Applied Intelligence IBERIA COE

A Global Hub of Data & Analytics experts

Víctor Aguado Martínez
Bernat Català Ulied
Joan Pau Gutiérrez Pascual

Context

Understanding situation

4 Datasets

+115K

Historical orders records

+35

European cities within supply chain

+700

Products

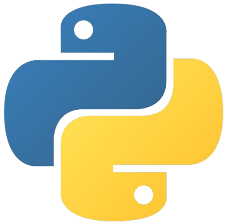
Different factors intervene in delay shipment cost and emission

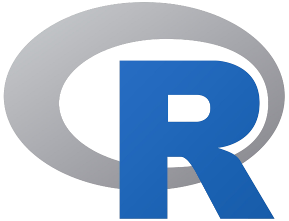
With delay in shipments, the perception of the company worsens

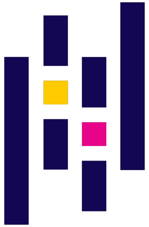
With data mining we are able to go beyond and **extract value** for our customers to **improve** their **services** and their **supply chain** in order to have less delays with a **better customer experience** and **added value** to **shipping services** through predictive **data-driven modelling**.

Methodology

Tools used







scikit
learn

statsmodels

Methodology

Process computed

The process we followed was EDA, Preprocessing, Training and Validation.

1.Data Interpretation

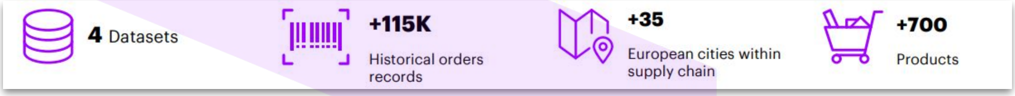
2.Data Preparation

3. Model prediction training & Validation

4. Data insights Business actions & Value proposition

Methodology

1. Data Interpretation

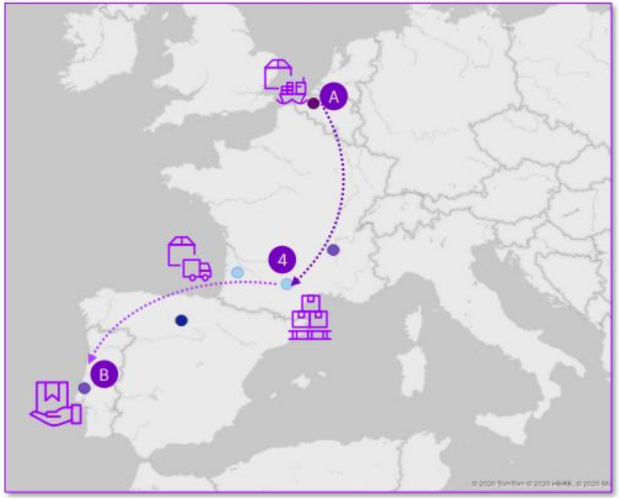


4 Datasets

+115K Historical orders records

+35 European cities within supply chain

+700 Products

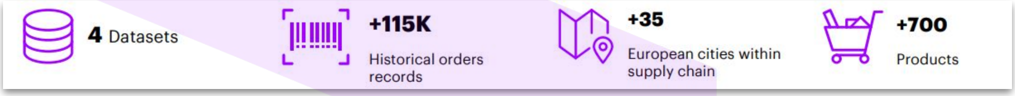


Data correction (missing values and incorrection) has been realized and **interpolated** in order **to understand the customers deliveries** and **observe** the **tendency between delays** or correct shippings

