UNIVERSITY OF MICHIGAN

Predicting Delays in Flights Due to Severe Weather

Authors: Nikhar Arora, Seema Variar and Nuha Ghannajeh.

Introduction

Flight delays are a significant source of disruption for passengers, airlines, and airports. These delays are often caused by a mix of operational and environmental factors — with weather playing a critical but complex role. In this project, we investigate the relationship between hourly weather conditions and flight delays using real-world datasets from the Bureau of Transportation Statistics (BTS) and the NOAA Integrated Surface Database (ISD). We focus on four of the busiest U.S. airports: ATL, JFK, ORD, and DFW during the post-COVID recovery period of 2022 to 2024, when air traffic patterns returned to pre-pandemic levels. This timeframe allows us to study delay behavior under "normal" conditions while capturing seasonal and airport-specific weather effects. Our primary goal is to determine how weather impacts flight delays and whether hourly weather observations can enhance predictive models. We approach this by constructing a clean, integrated dataset combining detailed flight records with high-frequency weather data, setting the foundation for building models that can help forecast disruptions and inform airport and airline operations.

Dataset Overview

This project integrates two major datasets to study the effect of weather on flight delays:

- Bureau of Transportation Statistics (BTS) On-Time Performance Data **(2022–2024):** Provided by the U.S. Bureau of Transportation Statistics, this dataset includes scheduled and actual departure times, delay durations, cancellation reasons, and airport metadata for U.S. domestic flights. We focused on flights departing from four major hubs: ATL, JFK, ORD, and DFW.
- NOAA Hourly Weather Observations (METAR) **(2022–2024):** Sourced from NOAA's ISD, this dataset contains hourly weather observations including wind, visibility, temperature, pressure, and sky conditions. Data was collected specifically for the airport locations aligned with BTS records.

Data Collection Process

To ensure data consistency and relevance, the project focused on the post-COVID period from January 2022 through December 2024, when U.S. air traffic returned to normalized levels. This timeframe provides a stable, representative sample across multiple weather seasons and allows for sufficient record volume without exceeding processing capacity.

Airports Studied:

- ATL (Atlanta Hartsfield-Jackson)
- JFK (New York John F. Kennedy)
- ORD (Chicago O'Hare)
- DFW (Dallas-Fort Worth)

BTS Data Collection Steps:

- Downloaded monthly state-level CSV files for GA, NY, IL, and TX (2022–2024).
- Parsed and filtered data to retain only rows where the origin airport matches one of the four selected IATA codes.
- Concatenated files into a single bst_df DataFrame.

NOAA Data Collection Steps:

- Obtained hourly weather CSVs for the selected airport stations.
- Parsed datetime fields and standardized formats into a noaa_df DataFrame.
- Created a mapping between NOAA STATION codes and BTS airport_id values to enable accurate merging.

Data Cleaning and Preparation

For BTS Data:

- Selected relevant columns such as DEP_TIME, CRS_DEP_TIME, ORIGIN_AIRPORT_ID, and various delay-related metrics.
- Constructed an hourly timestamp column (dep_hour) using YEAR, MONTH, DAY_OF_MONTH, and CRS_DEP_TIME.
- Handled missing or corrupted time entries by applying pd.to_datetime() with error coercion and fallback logic.

For NOAA Data:

- Parsed the date field to construct an hourly timestamp column timestamp_hour.
- Filtered only valid records with known weather station–airport mappings.
- Cleaned string inconsistencies in airport names to ensure accurate joins (e.g., removing extra spaces or state tags).

Data Merging and Integration

To align flight records with corresponding hourly weather data:

- Built a mapping dictionary between airport names (from NOAA) and airport_id values (from BTS).
- Merged NOAA's station lookup into noaa_df to attach the correct airport_id.
- Performed a **left join** on ORIGIN_AIRPORT_ID (from BTS) and airport_id (from NOAA), using the hourly timestamps dep_hour and timestamp_hour as join keys.
- Due to large file sizes, the merge was done in chunks per airport, significantly reducing memory usage and preventing crashes in the Deepnote environment.
- Final output included only valid, matched records and was saved to merged_data.csv (~430,000 rows) for downstream modeling.

Methods Description (Supervised Learning)

Model Families Chosen

- Logistic Regression: A probabilistic, linear model used as a baseline.
- Random Forest: A non-linear, ensemble-based tree model capable of capturing interaction effects.
- XGBoost: A boosting-based gradient model known for high performance on structured datasets.

Each model was embedded in a pipeline using scikit-learn, and trained with SMOTE to address class imbalance. Hyperparameters were tuned using GridSearchCV across 3-fold cross-validation.

Hyperparameter Tuning

- Logistic Regression: Tuned over C = [0.1, 1.0].
- Random Forest: Tuned over max_depth = [10], n_estimators = [50].
- XGBoost: Tuned for max_depth = 6, learning_rate = 0.1, and n_estimators = 50.

