Final Report

Topic:

Predicting Rainfall in London-Heathrow Using Machine Learning on 2000-2010 Meteorological Data

Business Problem:

In this project, my goal is to tackle a significant business challenge faced by sectors that rely on accurate weather forecasting, such as aviation, logistics, and outdoor event planning at Heathrow. The primary problem I aim to solve is improving the precision of rain predictions for Heathrow, a crucial international airport where even minor weather miscalculations can lead to flight delays, disruptions, and significant financial losses. For airlines and airport management, accurate forecasts of rain are essential for optimizing flight schedules, ensuring passenger safety, and reducing operational inefficiencies. The research questions driving my project include: How can historical weather patterns be leveraged to enhance rain prediction accuracy for Heathrow? What key meteorological factors contribute most to predicting rain in this specific location? By addressing these questions, I hope to create a robust predictive model that helps stakeholders anticipate rainy conditions more effectively, enabling proactive measures that mitigate disruptions and improve customer satisfaction. This approach will also contribute to more strategic decision-making and resource allocation, potentially reducing costs and enhancing operational resilience in the face of weather unpredictability.

Background/History:

Weather forecasting has a long and fascinating history that dates back thousands of years, beginning with the Babylonians around 650 BCE, who studied cloud patterns to predict weather changes. The field evolved significantly over time, especially with the inventions of the thermometer and barometer. In the 19th century, British Vice Admiral Robert FitzRoy advanced the practice by pioneering daily weather forecasts and issuing storm warnings, laying the groundwork for modern meteorology. By the 20th century, further progress came with the development of mathematical models and the use of computers to predict atmospheric behavior, allowing forecasts to become more accurate and reliable.

Weather forecasting is essential because it helps us prepare for and respond to the changing conditions of our environment. Accurate forecasts support daily activities, improve safety by predicting severe weather, assist farmers with agricultural planning, and guide various industries to optimize operations. Overall, having insight into upcoming weather allows individuals and communities to make informed decisions and enhance their overall resilience against nature’s unpredictability.

Data Explanation (Data Prep/Data Dictionary/etc):

Final Report

For my project, I am using a rich dataset I obtained from Kaggle, which was originally sourced from the European Climate Assessment & Dataset (ECA&D). The ECA&D project is known for providing comprehensive daily weather observations from meteorological stations spread across Europe and the Mediterranean. My dataset specifically covers data from 18 European locations, including cities like Basel, Budapest, Munich, and crucially for my project, Heathrow in the UK. This dataset spans the decade from 2000 to 2010, capturing 3,654 days of observations that include a diverse range of weather variables. Common data points in the dataset include 'mean temperature,' 'max temperature,' and 'min temperature,' while other variables such as 'cloud cover,' 'wind speed,' 'humidity,' 'pressure,' 'global radiation,' 'precipitation,' and 'sunshine' were included wherever available. The data was cleaned by the author to ensure quality: columns with more than 5% missing values were removed, while those with up to 5% missing values had their gaps filled using mean values. This meant that I had zero data cleaning to do on my part.

Methods:

The precipitation column had continuous values in it and I needed the column to have discrete values for me to be able to run the logistic model and the random forest model. So, I created a new column in my dataset that will indicate whether it rained or not, based on the amount of precipitation. If the precipitation is 0, it means there was no rain, and if it's greater than 0, it means it rained. After creating this column, I deleted the "Heathrow_precipitation" column. I then used logistic regression and random forest models to predict the chances of rain. To measure how well the models are performing, I used several metrics: accuracy (how often the models are correct), precision (how many times the model correctly predicted rain), recall (how many actual rain events the model was able to identify), and F1 score (a balanced measure of precision and recall). Lastly, I created a bar graph that showed the feature importance level of each feauture.

