Cab spotting data challenge

Michele de Gruttola
https://www.linkedin.com/in/michele-de-gruttola-2238808b/

Data and challenge

Input: a dataset containing mobility traces of ~500 taxi cabs in San Francisco collected over ~30 days, as [latitude, longitude, occupancy, time]

Challenge:

1. To calculate the potential for a yearly reduction in CO2 emissions, caused by the taxi cabs roaming without passengers. In your calculation please assume that the taxicab fleet is changing at the rate of 15% per month (from combustion engine-powered vehicles to electric vehicles). Assume also that the average passenger vehicle emits about 404 grams of CO2 per mile.

2. To build a predictor for taxi drivers, predicting the next place a passenger will hail a cab.

3. (Bonus question) Identify clusters of taxi cabs that you find being relevant from the taxi cab company point of view.

Output: this presentation and code here https://github.com/degrutto/cabspotting

Exploratory analysis (1) : checks

Check missing values: OK

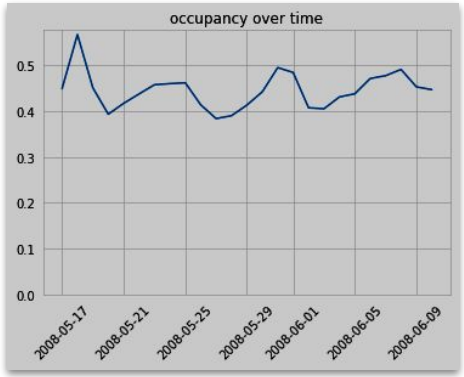
Check for suspicious outliers in the variables: OK

Plot occupancy over time and check also the time range

check we have 537 different taxi

```
df.describe(include='all')
```

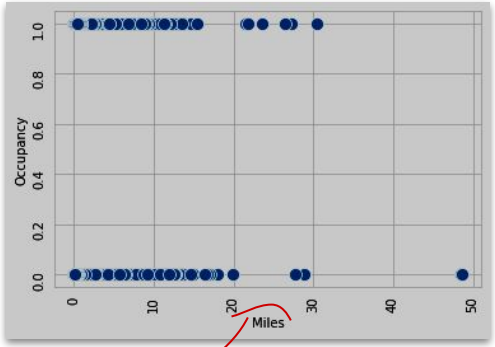
	Latitude	Longitude	Occupancy	Timestamp
mean	37.768006	-122.416266	0.479951	1.211504e+09
std	0.043735	0.030400	0.499618	2.743595e+05
min	37.513300	-122.557930	0.000000	1.211030e+09
25%	37.762060	-122.429170	0.000000	1.211250e+09
50%	37.780750	-122.415010	0.000000	1.211520e+09
75%	37.790260	-122.403770	1.000000	1.211739e+09
max	37.910430	-122.190330	1.000000	1.211960e+09



Exploratory analysis (2): preprocessing

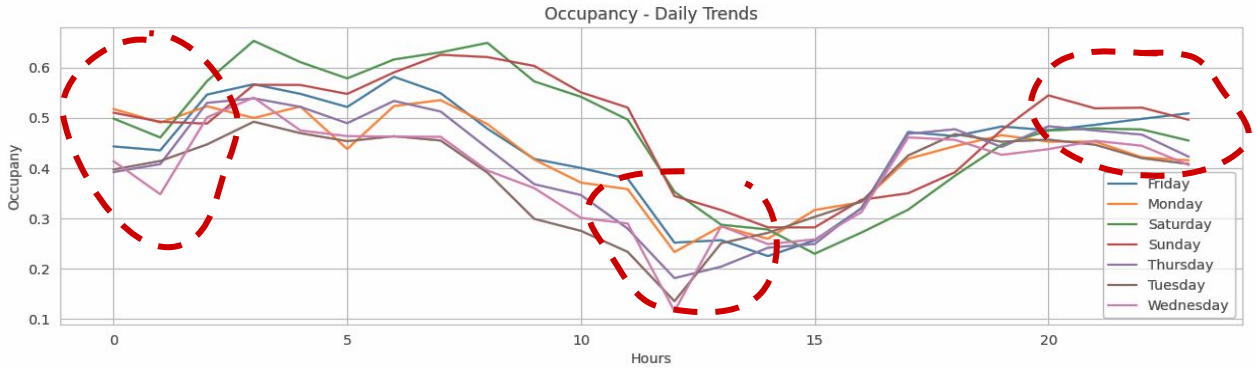
Add:

- taxi name
- previous time coordinate
- miles travelled from previous location
- day and hour from timestamp



