VU x KLM Case Study Kick-off

Predicting Standby (IPB) Passenger Show-Ups for KLM Operations

Case problem description

Predicting standby (IPB) passenger show-ups

Predicting the amount of passengers on a flight

=

Number of bookings

−

- People purchasing a seat in business / premium comfort / economy class

Cancellations / No-shows

+

- Cancellations
- Missed connections
- Missed flight (oversleeping?)
- Etc…

Standby (IPB) passengers

- IPB: *Indien Plaats Beschikbaar (In Case of Availability)*
- Airline employees in partnership have the privilege of flying standby
- This means flying for a reduced fee in case there are empty seats left

KLM Royal Dutch Airlines

Why are Standby (IPB) passengers relevant?

Some examples

In-flight operations

- Catering: minimize the amount of food and drink wasted
- Potable water: Minimize the amount of water loaded on board to reduce fuel consumption
- …

Pre-flight planning

- Aircraft weight and balance: relevant for e.g. the amount of cargo to take
- Fuel: Optimize the amount of fuel given the weight and flight duration of the aircraft
- …

Operational efficiency

- Number of gate agents needed to check-in and assist people: increase on-time and customer satisfaction
- Number of baggage handlers needed to ensure timely handling: increase on-time and customer satisfaction
- Security staff needed at Schiphol: increase customer satisfaction
- …

Your challenge:

Build a machine learning model to predict the number of IPB (standby) passengers who will show up at the gate for each KLM flight, based on information available.

Subtasks

1. **Data Exploration:**
   - Analyze feature distributions and relationships with the target and other features
   - Identify missing values, outliers, and potential data quality issues, and decide how to tackle these.
2. **Feature Engineering**
   - Propose and implement additional features that may help improve prediction based on the data available. Think creatively.
3. **Model Selection**
   - Evaluate different Machine Learning / Forecasting models/methods using a valid and fair validation method.
   - Consider model explainability as well as performance.
4. **Model Evaluation**
   - Select appropriate evaluation metrics for the business context
   - Analyze errors and potential biases (e.g., by route, day of week), think of the asymmetry in costs associated with over- and underpredicting

Dataset Walkthrough

Real numbers for real solutions

Data description: targets

Column name	Description
rowkey	Unique flight identifier formatted as: "{yyyy-mm-dd}|{carrier}|{flight_number}| |{departure_station}"
target_standby_leisure_pax	# of standby passengers (**This is what you need to predict**)

Notes:

- File name: targets.csv
- Size: 530k rows x 2 columns
- If a flight had 0 standby passengers, they are **not** in this dataset. Be considerate of this when joining the datasets.

	rowkey	target_standby_leisure_pax
0	2019-01-01|KL|422||DMM	1
1	2019-01-01|KL|422||MCT	1
2	2019-01-01|KL|427||AMS	6
3	2019-01-01|KL|428||DXB	9
4	2019-01-01|KL|445||AMS	2
...	...	...
529893	2025-12-27|KL|1957||AMS	4
529894	2025-12-27|KL|1959||AMS	2
529895	2025-12-27|KL|1969||AMS	1
529896	2025-12-27|KL|1982||BEG	1
529897	2025-12-27|KL|1985||AMS	1

529898 rows × 2 columns

Data description: Dataset_group_#

Notes:

- File names: dataset_group_#.csv
  - Group 1: query_moment = “2d”
  - Group 2: query_moment = “4d”
  - Group 3: query_moment = “6d”
  - Group 4: query_moment = “8d”
- Size: 1.46M rows x 21 columns
- Every row is 1 distinct flight

- More notes on next slide

Column name	Description
rowkey	Unique flight identifier formatted as: "{yyyy-mm-dd}|{carrier}|{flight_number}| |{departure_station}"
query_moment	Describes moment of rendering booking statistics (e.g. 6d)
aircraft_type	Aircraft type by IATA code (e.g. 73H)
aircraft_registration	Registration ID of aircraft (e.g. PHBXD)
flight_group	KLC/EUR/ICA indicator (e.g. EUR)
flight_cancelled	Indicator if flight is cancelled (true) or not (false)
departure_airport	IATA code of departure airport (e.g. AMS)
arrival_airport	IATA code of arrival airport (e.g. JFK)
departure_gate	Gate of departure (e.g. D23)
arrival_gate	Gate of arrival (e.g. D23)
scheduled_departure_time	Scheduled departure time (e.g. 2019-01-01T07:20+01:00)
scheduled_arrival_time	Scheduled arrival time (e.g. 2019-01-01T07:20+01:00)
{seat_class}_capacity	Total seat capacity in {seat_class} (e.g. 33)
total_blocked_seats_{seat_class}_class	Total blocked seats in {seat_class} (e.g. 5)
total_{seat_class}_class_staff_standby_bookings	Total booked standby seats in {seat_class} (e.g. 1)

Datasets

Data description: Dataset_group_#

Notes:

- **Query moment**: This indicates how long before scheduled_departure_date this data was generated. E.g. “6d” means the dataset was generated 6 days before the scheduled departure date. The target is always the latest, actual number.
- **Data Cleaning:** The dataset has been cleaned for the most part, but still contains a lot of missing values. Based on the available data, you should be able to identify whether it is possible to infer missing data or not.
- **Cancelled flights:** all cancelled flights have already been removed
- **Flight groups:**
  - KLC: KLM Cityhopper (short-haul)
  - EUR: Europe (short/medium-haul)
  - ICA: Intercontinental (long-haul)
- **There are 3 seat_classes:**
  - Business class
  - Premium Comfort class (only on ICA)
  - Economy class

	rowkey	query_moment_value	aircraft_type	aircraft_registration	flight_group	flight_cancelled	departure_airport	arrival_airport	departure_gate	arrival_gate	schedu
0	2019-01-01|KL|1000||LHR	2d	73H	PHBXD	EUR	false	LHR	AMS	NaN	NaN	
1	2019-01-01|KL|1001||AMS	2d	73H	UNASSIGNED	EUR	false	AMS	LHR	NaN	NaN	201
2	2019-01-01|KL|1002||LHR	2d	73H	UNASSIGNED	EUR	false	LHR	AMS	NaN	NaN	
3	2019-01-01|KL|1007||AMS	2d	73J	UNASSIGNED	EUR	false	AMS	LHR	NaN	NaN	201
4	2019-01-01|KL|1008||LHR	2d	73J	UNASSIGNED	EUR	false	LHR	AMS	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	
1461629	2025-12-27|KL|973||AMS	2d	E7W	PHEXJ	KLC	false	AMS	HUY	NaN	NaN	202
1461630	2025-12-27|KL|974||HUY	2d	E7W	PHEXJ	KLC	false	HUY	AMS	NaN	NaN	
1461631	2025-12-27|KL|975||AMS	2d	E90	PHEZF	KLC	false	AMS	HUY	NaN	NaN	202
1461632	2025-12-27|KL|983||AMS	2d	E90	NaN	KLC	false	AMS	LCY	NaN	NaN	202
1461633	2025-12-27|KL|984||LCY	2d	E90	NaN	KLC	false	LCY	AMS	NaN	NaN	

1461634 rows × 21 columns


KLM
Royal Dutch Airlines