TECH-TRIATHLON

A competition by rootcode

DATATHON CHALLENGE

The Challenge

Congratulations on your amazing work in the Designathon and Hackathon phases! You've already created a robust booking system that helps citizens across the country schedule appointments with government services. Now, it's time to supercharge your solution with the power of AI.

In today's fast-paced world, time is everything. Citizens visiting government institutions often have no idea how long it will take to complete their service - from when the staff starts working on their request to the moment it's done. What if your system could tell them in advance?

Task 1 - Predict Service Completion Time

Your first challenge is to predict the processing time for a booked service before the citizen even sets foot in the office. Please note that this is not the time they spend waiting in the queue - it's the time the staff will take to complete the task once started.

For example:
If someone books a selected service for 29 August 2025, your model might predict:
Expected completion time: 75 minutes.

Exact Inputs & Outputs

- **Input**

1. date (string, YYYY-MM-DD) — appointment date
2. time (string, HH:MM, 24h, local office time) — appointment start time
3. task_id (string) — the existing task ID from the Tasks Dataset

- **Output**

expected_completion_time_minutes (integer) — predicted processing time (minutes) for the staff to finish the task once started

Task 2 - Predict Staffing Needs

Government offices are divided into sections, each handling specific services. One of their biggest challenges is workforce planning. Therefore, your second challenge is to forecast the number of employees needed in each section for a given date. This helps managers allocate resources efficiently.

Exact Inputs & Outputs

- **Input**

1. date (string, YYYY-MM-DD) — the day to forecast staffing for
2. section_id (string) — the existing section ID from the Tasks/Staffing datasets

- **Output**

predicted_employee_count (integer) — employees needed in that section on that date for smooth operation

Note: The datasets required for these tasks, along with their descriptions can be found in the "Dataset Documentation" section at the end of this booklet.

The Dataset (Training Data You'll Receive)

In a real deployment, your AI will gather valuable new data over time. For this challenge, you'll receive **three linked training datasets** (joined via task_id and section_id):

1. Bookings Dataset – Citizen appointments & journey details
2. Tasks Dataset – Task details & assigned section (you will fill names)
3. Staffing Dataset – Daily staffing and workload records for each section

Important:
In the Tasks Dataset, you will see: task_id, task_name, section_id, section_name.

- We have pre-filled all task_id and section_id values.
- Your job: Fill in task_name and section_name to match your own use case.
- Do not change the IDs or any other dataset fields.
- You must define exactly 6 sections, with related tasks assigned to each section.
- All times are in Asia/Colombo and use a 24-hour clock.

Example Sections & Tasks

You must define exactly 6 sections for your solution, each with related tasks.
We have mentioned some example sections with relevance to the Department of Immigration & Emigration and tasks they might do down below, feel free to adapt or rename to match your scenario.

Department of Immigration & Emigration

First-time Passport Applications

- Accept and verify new passport applications
- Capture applicant biometrics (photo, fingerprints)

Renewals

- Process passport renewal requests
- Verify and update biometric data if required

Corrections & Amendments

- Correct name or date of birth errors
- Update address or legal name change details

Lost/Stolen Passport Reissue

- Record lost/stolen passport incident reports
- Issue replacement passports after verification

Document Verification

- Check authenticity of submitted documents
- Cross-verify documents with government databases

Special Cases

- Process diplomatic or official passports
- Handle urgent/emergency passport requests

Rules and Regulations

- **Deadline:** The submission form closes after the deadline.
- **Tasks Dataset Restriction:** You must only edit task_name and section_name.
- **Model Restrictions:** You are restricted from using any pre-trained models, except for synthetic data generation or pre-processing
- **API Usage:** Proprietary API-based modelling/preprocessing is prohibited.
- **Integrity:** Cheating, plagiarism, or rule violations will result in disqualification.

Terms and Conditions

1. **Use of Data:** The provided datasets may be used solely for the purpose of this competition. Any other use, including but not limited to commercial purposes, academic research, or personal projects, is strictly prohibited.
2. **Data Sharing:** The datasets must not be shared, distributed, or transmitted in any form - whether publicly or privately - to any third party. This includes uploading the datasets to external websites, forums, or social media platforms.
3. **Publication and Disclosure:** You are not permitted to publish, disclose, or make the datasets or any derivatives publicly available unless explicitly authorized by the competition organizers.
4. **Data Confidentiality:** By participating in the competition, you agree to maintain the confidentiality of the datasets and any sensitive information contained within them.
5. **Violation of Terms:** Any violation of these terms and conditions may result in disqualification from the competition.

Judging Criteria

- Data Wrangling - 15%
- Model and architecture Implementation - 35%
- Performance Score - 30%
- Creativity of the Solution / Out of the Box Thinking - 10%
- Demo Video - 10%

Deliverables

1. **Architecture Diagrams**

Short diagrams showing your model, preprocessing pipeline, and deployment idea (high level is fine).

2. **Data Pre-Processing Document**

A brief write-up of your data cleaning, feature engineering, and rationale.

3. **Model File**

Your final trained model in .h5 or .pkl format (saved alongside the notebook).

4. **Final Notebook (TeamName_FinalNotebook.ipynb)**

Notebook requirements:

- Keep all cells you used for data preprocessing, training, and evaluation exactly as they were - do not remove or alter them.
- Add one new cell at the very end that:
    - Imports your .h5/.pkl model file (assume it's in the same folder as the notebook).
    - Runs two clear inference demos:
        - Task 1: predict service completion time for a given date, time, and task_id.
        - Task 2: predict staffing needs for a given date and section_id.
    - Prints the inputs and predicted outputs clearly.

