ACCURAT

Places

Analytics proposal

A C C U R A T
/ | | | | / |

Index

- Data preprocessing

- Optimisation and automation ([1], [2])

- Lime ([3]

- Importance matrix

- Forecasting

- t-SNE ([4])

A C C U R A T
/ | | | | / |

Data Processing

Issues:

- Categorical or continues variable
- String variables
- “Target” variables
- Polishing and regex
- Scaling: when to scale? By algorithm or by feature type. Good question. Interpretability and backwards-scaling

Approach:

Identify some category of variables and apply specific regex

E.g.:

- Dates to Unix
- One-hot encoding
- Y and X identification (with aid)

ACCURAT

Processed data with the Places mock dataset

IT-2016-152156	2017-07-18	2017-08-05	Second Class	CG-12520	Consumer	Italy	FUR-BO-10001798	Furniture	...	261.9600	2	0.00	41.9136	Milan
IT-2016-138688	2017-06-26	2017-07-10	Second Class	CG-12520	Consumer	Italy	FUR-CH-10000454	Furniture	...	731.9400	3	0.00	219.5820	Naples
FR-2015-108966	2016-12-26	2017-01-13	Second Class	DV-13045	Corporate	France	OFF-LA-10000240	Office Supplies	...	14.6200	2	0.00	6.8714	Toulouse
FR-2015-108966	2017-10-30	2017-11-14	Standard Class	SO-20335	Consumer	France	FUR-TA-10000577	Furniture	...	957.5775	5	0.45	-383.0310	Marseille
IT-2014-115812	2016-11-15	2016-11-21	Standard Class	SO-20335	Consumer	Italy	OFF-ST-10000760	Office Supplies	...	22.3680	2	0.20	2.5164	Naples

	Order Date	Ship Mode	Segment	Country	Quantity	Discount	Profit	Latitude	Longitude	delay	Furniture	Office Supplies	North	Center	South
0	1.500329e+09	0	1	1	2	0.00	41.9136	45.466797	9.190498	1555200.0	1	0	1	0	0
1	1.498428e+09	0	1	1	3	0.00	219.5820	40.835934	14.248783	1209600.0	1	0	0	0	1
2	1.482707e+09	0	0	0	2	0.00	6.8714	43.604462	1.444247	1555200.0	0	1	0	0	0
3	1.509318e+09	0	1	0	5	0.45	-383.0310	43.296174	5.369953	1296000.0	1	0	0	0	0
4	1.479164e+09	0	1	1	2	0.20	2.5164	40.835934	14.248783	518400.0	0	1	0	0	1

A C C U R A T

Optimisation and automation

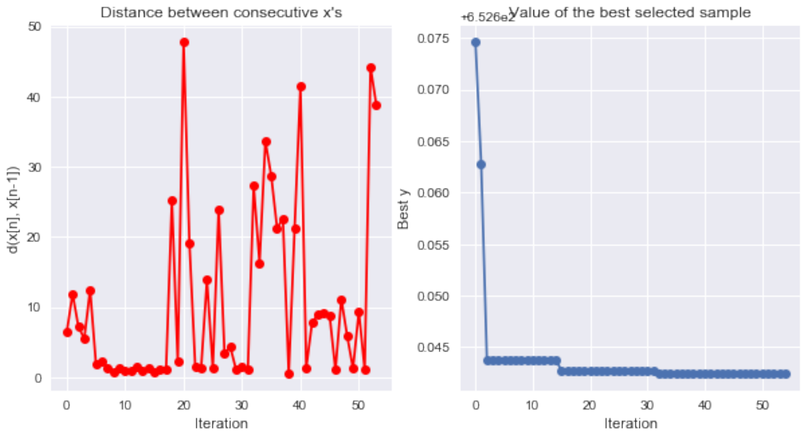
Issues:

- Auto ML problem
- Different solutions for different problems

Approach:

Identify some macro-areas and build solutions for those, then use auto-optimisation to fit the single cases. Technically speaking:

- Search grid
- Evolutionary algorithm
- GPyOpt

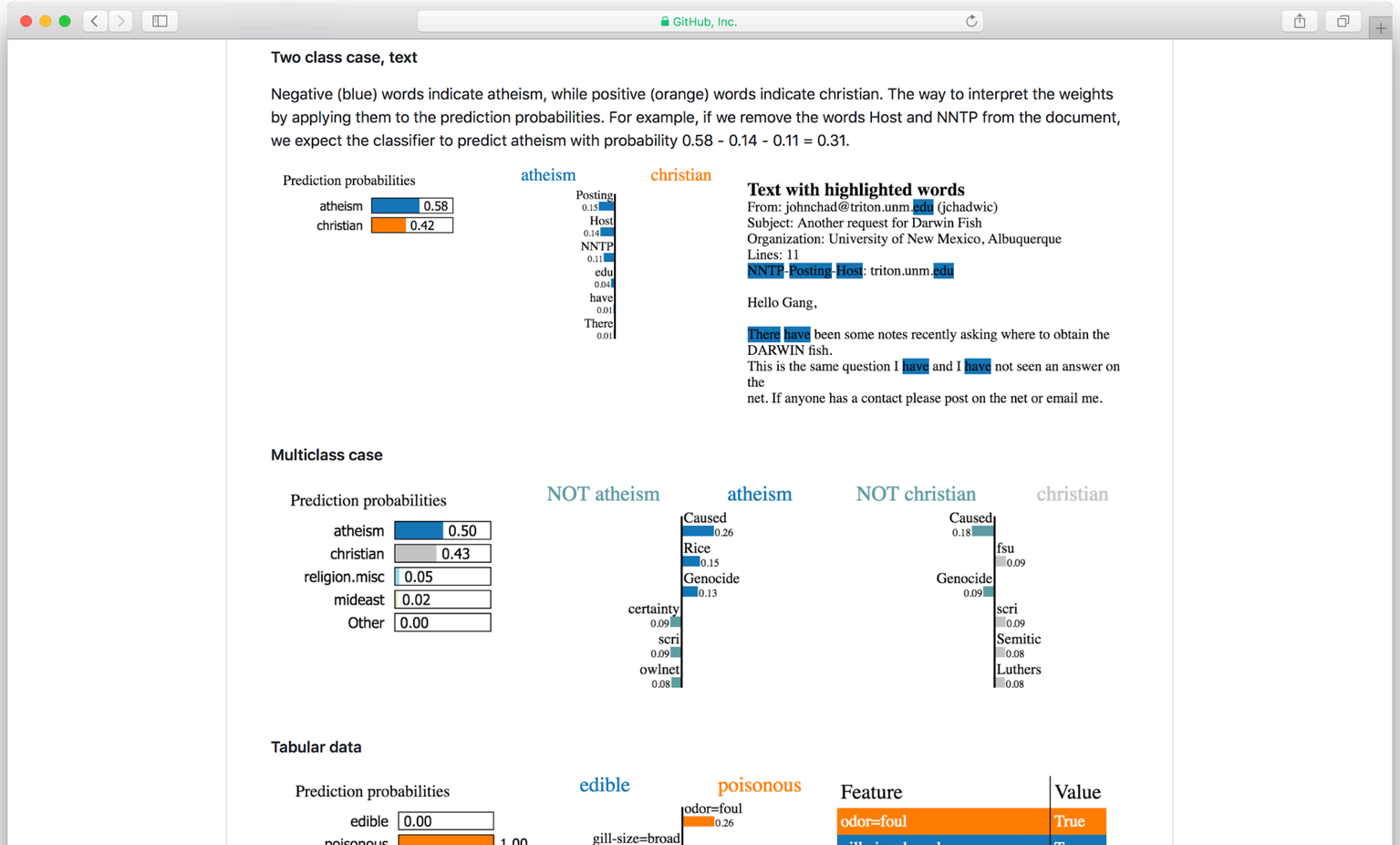


ACCURAT

Lime

Quoting the github page: *“This project is about explaining what machine learning classifiers (or models) are doing. At the moment, we support explaining individual predictions for text classifiers or classifiers that act on tables (numpy arrays of numerical or categorical data) or images, with a package called lime (short for local interpretable model-agnostic explanations).”*

