1. Predicting Flight Cancellations

• Question: Can we predict whether a flight will be canceled based on factors like carrier, weather conditions, NAS issues, and airport location?

• Model to Use: Logistic Regression or Random Forest Classifier

o Why: Logistic Regression is a good baseline model for binary classification problems (e.g., canceled vs. not canceled). However, if you want to capture complex relationships and interactions between features, a Random Forest Classifier would be more effective. It handles categorical variables well, is robust to missing data, and provides feature importance metrics.

2. Predicting Flight Delays

• Question: Can we predict the likelihood of a flight being delayed by more than 15 minutes using variables like the airline, month, weather conditions, and NAS issues?

• Model to Use: Gradient Boosting Classifier (e.g., XGBoost) or Support Vector Machine (SVM)

o Why: Gradient Boosting models like XGBoost can handle non-linear relationships and are very effective for structured/tabular data. They are also great for feature importance and interpretability. SVM with a non-linear kernel (e.g., RBF) can also be used for classification when the data is not linearly separable.

3. Classifying Root Causes of Delays

• Question: Can we classify the root cause of flight delays (carrier, weather, NAS, security, or late aircraft) based on flight data?

• Model to Use: Decision Tree Classifier or Random Forest Classifier

o Why: Decision Trees are interpretable and work well for multi-class classification tasks. Random Forests provide better performance due to ensembling and can also help identify the most important factors contributing to each type of delay.

4. Predicting Delay Duration

• Question: Can we predict the duration of a flight delay (in minutes) based on factors such as the airline, weather, month, and time of day?

• Model to Use: Linear Regression or Gradient Boosting Regressor (XGBoost)

o Why: Linear Regression serves as a good baseline for predicting continuous variables. However, Gradient Boosting Regressors are better suited for capturing non-linear relationships and interactions between features, making them ideal for more accurate predictions of delay durations.

5. Identifying Flights at Risk of Cascading Delays

• Question: Can we predict whether a delayed flight will cause subsequent delays in other flights (cascading delays) using historical delay patterns and airline schedules?

• Model to Use: Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) Networks

o Why: Cascading delays involve temporal dependencies (i.e., previous delays affect future ones), so models like RNNs and LSTMs are well-suited for capturing patterns over time. These models can learn sequential dependencies in data, which is crucial for predicting knock-on effects.

6. Clustering Airports Based on Delay Patterns

• Question: Can we cluster airports based on their delay patterns to identify those with the highest likelihood of delays due to specific factors (weather, congestion, etc.)?

• Model to Use: K-Means Clustering or DBSCAN

o Why: K-Means is a straightforward clustering algorithm for partitioning airports into groups based on similar delay profiles. If the data has noise or varying densities, DBSCAN (Density-Based Spatial Clustering) is more robust.

7. Predicting Peak Seasons for Flight Cancellations

• Question: Can we predict the peak months for flight cancellations using historical data on cancellations, weather patterns, and flight volumes?

• Model to Use: Time Series Analysis (ARIMA) or Prophet

o Why: ARIMA models are great for univariate time series forecasting. However, if you have multiple factors influencing cancellations, the Prophet model (developed by Facebook) is easier to use, handles missing data well, and accounts for seasonality and trend changes.

8. Determining Factors Contributing to Late Aircraft Delays

• Question: Can we predict whether a flight will experience a "late aircraft" delay based on the scheduled arrival time, departure airport, and previous flight delays?

• Model to Use: Logistic Regression or Naive Bayes Classifier

o Why: For a quick and interpretable solution, Logistic Regression works well. If you want a probabilistic approach that works with categorical variables and assumes feature independence, Naive Bayes is a good fit.