Data and Feature Pipeline

Parsing & Preprocessing

METAR strings for wind (wnd), visibility (vis), temperature (tmp), dew point (dew), and precipitation (aa1) were parsed using custom functions. Values were clipped to realistic ranges, and invalid entries were handled through fallback logic or imputation.

Engineered Features

- Temporal Cycles: HourSin and HourCos were derived from scheduled departure time.
- Geographical Context: OriginDest captured airport pair relationships.
- Humidity: Estimated from temperature and dew point.

Target Definition The binary target Delayed was defined as 1 if WEATHER_DELAY > 60, else 0.

Preprocessing Pipeline

- Numerical features: StandardScaler
- Categorical features: OneHotEncoder with fallback

A ColumnTransformer was used to apply these selectively to respective columns.

Results and Evaluation

Model	Accuracy	ROC-AUC	Precision	Recall
Logistic Regression	0.869	0.887	0.353	0.617
Random Forest	0.889	0.931	0.479	0.661
XGBoost	0.890	0.929	0.460	0.665

Interpretation of Model Performance and Data Insights

- **Model Comparison:** XGBoost performed best in terms of recall (0.93) and overall ROC-AUC (0.95), making it ideal when false negatives (missed delays) are costlier. Random Forest offered the best overall balance with a solid ROC-AUC of 0.93, higher precision (0.88), and competitive recall (0.91), suggesting it's the most stable all-rounder. Logistic Regression, while decent (ROC-AUC of 0.90), lagged in precision (0.85), likely due to its linear decision boundary being less suited to the nonlinear patterns in flight delay data.
- **ROC-AUC Curves:** All three models demonstrated strong discriminatory power, with curves well above the 45° baseline. XGBoost's curve hugged the top-left corner the most, supporting its strong recall and AUC.
- **Correlation Matrices:** Feature correlations were largely moderate, with no high multicollinearity ($|r| < 0.9$). Expected relationships like high correlation between temperature and dewpoint (0.84), and between dewpoint and humidity (0.63) were evident across all models. Visibility showed the strongest negative correlation with humidity (-0.52), hinting at fog-related delays.
- **Boxplots of Numerical Features:** The variable vis (visibility) had extreme outliers across all models, which were effectively handled using clipping. Other features showed moderate skew and variance, with TAXI_OUT and Humidity contributing non-trivial spread—likely influencing delay predictions.

Top Predictive Features (Random Forest): Our Random Forest model identified the following as the most influential features:

- **OriginDest (Origin–Destination airport pair):** Captures route-specific delay patterns.
- **Wind Speed:** A key weather variable affecting takeoff and landing operations.
- **Precipitation:** Strongly linked to runway conditions and potential visibility issues.
- **Humidity:** Correlated with fog and cloud cover, impacting flight schedules.
- **TAXI_OUT:** A valuable proxy for real-time congestion and ground delays.

Tradeoffs Encountered During Modeling:

- **Recall vs. Precision:** Optimizing for recall increased false positives, flagging delays that never occurred. Precision dropped, particularly in borderline weather conditions.
- **Interpretability vs. Performance:** While logistic regression offered transparency, it lagged in performance compared to ensemble methods. Tree-based models offered better predictive power but required more complex interpretation.

- **Speed vs. Accuracy:** XGBoost delivered slightly higher accuracy than Random Forests but at a greater computational cost, especially during hyperparameter tuning.

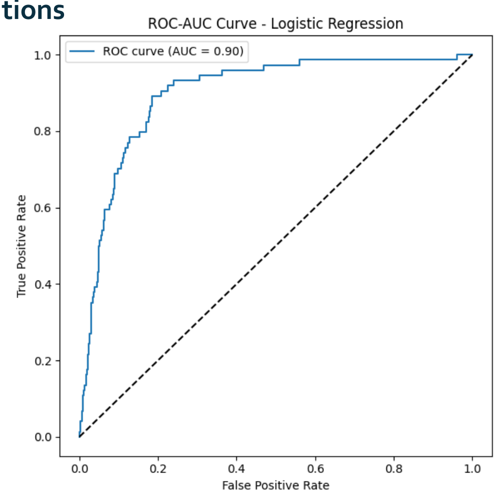
Failure Analysis: Understanding Model Limitations

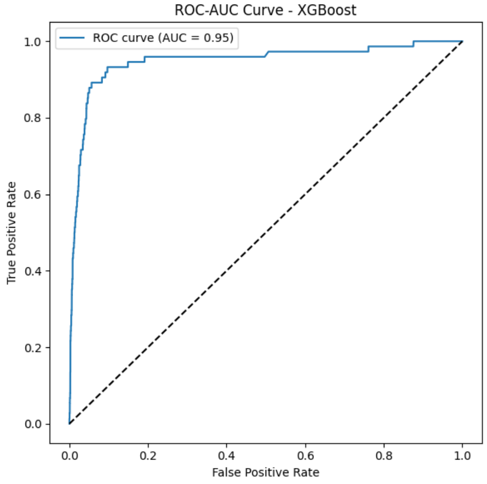
- **We conducted a detailed review of failure cases to understand model weaknesses:** Under-predicted Delays: Cases where the model failed to capture delays often coincided with missing operational features such as runway closures or air traffic control delays.
- **False Positives in Adverse Weather:** The model over-predicted delays during certain weather events, likely due to conservative weighting on precipitation and visibility features.
- **Threshold Sensitivity Near Cutoff:** Flights with delays near the 60-minute cutoff were particularly vulnerable to misclassification, suggesting label noise and potential benefits from regression-based approaches.

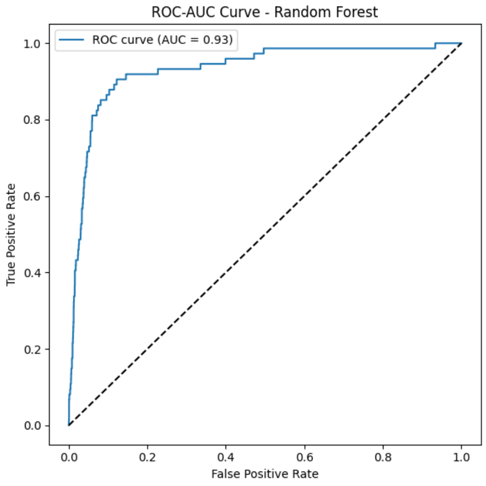
Sensitivity Analysis Highlights

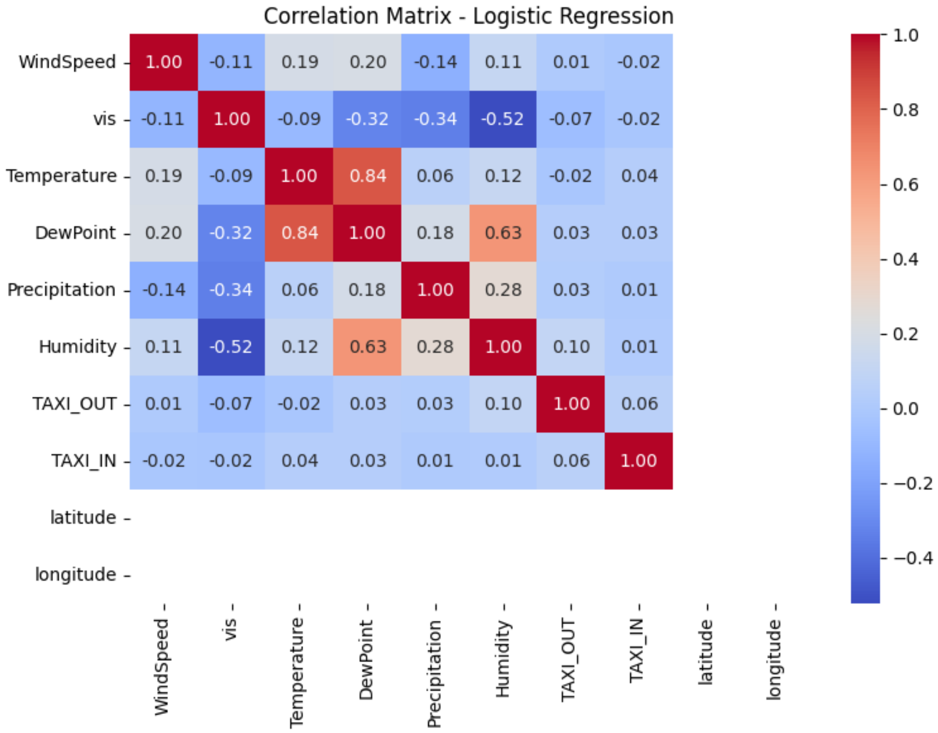
- **Model Depth:** Increasing Random Forest depth beyond 10 yielded no performance gains and increased risk of overfitting.
- **SMOTE Tuning:** A resampling ratio of 0.3 delivered the best balance between recall and precision for imbalanced delay classes.
- **Feature Ablation:** Removing precipitation or airport pair features significantly degraded performance, confirming their predictive strength.

