 | 333 | 336 | 339 | 342 |
| mean | 5270.002907 | 10.447674 | 45.494214 | 18.018318 | 197.764881 | 4175.463127 | 2008.035 |
| std | 1067.959116 | 2.265895 | 10.815787 | 9.241384 | 27.764491 | 858.713267 | 0.816938 |

* Missing Values :

|  |  |
| --- | --- |
| Column Name | Is Null sum |
| species | 11 |
| island | 10 |
| calorie\_requirement | 0 |
| average\_sleep\_duration | 0 |
| bill\_length\_mm | 7 |
| bill\_depth\_mm | 11 |
| flipper\_length\_mm | 8 |
| body\_mass\_g | 5 |
| gender | 17 |
| year | 2 |

Graphs:

Box plots to detect the outliers in the data, so that further processing could be done. The majority of the numerical datasets showed outliers when box plot was constructed. Therefore, in order to zero in, Z -score was calculated and Z-score > 3 were classified as outliers.



A black screen with white lines

Description automatically generated

A graph with a bar and numbers

Description automatically generated with medium confidence

A black screen with white lines

Description automatically generated

A graph of a body mass

Description automatically generated

A graph with a bar and numbers

Description automatically generated with medium confidence

Z-score statistics: few outliers were found in bill\_length\_mm and bill\_depth\_mm using z-score. All were removed for further processing.

* Number of outliers in bill length: 9
* Number of outliers in bill depth: 3

Correlation Matrix:

A screenshot of a computer

Description automatically generated

Bar plots: The plot is self-evident and show cases that the greatest number of penguins are Adelie and the least is Chinstrap.

A graph of penguins in different colors

Description automatically generated with medium confidence

Average Sleep Duration between species: We observe that the average sleep durations are similar between species, which indicates that certain characteristics are common across the species.

A graph with blue squares

Description automatically generated

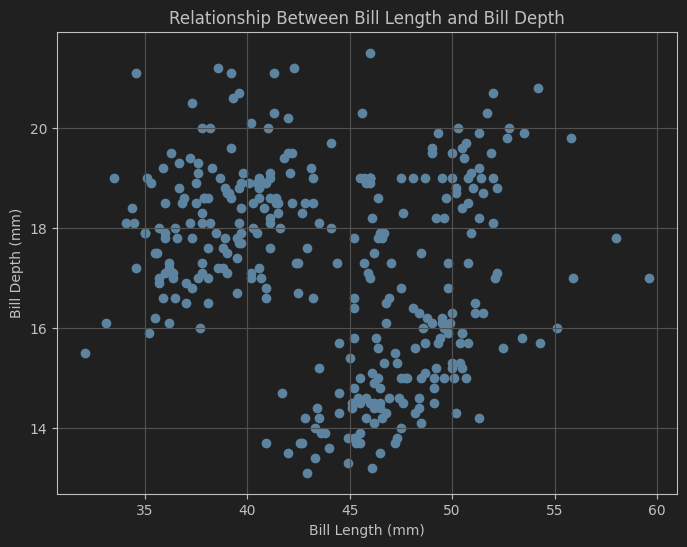
Pie Chart: It show cases the distribution of penguins by island, with maximum residing in Biscoe at 51.2% and the least in Torgersen at 14.5%

A pie chart with numbers and text

Description automatically generated

Scatter plot

Bill length and depth relationship: There appears to be clustering of data points into 3 groups. This is in line with our observation that there exist 3 distinct groups. Here we are drilling down to the data to corroborate the fact that the three groups might have different biological features/ characteristics.



* + Dataset 3 –
    - Domain: Impact of Minerals on Temperature per Country, scaled over time.
    - Type of Data: Structured, tabular with Categorical and numerical columns
    - Sample size: 63104, 12
    - Key Statistics:

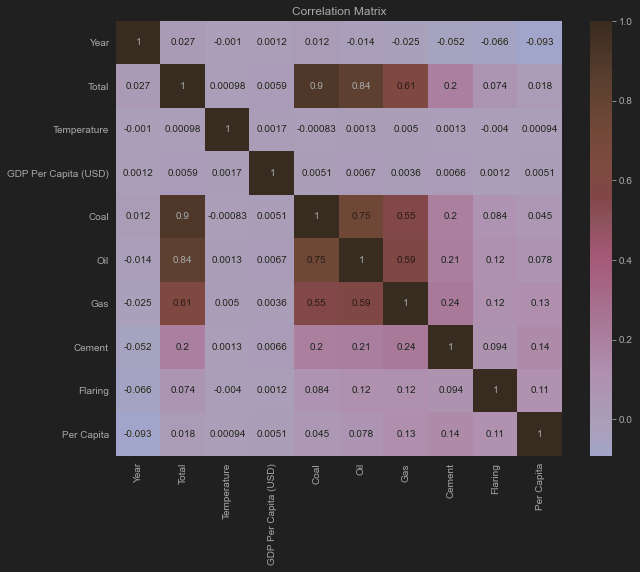
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year | Total | Temperature | GDP Per Capita (USD) | Coal | Oil | Gas | Cement | Flaring | Per Capita |
| count | 62307.000000 | 62381.000000 | 63104.000000 | 63104.000000 | 21797.000000 | 21774.000000 | 21717.000000 | 20488.000000 | 21338.000000 | 19392.000000 |
| mean | 1888.267097 | 73.683456 | 49.497813 | 39026.539015 | 127.387271 | 153.480038 | 125.162671 | 62.599364 | 56.074327 | 121.565443 |
| std | 122.651184 | 843.930381 | 17.292092 | 10975.539432 | 677.951392 | 670.830891 | 514.391435 | 353.918064 | 337.629062 | 489.339877 |

* Missing Values:

|  |  |
| --- | --- |
| Column | Null Count |
| Country | 2017 |
| ISO 3166-1 alpha-3 | 3621 |
| Year | 797 |
| Total | 723 |
| Temperature | 0 |
| GDP Per Capita (USD) | 0 |
| Coal | 41307 |
| Oil | 41330 |
| Gas | 41387 |
| Cement | 42616 |
| Flaring | 41766 |
| Other | 60419 |
| Per Capita | 43712 |

* Graph:

Correlation matrix: This showcases the correlation between all the features in the dataset. We can see that there exists a huge correlation between coal, gas, oil and cement data. With temperatures it is having less correlation. We will need to drill down to find the cause for this.



Line Chart: We plot the temperature to year graph for USA and find that their exist outliers, as in future data. We will remove this.

A graph showing the temperature of a year

Description automatically generated

A graph showing the temperature of a year

Description automatically generated

Scatter plot: next we plot a scatter plot between year and per capita, only to find that them exist an anomaly, that is at year 0, there is a peak of these values. So we effectively delete those, since we also see that there is a huge gap from year 0 till around year 1000.

