IS525

Analysis of New York City Weather Data

YaoYuan Sara Kiel

Problem Statement

The goal of this project is to integrate New York City's historical weather data, including variables such as temperature, humidity, wind speed, precipitation, and weather conditions, by building a structured data mart (Data Mart). We hope to provide city managers and researchers with a centralized and unified data platform to facilitate the analysis of the impact of weather on urban life (such as traffic, air quality, etc.).

Intended Audiences

1. Meteorological researchers and analysts: They need accurate humidity forecasts for weather forecasting and data analysis.
2. Agricultural producers and decision makers: Arrange irrigation and crop management based on humidity forecasts.
3. Urban management departments: Take preventive measures when air humidity is high to reduce traffic accidents on slippery roads.
4. The public and community organizations: Know weather changes in advance and make corresponding life arrangements.

Data Source

The weather data used in this project comes from historical meteorological observation records on NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION GLOBAL SURFACE SUMMARY OF DAY DATA (GSOD), including temperature, humidity, wind speed, visibility and other features. After data cleaning and preprocessing, it is used for model training and prediction. Since this is the most commonly used data attribute, it can also be easily generalized in the face of more professional enterprise-level data.

Client Details

Client Name and Contact Information

The client of this project is Feng Jing, acting Data Scientist of the Data Analysis Department of the Municipal Meteorological Bureau.

Contact Information: pefeng@ium.cn or (13816110833@139.com, personal)

Client Requests

Provide a visual instrument that displays weather data according to time changes

Use current data to predict the possibility of future rain and make a visual reliability and result analysis of the model performance.

Data Cleaning, Feature Correlation, and Exploratory Visualizations:

To prepare a unified dataset for weather analysis and ensure the accuracy and reliability of weather data, we conducted a comprehensive data cleaning process. Since some data had statistical requirements and the customer needed to further analyze the performance of different machine learning models on the data set, it was necessary to use Python to process the data. The steps taken are outlined below. But for the final visualization tasks, we used Tableau to present the processed data results and the model performance.

Data Cleaning

1. **Initial Data Inspection and Column Removal:** The primary dataset was loaded in and irrelevant metadata columns were removed, such as station ID (all of the data came from one NYC station).
2. **Feature Engineering - Data and Season**: The *DATE* column was converted to datetime formatted and new features were derived from it: Month and Season, with Season later mapped to season labels. This allows for potential seasonal trend analysis and grouped comparisons across time.

3. **Handling Missing Values:** GSOD-specific missing value codes (for example, 9999.9 and 99.99) were replaced with NaN across all applicable weather variables. This ensured accurate missing value handling for later modeling and visualization.
4. **Humidity Data Acquisition and Cleaning:** The primary dataset did not contain a key variable we wanted to analyze - humidity. A separate historical weather dataset containing humidity records, also sourced from NOAA, was retrieved to solve this problem. Missing values were filtered out, and the humidity column was concatenated with the primary dataset.

Feature Correlation:

1. We computed a Pearson correlation matrix between temperature, humidity, windspeed, and visibility, and visualized it using Seaborn. We wanted to highlight the relationship between variables to inform future predictive modeling choices.
2. We also computed pairwise Pearson correlation and p-values across weather variables and over time, filtering for statistically significant relationships. These results would similarly inform our predictive modeling in the future.

Exploratory Visualizations:

1. While most visualizations were done in Tableau with the finalized and cleaned dataset, a quick visualization of each key weather variable - temperature, humidity, windspeed, and visibility - over time was done with python. This was done for an initial view of potential seasonal trends and anomalies.

```
# addtl time features
df_final['DATE'] = pd.to_datetime(df_final['DATE'], errors='coerce')

df_final['Month'] = df_final['DATE'].dt.month
df_final['Season'] = df_final['DATE'].dt.month % 12 // 3 + 1  # Winter = 1, Spring = 2, Summer = 3, Fall = 4

season_map = {
    1: 'Winter',
    2: 'Spring',
    3: 'Summer',
    4: 'Fall'
}
df_final['Season_Label'] = df_final['Season'].map(season_map)

df_final.drop(columns=['Season'], inplace=True)
df_final.rename(columns={'Season_Label': 'Season'}, inplace=True)
```

