Final

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 delay prediction model that contains an end-to-end pipeline including data preprocessing/ cleaning, feature engineering and feature selection from the flight and weather features, and then finally inputting the cleaned data into various regression models to see if we can get relatively accurate departure delay predictions. The combined feature 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 models, 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

For this project we will be using the “Historical Flight Delay and Weather Data USA” dataset. This data set merged flight data with weather data across US airports [8]. The flight data within the set is filled with airport and plane details containing flight numbers, carrier codes, scheduled arrival/

departures, etc. The weather data is the average hourly weather forecast across the weather stations closet to the original airport and the destination airport, and it includes precipitation, visibility, wind speed, etc. Every row contains both the flight and weather data along with the flight's departure delay time, which was used for ground truth prediction values. The dataset was collected across seven months in 2019 with each month having its own csv file with up to 700,000 rows of data. For the project we only used the May 2019 file.

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 see if one of the isolated datasets 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 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 feature selection was also done to see which features were statistically more valuable for prediction.

Supervised Methods

Linear Regression

To model flight delays, we applied linear regression using both closed-form and gradient descent (GD) solutions. We trained the models on multiple datasets: cleaned airline data, cleaned weather data, PCA-transformed airline data, PCA-transformed weather data, and a combination of both cleaned and PCA-transformed datasets.

For the closed-form solution, we computed the regression coefficients using the normal equation, minimizing the least squares error. In addition, we used gradient descent to iteratively optimize the model, adjusting the learning rate and convergence criteria to enhance stability and performance. We evaluated the models based on root mean squared error (RMSE), with the goal of comparing the predictive accuracy across different data transformations and combinations.

All datasets were preprocessed to handle missing values, normalize numerical features, and properly encode categorical variables before training the models.

Ridge Regression

To further enhance our linear regression model, we introduced a constraint term. This is also known as ridge regression and aims to combat overfitting by reducing the size of our coefficients, or weights, in the linear regression model. We introduce a bit of bias with this term, but often greatly reduce variance, improving out-of-sample predictive accuracy.

To implement this effectively, we used cross validation to find the best lambda, which equates to our constraint size. This involved training our model using a certain value from a list of possible lambdas and seeing which one has the lowest root mean squared error (RMSE). This was done on both closed form and gradient descent approaches.

Just as we did with our linear regression model, all datasets were preprocessed to handle missing values, normalize numerical features, and properly encode categorical variables before training the models.

Results + Discussion

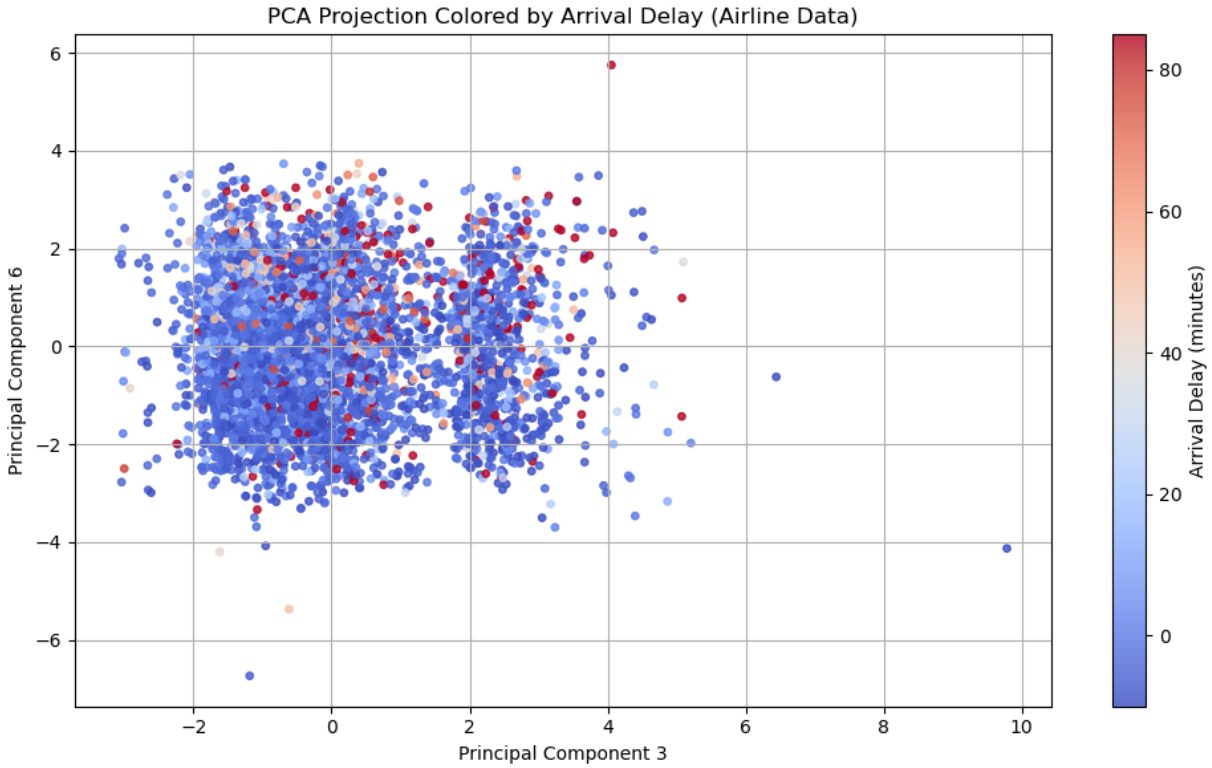
By using PCA for initial preprocessing, we were able to engineer the most relevant features that contribute to departure delays. We will be using MSE and RMSE metrics to evaluate the capability of our linear regression models to predict flight arrival delays. An indication that our models performed well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

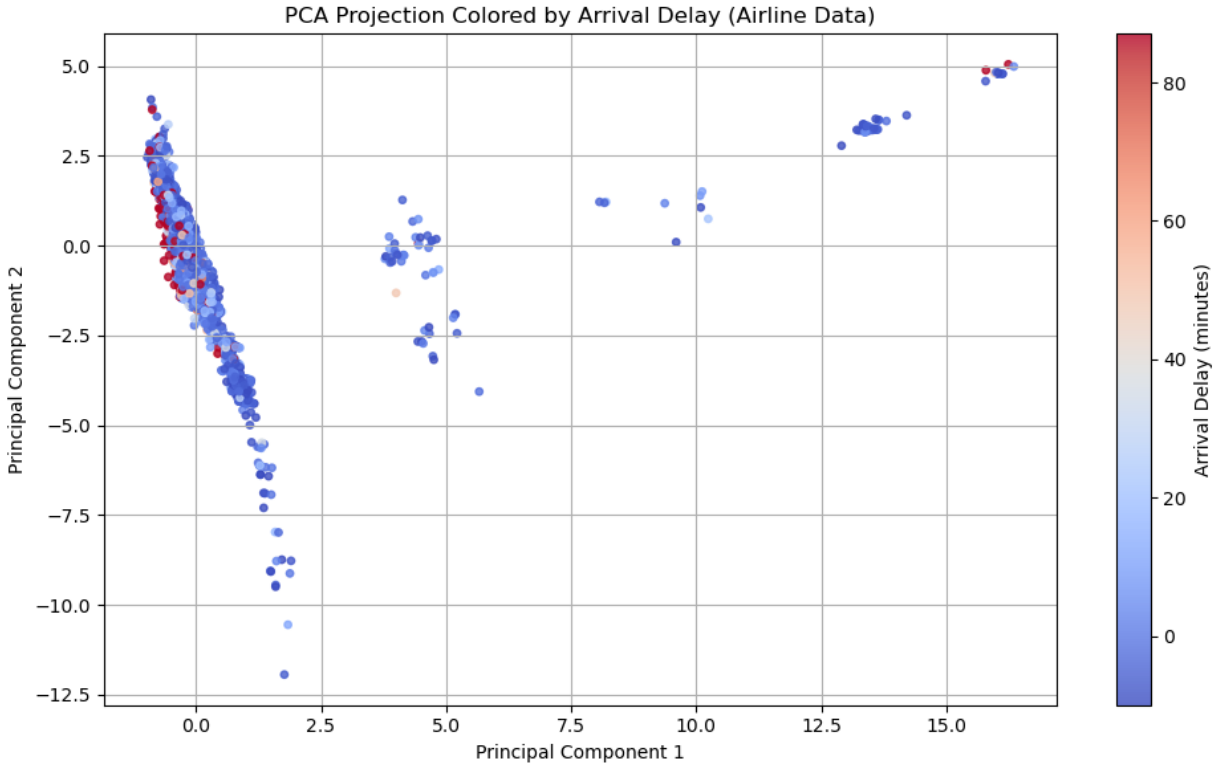
Visualizations

Unsupervised - Principal Component Analysis (PCA)

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.