A screen shot of a graph

Description automatically generated

A screen shot of a graph

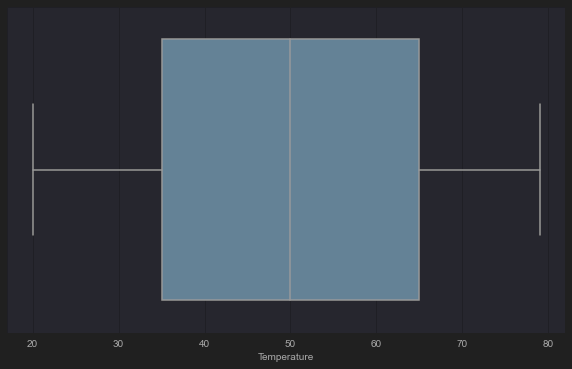
Description automatically generated

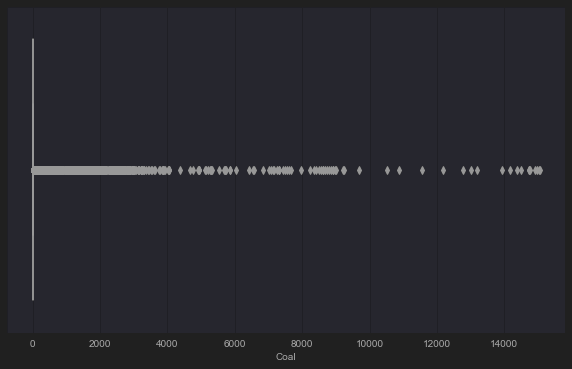
Box plot: We plot all numerical columns in this and try to find the outliers. The distribution of coal, gas, oil etc. is wide, so we take the next step, that is to calculate the z score to find and eliminate outliers more than 3 standard deviations away.

Outliers count (Z Score > 3):

* + - * Number of outliers in Total: 235
      * Number of outliers in Coal: 215
      * Number of outliers in Oil: 227
      * Number of outliers in Cement: 186
      * Number of outliers in Flaring: 167
      * Number of outliers in Per Capita: 246
      * Number of outliers in Gas: 256

We eliminate these values.





Bar chart: We plot the mean chart for all numerical for coal, gas, oil, cement and flaring and only showcase the top 10 values out of it.

A graph of a graph showing the number of countries/regions

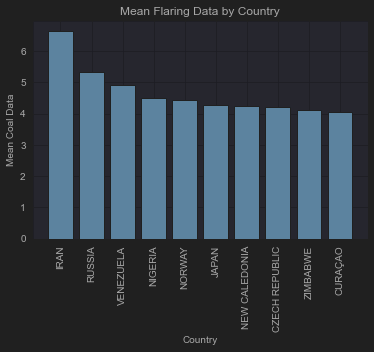
Description automatically generated A graph of oil prices

Description automatically generated

A graph of gas prices

Description automatically generated A graph of cement in a graph

Description automatically generated



PART 2: Penguin Dataset

Best Accuracy – 89.55 %

Loss Graph -

A graph with a blue line

Description automatically generated

Analysis – The loss graph for the model with a learning rate of 0.005 and 100,000 iterations shows that the model started with an initial loss of approximately 1.2 After a few iterations, the loss dropped significantly to 0.7, indicating rapid initial learning. The subsequent result suggests the model has reached a point where further training yields minimal improvements

Graphs for various Learning Rate and Iterations –

* Learning Rate – 0.001, Iterations – 100000, Accuracy – 85.07%

A graph with a line

Description automatically generated

* Learning Rate – 0.02 and Iterations – 200000, Accuracy – 89.55%

A graph with a blue line

Description automatically generated

* Learning Rate - 0.01 and Iterations – 250000, Accuracy – 88.06%

A graph with a line

Description automatically generated

* Learning Rate - 0.4 and Iterations – 300000, Accuracy – 88.06%

A graph with numbers and lines

Description automatically generated

* + A lower learning rate of 0.001 results in gradual convergence, which can prevent overshooting the minimum but may require more iterations
  + A medium learning rate of 0.005 often provides a good balance, leading to quicker convergence and improved accuracy
  + A high learning rate of 0.4 can lead to rapid learning but risks overshooting the optimal solution, causing fluctuations in loss and potentially lowering accuracy.
  + Increasing the number of iterations allows the model to refine its weights further, leading to lower loss. However, after a certain point, the benefits diminish, especially if the learning rate is not set appropriately.

