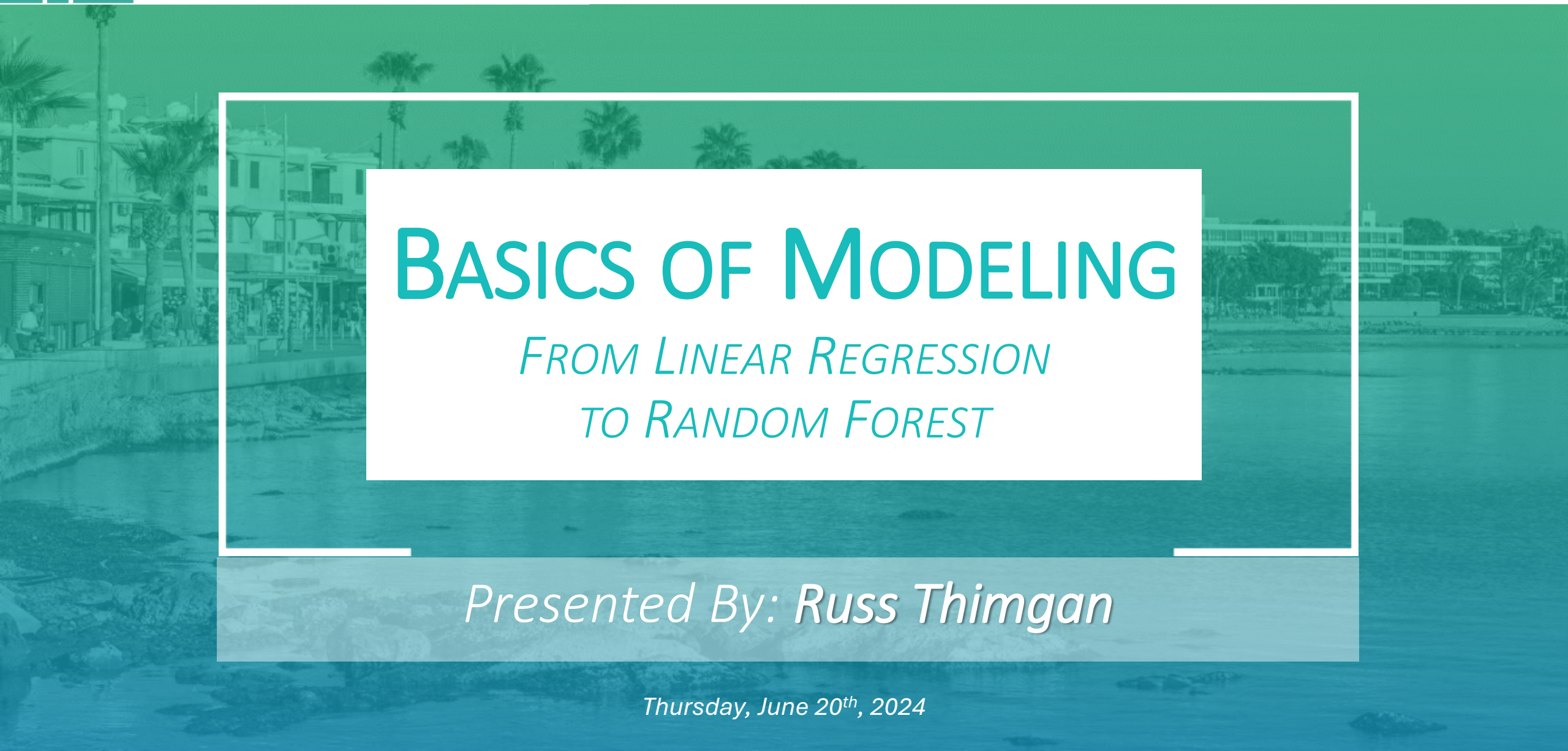


BASICS OF MODELING

FROM LINEAR REGRESSION TO RANDOM FOREST

Presented By: ***Russ Thimgan***

Thursday, June 20th, 2024

Key Speaker

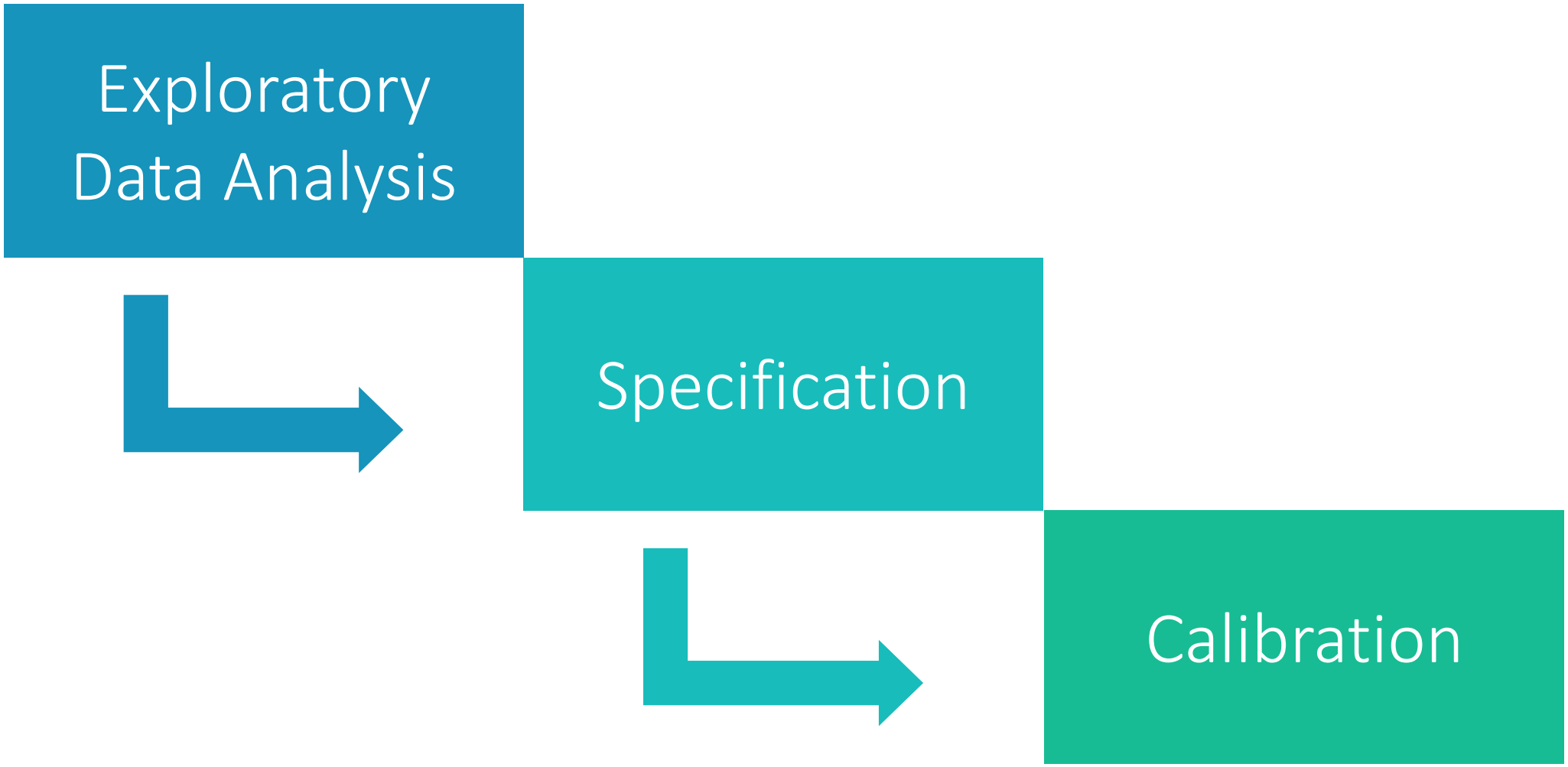
Russ Thimgan

(pronunciation: *timi-yin*)

Mass Valuation Consultant with over 35 years of experience in over 40 jurisdictions in the U.S., Canada, & abroad.

MODEL BUILDING BASICS

TRADITIONAL MODEL DEVELOPMENT PROCESS



WHAT IS AN “AVM”?

AVM stands for “Automated Valuation Model,” and can be broadly defined as a system that uses mathematical modeling to estimate a property’s current or future value.

MODELING TECHNIQUES

LINEAR REGRESSION

MACHINE LEARNING

MODEL BUILDING ASSUMPTIONS

- **Complete & Accurate Data:** All characteristics that influence value have been captured, are correct, and ready to be utilized

- **Representativeness:** Sample data for dependent variable represents the population of properties being valued

