Data Scientist – Learning Path

Programming background – known syntax basics:
C (university),
C++ (high school, self-taught),
C# (self-taught)

Technologies & Tools:
Programming language: Python 3.10,

Libraries: Pandas, NumPy, Matplotlib, Seaborn, SciPy, Scikit-learn, Pickle, NLTK, Gensim, SpaCy, SciKeras, TensorBoard, Streamlit, BeautifulSoup4, lxml, requests,

Frameworks: Tensorflow (CPU & GPU),

Databases: MySQL Workbench,

Analytical tools: Tableau Public, Power BI, Excel,

Version control & containers: Git, Docker

Operating system: Windows 10,

IDE & environments: PyCharm, Anaconda, Jupyter Notebook, VS Code,

Learning sources:
YouTube: Corey Schafer, Mosh, Keith Galli, Josh Starmer, Derek Banas, Alex The Analyst, Tech With Tim, 3Blue1Brown, Alejandro AO - Software & Ai, Krish Naik,

Udemy: Mosh, Paweł Krakowiak, Kirill Eremenko, Krish Naik (**Data Science Bootcamp ->** https://www.udemy.com/course/complete-machine-learning-nlp-bootcamp-mlops-deployment/?couponCode=KEEPLEARNING)

Documentation:
https://docs.python.org/
https://docs.pypi.org/
https://www.w3schools.com/python/
https://pandas.pydata.org/docs/user_guide/
https://matplotlib.org/stable/api/index.html
https://seaborn.pydata.org/api.html
https://scikit-learn.org/stable/api/index.html
https://xgboost.readthedocs.io/en/release_3.0.0/
https://www.nltk.org/api/nltk.html

Databases:
https://github.com/
https://www.kaggle.com/
https://stackoverflow.com/
https://archive.ics.uci.edu/

Python syntax, data processing and visualization, SQL

✅Python:

- Python (syntax, variables, data types, loops, functions),
- Data structures (lists, tuples, sets, dictionaries), comprehensions (map, filter, reduce, lambda), generators, datetime,
- Import modules, LEGB (Local, Enclosing eg. function nested in function, Global, Built-in),
- Try, except, finally, raise, type hints, *args i **kwargs,
- OOP, first-class functions, closures, decorators (property decorators), inheritance, dunders,
- File operations (csv, json, txt, xlsx),
- Threading, multiprocessing,

✅Data processing:

- NumPy (arrays, indexing, mathematical operations),
- Pandas (data loading, filtering, grouping, aggregation, apply, lambda, imputation, deletion),

✅Visualizations:

- Matplotlib & Seaborn (creating and editing plots),
- Statistical data analysis (mean, median, variance, standard deviation, correlation, covariance, distributions, statistical tests),

✅SQL:

- SQL – basics (select, where, having, limit, aliasing, group by, join, union, case, strings),
- Advanced SQL (subqueries, window functions, CTEs, temp tables, stored procedures, triggers and events, data cleaning),

Visualization cont., Machine Learning, NLP

✅Visualizations:

- Excel – conditional formatting, pivot tables, visualizations, formulas and functions, lookups, data cleaning,
- Power BI – prepering data/formatting, joins and relationships, DAX, drill down, groups, conditional formatting, visualizations,
- Tableau – visualizations, joins and relationships,

✅ML:

- Supervised ML: linear regression, logistic regression, kNN, SVM P.2(kernel) – theory + sklearn implementation,
- Unsupervised ML: K Means Clustering – theory + sklearn implementation,
- Hyperparameter tuning (GridSearch, Random Search),

- Supervised ML: decision trees (sklearn implementation, Gini and entropy, manual implementation with entropy), random forests,

✅ML cont.:

- Unsupervised ML: Dbscan Clustering,
- Feature Selection – Variance Threshold, Pearson Correlation, Information Gain, Chi-Square Test, Recursive Feature Elimination (RFE/RFECV), L1 Regularization, Mutual Information,
- Gradient Boosting (XGBoost, AdaBoost),
- Model evaluation metrics (Accuracy, MSE, RMSE, Confusion Matrix, ROC AUC, F1 Score – Recall/Precision),
- Dimensionality reduction PCA,
- Supervised ML: MultinomialNB,

✅NLP:

- NLP Text Preprocessing: cleaning the input – Tokenisation, Stemming, Lemmatization, Stopwords, Parts of Speech, Named Entity Recognition,
- NLP Text Preprocessing cont.: conveting text to vectors – One Hot Encoding, Bag of Words, N-Grams, TF-IDF, Word Embedding, Word2vec (CBOW, Skipgram / pre-trained by Google and from scratch), Avgword2vec,

Deep Learning

✅DL:

- Deep Learning theory:
  - ANN, Perceptron, Propogation, Weights,
  - Exploding Gradient Problem (Uniform Distribution, Xavier/Glorot Initialization, Kaiming He Initialization),
  - Vanishing Gradient Problem, Activation Functions (Sigmoid, Tanh, Relu, ELU, Softmax),
  - Loss and Cost Function (Regression – MSE, MAE, Huber Loss, RMSE, Classification – Binary/Categorical/Spas Cross Entropy),
  - Gradient Descent Optimisers (SGD, Mini Batch SGD, SGD with Momentum, Adagard, RMSPROP, Adam),
  - Drop Out Layer,
- CNN (image processing), Padding, Relu Operation, Max/Min/Mean Pooling, Flatting, Fully Connected Layer,
- End to End Deep Learning Project Using ANN – Customer Churn Prediction + Streamlit

- RNN,
- End to End Deep Learning Project With Simple RNN,

- LSTM, GRU

- LSTM and GRU End To End Deep Learning Project – Predicting Next Word,

- Bidirectional RNN/LSTM, Encoder/Decoder, Seq2Seq, Attention Mechanism,
- In depth Transformers Architecture,

- Beautifulsoup and requests basics,