Decision Trees | Fundamentals and Implementation

Thursday Learning Hour – Dec 10,2020

Do The Math

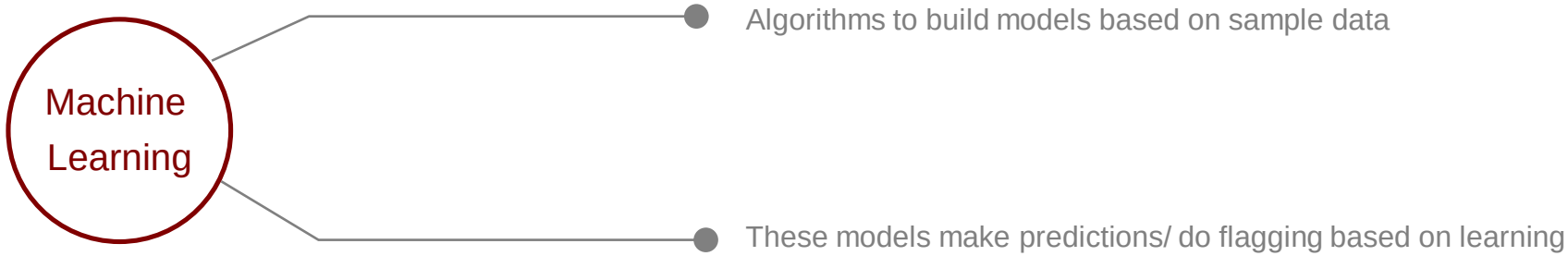
Chicago, IL
Bangalore, India
www.mu-sigma.com

December 11, 2020

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- Applications
- Cross Validation Techniques

ML Introduction



Lifecyle of an ML model

01

Data Selection and Transformation

- ✓ Procuring data
- ✓ Selecting most relevant attributes
- ✓ Log transformations
- ✓ Sales = Units* Price

02

EDA

- ✓ Understanding patterns through EDA

03

Data Standardization and Normalization

- ✓ Bring variables on same scale

04

Selection of an ML Algorithm

- ✓ Clustering/Classifier/Regression
- ✓ Choosing best algorithm

05

Model Training

- ✓ Make the model learn to predict based on train data

06

Model Accuracy

- ✓ % Correct predictions
- ✓ Confusion Matrix
- ✓ MAPE
- ✓ F1 Score

07

Model Selection

- ✓ Selecting best model based on best accuracy

08

Hyper-Parameter Tuning

- ✓ Find the best settings for max accuracy

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- Applications
- Cross Validation Techniques

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- Applications
- Cross Validation Techniques

Supervised Learning

Machine Learning

Unsupervised Learning

- ✓ There is no reference output variable, thus no ground truth
- ✓ Datapoints are “flagged” as one of the homogeneous clusters

Supervised Learning

- ✓ There are input variables and an output variable
- ✓ Model learns to predict output based on existing input-output combinations in train data
- ✓ Historical (train) data can be a ground truth (reference point) for the model

Output is Categorical

Classification

- ✓ Decision Trees
- ✓ Random Forest
- ✓ KNN
- ✓ XGBoost
- ✓ SVM
- ✓ Logistic Regression

Output is continuous

Regression

- ✓ Decision Trees
- ✓ Linear Regression
- ✓ Polynomial Regression
- ✓ Lasso Regression
- ✓ Ridge Regression
- ✓ Elastic Net Regression

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- Applications
- Cross Validation Techniques