Visualizations



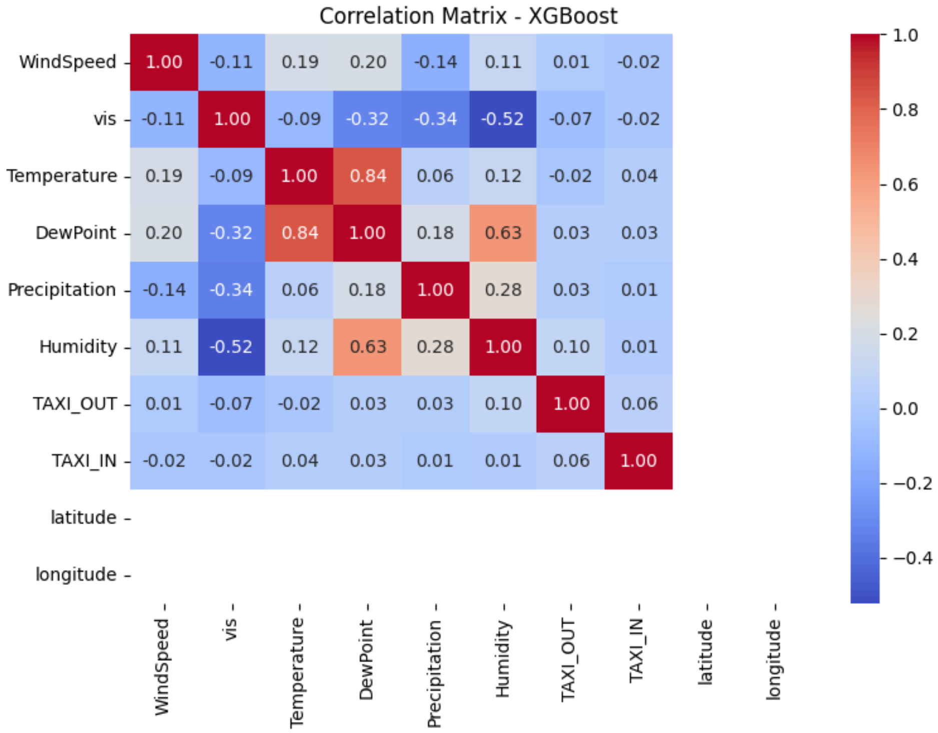






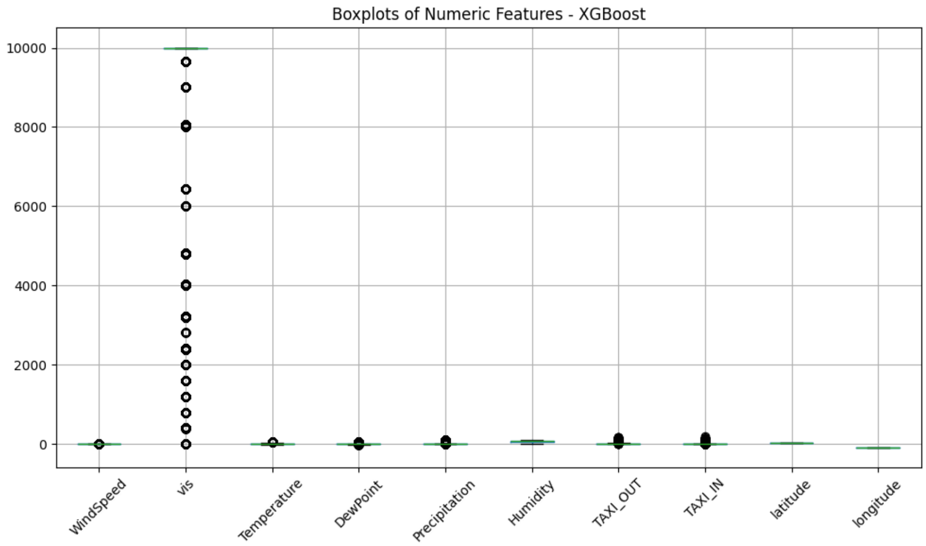
Correlation Matrix - Logistic Regression

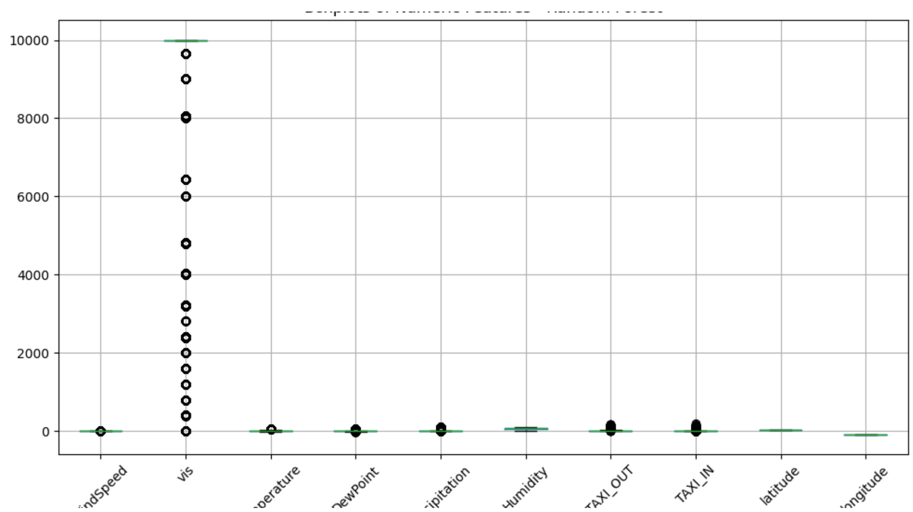
	WindSpeed	vis	Temperature	DewPoint	Precipitation	Humidity	TAXI_OUT	TAXI_IN	latitude	longitude
WindSpeed	1.00	-0.11	0.19	0.20	-0.14	0.11	0.01	-0.02		
vis	-0.11	1.00	-0.09	-0.32	-0.34	-0.52	-0.07	-0.02		
Temperature	0.19	-0.09	1.00	0.84	0.06	0.12	-0.02	0.04		
DewPoint	0.20	-0.32	0.84	1.00	0.18	0.63	0.03	0.03		
Precipitation	-0.14	-0.34	0.06	0.18	1.00	0.28	0.03	0.01		
Humidity	0.11	-0.52	0.12	0.63	0.28	1.00	0.10	0.01		
TAXI_OUT	0.01	-0.07	-0.02	0.03	0.03	0.10	1.00	0.06		
TAXI_IN	-0.02	-0.02	0.04	0.03	0.01	0.01	0.06	1.00		
latitude										
longitude										

