Midterm Checkpoint

Intro + Background

Flight delays and cancellations have become a critical issue in the aviation industry, causing significant economic losses and passenger inconvenience. In 2022 alone, flight disruptions generated an economic impact of $30-34 billion in the US [4]. For our project, we will be creating an arrival delay prediction model using PCA to find the best features from flight data and weather data then combining them into a Linear Regression to see if we can get more accurate arrival delay predictions. The combined feature Linear Regression model will be compared against models only using one of the datasets (either flight, weather, best flight feature, best weather features) to see which performs better.

Literature Review

A recent study used various supervised models to see what factors have the biggest impact on delay predictions [2]. Across the various model, they found that the features affecting the delay most are visibility, wind, and departure time, a combination of flight and weather feature with weather being used the most. Similarly, a study by Y. Tang also examining several flight delay algorithms using a mixed dataset found that categorical flight data such as “TAIL_NUM” have little impact on predicting flight delays [1]. In contrast, a recent study by Go Nam Lui et al also emphasized the importance of integrating both weather data for accurate arrival delay predictions. Go Nam Lui et al utilized a Bayesian statistical approach to quantify the impact of severe weather on airport arrival on-time performance, highlighting the complex relationship between weather conditions and flight delays [3]. Across their three key performance metrics, they discovered a non-linear relationship with the weather score, akin to a phase transition, proving severe weather’s effect on an airport’s arrival performance metric.

Dataset Description

Two datasets will be used for this project with the first being the Airline Flight Dataset [8]. This dataset contains information about the flight schedules, and aircraft types. We will be relying on the Weather Data: Explore Weather Patterns & Predictions in 10 Locations for our weather data [9]. This dataset contains weather conditions of past flights. After pre-processing the data, we found that there’s no way to line-up the flight and weather data as the years do not match up. A new dataset was found that has both flight and weather data but that’ll be discussed in the next steps. As for now these two datasets were used to test the PCA code for our data preprocessing, more on this in the methods section.

Problem Definition

The disruptions to airline operations discussed above arise from multiple factors—weather, air traffic control, crew availability, and technical issues. Notably, weather contributes substantially, with reduced visibility accounting for 52% of weather-related delays [7]. Traditional methods struggle to capture the complexity of these variables, underscoring the need for a robust machine learning model that can more accurately forecast potential delays and help airlines optimize operations. While weather data has been used in existing flight delay prediction models, our project focuses on evaluating its relative predictive value by comparing custom machine learning models trained exclusively on airline data versus those trained exclusively on weather data. By isolating these feature sets, we aim to show that weather data generally contributes more significantly to accurate delay prediction. Ultimately, we will combine the most informative features from both datasets to develop a third model and assess whether this integrated approach yields improved performance. These insights can help airlines and data engineers prioritize the inclusion of high-impact features—particularly weather data—when building or improving flight delay prediction systems.

Methods

Datacleaning and Unsupervised Method

The data preprocessing pipeline developed for this project begins with feature engineering to remove irrelevant data, such as cancelled flights and flight numbers, and to convert categorical string values—specifically airline, origin, and destination—into one-hot encoded vectors. Principal Component Analysis (PCA) is then applied to reduce dimensionality and mitigate multicollinearity by transforming the data into a set of uncorrelated components, while retaining 90% of the original variance. This preprocessing setup is particularly well-suited for linear models such as linear regression and ridge regression, which are sensitive to highly correlated features and benefit from the improved generalization and performance that dimensionality reduction provides.

Supervised Method

Linear Regression will serve as a baseline model using data pre and post PCA. Linear regression will be used to output predicted arrival delay times, and both the closed form solution and gradient descent will be explored to see if overfitting is an issue. To avoid using the same data for training and testing, we will be splitting the data—70% for training and 30% for testing. We will tune the learning rates of the gradient descent model for performance as well as system capabilities.

Potential Results + Discussion

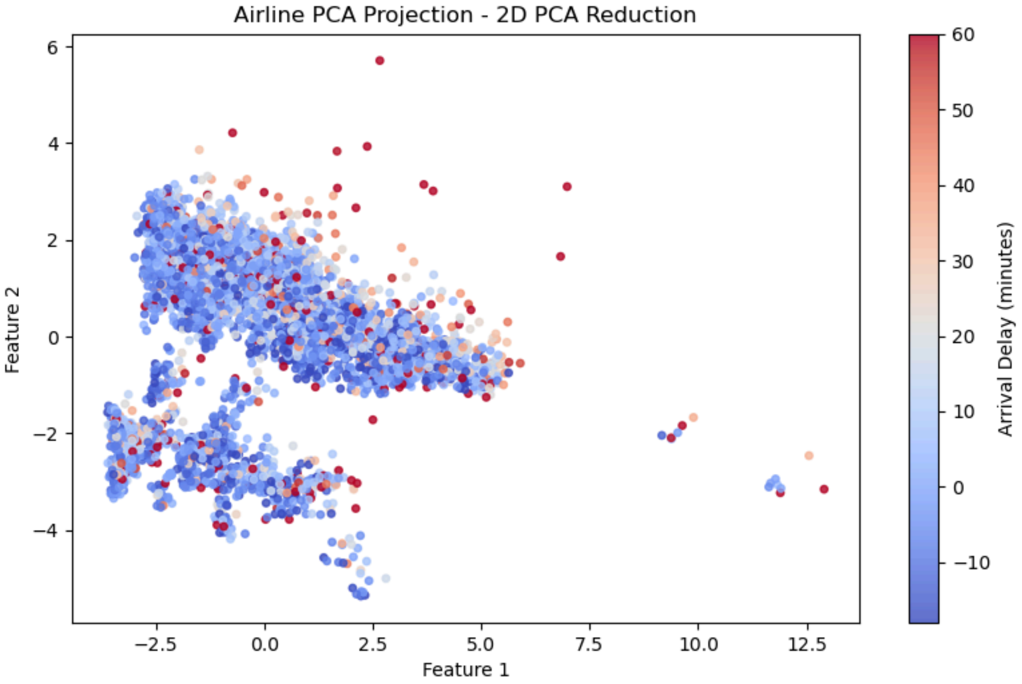
By using PCA for initial preprocessing, we will be able to extract the most relevant features that contribute to arrival times. We will be using MSE and RMSE metrics to evaluate the capability of our linear regression models to predict flight arrival delays. An indication that our models performed well would be lower MSE and RMSE values.

In this section we will be discussing the PCA visualizations generated on our datasets and the results of our linear regression models.

