End to end competitive data science

DataWeek 2020

About myself

I am an AI researcher with specialization in computer vision. I am also the current community manager at Data Science Nigeria.

I love building technologies. You can connect with me on any of these platforms.

Twitter: @elishatofunmi

Github: @elishatofunmi

Medium: @elishatofunmi

Content

We will be looking at the following:

1. About ML and DS competitions.
2. Getting started.
    a. Data understanding.
    b. Data visualization/ relation.
    c. Data preprocessing.
    d. Encoding and transformation.

Content

3. Modelling

    a. Simple modelling.
    b. Ensembles/robust models.
    c. Hyperparameter tuning.

4. Re-evaluation/re-modelling

6. Conclusion.

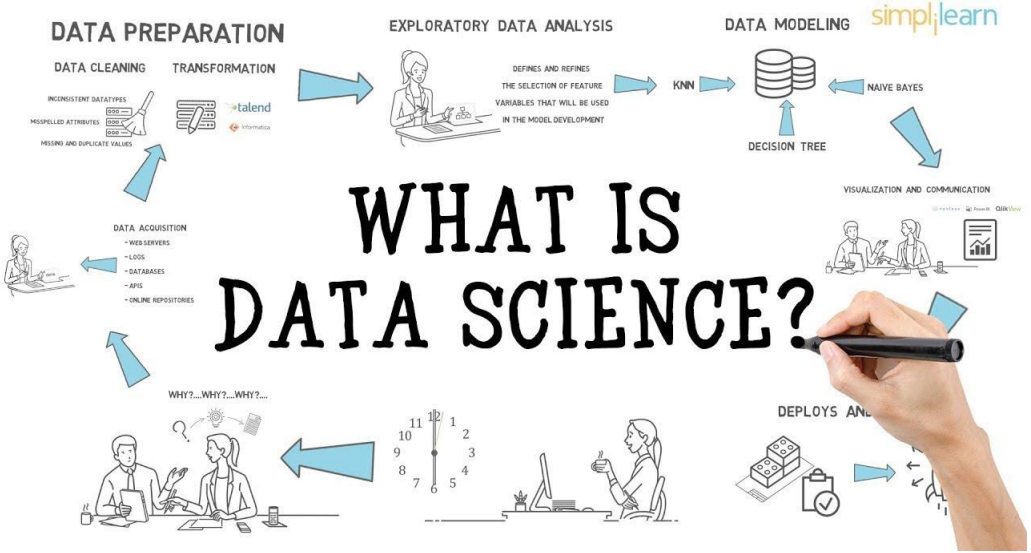
1. About ML and DS competitions

1. About ML and DS Competition

What is data science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

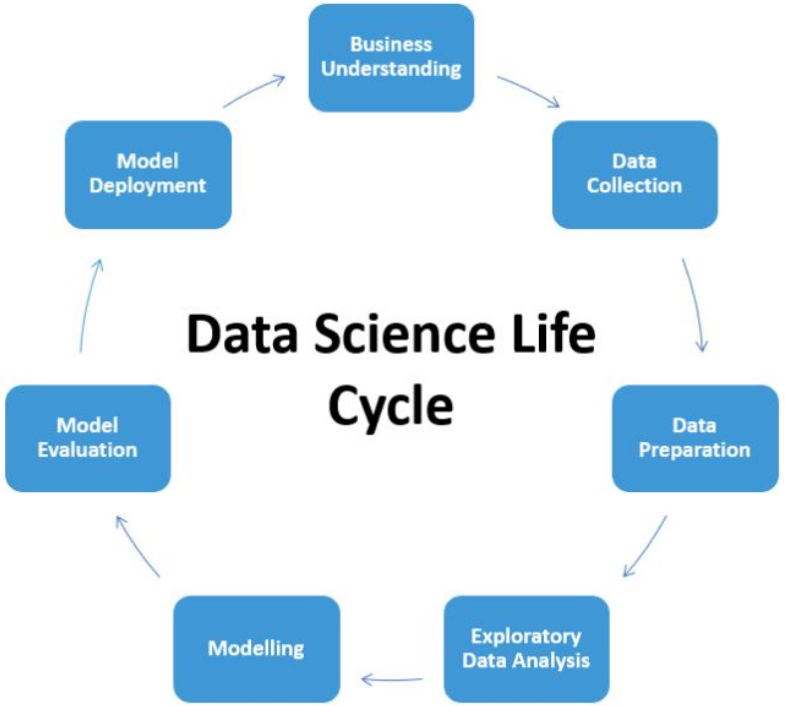
Data Science deals with the scientific understanding of the data itself.



1. About ML and DS Competition

Data Science Life Cycle

- Business understanding
- Data Collection
- Data preparation
- Exploratory data analysis
- Modelling
- Model evaluation
- Model deployment



2. Getting Started - Data Understanding

2. Getting Started - Data Understanding

Tips to understanding your data

- Analyze the data description.
- Create an atmosphere of how the data was collected.
- Understand your feature parameters.
- Visualize the data and view the relationship.
- Know the metric you are working with.

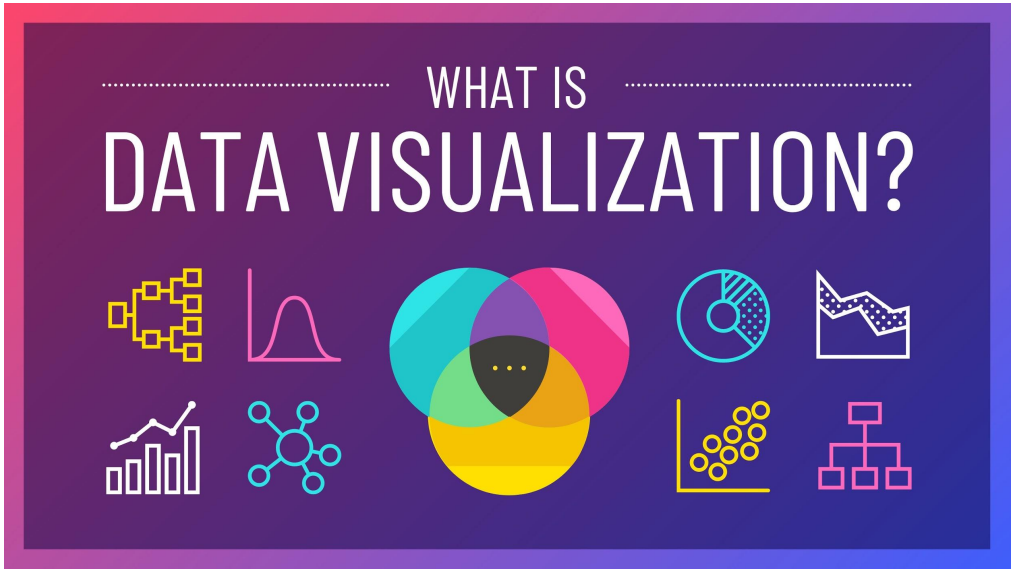
2. Getting Started - Data Visualization/Relation

2. Getting Started - Data Visualization/Relation

Definition

Data Visualization is a storytelling approach to understanding your data.

We can basically say, this is concretely insight based on understanding what your data is about.

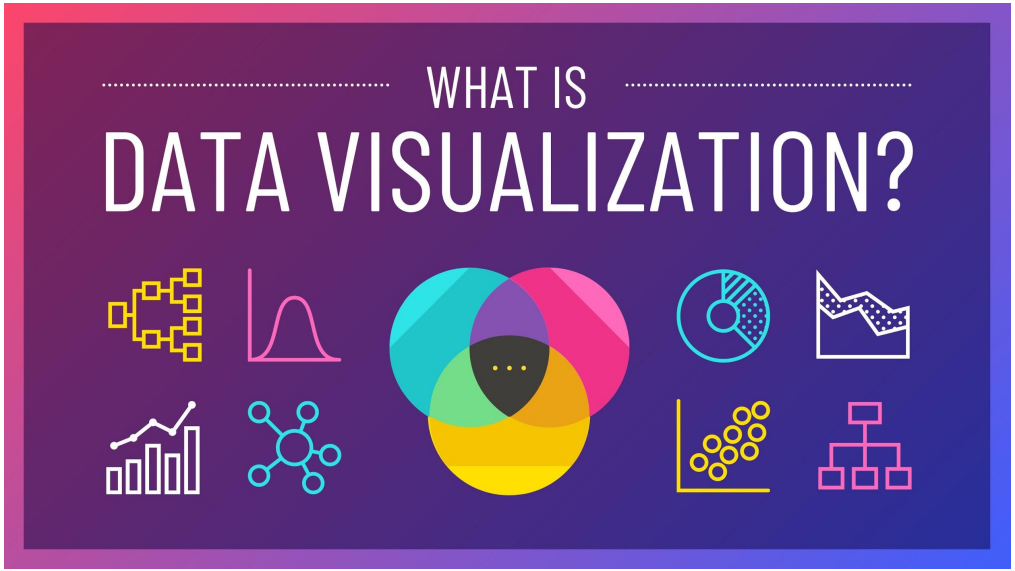


2. Getting Started - Data Visualization/Relation

Definition

To do this you communicate data understanding and insight with the use of visuals like:

- Bar charts.
- Histograms.
- Heat maps.
- Pie charts.
- Tree diagrams.
- Box plots.
- Line/rel plots. etc.

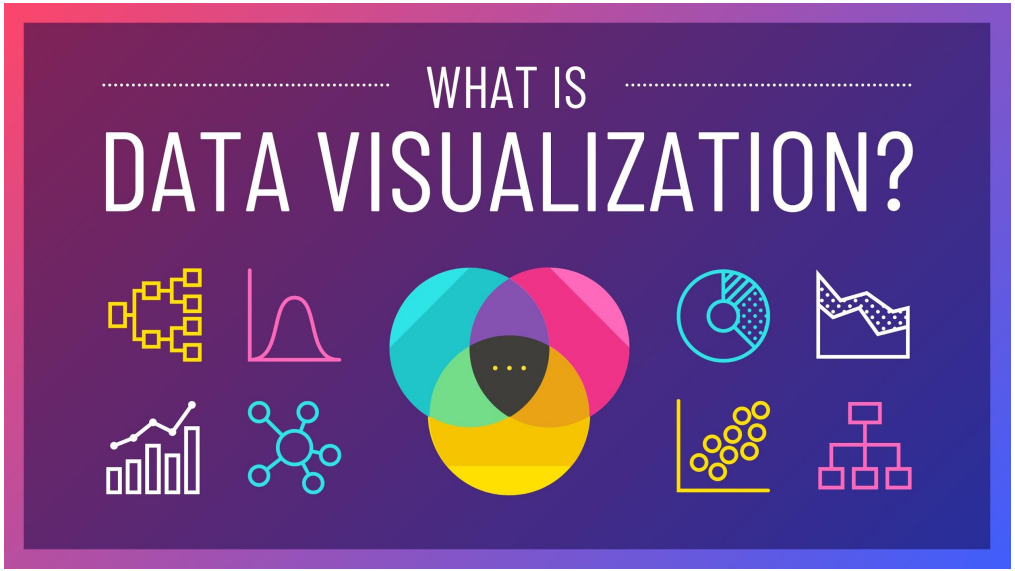


2. Getting Started - Data Visualization/Relation

Tips to Data Visualization/Relation

- Understand how to work with state of the art tools like:
  - Plotly
  - Seaborn
  - Pylab
  - Matplotlib e.t.c.
- Professional tools.
  - Tableau.
  - Power BI.

This gives you insight on what to do, how to approach your problem.



2. Getting Started - Data Preprocessing

2. Getting Started - Data Preprocessing

Data preprocessing is a stage of data refinement that deals with data cleaning, analytics (gaining insight into your data) and drawing out conclusion.

This process takes place before the actual processing/ data transformation.

2. Getting Started - Data Preprocessing

This include the following:

- Reading in the file.
- Viewing the columns/feature entries.
- Descriptive statistics(.describe, .info).

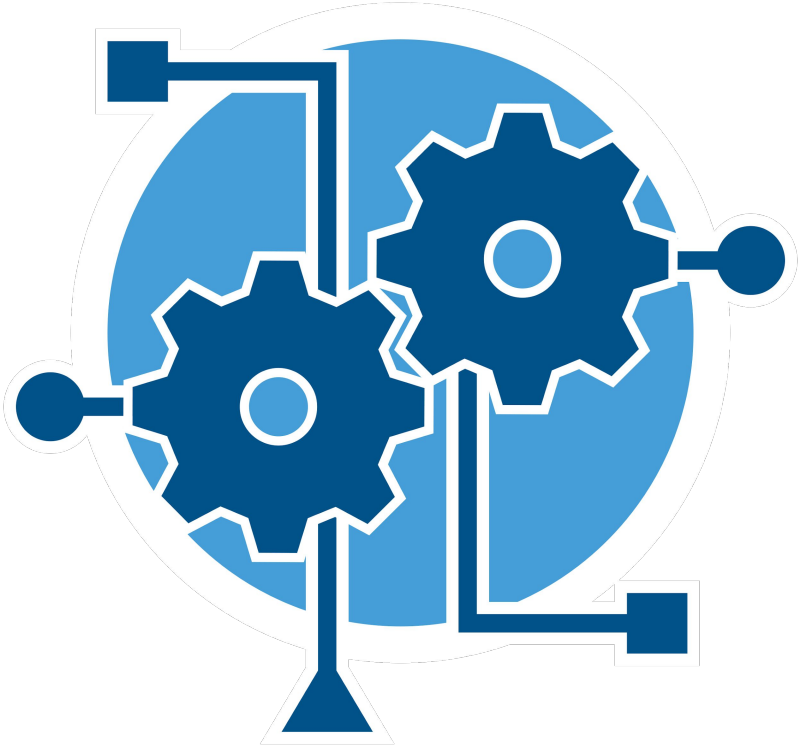
And others.

2. Getting Started - Data Encoding/ Transformation

2. Getting Started - Data encoding/transformation

This aspect of working through the data deals with how you execute decisions on your features. This is broadly divided into two:

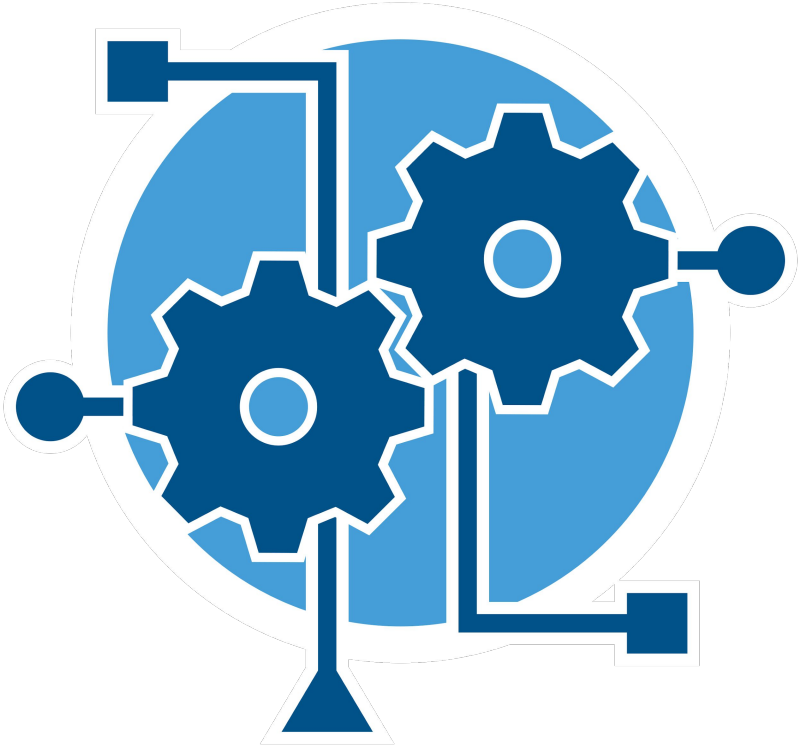
1. Normal transformation.
2. Feature engineering.



2. Getting Started - Data encoding/transformation

Normal transformation.

1. Filling missing values.
    a. Mean
    b. Median
    c. Mode
    d. Temporary filling (forward and backward filling).
2. Outliers detection/box plotting.

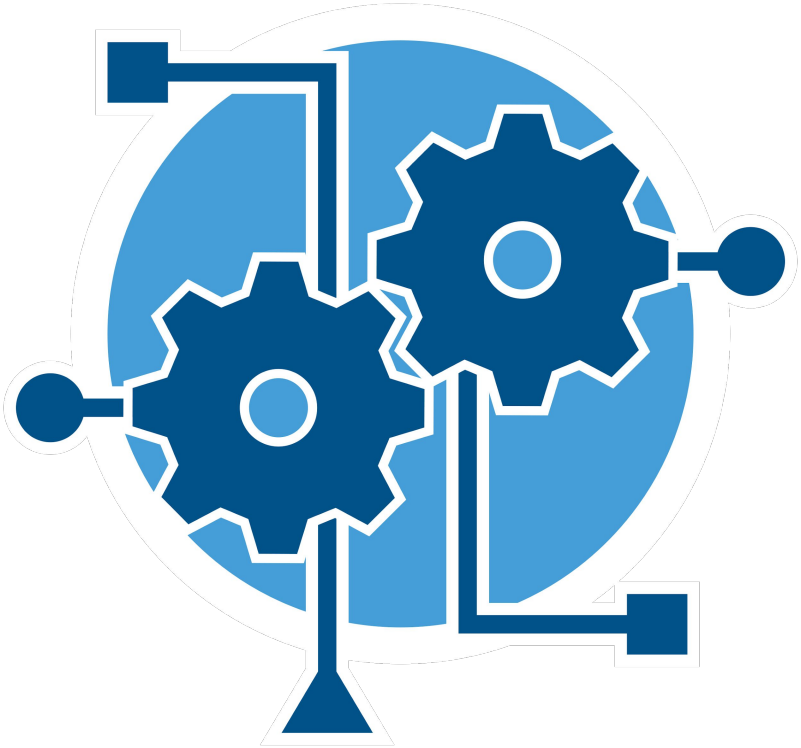


2. Getting Started - Data encoding/transformation

Feature engineering

1. Label encoding.
2. One-hot encoding.
3. Polynomials.
4. Feature normalization.
5. Feature Scaling.

And others.

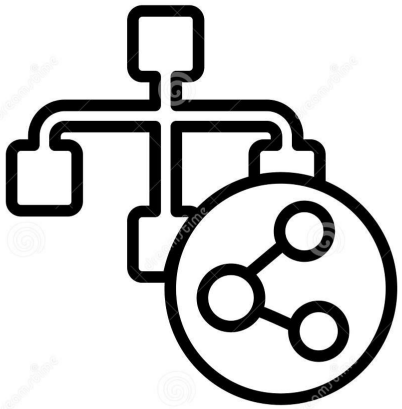


3. Modelling - Simple Modelling

3. Modelling - Simple Modelling.

Simple Models:

1. Classification
    a. Logistic regression
    b. Support vector classifier.
2. Regression
    a. Linear Regression (ridge and lasso).
    b. Support vector regressor.



data modelling

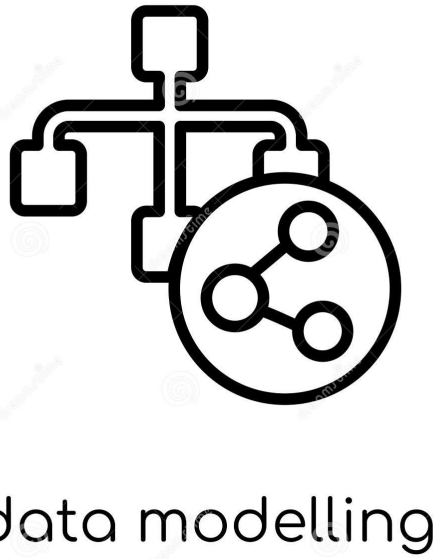
3. Modelling - Ensembles/robust ML models

3. Modelling - Ensembles and robust ML models.

Ensembles and robust ML models.

Classification and regression

1. Basic robust models
    a. Decision Tree.
    b. Random forest.
    c. Xgboost.
    d. LightGBM
    e. Cat boost
2. Neural Networks



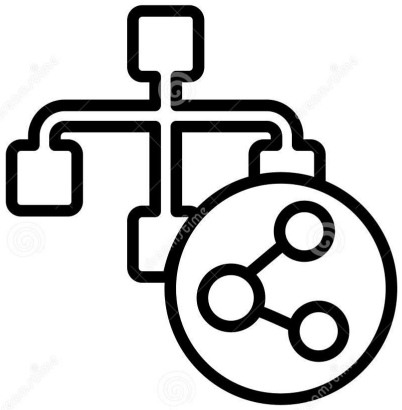
3. Modelling - Hyperparameter tuning

3. Modelling - Hyperparameter tuning.

Hyperparameter tuning.

This is done mostly to optimally decide model performance.

1. GridSearchCV
2. RandomizedSearchCV
3. Cross Validation



data modelling

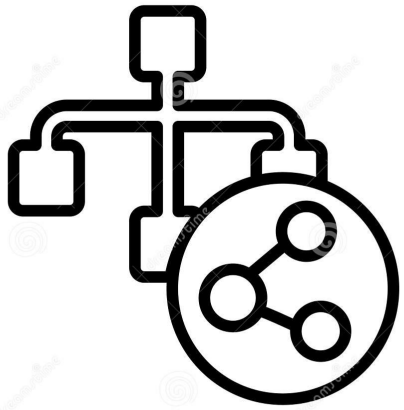
4. Re-evaluation/modelling

4. Modelling - Re-evaluation and modelling.

Re-evaluation and modelling.

This process deals with going over the entire process you've gone through.

Moreso, this also could mean you go the extra mile to stacking models/networks into build a more robust model.



data modelling

5. Conclusion

5. Conclusion

In conclusion, competitions helps to grow your data science and ML skills.

It also gives you insight on how to solve problem or enhance your problem solving strategies.

6. Questions and answers

6. Questions and answers