Modeling Error Types in Aviation Weather Forecasts: Machine Learning Approach

Samuel Choi
Valencia High School, Yorba Linda, CA

Abstract

The accurate prediction of aviation weather conditions is crucial for the safety and efficiency of air travel. METAR (Meteorological Aerodrome Reports) and TAF (Terminal Aerodrome Forecasts) serve as key tools for aviation professionals to assess actual and forecasted weather conditions at airports. This study explores the application of machine learning techniques to evaluate the accuracy of weather forecasts using data from aviationweather.gov. Historical weather data are utilized to train and assess various machine learning models, including random forests, naïve Bayes, artificial neural networks, gradient boosting, support vector machines, and k-nearest neighbors to classify weather forecasts as correct, false alarms, or missed detections. The primary objective of this research is to identify the most influential features, such as specific airports or weather characteristics, that impact the accuracy of forecasts.

Keywords: weather forecast, aviation, machine learning models, accuracy of prediction, feature importance

1 Introduction

1.1 Background

Aviation safety and operational efficiency rely heavily on the accuracy of weather forecasts, making it crucial to understand and mitigate forecast errors. The present study utilizes a dataset combining weather data from METAR (Meteorological Aerodrome Reports) and TAF (Terminal Aerodrome Forecasts), two key sources for aviation meteorology. These sources provide real-time observational data and forecast information, enabling a comprehensive analysis of prediction accuracy and associated errors.

The dataset includes diverse features that span both observational and forecast variables. Observational data comprise temperature, visibility, and wind speed, providing real-time weather conditions at airports. Forecast variables, such as TAF visibility and ceiling, offer predictive insights into future weather scenarios. To enrich the analysis, spatial and climatological attributes, including latitude, longitude, and climate type, are incorporated. These attributes capture the geographical and environmental context of each airport.

The study focuses on categorizing forecast errors into three distinct types:

- **No error (0):** Cases where the forecast accurately predicts the observed conditions.
- **Type I error (1):** False alarms, where the forecast overpredicts adverse weather conditions that do not materialize.
- **Type II error (2):** Missed detections, where the forecast fails to predict adverse weather conditions that do occur.

Accurately predicting these error types is critical, as their implications extend beyond operational delays and cost inefficiencies. Type I errors can lead to unnecessary disruptions, such as flight cancellations, rerouting, or unnecessary hire of a more experienced pilot, while Type II errors pose direct safety risks by exposing aircraft to unanticipated adverse weather.

2 Literature Review

Early weather forecasting models primarily relied on traditional statistical methods. For instance, autoregressive integrated moving average (ARIMA) models have long been used for time series forecasting under the assumption of linearity (Hyndman & Athanasopoulos, 2021; Box, Jenkins, Reinsel, & Ljung, 2015). However, such models can struggle to capture the non-linear dynamics inherent in meteorological data. To address abrupt changes in weather patterns—such as the sudden onset of a cold front—change-point detection techniques have been introduced. Killick, Fearnhead, and Eckley (2012) developed an optimal method for detecting changepoints in time series, which has been applied in various atmospheric studies.

The integration of machine learning has considerably enhanced forecast performance by accommodating complex, non-linear relationships among variables. For example, Zhang (2003) demonstrated improved forecasting accuracy by combining ARIMA with artificial neural networks (ANNs), thereby leveraging both linear and non-linear modeling strengths. Random Forests, introduced by Breiman (2001), have also been applied to meteorological classification tasks due to their robustness in handling complex datasets. In addition, support vector machines (SVM) have shown effectiveness in short-term forecasting of wind speed (Wang & Li, 2009).

Deep learning methods have further advanced the field. Although traditional ANNs require extensive data and computing power, architectures such as Long Short-Term Memory (LSTM) networks—first proposed by Hochreiter and Schmidhuber (1997)—are particularly well-suited for sequential data. Building on this, Shi et al. (2015) introduced the Convolutional LSTM network, which has proven effective for precipitation nowcasting by capturing both spatial and temporal dependencies.

Ensemble techniques provide another route to improved accuracy. Methods like Gradient Boosting (Friedman, 2001) and ensembles of Random Forests (Breiman, 2001) aggregate predictions from multiple models to reduce error. Hybrid approaches that merge machine learning with numerical weather prediction systems (e.g., the Weather Research and Forecasting [WRF] model) capitalize on both empirical data and physical modeling principles.

