Background

One of monday.com product managers comes to you and say that they need your help:
“We get thousands of new customers every day, and only have a few dozen consultants to work with, you have to carefully pick which accounts get that special VIP consulting services”

The problem

You are assigned to design and implement a lead scoring classification model (with binary target - lead_score) - in order to choose which accounts should receive the vip consulting

The data

Here is a description of the data available and their links to download them:

Users:

- **account_id:** Unique identification of the account
- **user_id:** Unique identification of the user
- **email:** Email of the user
- **name:** Full name of the user
- **created_at:** User registration date
- **is_admin:** Indication if the user is an admin in the account (admin privileges)
- **pending:** Pending invitation (whether the user accepted the invitation)
- **enabled:** Enabled to use the platform
- **became_active_at:** When the user became active
- **time_diff:** Time diff in relation to UTC
- **city:** IP-based city
- **region:** IP-based region
- **country:** IP-based country
- **serial_number:** User registration order in the account (the first user in the account will be 1, second 2 etc.)
- **has_photo:** Whether the user uploaded a photo to his profile
- **device:** User registration device
- **os:** User registration os
- **browser:** User registration browser
- **language:** User system language at registration
- **seniority:** Seniority of the user (Executive, manager, etc.)
- **has_phone:** Whether the user added his phone number

Accounts:

- **account_id:** Unique identification of the account
- **account_name:** Company name
- **created_at:** Account creation date
- **plan_id:** If the account is paying, this is the plan identifier
- **trial_start:** Start of the trial date
- **churn_date:** When did the account terminate the contract with us
- **churn_reason:** Cause of termination of the contract with us
- **time_diff:** Time diff in relation to UTC
- **region:** Region of the first user
- **country:** Country of the first user
- **subscription_started_at:** When did the account first start paying
- **paying:** Is the account currently paying
- **has_logo:** Whether the account uploaded a logo
- **device:** The device associated with the first user
- **os:** The OS associated with the account
- **browser:** The browser associated with the account
- **collection_21_days:** How much money the account paid in the first 21 days
- **company_size:** The size of the company by the “know the customer” survey
- **payment_currency:** Payment currency
- **max_team_size:** The size of the company by the “know the customer” survey
- **min_team_size:** The size of the company by the“know the customer” survey
- **industry:** The kind of industry of the account
- **utm_cluster_id:** The type of work the account was targeted for
- **mrr:** Monthly recurring revenue of the account
- **user_goal:** A user-reported role through a survey
- **user_description:** A user-reported specific description through a survey
- **team_size:** Truncated team size from survey
- **lead_score:** Accounts that could use consulting services (yes-1, no-0) OUR TARGET

Subscriptions:

- **event_happened_at:** When the contract was made
- **subscription_id:** Unique identification of the transaction
- **account_id:** Unique identification of the account
- **plan_id:** Unique identification of the plan
- **event_type:** Type of payment/credit
- **invoice_charge_amount:** The charge amount in the user currency
- **prev_plan_id:** Previous plan in case of a plan change
- **status:** If the payment was successful or failed
- **status_reason:** The cause of failure if any
- **currency:** Currency
- **invoice_charge_amount_usd:** Conversion to dollars value

- **mrr_gain:** Monthly recurring revenue for these transactions
- **next_charge_date:** If a recurring transaction, when is the next one
- **payment_type:** Payment method
- **transaction_date:** Transaction date (Important for future transactions)

Events:

- **date:** The date by which the events were aggregated
- **user_id:** Unique identification of the user
- **account_id:** Unique identification of the account
- **total_events:** Number of events for the user in that day
- **column_events:** Aggregate of column related events
- **board_events:** Aggregate of board related events
- **num_of_boards:** Number of boards the user used that day
- **count_kind_columns:** Number of column types the user used that day
- **content_events:** Aggregate of content modification events
- **group_events:** Aggregate of group related events
- **invite_events:** Aggregate of invites sent that day by the user
- **import_events:** Aggregate of import related events
- **notification_events:** Aggregate of events about notifications
- **new_entry_events:** Number of new sessions by the user that day
- **payment_events:** Aggregate events that contain data about the payments
- **inbox_events:** Aggregate events that contain data about the user inbox
- **communicating_events:** Aggregate communication events within the account
- **non_communicating_events:** Aggregate events that contain general usage in the platform
- **web_events:** Aggregate events that happened in the web interface
- **ios_events:** Aggregate events that happened in the ios app
- **android_events:** Aggregate events that happened in the android app
- **desktop_app_events:** Aggregate events that happened in the desktop app
- **empty_events:** Aggregate of general events with no specific category

Links for datasets:
https://drive.google.com/uc?id=1Of2nYW3tZvLkxRZfFlt4ogFU6CoibU27&export=download
https://drive.google.com/uc?id=1O_ccCjbemIsAmTrKmcBHMt92QY1oegWA&export=download
https://drive.google.com/uc?id=1ObK_sookmQkSBJJ2afyoLmqs0QBT0tlZ&export=download
https://drive.google.com/uc?id=1OewkTcswcsEiMHlUNFQuttuXeB7KsVje&export=download

The solution

These are the deliverables expected:

1. dev design doc:
    a. Based on the data you have, suggest 2 models that are relevant here an explain your choice
    b. Performance metrics: define how you are going to measure their performance and explain why

2. Code and ML (python):
   Implement the models you suggested with all relevant data and ML development steps, and select the best model according to your metrics

Notes:

- Track and document your steps, explain the logic and thought
- Make assumptions - but pls document them
- **The task is meant to be completed in a one-day effort**; please use that as a benchmark for managing your time, and be assured we will adjust our expectations accordingly - we are aware you are busy with life, so you can spread the work and return it in several days

If you have any questions feel free to reach out to ohad@monday.com.