Machine Learning Project

Priyanshu Bansal

2019ume0196@iitjammu.ac.in

Problem Statement

To Find out new covid test cases using regression decision tree from mobility data

1st table 6195 rows and 9 columns: mobility data

2nd table 9332 rows and 59 columns: covid data

What I have done?

- Data Preprocessing
- Build Decision Regression Tree from Scratch
- Compared standard deviation with my implemented tree and sklearn library tree
- Compared result with linear regression
- Improve performance

Data Preprocessing

- Filter data set 1 by Entity=’India’
- Filter data set by iso_code=’IND’
- match their dates in data set 2 because data set one has low data compared to dataset 1 for india
  data_2=data_2[(data_2.date>='2020-02-17') & (data_2.date<='2021-06-01')]
- Take important parameters from data set 1 such as 'Day' ,'retail_and_recreation' ,'grocery_and_pharmacy', 'residential', 'transit_stations', 'parks' ,'workplaces'
- Decompose date into day and month because date is same
- Take new cases from data set 1

Creating Regression Tree from scratch:

Code Flow Chart:

Regression tree: types of node:

Solution:

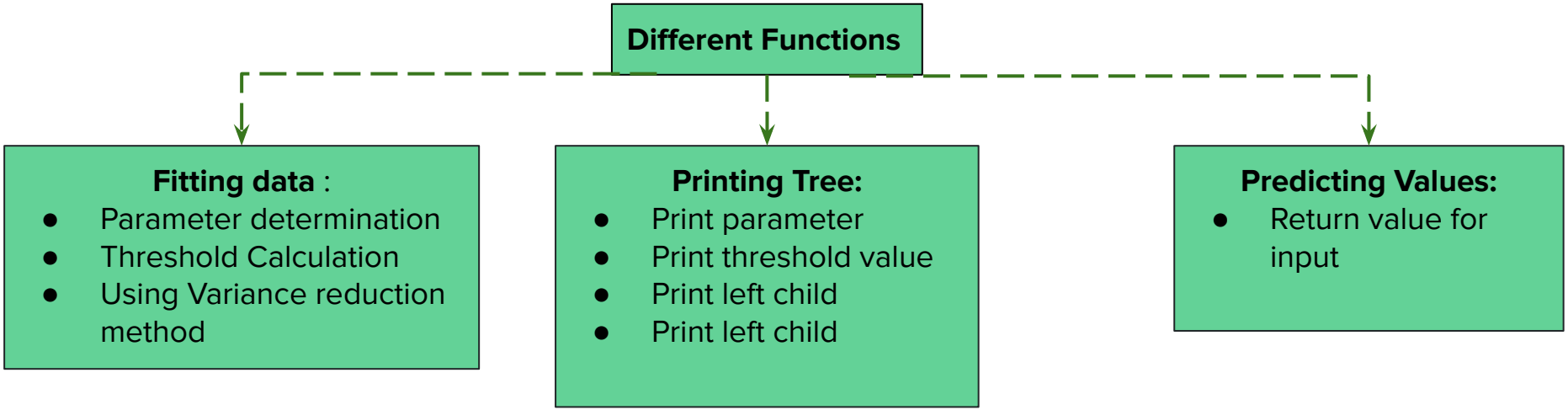
Leaf Node
Store predicted avg value

Internal node
Store at what parameter we need to split and at what point

For Leaf node:
value

For internal node:
Feature Index
Threshold
Variance_reduction
Left_child
Right_ child