Furthermore dummies have been created to **start training** the model

Methodology

1. Data Interpretation



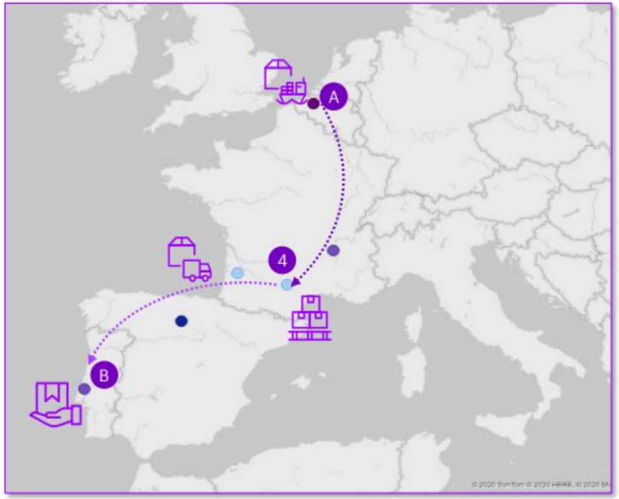
4 Datasets

+115K Historical orders records

+35 European cities within supply chain

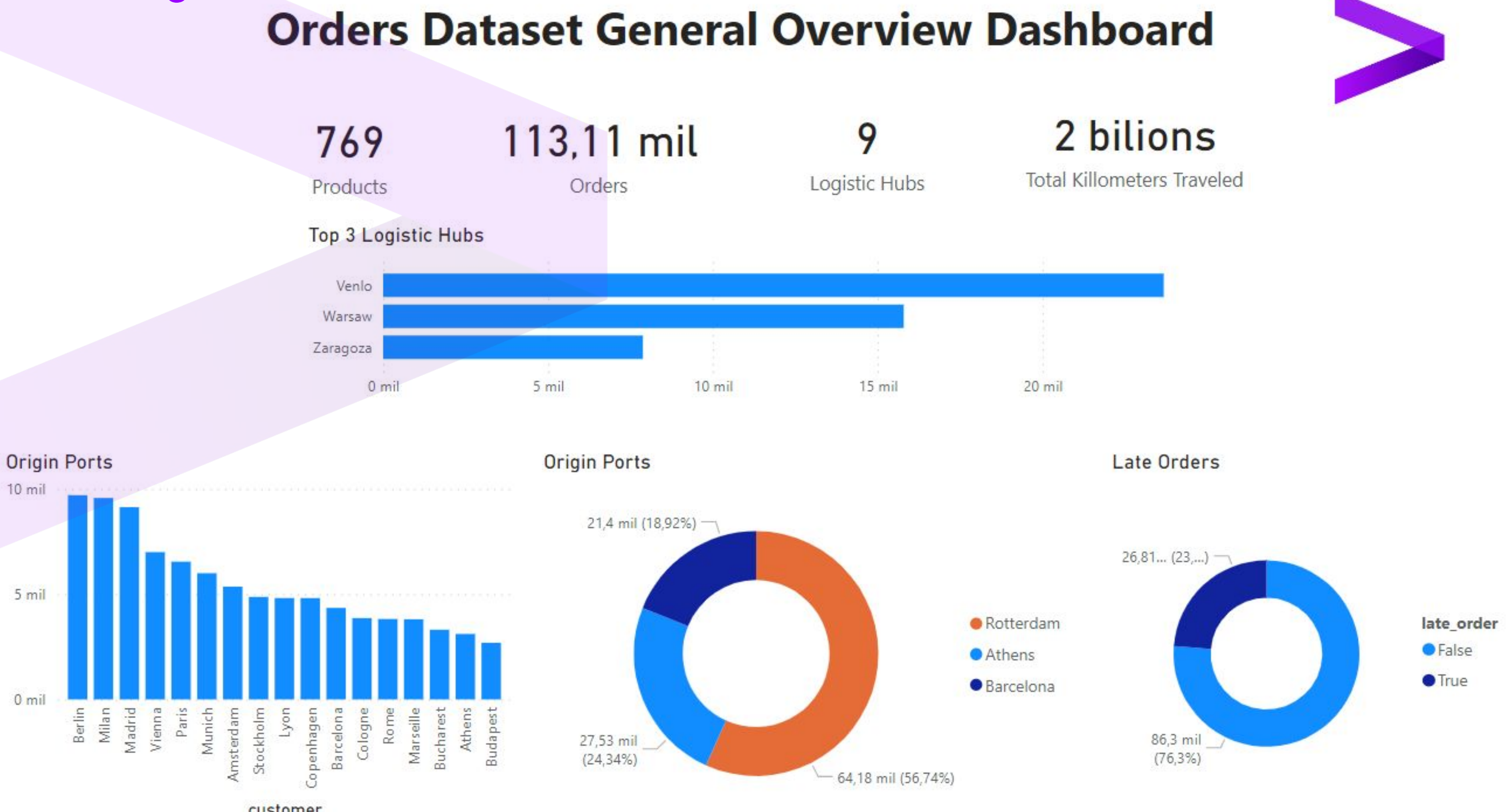
+700 Products

Once deliveries has been aggregated to our prepared dataset for trainning the model, **warehousings** and **final customers** has been **computed** to obtain delivery distances and **predict future data** with a **better performance**



Methodology

Some BI Insights…



Orders Dataset General Overview Dashboard

769
Products

113,11 mil
Orders

9
Logistic Hubs

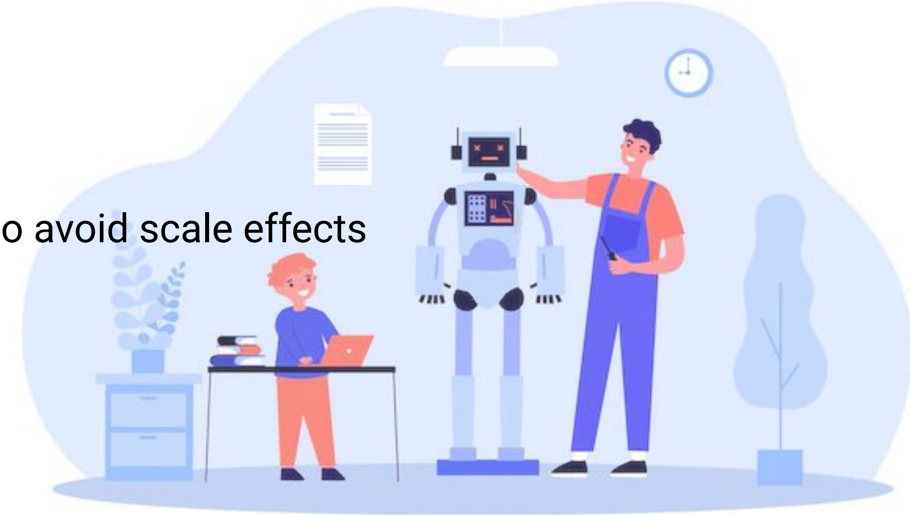
2 bilions
Total Killometers Traveled

Methodology

2. Data Preparation

- Furthermore dummies variables have been created to **learn** from categorical data
- Dataset has been **splitted*** for designing **train and evaluation sets** that will ensure model's validity
- **Normalization** of data set has been realized to avoid scale effects

*Splitting has been randomized with 80% train & 20% test



Methodology

3. Model prediction training & validation

- We have used a Generalized Linear Model (GLM) for solving this challenge.

  The main reason is that this challenge requires **explainability** and a GLM is a Statistical model that enables us to **explore the reasoning behind the predictions**.

- Dataset has been **splitted*** for designing **train and evaluation sets** that will ensure model's validity

- **Normalization** of data set has been realized to avoid scale effects

Methodology

4. Validation

For validation we have used the proposed ROC Curve and it's AUC. We also provide an accuracy metric for further evaluations.

