Exploratory Data Analysis

Daniel García Hernández

Master Big Data

March 2, 2020

: Overview

Why to learn EDA: Difference between success and failure

- Good EDA can be the difference between a good model and a great model
- By adressing all the available tools and the domain knowledge you will the get the most out of the data

Why to learn EDA: What's EDA?

NIST defines EDA as the *approach/philosophy for data analysis that employs a variety of techniques to:*[1]

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings

[1] https://medium.com/swlh/eda-exploratory-data-analysis-e0f453d97894

Variables and Feature engineering: Creating new information

- Converting from one type to another i.e. string to date
- Categorical to numeric and viceversa: `pd.cut`
- Using your domain knowledge to increase the information
- Statistical properties of the variables
- Use lags, rolling windows, etc
- Use accumulated values
- ...

It's the secret of the trade: the more you do it, the more tricks you learn and apply in your EDA process!

Dealing with missing data: NaN, NULL, NaT, ...

`pandas`, as usual, comes to our rescue when we have to deal with missing data:

- `pd.isnull(df)`: returns `True/False` for values in each column
- `df.fillna(value)`: fills each `NaN` with `value`
  - fill with average, median (ML)
  - fill with zero
  - fill with min/max
  - ...
- `df.dropna(axis)`: eliminates the rows/columns with a `NaN`

Data analysis and EDA: take your time!

When using all the previous steps, plus a correct visualization, we can get insights from the original data that was not there.

It takes time to create you own process, and there's plenty of literature about how to proceed. My recommendation: follow one, and with experience start modifying your process!

Data analysis and EDA: Process

- Preview data: `head, tail`
- General information: `describe(), sns.pairplot(df)`
- Check for NaN `df.isnull, df.dropna`
- Apply your stats knwoledge: outliers, distributions
- Create new information: out of your know how, out of datetimes, numeric to categorical, categorical to numeric, convert distributions
- Analyze relations with time, relations with categories, relations between features
- Create visualizations that reflect your analysis and insights

The End