Benefits / Drawbacks of Linear Regression –

* + Linear Regression performs requires less computational power and memory, making it suitable for smaller datasets or when computational resources are limited.
  + It performs well for linearly separable features making it an effective choice for binary classification tasks
  + Logistic regression is primarily used for binary classification problems
  + They are very sensitive to outliers which will affect the model’s accuracy

PART 3:

* Dataset Used – Diamonds
* Number of Samples – 53940
* Number of Features – 12
* Characteristics –
  + The dataset had 2 numerical columns and 10 categorical columns
  + “average us salary” and “number of diamonds mined (millions)” had no null values whereas rest columns had some null values
  + Number of outliers in each column A screenshot of a computer code

    Description automatically generated
  + After pre-processing I was left with 42190 samples and 16 features.
* Model Evaluation
  + Mean Squared Error
    - Linear Regression – 0.0037062
    - Ridge Regression – 0.0037061
  + Prediction v/s Actual Graphs
    - Linear Regression A graph with red and blue dots

      Description automatically generated
    - Ridge Regression – A graph with green and red lines

      Description automatically generated
  + Discussion
    - **Benefits and drawbacks of using OLS for weight computation** –
      * It is straightforward to implement and understand and it is very optimized for smaller datasets.
      * It is sensitive to outliers
      * If variables are highly correlated it can lead to very high standard errors and unstable coefficient estimates
      * It can overfit the model to the training data and ideally it is designed for linear relationships
    - **Evaluate the strengths and weaknesses of linear regression in general**
      * The computations involved in linear regression is efficient and simple
      * It can handle multiple independent variables
      * If the relationship is nonlinear, the model may be inadequate, and it is highly sensitive to outliers
      * With many predictors relative to the number of observations, linear regression can easily overfit the model
    - **Explain the motivation for using L2 regularization and how ridge regression improves upon linear regression. Discuss its benefits and limitations compared to linear regression**
      * L2 regularization can prevent overfitting as it adds a penalty for larger co-efficient
      * It can easily handle multicollinearity by shrinking co-efficient to zero
      * L2 regression improves upon linear regression as it adds penalty term this makes L2 better than linear
      * Limitations of L2 –
        + It introduces bias to model
        + It is more complex to implement
        + It is computationally more intensive than linear regression

PART 4:

* Dataset Used – Wine
* Number of Samples – 6497
* Number of Features – 12
* Target Variable – Quality
* Characteristics –
  + All are numerical columns
  + All columns show no null or na values
  + Sulphates and alcohol, have the largest positive correlation with target variable
* We run the combined dataset, first without a stopping criteria and then with criteria/ threshold of 0.00001. We have chosen this, since upon multiple inspections of trying to find a convergence point for 0.01, even after 1.75 million iterations, it was not stopping and at this point we were checking if the gradient fell below threshold for all values of the gradient vector.

|  |  |  |
| --- | --- | --- |
| Initialization Method | Stopping Criteria enabled | MSE metrics |
| Random | No | 12.72984 |
| Zero | No | 12.72984 |
| Xavier | No | 12.76704 |
| Random | Yes | 12.52427 |
| Zero | Yes | 12.61831 |
| Xavier | Yes | 12.77234 |

|  |  |  |
| --- | --- | --- |
|  |  | Random weight initialization with stopping criteria not activated |
|  |  | Zero weight initialization with stopping criteria not activated |
|  |  | Xavier weight initialization with stopping criteria not activated |
|  |  | Xavier weight initialization with stopping criteria activated. Stopped at 24th iteration |
|  |  | Zero weight initialization with stopping criteria activated. Stopped at 29th iteration |
|  |  | Random weight initialization with stopping criteria activated. Stopped at 15th iteration |