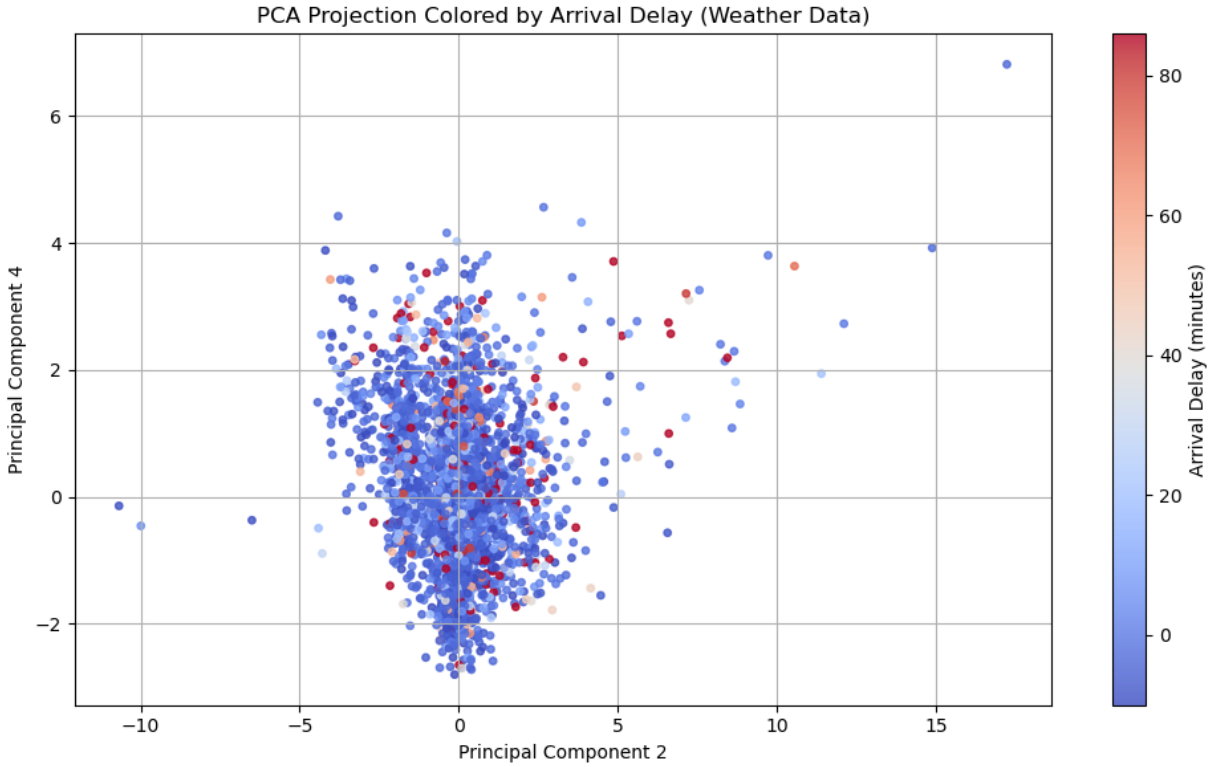


Airline PCA top 2 principal components correlated with departure delay

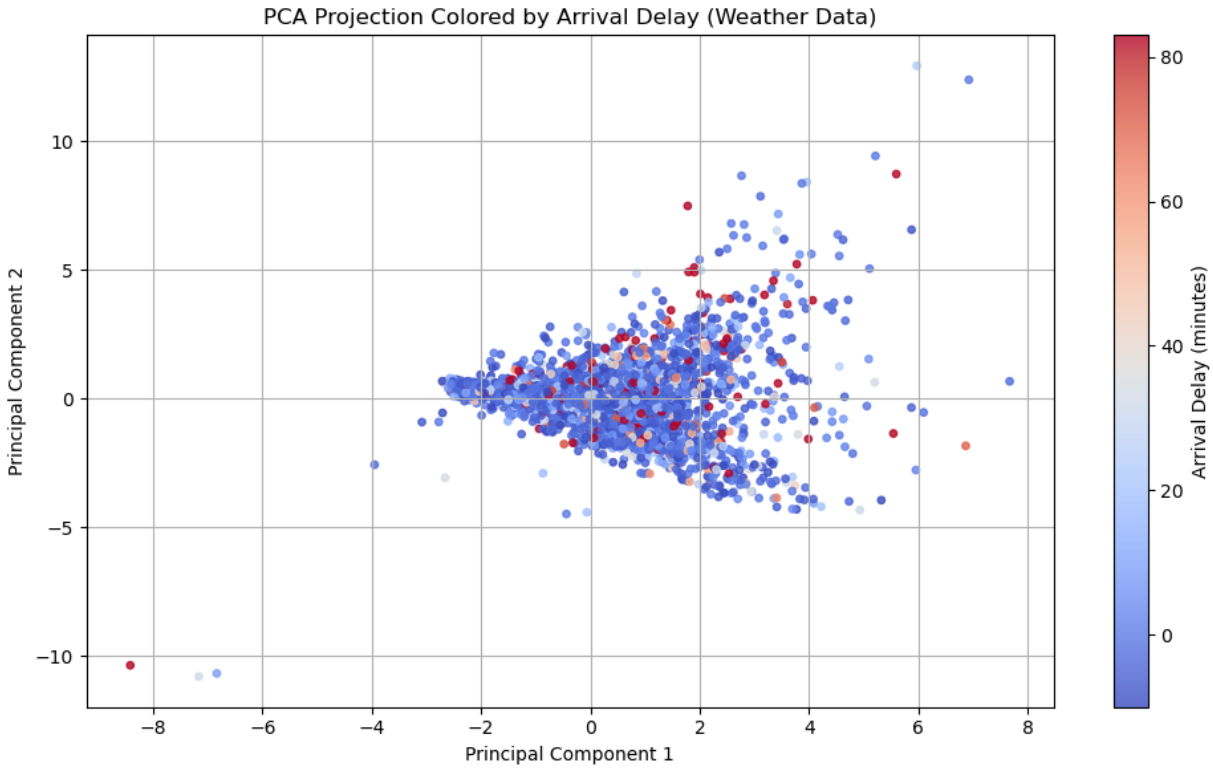


Airline PCA top 2 principal components with highest variance

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. Standard scaling was applied to normalize the features before PCA and the data was processed with PCA to retain 90% of the variance. The weather PCA projection visualizations demonstrate that the pipeline successfully captured meaningful structure within the weather data. The only columns in the weather data removed were the weather stations IDs while the dry bulb temperature, precipitation, pressure, visibility, etc., are all used as they have a high correlation with flight delays.



