Internship Pre-Task
Deadline: 2 July, 2023, 11:59 PM IST

Welcome to the year 2912, where your data science skills are needed to solve a cosmic mystery. The Spaceship Andromeda, With almost 13,000 passengers on board, set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.
While rounding Alpha Centauri en route to its first destination, the unwary Spaceship Andromeda collided with a spacetime anomaly hidden within a dust cloud. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension. To help rescue crews and retrieve the lost passengers, **you are challenged to predict which passengers were transported by the anomaly using records recovered from the spaceship's damaged computer system.**
Help save them and change history!

Dataset

- **train.csv** - Personal records for about two-thirds (~8700) of the passengers, to be used as training data.
    - PassengerId - A unique Id for each passenger. Each Id takes the form gggg_pp where gggg indicates a group the passenger is travelling with and pp is their number within the group. People in a group are often family members, but not always.
    - HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
    - CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
    - Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard.
    - Destination - The planet the passenger will be debarking to.
    - Age - The age of the passenger.
    - VIP - Whether the passenger has paid for special VIP service during the voyage.
    - RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
    - Name - The first and last names of the passenger.
    - Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.
- **test.csv** - Personal records for the remaining one-third (~4300) of the passengers, to be used as test data. Your task is to predict the value of Transported for the passengers in this set.
- **sample_submission.csv** - A submission file in the correct format.
    - PassengerId - Id for each passenger in the test set.

    - Transported - The target. For each passenger, predict either True or False.

- A sample image with a description of various columns is provided in the folder named train_details.png

Note: Due to the large size of the data, you are free to use the following: Google Colaboratory or Kaggle workspace.

Analysis

1. Perform necessary preprocessing and explore the data and come up with at least three insights about the data. Demonstrate your data exploration/analysis skills.
2. This analysis would be one of the main focus of our discussions in the interview. Be the spark that ignites our imagination, surprising us with your unique and visionary ideas.
3. Plot your insights strategically, like a master storyteller, to captivate your audience and ensure your message resonates.
4. Perform some dimensionality reduction and visualize the reduced feature space in 2D or 3D (use TSNE or PCA).
5. Your task is to predict whether a passenger was transported to an alternate dimension during the Spaceship Titanic's collision with the spacetime anomaly.
6. Do all the necessary steps for building a classifier and generating insights. Following is the indicative (but not complete) list of steps that you should implement
    a. Necessary data engineering
    b. Data split (train/val and perform cross-validation)
    c. Trying different models
    d. Reporting performance measures

Submission

- Put all your notebooks/code scripts in a GitHub repository.
- Maintain a README.md explaining your codebase, the directory structure, commands to run your project, the dependency libraries used, and the approach you followed.
- The link to the GitHub repository will be asked for during the interview.

Evaluation

It will be done based on the following:

- Share your strategies for effectively handling and processing sizable real-world datasets.
- Share examples of how your insights go beyond the conventional and bring a fresh perspective to the table.
- Share your approach to presenting code and findings in a clear, organized, and visually appealing manner.

Bonus Task

We all are crazy fans of cricket and cricketers in India. For the 2nd task, we have provided IPL data from 2008-2017, and ball-by-ball data of all the IPL cricket matches till season 9. The dataset contains 2 files: *deliveries.csv* and *matches.csv*.

Data

- *matches.csv* contains details related to the match, such as location, contesting teams, umpires, results, etc.
- *deliveries.csv* is the ball-by-ball data of all the IPL matches, including data of the batting team, batsman, bowler, non-striker, runs scored, etc
- Source: http://cricsheet.org/ (data is available on this website in the YAML format.)

Do same analysis as done in main task.

Disclaimer: Attempt your task independently, Plagiarism will be checked, Properly cite any resources that you may use.

The dataset is available on one drive link attached in mail.
If you have any queries, please reach out to us at nidhi.goyal@mahindrauniversity.edu.in