In aviation, reliable weather forecasts are critical for ensuring flight safety. Standard meteorological products such as METAR and TAF reports supply essential real-time and forecast data. Recent studies have applied machine learning to refine these forecasts. For instance, Wang and Zeng (2018) employed support vector machine techniques to predict runway visual range, thereby reducing prediction errors and enhancing decision-making for air traffic controllers.

3 Description of Data Set

METAR and TAF raw reports are provided as a single string in a standard format understood by pilots worldwide. We obtained the raw data from aviationweather.gov, selecting a time frame from September 6, 2024, to November 25, 2024. Observations and forecasts were recorded every six hours, resulting in four entries per day. The dataset was limited to 50 randomly chosen airports across the United States, yielding a total of 10,990 rows. From the raw data strings, we extracted individual measurements for the variables listed below and incorporated airport characteristics.

Observed and Predicted Weather Characteristics

- **Visibility_METAR (Visibility_TAF):** The observed (predicted) horizontal visibility reported (in miles).

- **Ceiling_METAR (Ceiling_TAF):** The observed (predicted) height of the lowest cloud layer covering more than half of the sky, measured in feet above ground level. If no such layer was reported, the value was set to 50,000 ft.

- **Wind_direction_TAF:** The observed (predicted) wind direction, measured in degrees from true north, reported in increments of 10 degrees (ranging from 10 to 360). A wind blowing from the true north is indicated by 360 degrees.

Airport Characteristics

- **Altitude:** The height of the airport above sea level (in feet).

- **Latitude**: The airport's global position relative to the equator (in degrees north).

- **Longitude**: The airport's global position relative to the prime meridian (in degrees west).

- **Distance_to_water:** Proximity of an airport to a large body of water such as a lake, bay, or ocean (in miles).

- **Koppen_class:** The climate classification of the airport's region, based on the Köppen climate classification system, which categorizes climates into five main groups: A (tropical), B (arid), C (temperate), D (continental), and E (polar). In the United States, all categories except A (tropical) are present.

The response (or target) variable was determined in two steps. First, we classified both the actual and predicted weather conditions into one of four categories:

- **Visual Flight Rules (VFR):** Ceilings above 3,000 feet and visibility greater than 5 miles (including clear skies).

- **Marginal Visual Flight Rules (MVFR):** Ceilings between 1,000 and 3,000 feet and/or visibility between 3 and 5 miles (inclusive).

- **Instrument Flight Rules (IFR):** Ceilings from 500 to less than 1,000 feet and/or visibility between 1 and less than 3 miles.

- **Low Instrument Flight Rules (LIFR):** Ceilings below 500 feet above ground level and/or visibility under 1 mile.

These categories follow a natural order from best to worst: VFR, MVFR, IFR, and LIFR.

Next, we constructed the target variable with three possible values:

- 0 if the predicted category matched the actual category (correct prediction).
- 1 if the predicted category was worse than the actual category (Type I error or false alarm).
- 2 if the predicted category was better than the actual category (Type II error or missed detection).

3.1 Methodology: Theory and Applications

3.1.1 Random Forest

A random forest is an ensemble of decision trees. Each tree is independently trained on a bootstrap sample of the data, on a random subset of features. Random forest algorithms have three main hyperparameters that need to be set before training. These are the number of bootstrap samples, how many leaves each decision tree should have, and the number of features sampled. The results are then aggregated. For a multinomial classification, the final prediction is determined by majority voting, where the most frequently predicted class is selected. To determine which features are the most relevant in a model, the importance score is computed for each feature, which measures how much the model's predictive accuracy decreases when a given variable is excluded.

To apply a random forest technique, we first randomly split the data into 80% training and 20% testing sets. Then we fit the model using the training set and compute the accuracy of prediction using the testing set. The model achieved a prediction accuracy of 92%. Table 1 presents the list of features along with their importance scores.

Feature	**Importance Score**
Ceiling_TAF (ft)	0.576
Distance (km)	0.121
Visibility_TAF (km)	0.111
Latitude (degrees)	0.061
Longitude (degrees)	0.056
Altitude (ft)	0.056
Koppen Climate Classification	0.018

Table 1. List of features and their important scores for the Random Forest model.

