# Predicting NYC Weather

## CIS 9660 - Final Project

Alexander Wendelborn
David Freitag
Kimberly Yee Tan
Neetu Kachawa
Tshering Sherpa

# Project Overview

We will attempt to predict the weather in NYC (Central Park)

- Temperature (predict value)
- Precipitation
  - Classification - raining/not raining?
  - Regression - amount of precipitation

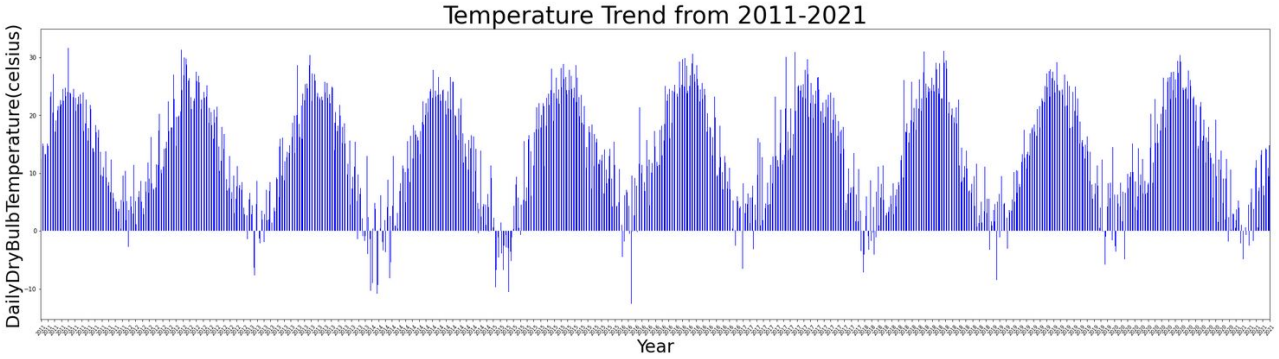
- Models created
  - Temperature regression tree and boosted tree
  - Precipitation regression tree and boosted tree
  - Linear regression for temperature
  - Logistic regression and tree-model classification for precipitation

# Project Data Background, Cleaning, and Issues

- Background:
  - Multiple Datasets from NOAA
    - Multiple cities' data: Houston, Miami, Minneapolis, Buffalo
    - NYC data: Central Park and LaGuardia Airport
- Cleaning
  - Remove NA's
  - Convert feature data types to numeric
- Issues
  - Mixed data type in each column
  - Alpha characters prevent float conversion

# Data Exploration and Patterns

- Seasonality:
  - Precipitation in some months of each year greater than others
  - Temperature follows cyclical pattern
- Data contained some autoregressive behavior where consecutive days were related
- Warming trend across years
- A number of the variables were skewed towards zero as they only have positive values



# Temperature Prediction (Regression)

- Lasso Regression
  - Predicting “Daily Dry Bulb Temperature”
  - Feature engineering - Weekly lag data, rolling average, transforming variables
- Tuning
  - LassoCV to find the optimal alpha value
  - CV = 5
- Evaluation
  - Training set RMSE value : 0.3428
  - Validation set RMSE value: 0.3503
- Potential Improvements
  - Combining feature engineering techniques
  - Specifying the parameter for alpha value

```
{col_name : coef for col_name, coef in zip(x_train.columns,
```

```
{'YEAR': 0.03408393767289274,
 'MONTH': -0.0,
 'DAY': 0.0,
 'PRCP(cm)': 0.006842082320079603,
 'SNOW(cm)': 0.0,
 'SNWD(cm)': 0.0,
 'TMAX(celsius)': 1.027485395620364,
 'TMIN(celsius)': 0.675019744933569,
 'WDF2(degrees)': -0.0,
 'WDF5(degrees)': -0.002114203191019785,
 'WSF2(metres/sec)': 0.0,
 'WSF5(metres/sec)': 0.0,
 'DailyDewPointTemperature(celsius)': 1.4105039738468168,
 'DailyRelativeHumidity': -0.0,
 'DailySeaLevelPressure(inches)': -0.024089306646680596,
 'DailyStationPressure(inches)': -0.0,
 'DailyVisibility': -0.2797345734597328,
 'DailyWetBulbTemperature(celsius)': 4.0288032588380975,
 'DailyWindSpeed(miles/hr)': 0.0,
```

**Transformed variables**

```
'rain_mean': 0.030754208837538807,
'Humdity/temp': -2.341614718479412,
'Dewpoint_sqrt': 1.4027409324695683,
'wetbulb/visibility': -0.07786813741701742}
```

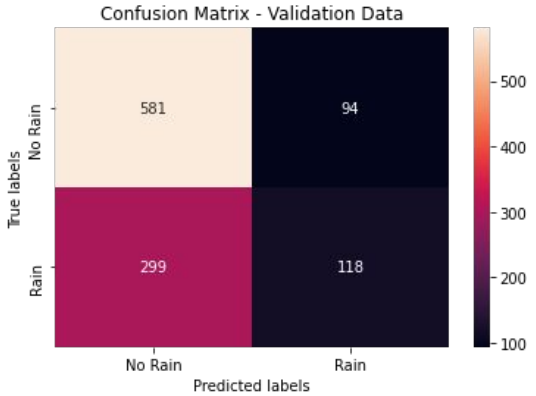
# Precipitation Prediction (Regression)

- Decision Tree Regression
  - Predicting Daily Precipitation - Label: ‘DailyPrecipitation(cm)(Hourlymean)’
- Analysis
  - Feature reduction, Feature engineering - added rolling average for 1 week
- Tuning
  - GridSearchCV to find the optimal parameters
  - n_jobs = 4, scoring = "neg_root_mean_squared_error"
- Evaluation
  - RMSE for Training Dataset: 2.8329157117527726
  - RMSE for Validation Dataset 2.9600148495378855
- Potential improvements given more time
  - Experiment by adding rolling average for varying time periods
  - Additional feature reduction and feature engineering combinations

| | Real Values | Predicted Values |
|---|---|---|
| 2059 | 0.000000 | 0.008230 |
| 1624 | 0.000000 | 0.008230 |
| 3508 | 0.000000 | 0.008230 |
| 2282 | 0.000000 | 0.008230 |
| 1265 | 0.000000 | 0.008230 |
| 383 | 0.000000 | 0.008230 |
| 1732 | 0.000000 | 0.008230 |
| 717 | 0.103673 | 0.490968 |
| 1310 | 1.648772 | 0.490968 |
| 1529 | 5.170714 | 5.431952 |
| 2738 | 0.000000 | 0.008230 |
| 966 | 5.380789 | 5.431952 |
| 2120 | 0.000000 | 0.008230 |
| 2945 | 0.000000 | 0.008230 |
| 2379 | 0.990169 | 5.431952 |
| 529 | 0.000000 | 0.008230 |
| 1712 | 0.000000 | 0.008230 |
| 829 | 7.268308 | 5.431952 |
| 466 | 4.421481 | 1.200738 |
| 96 | 2.813538 | 3.186127 |
| 2662 | 0.000000 | 0.145369 |
| 815 | 0.158750 | 0.008230 |
| 3563 | 11.784419 | 11.523275 |
| 591 | 0.000000 | 0.008230 |

# Classification: Will it Rain?

- Models/Analysis: Logistic Regression, Decision Tree
  - Target variable: PRCP - transformed into 0 or 1 (non-zero values)
  - Lagged data - use today's observations to predict tomorrow's rain
  - Most predictive variables: wind direction and speed, amount of rain yesterday, sea level pressure
- Tuning
  - Tested multiple values for the regularization parameter for Logistic Regression (L1 penalty - lasso)
  - Grid search to identify the optimal parameters for the tree model
- Evaluation
  - Confusion Matrix (challenge: false negatives)
  - AUC Score: 0.65 out of 1.0
- Potential improvements given more time
  - Random Forest
  - Larger parameter grid for grid search
  - Adding additional lags to the dataset



# Deploying the Model in a Production Environment

- Modify the code - Jupyter Notebooks to a single runnable Python script
- Host on a cloud server
- Inputs: weather data
- Outputs: weather prediction (temperature, precipitation)
- Nightly: update the training dataset with new observations and re-run the model
- Endpoints:
  - HTTP API using Flask
  - Create a data pipeline to generate a report based on predictions
  - Data pipeline to feed a Business Intelligence dashboard displaying predictions

Thank you.

At this time, we welcome any questions.