Pollution prediction using LSTM

Isam Al Jawarneh College of Computing and Informatics

University of Sharjah  
Sharjah, United Arab Emirates  
[ijawarneh@sharjah.ac.ae](mailto:ijawarneh@sharjah.ac.ae)

Madyan Omar Ibrahim Bagosher  
*College of Computing and Informatics*

*University of Sharjah*Sharjah, United Arab Emirates  
U23200049@sharjah.ac.ae

Fatemeh Mohammadi Aghjehmashhad  
 *College of Computing and Informatics*

*University of Sharjah*Sharjah, United Arab Emirates  
U23200047@sharjah.ac.ae

*Abstract*— With the advancement of industrialization, air pollution has emerged as a significant concern.

Keywords— Correlation coefficient, correlation matrix, forecasting, Geohash, LSTM, pollution, RMSE.

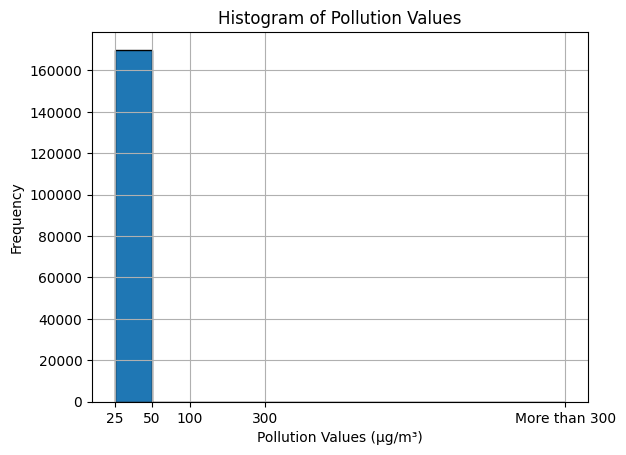
# Introduction

This ….

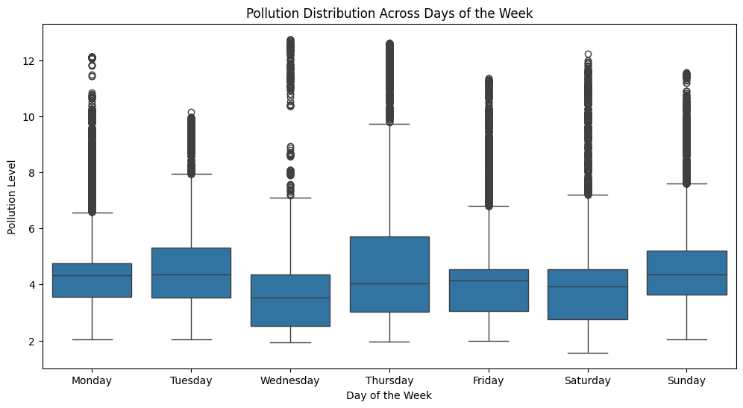
# BBasic Information about our data

## Basic Information about fine-grained low cost air quality (AQ) data

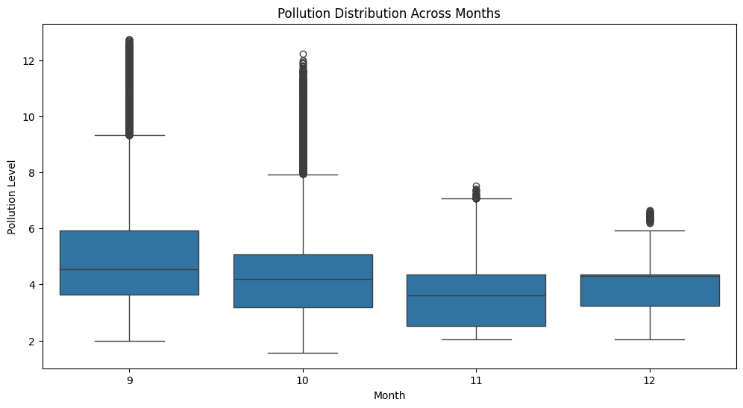
Various components can be considered when determining air quality. In our data, we focus on “pollution” and the geographic locations in New York City. The maximum pollution value and the minimum value are 12.74 and 1.57, respectively. The following picture is a histogram of pollution values:



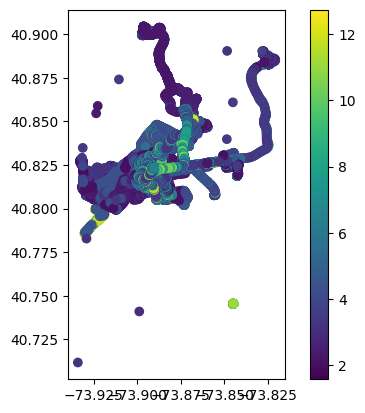
Also, we can see the pollution distribution across days of the week in the following picture:



Additionally, we can see the pollution distribution across months in the following picture:



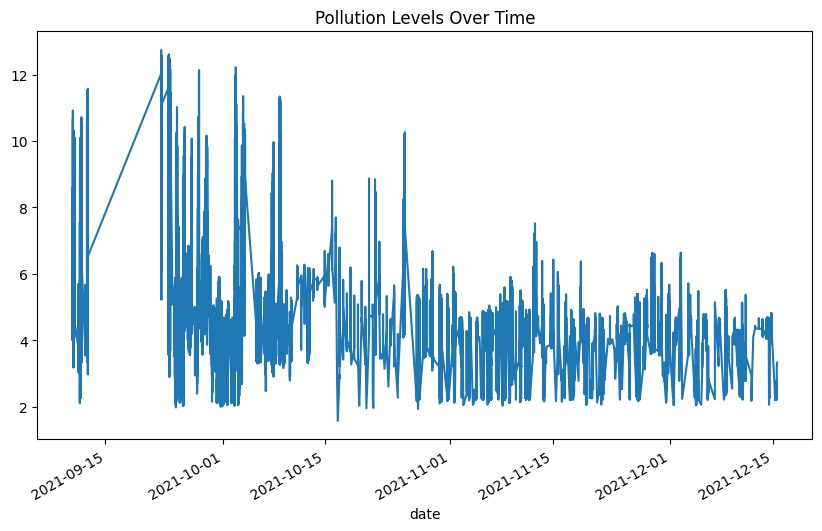
Also, the following map can help us to visualize the spatial distribution of pollution levels:



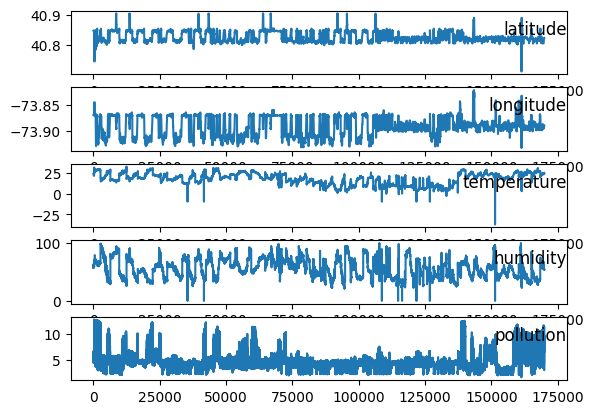
We can see areas with high and low pollution concentration.

The data was covered from 2021-09-10 at 12:29:09 to 2021-12-15 at 14:35:55 which makes 96 days and 02:06:46.

For investigating temporal trend of pollution, we plot the pollution levels over time in the following figure:



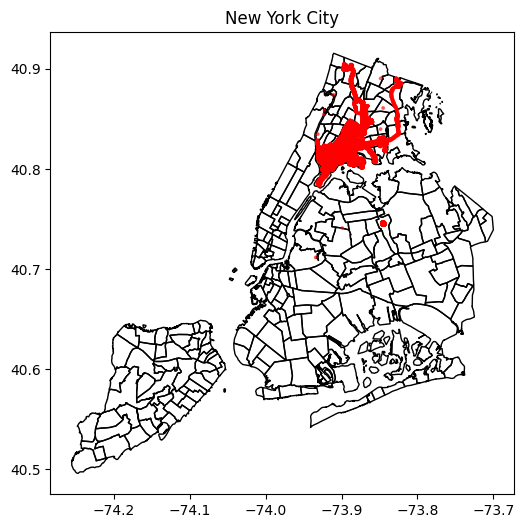
Also, we provide the graph for other variables in the data such as: Latitude, Longitude, temperature, humidity and pollution as follow:



# VVisualization

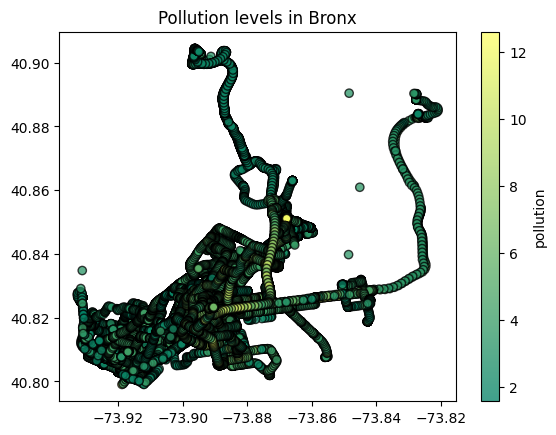
## Pollution Distribution in New York City

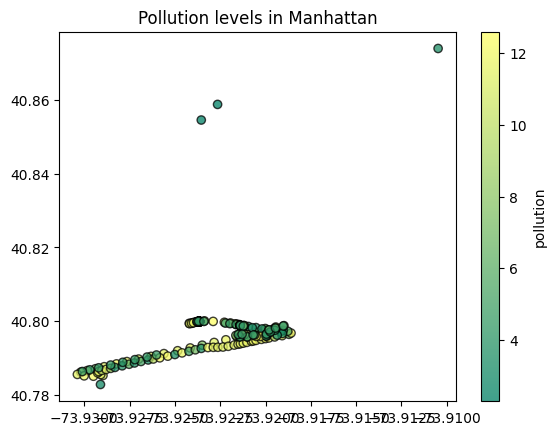
First, by creating the scatter plot on top of a map of New York City (nyc\_map), we managed to visualize pollution data. In fact, the following map provides spatial context for understanding the *distribution of pollution* within New York City, where each point on the map represents a pollution data point, and the size of the point reflects the pollution level.

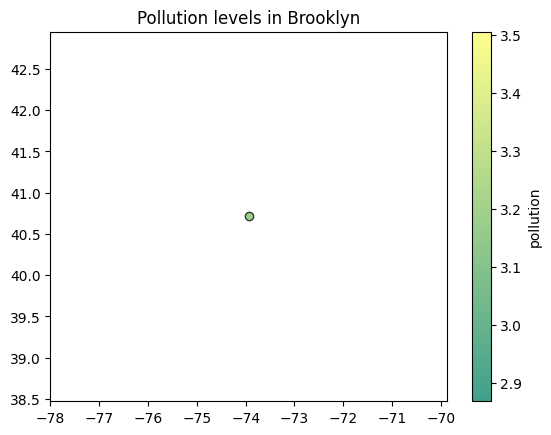


From this map, we can see where pollution is concentrated and how it spreads across different areas within the city.

Additionally, the following scatter plots help us to visualize pollution levels in each area.







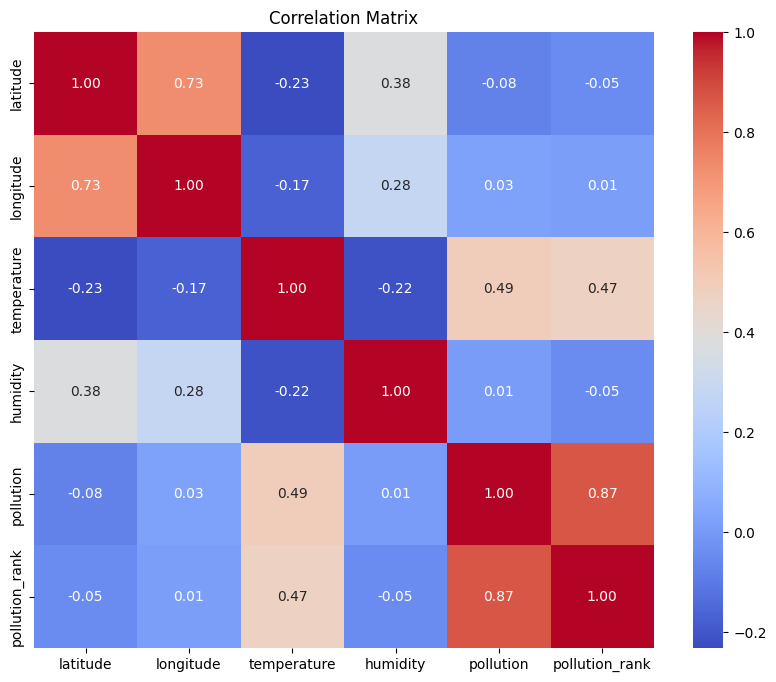


## Correlation Matrix

We recall that a correlation matrix is a table that shows the correlation coefficients between variables in a dataset. Each cell in the table represents the correlation coefficient between two variables. The correlation coefficient measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

* 1 indicates a perfect positive linear relationship
* -1 indicates a perfect negative linear relationship
* 0 indicates no linear relationship

We calculated the correlation matrix for our data and it looks like the temperature is the variable that has the highest correlation with the response variable pollution, at the par with 0.49, other variables are less significant.