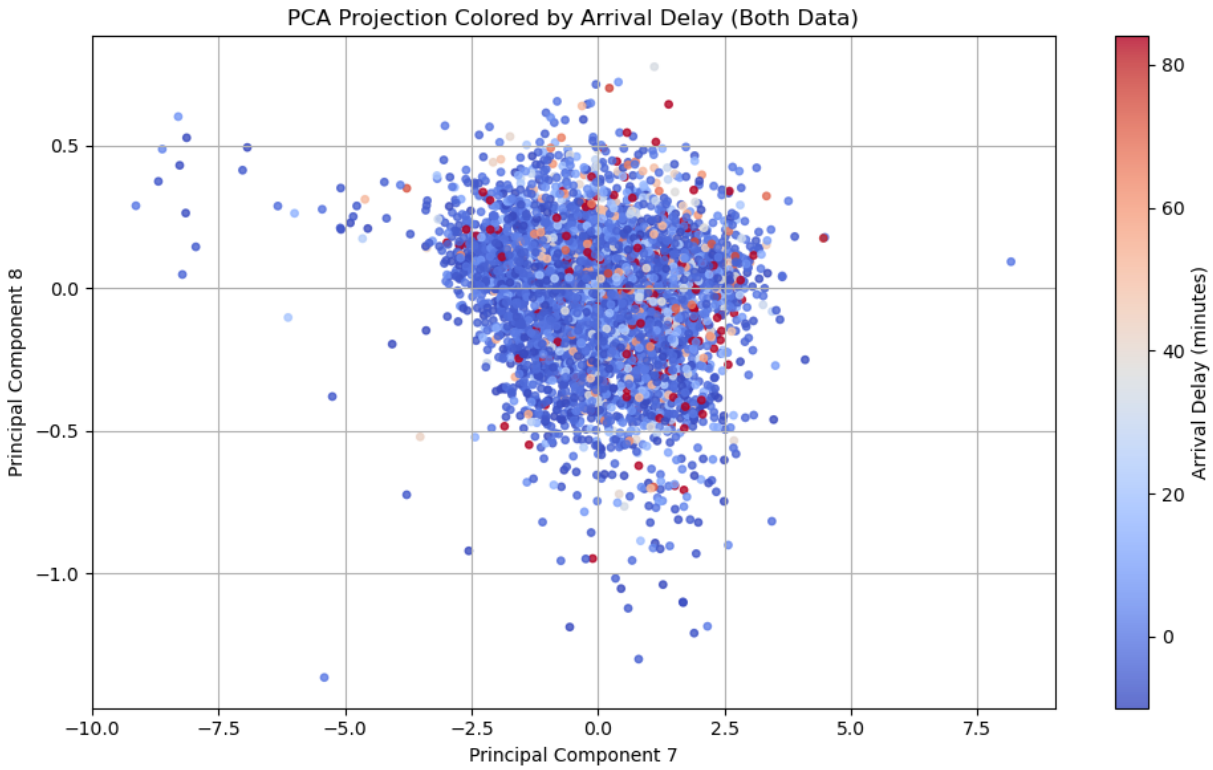
Weather PCA top 2 principal components correlated with departure delay



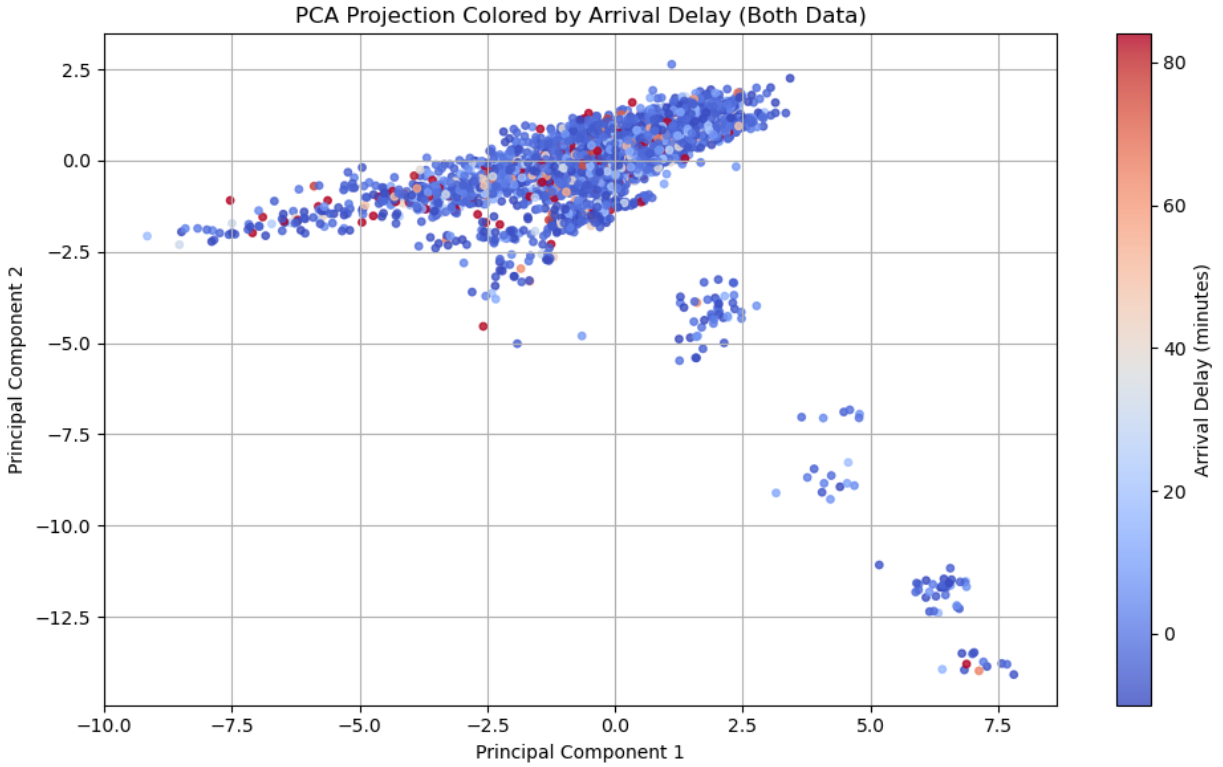
Weather PCA top 2 principal components with highest variance

Principal Component Analysis (PCA) was conducted using features from both airline and weather datasets to explore underlying structure in relation to arrival delays. The projections reveal that even when combining both domains, clusters and gradients are visible, indicating meaningful

variance across principal components. However, it's important to note that variance structure alone does not equate to predictive utility; components that explain the most variance (as shown in the top variant PCs) are not necessarily those most correlated with arrival delay (as seen in the delay-colored plots). This insight underscores the need for regularization techniques like ridge regression, which can better handle collinearity and focus on predictive signal over raw variance, compared to standard linear regression.



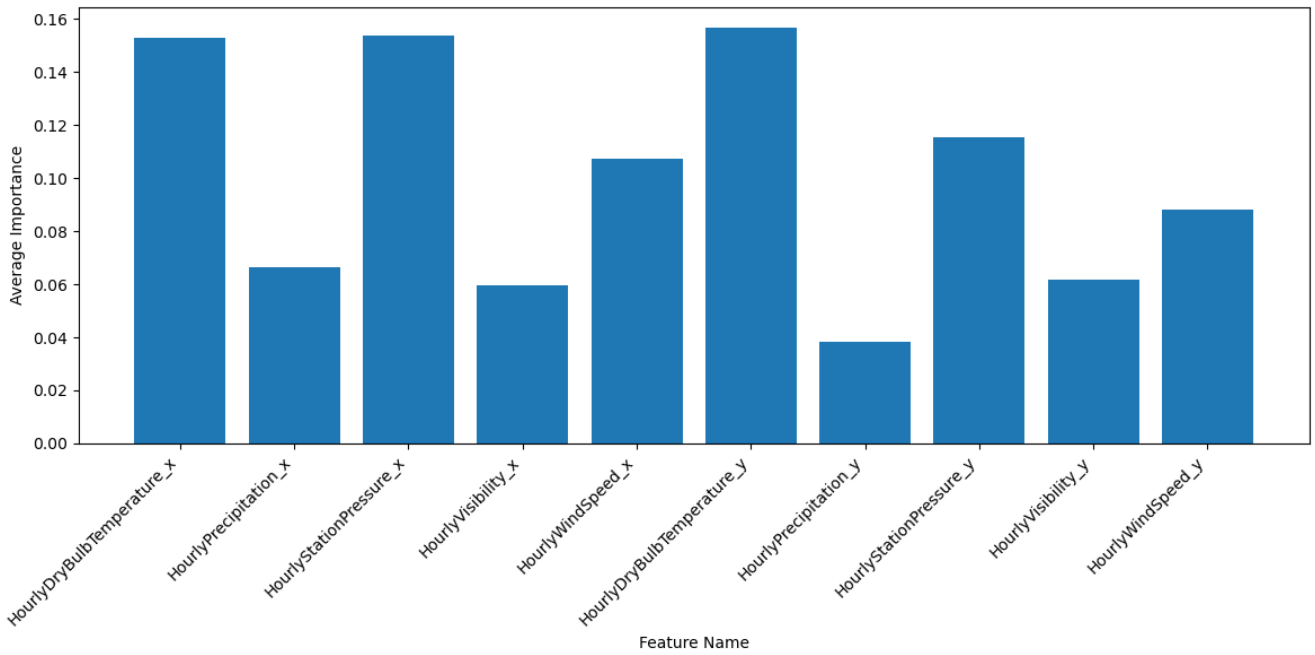
Both PCA top 2 principal components correlated with departure delay