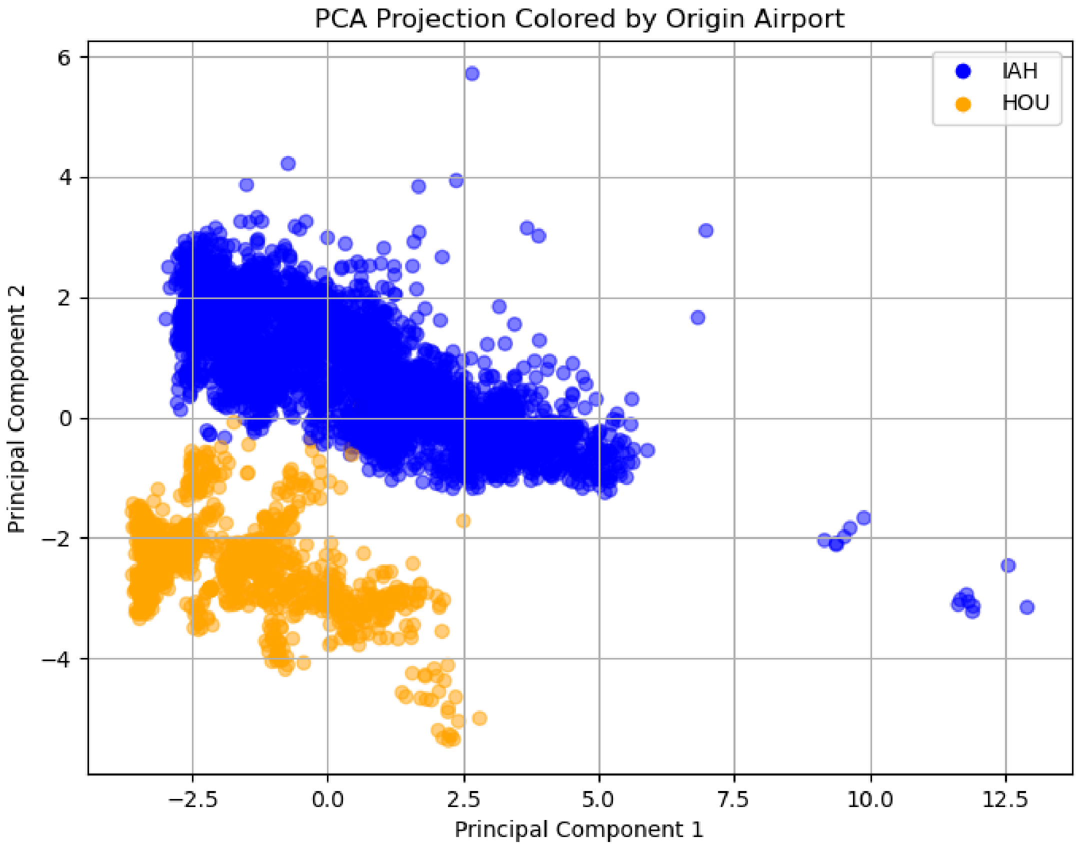
Visualizations

Unsupervised

The following PCA visualizations serve two purposes: first, to validate the functionality of the data preprocessing pipeline by demonstrating clear structure in the reduced feature space; and second, to highlight the nature of variance within the airline dataset. The clustering primarily reflects logistical factors—such as flight distance and airport of origin—rather than indicators of potential delay. This suggests a limitation in the predictive power of airline-only data, motivating the need to incorporate weather-related features that may capture more meaningful patterns linked to flight delays.



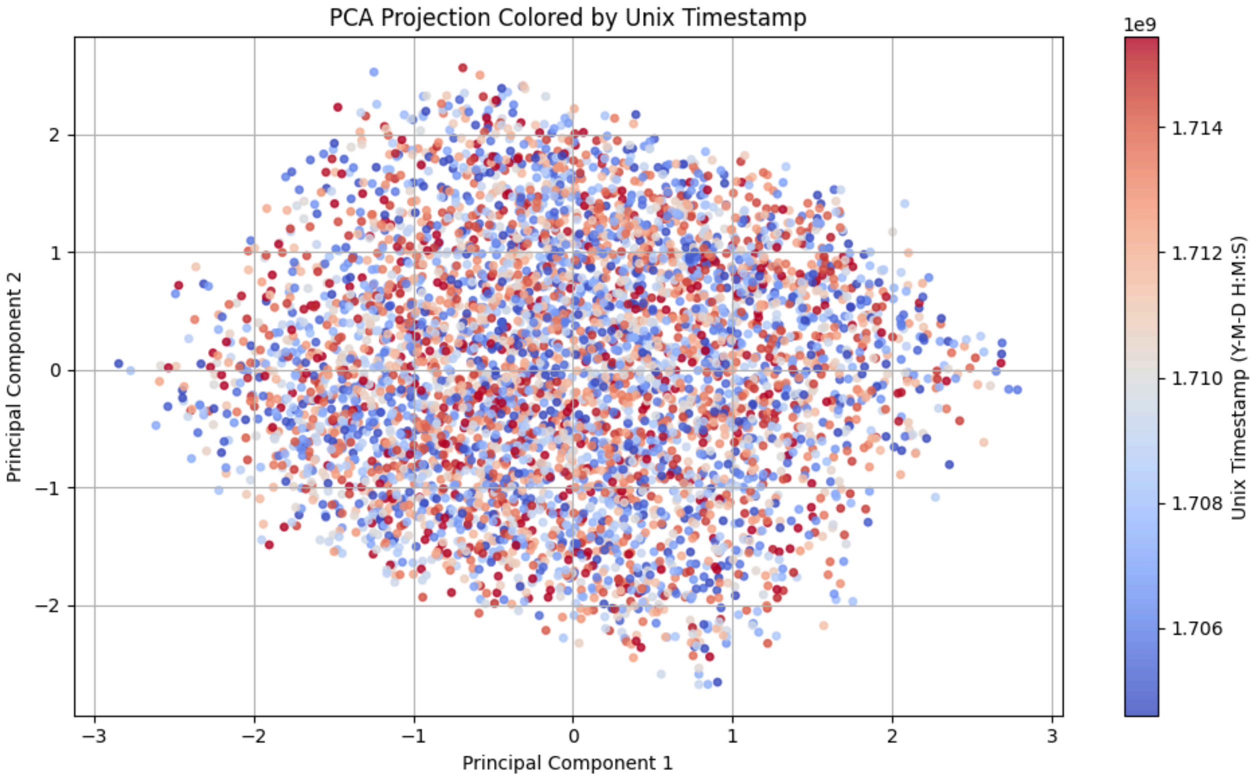
PCA scatter plot of airline data with arrival delay color bar



PCA scatter plot of airline data with two airports labeled by color

In parallel with the airline data pipeline, a separate preprocessing pipeline was developed to handle a weather dataset using the same methodology—feature engineering, one-hot encoding, and PCA for dimensionality reduction. Although the weather dataset could not be directly aligned with the airline data due to differences in time coverage, we still applied PCA and generated a visualization to validate that the pipeline was functioning correctly and capable of capturing meaningful structure within the weather features. The original plan was to train regression models separately on the airline and weather datasets, then combine the most influential features from both to create an integrated model. However, the misalignment in time periods prevented meaningful comparison or combination. This challenge, along with the future direction of the project, is discussed further in the Next Steps section.

