**Contents and Links in DataScience**

Transition towards DataScience

\*\*\*\*\*\*\*\*\*\*\*\*\*\***Step By Step Action Plan For Learning Data Science**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **Programming Language** : \*Python / R - (Java, Scala, Julia) - Reason: Huge Inbuilt libraries

\* **Python** :

IDE -

1. PyCharm

2. Jupyter\*

3. Spyder

Basic Syntax : variables, string, lists, tuples, dictionaries

Functions

Oops concept

Exception Handling

Libary: Numpy (to create arrays)- import, basic operations on arrays (Indexing,reshaping, broadcasting)

--------------------

\* **Data Preprocessing**

Library: Pandas (Data preprocessing) - DataSeries, DataFrames Operations - Convert into array - Given to ML Machine Learning model to train the Data,

Handling Missing Data

Scaling Data - normal distribution, Std ND

Library: scikit - All kind of ML algorithms - Superwise and Unsuperwise algorithms -

Seaborn: Python data visualization library based on matplotlib (correlation, pieplot, count plot, histogram, probabilty density functions..)

--------------------

2. **Visualization Tools**: Reporting Tools to publish report

PowerBI\*

Tableau

Qlik Sense\*

MATPLOTLIB, SEABORN in python

--------------------

3. **Mathematics**

Statistics - mean, median, mode, normal distribution, std ND, chernobyl inequality statistics, central limit theorem, IQR range, percentile, correlation, covariance, subNormalDistribution,std deviation, skew - Useful for substituting missing value

Differential Equations

Linear Algebra

statistics

Differential Calculus

Discrete Maths

Numerical Analysis

Matrices

Scalars & Vectors

Probabilty

------youtube - khanAcademy-----------

\*\*\*\*\*\*\*\*Processed Data Out for Algorithms\*\*\*\*\*\*\*\*

--------------------

4. **Exploratory Data Analysis EDA**

\*Consider a Dataset - Analyise What are Continuous & Discrete variables, Categorical variables,

\*For Suppose Data in form of Normal Distribution - How to scale down data - 2 ways - i. min max scalar ii. log normal scaling technique iii. std scalar

\*Scaler library -

--------------------

5. **DataStructures and Algorithm** - Use for writing Efficient Code

To store data, values

DataFrames, Tuples, Lists...

Sorting Algorithm

How fast our program runs

--------------------

6. **Machine Learning**

Classification

Regression

Reinforcement

Deep Learning

Dimensionality Red

Clustering

\* Algorithms - 15 to 20 compulsory -

Linear Regression algorithm

lasso regression algorithm

ridge regression algorithm

logistic regression algorithm and classifier

decision tree algorithm

random forest regression and classifier

adaboost regression and classifier

xgboost regression and classifier.............

\* Hyper Parameter tuning for above algorithms : Understand Maths - GridSearchCV, RandomSearchCV difference.

\* Knowledge on pipeline mechanism - Step by step implementation

Step1 : Doing Scaling Mechanism

Step2 : Doing principle component analysis on dimensional reduction

Step3 : Applying Algorithms

\*\*\*\*\*\*\*\*\*Machine Learning Model Created\*\*\*\*\*\*\*\*

--------------------

7. **Deployment**

\*\*\*\*\*\*\*We go for Deployment\*\*\*\*\*\*

Deployment in Azure

Deployment in Heroku - Platform as a service - Preferable - Free of Cost

Deployment in aws ec2 - Infrastructure as a service

Deployment in google cloud

\* Deveops - \*dockers and kubernetes - How to deploy code in a much efficient and scalable way

--------------------

8. **Deep Learning**

Algorithms

artificial Neural Network - ANN -

convolutional Neural Network - CNN - Image processing and classification

recurrent Neural Network - RNN - Time series data

All new algorithms are derived from above 3 algorithms only

Advanced Convolutional Neural Networks like vgg16 neural network, resnet- 50 neural network

Master Region-based convolutional neural network (RCNN), YONO cnn neural network, object detection

Deployment:

Same as ML

Libraries:

Keras : Wrapper on Tensor Flow

Tensor Flow\*

pytorch\*

**fast.ai** - fast.ai is a non-profit research group focused on deep learning and artificial intelligence.

- free open source library for deep learning called fastai (without a period), sitting atop PyTorch

--------------------

9. Framework : **Flask, Django**

Flask is a micro web framework written in Python.

Django is a Python-based free and open-source web framework that follows the model-template-view architectural pattern.

Model that is deployed in production is exposed as API to FrontEnd User

Create API

--------------------

10. **Databases**

SQL : Structured data : Merge data from Tables

Mongo Db (No sequel database) : Semi structured data(JSON) && Unstructured data(Image) - Install pymongo

BigData (hadoop) - hadoop cassandra - PEG, Hive, Flume are bigdata Architecture - All type of data (Structured and Unstructured data also)

django database (AWS)

\* Cloud Computing Services - AWS, Azure, cloud foundry

\*\*\*Train Live data continuously

--------------------

11**. Natural Language Processing** (NLP)- Deals with Text Data

Chatbot

eg: Political Sentiment Analysis, When Stock Market collapse

Chatbot Tool- AI tool to Quickly train the data

--------------------

12. **GIT repository**

Create branch, Merge branch, Master branch....

--------------------

13. **Excel**

--------------------

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

------------------------------

**Freelancing work**

https://www.upwork.com/

https://www.freelancer.in/jobs/

------------------------------

**Blog References** :

https://www.kdnuggets.com/2019/01/automated-machine-learning-python.html

https://towardsdatascience.com/

https://medium.com/topic/machine-learning

------------------------------

**Books :**

**Paid**

1. Book on Finance with Machine Learning and Deep Learning from the below url amazon url: https://www.amazon.in/Hands-Python-Fi...

https://www.amazon.in/Hands-Python-Finance-implementing-strategies/dp/1789346371/ref=as\_sl\_pc\_qf\_sp\_asin\_til?tag=krishnaik06-21&linkCode=w00&linkId=ac229c9a45954acc19c1b2fa2ca96e23&creativeASIN=1789346371

Hands-On Python for Finance: A practical guide to implementing financial analysis strategies using Python

2. Best book of Machine Learning, Deep Learning with python sklearn and tensorflow from below amazon url: https://www.amazon.in/Hands-Machine-L... https://www.amazon.in/Hands-Machine-Learning-Scikit-Learn-Tensor/dp/9352135210/ref=as\_sl\_pc\_qf\_sp\_asin\_til?tag=krishnaik06-21&linkCode=w00&linkId=a706a13cecffd115aef76f33a760e197&creativeASIN=9352135210

Hands-On Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems

3. Ethics and DataScience: https://www.oreilly.com/library/view/ethics-and-data/9781492043898/

**Free**

\* Python Data Science Handbook: https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf

\* Think Stats 2e: https://greenteapress.com/wp/think-stats-2e/

\* R For Data Science: https://r4ds.had.co.nz/index.html

\* Rules For Machine Learning: http://martin.zinkevich.org/rules\_of\_ml/rules\_of\_ml.pdf

\* Deep Learning: http://www.deeplearningbook.org/

------------------------------

**Datasets**:

Kaggle: https://www.kaggle.com/datasets

UCI: https://archive.ics.uci.edu/ml/index.php

Quandl: https://www.quandl.com/

US Government Open Dataset: https://www.data.gov/

Indian Government OpenDataset:https://data.gov.in/

World Bank Dataset: https://data.worldbank.org/

Group Lens dataset: https://grouplens.org/

Awesome Public Dataset: https://github.com/awesomedata/awesome-public-datasets

Google: https://cloud.google.com/public-datasets (One Year Free)

**Webscraping**

1. Beautiful SOUP

2. SCRAPY

3. URLLIB

------------------------------

**System Configuration**:

RAM - 8 GB

HardDisk - 500GB

Processor - intel i5

Graphic Card - GPU NVDIA - GTX 1050/ 1060 / 1070 / P5000 / P6000 & Virtual VRAM - 4 GB in GPU

(NVDIA - CUDA library Inbuilt - Works good for Tensor Flow)

