12GO - Data Scientist - Test case

2025

Task

You are provided with a dataset, “**data.csv**”, containing over 3 million trip bookings, some of which are labeled as fraudulent.

Objective:
Your task is to develop a fraud detection model to predict whether a booking is fraudulent or not. While the dataset is anonymized and partially synthetic, key patterns relevant to fraud detection have been preserved.
We do not expect a perfect F1 score, instead, we value your approach, methodology, and the reasoning behind how you achieve the results :)

Deliverables:
You are free to deliver your results in any format of your choice, link to Git repo, Google Colab, or just Jupyter Notebook (.ipynb). Please ensure the following:

- Your work is clear and easy to follow.
- Include a conclusion section summarizing your key findings. This can be part of the notebook, a README file, or a separate presentation — whichever you prefer.

Additional Information:
A detailed description of the dataset’s columns can be found in the next slide.

Data fields

Booking & Transaction Information

1. **bid** - Booking ID (unique identifier for each booking).
2. **channel** - User acquisition channel:
    a. **direct**: Direct visits from browsers
    b. **organic**: Search engine results
    c. **affiliate**: Traffic from affiliate partners
    d. **referral**: Referrals from non-affiliate websites
    e. **paid**: Paid marketing campaigns (PPC)
3. **createdon** - Timestamp of booking creation.
4. **paidon** - Timestamp of payment confirmation.
5. **godate** - Trip departure date and time.
6. **cust_name** - Traveler's name (as per booking).
7. **payer_name** - Name of the account holder used for the transaction.
8. **payer_country** - Country of the payer.
9. **usr_name** - Registered user's name.
10. **role_id** - Role of the user (e.g., admin, regular user).
11. **vehclass_id** - transport type.
12. **seats** - Number of seats booked.
13. **netprice_thb** - Net price of the trip in Thai Baht (excluding commissions).
14. **insurance_flg** - Flag indicating if insurance was purchased with the booking.
15. **p_attempts** - Number of payment attempts made by the user.

....**Next slide**

Data fields

User Information

1. **date_of_birth -** User's date of birth**.**
2. **email -** User's email address.
3. **nationality -** User's nationality.

3rd Party Risk Data

1. **email_domain_score** - Risk score based on the email domain reputation.
2. **email_score** - Overall risk score of the email.
3. **passenger_passport_score** - Risk score based on passport information.
4. **passenger_score** - Overall risk score of the passenger.

Security & Device Information

1. **ip** - IP address of the user.
2. **proxy** - Indicates if a proxy was detected (binary flag).
3. **tor** - Indicates if the connection was through the Tor network (binary flag).
4. **vpn** - Indicates if a VPN was detected (binary flag).
5. **useragent** - User-Agent string identifying the device used.
6. **recent_abuse** - Indicates recent suspicious or abusive activity associated with the user or IP.

Target Variable

1. **isFraud** - Target variable: flag indicating whether the transaction is fraudulent (1) or not (0).