Analysis:

When I analyzed the results, I found that the Random Forest model performed better than the Logistic Regression model in predicting rainfall at Heathrow. The Random Forest model achieved an accuracy of about 78%, compared to around 73% for the Logistic Regression. This means that the Random Forest was correct about 78 out of 100 times, while the Logistic Regression was accurate 73 out of 100 times. The confusion matrices tell me that the Random Forest model was more balanced in correctly identifying both rain and no-rain days, while the Logistic Regression had more difficulty predicting no-rain days accurately. Both models showed decent precision and recall, but the Random Forest had a slight edge, producing higher F1 scores, indicating better overall performance.

Looking at the feature importance from the Random Forest model, it's clear that 'pressure' and 'humidity' were the most critical factors in predicting rain, followed by 'sunshine' and 'global radiation.' This suggests that atmospheric pressure and humidity levels play the biggest roles when

forecasting rain at Heathrow. Understanding these important features helps clarify which weather patterns most influence rainfall predictions, giving stakeholders more reliable information for decision-making.

Conclusion:

In conclusion, this project successfully demonstrated that machine learning can be a powerful tool for predicting rainfall at Heathrow, an airport where accurate weather forecasts are critical for operational efficiency and safety. By using data from 2000 to 2010 and applying logistic regression and random forest models, it became clear that the random forest model was more effective, achieving a prediction accuracy of 78%. This level of accuracy is significant as it helps airport management and related sectors better anticipate rain, allowing them to make informed decisions and proactive adjustments to flight schedules and logistics. The analysis highlighted that 'pressure' and 'humidity' were the most influential features in forecasting rain, which aligns with our understanding of meteorological patterns. Overall, the project showed that with accurate data and proper model training, stakeholders can benefit from more dependable rain prediction models that enhance planning and minimize disruptions.

Assumptions:

Throughout the project, several assumptions were made to simplify the modeling process. One primary assumption was that the cleaned dataset obtained from Kaggle was representative of real-world weather conditions and did not require further data cleaning. It was also assumed that using daily weather data was sufficient for predicting rainfall, even though more granular hourly data could potentially yield different results. Additionally, it was presumed that weather patterns over the decade from 2000 to 2010 would be consistent enough to allow for the training of models that can generalize well for similar future scenarios. The creation of a binary rain indicator column assumed that any non-zero precipitation level, regardless of amount, would count as rain, simplifying the target variable for classification.

Limitations:

Despite its strengths, this project had some limitations. One major limitation was the decade-long time frame of the dataset, which may not fully represent long-term climate changes or anomalies beyond 2010. The analysis also only considered Heathrow's data without comparing results to other nearby airports or broader geographical regions. The models used daily weather observations, which may not capture the variability within a day, potentially impacting prediction accuracy. Furthermore, the project only used logistic regression and random forest models; more advanced models like gradient boosting or deep learning could have been explored for potentially

higher accuracy. Finally, the project was limited by the fact that only certain meteorological features were available, excluding potentially valuable data such as wind direction or more complex atmospheric readings.

Challenges:

During the project, several challenges were encountered. One of the initial challenges was understanding and processing the dataset to ensure it was well-prepared for machine learning models. Although the dataset was pre-cleaned, verifying its quality and structure took significant time. Another challenge was fine-tuning the logistic regression and random forest models to achieve the best performance, as different parameter adjustments yielded varying results. Interpreting the importance of each feature and understanding how they contributed to the prediction added another layer of complexity. Finally, analyzing the results required careful evaluation of the trade-offs between precision, recall, and accuracy to choose the best model for practical use.

Future Uses/Additional Applications:

The insights and models developed in this project could have further applications beyond Heathrow and aviation. For instance, similar predictive models could be adapted for other airports or transportation hubs where weather plays a crucial role in scheduling. Additionally, industries such as outdoor event planning, agriculture, and emergency response could leverage these models to predict rain and plan accordingly. The methodology could be expanded to include different types of weather predictions, such as snow or storms, enhancing overall readiness and safety measures in various sectors. With more comprehensive datasets, the models could also support long-term climate analysis, helping businesses and local governments plan for changes in weather patterns over time.