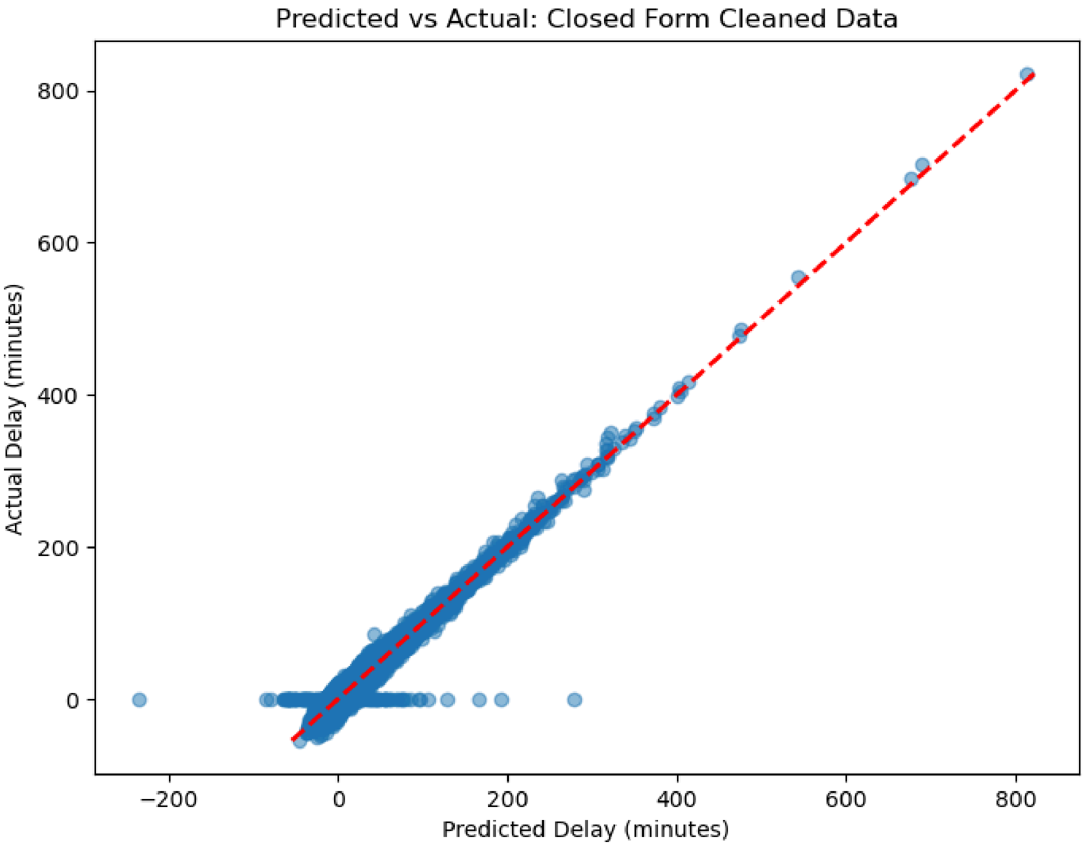
Unlike the airline PCA the weather data didn't have arrival delay times to compare the features too. There is no way to algin the airline and weather datasets; despite sharing locations the data was captured years apart and there is no way to match up flights with weather conditions at the time of flight. Because of this, the above plot is colored by the Unix Timestamp just to see how well the PCA's cluster against a timed feature. As expected no cluster was seen as the each Unix Timestamp is unique also the datasets features (Temperature C, Humidity %, Precipitation mm, Wind Speed km/h) all heavily affect each other removing any chance of clustering by space as well.