https://github.com/marcotcr/lime



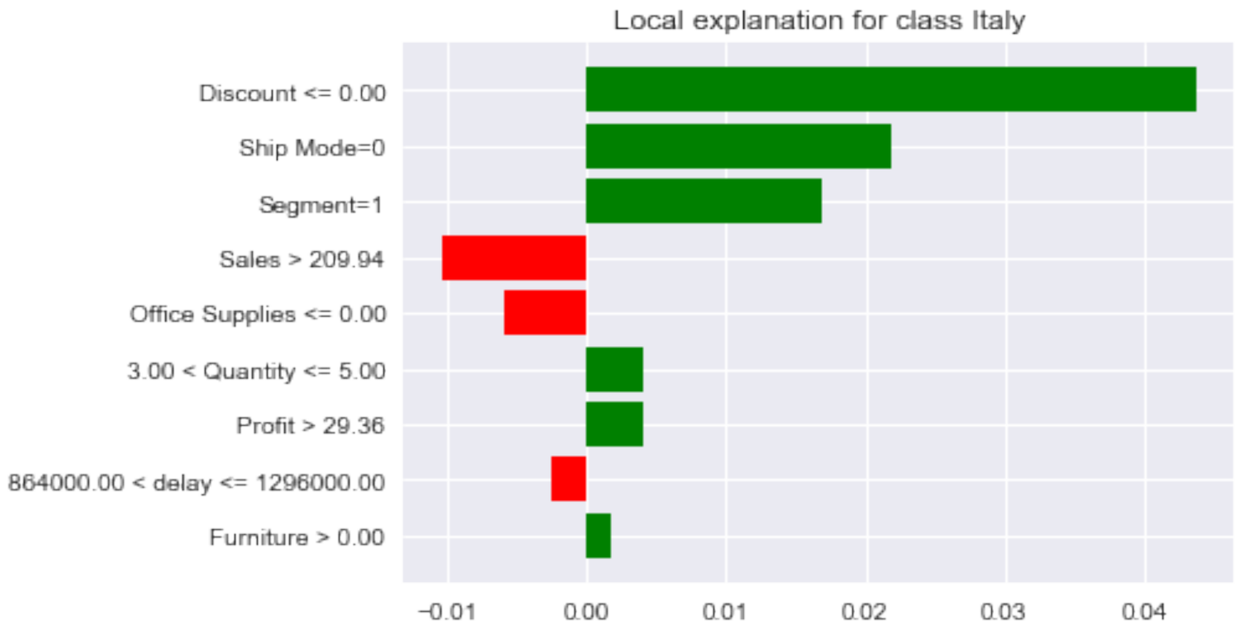
A C C U R A T

Lime, explain analytics and ML intuitively

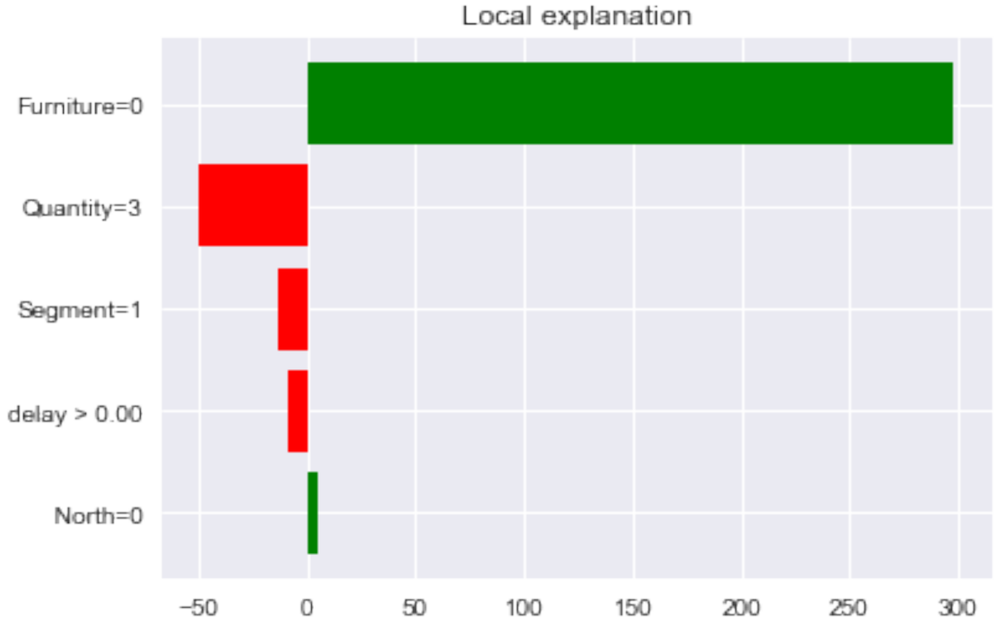
Issues:

- How to visualise analytical tools and what they do

Classification, classifying Italian buyers and sellers



Regression, explaining amount of "Sales"



A C C U R A T

