Exploratory data analysis

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA

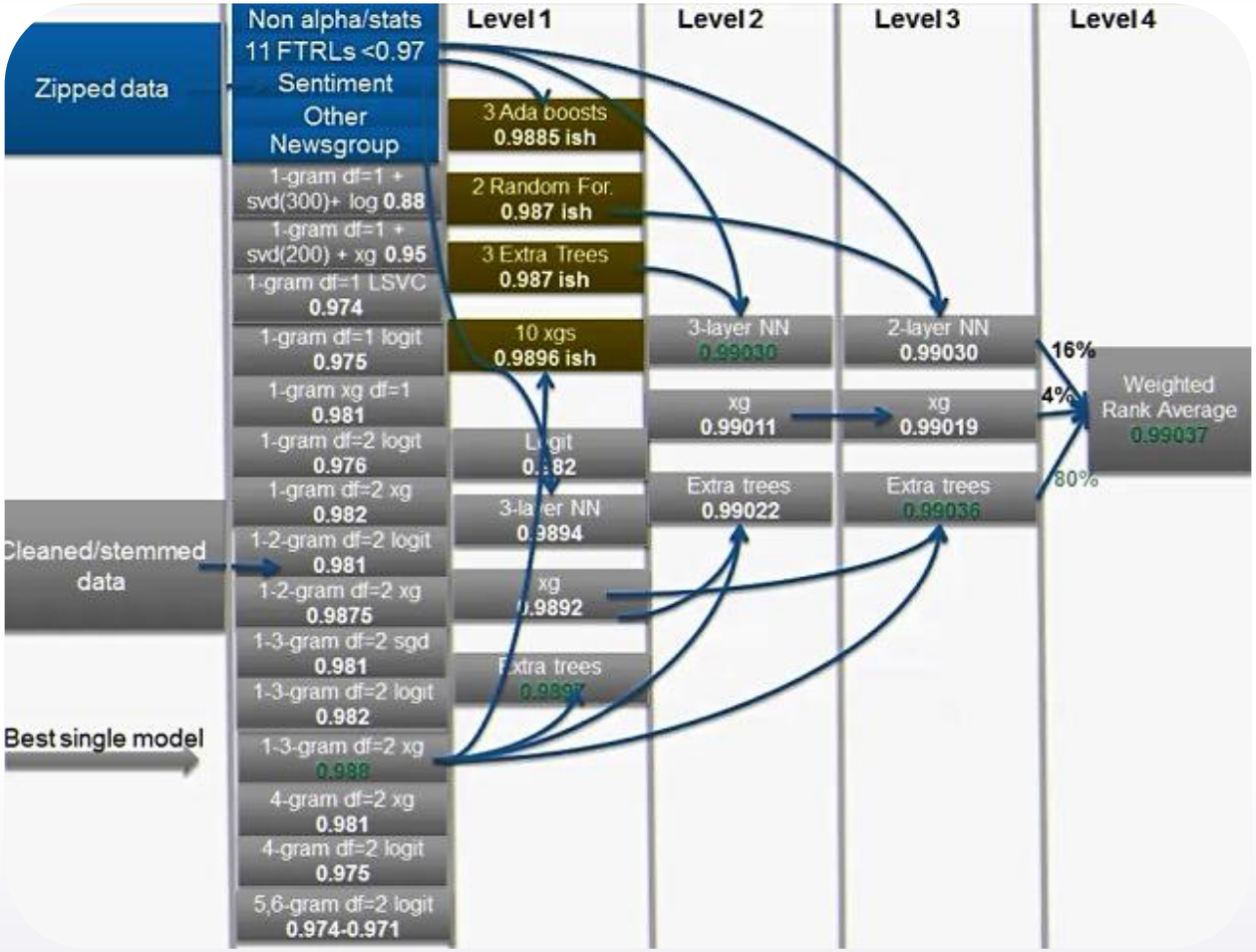
Exploratory Data Analysis (EDA)

EDA allows to:

- Better understand the data
- Build an intuition about the data
- Generate hypothesizes
- Find insights

Exploratory Data Analysis (EDA)

- Please, do not start with stacking...



Kaggle Team, http://blog.kaggle.com/2015/12/03/dato-winners-interview-1st-place-mad-professors/

Visualizations

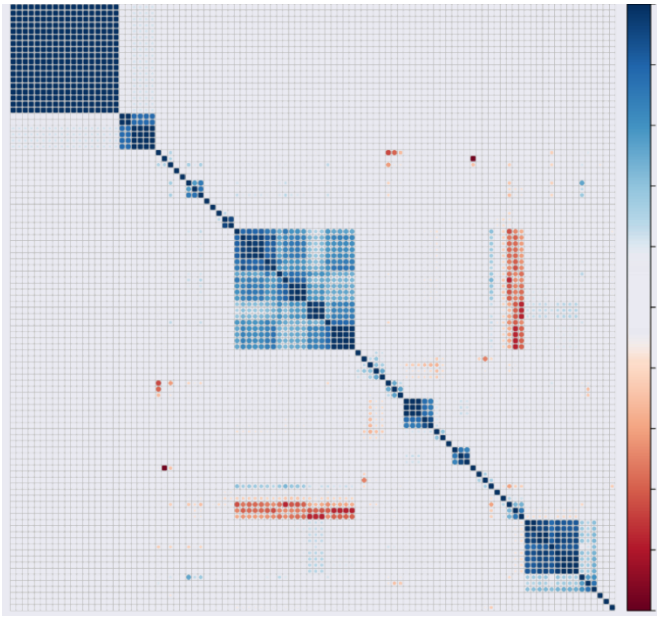
One of the main EDA tools is Visualization. When we visualize the data, we immediately see the patterns.

