Supervised classification for cleaning and verifying aircraft model codes in flight offer database

Sara Rool, 5th year ModIA student at INSAxENSEEIHT & data analyst apprentice at Airbus Blagnac. 20/09/2024

Context & aim

The Global Market Forecast provides a **projection of future aircraft demand** and to achieve this, it is essential to analyze past demand. **OAG** is a 20-year database of flight offer (planned operations). This database contains informations such as operating airline, departure airport, aircraft family, seating... of each flights but it contains some errors. In fact, the aircraft model is sometimes **unknown** or **incorrect**.

The aim of this project is to **predict** unknown aircraft codes using **supervised classification** method and to **clean up** incorrect codes.

Methods

From OAG, we keep the **aircraft seating** and the **hours flown** (flight time). We multiply each seats type with a weight, this is necessary to take into account the level of comfort of aircraft.

AIRCRAFT_CODE	FIRST	BUSINESS	PREMIUM	ECO	HOURS_FLOWN
str	i32	i32	i32	i32	f64
"332"	0	38	0	261	20957.083333
"unknow"	0	40	0	168	30792.666667
"333"	0	44	0	267	1.6498e6

Fig 1 : Extract of OAG for A330 family before weighting

SEATS	WEIGHT
FIRST	4
BUSINESS	2.5
PREMIUM	1.5
ECO	1

Fig 2 : Weight by seat type

We apply the **Support Vector Machines** method with a **Gaussian kernel**. We weight each of our individuals (aircraft) by their **contribution of hours flown**. As a matter of fact, an aircraft with few flight hours will tend to be an isolated/false case and therefore unlikely to be repeated.

$$\text{CONTRIBUTION}_{\text{aircraft\_configuration}} = \frac{\sum \text{HOURS\_FLOWN}_{\text{aircraft\_configuration}}}{\sum \text{HOURS\_FLOWN}_{\text{family}}}$$

Fig 3 : Contribution formula for an aircraft configuration within its family

Then, we **predict** the unknown and incorrect aircraft codes after training and calibrating our SVM model.

Example A330 family before cleaning

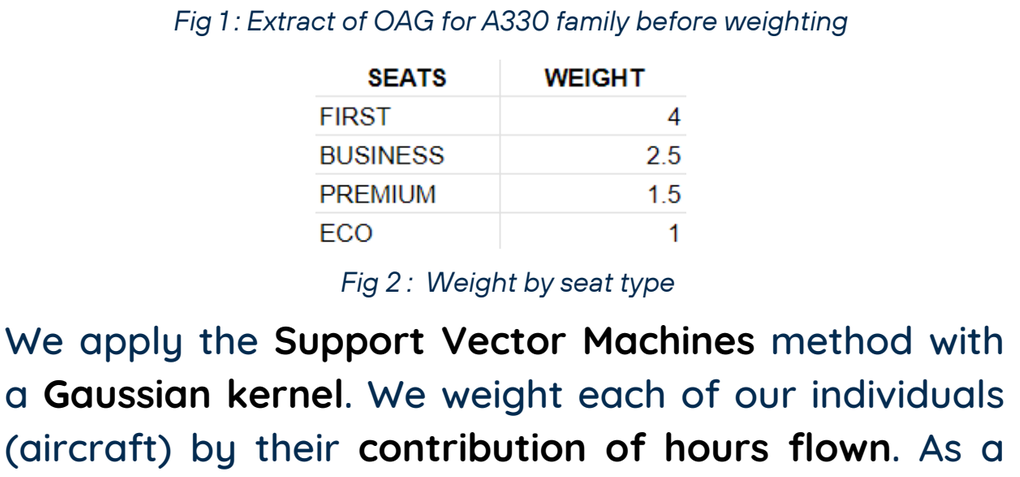

Fig 4 : Example of data before cleaning. Each dot represents an aircraft configuration. The size of the dot is the contribution. In this case, this is data from A330 family. The dots in red are to be determined.

Results

A330 family after cleaning

The predictions give us consistent results. We observe, in our example, that the two **major dynamics** are **well preserved**.