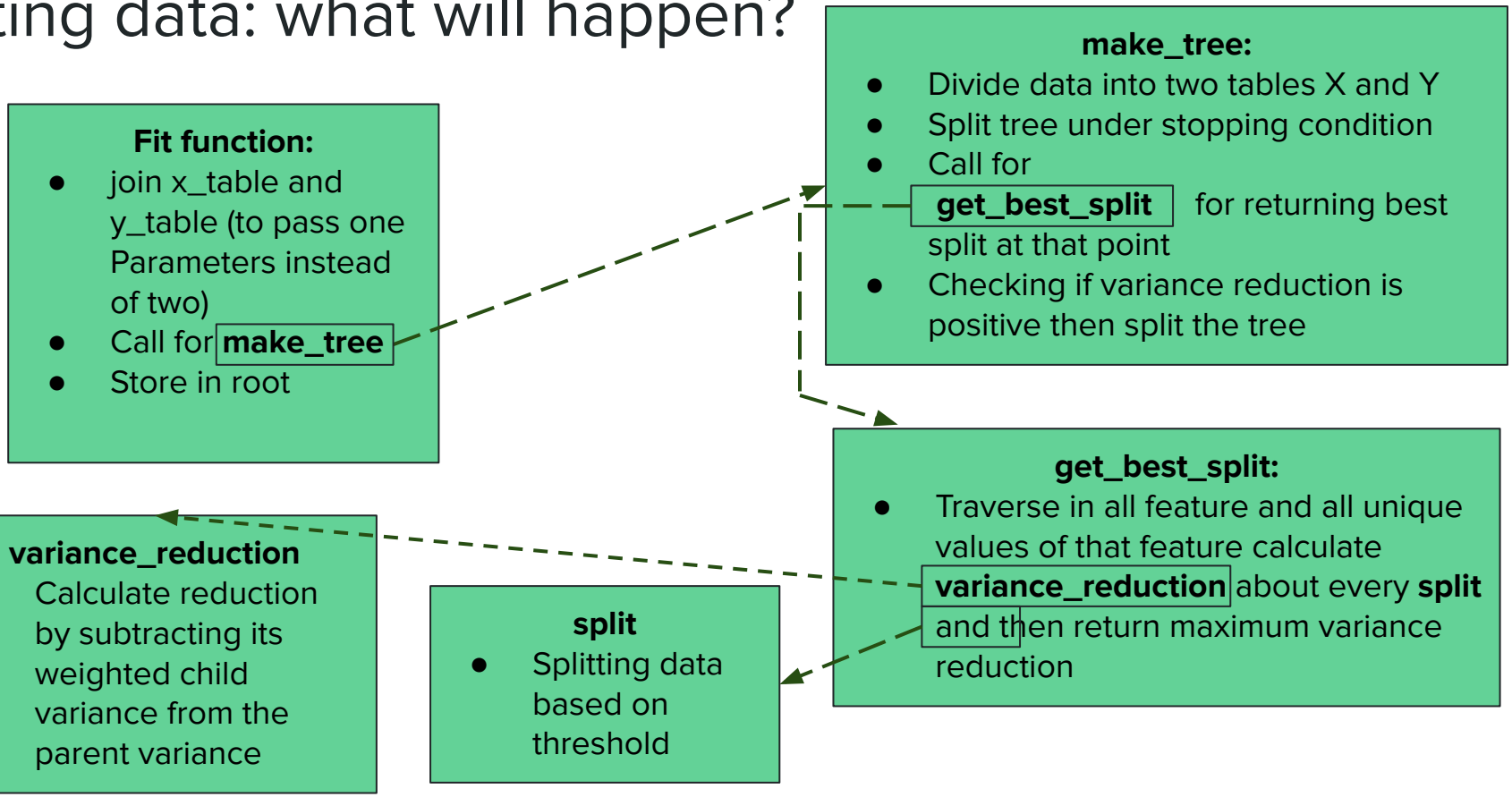
Code Flow Chart:



Fitting Data

- Fit (for fitting data)
- Make Tree (will make structure within parameters)
- Get best split (find best split for given parameters)
- split (for splitting data set about value and parameter)
- variance_reduction (for calculate variance reduction)

Fitting data: what will happen?



Printing Tree

Function : print_tree

- Recursively traverse in the tree
- pre -order traversal
- Printing parameter and threshold
- If value is not NULL print value at the leaf

Predicting Values

Functions:
predict, modal_prediction

- **predict :**
  - Iterate over input and call model_prediction for each value
  - Store all values in list and return

- **model_prediction :**
  - For one instance calculate output from design regression tree

Applying model:

Comparing performance

Model	Our model	Sklearn decision tree_regressor	sklearn linear regression
Standard deviation For training dataset	17232.6068451	20954.822678	58413.39641851
Standard deviation for testing dataset	36585.70565	24450.087227	68835.84764099
performance	Well performed	Excellent	poor

Improving performance

For improving performance we apply the fact that covid will affect after 14 days so we just simulate this thing and improved our model

Final improved model:

Comparing performance

Model	Our model	Sklearn decision tree_regressor	sklearn linear regression
Standard deviation For training dataset	13520.97994586	20268.97	60659.10985067
Standard deviation for testing dataset	20699.43457	20463.653	63018.47001069
performance	Improved from previous	Excellent from previous	Poor from previous

Thank you for listening.....