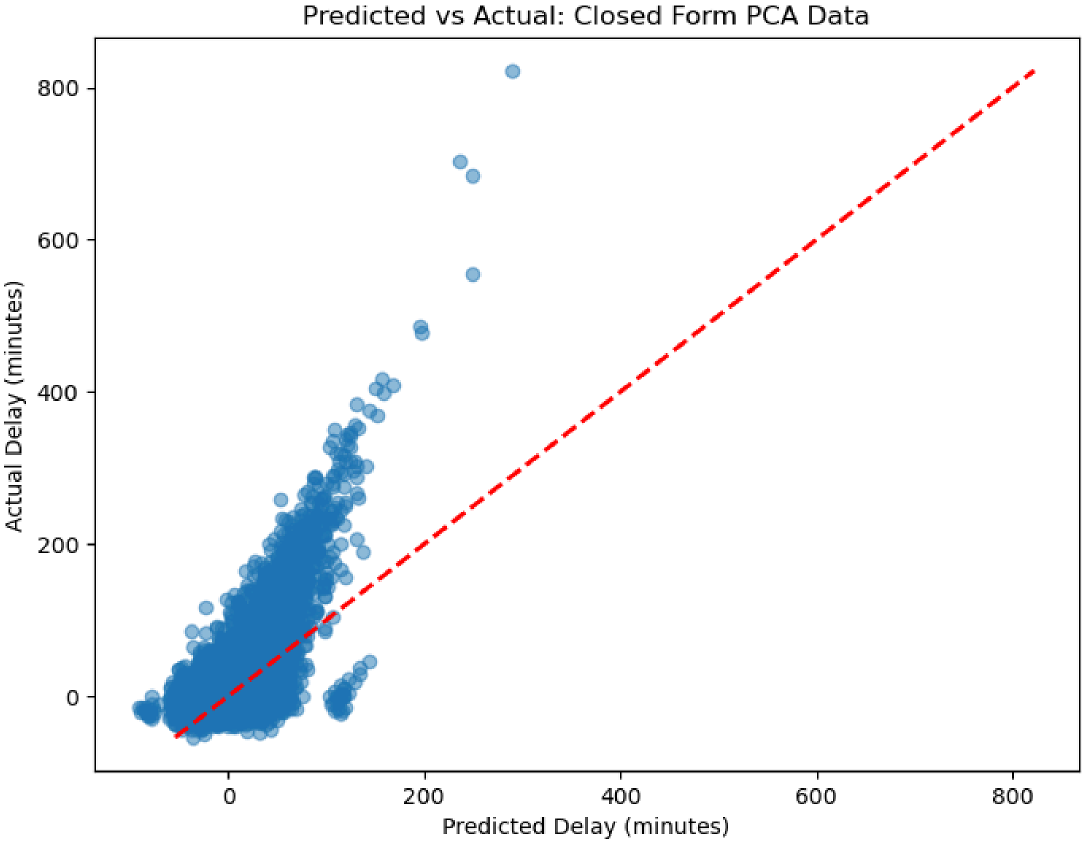
PCA scatter plot of weather data with timestamp color bar