* We observe that with mixed data, we don’t see significant differences between initializations and due to the random nature of initialization process, we tend to get skewed results. In order to mitigate that we conducted experiment on recued dataset where we could have much greater control. Defined below is the methodology and results of that experiment.
* Methodology: The dataset was cleaned and processed. Further, since the dataset was of smaller size and showed outliers in the dataset, we have approached the problem by implementing the same models twice. One without removing the outliers and the other by removing outliers whose z score was greater than 3. We calculate the MSE and report it here. We also implement mini batches, so that we can also handle bigger size samples for other datasets. In the end, since we are aware theoretically that Xavier initialization yields better results, we introduce a method to quickly find optimal lambda values for L1 and L2 regularization. Here we run the Gradient descent for 1000 iterations with another 1000 random lambda terms and find the lowest one with MSE. Refer code at additional\_experiment.ipynb
* Model Evaluation:

|  |  |  |
| --- | --- | --- |
| Initialization method | Data state | MSE metrics |
| Random Initialization | Original Data | 4.17 |
| Zero Initialization | Original Data | 4.16904 |
| Xavier Initialization | Original Data | 4.71719 |
| Random Initialization | Outliers removed | 0.67109 |
| Zero Initialization | Outliers removed | 0.67109 |
| Xavier Initialization | Outliers removed | 0.66577 |
| Xavier Initialization | Outliers removed & optimization run to find optimum lambda | 0.66563 |

We can observe that the metrics were skewed when we were using the original dataset (red wine). This was a conscious decision considering the sample size. We observe that none of the methods provide better results and is completely based on randomness of weight initialization. After this, we proceed to work on outliers removed data. We remove all the datapoints greater than 3 standard deviations on each side using the z score method. Once the process was completed, we observe that the best result was given by the Xavier Initialization method. To further tune this, we in fact ran a test to find the best lambda values in over 1000 experiments.

|  |  |  |
| --- | --- | --- |
|  |  | Random Weights |
|  |  | Zero Weights |
|  |  | Xavier Weights |
|  |  | Xavier Weights with tuned hyperparameters |

**Stopping criteria considerations:**

We observe that when we use gradient descent threshold limit of 0.00001 for any of the gradients, we can see the process stop in under 30 iterations with clear stable cost values being achieved. We have taken the threshold to be very low, since during the experiments we had ran to find out locations where the loop would exist when we took threshold of 0.01, but even after 1.75 million iterations on Xavier initialization, we were not able to seek convergence, thus we resorted to lowering the threshold point.

**Discussion:**

* L1 Regularization: encourages sparsity in the model, driving some feature weights to zero. This results in feature selection as it removes irrelevant features.
* L2 Regularization adds penalty proportional to squared magnitude of the coefficients. It prevents large coefficients, making the model more robust to noise and the risk of overfitting is reduced.
* Elastic Net regularization: Combines both L1 and L2 regularization. That is by adding L1, it gains the characteristics of L1, that encourages sparsity in model and by adding L2, it adds penalty to large coefficients.

**Key Advantages with respect to application:**

Since in the dataset, the correlation between features is predominant, Elastic net tends to handle multi collinearity better than lasso. Here Elastic net allows for feature selection, but at the same time stabilizing the model with L2. The combination of L1 and L2 allows the model to be more robust to handle real world data and reduce risk of overfitting.

**Elastic net regression:**

Benefits:

* Balanced regularization: since we combined L1 and L2
* Regularization strength control: controlled using lambda values
* Improved generalization: we often see that elastic net regression performs better than L1 and L2 alone in terms of generalization.

Potential Drawbacks:

* Hyperparameter tuning becomes a cumbersome task. We in fact approach this problem by running a short simulation when we just use the red wine dataset. But having to decide on lambda values adds complexity to the model.
* Potential overfitting if not properly tuned. For example, if L2 is too small, then it might not handle correlated features properly and if L1 is small, then feature selection might not yield good results.
* Higher computational Cost: Due to additional tuning and calculations required, the computational cost is on the higher side.

**Gradient Descent:**

Benefits:

* Efficiency with large datasets: It scales much better with large datasets.
* Flexibility: works well with multiple types of cost functions.

Drawbacks:

* Convergence Issues: it can get stuck in local minima if learning parameter is not properly set.
* Hyperparameter sensitivity: introduces own set of hyperparameters, which need to be tuned alongside other regularization parameters.
* Normalization: Requires the input features to be normalized. This adds additional complexity to the workflow.