**DS基础知识**

**Python**

Open source packages: pandas, numpy, matplotlib

ML packages: scikit-learn

DL packages: tensorflow, pytorch

Math: linear algebra, probability

ML: xgb, lgb, catboost

DL: MLP, CNN, Transformer

Model做完后，Logistic regression做最后的集成，有时会有意想不到的惊喜

ICR – Identifying Age-Related Conditions

Use ML to detect conditions with measurements of anonymous characteristics.

Target: predict whether a person have one or more medical conditions out of the three conditions (Type 1) or have none of the three conditions (Type 0).

Create a well-trained model that measures relationships b/t some specific characteristics and potential patents conditions.

快速确定患者是否患病以支持后续治疗给予

医疗数据的数据量小是正常现象。以及政府数据、宏观层面金融数据(e.g., CPI).

𝐁𝐮𝐢𝐥𝐝𝐢𝐧𝐠 𝐚 𝐒𝐭𝐫𝐨𝐧𝐠 𝐅𝐨𝐮𝐧𝐝𝐚𝐭𝐢𝐨𝐧  
  
1. 𝑫𝒆𝒔𝒄𝒓𝒊𝒑𝒕𝒊𝒗𝒆 𝑺𝒕𝒂𝒕𝒊𝒔𝒕𝒊𝒄𝒔  
→ Measures of Central Tendency: Mean, Median, Mode  
→ Measures of Dispersion: Range, Variance, Standard Deviation  
→ Percentiles, Quartiles, Interquartile Range (IQR)  
→ Data Types: Nominal, Ordinal, Interval, Ratio  
  
2. 𝑫𝒂𝒕𝒂 𝑽𝒊𝒔𝒖𝒂𝒍𝒊𝒛𝒂𝒕𝒊𝒐𝒏  
→ Histograms  
→ Box Plots  
→ Bar Charts  
→ Scatter Plots  
  
3. 𝑩𝒂𝒔𝒊𝒄 𝑷𝒓𝒐𝒃𝒂𝒃𝒊𝒍𝒊𝒕𝒚 𝑪𝒐𝒏𝒄𝒆𝒑𝒕𝒔  
→ Basic Probability Rules (Addition, Multiplication)  
→ Conditional Probability and Independence  
→ Bayes’ Theorem  
→ Probability Distributions: Binomial, Poisson, Normal  
  
𝐒𝐭𝐫𝐞𝐧𝐠𝐭𝐡𝐞𝐧𝐢𝐧𝐠 𝐘𝐨𝐮𝐫 𝐊𝐧𝐨𝐰𝐥𝐞𝐝𝐠𝐞  
  
1. 𝑫𝒊𝒔𝒕𝒓𝒊𝒃𝒖𝒕𝒊𝒐𝒏𝒔 & 𝑫𝒂𝒕𝒂 𝑪𝒉𝒂𝒓𝒂𝒄𝒕𝒆𝒓𝒊𝒔𝒕𝒊𝒄𝒔  
→ Normal Distribution  
→ Binomial, Poisson, and Exponential Distributions  
→ Skewness and Kurtosis  
  
2. 𝑺𝒂𝒎𝒑𝒍𝒊𝒏𝒈 𝑻𝒆𝒄𝒉𝒏𝒊𝒒𝒖𝒆𝒔  
→ Random, Stratified, and Systematic Sampling  
→ Addressing Sampling Bias  
  
3. 𝑰𝒏𝒇𝒆𝒓𝒆𝒏𝒕𝒊𝒂𝒍 𝑺𝒕𝒂𝒕𝒊𝒔𝒕𝒊𝒄𝒔  
→ Hypothesis Testing: Null vs. Alternative Hypotheses  
→ p-values, Significance Levels, and Confidence Intervals  
→ Type I and Type II Errors  
→ Parametric Tests: T-tests, Z-tests, ANOVA  
→ Non-Parametric Tests: Mann-Whitney U, Kruskal-Wallis, Wilcoxon Rank-Sum  
  
4. 𝑪𝒐𝒓𝒓𝒆𝒍𝒂𝒕𝒊𝒐𝒏 𝒂𝒏𝒅 𝑹𝒆𝒈𝒓𝒆𝒔𝒔𝒊𝒐𝒏  
→ Pearson, Spearman, and Kendall Correlations  
→ Simple and Multiple Linear Regression  
→ Logistic Regression  
→ Multicollinearity and Residual Analysis  
→ Understanding Correlation vs. Causation  
  
𝐌𝐚𝐬𝐭𝐞𝐫𝐢𝐧𝐠 𝐭𝐡𝐞 𝐊𝐞𝐲 𝐂𝐨𝐧𝐜𝐞𝐩𝐭𝐬  
  
1. 𝑩𝒂𝒚𝒆𝒔𝒊𝒂𝒏 𝑺𝒕𝒂𝒕𝒊𝒔𝒕𝒊𝒄𝒔  
→ Prior, Posterior, and Likelihood  
→ Bayesian Inference and its Applications  
→ MCMC (Markov Chain Monte Carlo) Methods  
  
2. 𝑺𝒕𝒂𝒕𝒊𝒔𝒕𝒊𝒄𝒂𝒍 𝑴𝒐𝒅𝒆𝒍𝒊𝒏𝒈  
→ Generalized Linear Models (GLMs)  
→ Logistic and Poisson Regression  
→ Model Selection: AIC, BIC  
→ Cross-Validation, Overfitting, and Underfitting  
→ Principal Component Analysis (PCA)  
  
3. 𝑪𝒂𝒖𝒔𝒂𝒍 𝑰𝒏𝒇𝒆𝒓𝒆𝒏𝒄𝒆  
→ Propensity Score Matching  
→ A/B Testing and Controlled Experiments  
→ Quasi-Experimental Methods: Instrumental Variables, Regression Discontinuity  
  
4. 𝑨𝒅𝒗𝒂𝒏𝒄𝒆𝒅 𝑷𝒓𝒐𝒃𝒂𝒃𝒊𝒍𝒊𝒕𝒚 𝑪𝒐𝒏𝒄𝒆𝒑𝒕𝒔  
→ Law of Large Numbers  
→ Central Limit Theorem (CLT)  
→ Markov Chains and Stochastic Processes