Workshop on Machine learning – Day2

ML development for classification Problems

Dayananda Ubrangala
Senior Data Scientist
VMware

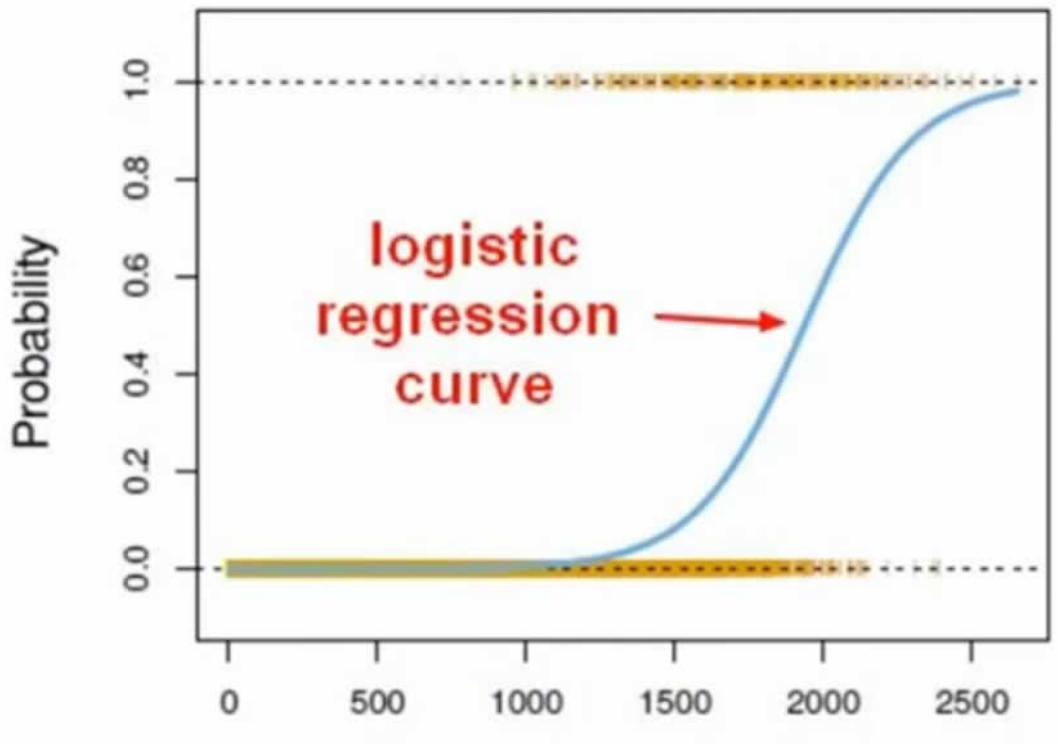
Date: 11th July 2021

Day 2 Agenda

- Model Development
  - Logistic Regression
  - Random Forest Model
  - XGBoost
- Model Comparison and Evaluation

Model Development

Logistic Regression



Machine Learning

Logistic Regression

Logistic Regression is a method for determining whether an entire set of independent variables has any functional relationship to a Qualitative dependent variable

Binary Logistic Regression

Multinomial Logistic Regression

LOGISTIC REGRESSION TERMINOLOGY

ODDS Ratio

MAXIMUM LIKELIHOOD ESTIMATOR

-2 Log Likelihood (-2LL)