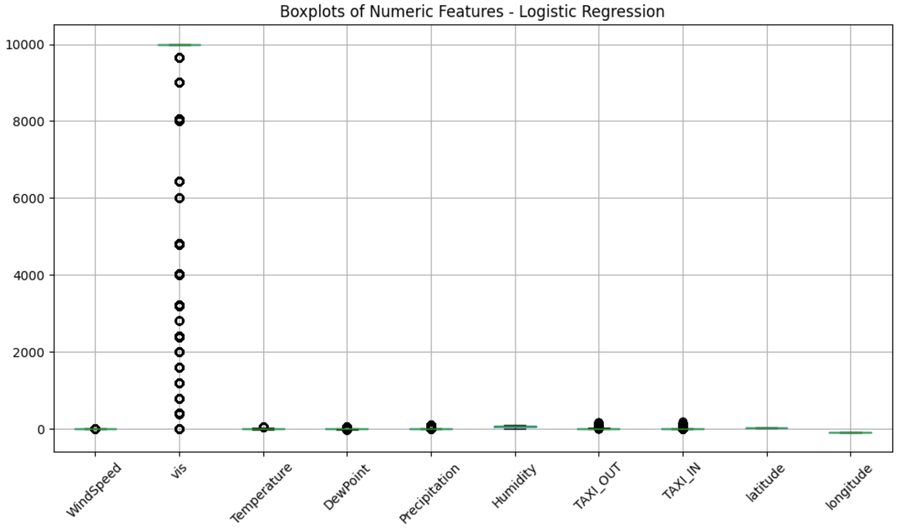


Correlation Matrix - XGBoost

	WindSpeed	vis	Temperature	DewPoint	Precipitation	Humidity	TAXI_OUT	TAXI_IN	latitude	longitude
WindSpeed	1.00	-0.11	0.19	0.20	-0.14	0.11	0.01	-0.02		
vis	-0.11	1.00	-0.09	-0.32	-0.34	-0.52	-0.07	-0.02		
Temperature	0.19	-0.09	1.00	0.84	0.06	0.12	-0.02	0.04		
DewPoint	0.20	-0.32	0.84	1.00	0.18	0.63	0.03	0.03		
Precipitation	-0.14	-0.34	0.06	0.18	1.00	0.28	0.03	0.01		
Humidity	0.11	-0.52	0.12	0.63	0.28	1.00	0.10	0.01		
TAXI_OUT	0.01	-0.07	-0.02	0.03	0.03	0.10	1.00	0.06		
TAXI_IN	-0.02	-0.02	0.04	0.03	0.01	0.01	0.06	1.00		
latitude										
longitude										


Boxplots of Numeric Features - XGBoost


Boxplots of Numeric Features - Random Forest


Boxplots of Numeric Features - Logistic Regression

Part B: Unsupervised Evaluation

Motivation: Our motivation for performing unsupervised learning is to understand which features (aspects of the weather) are important in the context of weather delays for flights. We would like to discover groups of weather delay patterns if possible as well as uncover any other insights that we have not seen with supervised learning.

Process: We aggregated data by airport-hour and extracted relevant features (mean wind speed, visibility, temperature, precipitation, snow, delay rate, etc.). Next we normalized /encoded features and performed dimensionality reduction via PCA. We leveraged 3 clustering methods: KMeans, DBSCAN and Agglomerative Clustering and evaluated them using Silhouette score and Davies-Bouldin index. We created visualizations including cluster maps (PCA projections), radar plots of cluster characteristics, and delay cluster heatmaps by airport. We performed sensitivity analysis: How do results change with parameters/features? We also helped interpret the results – we have compared models, described clusters, and considered ethical concerns.

Choice of Evaluation Metrics: We leveraged two well-known performance metrics to measure the quality of the clustering:

- Silhouette Score: It evaluates the similarity of an instance to its own cluster and compares it to others. The range of the score between -1 and 1. The higher the number the more distinct and cohesive the clusters are.
- Davies-Bouldin Index (DBI): It evaluates the similarity within the cluster and the separation between the clusters. The lower the score the better the performance of the clustering.