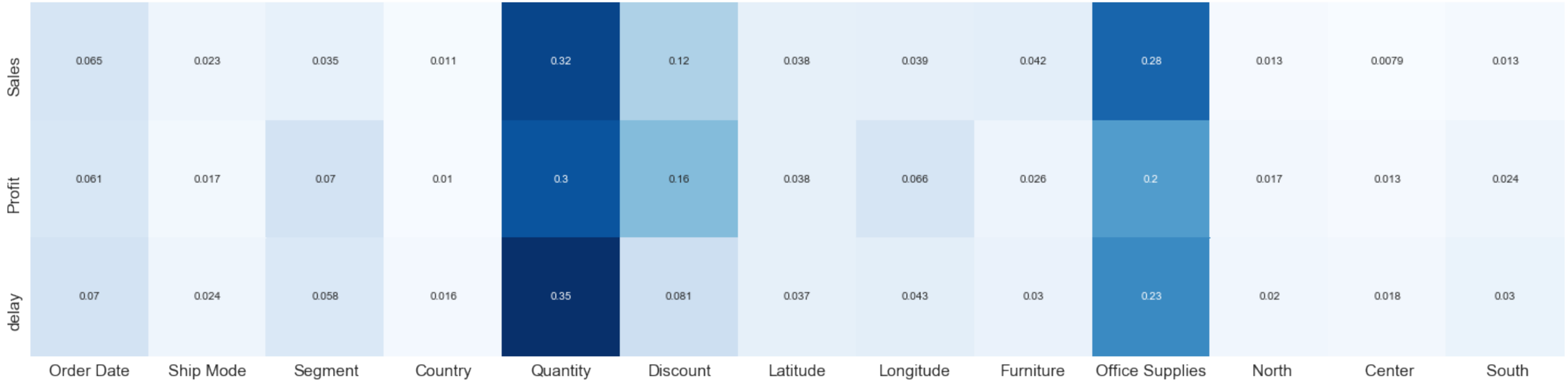
Importance Matrix

Issues:

- Visualising and computing relations from explanatory variables to target variables (e.g. from

Approach:

- Decision tree per target variables (*profit, delay, sales)* with explanatory variables
- Compute feature importance, plot it in the a pseudo-correlation matrix (sum of the row is 1)



