- Download Data for linear models practice exercise from LMS [under the module linear models]

Linear Regression Practice Exercise

1. Read data [Facebook comments] to python using pandas
2. Create pipeline to process data [these are some of the suggestions , you independently decide what you would like to do with rest of the variables]
    1. Create dummy vars for column page_category with frequency cutoff 200 .
    2. For columns 'Post Published Weekday'and'Base Date Time Weekday' replace ['Sunday', 'Monday'. . . . .] with [1,2, ...]
    3. Instead of creating dummies for datetimetype columns its better to represent them with values which are cyclic in nature themselves . Create sin and cos columns for both the columns mentioned in (above) as follows :

    df[col_sin]=np.sin(2*pi*df[col]/7)

    df[col_cos]=np.cos(2*pi*df[col]/7)
3. Build simple linear model for the data , check its performance using cross validation [target = Comments_in_next_H_hrs]
4. Build linear regression model using lasso , use parameter tuning to find appropriate value for hyper parameters . check how many coefficients are made zero.
5. Build linear regression model using ridge , use parameter tuning to find appropriate value for hyper parameters
6. Linear regression outcome, theoretically can take negative values as well, how can you ensure that your outcomes are positive

Logistic Regression Practice Exercise

Logistic Regression Exercise

Data dictionary is as follows (ignoring the first column which is id): This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

1. Read data 'default of credit card clients.xls' [downloaded and unzipped from LMS, module : linear models]. Use function pd.read_excel from package readxl. skip first row while reading the data
2. Build a pipeline to process data, base your decisions [or experiment with them] on the data dictionary given above.]
3. default payment next month is your target , build a tuned logistic regression model. [score using roc_auc]
4. Find the cutoff on the basis of F beta score (beta=2) [For hard class prediction from the predicted probabilities]