Project Title: Weather-Prediction-using-Ridge Regression

Introduction:

This report summarizes a project that aimed to predict the maximum temperature for the next day in New York City using a ridge regression machine learning model. The data used was historical weather data for New York City from 1970 to 2020.

The report provides an overview of the project including the data preparation, model development, evaluation, and enhancement steps. It also includes the theory behind ridge regression, results, and conclusions.

Data Preparation:

The historical weather data was loaded into Python and exploratory data analysis was conducted. This involved checking for missing values which were handled using forward fill. Columns with many missing values were dropped.

The target variable, maximum temperature for the next day (tmax), was created by shifting the original tmax column down by one day. This created the variable we wanted to predict.

The data was then prepared for machine learning by splitting into training and test sets. The first 10 years were used for training and the next 90 days for testing. This was repeated by adding the 90 test days back into training, and predicting the next 90 days to simulate backtesting.

Ridge Regression Model:

A ridge regression model was developed to predict the daily maximum temperature. Ridge regression is similar to linear regression but penalizes large coefficients to account for multicollinearity in the data. This is done by introducing a complexity parameter λ which controls how much the coefficients are shrunk.

The key advantage of ridge regression is it reduces overfitting on the training data by restricting the magnitude of the model coefficients. This improves the prediction accuracy on new unseen data.

The scikit-learn library in Python was used to implement the ridge regression model. The model was trained on the first 10 years of data and tested on the next 90 days. This was repeated in a backtesting approach to evaluate performance over the full dataset.

The initial ridge regression model achieved a mean absolute error of 5.13 and R-squared of 85.67% on the test set. This indicates the model explains 85.67% of the variance in maximum temperature.

Model Improvement:

To improve the model, new parameters were engineered based on rolling averages and means of related weather variables like minimum temperature and precipitation.

Specifically, 3-day and 14-day rolling averages and percentage differences were calculated for tmax, tmin, and prcp. The expand_mean() function was applied to find the historical monthly and yearly averages for each day.

Adding these new parameters improved the model performance. The mean absolute error decreased to 4.7 and the R-squared increased to 87.5%. The new features provided more context to the model through rolling trends and historical averages.

Results:

The final ridge regression model with the new engineered features demonstrated improved ability to predict daily maximum temperatures. The mean absolute error was reduced by 8.6% from 5.14 to 4.7. The R-squared increased 1.9% from 85.6% to 87.5%, indicating more variance in maximum temperature was explained.

These results show the new rolling average and mean features helped improve the predictive accuracy. The model is better able to account for temperature trends and historical data.

Conclusions

This project demonstrated an effective application of ridge regression for weather prediction. The model was able to accurately predict maximum daily temperatures in New York City using historical weather data. Feature engineering with rolling averages and historical means helped boost model performance.

The model achieved reasonably low error and high explained variance. Further improvements could potentially be made by tuning the ridge regression lambda parameter and testing additional feature engineering. Overall, the project shows the feasibility of using machine learning for weather forecasting.