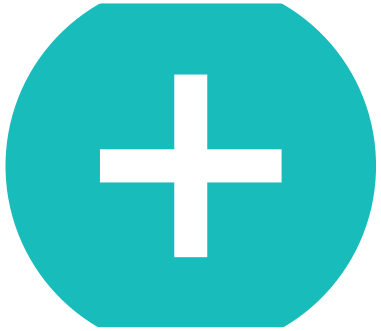
LINEAR REGRESSION BASICS

FIVE MULTIPLE REGRESSION ANALYSIS ASSUMPTIONS

1

Linearity

2



Additivity

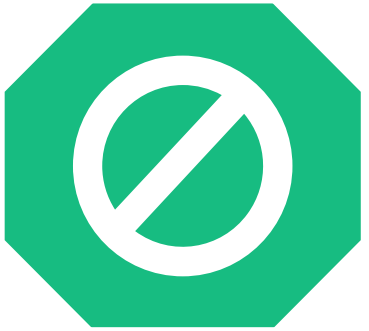
3

Normally Distributed Errors (Residuals)

4

Constant Variance of the Errors

5



Uncorrelated Independent Variables

NON-LINEAR REGRESSION

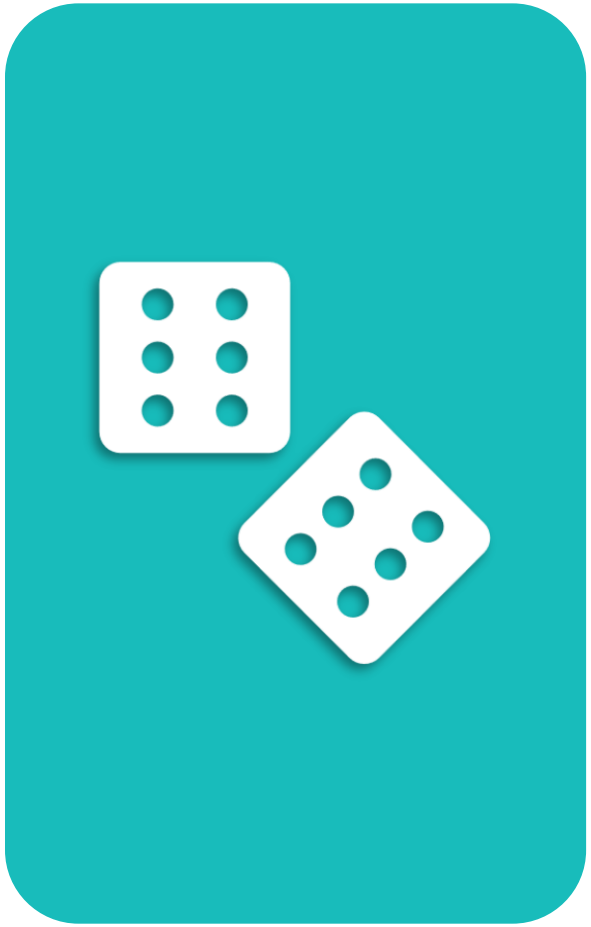
Hybrid Structure

Format: $SP = B_0 + [(B_1 * X_1) * X_2^{B2} * B_3^{X3}] + (B_4 * X_4)...$

LINEAR REGRESSION CONCLUSION

- Linear Regression has a long history in model building and is considered an industry standard
- Natural logarithms improve the models to the point of acceptance as a superior model building tool
- However, this method can violate some, if not all, of the assumptions of regression analysis
- Such weaknesses push modelers to explore alternatives that produce a better product

REGRESSION AND PROBABILITIES



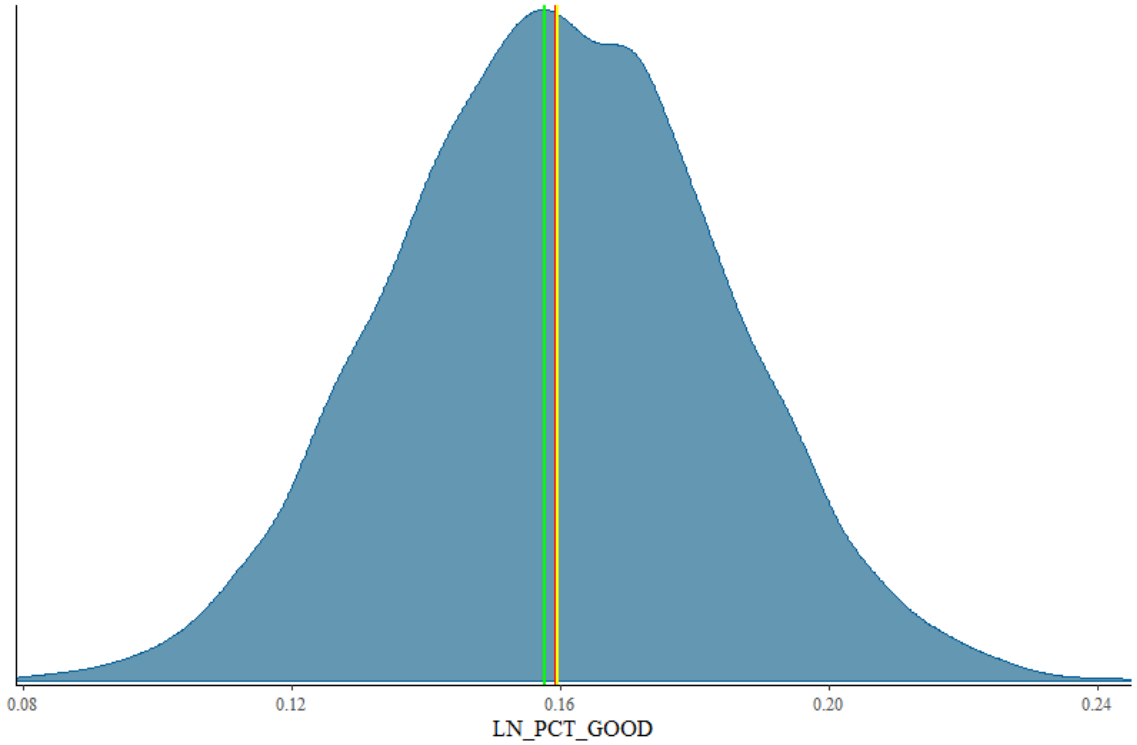
- Regression Model Values are estimates
- The coefficients of the predictors are simply the most probable number the model thinks is the coefficient
- Modelers often employ confidence intervals to determine the probable range of a coefficient

BAYESIAN LINEAR BASICS

BAYES' THEOREM

- $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- $P(A)$ -> Prior Probability
- $P(B)$ -> Evidence
- $P(B|A)$ -> Likelihood
- $P(A|B)$ -> Posterior Probability

BAYES' THEOREM AND BAYESIAN REGRESSION



Provides probabilities to quantify the uncertainty of a hypothesis

- θ *is the population parameter*
- D is some data randomly sampled from the population
- $P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}$
- $P(\theta)$ = how likely is the hypothesis true
- $P(\theta|D)$ = how likely the hypothesis is true now knowing some sampling D

BAYESIAN REGRESSION CONCLUSION

- Linear Bayesian Regression gleams a deeper understanding of the model coefficients
- Results are simple to interpret since they're viewed as probabilities
- However, can be difficult to perform since the choice of the Prior probability can be subjective

MACHINE LEARNING MODELS

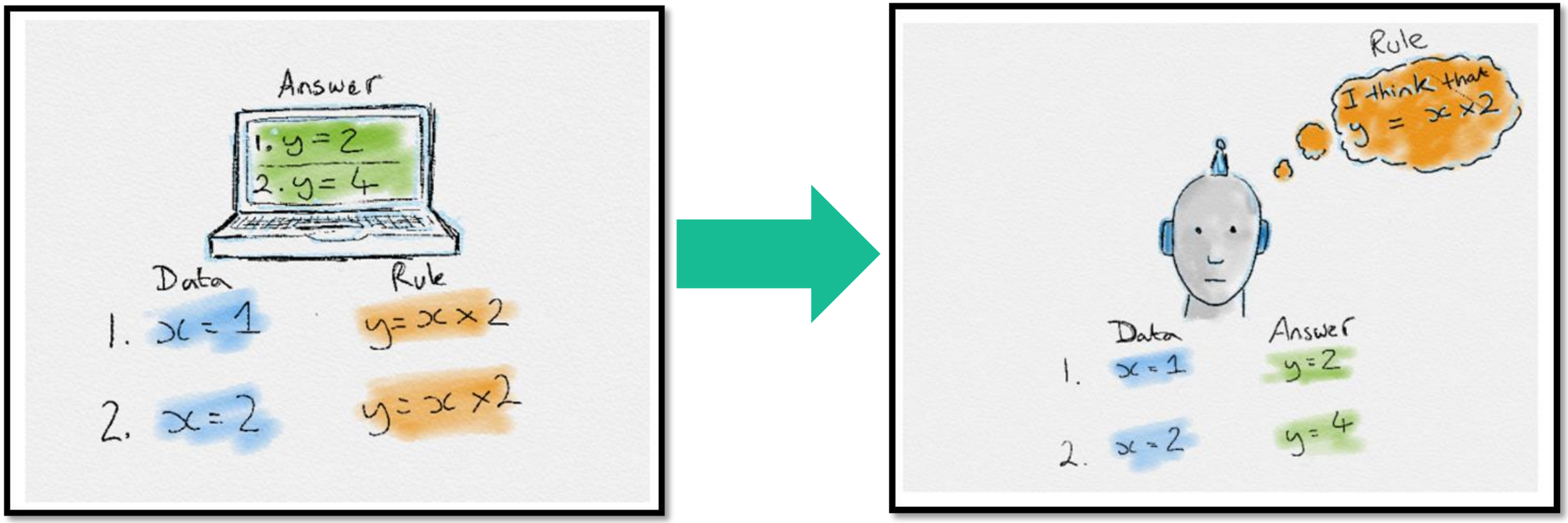
Special thanks to:

Josh Jorgensen

&

Luke Jorgensen

WHY IS MACHINE LEARNING SO EXCITING?



we are stepping away from rule-based systems (specification)

If(x = y): do z

Traditionally, software engineering combined human created rules with data to create answers to a problem. Instead, machine learning uses data and answers to discover the rules behind a problem. (Chollet, 2017)

MACHINE LEARNING MODEL DEVELOPMENT PROCESS

Exploratory Data Analysis

Specification

Calibration

Data Review and Data Engineering

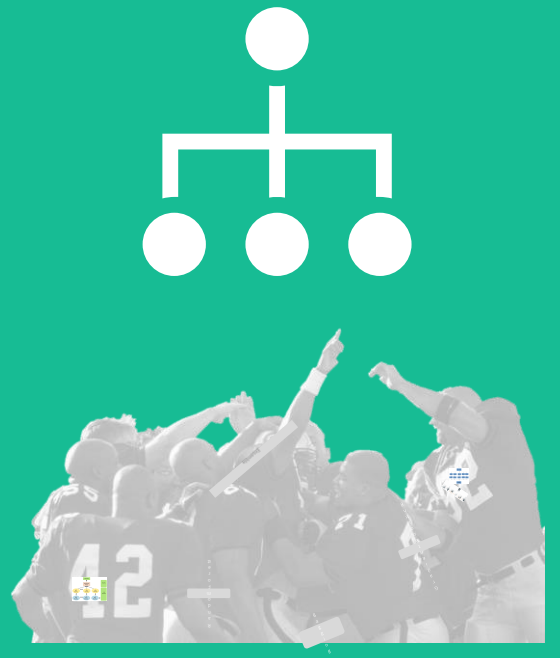
Exploratory Data Analysis → Feature Selection/ Engineering (Prepping data to go into the model) → Model Creation/ Development → Model Review → Model Application/ Deployment → Model Monitoring

SUPERVISED MACHINE LEARNING

Supervised Machine Learning is when human experts act as the teacher

We show the computer the correct answers (output)

From the data the computer is able to learn the patterns

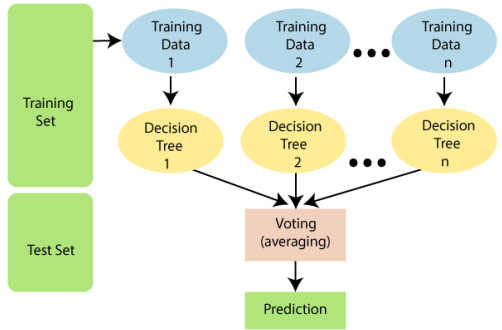


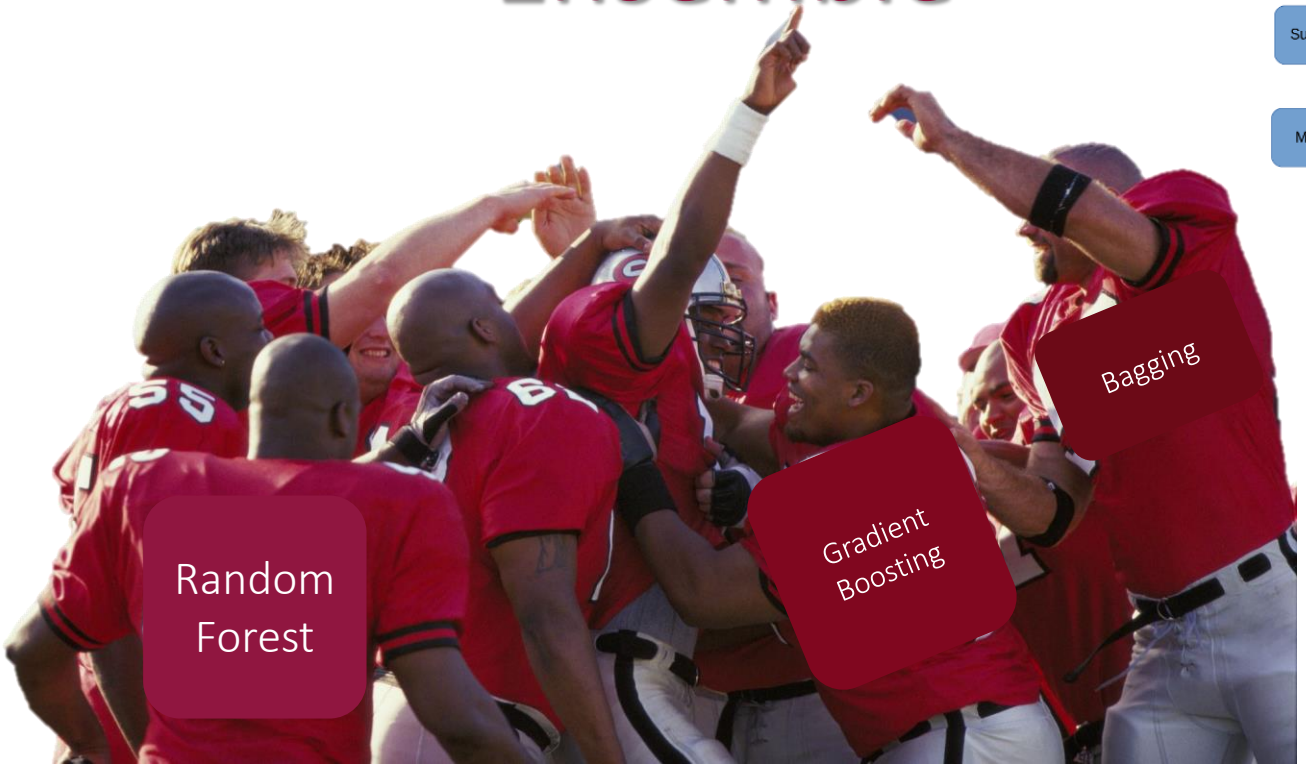
Common learning types

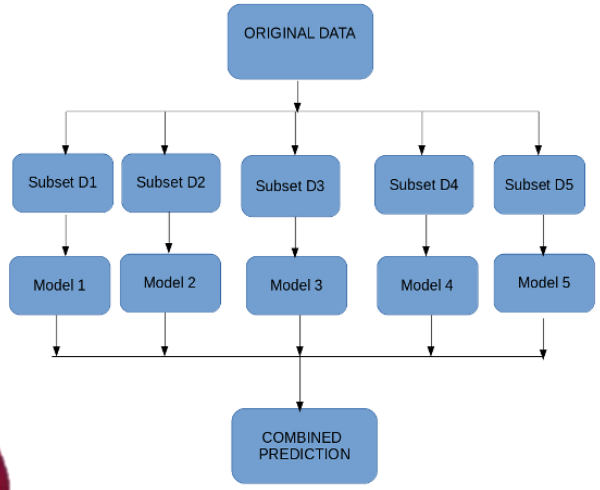
- Decision Trees
- **Ensemble Methods**
- Neural Networks

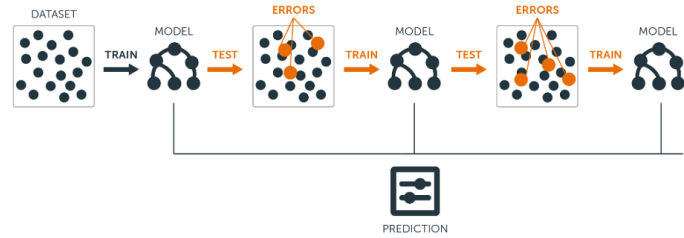
WHY USE ENSEMBLE METHOD?

The Ensemble Method is a machine learning technique that combines several base models in order to produce one optimal predictive model

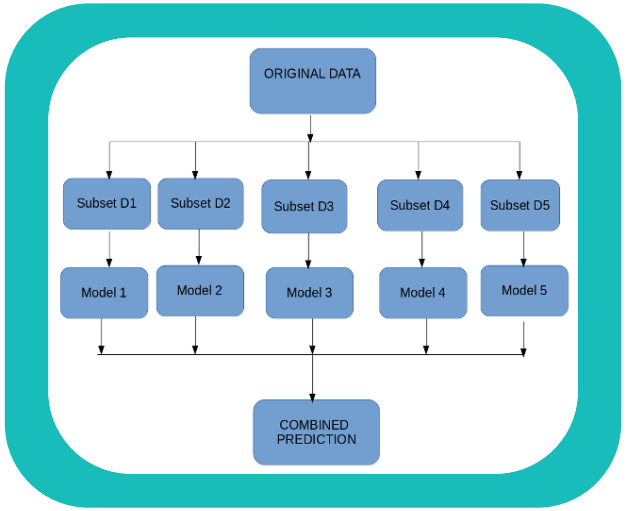




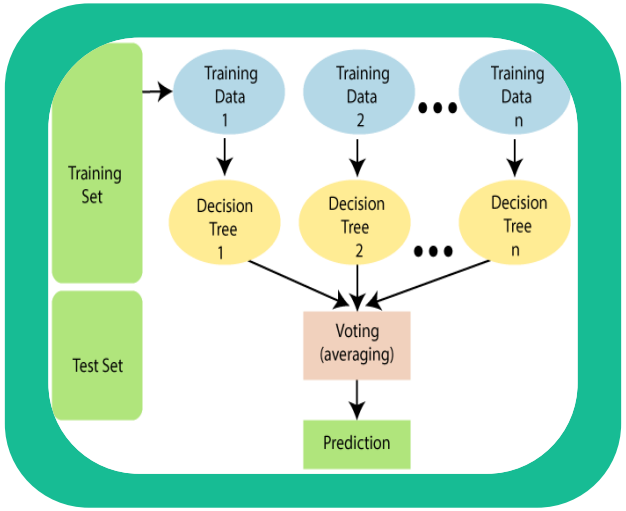




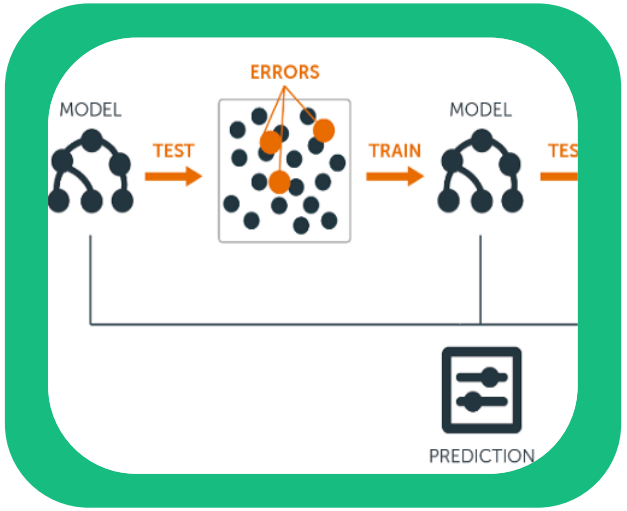
MOST COMMONLY USED TECHNIQUES IN MACHINE LEARNING



Bagging



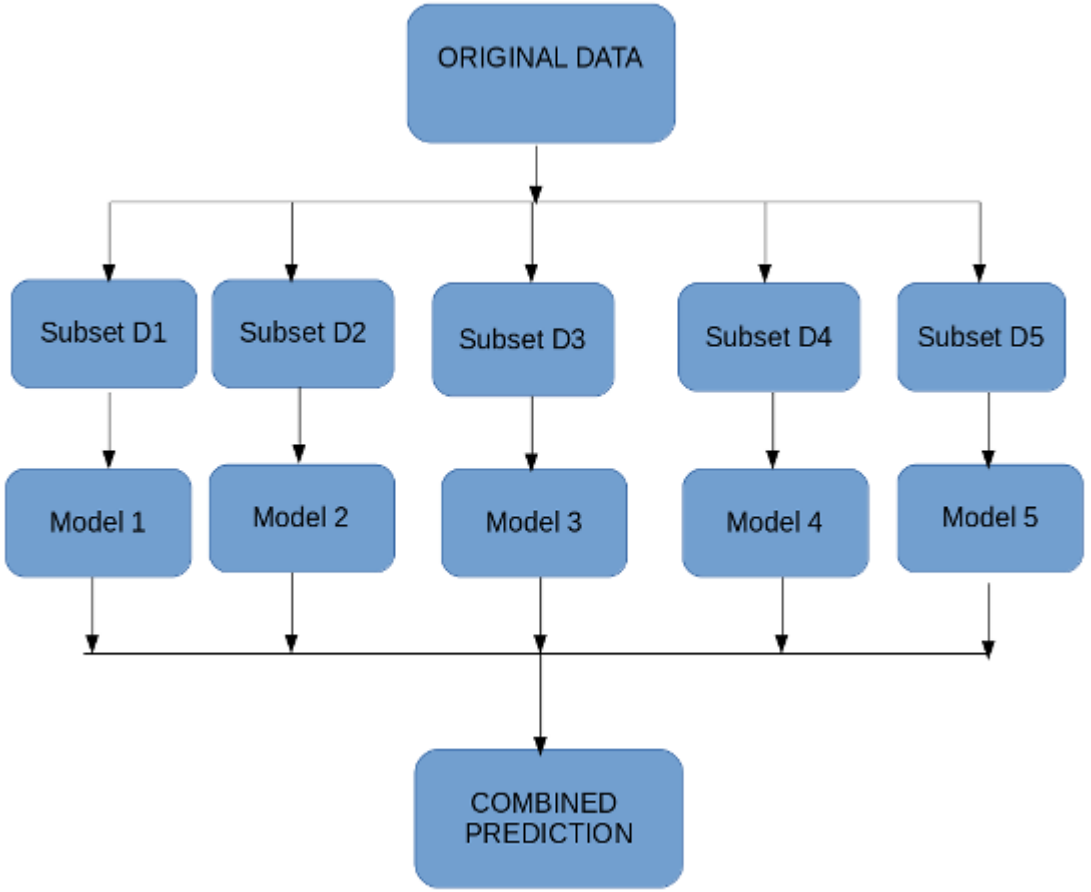
Random Forest



Gradient Boosting

MACHINE LEARNING TECHNIQUES: BAGGING

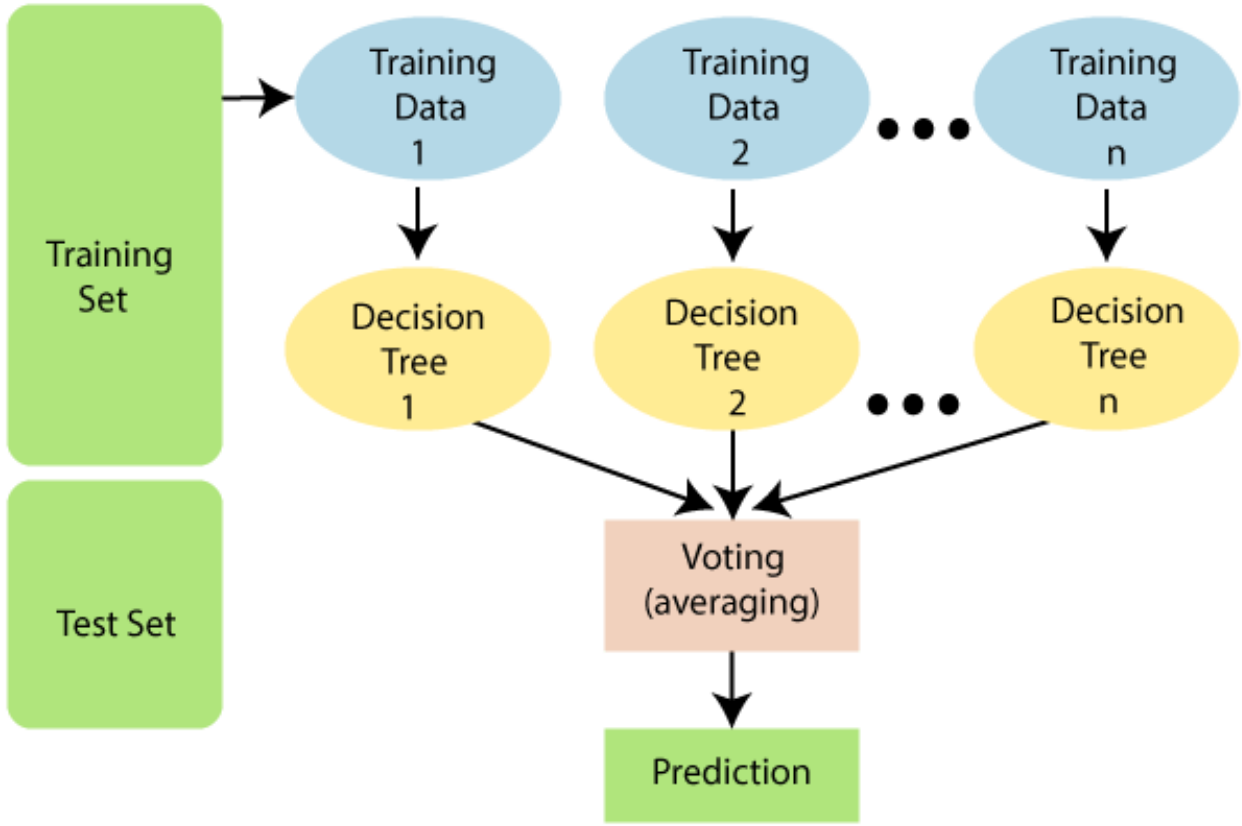
Bagging combines the results of multiple models. To ensure a good result, *bagging* utilizes Bootstrapping, a sampling technique (creating subsets of data to re-run the analysis).



MACHINE LEARNING TECHNIQUES:
RANDOM FOREST

Random Forest, a classifier, contains multiple decision trees on various subsets of the given dataset and takes their average to improve the dataset's predictive accuracy.

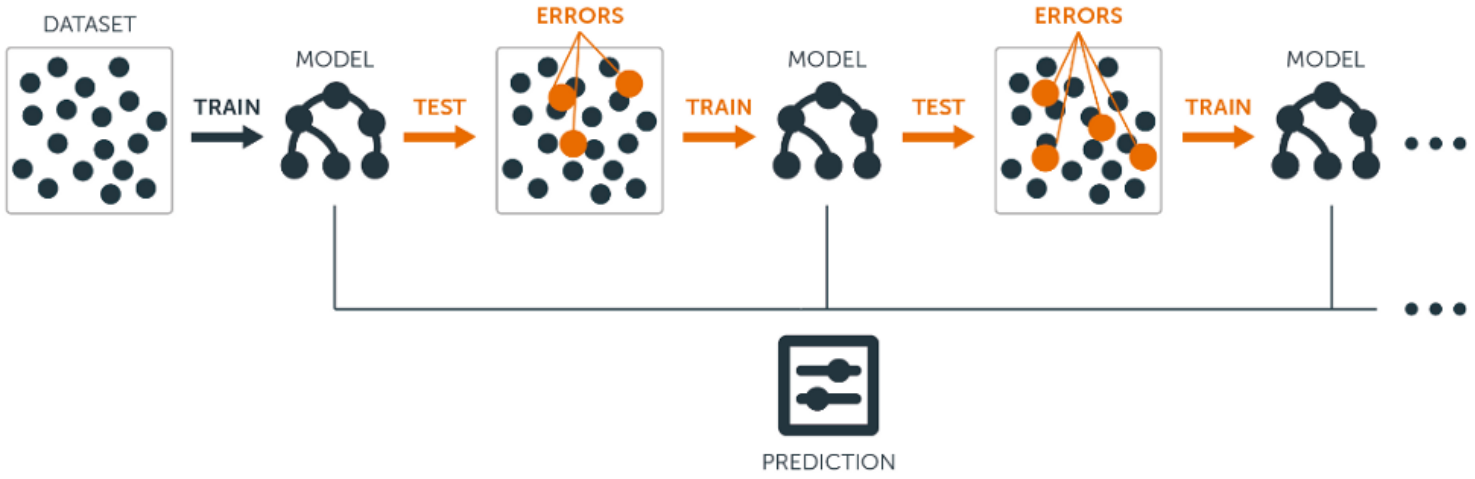
It doesn't rely on one decision tree, instead taking the prediction from each tree, and based on the majority votes of predictions, predicts the final output.



MACHINE LEARNING TECHNIQUES: GRADIENT BOOSTING

Gradient Boosting, or Boosting, is a sequential process where each subsequent model attempts to correct the errors of the previous model.

Each model will contribute to boosting the performance of the overall ensemble

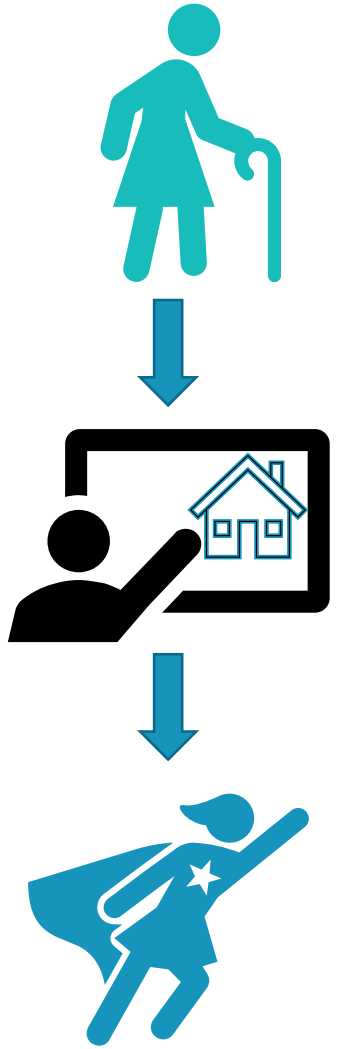


CONTROLLING HOW YOUR MODELS LEARN: THE ART OF HYPERPARAMETER TUNING

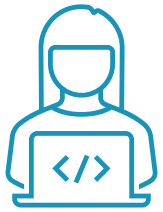
Parameters are what the “model” uses to make predictions

Hyperparameters are what a Machine Learning model learns

- Comes with default hyperparameters, they might not be optimal for your given process
- Models can consist of a multitude of hyperparameters
- Some of the hyperparameters can take on an infinite number of values



CONTROLLING HOW YOUR MODELS LEARN:
CHOOSING HYPERPARAMETERS



Hyperparameters

Parameters

Performance

Max Depth = 5
Sup sample = .5
Number of trials = 250

Weight Optimization

80%

Max Depth = 3
Sup sample = .7
Number of trials = 1,000

Weight Optimization

87%

Max Depth = 8
Sup sample = .9
Number of trials = 2,500

Weight Optimization

92%

- Allows for optimizing your model on any metric; such as R squared, COD, PRD, or any combination
- Adjusting hyperparameters and controlling how your model learns can stop overfitting

CONTROLLING HOW YOUR MODELS LEARN: HYPERPARAMETER TUNING FRAME WORKS

In python there are a magnitude of libraries to assist in tuning models

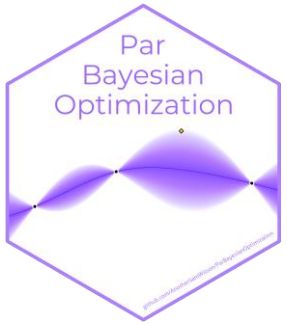
- Optuna
- Hyperopt
- Scikit learn – Grid and Randomized Search

R also supports hyperparameter optimization

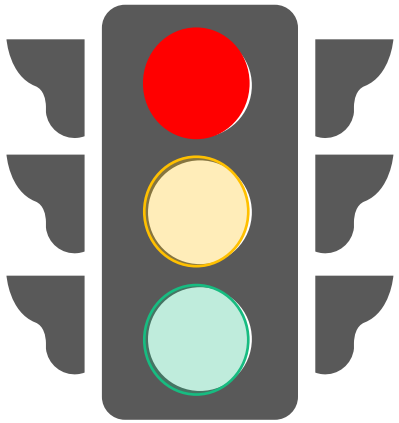
- Bayesian optimization

OPTUNA

HYPEROPT



WHAT STOPS SOME PEOPLE FROM ADOPTING MACHINE LEARNING?



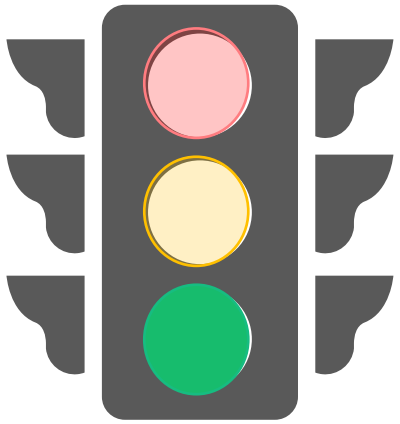
The strength of using algorithms without hard coding fixed rules also creates a weakness

- Explaining how these models work always poses its own set of challenges
- They change dynamically depending on what data point you are predicting; making them harder to explain

Many individuals think of Machine Learning models as a black box

- In turn, people do not trust the predictions provided by the model

HOW DO YOU OVERCOME THIS OBSTACLE?

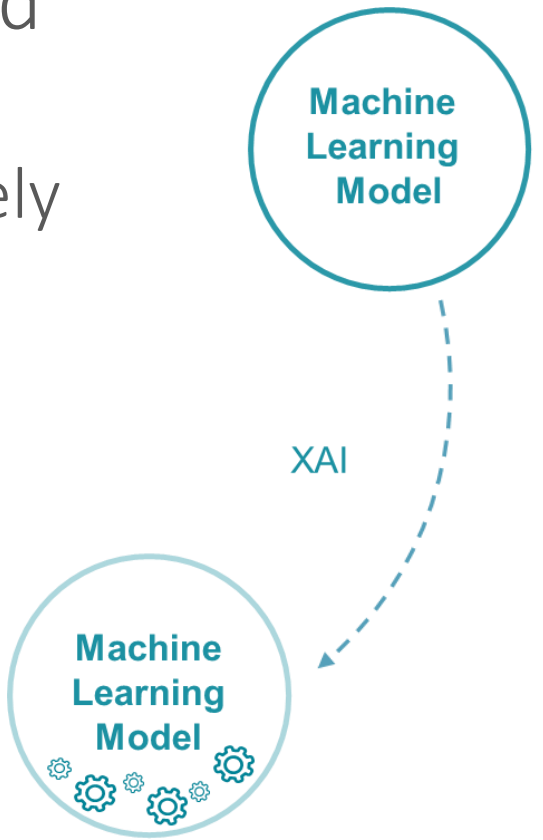


Explainable AI (XAI) helps us to understand how a model is making its predictions

- Ex. What features are positively or negatively impacting the outcome of a prediction?

With Explainable AI, we can answer the following questions:

- Why does the model predict that result?
- What are the reasons for this prediction?
- What are the strongest contributors to the prediction?
- How does the model work?



HOW DO YOU OVERCOME THIS OBSTACLE?

INTRODUCING SHAP

- SHAP stands for SHapley Additive exPlanations
- Is a method for explaining Machine Learning by using concepts of game theory to reverse engineer the output of any predictive model
- SHAP is quantifying the contribution that each feature brings to the prediction made by the model
- Using the outcome of each possible combination of features it determines the importance of a single feature
- SHAP offers unified global and local model interpretability

MACHINE LEARNING CONCLUSIONS

- There are several advantages machine learning models can have.
- Not pre-defined (specified model structure)
- Multiple models utilized to optimize the results
- Machine learning isn't a magic eight-ball for values
- We, as **valuers, can learn** from, well, machine learning!

CONCLUDING REMARKS

THANKS FOR ATTENDING!

Russ Thimgan

russ@prognose.us