A C C U R A T

Forecasting

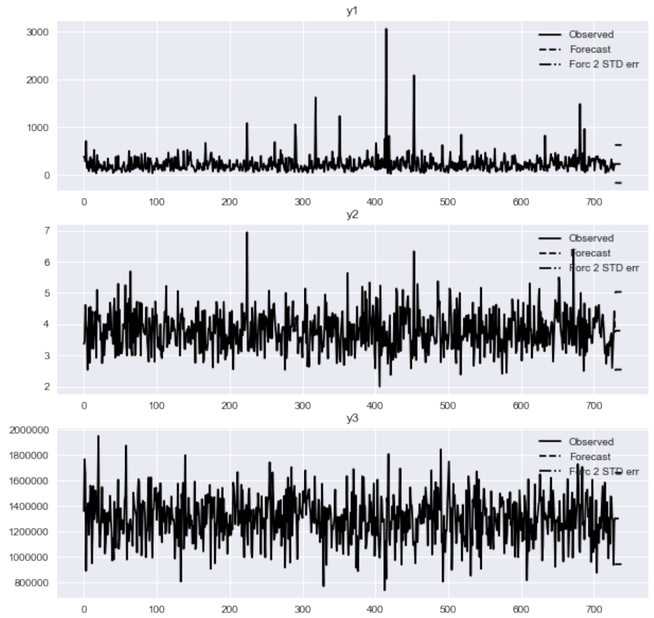
Order Date	Sales	Discount	Profit	delay	Quantity	Ship Mode	Segment	Country	Furniture	Office Supplies	North	Center	South	count
2016-01-02	377.852421	0.252632	-219.590716	0.512949	3.368421	3	7	16	6	10	5	6	5	19
2016-01-03	382.884889	0.255556	88.011978	0.849206	3.444444	2	3	8	2	6	6	0	2	9
2016-01-04	282.656154	0.176923	35.852985	0.751526	4.615385	2	6	9	2	7	4	2	3	13
2016-01-05	708.228708	0.163077	49.023400	0.125153	3.923077	3	8	13	5	7	8	2	3	13
2016-01-06	159.405875	0.200000	33.099137	0.464782	3.812500	3	6	14	4	5	5	3	6	16

Issues:

- Only aggregated data
- Various techniques (I have used VAR and LSTM but usually simpler models are better like ARIMA, or Linear Regression. We can give different methods for more experienced users.
- Optimisation, what makes sense to forecast? Are the models working? (This can be pseudo automatised)

Approach

- General model
- Forecast in the bubble view (only aggregates like daily or monthly)
- Going from this, to something edible:

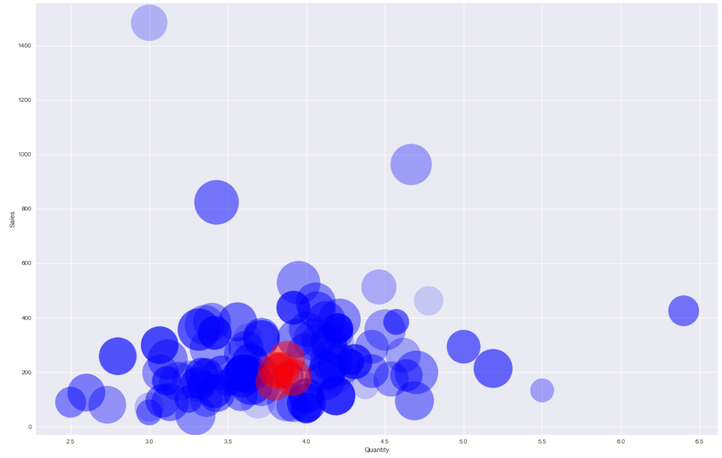


A C C U R A T

Forecasting with bubbles

- **Bubble chart:**
Automatic forecasting (in this daily) for the average values and then plots on a the aggregate granularity

- **Issue:**
A lot of data can lead to confusion and how to display uncertainty? Not my problem I output csv



