**Weather Prediction for J.F.K international Airport, New York**

**By Jaspreet Kaur (A04284632)**

**MASTER OF SCIENCE**

**in**

**COMPUTER SCIENCE**

**SUPERVISOR:**

**PROF. Dr. Scott King**

**Introduction and Problem Statement:**

* For this project, I used various Regression models to predict the maximum temperature. I took the dataset from website <https://www.ncdc.noaa.gov/cdo-web/datasets> in CSV format from 01/01/1990 to 04/09/2024 on daily basis.
* Accurately predicting maximum temperatures is crucial for numerous operational aspects, including flight scheduling, passenger comfort, and logistical planning.
* This project aims to develop a tailored predictive model using historical weather data and local factors to improve temperature forecasts.
* It seeks to enhance decision-making for airport authorities, optimize resource allocation, and mitigate weather-related risks effectively.

**TECHNOLOGY USED:**

I have used Google Collab as a platform to write and run our code.

I have used Google drive to store my dataset.

I have used python language and python libraries to write algorithms as well as entire code.

**DATA ANALYTICS PROCESS:**

Following steps were implemented for our dataset:

**a) Data Collection:**

We collected our dataset from NOAA. Our dataset consists of 12518 rows and 11 columns including the null values. We are considering the following attributes in our dataset:

A screenshot of a computer

Description automatically generated

**b) Data Cleaning:**

First calculated the Null values and then also calculated the percentage of missing values in each column because selecting only columns whose missing values are less than 5%.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

I choose the linear interpolation method to fil in the missing values because it provides a simple yet effective way to estimate the values between known data points. assumes a linear relationship between consecutive data points. While this assumption might not always hold true, it often provides a reasonable approximation, especially for short intervals between data points. Linear interpolation produces a smooth transition between known data points, resulting in a visually appealing curve.

A screenshot of a computer program

Description automatically generated

Then, converted the index to Date time. This indexing facilitates data retrieval, slicing, and subsetting operations, making it easier to extract relevant information for analysis or visualization.

A screenshot of a computer

Description automatically generated

Checked if there are any gaps in the data set.

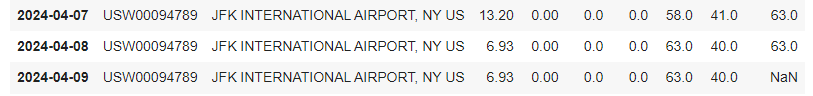
A screenshot of a computer

Description automatically generated

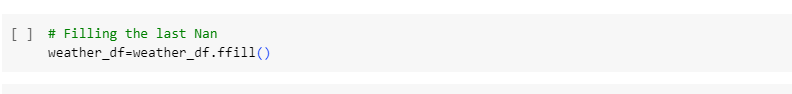
Selected the target column in the collected dataset which is maximum temperature: tmax-

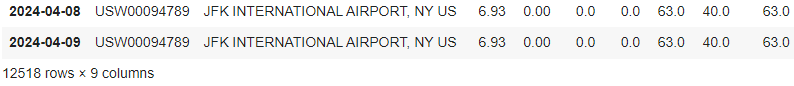
A screenshot of a computer

Description automatically generated

****

As we see, there was NaN value in the last row of the dataset, I filled it with the fast forward method.





**c) Data visualization:**

**A graph showing the maximum and minimum temperature for the data

Description automatically generated** **A screenshot of a chart

Description automatically generated**

By visually inspecting the heatmap, we can quickly discern which variables are positively correlated, negatively correlated, or not correlated at all. Strong positive correlations are represented by darker shades of one color, while strong negative correlations are represented by darker shades of another color. Weak correlations are depicted by lighter shades or absence of color.

**d) Models:**

**1. RIDGE REGRESSION:** Ridge regression is a regularization technique used in statistical modeling and machine learning to mitigate the problem of multicollinearity and overfitting in linear regression models. It is particularly useful when dealing with datasets where predictor variables (features) are highly correlated or when the number of predictor variables is large relative to the number of observations. Ridge regression shrinks the coefficients towards zero, reducing their variance and mitigating the effects of multicollinearity. This helps in improving the stability and generalization performance of the model. It trades off a small increase in bias for a significant reduction in variance. Ridge regression can automatically perform feature selection by shrinking less important features towards zero.

