Data Science with Python

Dr. Gopal Kumar

gopalkumar@iimraipur.ac.in

31 May 2020

Machine Learning

Features

Target

Org_type	Partner	Age	Designation	Supplier_gc	Customer_gc	Motivation_1	Operation_type
Manufacturer	Manufacturer	between 3	Manager	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer	between 3	Manager	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer	between 3	Others	Yes	Yes	4	Food
Manufacturer	Manufacturer; Se	<=30	Others	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer; Se	<=30	Others	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer; Se	50 and abo	Vice President/ F	Yes	Yes	5	Pharmaceutical Industry
Manufacturer	Service Provider;	between 3	Manager	Yes	Yes	2	Electricals
Service Provider	Service Provider	between 3	Others	No	Yes	4	Others
Manufacturer	Manufacturer	<=30	Others	No	No	3	Automobile industry
Manufacturer	Manufacturer	between 3	Manager	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer	<=30	Manager	No	No	4	Metals
Manufacturer	Manufacturer	<=30	Consultant/Analy	Yes	Yes	5	Food
Service Provider	Manufacturer; Se	<=30	Director	No	No	4	Others
Service Provider	Manufacturer; Re	between 3	Manager	Yes	No	4	Electricals
Service Provider	Manufacturer; Se	between 3	Others	No	No	4	Others
Service Provider	Service Provider	<=30	Others	Yes	No	2	Automobile industry
Wholesaler/Distri	Manufacturer; Re	50 and abo	Manager	Yes	Yes	5	Others
Manufacturer	Manufacturer; Se	<=30	Manager	Yes	Yes	4	Automobile industry
Manufacturer	Manufacturer; Wh	<=30	Others	Yes	Yes	2	Automobile industry
Service Provider	Service Provider;	<=30	Others	Yes	Yes	5	Software Industry
Manufacturer	Manufacturer; Se	between 3	Manager	Yes	Yes	3	Automobile industry
Service Provider	Manufacturer; Wh	50 and abo	Others	No	Yes	4	Software Industry
Wholesaler/Distri	Manufacturer; Se	<=30	Others	Yes	No	2.089715758	Others
Manufacturer	Manufacturer		Manager	No	Yes	3	Others

X_train

y_train

X_test

y_test

Classification

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Classification

- The goal is to predict a class label (categorical variable)
- Binary and multiclass classification

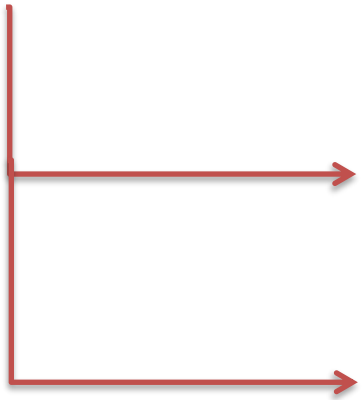
Yes/no

Spam/not spam

Rain/no rain

Classification

- The goal is to predict a class label (categorical variable)
- Binary and multiclass classification



Website language (Hindi/English/Spanish)

Irish example

Classification and Regression

Regression

- The goal is to predict a continuous number
- Predicting income from education, age, location, etc.
- Predicting yield of rice farm from previous yields, weather, number of employees working on the farm

Types of learning

- **Supervised Learning**
  - Learning based on training data + desired outputs (labels)
  - Classification and regression

- **Unsupervised Learning**
  - Learning based on training data (without desired outputs (labels))

Types of learning

- **Supervised Learning**
  - Logistic regression
  - Neural Network
  - K-Nearest Neighbors (KNN)
  - Support vector machines (SVMs)
- **Unsupervised Learning**
  - Clustering

Linear classification algorithms

- Logistic regression
- Support vector machines (SVMs)

Logistic Regression

It estimates Relationship Between

- ❑ (one) binary dependent variable and
- ❑ independent variables

It computes probability of occurrence of an event.

- ❖ Regulation parameter, C, controls the tradeoff between model complexity (overfitting) and model simplicity (underfitting)
- ❖ If the C increases, the model becomes more complex.

Logistic Regression

$$P(y_i = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

- $P(y_i = 1 \mid X)$ *is* the probability of the *ith* observation's target value, y_i, being class 1
- X is the training data
- β_0 and β_1 are the parameters to be learned.
- If $P(y_i = 1 \mid X) > 0.5$, *class 1 is predicted, otherwise 0 is predicted.*

Bank marketing campaigns dataset

- Portugal bank marketing campaigns results
- Conducted campaigns were based mostly on direct phone calls, offering bank's clients to place a term deposit.
- If client had agreed to place deposit - target variable marked 'yes', otherwise 'no’

Task

- predicting the future results of marketing companies
- formulating recommendations for such companies in the future

Bank marketing campaigns dataset

Features

1. **age** (numeric)
2. **job** : type of job (categorical: 'admin.','blue-collar', 'entrepreneur', 'housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. **marital**: marital status (categorical: 'divorced','married','single', 'unknown'; note: 'divorced' means divorced or widowed)
4. **education:** (categorical: basic.4y','basic.6y','basic.9y', 'high.school','illiterate','professional.course','university.degree', 'unknown')
5. **default**: has credit in default? (categorical: 'no','yes','unknown')
6. **housing**: has housing loan? (categorical: 'no','yes','unknown')
7. **loan**: has personal loan? (categorical: 'no','yes','unknown')
8. **contact**: contact communication type (categorical: 'cellular', 'telephone')

Bank marketing campaigns dataset

Features

9. **month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. **day_of_week**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. **duration**: last contact duration, in seconds (numeric).
12. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. **previous**: number of contacts performed before this campaign and for this client (numeric)
15. **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Bank marketing campaigns dataset

Features

16. **emp.var.rate**: employment variation rate - quarterly indicator (numeric)
17. **cons.price.idx**: consumer price index - monthly indicator (numeric)
18. **cons.conf.idx**: consumer confidence index - monthly indicator (numeric)
19. **euribor3m**: euribor 3 month rate - daily indicator (numeric)
20. **nr.employed**: number of employees - quarterly indicator (numeric)
21. **y** - has the client subscribed a term deposit? (binary: 'yes','no')

Thanks!