# SSEVERAL CORRELATION COEFFICIENTS

In this section, we want to investigate any association between pollution (also pollution rank) with geohash variable using several methods:

## Reverse Kendall's correlation coefficient

We calculated the average rank of pollution values per geohash category as well as the size of each geohash category (i.e., the number of observations in each category).Since both are ordinal data, it is suitable to use Reverse Kendall's correlation coefficient in order to find the association between the average rank of pollution values (ordinal data) and the size of each geohash category (ordinal data). Also, we repeated this strategy by using the neighborhood variable and borough variable instead of geohash variable, and we got the following results:

* Using “geohash” variable:
* Reverse Kendall's correlation coefficient: 0.18607703406282353
* Using “neighborhood” variable:
* Reverse Kendall's correlation coefficient: 0.3031491751402424
* Using “borough” variable:
* Reverse Kendall's correlation coefficient: 0.19999999999999998

One can see that the association between the average rank of pollution values per neighborhood category and the size of each neighborhood is positive and is higher than the other. This would indicate that larger neighborhood areas tend to have higher pollution levels on average.

Also, we repeated the above algorithm with different precision=[3, 4, 5]. We got the following results:

* Reverse Kendall's correlation coefficient for geohash: 1.0
* Reverse Kendall's correlation coefficient for neighborhood: -0.10734744162189888
* Reverse Kendall's correlation coefficient for borough: -0.39999999999999997

We got the same result for each precision in [3,4,5], and this can suggest that additional precision beyond a certain point doesn't significantly alter the ordering of the geohash (respectively, neighborhood and borough) categories concerning pollution ranks. As we can see the association between the average rank of pollution values per geohash category and the size of each geohash is positive and perfect. It suggest that there is a linear relationship between them.

## Spearman's Rank Correlation Coefficient

This measures the strength and direction of association between the ranks of 'geohash' (respectively, neighborhood and borough) and 'pollution'. It is suitable when the relationship is monotonic but not necessarily linear. We remind that for calculating the Spearman's Rank Correlation Coefficient, one can use the following formula:

, (1)

where

ρ (rho) is Spearman's rank correlation coefficient.

Σd² is the sum of the squared rank differences.

n is the number of observations.

First, we calculated the Spearman's Rank Correlation Coefficient between 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision, and we got the following:

* Spearman's Rank Correlation Coefficients:
* Geohash vs. Pollution: -0.06654169631677545
* Neighborhood vs. Pollution: 0.011637772263997271
* Borough vs. Pollution: 0.08611017903128607
* P-values:
* Geohash vs. Pollution: 4.454497640901313e-166
* Neighborhood vs. Pollution: 1.5984415142310416e-06
* Borough vs. Pollution: 4.068933168365492e-277

From the result, one can see that overall there is a weak

positive monotonic relationship between “pollution” and

geohash (respectively, neighborhood and borough).

Second, we calculated the Spearman's Rank Correlation Coefficient between the ranks of 'geohash' and 'pollution' with different precision=[3, 4, 5], however the results did not differ so much. The results are as follow:

* Spearman's Rank Correlation Coefficients for Geohash:
* Precision 3: -0.06654169631677545
* P-value: 4.454497640901313e-166
* Precision 4: -0.03478746127459427
* P-value: 1.1042075713576446e-46
* Precision 5: -0.005006042655342761
* P-value: 0.03901315711545015

## ANOVA (Analysis of Variance)

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences among means of three or more groups. It assesses whether the means of different groups are statistically significantly different from each other. ANOVA tests the null hypothesis that all group means are equal against the alternative hypothesis that at least one group mean is different.

