Milestone 2: Draft White Paper

Felipe Rodriguez

Bellevue University

DSC 680 Applied Data Science

Professor Amirfarrokh Iranitalab

April 28, 2024

Project Overview/Background

A local company wants to understand weather data in London. Additionally, the company wants to build a model to accurately forecast temperature. The model and predictions will be used to plan outdoor events. These events can include birthday parties, weddings, charity events and many others. By being able to predict the weather, the company can help costumers plan events and make date recommendations. The specific values that be used to make predictions are mean temp, precipitation, and snow depth.

Data Overview

The data is being obtained from Kaggle and is sourced from the European Climate Assessment. The data contains the following information: date, cloud_cover, sunshine, global_radiation, max_temp, mean_temp, min_temp, precipitation, pressure, and snow_depth. The data will allow for the analysis of weather based on the features provided. The fields in the data contain the following information:

Data Dictionary

column	description	data_type
date	Date of record in YYYYMMDD Format	object
cloud_cover	Cloud cover measurement in oktas	float64
sunshine	Sunshine measure in hours	float64
global_radiation	Irradiance measurement in Watt per square meter (W/m2)	float64
max_temp	Maximum temperature recorded in degrees Celsius (°C)	float64
mean_temp	Mean temperature recorded in degrees Celsius (°C)	float64
min_temp	Min temperature recorded in degrees Celsius (°C)	float64
precipitation	Precipitation measurement in millimeters (mm)	float64
pressure	Pressure measurement in Pascals (Pa)	float64
snow_depth	Snow depth measurement in centimeters (cm)	float64

The data preparation phase requires cleansing and prepping the data so that it is fit for use. The first portion involves removing null values. This piece will ensure that when creating models there are no errors. Prepping the data for modeling also involves creating specific date fields such as year, month, and day.

Methods

The methods used in this project will include various visualizations of prices, correlation matrices of the variables, and an ARIMA model to predict weather. The visualizations will include various scenarios of the weather data, which will include charts illustrating the weather trends. The correlation matrix that will be created will help understand the correlation between the weather features.

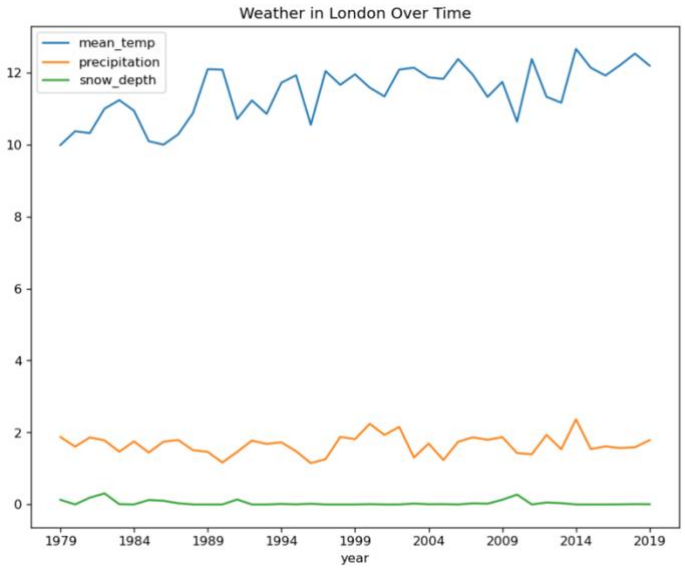
The initial model chosen for this project was an Autoregressive Integrated Moving Average or ARIMA model. ARIMA is used because it "is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends" (Hayes, 2024). In the first model created, the column used was mean temp. This model performed very well with a high R-square score (see Appendix A). R-square is used as the metric for determining model accuracy because R-square shows how well the data fits the regression model (Taylor, 2023). Since this model gave high results, the same model was conducted for the precipitation and snow depth variable. The model results for precipitation and snow depth did not perform as expected. While the same type of model was created, these two provided lower R-square scores. Because of this a Seasonal Autoregressive Integrated Moving Average or SARIMA model was used instead (see appendix B). This additional model on the other two variables provided low accuracy as well.