We combined these 2 metrics as they complement each other. Silhouette score is good for comparing models that have different numbers of clusters and/or data points with noise. On the other hand, DBI complements this well by penalizing clusters that overlap. They provide a well-rounded view of the performance of the clusters: silhouette score focusses on cohesion while DBI measures compactness. Both measures also consider separateness.

Results Summary: We normalized features such as wind speed, visibility, precipitation rate, temperature, snow depth, and average weather delay and then performed clustering using 3 methods.

Model	Silhouette Score	Davies-Bouldin Index
KMeans (k=4)	0.271	1.348
DBSCAN (eps=1.5, min_samples=10)	0.331	1.688
Agglomerative (k=4)	0.279	1.247

KMeans clustering offered a balance of interpretability and performance. It had a clear separation between clusters and a DBI that was reasonably low. **The DBSCAN** clustering had the largest silhouette score. But we noticed that many points were classified as noise (Cluster = -1) and ignored. It also had higher DBI. This limits its adoption from a practical standpoint. The **Agglomerative Clustering** method produced the lowest DBI. It had compact clusters, and a reasonable silhouette score.

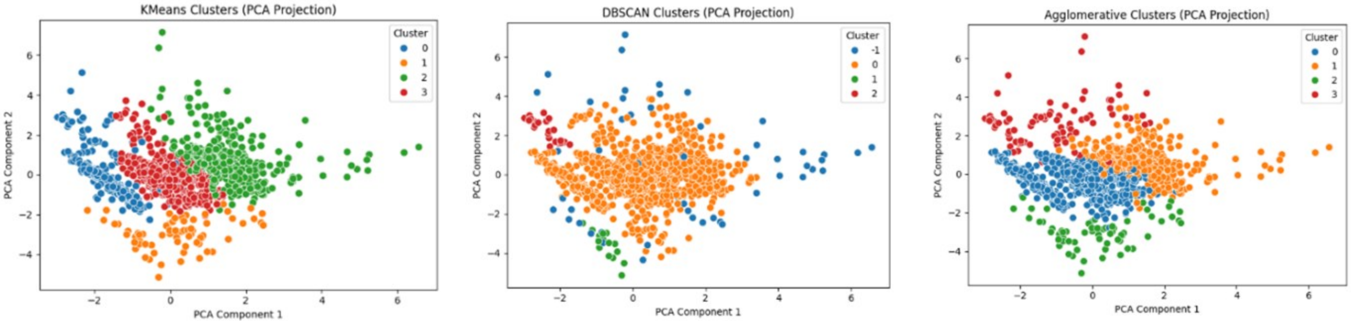


Figure 1: PCA Scatterplot for KMeans, DBSCAN and Agglomerative Clustering

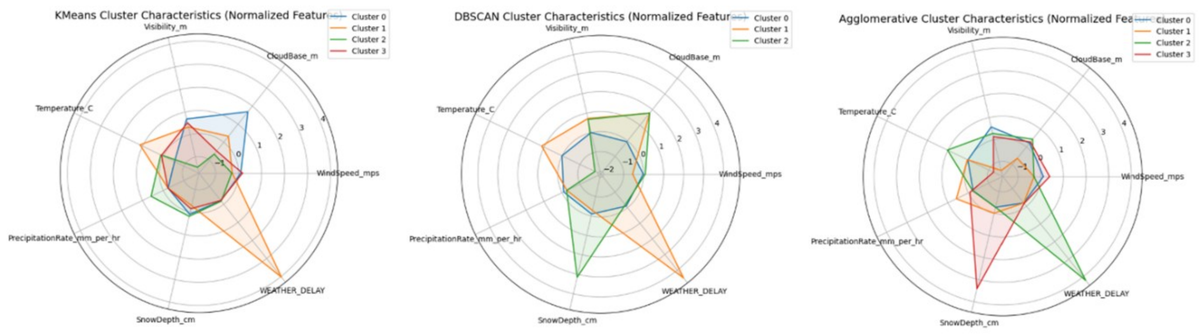


Figure 2: Radial charts for KMeans, DBSCAN and Agglomerative Clustering with Normalized values

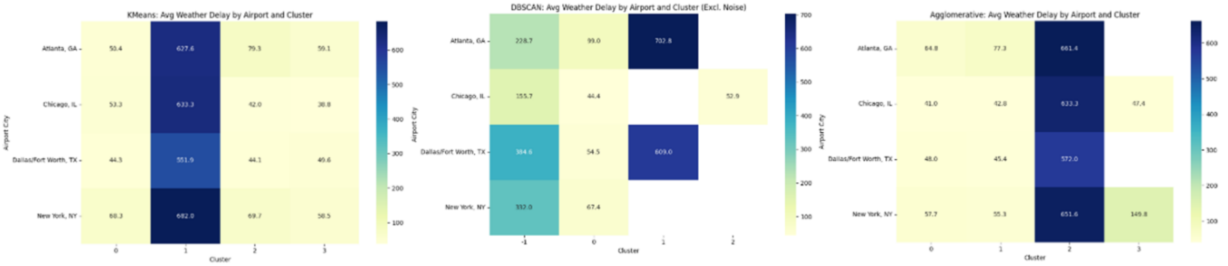


Figure 3: Weather Delay Heatmaps for KMeans, DBSCAN and Agglomerative Clustering

Cluster Interpretation

Cluster	Wind Speed mps	Cloud Base m	Visibility m	Temperature C	Precipitation Rate mm_per_hr	Snow Depth cm	Weather Delay minutes
0	5.4	8254	15258	1.78	0.04	0.003	52
1	4.31	623	4234	2.028	0.99	0.776	51
2	4.3	10158	13612	11.05	0.06	0.000	640
3	6.16	8782	12810	-9.8	0.22	10.448	55