Although predicted visibility and ceiling were used to determine the predicted flight category classifications (VFR, MVFR, IFR, or LIFR), their raw values were still included as features (predictors) in the model. Ceiling and visibility emerged as strong predictors, along with proximity to a large body of water (Distance). Among airport characteristics, the most important features, in decreasing order of significance, were latitude, longitude, altitude, and Köppen climate classification.

3.1.2 Naive Bayes Classifier

Naïve Bayes classifier is a probabilistic machine learning method based on Bayes' theorem under a *naive* assumption that the features are conditionally independent given the class label. By Bayes' theorem, the

posterior distribution of the target variable is given by

$$\mathbb{P}(y \mid x_1, x_2, \ldots, x_n) = \frac{\mathbb{P}(y)\mathbb{P}(x_1, x_2, \ldots, x_n \mid y)}{\mathbb{P}(x_1, x_2, \ldots, x_n)}.$$

Under the conditional independence assumption, we have:

$$\mathbb{P}(x_1, x_2, \ldots, x_n \mid y) = \prod_{i=1}^{n} \mathbb{P}(x_i \mid y).$$

In a multinomial classification problem, the posterior probability of each class is computed, and the class with the highest probability is predicted. Since the denominator $\mathbb{P}(x_1, x_2, \ldots, x_n)$ is the same in all classes, it can be ignored when comparing posterior probabilities. Thus, the decision rule is simplified to

$$\hat{y} = \arg\max_{y} \ \mathbb{P}(y) \prod_{i=1}^{n} \mathbb{P}(x_i \mid y).$$

The prior probabilities $\mathbb{P}(y)$ are estimated using the empirical proportions of each class in the training set. To compute the empirical conditional probability $\mathbb{P}(x_i = x \mid y)$ for categorical predictors, we compute the fraction of observations in class y in the training set where $x_i = x$. For continuous predictors, we use the normal density function with estimated mean $\hat{\mu} = \bar{x}$ and variance $\hat{\sigma}^2 = s^2$.

Fitting the Naive Bayes classifier to the training set yielded a prediction accuracy of 91.75%.

Naive Bayes Contextual Interpretation:

- Naive Bayes assumes feature independence, which is unrealistic in meteorological contexts. For example, visibility, wind speed, and weather conditions are often interdependent (e.g., reduced visibility is often accompanied by precipitation and lower ceiling heights).
- The model's weaker performance underscores the importance of accounting for feature relationships. This limitation highlights how weather variables interact in real-world forecasting, emphasizing the need for models that capture these complexities.
- Despite its simplicity, Naive Bayes provides a baseline understanding of the data, confirming that even basic assumptions can yield reasonable insights.

3.1.3 Artificial Neural Network

An ANN approximates a function by composing multiple layers. For a feed-forward network with L layers, the operations are as follows:

Input Layer:

$$a^{(0)} = x.$$

For layer $l = 1, 2, \ldots, L$:

$$\begin{aligned} z^{(l)} &= W^{(l)} a^{(l-1)} + b^{(l)}, \\ a^{(l)} &= \sigma\big(z^{(l)}\big), \end{aligned} \tag{1}$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the l-th layer, and $\sigma(\cdot)$ is an activation function (e.g., sigmoid, ReLU).

The network output is

$$f(x; \theta) = a^{(L)}, \tag{2}$$

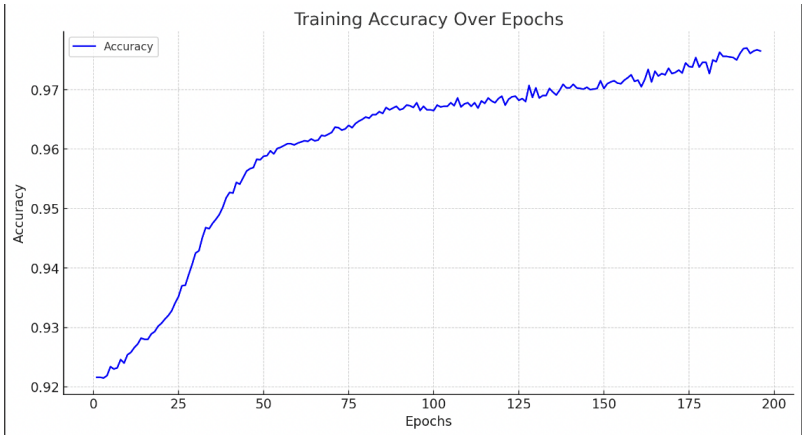
with $\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^{L}$.

