foxintelligence

Technical Test Data Quality Engineer - DataScience Team

Business Test

The client wants us to evaluate the competitor's "Deliveroo Plus" program using a tag to estimate profitability thresholds and potential impact on order frequency (both on their platform and competitor's) to decide if a similar program is worthwhile.

1 - Overview

The dataset consists of customer food order information in France. Each row represents a single order and contains the following attributes:

id_customer: Unique identifier for the customer.

merchant_name: Name of the food delivery platform used (e.g., Uber Eats, Deliveroo, Just Eat).

id_order: Unique identifier for the order.

order_date: Date the order was placed.

order_total_paid: Total amount paid by the customer for the order.

order_delivery_fee: Delivery fee charged by the platform.

order_total_promo: Amount of any promotional discount applied to the order.

order_tip: Optional tip amount given to the delivery driver.

order_delivery_zipcode: Delivery zip code for the order.

2- Technologies used

Python

Pandas: pandas is a powerful Python library specifically designed for data manipulation and analysis.

Matplotlib: matplotlib is a popular Python library for creating static, publication-quality visualizations.

Visual Studio Code (with Jupyter Notebook extension)

3 - Compute the market share per merchant per month for 2018 (in terms of number of orders)

Prior to directly addressing the question, I undertook 4 essential steps :

. Making A Copy Of The Dataset

. Exploratory Data Analysis (EDA)

I began by conducting an Exploratory Data Analysis (EDA) on the dataset. This initial phase involved getting acquainted with the data by:

Identifying the data types for each attribute.

Checking for missing values and outliers in each column.

Checking for duplicated values.

Analyzing summary statistics like mean, median ...

. Data Cleaning

Based on the insights gained from the EDA, I performed data cleaning tasks to ensure the accuracy and consistency of the analysis. This involved:

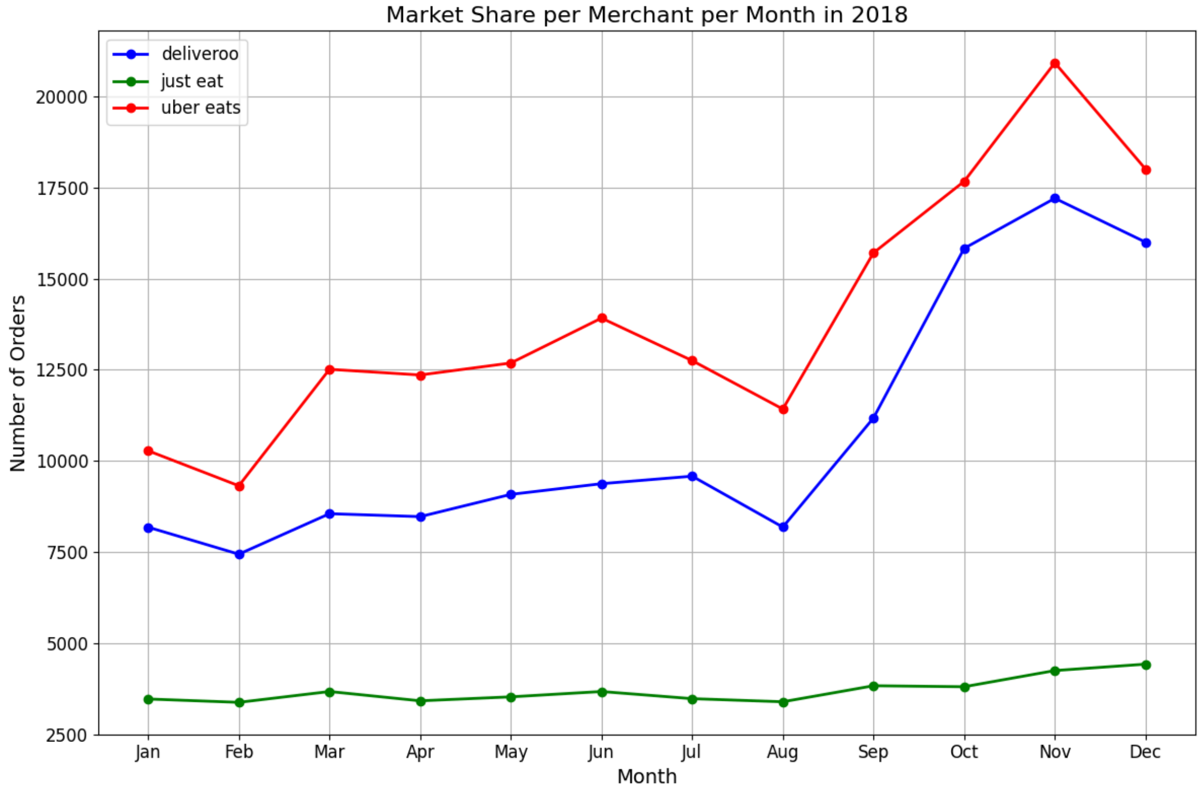
Handling missing values by imputing missing entries.