A screenshot of a computer program

Description automatically generated

**2.LINEAR REGRESSION:**

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables (predictors) and the dependent variable (outcome). The goal of linear regression is to find the best-fitting linear equation that describes this relationship. Linear regression provides easily interpretable coefficients that represent the relationship between each independent variable and the dependent variable. Linear regression assumes that the relationship between the independent and dependent variables is linear, that the errors are normally distributed with constant variance (homoscedasticity), and that the errors are independent of each other.

A screenshot of a computer program

Description automatically generated

3.Lasso Regression: Lasso regression, short for Least Absolute Shrinkage and Selection Operator, is a regularization technique used in regression analysis to reduce overfitting and perform variable selection by imposing a penalty on the absolute size of the coefficients. It is particularly useful when dealing with high-dimensional datasets with many predictor variables, some of which may be irrelevant or redundant. One of the main advantages of lasso regression is its ability to perform variable selection by driving some coefficients exactly to zero. Lasso regression can handle multicollinearity (high correlations between predictor variables) by shrinking the coefficients towards zero. It trades off a small increase in bias for a significant reduction in variance.

A screenshot of a computer program

Description automatically generated

**e) Plots:**

**Line Plot:** It is commonly used to visualize trends or patterns in data over time or across ordered categories. Line plots are particularly useful for showing continuous data or data that has a natural order.

**Violin Plot:** A violin plot is a type of graphical representation used to visualize the distribution of numeric data and its probability density. Each data point or group of data points is represented by a "violin" shape, which displays the distribution of the data. The width of the violin at each point represents the density of data values at that point. Violin plots are useful for comparing distributions between different groups or categories of data.

**Error Graph:** Error bar plot, is used to display the variability or uncertainty associated with each data point in a dataset. Error bars provide valuable information about the precision and reliability of the data, helping viewers assess the significance of differences between groups or conditions.

A screenshot of a computer program

Description automatically generated

**f) Errors as Evaluation Metrics:**

**1) R-Squared:** R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model.It ranges from 0 to 1, where:

R2 = 0 indicates that the model does not explain any of the variability of the dependent variable around its mean.

R2 = 1 indicates that the model perfectly explains all the variability of the dependent variable around its mean.

**2) Root Mean Square Error (RMSE):** It is a measure of the average magnitude of the errors (residuals) between predicted values and actual values in a regression model.

It is calculated by taking the square root of the mean of the squared differences between predicted and actual values.RMSE provides a measure of the typical deviation of the predictions from the actual values in the same units as the dependent variable. Lower values of RMSE indicate better model performance, with a value of 0 indicating perfect predictions.

**3)Mean Absolute Error (MAE):** Mean Absolute Error (MAE) is a useful metric for assessing the accuracy of regression models, providing insights into the average magnitude of errors between predicted and actual values. MAE is commonly used in regression tasks to evaluate the performance of models, particularly when the absolute magnitude of errors is important and needs to be directly interpretable. Lower values of MAE indicate better model performance, with a value of 0 indicating perfect predictions.

A screenshot of a computer program

Description automatically generated

**g)Models Evaluation:**

**RIDGE REGRESSION MODEL 1:**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

The rolling mean, also known as the moving average, is a statistical calculation used to analyze data points by creating a series of averages of different subsets of the full data set. It helps to smooth out fluctuations or noise in a dataset, making underlying patterns or trends more visible.

**A screenshot of a computer

Description automatically generated**

Starting values have NaN since it does not have past values to take rolling mean, so I cut the first 15 rows.

**A screenshot of a computer

Description automatically generated**

**A colorful squares with numbers

Description automatically generated with medium confidence**

**Ridge Regression Model 2:**

**A screenshot of a computer program

Description automatically generated**

**A graph of blue and orange lines

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A graph with a line

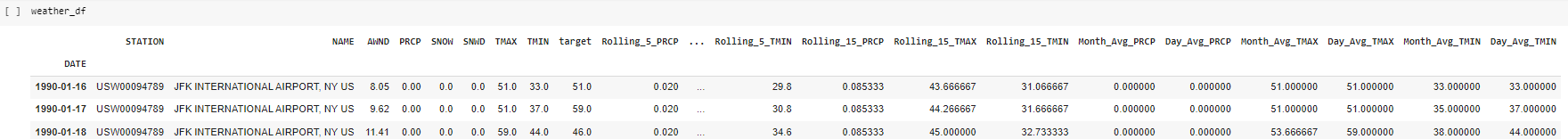
Description automatically generated**

Expanding mean function to add more columns:

It calculates the mean of all the values up to a certain point in the dataset. Expanding mean with more columns allows for a more comprehensive analysis of the dataset, leading to deeper insights and a better understanding of the underlying patterns and trends. It gives Richer Analysis, Improved Context, Enhanced Visualization and Increased Flexibility.