Obs

1. At night more occupancy for weekend
2. Dips during core of the working hs
3. raise in the evening



C02 emission potential reduction

The ratio between sum of distance with vacant taxi vs distance with passengers is about 44%

- this sets the max we could in principle save
- but we have the 15% monthly reduction in favour of EV

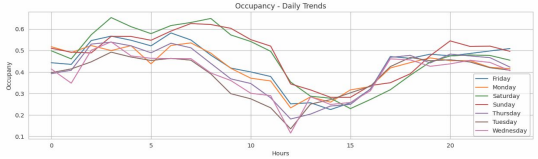
By assuming

- the year evaluation starts with this month
- we have an entire month
- each month, starting from the next, we decrease the emission by 15%

Results:

1. Co2 emission without passenger is 2921 tons !
2. total C02 emission is 13.929 tons !
3. potential C02 emission is still **21.0%**

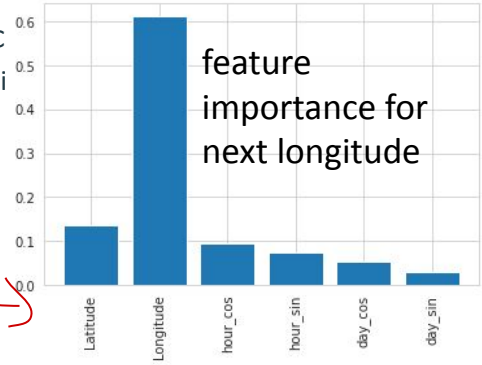
Predict where the taxi will onboard a passenger (1)



Modelling strategy:

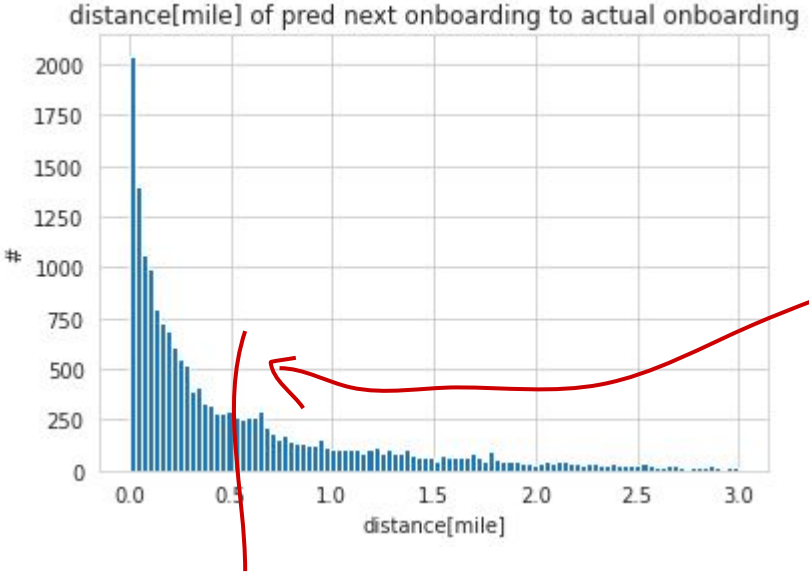
- treat the problem as a time series strategy
  - for each taxi separately, knowing (coordinates, time info) when occupancy is 0 we attempt to predict the coordinate of the next onboarding (which is our target)
  - for each taxi separately the target is the coordinate of next onboarding for the last 20% of time
  - we add hourly and day dependency as independent oscillatory waves

- Use an ensemble of Boosted Decision Tree **regressor** for the model architec
  - two separate model but same architecture for next onboarding latitude and longi
  - allows to check feature importance



Predict where the taxi will onboard a passenger (1)

- Check for each model the usual scores:
  - mean squared error (to select the best model per case)
  - Mean absolute error: mean(abs(pred -real))
  - Mean error : mean(pred - real)
- Check for the 20% of the sample we use for testing

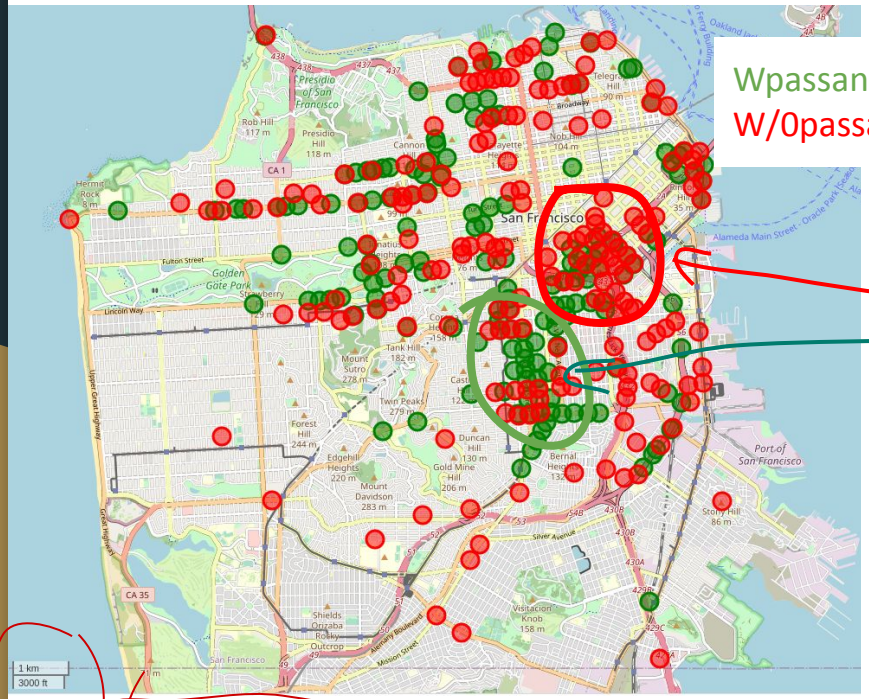


We can predict for 55% of the cases the next onboarding within half a mile

how good is that? see also next slide

Bonus 1: Cluster of taxi

Group the taxi locations dividing vacant and occupied taxi (integrating on all time)



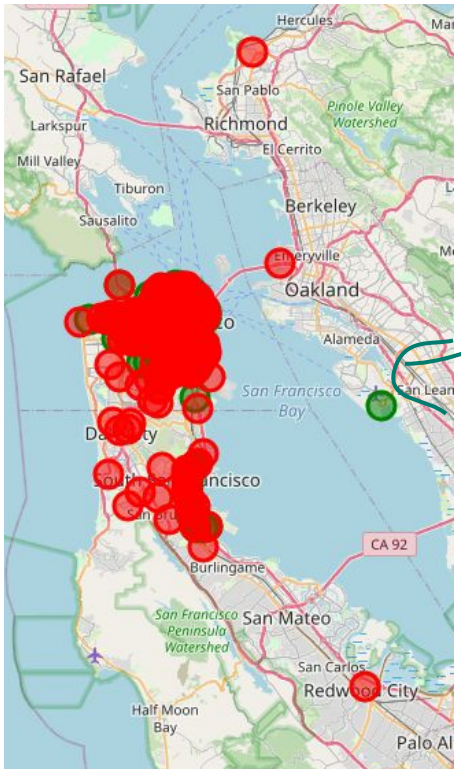
Obs (zoom on city center):

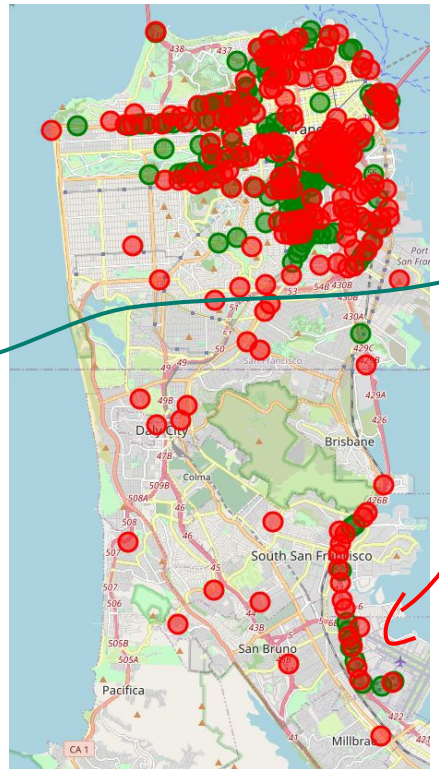
- cluster of empty taxi at S-E
- higher density for the two categories slightly shifted by about 1 mile

Bonus 1: Cluster of taxi

Wpassangers
W/0passangers

Group the taxi locations dividing vacant and occupied taxi ((integrating on all time)





Obs (unzoom):

- only 4 over 400 clusters are more isolated
- as expected high circulation in airport direction

Bonus 2: Taxi driver segmentation

Businesswise Segmentation (a la RFM Segmentation) of the taxi drivers using:

1. Miles Coverage
2. Distance covered with passengers
3. Active minute per day

Segmenting using percentiles for the three quantities in 4 segment each (less than 25 ,50,75 and more than 75 percentile).