Removing the zip code column.

Check if there's any duplicated rows.

Changing order_date from object to datetime

. Data Viz

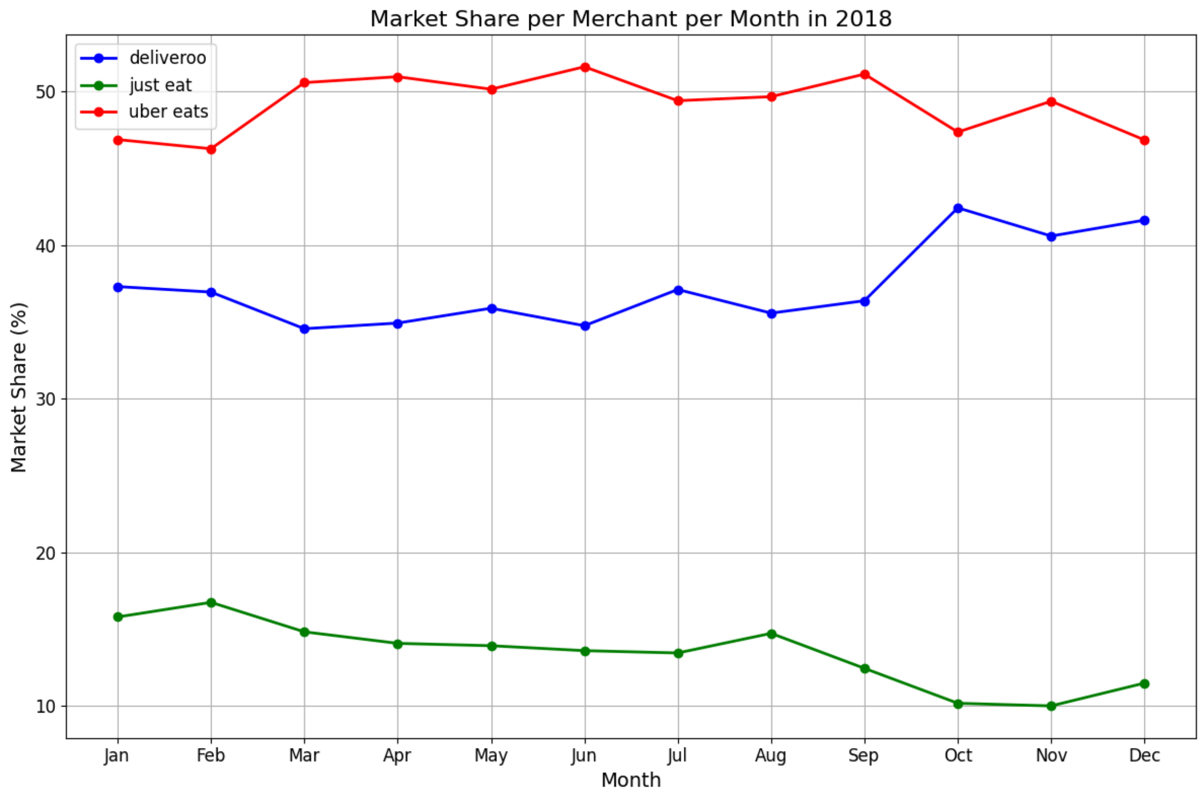


The visualization type: Line plot

The data it represents: Number of orders per merchant per month for 2018

Key insights :

> The most popular merchant by month in 2018 is “uber eats”.
>
> The periode with the most number of orders is between October and December.



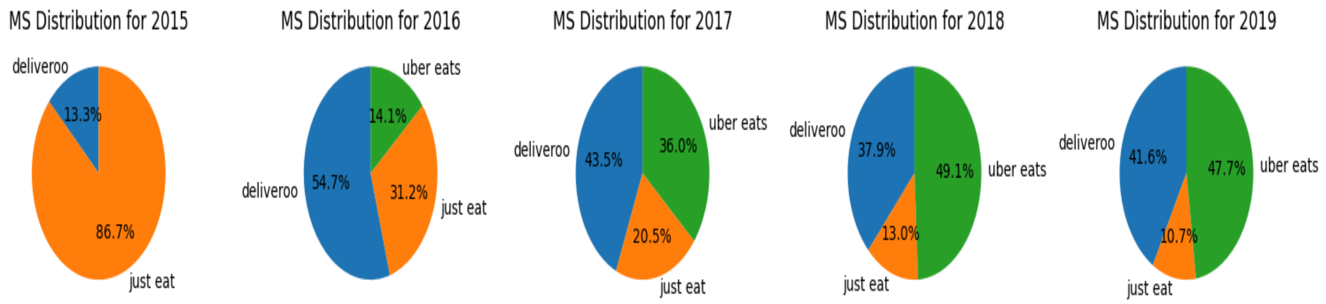
The visualization type: Line plot

The data it represents: Market share per merchant per month for 2018

Key insights :

The dominant merchant by month in 2018 is “uber eats” with approximately 50% of the whole market .

Since October deliveroo is gaining more popularity and increasing their market share (maybe due to the new deliveroo plus program).



The visualization type: Pie chart

The data it represents: Market share per merchant per year

Key insights :

“Just eat” dominated the market for the year 2015 but gradually losing market share especially when “uber eats” started its services in France starting from 2016.

“Deliveroo" had 13.3% market share in 2015 but in 2016 and 2017 “deliveroo” dominated the market with 54.7% and 43.5% repectevely .

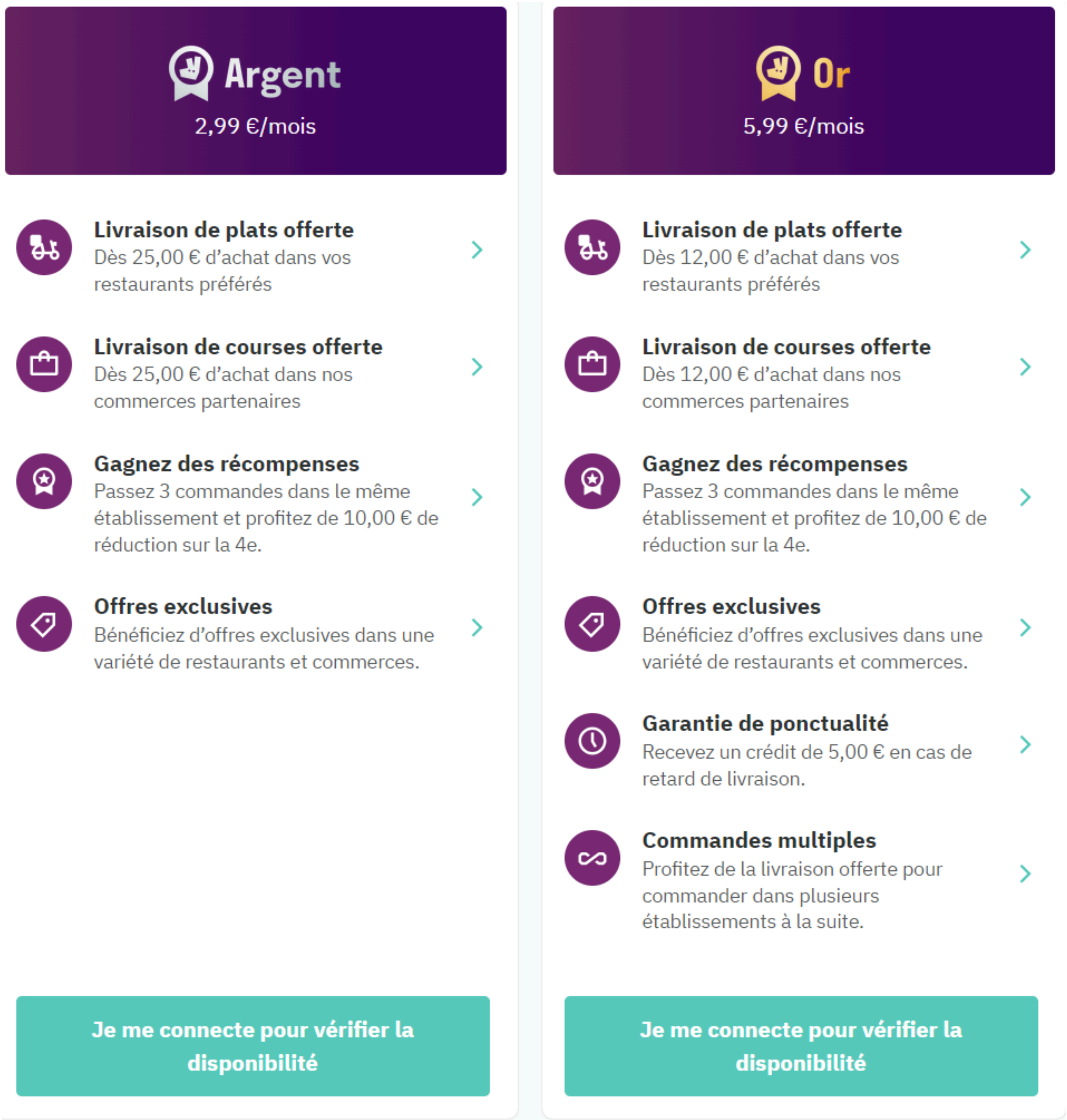
4 - Define and implement a methodology to identify users that have subscribed to "Deliveroo Plus" loyalty program and tag their orders.

