PlaceSense - Data Science Test

Instructions:

1. This test is confidential, and is usually given as hard copy in our offices. Due to the remote nature of the job, we're conducting it online – nevertheless, please make sure to delete this file once the test is over.

2. This is a closed test, thus no external references should be used except for references for the Python language.

3. You are not only tested on the final result, but on the way you've gotten there. At each step don't forget to document and explain your process and choices. Explanations using visuals are greatly encouraged. Make sure your solution is elegant and efficient, written in clear and clean code and sufficiently commented.

4. Please use JupyterNotebook, it makes the whole process easier

5. Before starting to code, make sure you fully understand the task and its expected results. If not, please don't hesitate to reach out to whoever is conducting the test! Otherwise, Good Luck!

Introduction:

One of the products of the company provides estimates for the number of people that visit any given location within a day, averaged on a monthly basis. In order to predict this number we use Supervised Machine Learning.

You are provided a dataset containing 5 training locations, a set of more or less descriptive features and the real number of visits counted in these locations. The goal of this exercise is to train a model to accurately predict the number of visits observed within each location using the given (and possibly extended) set of features. Please find further details about the features below.

Step 1 - Exploration

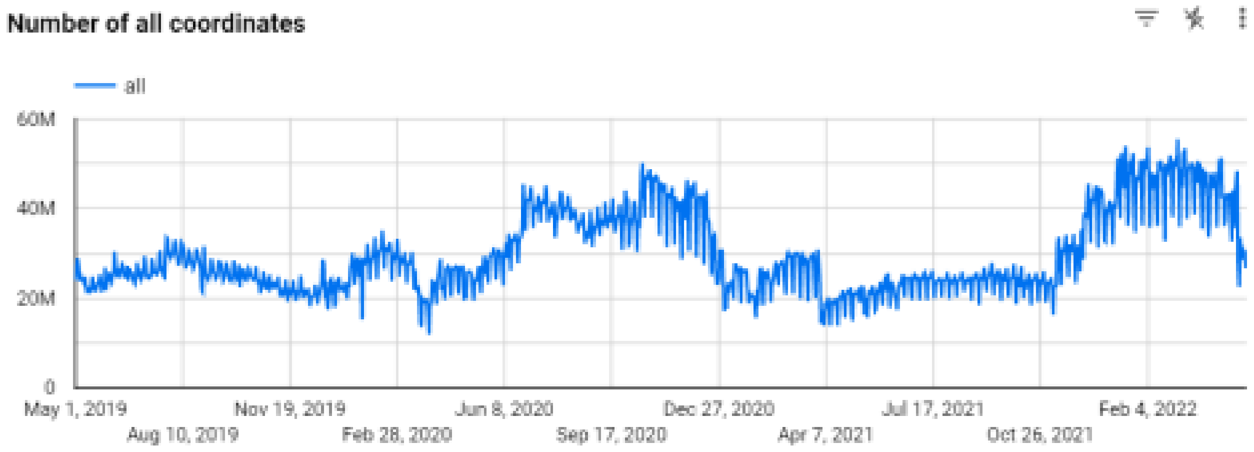
- Explore the data. Document your findings and ideas. What kind of problem is it? What test and metric would you use in order to evaluate the performance? Why?

Step 2 - Feature Engineering and Selection

- Apply whatever preprocessing steps you think would be necessary. This could include anomaly detection or normalization methods. Explain your decisions.
- Examine the features in more detail. Which are the best features? Choose a selection method and apply it. Do the results make sense? Try to explain why the features were selected.

Step 3 - Handling data inconsistencies

- A big challenge when working with GPS data are inconsistencies over time and region. This plot shows the number of all GPS coordinates in Germany over time, on which we base our estimates. This inconsistency also transfers to the number of visits, users, etc.



- Try finding solutions for this problem. For example by introducing new features based on existing ones. Also here, explain your choices or ideas.

Step 4 - Modeling

Choose one (or several) models and evaluate your results using the performance metric of your choice. Reflect on advantages and problems that might come with choosing certain models. Also reflect on overall problems during the process and with the results. Consider limitations that might arise when applying this model to non training locations all over Germany.

Bonus - Step 5 - Hyperparameter Tuning

If you still have time you can try tuning your model of choice to increase performance.

Features & Target

All features are either averaged or aggregated on a daily level

- “Location_id” : ID of the location
- “Local_date” : Date of the current row that all other features are associated with
- “Weekday” : Number of the day within the week
- “Week” : Number of the week within the year
- “Month” : Number of the month
- “Year” : Number of the year
- “Location_area” : Area of the location
- “Location_avg_user_activity” : Average number of visits per user
- “Location_avg_visit_duration” : Average visit duration
- “Location_weighted_coordinates” : Number of coordinates
- “Location_clustered_weighted_visits” : Number of visits
- “Location_clustered_weighted_users” : Number of users
- “Radius_avg_user_activity” : Average number of visits per user within radius (150m)
- “Radius_users” : Number of users within radius (150m)
- “Radius_visits” : Number of visits within radius (150m)
- “Postal_code_area” : Area of the postal code that the location is situated in
- “Postal_code_population” : Population of the postal code
- “Postal_code_users” : Number of users within postal code
- “Postal_code_visits” : Number of visits within postal code
- “City_area” : Area of the city that the location is situated in
- “City_population” : Population of the city

- “City_users” : Number of users within city
- “City_visits” : Number of visits within city
- “Region_visits” : Number of visits within region
- “State_users” : Number of users within state
- “State_visits” : Number of visits within state
- “Country_users” : Number of users within country
- “Country_visits” : Number of visits within country
- “Visits” (Target) : Real number of visits observed within location

Good Luck!