Supervised

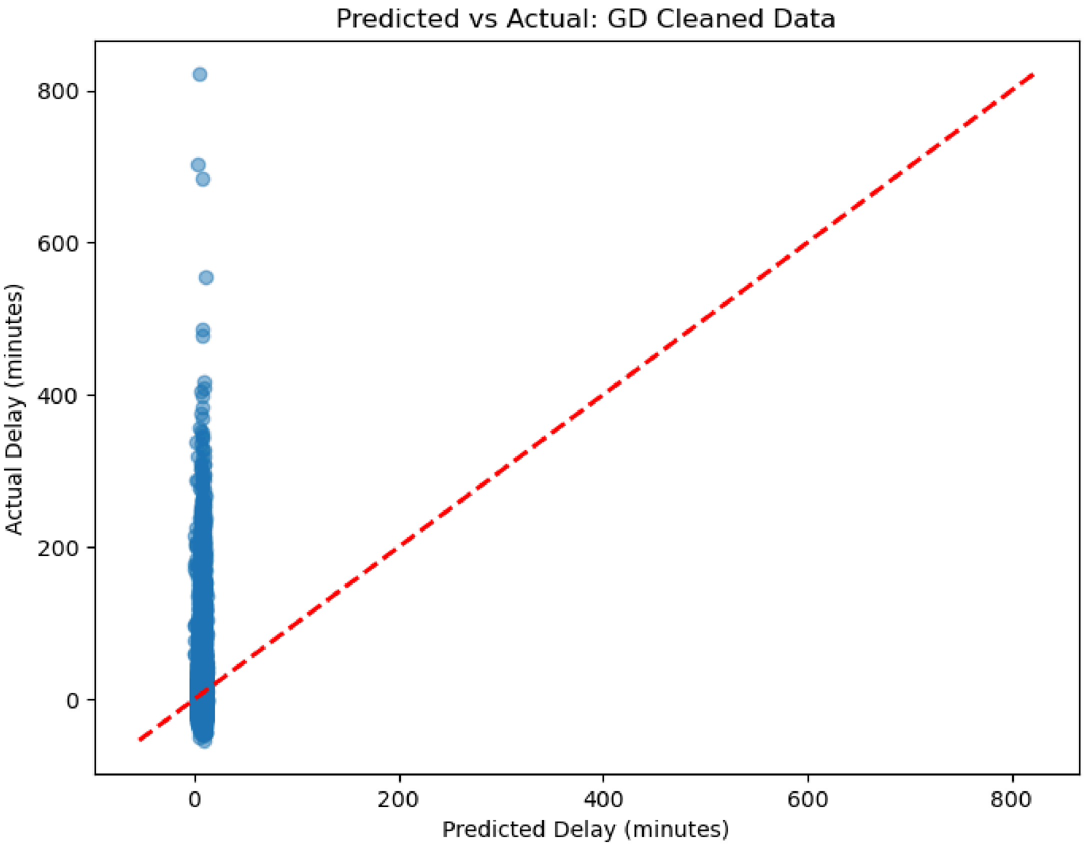
The scatter plots of predicted versus actual flight delays across four methods—closed-form and gradient descent (GD) on both cleaned and PCA-transformed airline data—reveal distinct performance patterns, with the red dashed line in each plot representing the ideal prediction scenario (y=x). The closed-form solution on cleaned data performs best, showing a tighter alignment with the ideal line, though it slightly overpredicts smaller delays and underpredicts larger ones, indicating a more effective capture of delay patterns. In contrast, both GD on cleaned data and the two PCA-based methods (closed-form and GD) exhibit significant underprediction, with predictions rarely exceeding 100-200 minutes despite actual delays reaching 800 minutes, as most points cluster below the ideal line. This consistent underprediction in PCA models suggests that dimensionality reduction may have discarded critical features, while GD's poor performance on both datasets highlights potential convergence issues or an inability to model larger delays, underscoring the need for more robust methods or additional data like weather to improve predictions.



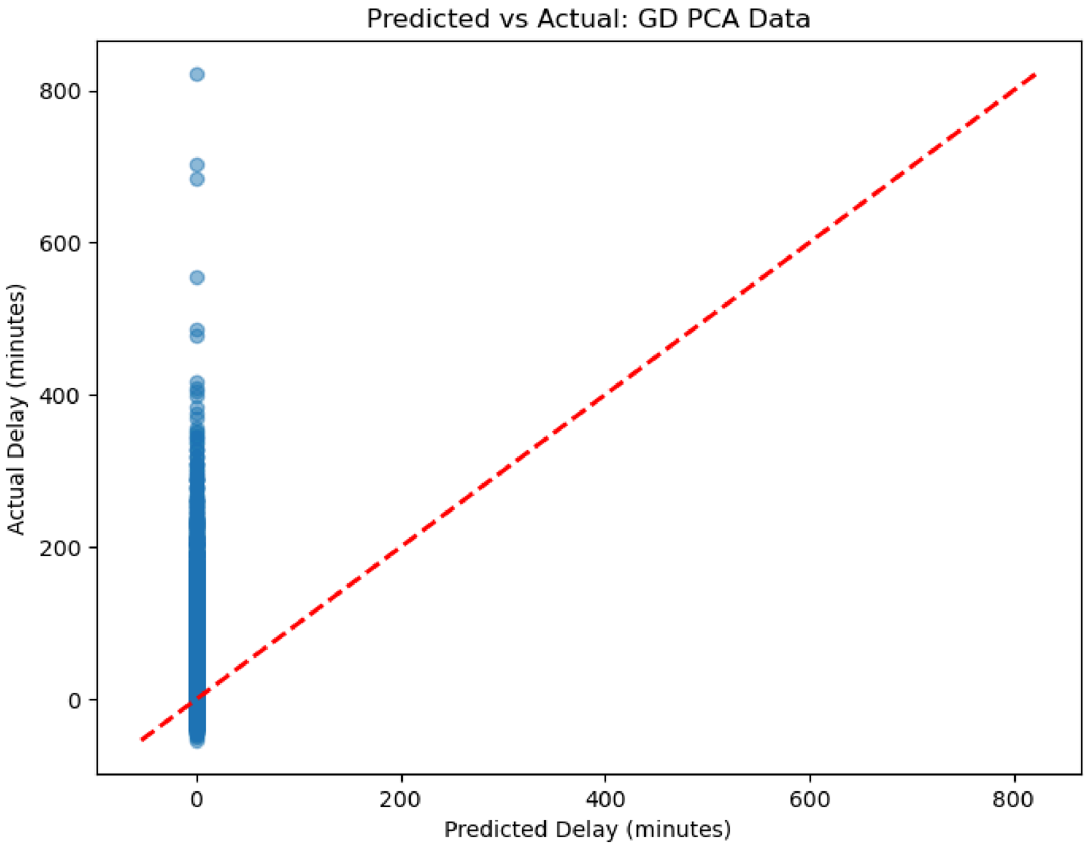
Closed-form solution on cleaned airline data: predicted vs. actual flight delays, demonstrating a tighter fit along the ideal line, though some overprediction occurs for smaller delays



