“Real-time prediction of flight arrival times using surveillance information” - SACBD@ECSA2018 Presentation
Madrid, September 25th 2018

David Scarlatti, Boeing R&T - Europe
TT Project Context

- EC H2020 funded Lighthouse project. +40 partners across Europe
- Demonstrator style: Focus on deployment of technology on business.
- Delivering “Pilots” in different Transport Industry Domains...
TT Project Context

- **Smart Highways**
  - AUSOL Load Balancing Pilot
  - Load Balancing Pilot – Norte Litoral

- **Sustainable Connected Vehicles**
  - Sustainable Connected Cars Pilot
  - Sustainable Connected Trucks Pilot

- **Proactive Rail Infrastructures**
  - Predictive Rail Asset Management Pilot
  - Predictive High Speed Network

- **Ports as Intelligent Logistics Hubs**
  - Valencia Sea Port Pilot
  - Duisport Inland Port Pilot

- **Smart Airport Turnaround**
  - Smart Passenger Flow Pilot
  - Smart Turnaround

- **Integrated Urban Mobility**
  - Integrated Urban Mobility: Tampere pilot
  - Valladolid Integrated Urban Mobility

- **Dynamic Supply Networks**

---

SACBD@ECSA2018
TT Project in Aviation

• Smart Airport Turnaround. Two Pilots:
  – Smart Passenger Flow Pilot – @Athens Airport

  – Smart Turnaround – @Milan Airport
Turnaround overview. Why Arrival Time matters?

Objective: Increase predictability of the process and airport situational awareness. As a result we expect an improvement and optimization of the turnaround.

While the Aircraft is on ground, the airline is not making money, and the airport can’t serve another aircraft.
Estimated Time of Arrival
Model/Rule Base vs. Data Driven

• Traditional approach: You model Aircraft behaviour and context (i.e. Winds, airspace use, etc...). Then you predict, knowing where is the aircraft, what will happen till it lands. This give you the ETA.

• Data Driven: Following current approach in many fields, let the data talk... ML can discover patterns, not apparent to humans, on what influence the arrival time.
But, which “Data“?

- We follow an iterative approach, starting with one source, and adding in each increment (Agile approach) a new source of data.
  1. First, just surveillance data (4D location) is used
  2. Weather conditions to be added
  3. Flight Plan data to be added
  4. Other sources???
High Level Architecture

- Architecture match a simple lambda architecture.

Data Collection: (surveillance, weather, etc...)

- Speed (Real-Time) Processing
- Batch Processing

Historical data wrangling and ML Training.

Real Time Predictions / Visualization

SACBD@ECSA2018
Data Collection

• Surveillance stream. We use FR24 live feed data, based mainly in ADS-B:
  – Sample records


– Latitude, Longitude, Altitude, Timestamp....
Data Collection

• We use Apache Flume for the collection. Current Flume Flow:

Source type = exec
selector.type = replicating

In Memory Channel

Sink type = Kafka

Sink type = HDFS
path = .../date=%Y-%m-%d
Batch Layer

- We use Hadoop and Hive for preparing the Training datasets:
• We use Kafka for the Data distribution and python/H2O framework for the real-time prediction.
• We adapted Virtual Radar Server to show the live ETA:

Machine Learning details

- Current version only uses surveillance derived features: latitude, longitude, altitude, speeds, timestamp...
- Dependent variable: Seconds to land
- Regression problem. Algorithm: Gradient Boost Machine (GBM)
- Training dataset: Ground truth for selected airlines. January 2016 to August 2017. +10 million observations. We take a 80 % split of this dataset for training, so that approximately 2.3 million observations are employed in the testing of the model.
This aircraft is about 85 minutes to land:
The ETA error is between -856 and +895 seconds,
i.e. if ETA is 8:30 it may land between 8:16 and 8:45

This aircraft is about 35 minutes to land:
The ETA error is between -507 and 389 seconds,
i.e. if ETA is 8:30 it may land between 8:21 and 8:36
In Figure 2 we compare the errors of the predictions issued by three services 1 hour before landing for 736 flights bound for Malpensa-Milan airport on the period ranging from September 2017 to December 2017 (new data never seen by TT algorithm).

Figure 2: Comparison of the error of estimated (ETA) and actual (ATA) times of arrival for different sources issuing predictions 1 hour before landing.
Future Work

• Adding new data sources:
  – METAR reports as weather data
    • New Flume Flow, new Hive tables for adding new features in training models, new Kafka topic with the METAR data
  – Flight Plans from Eurocontrol
    • Adding Mongo to Batch layer (HDFS no good for millions of small .xml files)

• Try other algorithms:
  – Xtreme Gradient Boost
Thank You!

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 731932