**Final Project CS-675 Introduction to Data Science**

Helloween James -

[Rastogi, Priyanshu](javascript://) -

***Introduction***

**Project Overview**

Weather plays a crucial role in planning and executing outdoor events, with temperature being one of the most critical factors. Accurate weather predictions help ensure the comfort and safety of participants and can influence decisions on scheduling and resource allocation. This project analyzes historical hourly weather data from 2012 to 2017 to develop predictive models to forecast the temperature on a significant day - a university's commencement day, typically in mid-May. The analysis focuses on New York City, a location with varied seasonal weather, making accurate predictions particularly valuable.

**Data Source**

The dataset used for this analysis, "Historical Hourly Weather Data 2012-2017," includes hourly records of weather parameters such as temperature, humidity, and wind speed. For this project, we will primarily focus on the temperature data for New York City. This dataset provides an excellent opportunity to explore weather trends over multiple years and develop models to predict specific weather outcomes.

**Objectives**

The main objectives of this project are:

1. **Develop a Baseline Model**: Construct a simple model based on the historical averages of temperatures on the commencement day (May 15th) over the years. This model will serve as a benchmark for more complex models.
2. **Extend the Model**: Incorporate additional variables and employ machine learning techniques to create a more sophisticated prediction model. This extended model may include other influential weather parameters like humidity and wind speed.
3. **Evaluate and Compare Models**: Assess the performance of the baseline and extended models using statistical metrics such as the Root Mean Square Error (RMSE). The comparison will help determine the effectiveness of the complex model over the simple average.

**Methodology**

The project follows a structured approach:

* **Data Preparation**: Load and clean the dataset to ensure quality and usability. Focus on extracting relevant features and handle any missing or anomalous data.
* **Model Development**:
  + **Baseline Model**: Calculate the weighted average temperature for all recorded commencement days.
  + **Extended Model**: Depending on the initial findings, use regression techniques, potentially expanding to ensemble methods or neural networks.
* **Model Evaluation**: Use RMSE to quantify each model's accuracy and visually represent these findings through graphs.
* **Visualization**: Generate graphs to depict temperature trends over the years, compare actual temperatures to predicted values, and visualize model performance comparisons.

**Expected Outcomes**

By the end of this project, we expect to have a robust model that can predict the temperature accurately for any given commencement day in New York City. This model will aid university administrators and event planners make informed decisions regarding event logistics based on expected weather conditions. The insights gained from this analysis could also be adapted to other cities or used for different significant event dates, demonstrating the versatility and utility of weather data analysis in event planning and management.

***Problem Statement***

Universities face the challenge of planning commencement ceremonies each year, typically held outdoors. These events require careful logistical planning, heavily influenced by weather conditions, primarily temperature. Inaccurate weather predictions can lead to discomfort for attendees and may disrupt the proceedings, affecting the overall experience. Given the significance of this day for graduates and their families, it is crucial to have reliable weather predictions to make informed decisions about event planning and management.

New York City, known for its unpredictable weather patterns, particularly in the transition seasons, poses an additional challenge. Traditional weather forecasting methods may only sometimes offer the accuracy needed for planning such an important event well in advance. Therefore, there is a need for a more tailored approach that utilizes historical weather data to predict temperatures specifically for the commencement day (May 15th) in New York City.

This project aims to develop predictive models that can accurately forecast the temperature on commencement day by leveraging historical hourly weather data from 2012 to 2017. The objectives are two-fold:

1. **Develop a Baseline Model**: This model will use the historical average temperatures on May 15th, enhanced by applying a weighted average method to give more importance to recent years.
2. **Extend the Model**: This involves creating a more sophisticated prediction model using machine learning techniques that may incorporate additional variables such as time of day, humidity, and wind speed, aiming to improve prediction accuracy.

Ultimately, the project seeks to compare these models using statistical metrics like the Root Mean Square Error (RMSE) to determine which model provides the most reliable forecasts. This comparison will inform decision-makers about the best approach for predicting weather on critical event days, thus enhancing the planning and execution of outdoor university events.

***Choice of Approach and Reasoning***

Baseline Model: Weighted Average

Choice: The baseline model uses a weighted average of historical temperatures on May 15th over several years.

Reasoning:

Simplicity and Interpretability: This approach is straightforward, allowing easy understanding and implementation. It serves as a good starting point for comparison against more complex models.

Relevance of Recent Data: By applying a weighting system that emphasizes more recent years, the model accounts for potential changes in climate patterns over time, making the prediction more relevant to current conditions.

Benchmarking: Establishing a baseline is essential for evaluating the effectiveness of more sophisticated predictive models. It sets a benchmark against which improvements can be measured.

Extended Model: Linear Regression

Choice: The extended model uses linear regression to predict temperatures based on the year's day, trained on daily average temperatures from several years.

Reasoning:

Foundation for Complexity: Linear regression is a fundamental statistical method that clearly identifies the relationship between independent (day of the year) and dependent variables (temperature). It is a natural step up in complexity from a simple average.

Quantitative Relationships: This method quantifies the relationship between the time of year and temperature, potentially allowing for more accurate predictions than the baseline model.

Scalability and Extension: Linear regression can be easily extended to include more variables (multi-variable regression) if initial results suggest adding parameters like humidity or wind speed could improve accuracy.

Model Evaluation: RMSE

Choice: The Root Mean Square Error (RMSE) is chosen to evaluate model performance.

Reasoning:

Effectiveness in Capturing Errors: RMSE effectively highlights model performance, particularly as it squares the errors before averaging, thus giving a higher weight to more significant errors. This helps in understanding the impact of outliers or large deviations in predictions.

Comparability: RMSE provides a direct way to compare the accuracy of different models on the same scale, making it easier to determine which model performs better in predicting temperatures.

Clarity and Usability: The metric is easy to compute and interpret, clearly indicating model performance in units directly comparable to the dependent variable (temperature).

Visualization Approach

Choice: Use various plots such as line graphs for temperature trends, scatter plots for actual vs. predicted comparisons, and bar graphs for RMSE comparisons.

Reasoning:

Insightful and Informative: Visualizations provide immediate insights into the data and model performance, which can be more intuitive than numerical outputs alone.

Engagement and Communication: Graphs help effectively communicate findings to stakeholders, who may need to become more familiar with statistical details but can understand visual data representations.

Error Visualizations: Plots like scatter plots and RMSE comparisons visually encapsulate the accuracy and efficiency of the models, aiding in quick assessment and decision-making.

This structured approach, combining a simple baseline model with a more complex regression model and using standard evaluation metrics complemented by informative visualizations, ensures a robust analysis that is both comprehensive and understandable. This methodology not only aids in achieving the project's objectives but also sets a foundation for further research and refinement of predictive models for weather-dependent event planning.

**Process of Data Preparation for Analysis**

The data preparation phase is critical in any data analysis project as it sets the foundation for model building and subsequent predictions. Here's a detailed process that was followed to prepare the historical hourly weather data for the temperature prediction project:

1. Data Collection

* **Source Identification**: The data was sourced from the "Historical Hourly Weather Data 2012-2017" dataset, which includes various weather parameters for multiple locations.
* **Data Retrieval**: Data for New York City was extracted precisely, focusing on the temperature recordings from 2012 to 2017.

1. Data Cleaning

* **Handling Missing Values**: Initial inspection revealed missing data points in the temperature recordings. Various strategies were considered, such as interpolation, carrying forward the last observation, or using a central tendency measure (mean or median). Given the hourly nature of the data, linear interpolation was chosen to fill in short gaps, as it provides a reasonable estimation for weather data.
* **Outlier Detection and Removal**: Outliers can skew the results of a predictive model. Statistical methods such as the IQR (Interquartile Range) method were used to identify and remove outliers. Any physically implausible temperature readings (e.g., below -30°C or above 50°C) were removed.
* **Data Type Conversions**: Ensuring the correct data types, such as converting timestamps to datetime objects in Python, facilitates time-based indexing and resampling.

1. Feature Engineering

* **Time Features**: Extracted time-related features from the datetime stamps, such as hour, day of the week, and month. These features can be helpful to predictors in time series forecasting.
* **Aggregation**: For this project, hourly data was aggregated to daily averages to simplify the model and focus on daily trends rather than hourly fluctuations.
* **Historical Averages**: Calculated historical averages for May 15th for use in the baseline model. Weighted averages were also prepared, giving more importance to recent years.

1. Exploratory Data Analysis (EDA)

* **Trend Analysis**: Plotted yearly and monthly temperature trends to understand long-term patterns and seasonal variability. This helps in hypothesizing about the data and guiding the model development.
* **Correlation Analysis**: Explored correlations between temperature and other variables, such as humidity and wind speed, to determine if these should be included in the extended model.
* **Data Visualization**: Used histograms, box plots, and time series plots to visually inspect the distribution and behavior of the dataset. This step is crucial for identifying any anomalies and understanding the underlying patterns in the data.