To identify Deliveroo Plus users and tag their orders, this is how i approached this challenge:

Defining the Marker:

The dataset lacks direct information about Deliveroo Plus subscriptions.

To define markers for identifying potential users in the dataset I used the official “deliveroo” website.



Argent

2,99 €/mois

Livraison de plats offerte
Dès 25,00 € d’achat dans vos restaurants préférés

Livraison de courses offerte
Dès 25,00 € d’achat dans nos commerces partenaires

Gagnez des récompenses
Passez 3 commandes dans le même établissement et profitez de 10,00 € de réduction sur la 4e.

Offres exclusives
Bénéficiez d’offres exclusives dans une variété de restaurants et commerces.

Or

5,99 €/mois

Livraison de plats offerte
Dès 12,00 € d’achat dans vos restaurants préférés

Livraison de courses offerte
Dès 12,00 € d’achat dans nos commerces partenaires

Gagnez des récompenses
Passez 3 commandes dans le même établissement et profitez de 10,00 € de réduction sur la 4e.

Offres exclusives
Bénéficiez d’offres exclusives dans une variété de restaurants et commerces.

Garantie de ponctualité
Recevez un crédit de 5,00 € en cas de retard de livraison.

Commandes multiples
Profitez de la livraison offerte pour commander dans plusieurs établissements à la suite.

Je me connecte pour vérifier la disponibilité
Je me connecte pour vérifier la disponibilité

key points to identify potential Deliveroo Plus users based on the website::

- **Free Delivery:**

  Deliveroo Plus offers two tiers: Argent (€2.99/month) and Or (€5.99/month).

  Both tiers have free delivery on orders above €25.00 from restaurants and grocery stores.

  OR tier have free delivery on orders above €12.00 from restaurants and grocery stores.

- **Rewards Program:**

  Deliveroo Plus users might benefit from a rewards program where they earn rewards for placing multiple orders at the same establishment.

- **Guaranteed On-Time Delivery:**

  Plus users might have some guarantee for on-time delivery.

- **Multiple Orders:**

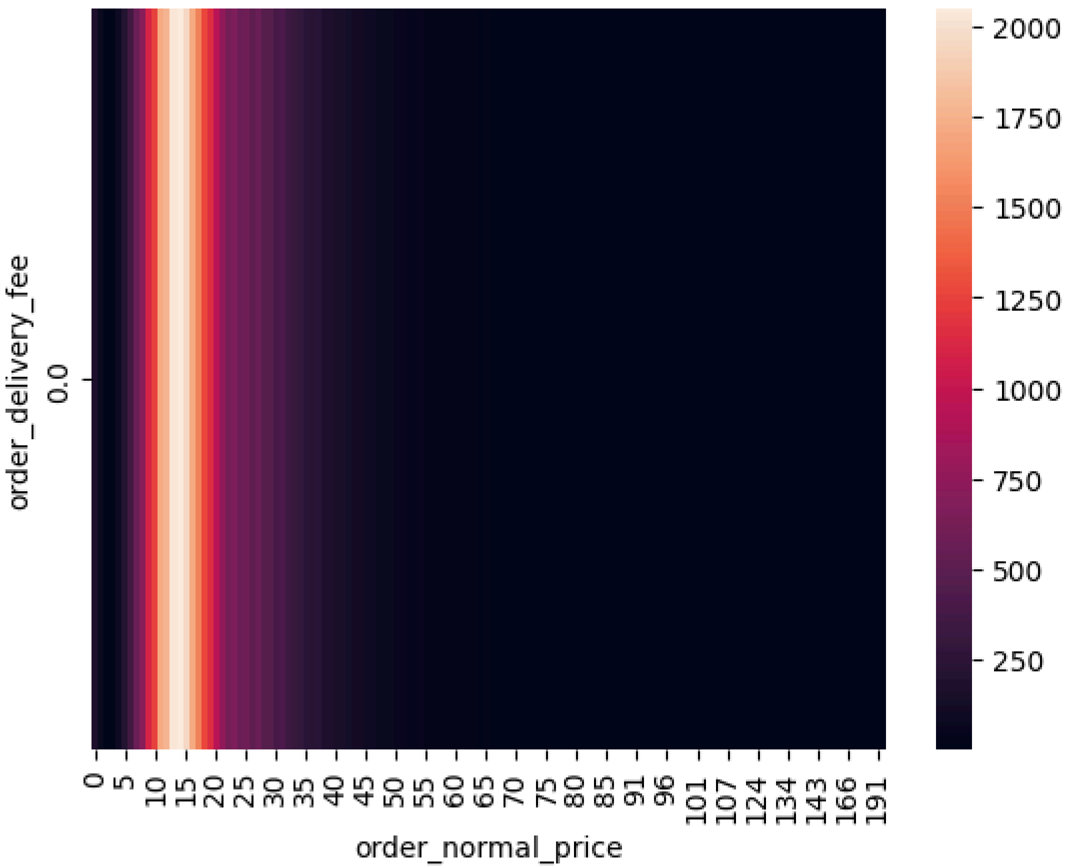
  Plus users might be able to place multiple orders from different establishments and potentially benefit from free delivery.

Because of the nature and the informations i have in the dataset i decided to identify deliveroo plus users using free delivery :

1. Filtered the dataset to include only the “deliveroo” orders.
2. Added a column named “ order_normal_price” that is equal to “order_total_paid” - “order_delivery_fee” + “order_total_promo”. That's because I want to avoid misleading results.
3. Filtered the “deliveroo” dataset to add two conditions (“order_date” >= 2018 september And “order_delivery_fee” = 0)
4. Converted the type of “order_normal_price” from float to int for visibility purposes.
5. Used a heatmap to identify the "order_normal_price" distribution for orders with free delivery fees after september 2018

Using this method I succeeded in identifying the value of “order_normal_price” when we have a significant number of orders that is **€12.**

The heatmap showed a spike after the €12 cap for "order_normal_price".



Based on this insight that matches the information from the website I could flag those orders by adding a column named "deliveroo_plus_order" that has 1 if the order is plus or 0 if the order is not.

- Neglected the orders that might be a plus order but are less than €12
- Finally I saved the dataset as csv that has the deliveroo plus order flagged.

Dealing with Unsubscriptions:

For users who display Plus behavior, then revert, i suggest to consider a grace period (e.g. 1 month) before removing the Plus tag.

Temporality Management:

Without direct subscription information, I cannot capture exact subscription start and end dates. All I can do is identify the deliveroo plus orders.

5 - Drawbacks of the Method for Identifying Potential Deliveroo Plus Users

Data Analysis Limitations:

Data bias: The method relies on the assumption that users only enjoy free delivery with Deliveroo Plus. One-off promotions or restaurant-specific offers can skew the results.

Marker Reliability:

Free delivery: Other platforms and restaurants offer free deliveries without a subscription, which can dilute the reliability of this marker.

Order frequency: Increased order frequency may have other causes besides a Deliveroo Plus subscription, such as changing dietary habits or a promotion.

Rewards program: Information about the rewards program is not available.

Generalizability of Results:

Service evolution: Deliveroo Plus offers and features are constantly evolving. Analyzing the data using one point of view may not reflect the current reality of the service.

6 - Method Improvement

Machine Learning Integration:

Supervised Learning: If user subscription status is available, train supervised learning models to identify patterns in user behavior that differentiate Plus members from non-members. Utilize features like order frequency, average order value, delivery fee behavior, and restaurant types frequented.

Unsupervised Learning: If user subscription status is unavailable, leverage unsupervised learning techniques like clustering to identify user groups based on their ordering habits. Analyze these clusters to see if specific clusters exhibit characteristics potentially linked to Deliveroo Plus usage.

Enriching Dataset Features:

Delivery Information: Including delivery timestamps to analyze delivery speed and potential correlations with Plus membership.

Order Content: Incorporating details about ordered items (categories, quantities) to explore potential trends in Plus member purchases.

User Reviews: Analyzing user reviews for mentions of Deliveroo Plus benefits or frustrations to gain qualitative insights into user behavior. Utilizing sentiment analysis techniques to identify positive mentions of Plus features.

7 - Conclusion

By addressing data limitations, incorporating machine learning algorithms, and enriching the dataset with additional features, we can develop a more robust and reliable method for identifying potential Deliveroo Plus users.