Scatter plot of predicted vs. actual flight delays using the closed-form solution on PCA-transformed airline data, revealing significant underprediction for larger delays as most points cluster below the ideal line



Predicted vs. actual flight delays for gradient descent on cleaned airline data, showing a tendency to underpredict larger delays with most points concentrated at lower actual delay values



Gradient descent on PCA-transformed airline data: predicted vs. actual flight delays, highlighting a consistent underprediction trend, with predictions failing to capture the full range of actual delays

Quantitative Metrics

Method	Dataset	RMSE (minutes)
Closed Form	Cleaned	6.135
Closed Form	PCA	24.769
Gradient Descent	Cleaned	30.298
Gradient Descent	PCA	31.256

Analysis

What is PCA and why did we use it? Are there other data preprocessing methods that could be used?

PCA stands for Principal Component Analysis and is a linear dimensionality reduction technique. The data is linearly transformed onto a new coordinate system such that the directions (principal components) capturing the largest variation in the data can be easily identified. These new directions can hopefully be used to see patterns in the data and create better visualizations. We additionally used feature scaling to standardize the data and ensure that features contribute equally to the analysis. Feature scaling helps RMSE by ensuring that all input features contribute equally to the model, leading to more stable and accurate predictions. This can also lead to better model convergence with gradient descent-based models, which we used in our project. Additional data preprocessing methods include feature selection as done in class, independent component analysis, and t-distributed stochastic neighbor embedding (t-SNE).

Did using PCA for data preprocessing (dimensionality reduction) help the regression model? Why or why not?

We found that the linear regression models with PCA components performed worse than the models without PCA. We attribute this to possible nonlinearities in the data that PCA or linear regression can’t capture. PCA is inherently a linear transformation, thus any nonlinear relationships in the data won’t be seen in the principal components.

Why use RMSE over MSE?

One reason to use RMSE over MSE is interpretability of the data. When one squares the error, this introduces a new unit that may be difficult to understand in the context of the data. Taking the square root of this provides the same unit as the target variable, making it more interpretable. RMSE also moderates the impact of outliers by taking the square root.

How did linear regression perform? What would be the benefits of using ridge regression instead?

Linear regression showed mixed performance across the datasets, with the closed-form approach outperforming gradient descent, particularly on the cleaned airline data, where it likely captured the underlying patterns more effectively. However, its performance dropped noticeably on the PCA-transformed data, possibly because the dimensionality reduction stripped away key details. Gradient descent struggled even more, suggesting it may not have been optimally tuned to converge well. The results point to linear regression’s limitations in handling nonlinear relationships, especially since we’re only using airline data without additional factors like weather, which could introduce more complexity. Ridge regression could improve things by adding regularization, which helps control overfitting and stabilizes the model, making it better suited to handle subtle patterns or noisy data—benefits that might shine through even more once we bring in richer datasets like weather information.

Next Steps

New Dataset

One of the key next steps in this project involves integrating airline and weather data into a single model to better evaluate their combined predictive power for flight delays. The original plan to merge insights from two separately processed datasets was hindered by a mismatch in their timeframes, which made it impossible to align records based on date and time. To resolve this, we have identified a new dataset that includes both airline and weather features natively aligned by timestamp. Unlike the original airline dataset, which was limited to just two airports—one domestic-focused and one an international hub—this new dataset spans a broader range of airports, helping to reduce potential bias in the model and improve the generalizability of our results. In the next phase of the project, we will update our preprocessing pipeline to accommodate this new dataset, perform joint PCA and feature selection, and assess whether the combined model outperforms those trained on airline or weather data alone.

Ridge Regression

Ridge regression can be used to help overfitting, but we also hope to try and capture some of the nonlinearities in the data. We hope to do this by transforming the input features into polynomial features before applying ridge regression. This essentially creates new features that capture non-linear relationships, while the ridge penalty still helps with overfitting.

Feature Selection Preprocessing

Feature selection preprocessing is an additional dimension reduction technique used in machine learning. We hope to use this to preprocess our data to find the features that contribute most to predicting the weather. We also hope to compare this approach with PCA to see which one is best on our specific datasets, as each approach has its own strengths and weaknesses.

Baseline Comparison w/ Other Prediction Models

We want to compare the models that we implement on the data with attempts that other machine learning scientists have used for the same problem. We found two distinct models that attempted to answer/predict the same question as us and we’re curious to see how we compare once we’ve tuned preprocessing our data and found a model that gives us a low RMSE without overfitting.

References

[1] Y. Tang, “Airline Flight Delay Prediction Using Machine Learning Models,” 2021 5th International Conference on E-Business and Internet, Oct. 2021, doi: https://doi.org/10.1145/3497701.3497725

[2] H. Khaksar and A. Sheikholeslami, “Airline delay prediction by machine learning algorithms,” Scientia Iranica, vol. 0, no. 0, Dec. 2017, doi: https://doi.org/10.24200/sci.2017.20020

[3] Lui, G. N., Hon, K. K., & Liem, R. P. (2022). Weather impact quantification on airport arrival on-time performance through a Bayesian Statistics Modeling Approach. Transportation Research Part C: Emerging Technologies, 143, 103811. https://doi.org/10.1016/j.trc.2022.103811

[4] “AirHelp Report: The impact of flight disruption on the economy and environment,” AirHelp, Sep. 26, 2023. Available: https://www.airhelp.com/en-gb/press/airhelp-report-the-impact-of-flight-disruption-on-the-economy-and-environment/

[5] J. Knutson, “Airline issues leading cause for flight delays, federal data shows,” Axios, May 11, 2023. Available: https://www.axios.com/2023/05/11/flight-delays-airlines-data

[6] H. Bhanushali, “Impact of Flight Delays,” ClaimFlights, May 15, 2023. Available: https://claimflights.com/impact-of-flight-delays/

[7] J. A. Algarin Ballesteros and N. M. Hitchens, “Meteorological Factors Affecting Airport Operations during the Winter Season in the Midwest,” Weather, Climate, and Society, vol. 10, no. 2, pp. 307–322, Apr. 2018, doi: https://doi.org/10.1175/wcas-d-17-0054.1

[8] Arun Jangir, “Airline Flight Dataset: Schedule, Performance etc,” Kaggle.com, 2023. Available: https://www.kaggle.com/datasets/arunjangir245/airline-flight-dataset-schedule-performance-etc. [Accessed: Feb. 22, 2025]

[9] P. Patil, “Weather Data,” Kaggle.com, 2024. Available: https://www.kaggle.com/dataset/prasad22/weather-data

Gantt Chart

https://gtvault-my.sharepoint.com/:x:/g/personal/clolley3_gatech_edu/EXdscbhyK1pFmTAfbqBlGB8BNwvi2sRJRQbh82muJ2Q8e=AXqmsq&nav=MTVfezAwMDAwMDAwLTAwMDEtMDAwMC0wMDAwLTAwMDAwMDAwMDAwMH0

Contribution Table

Name	Midterm Checkpoint Contribution
Allen Gao	Linear Regression, Supervised Visualizations, Results + Discussion Intro, & Supervised Methods
Chase Lolley	Data Cleaning, PCA w/ Airline Data, Handling Github Pages, Unsupervised Methods, Problem Definition, & Unsupervised Visualizations
Shahameel Naseem	Linear Regression, Supervised Methods, Supervised Visualizations
Sidney Wise	PCA w/ Weather Data, Intro + Background, Literature Review, Dataset Description, Unsupervised Visualizations, & References
Steven Haener	Analysis, Next Steps, & Gantt Chart

YouTube Video/Slideshow

https://youtu.be/epY78fKEqLc

ml_cs7641_group30_project_spring2025

ml_cs7641_group30_project_spring2025