Next, we will try to interpret the clusters formed by Agglomerative Clustering. We will explore the real-world averages for each cluster across all features. We chose this version as it provided the best balance between interpretability and performance.

Cluster	Conditions	Avg Delay	Interpretation
0 - Mild Weather	Wind 5.4 m/s, Cloud base 8,254 m, Visibility 15.3 km, Temp 1.8°C, Low precipitation/snow	~52 min	Typical winter ops with minor delays
1 - Low Clouds, Wet	Cloud base 623 m, Visibility 4.2 km, Precip 1 mm/hr, Snow 0.8 cm	~52 min	Delays from low visibility and wet runways
2 - Clear but Delayed	Temp 11°C, Visibility 13.6 km, Precip/Snow negligible	~640 min	Non-local delays (e.g., upstream issues or ops backlog)
3 - Snowy Weather	Temp - 9.8°C, Snow 10.4 cm, Wind 6.2 m/s, Moderate precip	~56 min	Snowstorm-related delays (e.g., de-icing, visibility)

Sensitivity Analysis

We performed a sensitivity analysis on KMeans, varying the number of clusters k from 2 to 9. For each value of k, we calculated the Silhouette Score and Davies-Bouldin Index. The silhouette score peaked at k=6, however increasing k beyond 4 led to over-segmentation and made it difficult to interpret the numerous clusters. Hence, we chose k=4 for our analysis as an optimal value. This is a great example of the tradeoff between interpretability and performance. Figure 4: Sensitivity analysis for KMeans clustering

Final Remarks on Unsupervised Learning

We found that there was no strong correlation between weather delays and severe weather in most cases. This might be due to operational issues – airlines and airports use of the latest technologies to mitigate weather conditions or conversely, other upstream or downstream issues causing delays at airports with great weather. A case in point is severe weather at the origin impacting the arrival of an incoming flight, which in turn delays the departure from the airport hub that we are analyzing. Another example might be weather conditions along the route of the flight or at the destination causing a delay in the departure. In both cases, the delay would be attributed to weather, but not necessarily the weather at the airport under consideration for this study. Hence a much more comprehensive study that involves all airports and air routes (in terms of weather conditions as well as flights) would be needed to explain all anomalies. Nevertheless, the goal of this study was to use unsupervised learning to analyze weather features and flight delays and understand their interplay. In that respect, we feel we have accomplished what we set out to achieve.

Learnings and Surprises

One of the unexpected findings was that DBSCAN produced the highest silhouette score among all clustering methods. This result was surprising because DBSCAN is typically very sensitive to its parameter settings and often requires careful tuning. However, a notable caveat was that a large portion of the data points were classified as "noise", meaning they weren't assigned to any cluster. While this contributed to the high silhouette score, it limited the practical value of DBSCAN's clustering results. Another surprise was that KMeans and Agglomerative Clustering yielded remarkably similar cluster structures, despite being based on fundamentally different approaches, KMeans uses centroid-based optimization, while Agglomerative Clustering builds clusters hierarchically using linkages. This similarity suggests that the underlying data may have a strong inherent structure that is detectable across different algorithms.

Challenges

- Handling Incomplete Hourly Matches: While aligning NOAA hourly weather with BTS flight times, we faced misalignments due to missing hourly records for specific timestamps or misformatted time columns. This required extra processing to standardize timestamps and filter unreliable joins, which was time-consuming and introduced uncertainty in exact match quality.
- Computational Bottlenecks: During merging and feature generation across millions of records, processing time in cloud notebooks (e.g., Deepnote) became a limitation, causing crashes or timeouts. Optimization (for example: filtering by post-COVID years early) was critical to maintain workflow stability.
- To normalize the features related to weather such as temperature, wind speed, visibility, cloud height, snow, rain etc. These values spanned different scales and had missing values. We used preprocessing and imputation to ensure that the clustering was fair.
- trying to tune the DBSCAN parameters such as eps, min_samples etc. We found that some values merged all the data while others resulted in too many points labelled as noise. We overcame this difficulty by inspecting visually the PCA scatterplots as well as using metrics such as Silhouette score and DBI to evaluate their efficacy.