The higher the better for all properties: assign score from 1 to 4

Results:

Best Taxi Drivers (Actively driving cab and working much more than other per day) 444: 51
Voyager (Actively driving cab, mile coverage is in the top %25): 4XX : 132
Most Occupied Taxi (mile coverage with passengers is in the top %25) X4X : 134
Most Active Tax (active time is in the top %25) XX4: 134
Fast and Vacant (Actively driving more miles but less minutes per day) 411: 0
Eco friendly/Lucky Stand-by (Driver is not driving without passenger but he/she drive more than other) 144: 0
Not Active Taxi Drivers (mile coverage, mile coverage with passanger and active minutes per day are in the bottom %25) 111: 83

BACKUP

Who am I: ask linkedin (1)

michele de gruttola
Senior Data Scientist and Machine Learning specialist

Research Fellow
CERN
Jan 2015 - Dec 2017 · 3 yrs
Geneva Area, Switzerland

>1k scientists

During my fellowship program at CERN I designed statistical (frequentist and bayesian) data analysis for the CMS experiments. I developed refined algorithm to detect unique signals, leading to the Hi …see more

Evidence for the decay of the Higgs boson to bottom quarks with CMS data
A search for the standard model (SM) Higgs boson (H) decaying to bb when produced in association with an electroweak vector boson (V) is presented for the following…

Postdoctoral Researcher
University of Florida
Jan 2011 - Nov 2014 · 3 yrs 11 mos
Gainesville, Florida Area

During my postdoctoral period I have conducting research for the Fermilab and CERN experiment leading in particular to the Higgs boson observation. I have also supervised three successful Phd students

47th Annual Fermilab Users Meeting
The annual gathering of the Fermilab Users Organization will be held on June 11-12, 2014. This meeting is an opportunity for discussion of recent physics results from the…

LPC fellow
Fermilab
2011 - 2013 · 2 yrs

Research for the Fermilab and CERN experiments.

LHC Physics Center
I will work on Higgs search, which will be a topic of great interest in 2011, when LHC integrated luminosity will start to be competitive with TeVatron's capabilities. Two Hig…

See all 8 experiences

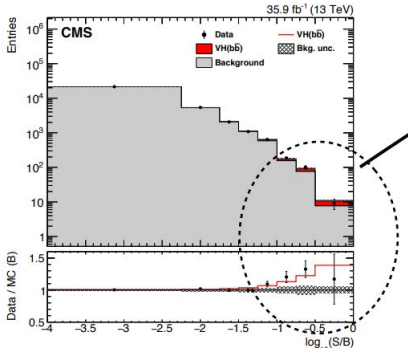
Education

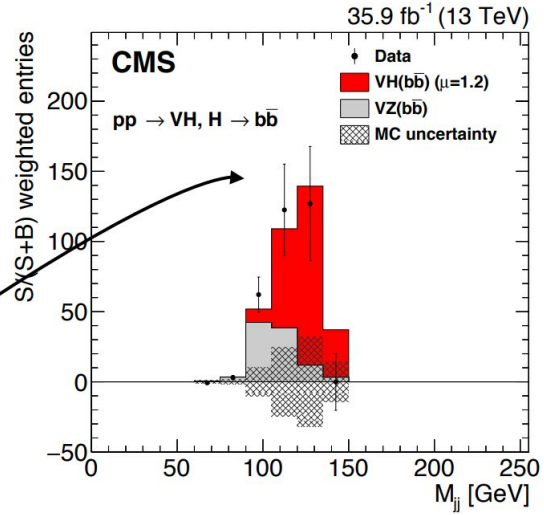
Università degli Studi di Napoli Federico II
Doctor of Philosophy - PhD, Physics
2008 - 2010

Thesis on analysis of physics collision data at CERN

Already mastering all the modern data science techniques while doing fundamental research at CERN: big data, advanced bayesian and frequentist statistics, machine learning ….