Both PCA top 2 principal components with highest variance

Supervised - Random Forest Feature Selection

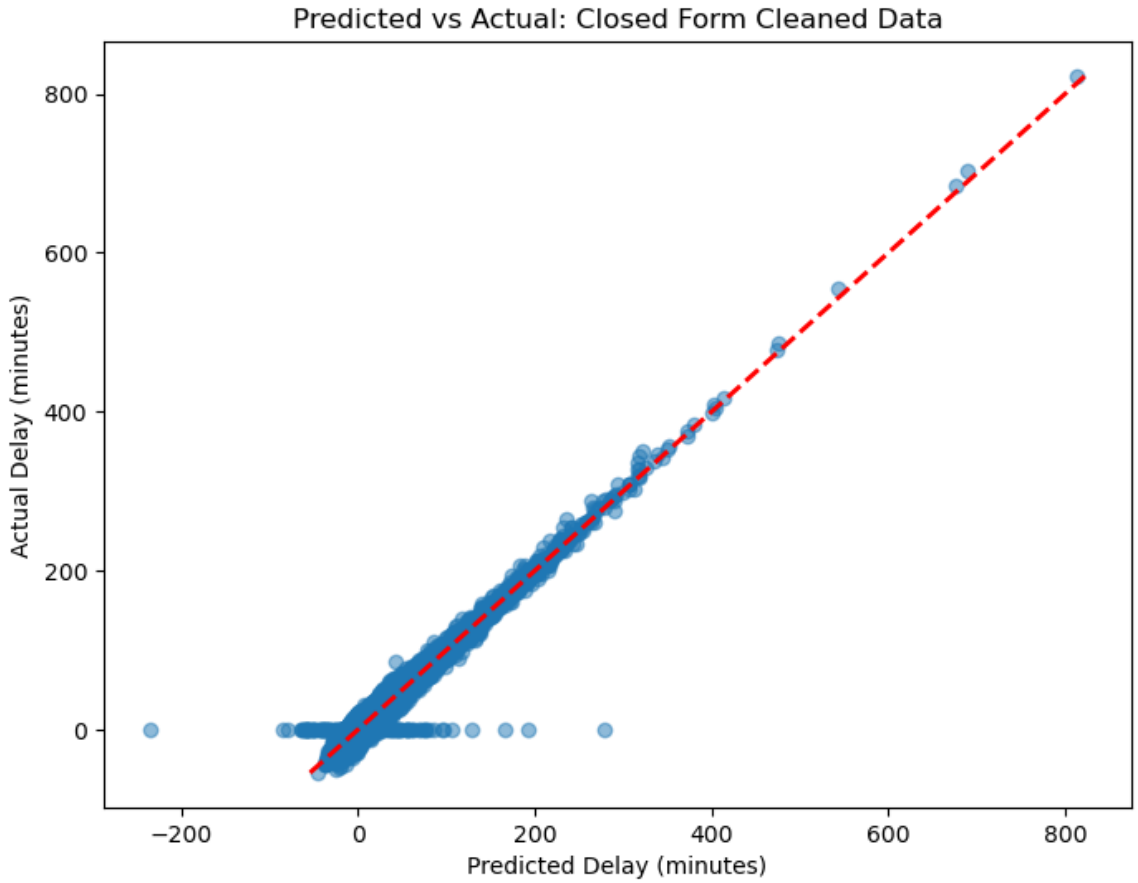
To gain an understanding of the impact of each feature on flight departure delays, random forest search was implemented. The importance values were averaged across all trees to avoid overfitting. Features such as temperature, pressure, and wind speed at both the origin and destination airports are, on average, more important in predicting departure delays. With this, there could be some potential in selecting only these features for regression models. However, these results are not intuitive, as visibility has been cited as a large contributor to flight delays [7].



Feature selection using weather data

Supervised - Linear Regression

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.



