greatlearning
Learning for Life

Project on Ensemble Methods

Problem statement

- We have data from a Portuguese bank on details of customers related to selling a term deposit
- The objective of the project is to help the marketing team identify potential customers who are relatively more likely to subscribe to the term deposit and this increase the hit ratio

Deliverables

greatlearning
Learning for Life

Exploratory data analysis	Address data challenge	Prepare the data for analytics	Create the ensemble model	Improve accuracy and recall
• Univariate analysis • Mutlivariate analysis	• Data pollution • Outliers • Missing values	• Load data • Scale, transform, normalize	• Classification model • Various ensembles	• Key hyperparameters • Regularisation techniques • Try out the different models to get to the maximum accuracy and Recall.

Learning Objectives

- Data Description
- Attribute information
- Steps to follow
- Insights based on feature importance
- Findings & recommendations
- Model selection
- Conclusion

Data dictionary

greatlearning
Learning for Life

Bank client data	Related to previous contact	Other attributes
• 1 - age • 2 - job : type of job • 3 - marital : marital status • 4 - education • 5 - default: has credit in default? • 6 - housing: has housing loan? • 7 - loan: has personal loan?	• 8 - contact: contact communication type • 9 - month: last contact month of year • 10 - day_of_week: last contact day of the week • 11 - duration: last contact duration, in seconds*	• 12 - campaign: number of contacts performed during this campaign and for this client • 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign • 14 - previous: number of contacts performed before this campaign and for this client • 15 - poutcome: outcome of the previous marketing campaign • Output variable (desired target): 21 - has the client subscribed a term deposit?

* Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Data Description

Goal

https://archive.ics.uci.edu/ml/datasets/bank+marketing#

Using the collected from existing customers, build a model that will help the marketing team identify potential customers who are relatively more likely to subscribe term deposit and thus increase their hit ratio.

The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

Steps to follow

- Import the libraries
- Get the data
- Count the instances of each class in the data to check if data is skewed towards a class
- Loop through all columns in the dataframe
- Only apply for columns with categorical strings
- Replace strings with an integer
- Split the data into training and test set
- Instantiate decision tree as the default model.
- Look at the class level scores for the overfit model.
- Use random forest which gives us ensemble instances which are very dissimilar.
- Try various bagging and boosting methods and compare model performance scores

greatlearning
Learning for Life

HappyLearning

Questions?