Training is performed by minimizing a loss function $L(f(x;\theta), y)$ using gradient-based methods. For example, using gradient descent:

$$\theta \leftarrow \theta - \eta \nabla_\theta L(f(x;\theta), y), \tag{3}$$

where $\eta > 0$ is the learning rate.

Performance: Accuracy of 92.57%.



Contextual Interpretation:

- The ANN's ability to capture non-linear relationships provides deep insights into the data. It effectively models the complex interactions between features such as the joint effect of visibility, ceiling, and weather on error likelihood.
- The confusion matrix reveals that the ANN excels at identifying correct forecasts (Class 0), but struggles with Type I (false alarms) and Type II (missed detections) errors. This discrepancy reflects real-world forecasting challenges, where adverse events (minority classes) are inherently harder to predict due to their rarity.
- The ANN's iterative improvement during training reflects the importance of nuanced relationships in meteorology, such as how rapid changes in wind direction or speed might signal shifts in weather conditions.

Class	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
0 (No Error)	92.8	99.7	96.1
1 (Type I)	60.0	0.039	0.0732
2 (Type II)	0.500	0.046	0.084

Table 1: Precision, recall, and F1-score for ANN predictions across error classes.

Gradient Boosting

Gradient boosting builds an additive model in a forward stage-wise fashion. The model is expressed as:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x), \tag{4}$$

where:

- $F_0(x)$ is an initial guess (often a constant),
- $h_m(x)$ is a weak learner (e.g., a decision tree) at iteration m, and
- γ_m is a step size.

At each iteration m, the pseudo-residuals for each training example are computed as

$$r_{im} = - \left. \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right|_{F(x_i)=F_{m-1}(x_i)}. \tag{5}$$

The weak learner $h_m(x)$ is fit to the pseudo-residuals, and the optimal multiplier γ_m is obtained by solving

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{N} L\Big(y_i, F_{m-1}(x_i) + \gamma\, h_m(x_i)\Big). \tag{6}$$

Performance: Test accuracy of 66.2% and cross-validation accuracy of 90% ± 1%.

Contextual Interpretation:

- Gradient Boosting's lower test accuracy suggests overfitting or challenges in generalizing minority classes (Type I and II errors), despite SMOTE balancing.
- Feature importance rankings highlight that visibility and weather conditions are dominant predictors. This aligns with aviation safety concerns, where reduced visibility or adverse weather significantly impacts flight decisions.
- The variable importance further suggests that spatial features like latitude and altitude play secondary roles, indicating regional variations in forecast accuracy (e.g., tropical climates vs. arid regions).
- Gradient Boosting's performance hints at the intricate dependencies between weather features, particularly in detecting rare events, underscoring the need for advanced ensemble methods to enhance predictive power.

Feature	**Importance Score**
Ceiling_TAF (ft)	0.519
Visibility_TAF (km)	0.327
Altitude (ft)	0.050
Distance (km)	0.044
Latitude (degrees) (C)	0.029
Longitude (degrees)	0.025
Koppen Climate Classification	0.006

Table 2: Top feature importance scores for the Random Forest model.

Support Vector Machines (SVM)

Given a training set $\{(x_i, y_i)\}_{i=1}^{N}$, with $x_i \in \mathbb{R}^d$ and labels $y_i \in \{-1, +1\}$, the **hard-margin SVM** finds the hyperplane

$$w^\top x + b = 0,$$

which maximizes the margin between the two classes. This is formulated as the following optimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(w^\top x_i + b) \geq 1, \quad i = 1, \ldots, N. \end{aligned} \tag{7}$$

Introducing Lagrange multipliers $\alpha_i \geq 0$, the **dual formulation** becomes:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \ \forall i. \end{aligned} \tag{8}$$

For the **soft-margin SVM**, slack variables $\xi_i \geq 0$ are introduced to allow for misclassification, and the optimization problem becomes

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i, \\ \text{subject to} \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, \ldots, N, \end{aligned} \tag{9}$$

where $C > 0$ is a regularization parameter.

Performance: Achieved 93.02% accuracy across various kernels, with minimal variation between Linear, Polynomial, RBF, and Sigmoid kernels.

Contextual Interpretation:

- SVM's high and consistent accuracy highlights the dataset's inherent separability. This suggests that weather variables like visibility, ceiling, and wind speed form clear boundaries between error categories in the feature space.
- The linear kernel's near-parity with non-linear kernels indicates that the relationships between features and error types are relatively linear. For instance, a drop in visibility or a lower ceiling height directly correlates with higher error likelihoods, making these features strong indicators of forecast quality.
- In a real-world setting, SVM could provide a reliable, interpretable method to flag potential forecast inaccuracies, particularly when computational efficiency is necessary.

K-Nearest Neighbors (KNN)

KNN is a **non-parametric** method. Given a query point x, we define a distance (commonly the Euclidean distance)

$$d(x, x_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{i,j})^2}, \tag{10}$$

and identify the set $\mathcal{N}_k(x)$ of the k points in the training set closest to x.

For Classification: The predicted class is the mode (majority vote) of the labels of the k nearest neighbors:

$$\hat{y} = \text{mode}\{y_i \mid x_i \in \mathcal{N}_k(x)\}. \tag{11}$$

For Regression: The predicted value is the average of the neighbors' responses:

$$\hat{y} = \frac{1}{k} \sum_{x_i \in \mathcal{N}_k(x)} y_i. \tag{12}$$

Performance: Achieved an accuracy of 92.61%.
Contextual Interpretation:

- KNN's reliance on local feature similarities aligns well with the spatial and temporal nature of meteorological data. For instance, weather patterns at a specific location and time are often similar to those observed at nearby locations or times.
- The model's success indicates that clustering of weather conditions in the feature space aligns with error categories. This clustering reflects the real-world scenario where specific combinations of low visibility, high wind speeds, or adverse weather often lead to systematic forecast inaccuracies.
- KNN's sensitivity to the number of neighbors (set at 31) suggests that meteorological patterns require an optimal balance between local and global information for accurate error detection.

4 Summary and Discussion

4.1 Summary of Results

This study compared multiple machine learning models in predicting the accuracy of aviation weather forecasts by classifying errors into correct predictions (no error), false alarms (Type I errors), and missed detections (Type II errors).

Across most models, ceiling height and visibility emerged as the most influential predictors. Proximity to water, latitude, and altitude also played non-negligible roles. Gradient boosting, despite its theoretical strength, underperformed on the test set, likely due to overfitting or class imbalance challenges. SVM and ANN demonstrated the highest performance, suggesting strong separability and non-linear relationships within the dataset.

4.2 Future Applications:

The methodology developed here has strong potential for integration into operational systems. For example, a model identifying likely forecast errors could serve as a second-layer alert system for air traffic controllers or automated flight planning software, enabling more cautious decision-making when forecast confidence is low.

5 Supplemental Materials

The raw and cleaned data sets and all relevant Python codes are available in the following github repository: `https://github.com/samrocksnature/Modeling-Error-Types-in-Aviation-Weather-Forecasts-Machine-Learning-Approach`.

Acknowledgements

This research would not have been possible without the guidance and mentorship of Dr. Olga Korosteleva. Her expertise in statistical modeling and invaluable support throughout the project are greatly appreciated.

I would like to thank Mr. Eric Huang for inspiring me to pursue this project by making AP statistics one of my favorite classes in high school, and my parents for supporting and encouraging me through all my mathematical endeavors.

References

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

5. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. `https://otexts.com/fpp3/`

6. Johnson, P., Lee, R., & Patel, S. (2021). Deep learning for rare event prediction in aviation. In *Proceedings of the International Conference on Machine Learning*.

7. Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590–1598.

8. Kroese, D. P., Botev, Z., Taimre, T., & Vaisman, R. (2019). *Data science and machine learning: Mathematical and statistical methods*. Chapman & Hall/CRC.

9. Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2* (3rd ed.). Packt Publishing.

10. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems* (pp. 802–810).

11. Smith, J., Brown, K., & White, L. (2020). *Machine learning models for runway visibility prediction*. Journal of Aviation Safety.

12. Wang, W., & Zeng, X. (2018). Application of support vector machine in runway visual range prediction. *Journal of Intelligent & Fuzzy Systems, 34*(4), 2397–2403.

13. Wang, Y., & Li, Q. (2009). Support vector machine-based short-term wind speed forecasting. *Energy Conversion and Management, 50*(4), 920–926.

14. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.