Assumptions

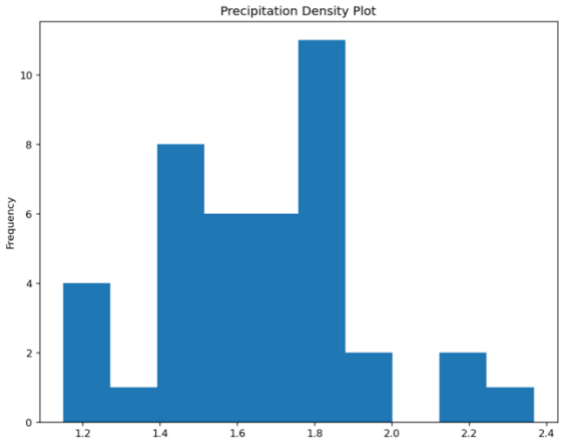
Prior to analyzing the data, the following assumptions are made:

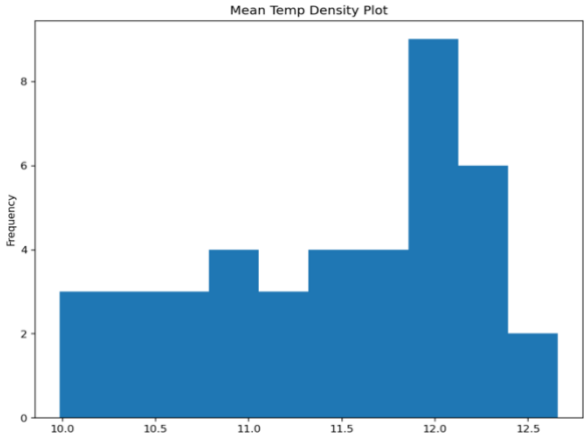
- The data provided and gathered reflects the weather in London.
- The model assumes there are no external factors that are impacting weather.
- While weather can be predicted, it will not be 100% accurate due to external behavior

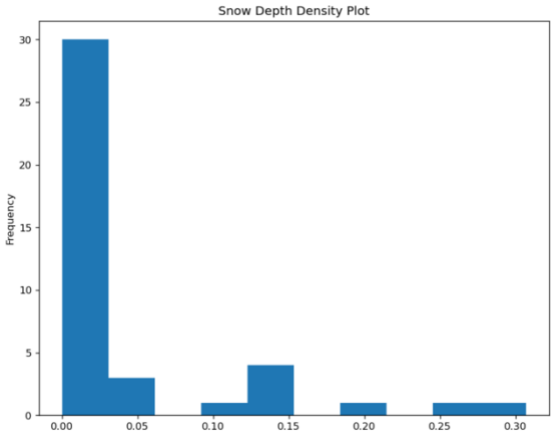
Analysis



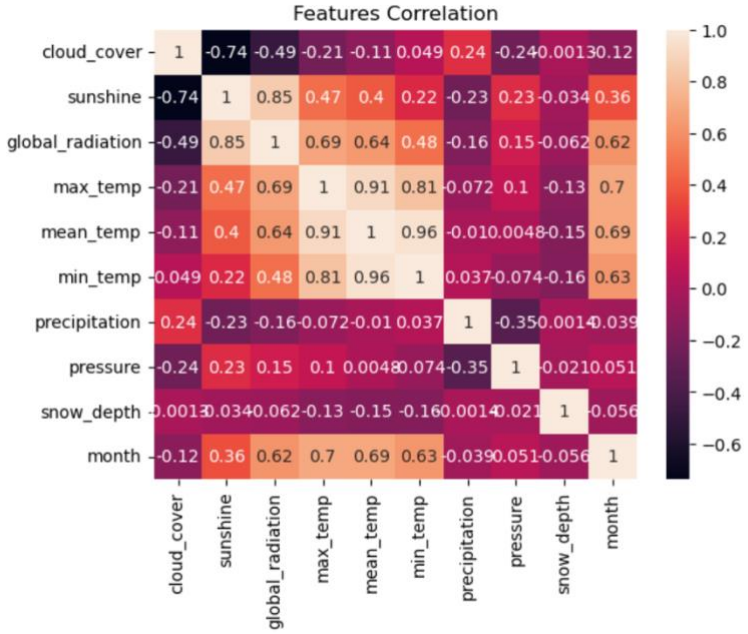
The data explored gives insights on how the weather changes over time. The first plot gives insight on how the target weather features have changed overtime. The chart to the right shows that the mean temperature has been increasing since 1979, while precipitation and snow depth have remained consistent. The next set of visuals give an idea of how the mean temp, precipitation, and snow depth are distributed.







To understand further how the variables relate to each other, a correlation matrix was created on the different features of the data. This analysis is done on the numeric features of the dataset. Below are how the variables relate to each other. The initial assumption on this data set was that month was going to play the biggest role in mean temp, precipitation, and snow depth. In a correlation matrix, the higher the value is to 1 the higher the correlation is. This analysis shows that the month field plays a big role on the temperature fields over precipitation and snow depth.



Challenges/Limitations/Ethical Considerations

The challenges in this data were found when creating the model. When using the proposed features, the model performed well for mean temp but not for precipitation and snow depth. Because of this finding, there was a decision to conduct a SARIMA model for these two,

to adjust for seasonality. However, that improved the model slightly for snow depth but not for precipitation.