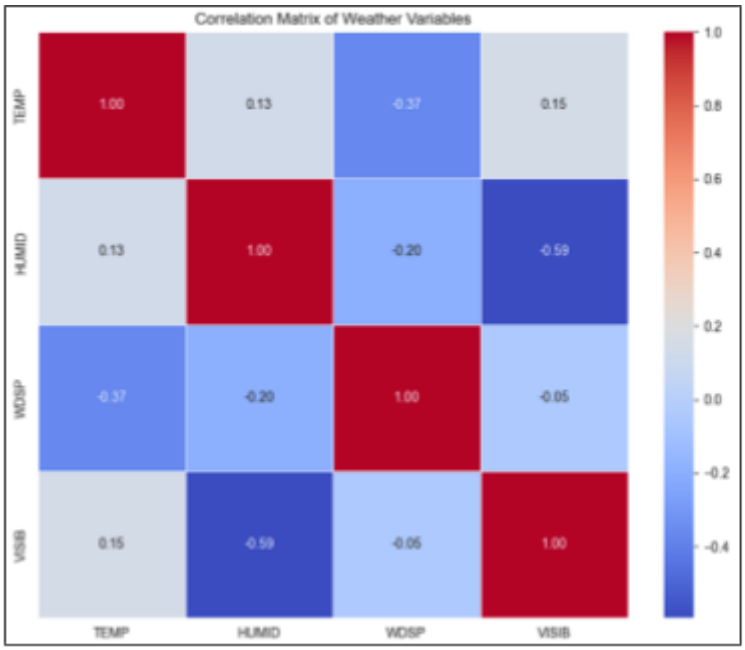
Converting the date column format and deriving season and month from it.

```python
import numpy as np

missing_replacements = {
    'TEMP': 9999.9,
    'DEWP': 9999.9,
    'SLP': 9999.9,
    'STP': 9999.9,
    'VISIB': 999.9,
    'WDSP': 999.9,
    'MXSPD': 999.0,
    'GUST': 999.9,
    'MAX': 9999.9,
    'MIN': 9999.9,
    'PRCP': 99.99,
    'SNDP': 999.9
}

df_final.replace(missing_replacements, np.nan, inplace=True)
```

Missing value handling.



Heatmap created to visualize feature correlation.

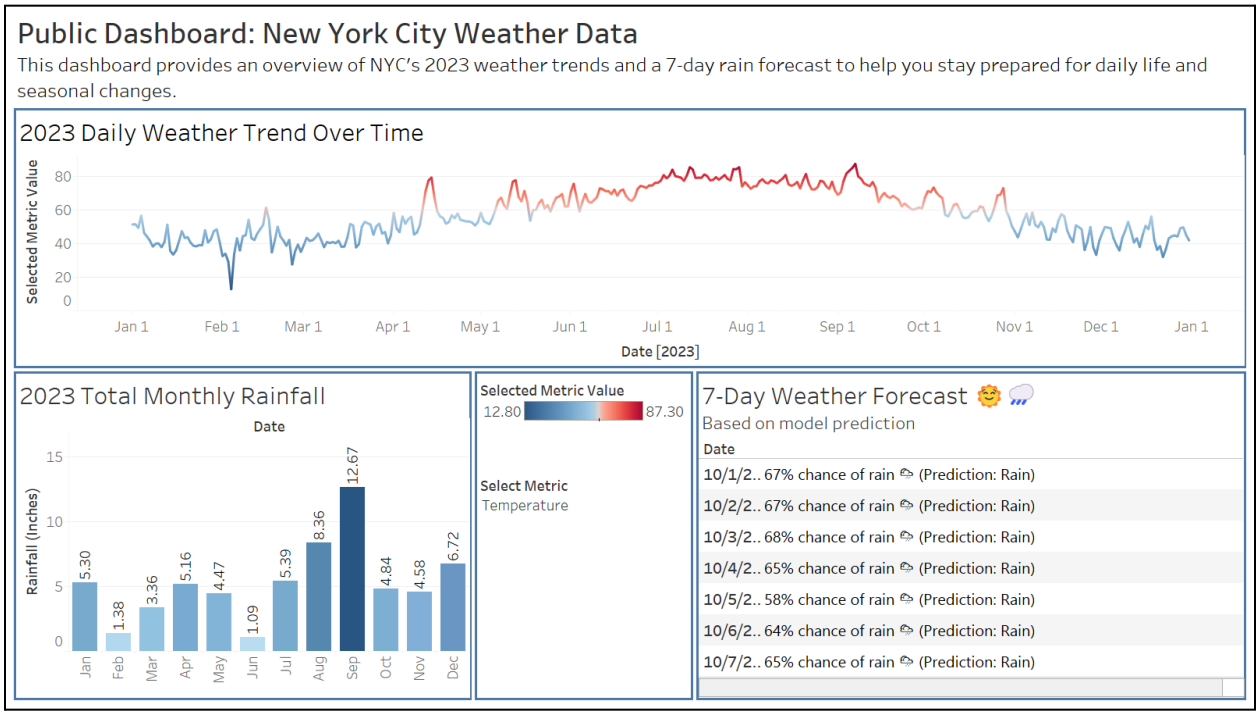
Descriptive Dashboards

We used Tableau to build two descriptive dashboards:

1. One for a public audience of average citizens and community organizers to show the 2023 weather trends in New York City and rain forecast for the next seven days;
2. And a second one for a more informed audience of city researchers, managers, and decision markers that might solely be interested in predicting weather in advance.