Papperspace GPU url: https://www.paperspace.com/ - (Preferable for Deep Learning) - Virtual Graphic Card Available (Paid)

Google colab - ML and small projects

Operating System OS in DataScience - **Linux**

Don't require AntiVirus

Open Source

Easy to integate GPU

Inbuilt Python installed

Fast

Huge libraries

Camera Support Libraries

Easy Deployment in Cloud

Working in Linux makes the problem solution Simple

MSI Laptops: https://www.amazon.in/MSI-GL63-9RCX-213IN-i5-9300H-Graphics/dp/B07TNMKVW8/ref=sr\_1\_3?keywords=msi&qid=1569687428&sr=8-3

Lenovo: Lenovo Legion Y 540 - 1 Lakh appx.

------------------------------

**Online Courses:**

**Paid**

1. Applied Data Science with Python Specialization Certificate from University of Michigan https://www.coursera.org/specializations/data-science-python

2. Data Science Specialization From JOHNS HOPKINS University https://www.coursera.org/specializations/jhu-data-science

3. Statistics and Data Science MicroMasters — MIT (edx) (1 lakhs or $1500) https://www.edx.org/micromasters/mitx-statistics-and-data-science

4. Udemy - Jose Portilla- Python for Data Science and Machine Learning Bootcamp https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/

Machine Learning A-Z – Hands on Python and R in Data Science by Kirill Eremenko, Hadeline de Ponteves

5. IBM Data professional Data Science certificate on Coursera https://www.coursera.org/professional-certificates/ibm-data-science

**PG DIPLOMA**

1. Upgrad https://www.upgrad.com/

2. Great Learning https://www.greatlakes.edu.in/e-learning/pgp-in-data-science-and-engineering-full-time

3. Applied AI course https://www.appliedaicourse.com/ (Preferable) - YouTube Free Videos - Srikanth

**Free (Order to Learn)**

1. Python

Corey Schafer https://www.youtube.com/user/schafer5

Sentdex https://www.youtube.com/user/sentdex

2. Machine Learning

Machine Learning with Maths, Statistics and Linear Algebra Andrew NG applied AI https://www.youtube.com/watch?v=PPLop4L2eGk&list=PLLssT5z\_DsK-h9vYZkQkYNWcItqhlRJLN

Krish Naik https://www.youtube.com/watch?v=EqRsD3gqeCo&list=PLZoTAELRMXVOnN\_g96ayzXX5i7RRO0QhL

Sentdex https://www.youtube.com/user/sentdex

Statquest with Josh Starmer https://www.youtube.com/user/joshstarmer - Explaination with Animations

3. Natural Language Processing

Krish https://www.youtube.com/watch?v=6ZVf1jnEKGI&list=PLZoTAELRMXVMdJ5sqbCK2LiM0HhQVWNzm

Sentdex https://www.youtube.com/user/sentdex

3. Deep Learning

Andrew Ng https://www.youtube.com/watch?v=CS4cs9xVecg&list=PLkDaE6sCZn6Ec-XTbcX1uRg2\_u4xOEky0

Krish Naik https://www.youtube.com/watch?v=DKSZHN7jftI&list=PLZoTAELRMXVPGU70ZGsckrMdr0FteeRUi

4. Data Science Projects https://www.youtube.com/watch?v=5Txi0nHIe0o&list=PLZoTAELRMXVNUcr7osiU7CCm8hcaqSzGw (Paid)

5. Feature Engineering Playlist https://github.com/aikho/awesome-feature-engineering

6. Feature Selection Playlist https://github.com/anujdutt9/Feature-Selection-for-Machine-Learning

7. Spring board India Youtube url: https://www.youtube.com/channel/UCg5UINpJgS4uqWZkv2Qh1Mw

------------------------------

\*\***AutoML** - AutoML consumes more time.

1. AutoKeras: An AutoML system based on Keras. It is developed by DATA Lab at Texas A&M University. The goal of AutoKeras is to make machine learning accessible for everyone. Useful for Domain Expertiee.

2. Auto-sklearn : Frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in Bayesian optimization.

3. TPOT : The Tree-Based Pipeline Optimization Tool (TPOT) was one of the very first AutoML methods and open-source software packages developed for the data science community.

4. H20AutoML : H2O is an open source

5. Google AutoML is Paid version

------------------------------

**Research Papers for Professionals** :

Visit IEEE

Hyperparameter Tuning -

optimizers -

Use case based Research Paper- Agriculture

- HealthCare

- Insurance

Problem Statements:

Changing Background according to Browsing History

A/B Testing using ML

ChatBoard dashboard Recommendation for Insurance Product

------------------------------

**Extra Skillset**

1. NLP (Alexa, Google Assistant) - ChatBoard (Sarcasm Recognition)

2. BigData

3. Computer Vision Data (raspberry pi Stand Alone device)

------------------------------

\* **Extra Points**:

Techincal:

\* To get ML results accurately - Train the Data continuously

\* gartner survey and Kaggle survey - Lack of Data Science Engineers

\* Reverse Engineering : Use Case - Mathematics Equations

\* Pipeline Architecture - Steps to Follow - Feature Engineering, Feature Selection and Training Output

\* Commit Code in Github

NonTechincal:

\* Writing Blogs

\* Update Linked In Profile

\* Salary : Visit GlassDoor Website, linkedin

\* Kaggle Competition Rank in between 1 to 100

\* Hackathon Participation

\* Certifications

\* YouTube as a Carrier Option

------------------------------

**Resume -**

Final Year Student - Motive: Internship

1. Create a POC concept

2. Have good skills in Data Structures and Algorithms also

3. Commit the code in the Github Repository

4. Add complete Project Description

5. Share the project or the problem statement in the blogs

6. Share the blogs in Linkedin Profile

7. Participate in Kaggle competitions and share the solution in the blogs

8. Research Papers

0-2 Years

0 Create a Data Science Project in ur company with the domain knowledge u have

1 Share the challenges that you faced and how you fixed in blogs

2 Share the blogs in Linkedin Profile

3 Look For Internal transfer

4 Show Your Manager and tell them about ur ideas with respect to the project and how it will add values to the project

5 Eg: A/B Testing

Projects Required For Professionals

1. HealthCare Domain- Medical Image Analysis, Disease Prediction, Pneumonia Detection,

2. Ecommerce Domain: Projects: Recommendation System, A/B testing on Products, Creating Chatbots for Recommendation

3. Bank : Chatbot for recommended products, Account Opening, Credit Card Fraudlent

3-8 Years

1. Create a Complete Project end to end Deployment

2. Try to get multiple Domain Expertiee

3. Impact of the project whether it has reduced the cost

4. Write Blogs with Advanced problem statement with end to end deplotment

5. Research with different platforms (servers, deployment mechanism)

6. Research Papers, File for Patents

------------------------------

Connect with **Krish Naik** here:

Twitter: https://twitter.com/Krishnaik06

Facebook: https://www.facebook.com/krishnaik06

instagram: https://www.instagram.com/krishnaik06

Support in Patreon: https://www.patreon.com/join/2340909?

YouTube Playlist Order

1. Statistics in Machine Learning

2. Complete Machine Learning Playlist

3. DataStructures and Algorithm - Use for writing Efficient Code

4. DataScience and Machine Learning with Python and R

5. DataScience Projects

6. Deployment of ML models

7. Complete Deep learning

8. DataVisualization Tutorial

9. PowerBI

10. Tableau

11. Natural language processing

12. Feature Enggineering

13. Computer vision tutorial

14. Data Science Interview Question............

\*\*To join as a member in Krish Naik channel to get additional benefits like materials in Data Science, live streaming for Members and many more Live Projects https://www.youtube.com/channel/UCNU\_lfiiWBdtULKOw6X0Dig/join

\*\* Mighty Data Science Bundle - http://bit.ly/37evQR7

\*\* AI and ML Suite - http://bit.ly/35xSRNI

Coupon KNBF20" to get Extra 20% Off

Thanks Krish ... I collected most of this Data from your Teachings

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

- Pradeep J