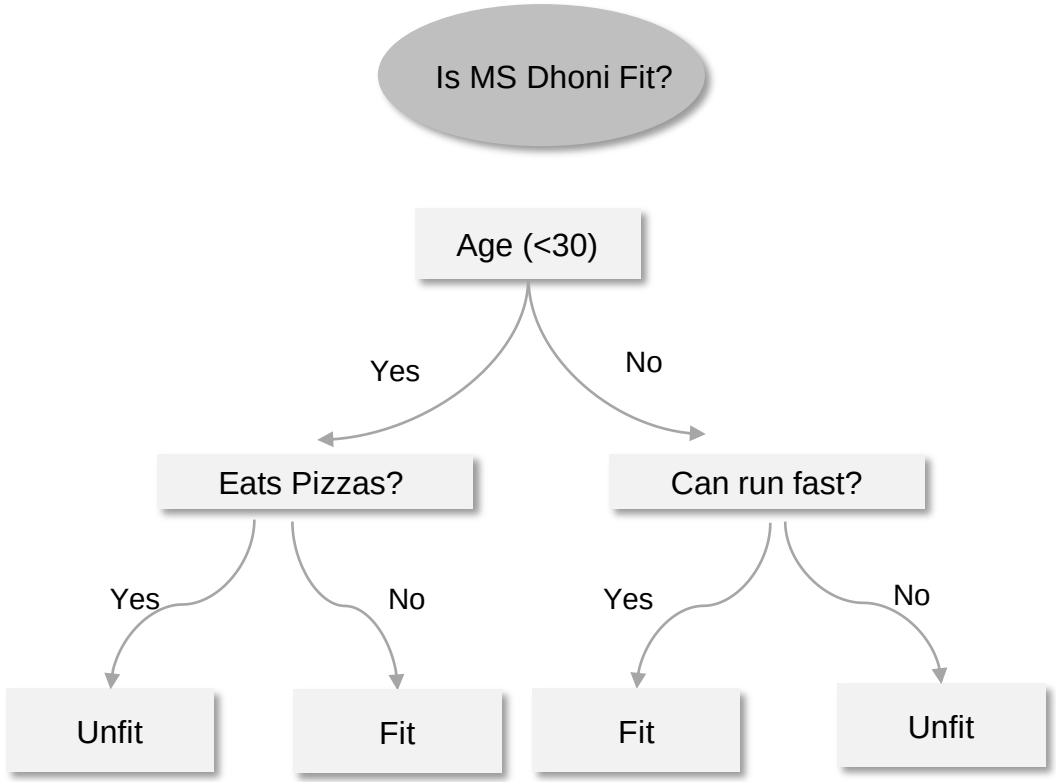
Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- **Decision Trees**
- Applications
- Cross Validation Techniques

Decision Trees | Overview

Divides whole dataset into Tree like structure for taking decisions

Done through Gini Index/ Information gain calculation



01 **Select a Root node**

Pick a variable and condition to split into branches

02 **Root node is split into Child Nodes**

These are called branches, and it is done based on root variable values

03 **Process is repeated till leaf nodes**

Tree can be pruned as well to get desired level of output

04 **In case of regression**

The same process is done, but ChiSquare or Reduction in variance method is used

Decision Trees | Pros vs Cons

Benefits

✓ Can be used for both Classification and regression

✓ Scaling and Normalizing is not required

✓ Easy to understand through visualization through an end-user

Drawbacks

Small changes in data cause instability in the model due to Greedy approach

Overfitting is a common issue

Time consuming

Decision Trees | Gini Index & Information Gain

Independent. Variable	Independent. Variable	**Target Variable**
Class	Gender	Stay in hostel
9	M	Yes
10	F	No
8	F	Yes
8	F	No
9	M	Yes
10	M	No
11	F	Yes
11	M	Yes
8	F	Yes
9	M	No
11	M	No
11	M	Yes
10	F	No
10	M	Yes

Decision Trees Notebook

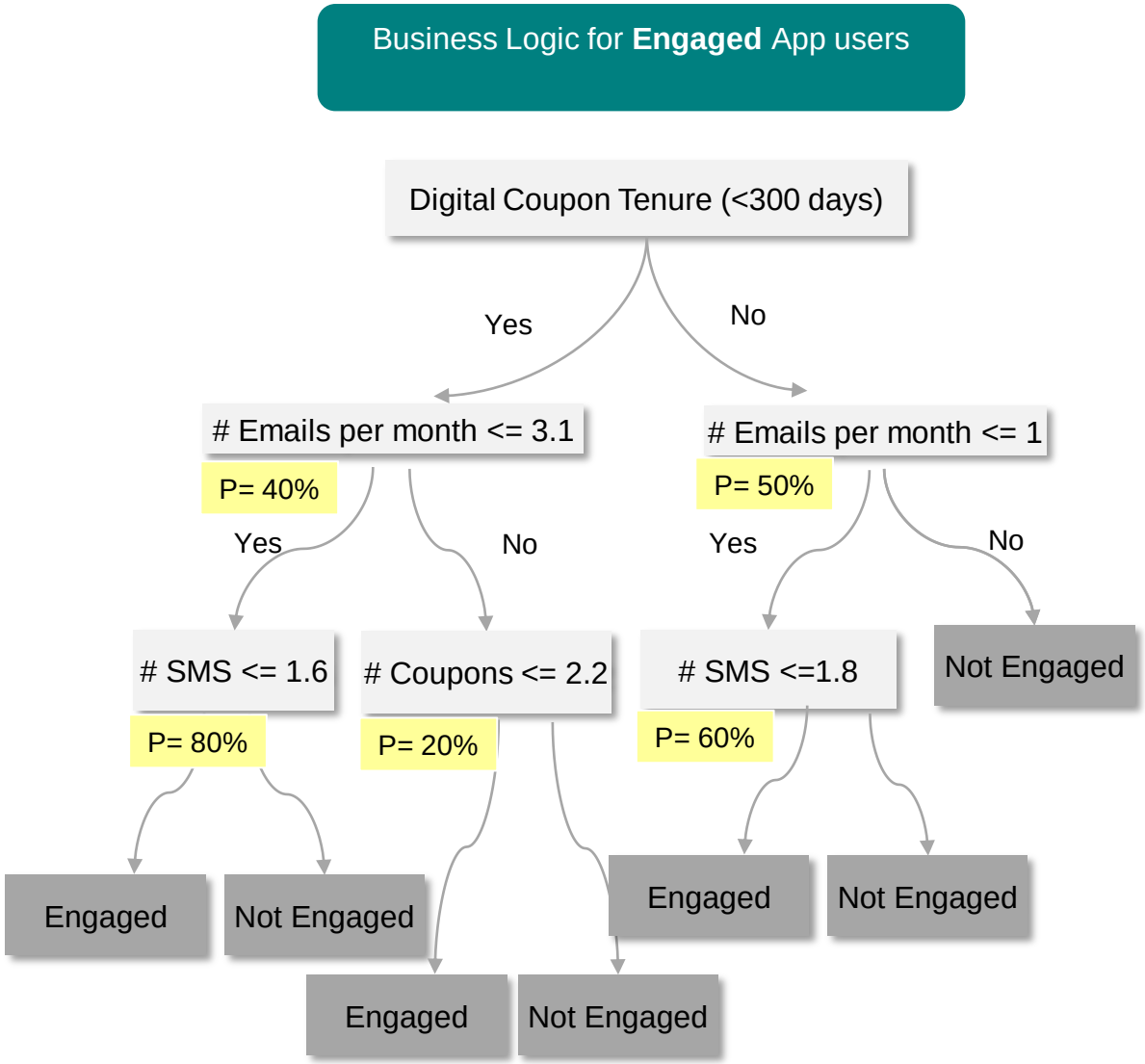
Agenda

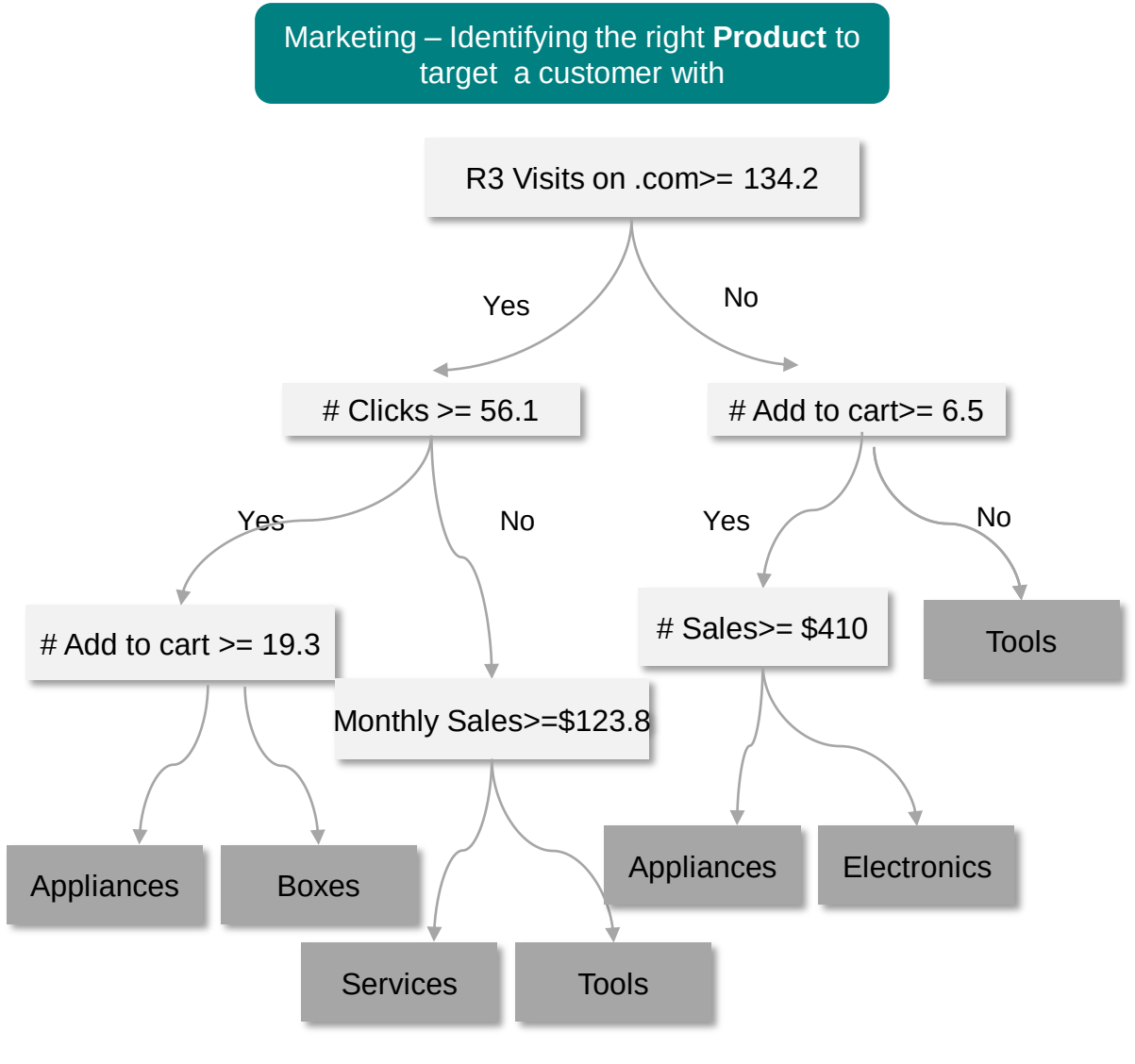
- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- **Decision Trees**
- Applications
- Cross Validation Techniques

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- **Applications**
- Cross Validation Techniques

Applications





Applications

Enhancing Customer Experience

Decision Tree can identify the important factors driving a business goal

Eg- Online – Transactions, Urgency , Offline – Price, # Resources

Energy Consumption

Important factors driving high/low use of energy

Eg- Number of people, Type of household, Income

Identifying Bank Fraud

Based on historical data, frauds can be detected in an early stage

Marketing Campaigns

Identifying which audience to target

Identifying important parameters which drive a campaign goal

Churn Prediction

A very powerful and intuitive process to predict customer churn

Healthcare

To check what all factors lead to a specific disorder

Eg – Symptoms deciding whether the tumor is serious or not

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- **Applications**
- Cross Validation Techniques

Agenda

- Pipeline of an ML Project
- Supervised Learning - Classification vs Regression
- Decision Trees
- Applications
- Cross Validation Techniques

Cross Validation Techniques

Accuracy of an ML model on Train dataset is 95% on Test dataset.

Is it good?

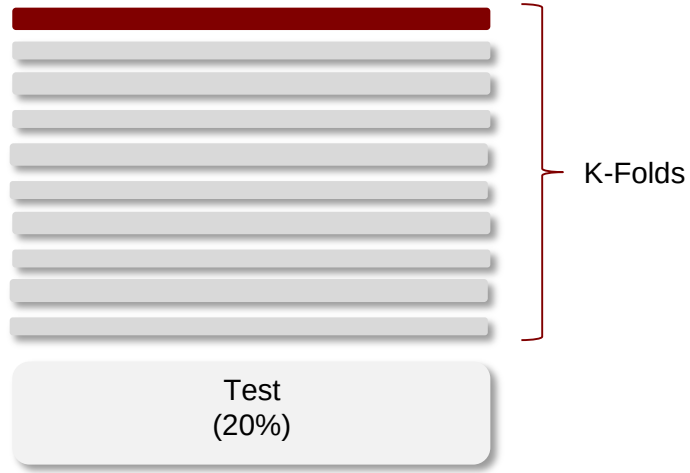
Hold Out Method

Hyper-tuning?

60/20/20

Train data- Only 60%

K-Fold



Q & A