Data Platform

Oracle
Teradata
MS SQL
HDFS
MongoDB

Data Model

Hierarchical Data Model
Entity Relationship Model
Dimensional Data Model
Graph Data Model

ETL Software

Informatica
DataStage
Ab Initio
SSIS
Spark/Pandas
RDBMS Utilities

Data Science Packages

- R
- Python
    - Scikit-learn
    - Tensorflow
    - Pytorch
- Java
    - Apache Commons Math
    - Weka
    - Deeplearning4

Data Visualization

Matplotlib
Seaborn
JavaScript D3
Tableau
Power BI
OBIEE

Transformation

- Normalization/Scaling
- Standardization/Z-score Normalization
- Log Transformation
- Box-Cox Transformation
- Binning/Discretization
- One-Hot Encoding
- Label Encoding
- Date/Time Features
- Imputation
- Interaction Terms
- Text Vectorization
    - Bag-of-Words
    - TF-IDF
    - Word Embedding
- Principle Component Analysis (PCA)
- Laplace
- Fast Fourier Transform (FFT)
- Wavelet

Smoothing Techniques

Moving Average
Kernel Smoothing (Kernel Density Estimation)
Local Regression (LOESS/LOWESS)
Spline
Smoothing Splines
Gaussian Processes (GP)
Exponential Smoothing
Tikhonov or Ridge Regression
Fast Fourier Transform Denoise

Variable Selection

Stepwise
LASSO
Ridge

Cross-Validation

Kfold
Leave-One-Out
Holdout
TimeSeriesSplit
Monte Carlos

Model Techniques

- Ensemble
    - Bayesian Model Averaging (BMA)
    - Stacking (Stacked Generalization)
    - Bagging (**B**ootstrap **Agg**regat**ing**)

ML Linear Models

- Linear Regression
    - Simple/Multiple Linear Regression
    - Polynomial Regression**
    - Ridge Regression (L2 Regularization)
    - Lasso Regression (L1 Regularization)
    - Elastic Net Regression
    - Bayesian Linear Regression
    - Generalized Linear Models (GLM)
    - Weighted Least Squares (WLS)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Linear Discriminant Analysis (LDA)
- Logistic Regression
- Linear Support Vector Machines (SVM)

ML Non-Linear Models

- Isotonic Regression
- Exponential Regression
- Logarithmic Regression
- Polynomial Regression
- Quadric Determinant Analysis (QDA)
- Decision Trees
    - Re-sampling:
        - Bagging
        - Random Forest
    - Re-weighting: Boosting
        - Gradient Boosting Machines
        - XGBoost (Extreme Gradient Boosting)
        - LightGBM
        - Adaptive Boosting (AdaBoost)
        - CatBoost
        - Stochastic Gradient Boosting (SGD)
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM) with Non-Linear Kernels
- Neural Networks
    - Convolutional Neural Network (CNN)
    - Capsule Network (CapsNet)
    - Long Short-Term Memory (LSTM)
    - Residual Neural Network (ResNet)
    - *Autoencoder*
    - *Generative Adversarial Network (GAN)*
    - *Transformer*
    - *Generative Pre-trained Transformer (GPT)*
- *Radial Basis Function (RBF) Networks*
- Gaussian Processes
- Kernelized Support Vector Machines (SVM)
- Generalized Additive Models (GAM) - Backfitting
- Kernel Ridge Regression
- Nearest Centroid Classifier

Statistical/Financial Models

Exponential Weighted Moving Average (EWMA)
Time Series Analysis
Monte Carlos Simulation
Copula Models
Stochastic Calculus
Markov Chain Monte Carlo (MCMC) Methods
AutoRegressive
Heterogenous AutoRegressive
SARIMAX
Generalized AutoRegressive Conditional Heteroskedasticity (GARCH)
Capital Asset Pricing Model (CAPM)
Mean Variance Optimization (MVO)
Potential Future Exposure(PFE)/Value at Risk (VaR) Methods
Black-Scholes
Arbitrage Pricing Theory (APT)
Factor Models

Distributions

Bernoulli
Binomial
Chi
Exponential
F
Log Normal
Poisson
Gaussian
Geometric
Student-t
Uniform
Weibull

Statistical Tests

R-squared
T-test
F-test
Durbin-Watson
Jarque-Bera
Skew
Kurtosis
Ljung-Box
Heteroskedasticity
ADF
KPSS
ACF
PACF