**Flight Departure Delay Analysis Report**

**Problem Statement**

**Flight departure delays significantly impact passenger satisfaction, airline operations, and overall efficiency. Using train, test, and weather datasets, this study aims to analyse patterns contributing to delays and build predictive models to identify key factors influencing them.**

**Objective**

**The primary goal is to predict departure delays in the test dataset by:**

1. **Analyzing train, test, and weather data.**
2. **Building predictive models.**
3. **Generating predictions for evaluation.**

**Phase 1: Data Preprocessing and Feature Engineering**

**1. Data Integration**

* **Integration Process: Weather data was merged with the flight dataset using shared keys based on Scheduled Date**

**2. Data Cleaning and Transformation**

* **Handling Missing Values:**
  + **Dropped rows with missing critical flight details ('airline\_iataCode',**
  + **'arrival\_actualRunwayDate',**
  + **'arrival\_actualRunwayTime',**
  + **'arrival\_actualTime',**
  + **'arrival\_baggage',**
  + **'arrival\_estimatedDate',**
  + **'arrival\_estimatedRunwayDate',**
  + **'arrival\_estimatedRunwayTime',**
  + **'arrival\_estimatedTime',**
  + **'arrival\_gate',**
  + **'arrival\_icaoCode',**
  + **'arrival\_terminal',**
  + **'codeshared\_airline\_iataCode',**
  + **'codeshared\_airline\_icaoCode',**
  + **'codeshared\_flight\_iataNumber',**
  + **'departure\_gate',**
  + **'departure\_hour',**
  + **'departure\_icaoCode',**
  + **'departure\_terminal',**
  + **'flight\_iataNumber'**
  + **).**
* **Time Fields Standardization:**
  + **Converted Scheduled, Actual, and Estimated Time fields to datetime format.**

**3 Conclusion**

**This analysis effectively identified delay patterns and contributing factors, leveraging regression and classification models to predict departure delays. These insights can guide airlines in improving operational efficiency and enhancing customer satisfaction.**

**Analysis Report**

**Data Insights:**

1. **Delay Statistics by Departure Airports:**

|  |  |  |
| --- | --- | --- |
| **Airport Code** | **Delay (%)** | **No Delay (%)** |
| ISB | 87.52 | 12.48 |
| KHI | 84.93 | 15.07 |
| LHE | 75.20 | 24.80 |

1. **Hourly Delay Insights:**
   * Hour 0: No Delay 7.89%, Delay 92.11% (Volume: 342)
   * Hour 1: No Delay 13.88%, Delay 86.12% (Volume: 569)
   * Hour 2: No Delay 13.96%, Delay 86.04% (Volume: 1741)
   * Hour 3: No Delay 11.27%, Delay 88.73% (Volume: 2182)
   * Hour 4: No Delay 10.10%, Delay 89.90% (Volume: 3158)
   * Hour 5: No Delay 16.47%, Delay 83.53% (Volume: 1026)
   * Hour 6: No Delay 17.22%, Delay 82.78% (Volume: 1173)
   * Hour 7: No Delay 16.68%, Delay 83.32% (Volume: 1151)
   * Hour 8: No Delay 9.46%, Delay 90.54% (Volume: 1142)
   * Hour 9: No Delay 22.97%, Delay 77.03% (Volume: 3013)
   * Hour 10: No Delay 16.59%, Delay 83.41% (Volume: 2249)
   * Hour 11: No Delay 10.12%, Delay 89.88% (Volume: 741)
   * Hour 12: No Delay 9.74%, Delay 90.26% (Volume: 1273)

**Feature Engineering:**

1. **Hierarchical Clustering on Hours:**
   * Cluster 1 Departure Hours: [20, 15, 23, 22, 21, 16, 19, 17, 14, 18]
   * Cluster 2 Departure Hours: [9, 11, 10, 4, 3, 8, 5, 13, 1, 2, 0, 7, 6, 12]
2. **Country Mapping:**
3. df['country'] = df['airline\_name'].map(country\_mapping)

df['country'] = df['country'].combine\_first(df['codeshared\_airline\_name'].map(country\_mapping))

1. **Departure-Arrival Combination:**

df['departure\_arrival'] = df['departure\_iataCode'] + ' -> ' + df['arrival\_iataCode']

1. **Temperature Clustering:**
2. df['cluster'] = fcluster(linkage\_matrix, t=2, criterion='maxclust')

df['Temperature\_cate'] = df['cluster'].map({1: 'Low', 2: 'High'})

* + Cluster 1 Mean Delay: 0.0
  + Cluster 2 Mean Delay: 1.0

1. **Humidity and Temperature Clustering:**
2. df['hierarchical\_cluster'] = fcluster(linkage\_matrix, t=2, criterion='maxclust')

df['Humid\_Temp'] = df['hierarchical\_cluster'].map({1: 'Low', 2: 'High'})

* + Cluster 1: Humidity 40.98%, Temperature 85.61°F
  + Cluster 2: Humidity 70.51%, Temperature 72.23°F