- S and B are expected signal and postfit background in the bin of the output of the BDT distributions in which each event is contained





Who am I: ask linkedin (2)

michele de gruttola
Senior Data Scientist and Machine Learning specialist

INAIT SA
3 yrs 10 mos

30 persons

- **Research Scientist (Machine Learning and Data Science)**
  Full-time
  Jan 2021 - Oct 2021 · 10 mos
  Lausanne, Vaud, Switzerland

- **Machine Learning Engineer**
  Jan 2018 - Jan 2021 · 3 yrs 1 mo
  Lausanne Area, Switzerland

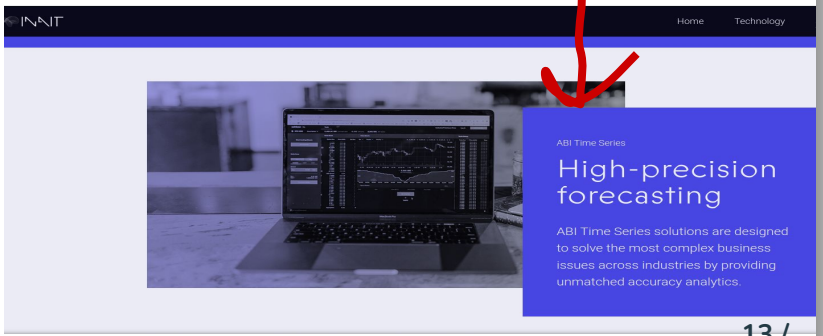
  I have designed and commissioned machine learning tools for image reconstruction, classification, time series predictions and anomalies using deep learning (auto-encoders, generative adversarial networks, (de)convolution networks, etc. etc.) working with clients to maximize their business outcome

  keywords: KPI, deep learning, computer vision, object detection, image reconstruction, decision trees

4+ years experience in industry in two different start-ups so far exploring many aspects such as algorithm prototyping, data engineering of the pipleine, product ownership andf full deployment from PoC, PoV to mockup to first product version

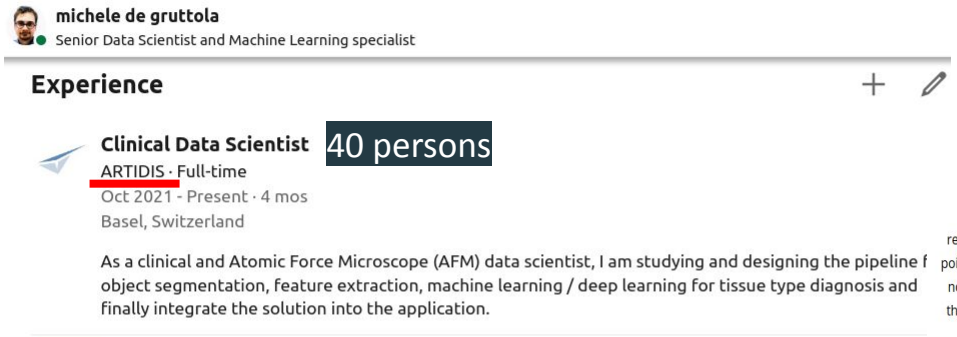
Fields:

1. Computer Vision with and w/o deep learning
2. Natural language processing
3. Text and image compression using deep learning
4. Time series predictions



Who am I: ask linkedin (3)

Computer Vision with and w/o deep learning



michele de gruttola
Senior Data Scientist and Machine Learning specialist

Experience

Clinical Data Scientist 40 persons
ARTIDIS · Full-time
Oct 2021 - Present · 4 mos
Basel, Switzerland

As a clinical and Atomic Force Microscope (AFM) data scientist, I am studying and designing the pipeline f object segmentation, feature extraction, machine learning / deep learning for tissue type diagnosis and finally integrate the solution into the application.

We are committed to improving the quality of life for patients while protecting their personal data.

The individual patient history on our portable platform forms the basis for holistic, comprehensive and communal solutions that promote long-term relationships with each individual. ARTIDISNET is an AI-powered algorithm integrating patient nanomechanical biomarkers and hundreds of relevant data points from the patient's own clinical data to provide the patient with personalized high-quality treatment. Together with clinical data like tumor size, lymph nodule status, etc. we integrate the nanomechanical biomarker data and combine them with imaging, genetic, and histopathology parameters, as well as therapy follow-up data to generate medical insights and support treatment decision-making. ARTIDISNET enables the clinician to provide the right patient with the right treatment and to maximize the positive treatment results.

What I am good at

I have demonstrated full coverage of all aspects of a AI/data science product development:

1. problem understanding after interaction with clients
2. **algorithm** development
   a. prototyping
   b. developing a data pipeline for the production (Azure MLOps, exoscale, python, in Jenkins/git)
   c. testing (CI/CD)
3. Product **management**
   a. Product ownership
   b. Scrum master using Jira
   c. interaction with devops, frontend developer, mobile app engineer
4. Business discussion of PoC , PoV, mVP, Mockup, …
5. Product prototype with solutions shown to **client**