We need the F-statistic and p-value for the ANOVA test. We recall that the F-statistic measures the ratio of the variance between groups to the variance within groups. And the p-value indicates the probability of obtaining the observed results (or more extreme results) if the null hypothesis is true. The decision rule is that if the p-value is less than a chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that there are statistically significant differences in pollution levels across different geohash (respectively, neighborhood and borough) categories. Otherwise, we fail to reject the null hypothesis, suggesting no significant differences in pollution levels between the groups.

First, we show the ANOVA test results for 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision:

* ANOVA Test Results:
* Neighborhood - F-statistic: 194.2967984990712, P-value: 0.0
* Borough - F-statistic: 1571.4452929273557, P-value: 0.0
* Geohash - F-statistic: 3520.2944091142294, P-value: 0.0

Second, we show the ANOVA test results for 'geohash'

and 'pollution' with applying precision=[3, 4, 5]:

* ANOVA Test Results:
* For geohash:
* Precision Level precision\_3: F-statistic = 3520.2944091142294, P-value = 0.0
* Precision Level precision\_4: F-statistic = 1762.629585813516, P-value = 0.0
* Precision Level precision\_5: F-statistic = 412.3721028769785, P-value = 0.0

One can see that the F-statistic values are large, however the p-values are very small. So, we reject Null hypothesis and we may say that there are statistically significant differences in pollution levels across different geohash (respectively, neighborhood and borough) based on ANOVA test.

## Kruskal-Wallis test

The Kruskal-Wallis test is commonly used when the assumptions of parametric tests like ANOVA are violated, such as when the data are not normally distributed or when the groups have unequal variances.

First, we show the Kruskal-Wallis test results for 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision:

* Kruskal-Wallis Test Results:
* Neighborhood: Statistic = 3194.7088275131123, P-value = 0.0
* Borough: Statistic = 1281.6427998977276, P-value = 3.178456778603967e-276
* Geohash: Statistic = 752.7255492814534, P-value = 1.025109961523071e-165

One can see that the p-values are very small. So, we reject Null hypothesis and we may say that there are statistically significant differences in pollution levels across different geohash (respectively, neighborhood and borough) based on the Kruskal-Wallis test results.

Second, we show the Kruskal-Wallis test results for 'geohash' and 'pollution' with applying precision=[3, 4, 5]:

* Kruskal-Wallis Test Results For geohash:
* Precision Level precision\_3: Statistic = 752.7255492814534, P-value = 1.025109961523071e-165
* Precision Level precision\_4: Statistic = 767.5797478376305, P-value = 2.0997884702223404e-167
* Precision Level precision\_5: Statistic = 1549.7445498734908, P-value = 0.0

It is clear the p-values are very small and we lead to reject Null hypothesis and we may say that there are statistically significant differences in pollution levels across different geohash based on the Kruskal-Wallis test results.

## Kendall's Tau Correlation Coefficient

This measures the strength and direction of association between the ranks of 'geohash' and 'pollution'. It is suitable for ordinal data or when the relationship is not necessarily linear.

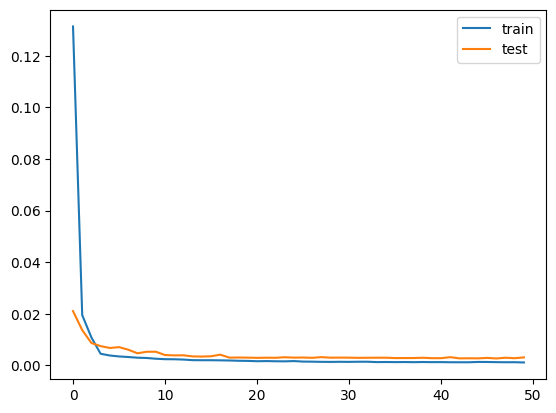
* Kendall's Tau is a rank-based correlation coefficient that measures the ordinal association between two variables. We calculated the Kendall's Tau correlation coefficient and the associated p-value. The correlation coefficient ranges from -1 to 1. The p-value indicates the significance of the correlation coefficient. If the p-value is less than a chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that there is a statistically significant association between 'geohash' and 'pollution'. Otherwise, we fail to reject the null hypothesis, suggesting no significant association. We got the following results by applying precision=[3, 4, 5]:
* Kendall's Tau Correlation Coefficient Results for neighborhood:
* For neighborhood: Correlation = 0.009259718338660744, P-value = 1.3397951600462843e-07
* Kendall's Tau Correlation Coefficient Results for borough:
* For borough: Correlation = 0.07034127731459229, P-value = 4.592161998189867e-276
* Kendall's Tau Correlation Coefficient Results for geohash:
* For geohash Precision Level 3: Correlation = -0.054383698700505266, P-value = 1.0251099615631986e-165
* For geohash Precision Level 4: Correlation = -0.028405622420667093, P-value = 1.2769262199553581e-46
* For geohash Precision Level 5: Correlation = -0.003214327504560284, P-value = 0.08134248790613795