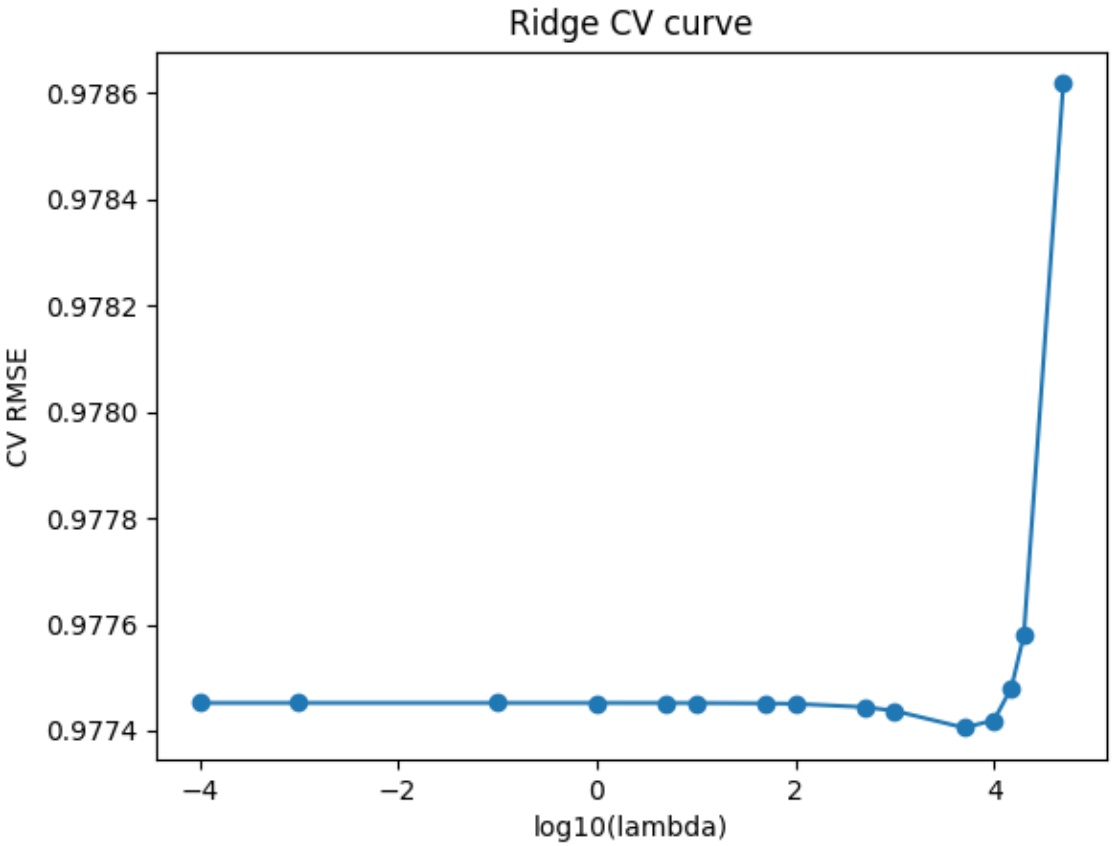
Closed-form solution on cleaned airline data: predicted vs. actual flight delays, demonstrating a tighter fit along the ideal line, though some overprediction occurs for smaller delays

To explore potential improvements, additional models were trained using PCA-transformed airline data, PCA-transformed weather data, and a combination of both. Despite these efforts, none outperformed the original closed-form solution on cleaned airline data. The best RMSE from the new methods—48.168 minutes using a closed-form model on combined weather and PCA-transformed airline data—was still significantly worse than the cleaned airline closed-form baseline (not shown in this table). Notably, all PCA-based models consistently hovered around similar RMSE values (~48–49 minutes), suggesting that dimensionality reduction may have limited their capacity to capture complex delay patterns. Even with the inclusion of weather data, performance gains were marginal, indicating that either more expressive modeling techniques or richer, non-linear models may be required to achieve substantial improvements.

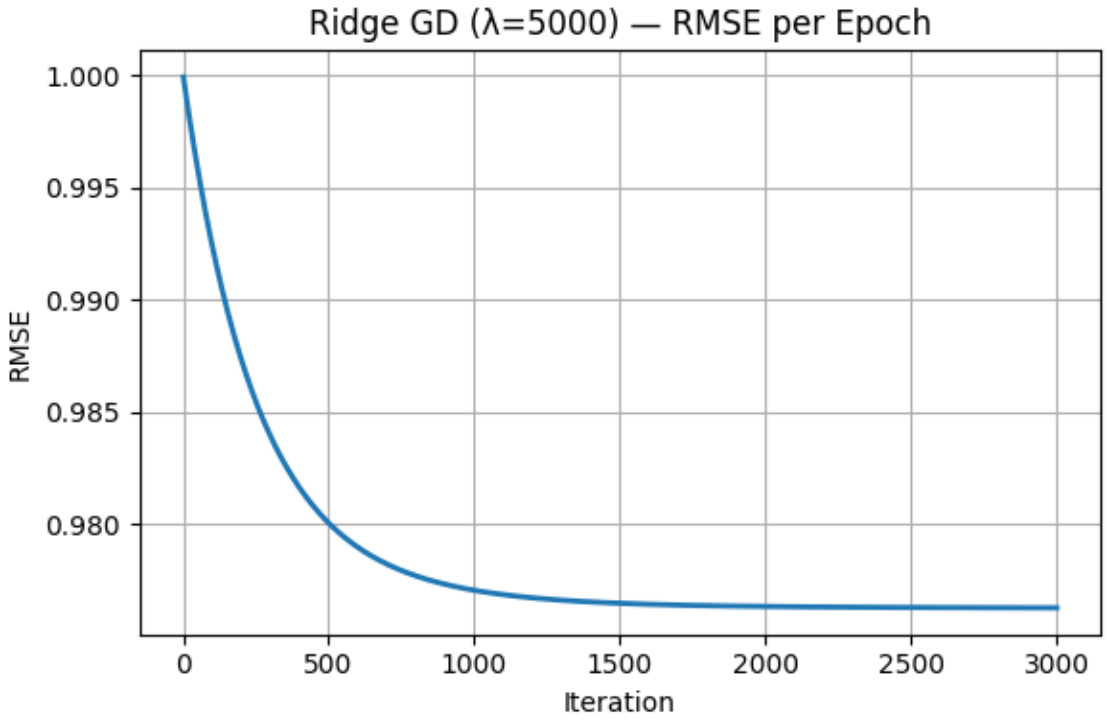
Supervised - Ridge Regression

We found that during our hyperparameter search using cross validation, larger lambdas led to a lower RMSE. The large majority of flights have a very small delay and therefore most data points will be close to zero. Creating a model that always predicts zero in this situation will mean that your RMSE will be low, even though there are still larger delays that are highly mislabeled. This leads to a larger error for the more delayed flights, which is what we were trying to predict. We

decided to lower the constraint value and accept a higher RMSE for smaller delays as well as to further preprocess the data. This involved removing flights with a delay of less than 5 minutes or more than 180 minutes. We saw some improvements in our RMSE values but a further look into our data and the current approach would be warranted for future works.