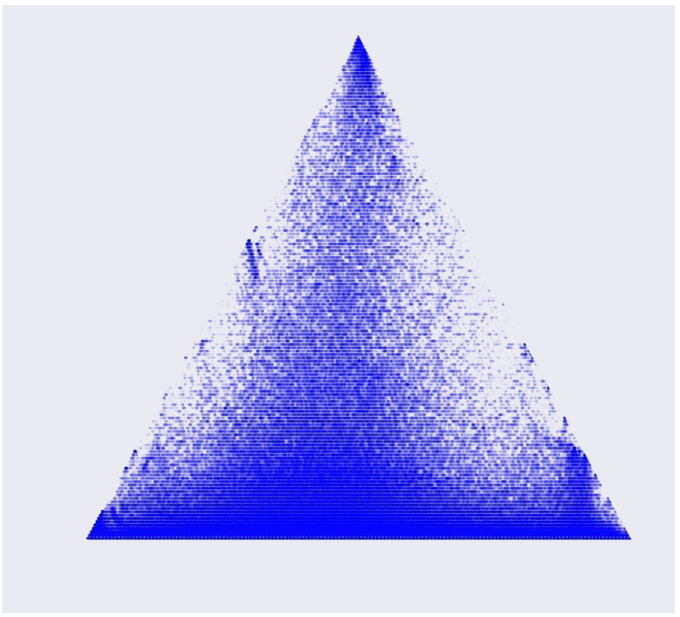
Visualization ⟶ Idea

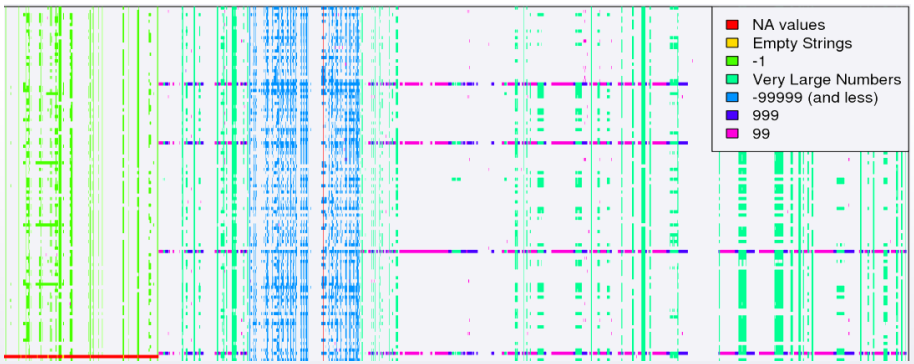
Patterns lead to questions

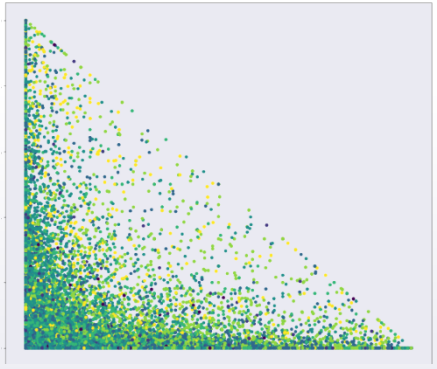
Idea ⟶ Visualization

Hypothesis testing









Motivating example

Alexander D'yakonov

Moscow, Russian Federation
Joined 7 years ago · last seen 21 days ago
http://alexanderdyakonov.narod.ru/english.htm

Followers 2

Competitions Grandmaster

Home | Competitions (36) | Kernels (1) | Discussion (104) | Followers (2) | Contact User | Follow User

Competitions Grandmaster

Current Rank	Highest Rank
199 of 60,591	1

9	14	4

Greek Media Monitoring M... · 3 years ago · Top 1%	1st of 120
dunnhumby's Shopper Cha... · 6 years ago · Top 1%	1st of 277
Large Scale Hierarchical Te... · 3 years ago · Top 2%	2nd of 119

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Contributor

Unranked

2	7	27

Code sharing · 3 years ago	21 votes
Thanks · 6 years ago	14 votes
congrats to the winners! · 2 years ago	10 votes

Motivating example

person id	person info	promo info	# promos sent	# promos used	*used this promo?*
14	...	...	13	4	**1**
3	...	...	43	35	**0**
0	...	..	6	0	**1**
32	...	...	15	13	**1**

Motivating example

id	...	# promos sent	# promos used	*diff*	*used this promo?*
13	...	0	0	1	**1**
13	...	1	1	0	**0**
13	...	2	1	1	**0**
13	...	4	2	1	**1**
13	...	5	3	1	**1**
13	...	6	3	NaN	**0**

1. For each person sort by '**# promos sent**'
2. Look at difference between consecutive rows in '**# promos used**' column ('***diff***' feature)

Conclusion

With EDA we can:

- get comfortable with the data
- find *magic features*

Do EDA first. Do not immediately dig into modelling.

In the following videos

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA