The Art of Data Science: Transforming Data into Actionable Insights

Premanand S

June 3, 2023

Preferable languages used for Machine Learning

Table: Tug of war between languages

Python	**R**	**Julia**
General purpose	Statistical analysis	Scientific computing
Good	Good	**speed** & **performance**
Huge community	Huge community	small community
200k libraries	15k libraries	3k libraries
In Billions	In Billions	13M downloads
-	-	**Compile just in time**
Jupyter, Pycharm	R Studio	Juno IDE
ijulia	-	-

Different IDE's for Machine Learning

- Thonny
- Visual Studio Code
- Atom
- Google Colaboratory
- Jupyter NB
- Spyder IDE
- Pycharm Community

Important libraries for Data Science

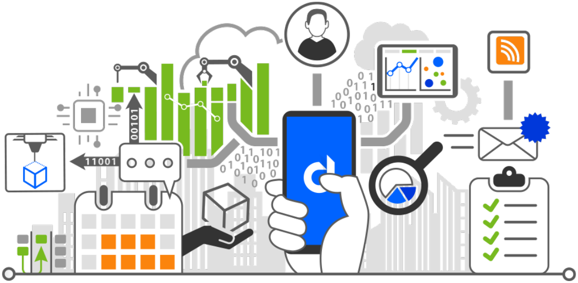
- **Numpy** - Scientific computing
- **Pandas** - Data analysis and manipulation
- **Scipy** - Maths, Science, Engineering
- **NLTK** - NLP
- **Matplotlib** - Data Visulaization
- **Seaborn** - Statistical Visualization
- **Bokeh** - High end Visualisation
- **Scikit-learn** -Machine Learning library
- **StatsModels** - Statistics
- **SymPy** - Symbolic maths, computer algebra system
- **Keras** - Neural Netowork
- **Tensorflow** - Fast numerical computing

Basic Steps for Data Science

- Data Collection or Acquisition
- Importing Libraries
- Loading Datasets
- Pre-processing and Exploratory Data Analysis (EDA)
- Splitting of Datasets
- Feature Scaling
- Feature Selection
- Dimensionality Reduction
- Modelling
- Metrics

Need an interesting read, Glimpse to Machine Learning- Hit me!

Step 1: Data Collection or Acquisation



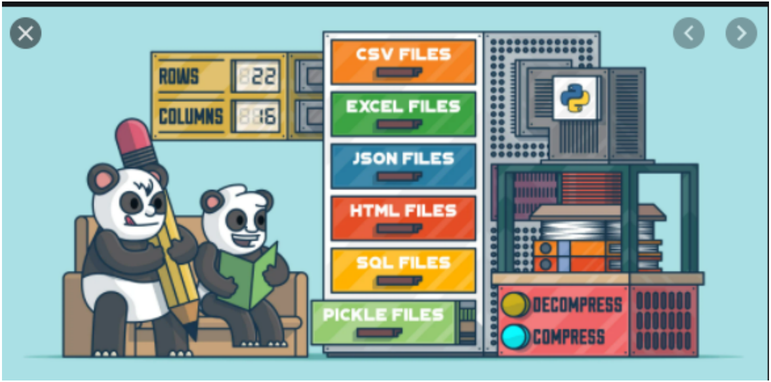
- By using Sensors, Medical devices like ECG, PPG...
- Google dataset search - link
- UCI Machine Learning Repository - link
- CMU libraries - link
- OpenML - link
- Fivethirtyeight - link
- Physionet - link
- Kaggle datasets - link
- Data.gov - link
- Academic torrents - link
- Awesome dataset by github - link and many more...

Step 2: Importing libraries

- Either installing or importing pacakages
- Through PIP install! - Through conda install

Step 3: Loading Datasets

- .csv, .json, .xlsx, .xml, .docx, .txt, .pdf, .png, .jpg, .mp3, .mp4



Step 4: Pre-Processing & Exploratory Data Analysis

- 60 - 70 work in this step
- Data information (.head, .tail, .shape, .columns)
- Overall data type (.info)
- Understanding basic statistics of data (.describe)
- Target details (.unique, .valuecounts)
- Checking for missing values (.isnull.sum)
- Solution for missing values (SimpleImputer)
- Outlier detection (Univariate: Box-plot, Grubbs test, Multivariate: PCA, Mahalanobis Distance, Cook's Distance...)
- Skewness (log transformation, square root transformation, box-cox transformation) and Kurtosis of data
- Correlation between features (.corr)
- Dependent and Independent variables
- Encoding categorical data for both dependent and independent variables (label encoder, OneHotEncoder, pd.getdummies)

Step 5: Splitting of datasets

Step 6: Feature Scaling

- Normalization Vs Standardization
- Normalization (z score, min-max, scaling to unit length, logarithmic scale) and Standardization
- Feature transformation (Scaling (Minmax scaler, standardscaler, normalizer, robustscaler), Discretization, Binning

Step 7: Feature Selection

If you put garbage in, you will only get garbage to come out

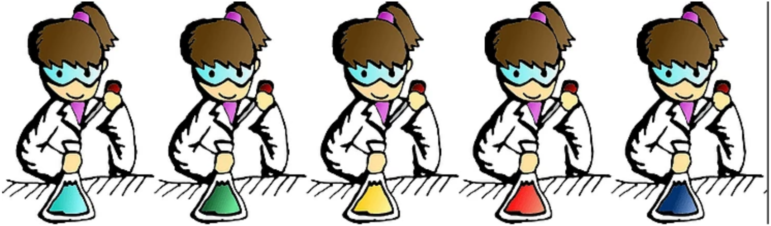
- Random forest classifier, chi-2, select from model, variance threshold, correlation threshold, Pearson's correlation (heatmap), chi squared, ANOVA f value, maximal information coefficient (MIC)
- wrapper based – forward search, backward search, recursive feature elimination (RFE)(rfe.support, rfe.ranking)
- sequential feature selector (SFB, SBS, SFFS, SBFS)
- Embedded methods – lasso regularization in linear regression, select k best in random forest, Gradient Boosting Machine (GBM) (univariate and multi variate feature selection)

Step 8: Dimentionality Reduction

Critic cat
is not impressed by the artist's utter lack of dimensionality
ICANHASCHEEZBURGER.COM

- What are Dimensionality Reduction Techniques?
- The Curse of Dimensionality
- Importance of Dimentionality Reduction
- Feature Selection
- Feature Extraction

Step 9: Modelling



- Regression / Classification algorithms
- Fit & Predict
- Hyper Parameter!
- Grid search
- Cross validation

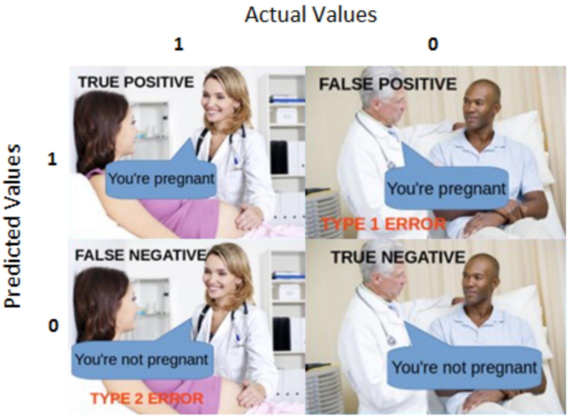
Step 10: Metrics

HOW DID OUR AD PERFORM ON YOUR NETWORK?
GREAT!
YOUR AWESOMENESS METRIC IS UP 22%,
YOUR COOL FACTOR METRIC IS UP 36%,
YOUR SMILEY-FACE METRIC IS UP 58%,
AND A.M.W.T.M.U. IS UP A RECORD 85.4%.
WHAT'S A.M.W.T.M.U.?
ANOTHER METRIC WE TOTALLY MADE UP.
TOM FISH BURNE
© marketoonist.com

Metrics - Regression

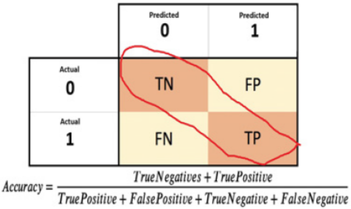
S No	Term	Criterion
1	R-Squared	High
2	Adj R-squared	High
3	F-Statistics	High
4	Std.Error	Close to zero
5	t-Statistics	>1.96 <0.05
6	AIC (Akaike Info Crit)	Low
7	BIC (Bayesian)	Low
8	Mallows cp	Should be close to no of target
9	MAPE (Mean Abs Per Err)	Low
10	MSE (Mean Squ Err)	Low
11	MPE (Mean Per Err)	Low
12	Min-Max Acc	High

Metrics - Classification - Confusion matrix 1



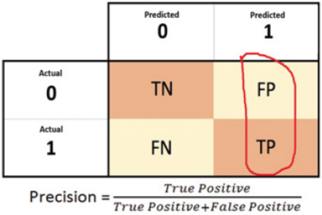
Metrics - Classification - Confusion matrix 2

- f1 score = 2*((precision*recall)/(precision+recall))



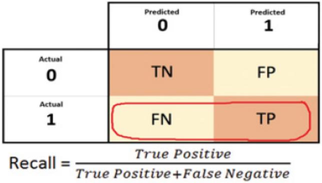
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$



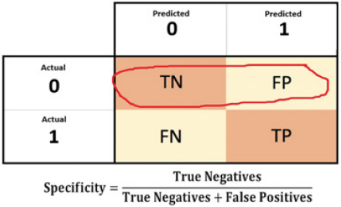
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Not convinced, need more clarity for DS ?

I always believe, if any concept can be conveyed through animation, it wont forget easily.
This is the link: Data Science @Simplilearn.

So Some tips to improve!

Some best course for Machine Learning across the globe

- Python (Dr. Charles Severance) & Coursera
- Springboard - Machine Learning Career Track
- Coursera - Machine Learning - Stanford University
- Udemy - Machine Learning A-Z (Python and R in Data Science)
- EdX - Data Science - Machine Learning
- Machine Learning Mastery - Machine Learning track
- Python for Everybody (youtube) - Charles Severance
- Datacamp - Machine Learning Scientist with Python - Free subscription!

Tips to improve Machine Learning coding & knowledge

- Open GitHub repository & and start coding from scratch for different dataset Github link - Details about Github - Hit me!
- Try to participate, competitions in Kaggle website for major attractions. Eg: Abhishek Thakur (Approaching (Almost) Any Machine Learning Problem) - Kaggle link
- Try to follow worlds top scientist, R & D, some reowned personalities in the field of Data science, Machine Learning and Deep Learning for their work and tips - Linkedin link

To summarize, Data Science

Data tells a story, But it needs a guide, Someone to uncover, The gems it tries to hide.
Data science is the key, To unlock the secrets within, With code as our magic, We let the insights in.
From clustering to regression, We let the algorithms run, And from the data we collect, We find patterns that were none.
With data at our fingertips, The world is ours to see, Data science is the lens, To unlock its mysteries.

mail me: premanand.s@vit.ac.in and er.anandprem@gmail.com
ring me: +91 73586 79961
follow me: Linkedin
Website: anandsdata

All of us don't have equal talent. But all of us have equal opportunity to develop our talent!