Predicción de la Cantidad de Aeronaves en el Espacio Aéreo Cubano

Informe Técnico Final

Equipo de Desarrollo

8 de enero de 2026

Resumen

Este documento detalla el desarrollo de un sistema de predicción de tráfico aéreo para la FIR Habana. Se aborda la problemática de gestionar eficientemente el flujo de aeronaves mediante el uso de modelos de aprendizaje automático (XGBoost). El informe describe la metodología completa, desde la recolección y limpieza de datos (incluyendo la corrección crítica de conteo de vuelos únicos), hasta la ingeniería de características y la evaluación de resultados. Se discuten las implicaciones de los hallazgos y se proponen líneas de trabajo futuro.

Índice

1. Introducción

1.1. Antecedentes y Contexto

Breve historia del control de tráfico aéreo en la región y la necesidad de herramientas predictivas modernas.

1.2. Definición del Problema

Descripción formal del problema de forecasting de series temporales en este dominio.

1.3. Objetivos

1.3.1. Objetivo General

1.3.2. Objetivos Específicos

- Implementar pipeline de procesamiento de datos ATFM.
- Desarrollar modelos de regresión robustos.
- Validar resultados con métricas cuantitativas y cualitativas.

1.4. Alcance y Limitaciones

2. Marco Teórico

El presente capítulo establece los fundamentos conceptuales necesarios para comprender el alcance y la metodología del estudio. Se describe el contexto de la gestión de tráfico aéreo, las particularidades de las series temporales en este dominio y se detallan los algoritmos seleccionados para el desarrollo de los modelos predictivos.

2.1. Gestión de Afluencia del Tráfico Aéreo (ATFM)

La Gestión de Afluencia del Tráfico Aéreo (ATFM) es un servicio establecido para asegurar un flujo de tránsito aéreo seguro, ordenado y expedito. Su objetivo principal es asegurar que la capacidad de control de tránsito aéreo (ATC) sea utilizada al máximo posible y que el volumen de tráfico no exceda las capacidades declaradas por la autoridad competente. En el contexto de este proyecto, la región de estudio es la Región de Información de Vuelo (FIR) de La Habana, un corredor estratégico para el tráfico norte-sur en las Américas.

Un componente crítico en la medición del volumen de tráfico es el Identificador Único Global de Vuelo (GUFI). A diferencia de los conteos tradicionales por sector que pueden inflar las cifras debido a que un mismo vuelo atraviesa múltiples sectores, el uso del GUFI permite contabilizar aeronaves únicas en el sistema. Esta distinción es fundamental para obtener una estimación precisa de la carga de trabajo real y la demanda del espacio aéreo [1].

2.2. Análisis de Series Temporales en Aviación

El tráfico aéreo se caracteriza por exhibir patrones temporales altamente definidos. Se observa una fuerte estacionalidad a múltiples niveles: ciclos diarios determinados por los horarios comerciales de las aerolíneas, ciclos semanales con variaciones entre días laborables y fines de semana, y ciclos anuales influenciados por temporadas turísticas y festividades. El problema de predicción se aborda analizando estas componentes (tendencia, estacionalidad y ruido) en la serie histórica para proyectar valores futuros en un horizonte definido [2].

2.3. Algoritmos de Predicción Utilizados

En este estudio se han implementado y comparado diversos enfoques de modelado, abarcando desde estadísticas clásicas hasta aprendizaje profundo.

2.3.1. Modelos Estadísticos y Aditivos

Los modelos ARIMA (AutoRegressive Integrated Moving Average) y su variante estacional SARIMA constituyen la línea base clásica para la predicción de series temporales. Estos modelos asumen que los valores futuros son una función lineal de los valores pasados y los errores pasados [3]. Por otro lado, Prophet es un modelo aditivo desarrollado por Facebook, diseñado para manejar series temporales con fuertes efectos estacionales y días festivos. Descompone la serie en componentes de tendencia, estacionalidad y eventos, siendo particularmente robusto frente a valores atípicos y cambios en la tendencia [4].

2.3.2. Aprendizaje Automático Basado en Árboles

Los métodos de ensamble basados en árboles de decisión han demostrado un rendimiento superior en datos tabulares estructurados. Random Forest construye múltiples árboles de decisión durante el entrenamiento y genera la media de las predicciones de los árboles individuales, reduciendo la varianza y el riesgo de sobreajuste [5].

XGBoost (Extreme Gradient Boosting), el modelo central de este estudio, es una implementación optimizada del algoritmo de Gradient Boosting. A diferencia de Random Forest que construye árboles independientes, XGBoost construye árboles de forma secuencial, donde cada nuevo árbol intenta corregir los errores de los anteriores. Utiliza una función objetivo regularizada que controla la complejidad del modelo, haciéndolo altamente eficiente y preciso para capturar patrones no lineales en grandes conjuntos de datos [6].