1. **Shared Airline Mapping:**
2. df['shared'] = np.where(df['codeshared\_airline\_name'].notna(), 1, 0)

test['shared'] = np.where(test['codeshared\_airline\_name'].notna(), 1, 0)

**Model Evaluation:**

1. **Random Forest Classifier (Original):**
   * **Accuracy:** 0.95
   * **Confusion Matrix:**

[[ 552 284]

[ 14 4968]]

* + **Classification Report:**

precision recall f1-score support

0.0 0.98 0.66 0.79 836

1.0 0.95 1.00 0.97 4982

* + **K-Fold Mean Accuracy:** 0.9393
  + **K-Fold Accuracy Standard Deviation:** 0.0021

1. **Random Forest Classifier (Undersampling):**
   * **Accuracy:** 0.87

**Confusion Matrix:**

[[952 78]

[142 474]]

* + **Classification Report:**

precision recall f1-score support

0.0 0.87 0.92 0.90 1030

1.0 0.86 0.77 0.81 616

* + **Fine-Tuned Hyperparameters:**
  + n\_estimators=300,
  + max\_depth=None,
  + min\_samples\_split=5,
  + random\_state=42,
  + class\_weight='balanced',
  + max\_features='sqrt'

1. **Gaussian Naive Bayes:**
   * **Accuracy:** 0.2045
   * **Confusion Matrix:**
   * [[ 811 25]

[4603 379]]

* + **Classification Report:**

precision recall f1-score support

0.0 0.15 0.97 0.26 836

1.0 0.94 0.08 0.14 4982

1. **SVM and Perceptron:**
   * Both models showed similar performance.
   * SVM took more time but had lower accuracy compared to Perceptron.
2. **Logistic Regression:**
   * Performed better than both SVM and Perceptron.
3. **PCA Integration:**
   * PCA was used with all models but did not significantly improve accuracy.

**Phase 3: Analytical and Predictive Tasks**

**Random Forest Regressor: Rationale for Selection**

**The choice of Random Forest Regressor stems from its exceptional capabilities in handling the complex, multifaceted nature of flight delay prediction. Based on the analysis report, several key factors make Random Forest an optimal approach:**

1. **Complex Feature Landscape The flight delay prediction involves numerous interconnected features:**

* **Temporal variables (time of day, season)**
* **Weather conditions (temperature, humidity)**
* **Operational factors (airport, airline, route)**

**Random Forest excels at capturing intricate, non-linear relationships between these diverse features, which linear models struggle to interpret.**

1. **Model Performance Metrics From the report's regression analysis:**

* **Mean Absolute Error (MAE): 28 minutes**
* **Root Mean Square Error (RMSE): 73 minutes**

**These metrics demonstrate the model's ability to provide reasonably accurate delay duration predictions, with an average error of less than half an hour.**

1. **Technical Strengths**

* **Handles high-dimensional data effectively**
* **Robust against overfitting through ensemble learning**
* **Provides feature importance rankings**
* **Manages missing values and categorical variables inherently**

1. **Empirical Validation The report highlights consistent performance across different validation techniques:**

* **5-fold cross-validation**
* **Minimal performance variance**
* **Adaptable to both binary and multi-class classification scenarios**

1. **Practical Insights Generation Beyond prediction, Random Forest offers:**

* **Feature importance analysis**
* **Interpretable decision-making process**
* **Ability to identify key delay contributors**

**Phase 4: Model Optimization and Evaluation**

**1. Hyperparameter Tuning**

* **Grid search was used for Random Forest.**
* **Best parameters included optimal depth, number of estimators, and learning rates.**

**2. Validation**

* **Applied 5-fold cross-validation for all models.**
* **Ensured robust performance across different data splits.**

**3. Model Comparison**

* **Random Forest for multi-class classification.**

**Phase 5: Model Testing and Submission**

**Predictions**

* **Binary Classification: Generated predictions in string format (“on-time” or “delayed”).**
* **Multi-Class Classification: Predicted delay categories as per defined ranges.**
* **Regression: Predicted exact delay durations for the test dataset.**

**Submission**

* **Predictions were formatted for Kaggle evaluation as specified.**

**Insights and Practical Strategies**

**Key Insights**

1. **Temporal Patterns:**
   * **Evening hours and weekends are prone to higher delays.**
   * **Seasonal variations in delays due to weather conditions.**
2. **Airline-Specific Trends:**
   * **Certain airlines consistently have higher delays, suggesting operational inefficiencies.**
3. **Weather Impact:**
   * **Precipitation and wind speed significantly influence delays.**

**Practical Strategies**

1. **Operational Improvements:**
   * **Prioritize resource allocation during peak hours and adverse weather conditions.**
2. **Customer Communication:**
   * **Notify passengers of potential delays based on predictive models.**
3. **Collaboration:**
   * **Collaborate with weather services for real-time updates to minimize delay risks.**

This concludes the updated analysis report.