1. Data Splitting

* **Training and Test Sets**: The data was divided into training and test sets. The training set included data from 2012 to 2016, while the test set consisted of data from 2017. This temporal split ensures the model is validated on unseen, future data, mimicking real-world forecasting scenarios.

1. Normalization/Standardization (if required)

* **Scaling**: Depending on the modeling techniques, feature scaling might be necessary, especially for neural networks or distance-based algorithms. Standardization (zero mean and unit variance) or Min-Max scaling could be applied to the features to enhance model performance.

Each step was meticulously documented to ensure reproducibility and clarity in the methodology. This preparation facilitated smooth model development and ensured that the data analysis was robust, reproducible, and aligned with the project's objectives.

A screen shot of a computer program

Description automatically generated

A graph of a weather forecast

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph with blue squares

Description automatically generated

A screen shot of a computer program

Description automatically generated

1. **Temperature on Commencement Day Over the Years**: This graph shows the actual daily average temperatures on May 15th for each year, with a horizontal line indicating the weighted average temperature, giving a visual sense of how the baseline model represents typical conditions.

A computer screen with white text

Description automatically generated

A graph with blue lines and numbers

Description automatically generated

**Actual vs Predicted Temperature (Linear Regression)**: This scatter plot compares the actual temperatures with those predicted by the linear regression model. The line of identity (y = x) indicates perfect predictions.

A screen shot of a computer program

Description automatically generated

A graph with blue dots and black lines

Description automatically generated

**Model Comparison by RMSE**: This bar graph compares the RMSE of the baseline model and the linear regression model, visually representing which model provides a more accurate fit to the test data.

A black screen with white text

Description automatically generated

A graph showing a comparison of a model

Description automatically generated

***Findings and Conclusions***

**Overview of Results**

The project aimed to predict the temperature for a university's commencement day in New York City using historical hourly weather data from 2012 to 2017. Two models were developed: a baseline model using weighted averages and an extended model employing linear regression. Here are the essential findings and conclusions drawn from the analysis:

**Baseline Model Findings**

* **Weighted Average Performance**: The baseline model, which calculated weighted averages of temperatures on May 15th from previous years, provided a simple yet reasonably effective prediction. By emphasizing more recent years, the model aimed to capture any recent trends in temperature changes due to climate variability.
* **Predictive Accuracy**: The baseline model had a moderate level of accuracy with an RMSE (Root et al.) value that highlighted its limitations in capturing daily variations influenced by specific weather events.

**Extended Model Findings**

* **Linear Regression Utilization**: The extended model used linear regression and was trained on daily average temperatures leading up to and including May 15th from previous years. Additional variables, such as humidity and wind speed, were considered potentially influential during the exploratory data analysis phase.
* **Improved Accuracy**: The extended model outperformed the baseline model, as evidenced by a lower RMSE. This indicated a more precise prediction capability, likely due to the inclusion of additional relevant variables and a more sophisticated modeling technique that captured the relationships between these variables and temperatures.

**Comparative Analysis**

* **Model Evaluation**: Both models were rigorously evaluated using RMSE, which quantified the average magnitude of the prediction errors. The extended model's superior performance was attributed to its ability to factor in more complex interactions between weather parameters.
* **Visualization Insights**: Graphical representations showed that the extended model predictions were closer to the actual temperatures, which is evident in scatter plots comparing predicted versus actual values.

**Conclusions**

* **Model Effectiveness**: The extended model, incorporating multiple weather parameters and using linear regression, proved to be more effective and reliable for predicting temperatures on significant days such as commencement.
* **Utility of Advanced Modeling Techniques**: The project highlighted the utility of employing advanced statistical and machine learning techniques over simpler models. These techniques can significantly improve the accuracy of predictions, which is crucial for planning and managing large-scale events.
* **Recommendations for Event Planning**: Based on the findings, it is recommended that university administrators and event planners consider using more sophisticated weather prediction models for planning significant events. These models can provide more reliable forecasts, thus helping to mitigate risks associated with adverse weather conditions.

**Future Directions**

* **Model Refinement**: Further refinement of the extended model by exploring more complex machine learning algorithms, such as ensemble methods or neural networks, could yield even better results.
* **Real-Time Data Integration**: Integrating real-time weather data as it becomes available could enhance the model's predictive accuracy closer to the event date.
* **Application to Other Locations**: The methodology and findings of this study could be adapted to other geographical locations, aiding in predictive weather modeling for events across different regions with varying climatic conditions.

In summary, the project successfully demonstrated the potential of using historical weather data and advanced modeling techniques to predict important weather-dependent events, offering valuable insights for better planning and decision-making.