Public Dashboard

The public-facing dashboard provides an overview of NYC's 2023 weather trends and a seven day rain forecast to help the average user stay prepared for daily life and seasonal changes. This dashboard includes an interactive line chart that tracks key weather variables (temperature, humidity, windspeed, and visibility) and their changes across time in 2023. A total monthly rainfall count in the form of a bar chart is present as well, in addition to the rain forecast, which is based on model prediction.



Expert Dashboard

The client asked us to analyze the prediction ability of LSTM for this dataset model. We used double discount visualization to visually present the prediction accuracy of the model and technically used different residual analysis for attributes and target values to further demonstrate the model training results of LSTM.

Due to the size of the data, we use a prediction method that predicts the next day every 10 days. As for the selection of the entire model, we choose the neural network model. Among the neural network models, LSTM is the best especially for time series prediction models.

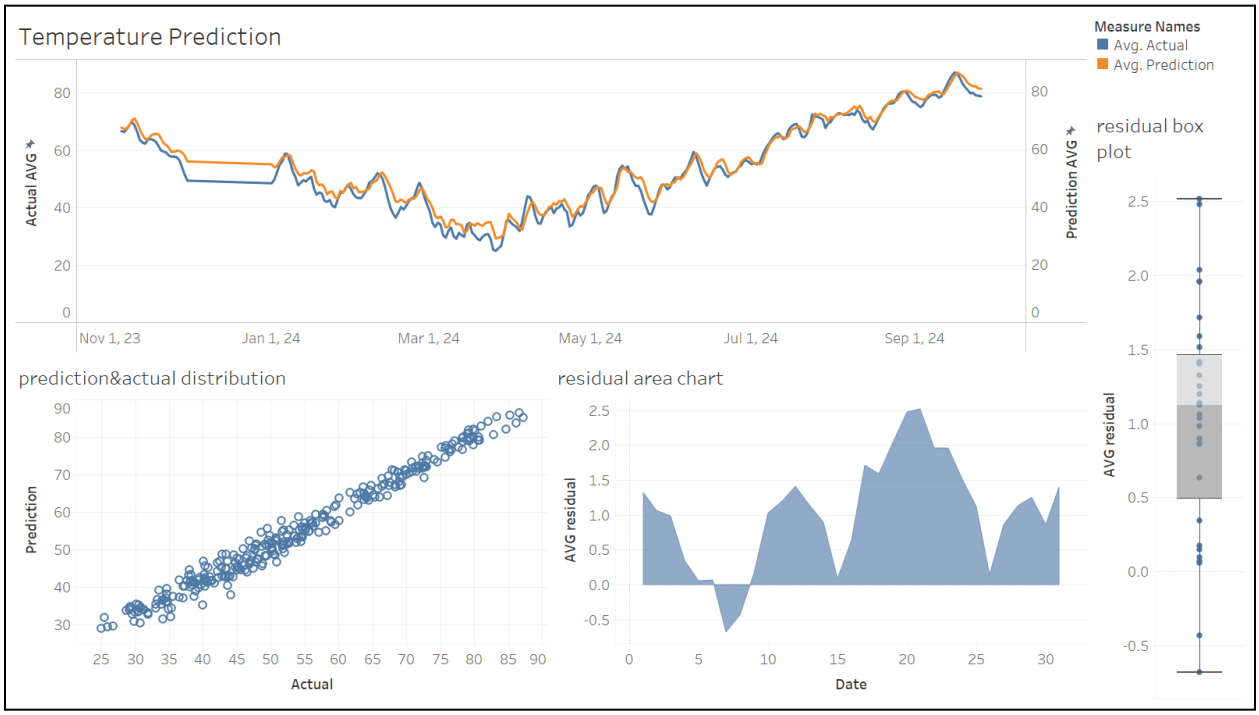
The trend fits well: Here we show the accuracy of the temperature predicted by the model. It can be seen that the accuracy is very high and the entire temperature curve can also be fitted very well.

High accuracy: In the scatter plot, the predicted value and the actual value show an obvious linear relationship, and the points are concentrated near the diagonal, indicating that the prediction is relatively accurate.

Small residual: The residual box plot shows that the residuals are mainly concentrated between 0 and 1, and a few outliers exceed 2, indicating that the prediction error is small and most of them are within a reasonable range.

Residual volatility: The residual area chart shows that the residuals fluctuate in different time periods, and some intervals have slight deviations, but the overall stability.

Overall, the LSTM model has a good prediction effect, and the trend is highly consistent with the actual, but there are certain deviations in some periods.



Visualization of the LSTM model's ability to predict rain conditions

Conclusion

Describe the steps taken in the project

1. Data collection and cleaning
2. Feature analysis and Model evaluation
3. Result export and visualization
4. Processed data display
5. Model comparison analysis
6. LSTM model performance display

Describe the analysis and Describe the discoveries

We have shown it above, here is a brief summary:

Time series analysis:

- Analyze the changing trend of meteorological variables at different time granularities to verify the periodicity and stability of the data.
- It is found that humidity fluctuates to a certain extent within the daily cycle, especially during periods with large temperature differences.

Model comparison analysis:

- Elasticnet model shows the worst prediction performance, with most predicted values concentrated at a single point, indicating underfitting. Lasso performs better in the low humidity range with concentrated predictions, but struggles with high humidity, where predictions become scattered and fluctuate significantly.
- Ridge model demonstrates a good linear fit, with predicted values generally following the true values. RandomForest performs the best, with predicted values closely matching true changes, showing high fitting accuracy and strong generalization, with evenly distributed data points.

LSTM performance analysis:

The LSTM model shows good alignment between predicted and actual values, as seen in the temperature vs. rainfall plot, indicating a reasonable fit. However, residual plots reveal systematic errors, with negative residuals at lower rainfall values and positive residuals at higher values, suggesting underestimation at low and overestimation at high rainfall levels.

Describe any challenges encountered and how you resolved the challenge

In commercial or industrial application scenarios, even if the model prediction accuracy is high, the interpretability and user acceptance of the model results are also key considerations.

Many business personnel or managers lack understanding of complex machine learning models (such as polynomial regression, random forest, etc.) and may be skeptical about the reliability of model results.

Solution:

Use feature importance analysis and sensitivity analysis to explain the specific impact of different variables (such as temperature, wind speed, visibility) on humidity prediction.

In the visualization chart, add explanatory annotations, such as marking key change points and possible influencing factors in the trend chart.

In the report, convert complex regression coefficients or model performance indicators (such as R^2, MSE) into intuitive business indicators, such as "the model can predict the humidity change trend with an accuracy of 80%". In the model introduction section, reduce the use of overly academic terms and explain the prediction effect in language that is closer to practical applications.

Describe any adjustments you had to make from the original plan

Plan adjustment:

The original plan was to use only linear regression, but due to the large impact of nonlinear characteristics, polynomial regression and LSTM was added as a supplement. And after that, we received feedback that some users might not be able to accept overly complex analytical visualizations. We adjusted our plan to create two dashboards to display data for different users (the public and the company's technical department).

Visualization improvement:

Based on customer suggestions, a model prediction comparison chart was added to clearly show the difference in the effects of each model. And the terminology used and the depth of introduction in the visualization process were adjusted to adapt to the understanding scope of business users.

Model optimization:

The initial model only considered temperature and wind speed. Visibility was introduced through feedback suggestions to improve the accuracy of humidity prediction.

Describe any feedback you received from your client and how you incorporated their feedback into your final product

Feedback 1: Insufficient visual comparison

Adjustment: Added prediction comparison charts of different models and differentiated them by color to make the results more intuitive.

Feedback 2: Unstable prediction accuracy

Adjustment: Added LSTM model to improve the performance under nonlinear features, and highlighted the improvement effect in the report.

Feedback3: Lack of dashboards from different audience perspectives

Adjustment: Based on customer suggestions, we are currently reducing the difficulty of understanding the dashboard and using the public dashboard to display it. In the future, we may have 1 dashboard for expert audiences and 1 for the public audience. The public dashboard focuses on displaying daily weather changes and short-term rainfall forecasts, and the interface is simple and easy to understand. The expert dashboard includes trend analysis, model residuals, and forecast confidence intervals, providing in-depth analysis capabilities

After completion, there will be a "toggle view" button on the dashboard, which can be switched according to the audience.