A limitation of the ARIMA model is that it does not take seasonality into effect. Seasonality occurs when there is a pattern in the data. To adjust for this, a SARIMA was used for the fields precipitation and snow depth. Additionally, month did not affect the two fields precipitation and snow depth as expected. This observation could potentially affect model performance since there is not many differentiating features, which will make it hard to predict weather trends.

Ethical considerations for this data and models are to ensure that this data is secured from unauthorized uses and distribution. This model is meant to aid this organization and should be carefully reviewed before sharing. Additionally, weather predictions can be conducted but there is a factor of uncertainty.

Future Uses/Recommendations/Implementation Plan/Conclusion

The ARIMA model performed well for the field mean temp with a high R-Squared score. It is recommended that this model be implemented for use. The models for precipitation and snow depth performed poorly and when adjusting for seasonality they improved slightly. A recommendation for these two fields is to continue to gather and analyze the data to see if there are other models that can be created to achieve better performance.

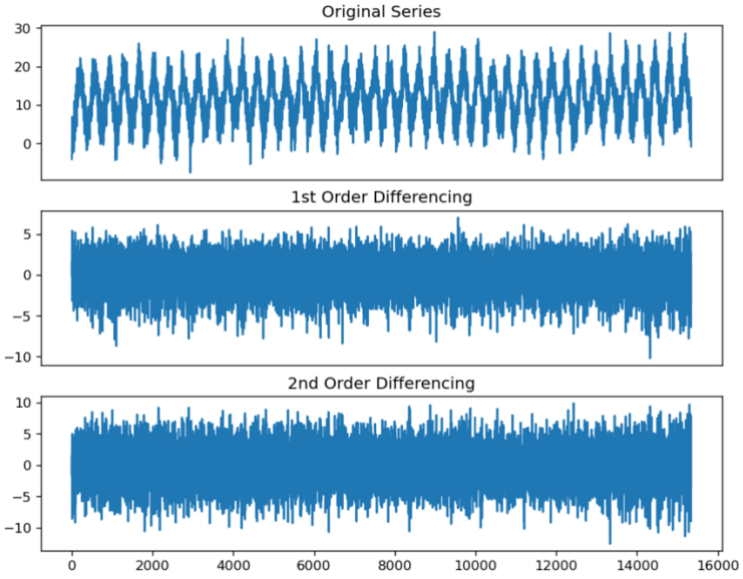
The best way to implement this model is to develop a web application that will allow a user or agent to input data and retrieve predictions based on the model that was developed. An example of a program that will allow for this ability is Streamlit. To use this, a model needs to be created, a csv of the timeseries, and lastly the Streamlit server (Shiledarbaxi, 2021). This method

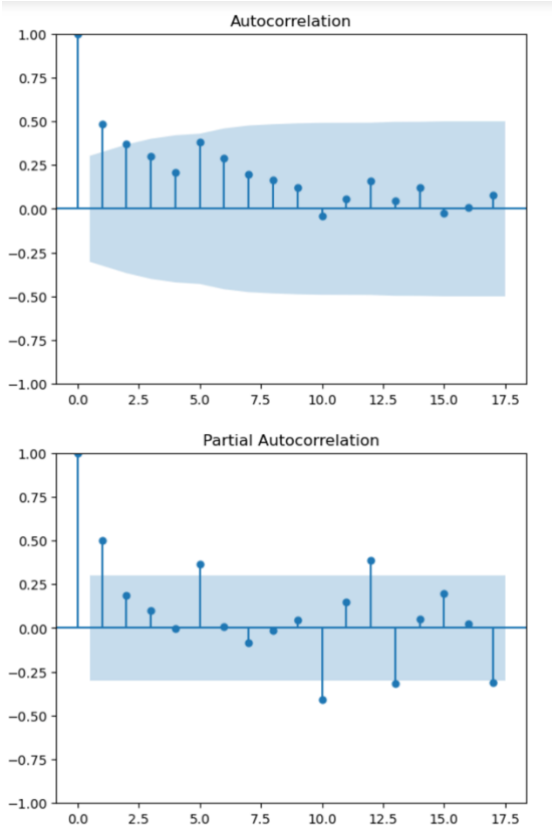
does not need knowledge of web application programming languages. The model when saved to a csv can be used by Streamlit to predict a certain number of days input by the user.

In this project, weather data was analyzed to understand which factors correlate and influence mean temp, precipitation, and snow depth. The main factors that resulted in a high correlation for price were months and temperature fields. The analysis also uncovered that precipitation was mainly influenced by cloud cover. Additionally, a model was created to predict mean temp, precipitation, and snow depth. The model predicting mean temp performed well while the other two still require more research.