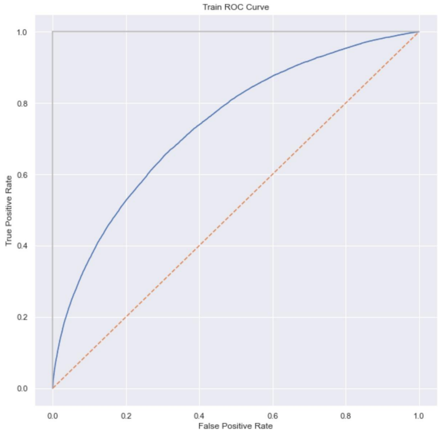
As we can see on the evaluations, the model does not present overfitting.

ROC score

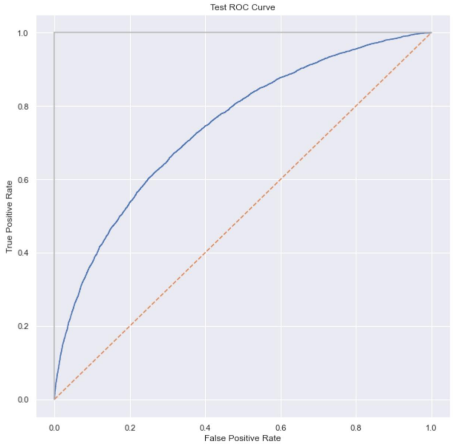
TRAIN AUC SCORE for GLM: 73.83%
TEST AUC SCORE for GLM: 74.28%