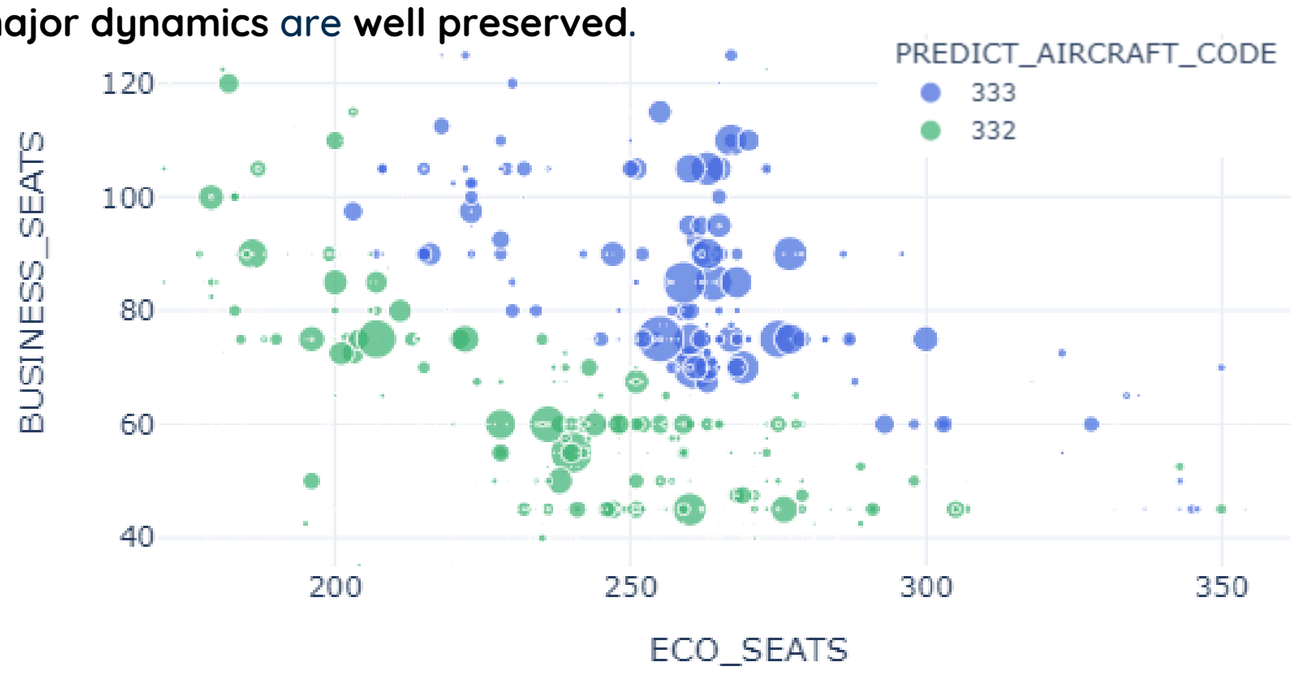

Fig 5 : Example of A330 data after cleaning

Classification score

Using SVM, we can **quantify the certainty** of our predictions. The prediction probabilities are generally good (>70%). Uncertainties are around the **separation hyperplanes**.

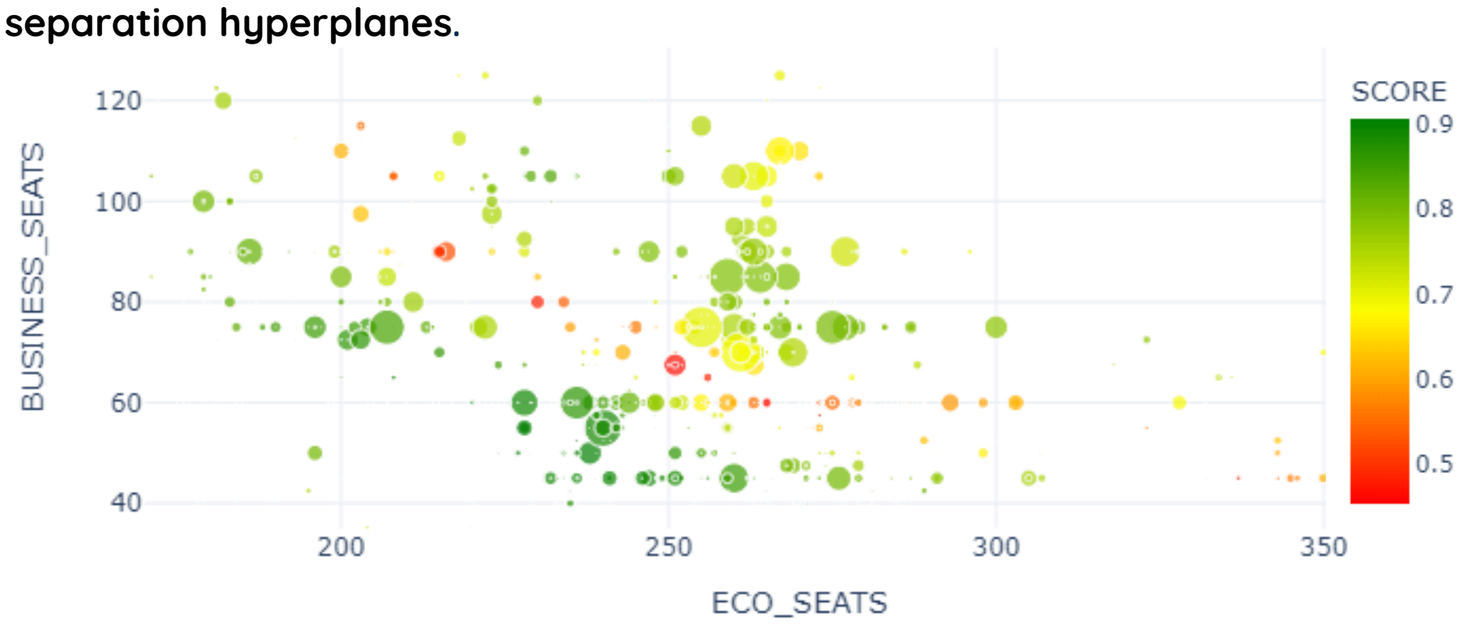

Fig 6 : Example of A330 data after cleaning and coloring by a prediction confidence score

Quantification of OAG error

We analyze OAG in terms of hours flown and not in number of operations. Indeed, the objective being to predict future dynamics, we need to **capture the major past dynamics** and not necessarily all operations of all airlines.

	Score		
	< 70%	>= 70 %	**Total**
Agree	0.04	0.68	0.72
Unknown	0.05	0.18	0.23
Disagree	0.03	0.02	0.05
Total	0.12	0.88	1

Fig 7 : Quantification of error for A330 data

In the A330 example, we find the **same code in 72%** of cases. SVM helped to classify **23% of unknown codes**. In 5% of cases, the initial code and the prediction disagree.

Final statement & future steps

The SVM approach give us a cleaned database and **several quantification elements**. Some predictions may be wrong but these errors will come up over time by using OAG and will be verified and then corrected by humans.

Having this database allows for **more precise analyses** and to **cross-reference** it with other databases. For example, it is interesting to put it into perspective with **CIRIUM**, which is a database containing **airline fleets**, and to look at the **evolution of aircraft configurations over time**.

Acknowledgements: I want to thank Thomas Fishman, my tutor, for guiding and supporting me.