LatAm Automation & Advanced Analytics (A&A)
Application Case Study (ML Ops)

Introduction
Thank you for dedicating your time on this case study! We're really glad you're interested in joining the A&A team and we've done our best to make sure this is directly applicable to the role at hand. Our goal is to ensure this is a solid fit for both you and the team, therefore we appreciate all the work put into this application.

Guidelines
● This exercise is meant to be completed in 5 days;
● This exercise is meant to be completed in English (feel free to use any packages; Python is mandatory);
● All assets (including scripts, supporting files/links, prediction CSV and the final presentation slides) must be submitted at the end of the last day;
● The assets should contain all assumptions/rationale and the slides should follow the below:
○ Slide #1: cover with your full name;
○ Slide #2: problem definition,
○ Slide #3: assumptions,
○ Slide #4:final model (and its configuration),
○ Slide #5: findings,
○ Slide #6: results and next steps;,
● All analysis/code will be reviewed previously to the interviews and discussed during those, you'll be expected to defend your findings and justify your recommendations;
● Be mindful of the way you present your findings: clarity of message, storyline, structure, business acumen and problem solving are important;

Good luck! We can’t wait to see what you'll come up with! :)

Uber Rides Trip Duration Prediction
When you request an Uber ride, you're likely to notice that our price is different from the cost of the same trip a few days earlier. That’s because of our dynamic pricing algorithm, which adjusts rates based on a number of variables, such as time and distance of your route, traffic, tolls, taxes, surcharges, fees and the current rider-to-driver demand.

In efforts to better understand fares behavior, the São Paulo (Brazil) Rides' Marketplace Health team provided a CSV of trips in the city, its schema/definitions (below), and

asked the LatAmA&A team to build a model to **predict trip duration in the city of São Paulo**. In this exercise, you will be using your skills to thoroughly to:
● Understand and cleanup the data (if applicable);
● Combine provided data with external data sources (if applicable);
● Select, manipulate and engineer data into new features (if applicable);
● Build model(s) to predict Uber Rides trip duration;
● Evaluate model(s) and, if applicable, compare their respective scores (which should be defined by the candidate);

Column Name	Data Type	Description
id	string	Trip UUID
pickup_ts	timestamp	Local timestamp when the trip began
pickup_lat	float	Latitude of the pickup point
pickup_lng	float	Longitude of the pickup point
dropoff_ts	timestamp	Local timestamp when the trip ended
dropoff_lat	float	Latitude of the dropoff point
dropoff_lng	float	Longitude of the dropoff point
eta	integer	Estimated request to pickup time (in seconds)
ata	integer	Actual request to pickup time (in seconds)
trip_distance	float	Total trip distance (in kms)
trip_duration	integer	Total trip duration (in seconds)
is_airport	boolean	If trip's pickup or dropoff location is an airport
pickup_airport_code	string	Pickup airport code
dropoff_airport_code	string	Dropoff airport code
is_surged	boolean	If trip surged on rider's side (surge_multiplier > 1.0)
surge_multiplier	float	Multiplier used for final fare calculation when on dynamic pricing
driver_rating	float	Driver's rating (shown in the app) at the trip's request
lifetime_trips	integer	Driver's lifetime completed trips at the trip's request

The dataset comes in the shape of 2.1 million individual trips observations: 1.5 million training observations (latam_aa_train_data_mlops.csv) and 600k test observations (latam_aa_test_data_mlops.csv).
Important
- The prediction CSV should be the latam_aa_test_data_mlops.csv data set with an

additional column named prediction, with the chosen model's output for each row (this new column should only receive integer numbers);
- Your solution's folder structure should be submitted zipped, and ready to be unzipped and ran by the A&A team;