Cross validation curve of our RMSE decreasing as our lambda increases



Converging

RMSE of our gradient descent approach

Quantitative Metrics

Method	Dataset	RMSE (minutes)
LR Closed Form	Cleaned Airline	6.135
LR Closed Form	PCA Airline	48.402
LR Gradient Descent	PCA Airline	49.333
LR Closed Form	PCA Weather	48.489
LR Gradient Descent	PCA Weather	48.507
LR Closed Form	PCA Combined	48.168
LR Gradient Descent	PCA Combined	48.838
RR Closed Form	PCA Airline	47.204
RR Gradient Descent	PCA Airline	46.755
RR Closed Form	PCA Weather	47.512

Method	Dataset	RMSE (minutes)
RR Gradient Descent	PCA Weather	47.988
RR Closed Form	PCA Combined	47.124
RR Gradient Descent	PCA Combined	47.355

Analysis + Comparision

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 available 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 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 reasonable performance in modeling flight delays, especially when using the closed-form solution on cleaned airline data. This configuration yielded the best results, with relatively low RMSE and predictions that closely aligned with actual values. However, the model struggled to capture extreme delays accurately, often underpredicting them, and showed signs of overfitting or poor generalization when trained on reduced or transformed datasets like PCA.

We also implemented ridge regression to address potential overfitting and multicollinearity, but it offered only marginal improvement. This minimal performance gain suggests that the limitations of linear models may stem not just from overfitting, but from the underlying nonlinear structure of the data. These results indicate that capturing complex interactions and patterns likely requires nonlinear modeling approaches to better reflect the true behavior of flight delays.

Ridge regression could potentially improve performance by introducing regularization, which penalizes large coefficient values and reduces model complexity. This is especially beneficial in high-dimensional datasets or when features are correlated, as is often the case in airline and weather data. By adding a regularization term, ridge regression can help mitigate overfitting, improve generalization to unseen data, and offer more stable solutions when the feature matrix is ill-conditioned or near-singular, which can occur after PCA or with noisy features.

Next Steps

Our findings suggest several promising directions for improving flight delay prediction. While PCA helped reduce dimensionality, it became clear that variance does not necessarily translate to predictive power. We performed preliminary feature selection, which identified features with strong relationships to the target variable, but these insights were not yet leveraged in training. A valuable next step would be to implement mutual information-based selection or recursive feature elimination (RFE) to explicitly train on the most relevant features, potentially preserving signal lost during PCA.

Since ridge regression offered minimal gains over standard linear regression, the persistent underprediction of extreme delays and overall high RMSE point to the need for nonlinear models that can better capture complex patterns and interactions. In future work, experiments could be done with models such as:

- Random Forest Regressors for handling non-linearities and feature importance interpretation.
- Support Vector Regression (SVR) with non-linear kernels to model complex delay behaviors.
- Neural Networks, especially shallow multi-layer perceptrons, for modeling intricate relationships if sufficient data and compute resources are available.

Furthermore, since combining airline and weather data didn't yield significant improvements, future work should prioritize intelligent feature engineering via creating derived features like

temperature-wind interactions, categorical time bins (e.g., rush hour vs. late night), or holiday indicators. These could capture nuances not visible in raw data.

Lastly, working with large datasets highlighted the importance of streamlined, scalable preprocessing pipelines. Automating tasks like outlier detection, feature normalization, and missing data imputation will help ensure future models remain robust and efficient, especially when expanding to even larger, more diverse datasets.

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] I. Gheorghiu, "Historical Flight Delay and Weather Data USA," Kaggle.com, 2020. Available: https://www.kaggle.com/datasets/ioanagheorghiu/historical-flight-and-weather-data/data.

Gantt Chart

https://gtvault-my.sharepoint.com/:x:/g/personal/clolley3_gatech_edu/EXdscbhyK1pFmTAfbqBlGB8BNwvi2sRJRQbh82muJ2Q8Yw?e=AXqmsq&nav=MTVfezAwMDAwMDAwLTAwMDEtMDAwMC0wMDAwLTAwMDAwMDAwMDAwMH0

Contribution Table

Name	Midterm Checkpoint Contribution
Allen Gao	Supervised Visualizations, Results + Discussion Intro, & Supervised Feature Selection
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, Next Steps
Sidney Wise	PCA w/ Weather Data, Intro + Background, Literature Review, Dataset Description, Unsupervised Visualizations, & References
Steven Haener	Ridge Regression, Analysis + Comparison, & Gantt Chart

YouTube Video/Slideshow

https://youtu.be/epY78fKEqLc

ml_cs7641_group30_project_spring2025

ml_cs7641_group30_project_spring2025