Data Scientist Challenge
LATAM Airlines

Instructions:

At Advanced Analytics we highly value teamwork and the constant interaction between the different roles that work in a data-based product, such as Data Scientists, Machine Learning Engineerings, and Data Engineers, to name a few. Our work is collaborative by nature, and for this reason we look after proper use of Git as an essential skill in our newest members. This challenge must be delivered through any git platform of your choice, with a small caveat: it must be public for us to see it and evaluate it.We are looking to understand how you developed your codebase, and how you improved it through time; additionally, if you have your own projects in this repository they will help us to better understand your experience based on your own previous work.

Git Instructions:
1) Create a repository on the git platform that suits you best. Remember that this platform should be public for us to see it.
2) Use a main branch for any official release that we should review, and a development branch for any increment. Optional, take up some GitFlow development practice.

Challenge Instructions:
1) You must send the link to the repository to the email from which you were contacted with the subject: **Challenge Data Scientist - [Name][Last Name],**
Example: Challenge Data Scientist - John Doe.
2) Changes will be accepted in the repository until the date and time indicated in the email.
3) In the following Google Drive folder you will find the instructions for the challenge, as well as the `dataset_SCL.csv` file that will be used during this exercise.
4) The repository should have a jupyter notebook called solution.ipynb using Python 3. Solutions in any other languages such as R will not be reviewed.
5) The solution.ipynb notebook should contain all of the answers. Partially developed challenges will not be reviewed
All the necessary files to replicate the challenge must be inside the repository. The notebook will be executed by a reviewer, and it is expected to run seamlessly
7) A copy of your Resumé in .pdf format in the repository

Challenge:

The problem consists in predicting the probability of delay of the flights that land or take off from the airport of Santiago de Chile (SCL). For that you will have a dataset using public and real data where each row corresponds to a flight that landed or took off from SCL during 2017. The following information is available for each flight:

Fecha-I: Scheduled date and time of the flight.
Vlo-I : Scheduled flight number.
Ori-I : Programmed origin city code.
Des-I : Programmed destination city code.
Emp-I : Scheduled flight airline code.
Fecha-O : Date and time of flight operation.
Vlo-O : Flight operation number of the flight.

Ori-O : Operation origin city code
Des-O : Operation destination city code.
Emp-O : Airline code of the operated flight.
DIA: Day of the month of flight operation.
MES : Number of the month of operation of the flight.
AÑO : Year of flight operation.
DIANOM : Day of the week of flight operation.
TIPOVUELO : Type of flight, I =International, N =National.
OPERA : Name of the airline that operates.
SIGLAORI: Name city of origin.
SIGLADES: Destination city name.

Challenge
1. How is the data distributed? Did you find any noteworthy insight to share? What can you conclude about this?
2. Generate the following additional columns. Please export them to a CSV file named synthetic_features.csv:
○ high_season : 1 if Date-I is between Dec-15 and Mar-3, or Jul-15 and Jul-31, or Sep-11 and Sep-30, 0 otherwise.
○ min_diff : difference in minutes between Date-O and Date-I .
○ delay_15 : 1 if min_diff > 15, 0 if not.
○ period_day : morning (between 5:00 and 11:59), afternoon (between 12:00 and 18:59) and night (between 19:00 and 4:59), based onDate-I .

3. What is the behavior of the delay rate across destination, airline, month of the year, day of the week, season, type of flight?What variables would you expect to have the most influence in predicting delays?
4. Train one or several models (using the algorithm(s) of your choice) to estimate the likelihood of a flight delay.
Feel free to generate additional variables and/or supplement with external variables.
5. Evaluate model performance in the predictive task across each model that you trained. Define and justify what metrics you used to assess model performance. Pick the best trained model and evaluate the following: What variables were the most influential in the prediction task? How could you improve the
Performance?

Additional aspects to consider in the evaluation
Order and clarity when proposing an analysis, idea, code, etc.
Creativity to solve the challenge.
Versioned code in Git.
We are not going to review excel, macros, R codes.
We will not review challenges that do not arrive on the indicated date
When in doubt, make your assumptions explicit.
Try to express yourself as best as possible to explain your decisions and answers.