Algorithm Selection Good interpretability No Poor interpretability Yes EDA: Data Pre-processing: Missing Large memory footprint exploration, Outlier value treatment, Data **Data** Data treatment, Data cleaning, Data visualization transformation Modeling Process Binary Classification Accuracy **Linear Boundaries** DT • GBM • **SVM** Post modeling: Pre-modeling: Sampling, RF XGB LR NBC Modeling Hyperparameter tuning, feature importance, model evaluation, cross Is it dimensionality reduction Requirement? validation Speed >100 features *Large **Time Series** dataset LR GBM Algorithm selection can be done as per modelling requirement from the flow chart Data? **SVM** SGD DT • NB (only for labelled data) -class locally For the common algorithms their characteristics mention under one type of **NBC** XGB **LGBM** deep SVM problem will hold good for other problems as well If you are dealing with Text Classification, please refer to multiclass and binary classification as the algorithms used are same If (no. of features > no. of samples) please get more data before proceeding #Classes = 2? **Predict** *For data too large, the limit beyond which it can be said that the data at hand can be **Category?** considered large or not completely depends on the problem at hand, type of dataset and other such specific factors. However, to give an example, a dataset containing 100k datapoints Speed, Accuracy Data too large? * can be considered large. **ARIMA EM** Trend Univariate Speed Accuracy Yes No Holt's Winter Auto ARIMA RF Seasonality LR **SGD** Linear Multivariable **Predict** DT • GBM • **SVM NBC** FB-Prophet Simple moving Average No trend No **Numerical** XGB **VARMAX** Multiclass Classification Seasonality Data? Fb- Prophet **TBATS** Time Series More Advanced techniques **Problem Adaptation** Ensemble **Problem Transformation** <100k Sample Speed, Accuracy Feature Importance **Binary Relevance** Multilabel adopted KNN Classifier Chain Is the data **Label Powerset** Multilabel Classification Multilabel Ridge Speed Accuracy **SGD** Lasso RF DT • Elastic net **SVR** LR- Linear/Logistic Regression, SVM- Support vector, DT- Decision tree, GBM • LR Regression RF- Random Forest, GBM- Gradient boosting machine, SGD- Stochastic gradient descent, XGB- Extreme Gradient Boosting, LGBM- Light GBM, References: 1. https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learningalgorithm-use/ 2. https://scikit-learn.org/stable/tutorial/machine_learning_map/ **NB**- Naïve Bayes, **EM**-Exponential Smoothening 3. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet