Reenie Christudass

DSC680-T301

I. **Introduction**

Predicting the average temperature in each area is critical in many industries, from agriculture to energy production. Accurate temperature forecasts can help businesses and governments make informed decisions about resource allocation, production planning, and disaster preparedness. This white paper presents an ARIMA model developed to predict the average temperature for the next three years in a specific geographic region. The model is based on historical temperature data and incorporates seasonal trends and long-term patterns to produce reliable temperature forecasts. This paper will describe the methodology used to develop and evaluate the ARIMA model and its potential applications and limitations.

The ARIMA (Auto Regressive Integrated Moving Average) model is a popular time series forecasting method used to predict future values based on past observations. The model consists of three components: the autoregressive (AR) component, the integrated (I) component, and the moving average (MA) component. The AR component models the relationship between the current value and the previous values, the MA component models the current value and the previous forecast errors, and the I component involves differencing the time series data to make it stationary. By combining these three components, the ARIMA model can capture both short-term and long-term trends in the data and produce accurate forecasts. The ARIMA model is widely used in various fields, including economics, finance, and climate science.

II. **Business Problem**

The task of predicting average temperature has a significant impact on a range of industries, including agriculture, energy production, and transportation. Accurate temperature forecasts can aid in resource allocation, crop planning, and energy demand forecasting. In this project, we aim to develop an ARIMA model to predict the average temperature for the next three years in a specific geographic region.

We used historical temperature data from the National Centers for Environmental Information (NCEI). The dataset includes monthly average temperature readings from a particular weather station. The dataset also includes meteorological variables such as wind speed, precipitation, fog, and thunderstorm occurrences. The ARIMA model was explicitly developed using historical data to predict the average monthly temperature.

The importance of accurate temperature prediction must be balanced. For example, temperature fluctuations can significantly impact crop yield and quality in the agricultural industry. Similarly, energy production is affected by temperature changes, with high temperatures leading to an increase in energy demand for cooling purposes. The transportation sector also relies on accurate temperature forecasts to anticipate disruptions caused by weather events such as snowstorms, heat waves, or hurricanes.

With accurate temperature forecasts, various industries can plan and make informed decisions. For example, farmers can use temperature forecasts in the agricultural industry to plan their planting and harvesting schedules, optimize irrigation and fertilizer usage, and reduce crop losses due to extreme temperature events. Energy producers can anticipate changes in energy demand caused by temperature fluctuations, enabling them to adjust their production schedules accordingly. The transportation sector can also use temperature forecasts to plan for potential disruptions caused by weather events, leading to better resource allocation and reduced downtime.

Overall, the ARIMA model can provide valuable insights into temperature patterns and help various industries make informed decisions, leading to improved efficiency, reduced waste, and better outcomes.

III. **Background/History**

The ARIMA model is a suitable solution for predicting average temperature because it can capture both short-term and long-term trends in the data. The model is designed to analyze and model time-series data, making it particularly useful for predicting temperature patterns over extended periods.

One advantage of the ARIMA model is that it can incorporate the effects of seasonal variations, such as temperature changes due to the summer or winter seasons. This makes it helpful in predicting temperature patterns over different seasons and identifying trends and patterns that seasonal variations may obscure.

Another advantage of the ARIMA model is that it can identify and remove data trends unrelated to the variable of interest, such as temperature changes caused by external factors like changes in data collection methods. This helps to ensure that the model only captures the temperature patterns relevant to the problem being solved.

Furthermore, the ARIMA model can identify and model non-linear relationships between the temperature data and other meteorological variables such as precipitation, wind speed, and thunderstorm occurrences. This makes it possible to produce more accurate and informative temperature forecasts.

Overall, the ARIMA model is a suitable solution for predicting average temperature due to its ability to capture seasonal variations, remove irrelevant trends, and identify non-linear relationships. The ARIMA model can help various industries make informed decisions by providing accurate and reliable temperature forecasts, leading to improved efficiency, reduced waste, and better outcomes.

Top of Form

Bottom of Form

IV. **Data Explanation**

The data for this project was obtained from the National Centers for Environmental Information (NCEI) at the National Oceanic and Atmospheric Administration (NOAA). The data set contains monthly average temperature readings from weather stations across the United States from 1895 to the present. For the current model I have built, I have taken the last 15 years of data. Each record in the data set includes the following variables:

1. STATION (11 characters) is the station identification code.
2. DATE is the year of the record (4 digits) followed by a month (2 digits). A dash separates year and month.
3. AWND = Monthly Average Wind Speed
4. DYFG = Total number of days in the month with fog reported.
5. DYTS = Total number of days in the month where one or more thunderstorms were reported
6. PRCP = Total Monthly Precipitation (Fahrenheit).
7. TAVG = Average Monthly Temperature (Fahrenheit).
8. TMAX = Monthly Maximum Temperature (Fahrenheit).
9. TMIN = – Monthly Minimum Temperature (Fahrenheit).

The data spans an extended period, making it suitable for studying long-term trends in temperature. The data set may contain missing or incomplete data due to various factors, such as missing or faulty temperature sensors or errors in data collection and recording.

Histogram

Description automatically generated with medium confidence

To address these issues, the data preprocessing step removes records with missing or flagged data and converts temperature readings to Celsius. This results in a clean and consistent data set that can be used for further analysis and modeling.

The missing values in the TAVG variable are substituted by using the average of the previous and next month's TAVG values as a substitute for missing TAVG values’ preserving the overall trend in temperature and avoiding introducing biases into your analysis. However, it's worth noting that this approach assumes that the temperature trend between adjacent months is relatively constant, which may only sometimes be the case.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedThe missing values in the DYTS variable with 0 is a reasonable approach, indicating that no

thunderstorms were recorded during the given month. Thunderstorms only occur for a few months, and it's common for some weather stations not to record any thunderstorm data during a given period. By using 0 as a substitute for missing values in DYTS, you are maintaining the integrity of the data and avoiding introducing any biases that might arise from using alternative values like mean, median, or mode. DYFG (Fog reported) and PRCP(Precipitation) can also be seen in a few months. The above approach was used for these variables also.Chart, bar chart

Description automatically generated

# **V. Methods**

* Stationarity Test: Before fitting an ARIMA model, it is essential to ensure that the time series data is stationary. This means that the statistical properties of the time series (e.g., mean, variance) should not change over time. I performed Augmented Dickey-Fuller (ADF) test on the original TAVG data and found that it was not stationary (p-value > 0.05).
* Differencing: To make the data stationary, performed seasonal differencing by subtracting the TAVG value at time t-12 (one year ago) from the TAVG value at time t. This helped to remove any seasonality patterns in the data. I did only seasonal differences because the weather data is seasonal, and trends are observed.
* Auto ARIMA: After differencing the data, auto\_arima functions automatically select the best ARIMA model based on the AIC (Akaike Information Criterion). This function tests different combinations of p, d, and q values and chooses the combination that minimizes the AIC. The best model was ARIMA (0,1,1) (2,1,0) [12]. The model includes a moving average order 1, two seasonal AR terms, and a seasonal difference of order 1.
* Train-Test Split: The data was split into a training set (all the data except the last 12 months) and a test set (the last 12 months). This allowed you to fit the model on the training data and evaluate its performance on unseen test data.

# **VI. Analysis**

* ARIMA Model evaluation was performed without exogenous variables: The performance of the ARIMA model on the training data was AIC 819.697, and RMSE (root mean squared error) was 17.09.
* Model Evaluation with exogenous variables like AWND, PRCP, DYTS, and DYFG: The performance of the ARIMA model on the training AIC was 819.697, and RMSE (root mean squared error) was 17.09.

The lower AIC indicates a better model fit, while a lower RMSE indicates better accuracy. But both the models gave the same score.

A picture containing chart

Description automatically generated

# **VII. Limitations**

The model is limited to only one weather station within Iowa.

# **VIII. Challenges**

The exogenous variables I was using may not be significant predictors of the dependent variable (A1C or RMSE), and hence they are not contributing to the model's performance. Re-evaluate the selection of your exogenous variables and consider using other variables more closely related to the outcome variable.

The exogenous variables you are using may need to be in the correct format or have suitable scaling. For example, if the units of measurement of the exogenous variables differ from the dependent variable, this may affect the model's performance. In this case, you may need to rescale or transform the variables to ensure they are comparable to the dependent variable.

# **X. Recommendations and Implementation Plan**

* Fairness and bias:
  + Recommendation: Conduct a comprehensive review of the data used to develop the model to ensure that it represents the full range of weather conditions experienced in Dubuque. Take steps to address any gaps or biases in the data.
  + Implementation steps: Hire a data scientist to review the data used in the model and conduct additional data collection as needed. Work with community stakeholders to identify any concerns or biases and address them in the model development process.
* Privacy and confidentiality:
  + Recommendation: Implement strong data security measures to protect the privacy and confidentiality of individuals in the community.
  + Implementation steps: Work with a cybersecurity expert to implement data security measures like encryption and secure storage. Develop a data use policy that clearly outlines how the data will be used and who will have access to it.
* Transparency and interpretability:
  + Recommendation: Ensure that the model's assumptions, limitations, and predictions are clear and transparent to stakeholders in the community.
  + Implementation steps: Develop a plain-language explanation of the model and its outputs and make this information available to the public. Hold community workshops and engagement sessions to solicit feedback from stakeholders and ensure that they understand the model's potential implications.
* Potential harms:
  + Recommendation: Conduct a risk assessment to identify potential negative impacts of the model's predictions and take steps to minimize these risks.
  + Implementation steps: Work with community stakeholders to identify potential harm from inaccurate or biased predictions. Develop a contingency plan to address these risks and ensure equitable and transparent decision-making around emergency preparedness and resource allocation.
* Stakeholder engagement:
  + Recommendation: Engage with community stakeholders throughout the model development and deployment process to ensure their perspectives and concerns are heard.
  + Implementation steps: Hold regular community meetings and engagement sessions to solicit stakeholder feedback. Develop a community advisory board to provide ongoing input into the model's development and deployment. Consider hosting training sessions for community members and decision-makers to ensure that they understand how to use the model effectively.

# **XI. Ethical Assessment**

* Fairness and bias: Depending on the data collected, the modeling targets only one weather station in Dubuque, Iowa. Other weather stations in Iowa are not included in the model. Also, are there any other weather stations in Dubuque not included? The model may not be accurate for all groups or ideal for just one station.
* Privacy and confidentiality: The weather data does not include personal information; it is still important to ensure it is protected and secure. This may involve preventing unauthorized access to the data or implementing data de-identification techniques to protect data authentication.
* Transparency and interpretability: It is essential to ensure that stakeholders in the community clearly explain and understand the model's assumptions, limitations, and predictions. There needs to be a process to provide plain-language explanations of the model and its outputs and solicit feedback from stakeholders to ensure they understand the model's potential implications.
* Potential harms: While it may seem unlikely that a weather prediction model could cause significant harm, it is still important to consider potential unintended consequences or negative impacts of the model's predictions. For example, if the model mispredicts a severe weather event, this could cause unnecessary panic or disruption in the community. Additionally, if the model were to contribute to decisions about resource allocation or emergency preparedness, it is essential to ensure that these decisions are made in a way that is fair and equitable for all members of the community.
* Stakeholder engagement: Given that the model is intended to serve the Dubuque community, it is essential to engage with stakeholders in the development and implementation of the model. This may involve seeking feedback from community members, local government officials, emergency responders, and others to ensure the model aligns with community needs and priorities.

# **XII. Conclusion**

In conclusion, the weather prediction model developed using data from a weather station in Dubuque, Iowa, has the potential to aid in weather prediction greatly. By leveraging advanced algorithms and machine learning techniques, this model can analyze historical weather data to identify patterns and trends that may inform future weather events. This can help meteorologists and other weather experts to make more accurate predictions about temperature, precipitation, wind patterns, and other weather-related factors.

The benefits of this model are many. For example, it can help improve the accuracy of weather forecasting, which can significantly impact industries such as agriculture, transportation, and tourism. It can also help emergency responders better prepare for severe weather events and take proactive steps to mitigate their impact.

Overall, the weather prediction model has the potential to significantly improve our understanding of weather patterns in Dubuque and provide valuable insights into how the weather may impact the community in the future. By leveraging this model alongside other weather forecasting tools and resources, meteorologists and other weather experts can make more informed decisions and better protect the community from the impacts of severe weather events.

# **XIII. References**

* Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
* Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
* Chatfield, C. (2019). The analysis of time series: an introduction. Chapman and Hall/CRC.
* Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer Science & Business Media.
* Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. Springer.
* Kang, S., & Kang, J. (2020). ARIMA and Regression Model-based Daily PM2. 5 Concentration Forecasting. Journal of Korean Society of Environmental Engineers, 42(10), 553-560.
* Akhtari, S., & Bouchard, M. (2019). Short-term traffic volume forecasting using ARIMA models with exogenous variables. Journal of Transportation Engineering, Part A: Systems, 145(11), 04019052.
* Wu, Y., Song, S., Huang, Y., & Li, Y. (2020). Research on the forecasting model of atmospheric visibility based on ARIMA and exogenous variables. Journal of Cleaner Production, 244, 118714.
* Tewari, A., Singh, A., & Arora, N. (2020). A comparative study of time series forecasting techniques for short term solar radiation prediction. Renewable Energy, 145, 139-149.
* Tian, Y., Chen, Y., & Wei, Y. (2019). Using time series models to analyze and forecast weather data: A case study of Hangzhou, China. Environmental Science and Pollution Research, 26(19), 19309-19320.
* Chen, C. H., & Chiu, Y. H. (2018). Development of an efficient machine learning model for accurate prediction of typhoon inundation areas. Water, 10(9), 1237.

<https://otexts.com/fpp2/>

<https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021>

<https://www.crcpress.com/The-Analysis-of-Time-Series-An-Introduction-7th-Edition/Chatfield/p/book/9781138713548>

<https://www.springer.com/gp/book/9783540401723>

<https://www.springer.com/gp/book/9783319298528>

<https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09820894&language=en>

<https://ascelibrary.org/doi/abs/10.1061/JTEPBS.0000336>

<https://www.sciencedirect.com/science/article/abs/pii/S0959652620352518>

<https://www.sciencedirect.com/science/article/pii/S096014811931335X>

<https://link.springer.com/article/10.1007/s11356-019-05319-2>

<https://www.mdpi.com/2073-4441/10/9/1237>