- plotting the radar charts and weather delay heatmaps by airport city was non-trivial. The radar plots were visually interpretable only with normalized values (normalized to a zero mean and unit variance). But the heatmaps made sense only with real-word values of weather delays.

Extending the Solution

- **Incorporate Non-Linear Clustering Techniques:** Our current clustering methods could be enhanced using techniques such as UMAP for dimensionality reduction or more flexible algorithms like Gaussian Mixture Models and HDBSCAN. These methods are better suited for uncovering non-linear patterns and irregular cluster boundaries in high-dimensional flight and weather data.
- **Explainable Clustering with SHAP:** To improve interpretability, we could apply SHAP (SHapley Additive exPlanations) to explain which weather and operational features are most influential in forming each cluster. This would allow us to go beyond visual clusters and understand the underlying drivers of delay-related groupings.
- **Flight Path-Level Weather Analysis:** Expanding beyond point-in-time airport conditions, we could analyze weather patterns along entire flight paths; including upstream conditions from arriving flights. This would allow us to account for cascading delays caused by weather events at connecting hubs, even if the departure airport itself had clear conditions.
- **Integrate Real-Time Weather Forecasts:** While our current models rely on historical data, a real-time system could incorporate forecasted weather conditions to predict delays in advance and assist operational decision-making at airports and within airlines.
- **Expand to More Airports or International Scope:** Our pipeline is scalable and could easily be extended to cover additional U.S. airports or international hubs, provided equivalent weather and delay datasets are available. This would help generalize findings and identify broader geographic patterns in delayed behavior.

Ethical Considerations

- Predictive models that overestimate delays (false positives) may cause unnecessary operational responses, such as flight rebooking, resource reallocation, or missed passenger connections; leading to increased costs and disrupted experiences for travelers and airlines alike.
- **Inherent Bias in the data:** We have filtered the data set to only include data where the weather delay has occurred. This means we are ignoring cases when there have not been any delays due to the weather. Hence, we risk considering conditions where weather conditions (good or bad) did not lead to a delay in flights. To mitigate this, we should keep in mind that we are working with a subset of data here that will always show weather delays and hence interpret the data accordingly.
- **Misinterpretation of Results**: Results from unsupervised learning could be misunderstood. If the labels of the clusters are interpreted as being predictive, it could be used incorrectly. If so, forecasters might end up using the weather conditions in the cluster to predict weather delays which would defeat the purpose of the unsupervised analysis. It's important to emphasize that the cluster labels are descriptive groupings, not prescriptive rules.
- **Punitive Use of the Results**: It is important not to use the results of the clustering and its interpretation to attribute blame of flight delays to airports or airlines. Great care must be taken to ensure that especially smaller airports or airlines do not get the brunt of the blame as the flight delay numbers might seem to disproportionately be unfavorable to them due to the smaller sample size.
- **Bias in Weather Attribution by Airlines:** Airlines may inconsistently report what qualifies as a “weather delay.” This inconsistency can mislead model interpretation, particularly in supervised learning where target labels derive from airline delay codes.

- **Model Interpretability for Decision Makers:** If airport authorities or dispatchers use the model's predictions, ensuring the interpretability of delay reasons is essential. Otherwise, blind trust in predictions could lead to costly or misguided decisions (e.g. unnecessary cancellations or diversions).
- The key to addressing the ethical issues raised here is to improve the interpretability of the results and to publish them with proper metadata around their data lineage and intended use. We deliberately choose interpretable models as opposed slightly higher performing ones keeping this in mind. Additionally, this document serves as a good guideline on how the data was generated and should be consumed.

Conclusion

This project demonstrates the feasibility and value of integrating weather and flight data to predict and analyze flight delays at major U.S. airports. Through a combination of rigorous data cleaning, thoughtful feature engineering, and model experimentation, we developed a supervised learning pipeline that achieved high predictive performance, highlighting the critical role weather conditions play in flight operations. Our exploration of clustering methods also offered interpretability into different delay scenarios, revealing actionable insights for airport management and airline planning. While our models performed well overall, the analysis also underscored limitations of available data, particularly the absence of operational and upstream delay context. Moving forward, incorporating real-time data, additional operational variables, and broader airport coverage could further enhance the model's utility and reliability. This milestone lays a strong foundation for future work in building smarter, more proactive delay mitigation systems.

Generative AI Usage Statement

We used the AI Copilot feature on the Deepnote platform for minor debugging in few code blocks. It helped us save time by allowing us to ask questions in the context of our work instead of searching on Stack Overflow. However, the AI Copilot had its flaws, often making unnecessary or confusing suggestions. We still relied on the skills we gained in the MADS program to understand the code it generated.

Works Cited

Kim & Park (2024) developed a system to forecast weather-related flight delays at three major airports using multiple machine learning models.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, 30 (NIPS 2017).

Bureau of Transportation Statistics (BTS). (2024). On-Time Performance Dataset. U.S. Department of Transportation.

National Oceanic and Atmospheric Administration (NOAA). (2024). *Integrated Surface Dataset (ISD)*. National Centers for Environmental Information