Accuracy score

Train Acc: **78.23%**
Test Acc: **78.64%**



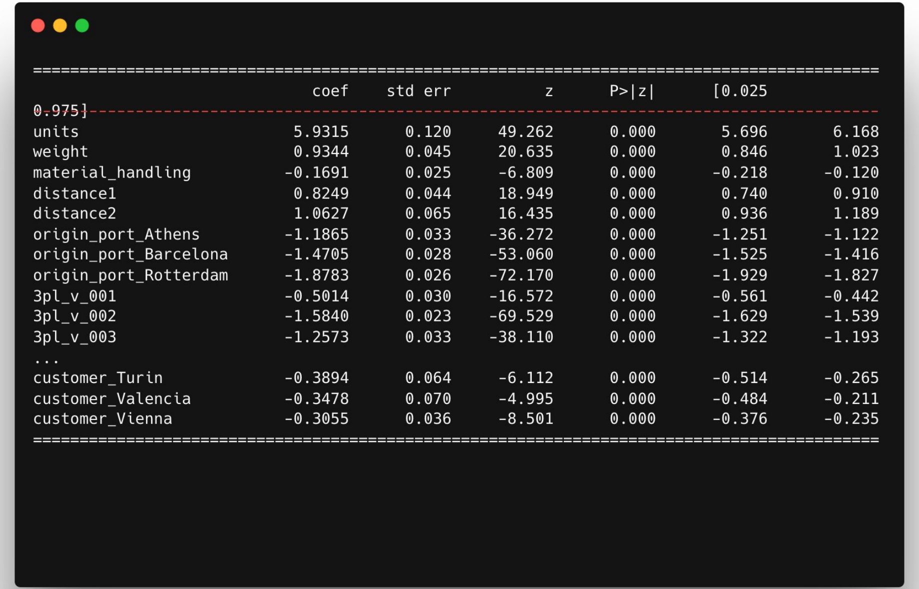
Train ROC curve



Test ROC curve

Methodology

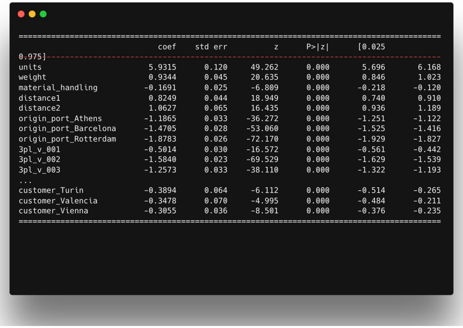
Model explainability



```
==============================================================================================
                              coef    std err          z      P>|z|      [0.025
0.975]----------------------------------------------------------------------------------------
units                       5.9315      0.120     49.262      0.000       5.696       6.168
weight                      0.9344      0.045     20.635      0.000       0.846       1.023
material_handling          -0.1691      0.025     -6.809      0.000      -0.218      -0.120
distance1                   0.8249      0.044     18.949      0.000       0.740       0.910
distance2                   1.0627      0.065     16.435      0.000       0.936       1.189
origin_port_Athens         -1.1865      0.033    -36.272      0.000      -1.251      -1.122
origin_port_Barcelona      -1.4705      0.028    -53.060      0.000      -1.525      -1.416
origin_port_Rotterdam      -1.8783      0.026    -72.170      0.000      -1.929      -1.827
3pl_v_001                  -0.5014      0.030    -16.572      0.000      -0.561      -0.442
3pl_v_002                  -1.5840      0.023    -69.529      0.000      -1.629      -1.539
3pl_v_003                  -1.2573      0.033    -38.110      0.000      -1.322      -1.193
...
customer_Turin             -0.3894      0.064     -6.112      0.000      -0.514      -0.265
customer_Valencia          -0.3478      0.070     -4.995      0.000      -0.484      -0.211
customer_Vienna            -0.3055      0.036     -8.501      0.000      -0.376      -0.235
==============================================================================================
```

Use cases

Model Explainability



```
==========================================================================================
                              coef    std err          z      P>|z|      [0.025      
0.975]------------------------------------------------------------------------------------
units                       5.9315      0.120     49.262      0.000       5.696       6.168
weight                      0.9344      0.045     20.635      0.000       0.846       1.023
material_handling          -0.1691      0.025     -6.809      0.000      -0.218      -0.120
distance1                   0.8249      0.044     18.949      0.000       0.740       0.910
distance2                   1.0627      0.065     16.435      0.000       0.936       1.189
origin_port_Athens         -1.1865      0.033    -36.272      0.000      -1.251      -1.122
origin_port_Barcelona      -1.4705      0.028    -53.060      0.000      -1.525      -1.416
origin_port_Rotterdam      -1.8783      0.026    -72.170      0.000      -1.929      -1.827
3pl_v_001                  -0.5014      0.030    -16.572      0.000      -0.561      -0.442
3pl_v_002                  -1.5840      0.023    -69.529      0.000      -1.629      -1.539
3pl_v_003                  -1.2573      0.033    -38.110      0.000      -1.322      -1.193
...
customer_Turin             -0.3894      0.064     -6.112      0.000      -0.514      -0.265
customer_Valencia          -0.3478      0.070     -4.995      0.000      -0.484      -0.211
customer_Vienna            -0.3055      0.036     -8.501      0.000      -0.376      -0.235
==========================================================================================
```

Units is the *root cause* that most affect to delays

Distance Logistical hub - customer destination is more critical than Distance from origin - Logistical hub

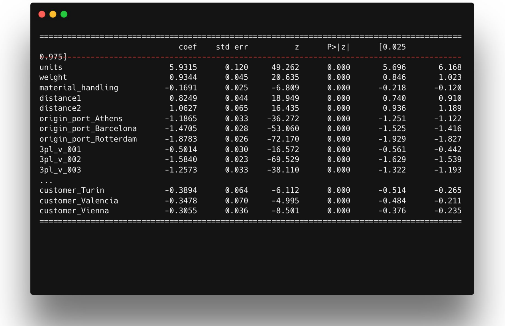
Package weight is the 3rd cause of delays

Rotterdam is the most confident logistic hub and Athenes the worst

👍 Rotterdam 😓 Rotterdam

Use cases

Model Explainability



```
==============================================================================
                            coef    std err          z      P>|z|      [0.025
0.975]------------------------------------------------------------------------
units                     5.9315      0.120     49.262      0.000       5.696       6.168
weight                    0.9344      0.045     20.635      0.000       0.846       1.023
material_handling        -0.1691      0.025     -6.809      0.000      -0.218      -0.120
distance1                 0.8249      0.044     18.949      0.000       0.740       0.910
distance2                 1.0627      0.065     16.435      0.000       0.936       1.189
origin_port_Athens       -1.1865      0.033    -36.272      0.000      -1.251      -1.122
origin_port_Barcelona    -1.4705      0.028    -53.060      0.000      -1.525      -1.416
origin_port_Rotterdam    -1.8783      0.026    -72.170      0.000      -1.929      -1.827
3pl_v_001                -0.5014      0.030    -16.572      0.000      -0.561      -0.442
3pl_v_002                -1.5840      0.023    -69.529      0.000      -1.629      -1.539
3pl_v_003                -1.2573      0.033    -38.110      0.000      -1.322      -1.193
...
customer_Turin           -0.3894      0.064     -6.112      0.000      -0.514      -0.265
customer_Valencia        -0.3478      0.070     -4.995      0.000      -0.484      -0.211
customer_Vienna          -0.3055      0.036     -8.501      0.000      -0.376      -0.235
==============================================================================
```

Third party logistic company	RANKING
V_002	1
V_003	2
V_001	3

TOP 3 - 3rd party logistic companies that provide better solutions against delays

Conclusions

How those insights can be translated to business actions and value proposition?

Dealing with root causes …

Try to optimize of units deliveries and control packages where amount of units is elevated

Try to optimize* and track units deliveries in logic hub and between final destination

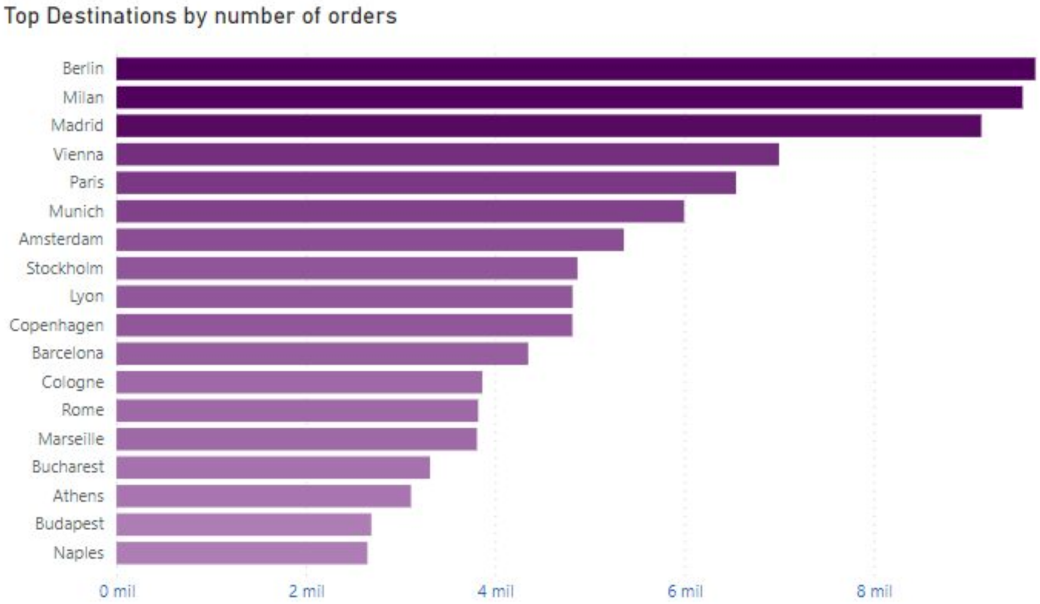
Try to optimize weight in packages by offering customers better solutions

Try to perform Athenes logic hub KPI or try to move location or reorganize it

Conclusions

How those insights can be translated to business actions and value proposition?

* We would propose to place new logistic hubs on locations based on top destinations, in order to optimize CO2 emissions and shipment distance.



Thank you 4 your attention !

Repository is available to check it out (⭐ are welcomed ! :))

GitHub

https://github.com/bcatala/Datathon