Recommendations:

Moving forward, a few key recommendations would improve the project's scope and effectiveness. Expanding the dataset to include more recent years and possibly integrating data from multiple weather stations would help create a more robust and current model. Exploring more complex models like gradient boosting machines or even deep learning approaches could offer better performance. Additionally, incorporating more meteorological variables, such as wind direction or seasonal trends, could provide deeper insights into weather patterns and improve prediction accuracy. Finally, considering different thresholds for defining rain (e.g., light versus heavy rain) might make the predictions more practical for specific applications.

Final Report

Implementation Plan:

To implement this predictive model in a real-world scenario, the process would begin with integrating the model into an airport's weather monitoring system. This would involve setting up an automated data pipeline that continuously feeds updated meteorological data into the model. The output would then be used to inform flight operations and ground staff, potentially through a dashboard that displays predictions and the likelihood of rain. Regular updates and model retraining would be essential to maintain its accuracy as more recent weather data becomes available. Collaborations with meteorologists and data scientists would also be needed to interpret and validate the model's predictions, ensuring that decision-makers are confident in its outputs.

Ethical Considerations:

When working on a project like this, I can't ignore the potential ethical concerns that might come up when analyzing weather data and predicting outcomes such as rain. Keeping in mind that this is just a project, if the results of my work were meant for use in the real world, there'd be certain concerns. One major concern is the reliance and responsibility that people or businesses might place on my predictions. If my model isn't highly accurate or is used beyond its intended purpose, there could be serious real-world consequences, like disrupted flight operations or economic losses for companies relying on my forecasts. This brings up the issue of accountability—who's responsible if a prediction based on my model leads to unexpected issues? Additionally, there is the challenge of transparency. Users of my predictive tool need to understand how these forecasts are generated, which means I have to make sure that my methods and the limitations of the model are clearly communicated.

Another concern is bias in the dataset or analysis. If the historical data has any inconsistencies or if it disproportionately represents certain weather patterns, this could skew the results and create unfair or unreliable forecasts. Lastly, as with any project that uses large datasets, there are potential privacy and data usage considerations. While weather data isn't sensitive in the way personal data is, ensuring that I'm complying with data-sharing agreements and using the data responsibly is still an important ethical aspect.

References:

- Kaggle. (n.d.). *Weather prediction dataset*. Retrieved October 28, 2024, from https://www.kaggle.com/datasets/thedevastator/weather-prediction/data?select=metadata.txt
- Klein Tank, A. M. G., & Coauthors. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology, 22*(12), 1441–1453. https://doi.org/10.1002/joc.773

Final Report

- Huber, F., van Kuppevelt, D., Steinbach, P., Sauze, C., Liu, Y., & Weel, B. (n.d.). *Will the sun shine? – An accessible dataset for teaching machine learning and deep learning*. DOI TO BE ADDED!
- Gaitán, C. F. (2018, May 20). *Only happy when it rains: A brief history of weather forecasting*. Arable. https://www.arable.com/blog/only-happy-when-it-rains-a-brief-history-of-weather-forecasting/

Illustration:

Logistic Regression Evaluation

```
Accuracy: 0.7250341997264022
Confusion Matrix:
[[242 124]
 [ 77 288]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.66      0.71       366
           1       0.70      0.79      0.74       365

    accuracy                           0.73       731
   macro avg       0.73      0.73      0.72       731
weighted avg       0.73      0.73      0.72       731
```

Random Forest Classifier Evaluation

```
Accuracy: 0.7811217510259918
Confusion Matrix:
[[286  80]
 [ 80 285]]
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.78      0.78       366
           1       0.78      0.78      0.78       365

    accuracy                           0.78       731
   macro avg       0.78      0.78      0.78       731
weighted avg       0.78      0.78      0.78       731
```

Final Report

Feature Importance		
Rank	**Feature**	**Importance**
1	HEATHROW_pressure	0.237404
2	HEATHROW_humidity	0.184932
3	HEATHROW_sunshine	0.111176
4	HEATHROW_global_radiation	0.103171
5	HEATHROW_temp_max	0.099372
6	HEATHROW_temp_min	0.089204
7	HEATHROW_temp_mean	0.083304
8	HEATHROW_cloud_cover	0.05172
9	MONTH	0.039718