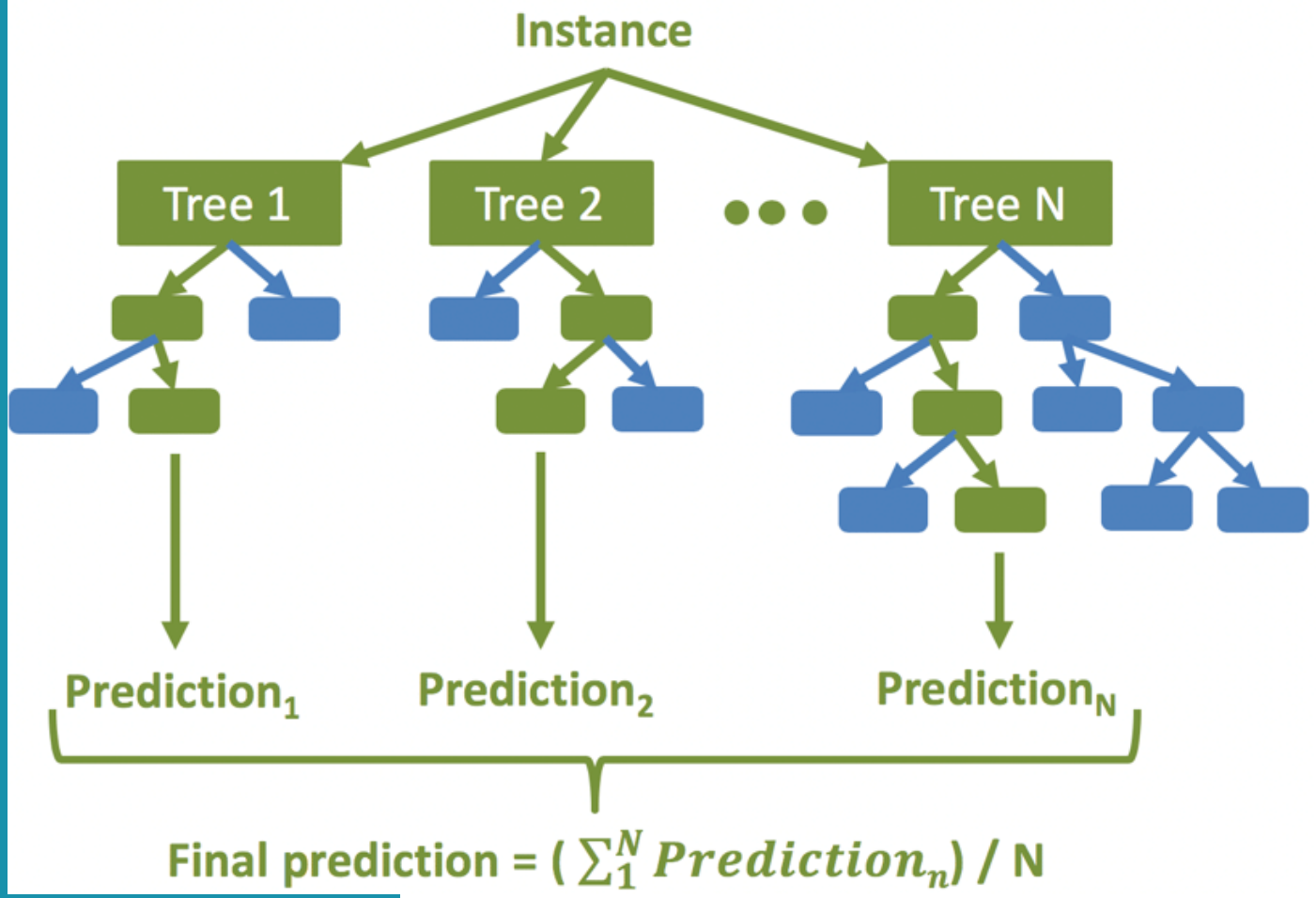
PSEUDO R2

WALD TEST

HOSMER & LEMESHOW TEST

A. Y response variable is RISK level with possible values (High, Medium, Low) for credit card appliers, which will be affected by quantitative variables like age, income, loans, and other qualitative variables also like marital status (yes/no), etc.
B. Y dependent variable is Success/failure of a student in the final exam based on predictors like Age, IQ levels, scores, yrs of education etc.
C. Y dependent variable is existence of Blood Pressure anomalies in patients, affected by age, height, weight, cholesterol levels etc.

Random Forest Model



Random Forest Classifier

What is Random Forest

Random Forest is a method that operates by multiple decision trees during training phase.
The decision of the majority of the trees is chosen by the random forest as the final decision.

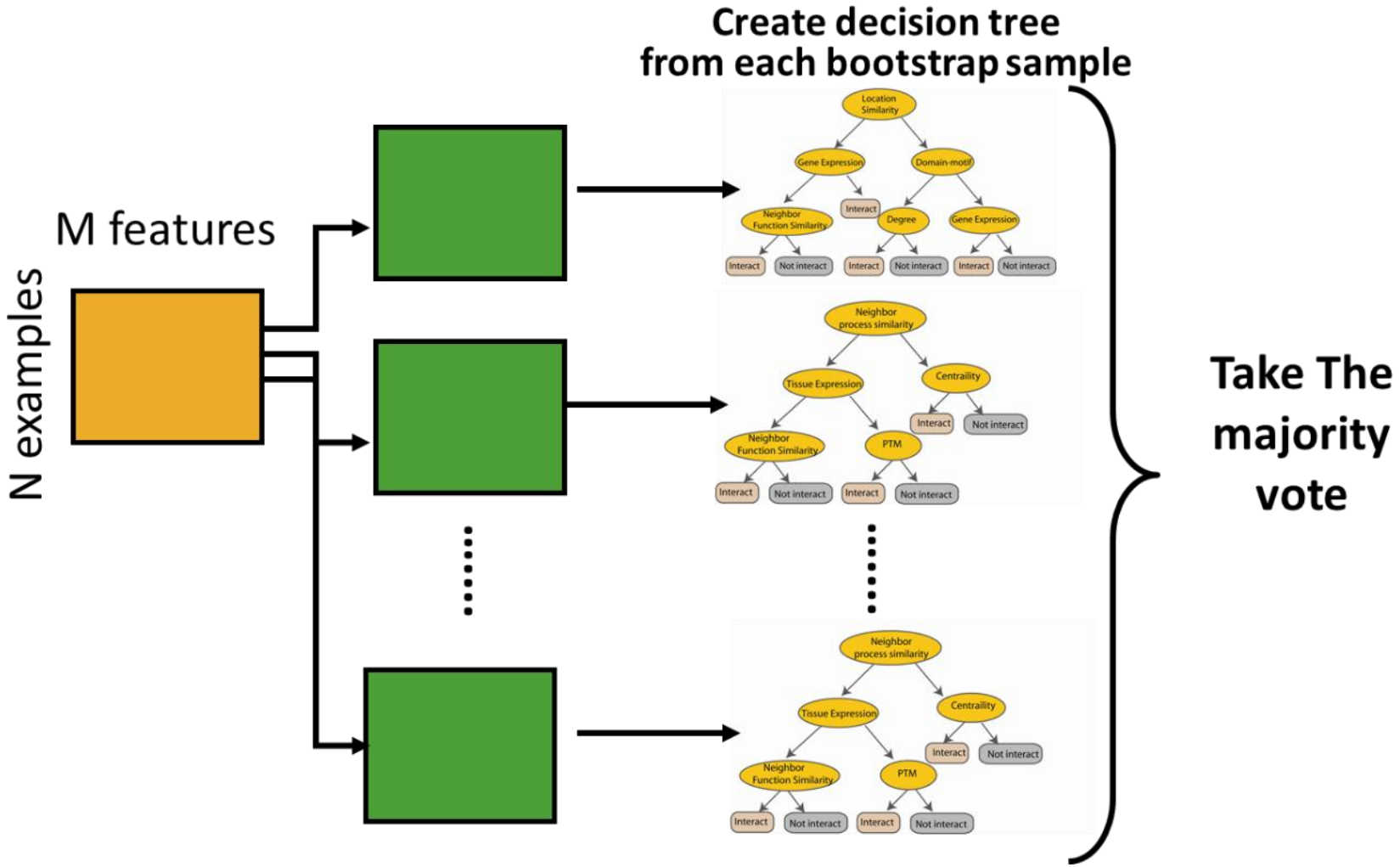
Why Random Forest

- ❑ No Overfitting
  - Use of multiple trees reduce the risk of overfitting
  - Training time is less
- ❑ High Accuracy
  - Runs efficiently on large database
- ❑ Estimates missing data
  - Random forest maintain accuracy when a large proportion of data is missing

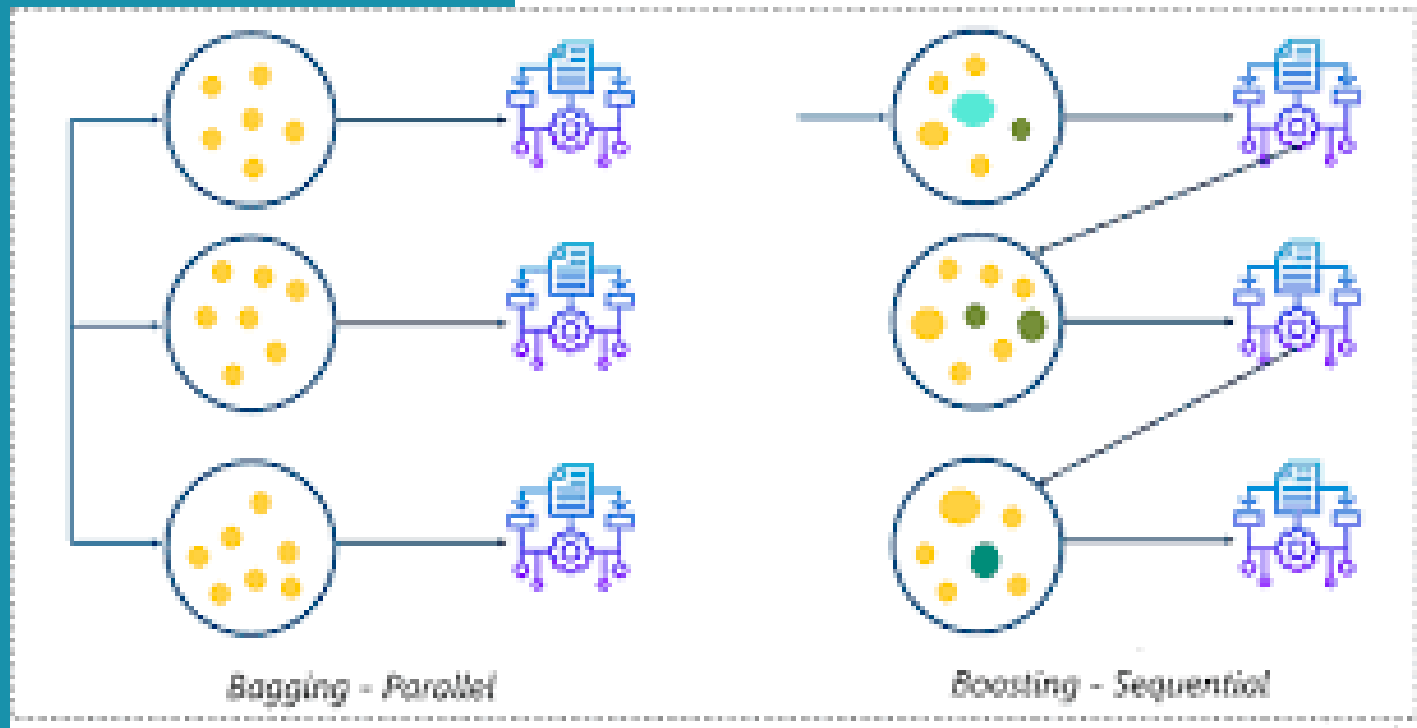
Random Forest Parameter

- n_estimators = number of trees in the forest
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

Random Forest

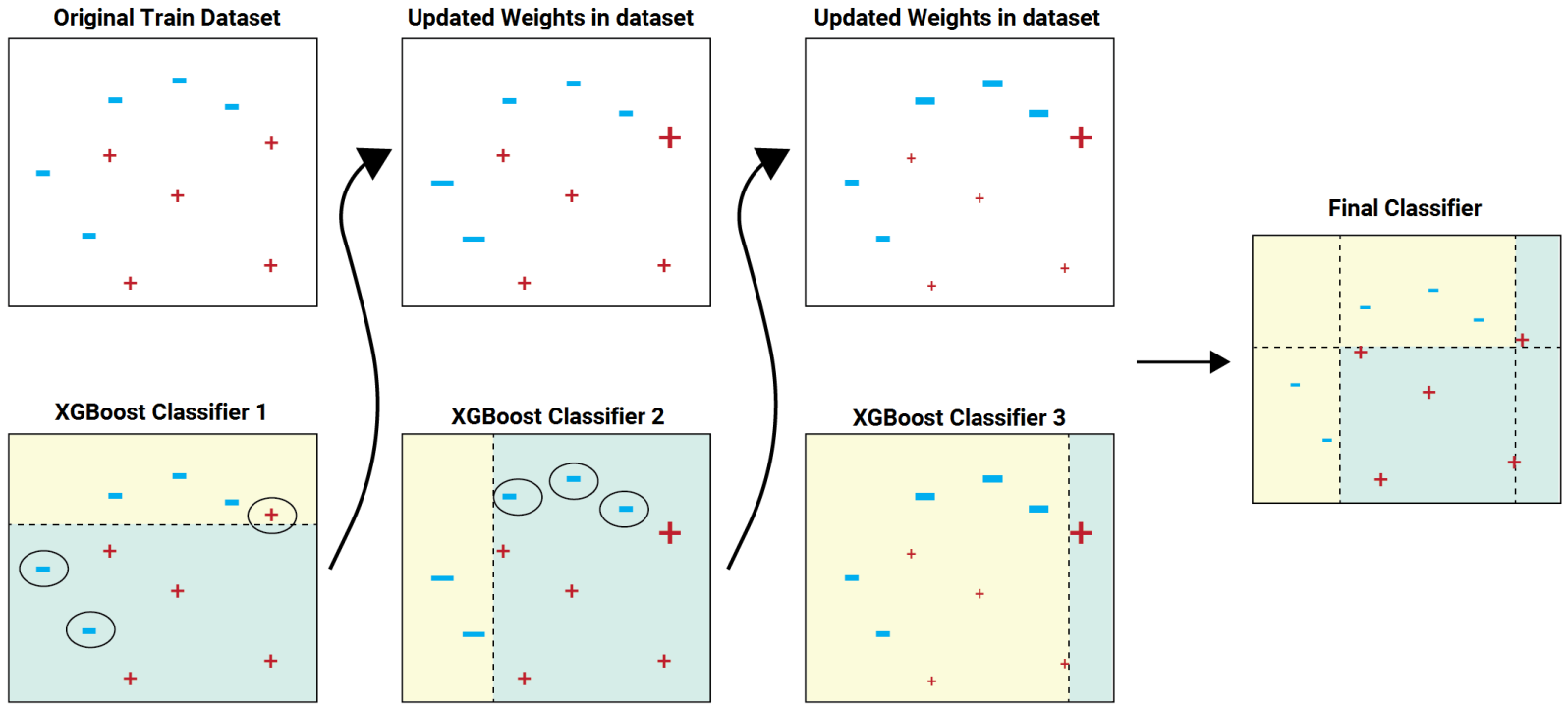


XGBoost Model



XGBoost

- What is XGBoost?
- What is boosting?
- What is gradient boosting?
- Why is XGBoost so good?



XGBoost hyperparameters

Generally, the XGBoost hyperparameters have been divided into 4 categories

General parameters	**booster** **nthread** **verbosity**
Booster parameters	**eta** ; **gamma**; **max_depth**; **min_child_weight** **max_delta_step** **Subsample**; **tree_method** **scale_pos_weight** etc..
Learning task parameters	**objective** **eval_metric** **seed**
Command line parameters	They are only used in the console version of XGBoost

Q&A