Note: Failure to include both the data preprocessing, training & evaluation details and the final model import + inference cell may result in a reduction of marks.

5. **Tasks Dataset (Filled Names)**

Submit the tasks.csv with your completed task_name and section_name values (IDs must remain unchanged).

6. **Evaluation CSV Files**

Along with the dataset, you will receive two additional files required for evaluation, located in the Evaluation Input folder inside the dataset directory. These files are named task1_test_inputs.csv and task2_test_inputs.csv, corresponding to Task 1 and Task 2 respectively. Once your models are finalized, you must run inference on these input files. For Task 1, task1_test_inputs.csv contains the following structure:

row_id	date	time	task_id
cadcsv30dfbab7586131ca2329207b9cff81d5d5	2025-08-02	12:21	TASK-001
cmm3066c865c7053bb39092977fc0e513e045159	2025-12-02	10.22	TASK-002

Deliverables

Your model should use the date, time, and task_id to predict true_processing_time_minutes and generate an output file like this:

row_id	true_processing_time_minutes
cadcsv30dfbab7586131ca2329207b9cff81d5d5	12
cmm3066c865c7053bb39092977fc0e513e045159	65

For Task 2, task2_test_inputs.csv has the following structure:

row_id	date	section_id
vk4201f75875bff49278c1328d019123f5e7d6f67	2025-04-03	SEC-002
tstb4944e3128a8dc35c448db69452b7e0e7fb6d	2025-02-11	SEC-003

Your model should use the date, time, and task_id to predict true_processing_time_minutes and generate an output file like this:

row_id	true_required_employees
vk4201f75875bff49278c1328d019123f5e7d6f67	12
tstb4944e3128a8dc35c448db69452b7e0e7fb6d	2

You must include both output files with your final submission.

7. Demo Video

- Demo Video: 3–5 minutes (Unlisted YouTube video) covering your model architecture, preprocessing, and challenges you faced.

Please add all requested deliverables into one folder. Afterward, compress this folder into a zip file, ensuring it retains the name *TeamName_Datathon.zip*. Upload the zip file to the submission form.

Deadline for submissions: **23rd August, 2025 at 11.59 PM IST**

🔗 **Submission Form:** https://forms.gle/yK7SiMDHRsXsq7b86

Dataset Documentation

This dataset represents real-world operations of a large government office, covering appointment bookings, the services provided, and staffing levels.
It is structured into three linked datasets:

- **Bookings Dataset** – Details of each citizen's appointment
- **Tasks Dataset** – A mapping of tasks to the section responsible for handling them
- **Staffing Dataset** – Daily staffing and workload records for each section

The datasets are linked through task_id and section_id.

1. Bookings Dataset

Purpose: Contains information about every booking made at the government office. This includes when it was booked, when the appointment was, which task was requested, how many documents were submitted, and the citizen's satisfaction rating.

Columns

- booking_id (string) — Unique booking reference number
- citizen_id (string) — Encoded (anonymized) ID of the citizen making the booking
- booking_date (date) — Date when the booking was made
- appointment_date (date) — Date of the appointment
- appointment_time (time) — Scheduled time of the appointment
- check_in_time (datetime) — Actual check-in time at the office
- check_out_time (datetime) — Actual check-out time from the office
- task_id (string) — Unique task ID linking to the Tasks Dataset
- num_documents (integer) — Number of documents submitted during the appointment
- queue_number (integer) — Queue position (ticket number) received upon check-in
- satisfaction_rating (integer, 1–5) — Feedback rating given by the citizen

Notes

- Processing time can be calculated as: check_out_time - check_in_time.
- Join task_id with the Tasks Dataset to determine which section handled the appointment.

Dataset Documentation

2. Tasks Dataset

Purpose: Provides details of the available tasks and maps them to the corresponding sections in the government office.

Columns

- task_id (string) — Unique ID for the task (pre-filled — do not change)
- task_name (string) — The descriptive name of the task (you must fill this)
- section_id (string) — Unique ID for the section handling the task (pre-filled — do not change)
- section_name (string) — The descriptive name of the section (you must fill this)

Instructions for Participants

- You must only fill in task_name and section_name to match your chosen use case (e.g., Passport Office, Healthcare Institute, Transport Board).
- Do not change task_id, section_id, or any other columns.
- You must have exactly 6 sections and a set of tasks assigned to each section.
- You may rename the sections and tasks to fit your scenario, but the structure and IDs remain unchanged.

3. Staffing Dataset

Purpose: Shows daily staffing levels and total productive task time for each section. This allows participants to derive the average time per task and estimate optimal staffing needs.

Columns

- date (date) — Date of staffing record
- section_id (string) — Unique ID for the section (matches section_id in Tasks Dataset)
- employees_on_duty (integer) — Number of employees present in that section on that date
- total_task_time_minutes (integer) — Total number of minutes spent by all employees completing tasks in that section on that date

Notes

- total_task_time_minutes represents the sum of actual task processing times for all bookings in that section on that date.
- From this, you can derive average time per task and estimate optimal staffing.

Click Here to Access the Datasets

WISH YOU ALL THE BEST!

For any questions, drop us an email at tech-triathlon@rootcode.io