**A close-up of a computer code

Description automatically generated**

****

**A colorful grid with numbers

Description automatically generated**

**Ridge Regression - Model 3:**

**A screenshot of a computer program

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a computer screen

Description automatically generated**

**A graph with a line

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**Linear Regression – Model 1**

**A screenshot of a computer

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**Linear Regression – Model 2: Adding Rolling Mean:**

**A screenshot of a computer program

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**Linear Regression – Model3: Expanding Columns:**

**A screenshot of a computer program

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Lasso Regression- Model 1:**

**A screenshot of a computer program

Description automatically generated**

**A graph with blue and orange lines

Description automatically generatedA screenshot of a computer screen

Description automatically generated**

**A graph with a line

Description automatically generated**

**Lasso Regression- Model 2- Adding Rolling Mean:**

**A screenshot of a computer program

Description automatically generated**

**A graph showing a graph of value

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**Lasso Regression- Model 3: Expanding Columns:**

**A screenshot of a computer program

Description automatically generated**

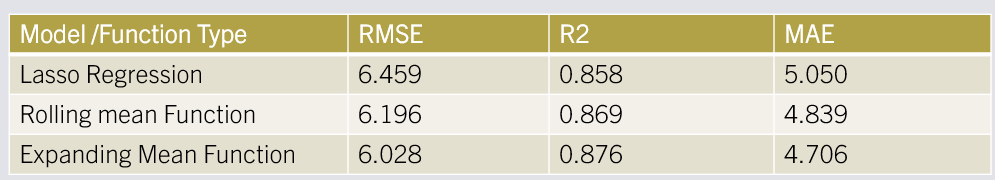
**A graph of blue and orange lines

Description automatically generated A diagram of a violin plot

Description automatically generated**

**A graph with a line

Description automatically generated**

****

**Splitted the dataset into Train and Test for the next two models:**

**A close-up of a computer screen

Description automatically generated**

**Elastic Net Regression Model:**

Elastic Net regression is a regularization technique that combines both Lasso (L1 regularization) and Ridge (L2 regularization) penalties in a linear regression model. It's particularly useful when dealing with high-dimensional data where the number of predictors (features) is much larger than the number of samples.

A screenshot of a computer program

Description automatically generated

A graph showing a line of blue and orange lines

Description automatically generatedA diagram of a violin plot

Description automatically generated

A graph with blue line

Description automatically generated



**Random Forest Model:**

Random Forest is a versatile and powerful ensemble learning method used for both classification and regression tasks. It belongs to the family of tree-based models and is known for its high accuracy, robustness, and ability to handle complex datasets with high dimensionality. Random Forest consists of a collection (or ensemble) of decision trees. Each tree is built independently, and the final prediction is made by aggregating the predictions of all the individual trees.

Advantages

\*Robust to overfitting, thanks to the ensemble of trees and randomization techniques.

\*Can handle both numerical and categorical data.

\*Requires minimal preprocessing of data (e.g., no need for feature scaling).

\*Provides feature importance scores, which can be helpful for feature selection.

A screenshot of a computer

Description automatically generated

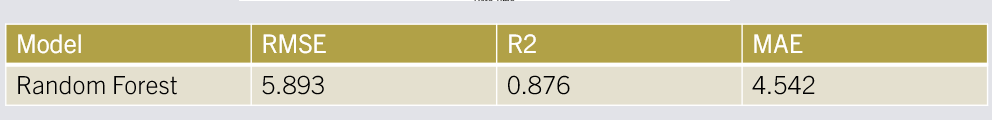
A graph showing a blue and orange line

Description automatically generated A diagram of a violin plot

Description automatically generated

A graph with a line

Description automatically generated



**h) Comparison of Models based on Error metrics:**

**A table with numbers and a few words

Description automatically generated with medium confidence**

**i)Conclusion:**

* A lower RMSE and lower MAE indicates a better fit. An R-squared value closer to 1 indicates a better fit.
* Random Forest has the lowest RMSE (5.893) and MAE (4.542), which means it has the least overall prediction error on average. Random Forest excelled with the lowest errors (RMSE, MAE) and explained a high portion of the data's variance (R-squared).
* It outperformed other models like Ridge, Lasso, and Linear Regression.
* Applying Rolling mean function and expanding mean function improves the accuracy and decreases the error.

**Thank You.**