Appendix A

This appendix includes the results and findings of the ARIMA models created. All the models are created using different p, d, and q variables. Checking the orders of differencing, autocorrelation and partial autocorrelation were used to find these values. The images below highlight the findings for the model created on mean-temp.

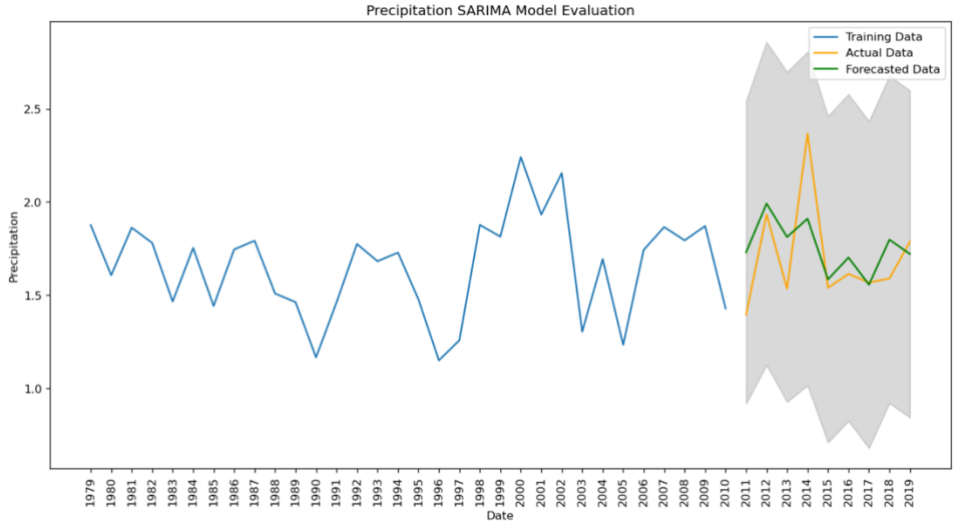




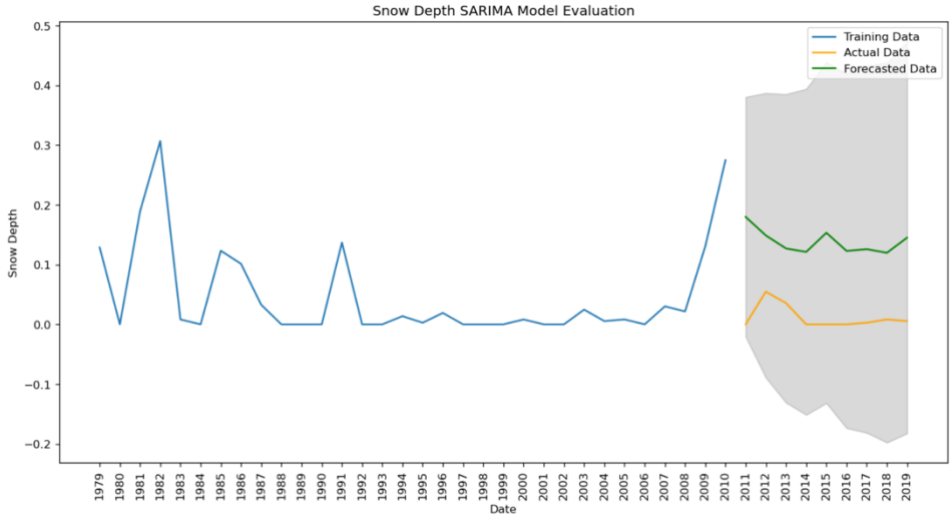
The first model performed well with a high R-Squared score of RMSE: 0.90. The models for precipitation and snow depth performed poorly with a score of .30 and .10. Because of this SARIMA models were conducted on these two variables.

Appendix B

This appendix includes the results and findings of the SARIMA models created for precipitation and snow depth. The same p, q and d values that were identified in the ARIMA model were used for the SARIMA model. For seasonal interval, 12 months was chosen. This model improved performance when adjusting for seasonality for snow depth but did not improve for precipitation.



RMSE: 0.22582083286525856



RMSE: 0.12910089674446118

Questions:

1. What can be done to improve precipitation and snow depth model accuracy?
2. Are there other models that can be explored?
3. How can this model for mean temperature be used?
4. What interval of time is being used for this model?
5. Would containing more features improve the performance in the model?
6. Can your team aid in model deployment?
7. What resources are needed for model deployment?
8. Is there any additional validation work needed for this?
9. Can you model based on daily data?
10. What needs to be done to maintain this model?

Reference:

Hayes, A. (2024, April 5). *Autoregressive integrated moving average (ARIMA) prediction model*. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Taylor, S. (2023, November 22). *R-squared*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/r-squared/