One can see that all the p-values are small except the case of “geohash” with precision 5. In this case, p-value=0.08 and is not less than the significance level= 0.05, and so we do not reject null hypothesis. This means that there is no significant association between 'geohash' and 'pollution'.

# Pollution Prediction using LSTM

## Training the data (To be completed…)

We split our data into train and test sets. We used the LSTM model and we get the following graph:



We calculated that for the prediction model we have RMSE=0.001 and for the SVR model is 0.009, and for linear regression, we got MSRE=0.004.

## Prediction

By grouping the data by location (geohash) and training data for each location, and training a forecasting model for each location, given the pollution (PM, particulate matters) for prior hours, we forecast the pollution at the next 24 hour as follow:

predictions = {

* 'dr5rt': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr5ry': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr5rz': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72j': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72m': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72n': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72p': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72q': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72r': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72w': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr72x': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
* 'dr782': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891],
*  'dr788': [4.71040125, 5.51816975, 5.73704911, ..., 5.96994891]

}

RESULTS COMPARISON

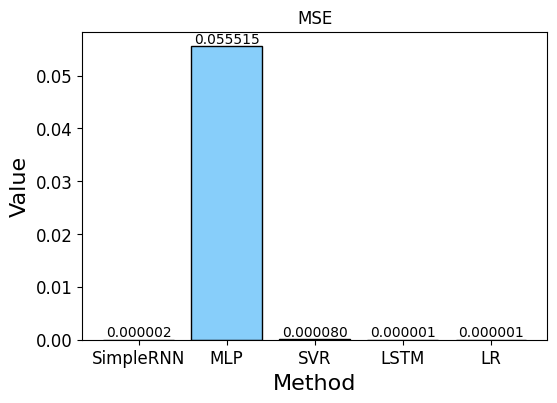
A graph of a bar graph

Description automatically generatedDuring the evaluation process, we calculated various metrics such as the mean absolute error (MAE) which is a measure that tells us the average magnitude of the errors in our forecasts. Our Linear Regression model achieves the lowest MAE score followed by the LSTM and RNN models.

A graph of different colored bars

Description automatically generated

The next metric we calculated was the mean squared error (MSE) which places a higher weight on larger errors. Once again our Linear regression model and LSTM model display a low MSE score, followed by our RNN model.



We proceeded to calculate the root mean squared error (RMSE) as well as other metrics such as: Accuracy, Correct prediction count, error rate etc.

A graph of different colored bars

Description automatically generated

A graph of a number of different colored squares

Description automatically generated with medium confidenceA graph of error rate

Description automatically generated

The table below displays the summary of all the metrics rounded up that have been calculated for each of the models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | Max prediction error | Correct Predictions | Accuracy | Error Rate |
| Long Short-Term Memory (LSTM) | 0.00060 | 0.00000 | 0.001 | 0.088 | 113890 | 70.63 | 29.37 |
| Recurrent Neural Network (RNN) | 0.00069 | 0.00000 | 0.001 | 0.097 | 108619 | 67.37 | 32.63 |
| Multi-Layer perceptron (MLP) | 0.16276 | 0.05552 | 0.236 | 0.793 | 20014 | 58.90 | 41.10 |
| Support Vector Regression (SVR) | 0.00722 | 0.00008 | 0.009 | 0.060 | 80691 | 50.04 | 49.96 |
| Linear Regression (LR) | 0.00011 | 0.00000 | 0.004 | 0.097 | 79007 | 49.00 | 51.00 |

The LSTM model seems to perform the best and achieved the best results.

##### Acknowledgment

We extend our heartfelt thanks to the referee for their thoughtful comments and constructive feedback.

##### References

1. J. Brownlee, (2017, August). Multivariate Time Series Forecasting with LSTMs in Keras. Machine Learning Mastery. Retrieved from <https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>
2. J. Brownlee, (2020, August). Feature Selection with Real and Categorical Data. Machine Learning Mastery. Retrieved from <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>