2.3.3. Redes Neuronales Recurrentes

Las redes LSTM (Long Short-Term Memory) son una arquitectura de redes neuronales recurrentes diseñada para aprender dependencias a largo plazo. A través de mecanismos de compuertas (entrada, olvido y salida), las LSTM pueden retener información relevante sobre secuencias temporales extensas mitigando el problema del gradiente desvaneciente, lo que las hace idóneas para modelar dinámicas temporales complejas [7].

2.4. Métricas de Evaluación

Para cuantificar el desempeño de los modelos se utilizan las siguientes métricas estándar:

El Error Absoluto Medio (MAE) mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

La Raíz del Error Cuadrático Medio (RMSE) es una medida cuadrática que penaliza más severamente los errores grandes, siendo útil para detectar predicciones con desviaciones significativas:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

El Coeficiente de Determinación (R^2) indica la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes, proporcionando una medida de la bondad de ajuste del modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

3. Metodología Experimental

3.1. Adquisición y Procesamiento de Datos

3.1.1. Datos Meteorológicos

Se utilizaron las herramientas `db-tool` y `etl-tool` para la extracción, limpieza y almacenamiento de variables climáticas históricas (viento, temperatura, visibilidad) que afectan la operaciones aéreas.

3.1.2. Noticias y Eventos

Implementación del módulo `Event_extractor` para el procesamiento de lenguaje natural (NLP) sobre fuentes de noticias, permitiendo identificar eventos disruptivos o de alta demanda turística.

3.1.3. Datos de Operaciones Aéreas

Archivos horarios (ATFM), diarios (ATC) y mensuales, estructurados por sectores y rutas.

3.2. Fusión e Integración de Datos

Estrategias para la unificación de series temporales con diferentes frecuencias (horaria vs diaria) y la incorporación de features exógenas (clima y noticias) al dataset principal de tráfico aéreo.

3.3. Preprocesamiento y Limpieza

3.3.1. Corrección de Agregación de Datos

Detalle del problema de doble conteo por sectores y la solución implementada (conteo de GUFIs únicos).

3.3.2. Manejo de Valores Faltantes y Outliers

3.4. Ingeniería de Características

3.4.1. Variables Temporales y Cíclicas

3.4.2. Variables de Calendario (Festivos Cubanos)

3.4.3. Variables Exógenas (Impacto COVID-19)

3.4.4. Variables de Retardo (Lags) y Ventanas Móviles

3.5. Estrategia de Validación

División Entrenamiento/Prueba y métricas seleccionadas (MAE, RMSE, R^2, MAPE).

4. Desarrollo e Implementación

4.1. Arquitectura del Sistema

Diagrama de flujo de datos: Ingesta (ETL/Extractor) -¿Preprocessor -¿Feature Engineer -¿Model Registry.

4.2. Modelos Evaluados

Se implementaron y compararon múltiples enfoques de forecasting definidos en `models/model.py`:

- **ARIMA/SARIMA:** Línea base estadística clásica.
- **Prophet:** Modelo aditivo para series con fuerte estacionalidad.
- **Random Forest:** Ensamble de árboles para capturar relaciones no lineales.
- **XGBoost:** Boosting de gradiente optimizado para alto rendimiento.
- **LSTM (Long Short-Term Memory):** Redes neuronales recurrentes para secuencias complejas.
- **Ensemble:** Combinación ponderada de los mejores modelos individuales.

4.3. Optimización de Hiperparámetros

Descripción del proceso de *hyperparameter tuning* (e.g., Grid Search, Random Search) realizado para maximizar las métricas de desempeño en cada modelo.

4.4. Configuración Final

Parámetros óptimos seleccionados para el modelo desplegado.

5. Resultados

5.1. Desempeño Global del Modelo

Tabla comparativa de métricas en el conjunto de prueba.

5.2. Análisis Temporal de Predicciones

Gráficos de series temporales (Real vs Predicho) para el horizonte de prueba.

5.3. Importancia de Características

Análisis de qué variables influyen más en la predicción.

6. Discusión

6.1. Interpretación de los Resultados

¿El modelo captura correctamente los picos y valles? ¿Cómo se comporta en días atípicos?

6.2. Análisis de Errores

Identificación de momentos con mayor error y posibles causas (e.g., eventos no capturados por las features).

6.3. Impacto de la Calidad de los Datos

Discusión sobre la reducción de volumen observada tras corregir la lógica de agregación y la persistencia de volúmenes altos en picos.

6.4. Implicaciones Operativas

Utilidad práctica del modelo para los controladores y planificadores.

7. Conclusiones y Recomendaciones

7.1. Conclusiones

7.2. Trabajo Futuro

Referencias

[1] International Civil Aviation Organization, *Manual on Collaborative Air Traffic Flow Management (Doc 9971)*. ICAO, Montreal, Canada, 2014.

[2] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th ed., 2015.

[3] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. Melbourne, Australia: OTexts, 2nd ed., 2018.

[4] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[5] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

A. Anexos

A.1. Fragmentos de Código Relevantes