A C C U R A T

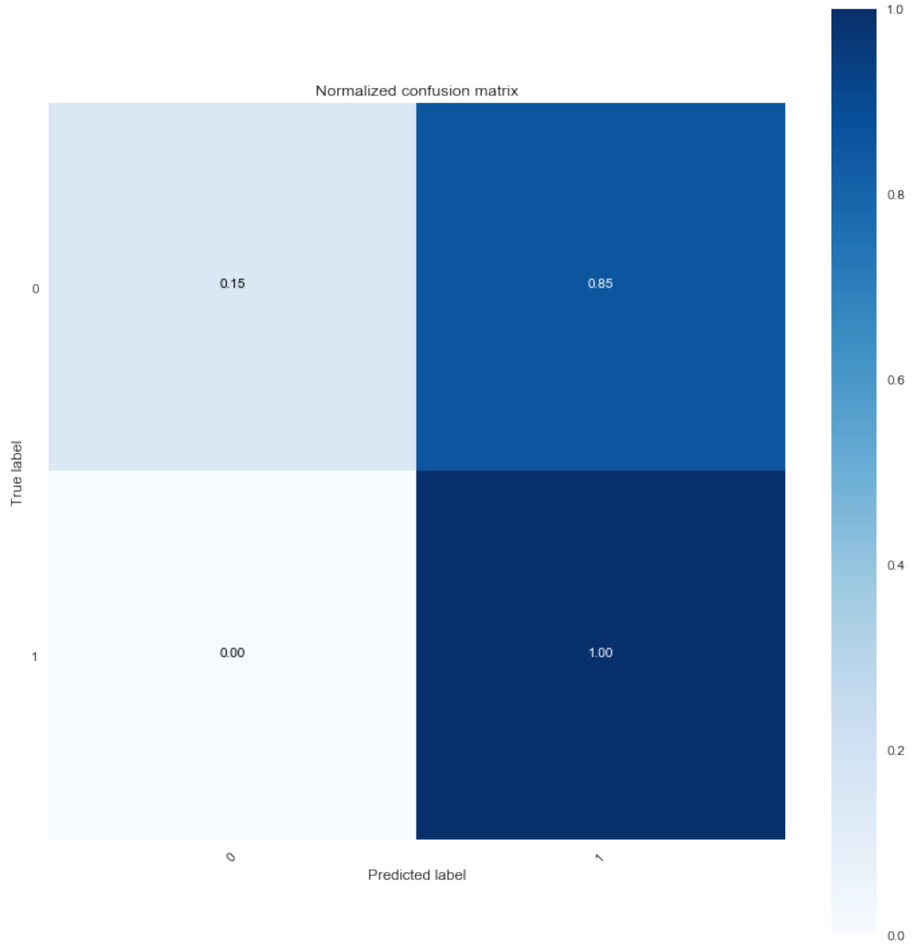
Forecasting with probabilities

- **Advantages:**
Compute a model to output probability (e.g. on Italy and France) to determine “ideal profiles” or on relevant variables to understand new data coming in. Can be done one and quickly (or slowly if you want better results)

- **Issue:**
Embedding this visually in a “story”

Ship Mode	Segment	Sales	Quantity	Discount	Profit	delay	Furniture	Office Supplies
0	1	261.9600	2	0.00	41.9136	1555200.0	1	0
0	1	731.9400	3	0.00	219.5820	1209600.0	1	0
0	0	14.6200	2	0.00	6.8714	1555200.0	0	1
0	1	957.5775	5	0.45	-383.0310	1296000.0	1	0
0	1	22.3680	2	0.20	2.5164	518400.0	0	1

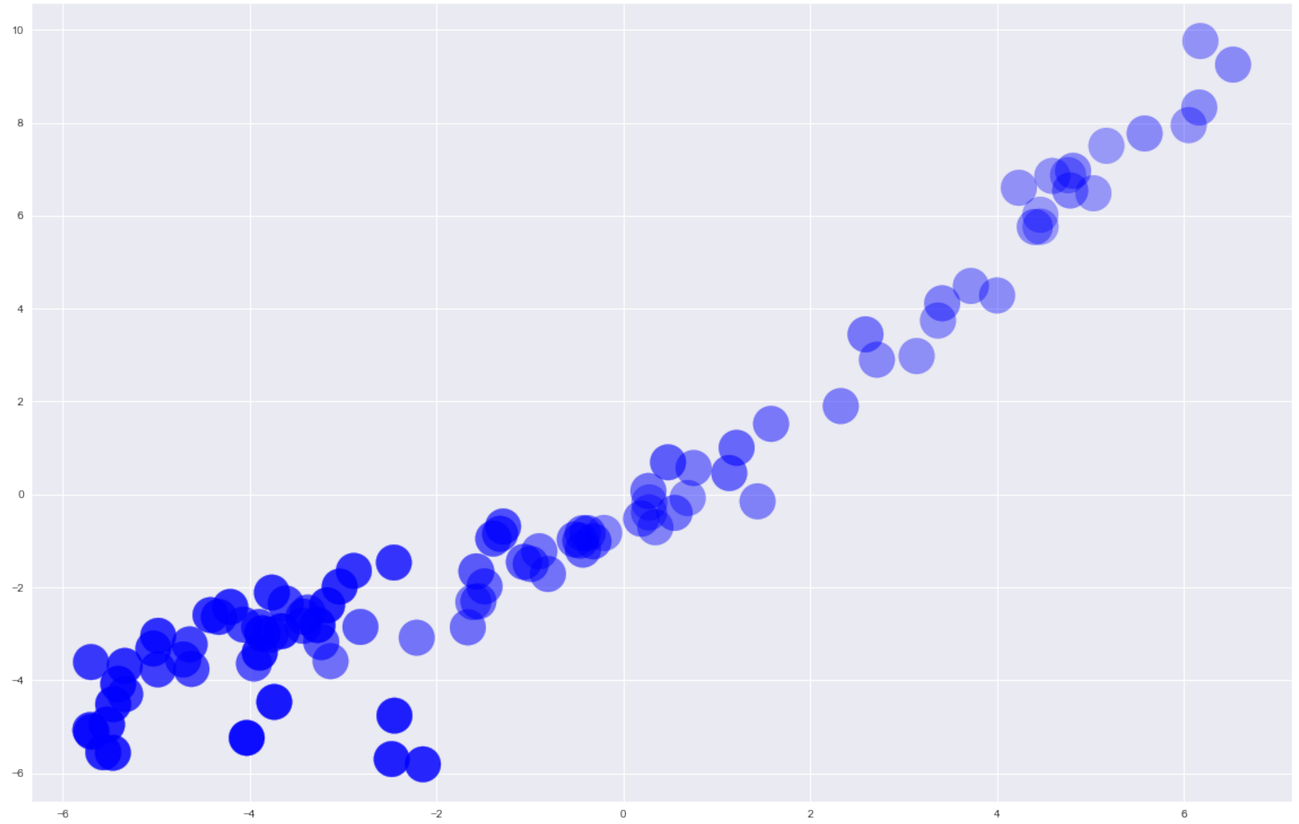
France	Italy
0.108486	0.891514
0.098739	0.901261
0.163130	0.836870
0.657867	0.342133
0.151367	0.848633



1 is Italy, 2 is France

A C C U R A T

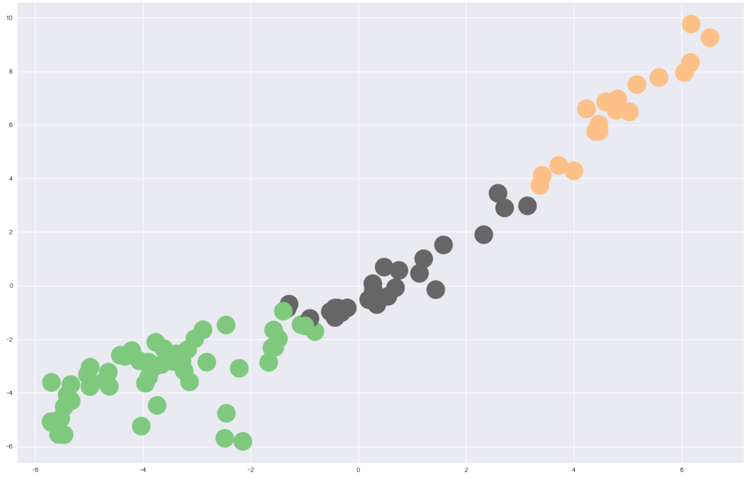
t-SNE for bubble chart



- **Experimental:**
Use of dimensionality reduction to make data “cluster itself” from the bubble chart in order to aid user selection. We can give different methods for more experienced users, that is easier.

- **Issues:**
Time costly and hyper parameters

- **Potential add-ons:**
Clustering methods, DBSCAN or K-Means etc., like in the example below



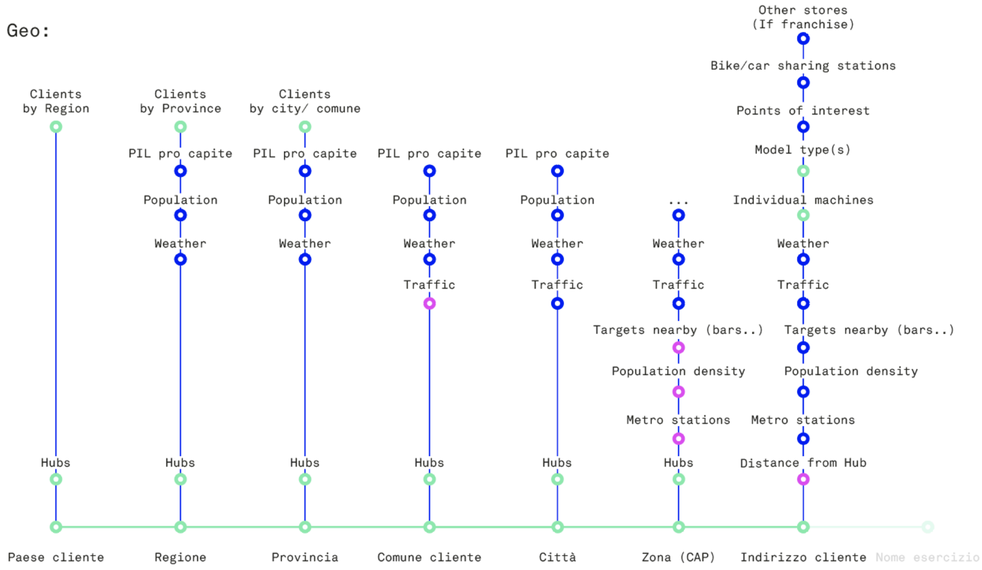
A C C U R A T

Other classical techniques

- Linear Regression and logistic regression [5]
- Decision tree plot [6]
- No average nor median nor frequency except in context

From last time, external datasets (external_data html.zip in slack/places)

- Eurostat
- World Bank
- Cities
- Weather
- Traffic



A C C U R A T
/ | | | | / |

Looking forward in analytics

- Work on automatising and putting on server (Python flask)
- Defining data
- Formalise analysis in notebook.html
- Try different datasets
- Create mock datasets
- We need loads of data for this so we need to put boundaries

Extra references:

http://www.bzarg.com/p/how-a-kalman-filter-works-in-pictures/
https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers
https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf
https://github.com/aloctavodia/Statistical-Rethinking-with-Python-and-PyMC3