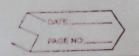
PART 1 - Data Preprocessing.
misorkflow with a more with a more with the most
Data Preprocessing
· import data
· clean duta -X-Important step IRL.
o splitinto train & test sets.
Modelling 201000 (il toring)
Build Model of Alles Delegan - page 19
Train Model and - dillipposition
Make Predictions 100 111 - 25/119 (2)
THE STATE OF THE S
Evaluation
· calculate performance matrix
. Have resplict.
He purious transfer and the land to the same and the same
Train-Test split. and an internal langui o
usually 80/20 percent
. Used trained model on tech
set that has known results
to compare with models result.
Feature-Scaling
Remembra (always applied to columns & never on
As one a section the stress smortalets to process
(1) Normalization x' = X - Xmin [Most values blo
Xmax-xmin /
where x is the value of column
X min is lowest & Xmax is higher
2). Standardization x1 2, x-4, where yis mean!
o tondar alzarium a average
[+3,-3] extremes or dutiess. It is std. deviation



- with unscaled features, the lower valued/unit column become irrelevant & make small difference to large scale values.

Practical.

allows to import a library or function or any module,

(1) impost librasies

Onumpy -works with arrays for inputs.

- 2) matplotlib for charts/graph/visualizations
- (3) Pandas import dataset / create matrix of features

o Libraries - ensumble of module containing modules &

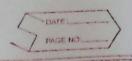
Tallas

- · impost numpy as no
 - we can call with no woow.
 - o impost matplotlib. pyplot as plt.

 interested in this anodate.
 - o impat pandos as pd.
- (2) import clataset
 - (1) dataset = Pd. read-CSV (Himname. extension).
 - · create a data frame with all values in data set.
 - (3) create 2 new entities
 - * features the column with which we are going to predict independent dependent variable. of dataset

* dependent variables - last column, that need to be predicted

fulltonn - index locates feature of pandas, allows us to take Index of columns I rows we wann a extract * x = dataset. iloc, [:, : -1]. values takes all rows (in pythonit means a range & with out LB OXUB means all all call except lost). * 42 dataset_iloc[:,-1 Handling Missing Values-Method I. ? Just delete the row of missing value. - works good for large datasets. with low y- of rovissing data. Hethod II: Replace missing value by average of all data in column. Scikitlearn - has lot of tools & preprocessing tools for this task. * we use a module called impute in scikitlearn & a class import the simple Imputer from it. imputer simple Imputer (missing values opnon, Now, this object class has functions that will be used to find missing values then input it; excluded in python (3-1)s imputer tit (x [:, 1:3] imputer, transform (walve 1:37) & returns new updated matrix of



So to change	oxiginal	as It is have	teature roatoix
· ×(::	1:37 =1	mouter trons	tom (x[:,1:3]).
o senso proces	(1)		

Encoding Categorical Data.

It will be difficult to co-relate rompute these features with categoring string clase, Thus we need to turn these categories to numbers.

Idea Helhod T: Gniode France to O Spain to 1 Germany to 2

model may interpret as order /numeric order & interpret as the order mutters.

To-DO: Furn Turn to one hot encoding i.c. turn the category into total columns with total no of distinct category.

France spain & germany we ge for France. Spain germany

Encode InDependent variable

. For this we use 2 classes

(i) column transformer class from composed mochele of scikit

(3) one hot encodes class from preprocess module of scikit what kind of transformation on cohich indexes

C+ 2 columnTranstormes (transformers remainder encoding reansformaling

be applied for * transformers = [('encoder', 'onettot Encoder'), transformations COI) indexes to transform benede

Leep columns as it is to count apply transformation.

DATE:

remainders 2' passthrough

If we don't apply the remaining columns wont be added.

with this class, we can fit and transform at once with function we can fit and transform.

don x = Ct. fit_transform (x)

need to convext it to np array

we call x = np-array (c+-fit_transform (x))

3 Encode Dependent Variable.

Scikit

o we use label encodes class from preprocessing from

(b) = Label Encoder ()

value, so its obvious what

otal filas gu

needs to be performed

4 c 161. fit transform (y).

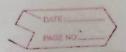
doesn't need to be numpy array since

(5) Splitting Dataset Into Train & Test set.

X' we apply feature scaling after splitting data.

because we only use train set for Mi model while test set represents seal would date. I shouldn't be attends

Et just needs to be test to model be get offe.



Also when applying feature scaling we use mean & standard deviat that might leak some into to train set which isn't supposed to happen. The ML model will get into on test set & may fit itself for it. - We scale test later after splitting & scaling trainset. Fort For this will use sklearn module that has. model = selection class & with train-test_split fn. · We split into 4. 2 feature matrices 8-· 2 depend variables. expects first teature mat x-train, people, y train, y test = train-test - split (x, u expects dep.var ~ 2 more args ~ spart or 1 test_size z 0.2 int type seed for random value.

(c) Feature Scaling (Not necessary for all models).

works well all the time-

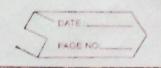
2 techniques () standardization (with mean(x) en deviation(x).

(2) Normalization (with Amin)

distribution among an our teatures.

(recommended for specific situations).

from sklearn preprocessing import standard scales



SC = Standard Scaler () = automatically applies Standardisation

· Pummy variables -> variable columns after encoding categorical data.

We don't need to apply standardisation to dummy variables since they are already blu (-3,3) or (0,1)

1 First we fit our scales on training sets. so x-train(00,3:) 2 8c-fit-transform (x-train C:,3:])

Next we tit x-test.

- we only apply transform method here since this data is like new data received IRL by model.

- The features of test set need to be scaled by the same

scalex used on training set.
- we do not calculate a new fit. we got that from fit_transform earlier x_test[:,3:] = sc-transform (x_test[:,3:])





random-state = 0)

SUHHDRY:
The state of the s
1) All Libraries top preprocessing
· import pandas as Pd. [General]
· II numpy CIG TP
· from Skleam-impute import simple Imputer (for Hissing)
" Il sklearn compose " columntantomes (for encoding)
· 11 Skkam. Preprocessing 11 OneHotEncoder caregorical
o 11 11 LabelEncedet (tabel/dependent)
11 Skleam model-sciection 11 train-test-split (turn splitting)
· 11 Skleam-prepagessing 11 StandardScalex (for Scaling).
(3) To import dataset (and assign dep. & indep variable).
dataset : Pd. read-Esy (' ·/loadfikname')
x = dataset-iloc[:,:-1]
4 : " [:,2]
The state of the s
3 Handle Hissing data
imp & Simple Impure (missing-values + np-nan, strategy 2 mcan)
imp. fit (x[:, 1:3]) [] first fit tocompute
x[:,1:3] > impurex. transform (x[:,1:3]) = 8" frontom to replace
A LITTER OF THE PROPERTY OF TH
(5) Encoding Categorical Data.
1 Indep. variable.
rolly: ColumnTransformer (transformers: [('emoding', oneHortmoder(),
[0])], remainder z 'passttmough')
* x z np-array (witr + fit - transform (x))
(2) Dependentale

ye Ibenc. fit transform (y)

Split Data

X-tr, X-tst, Y-tr, Y-tst. train-test-split (2,4, test-size:0.2,

Scaling (only when necessary)

Scaling (only when necessary)

Scaling (shandardScaler()

x-train (shandardScaler())

y-train (shandardScaler())

y-train (shandardScaler())

y-train (shandardScaler())