Regression Problem

1. Download Data Facebook_comments.csv zip file and unzip from LMS (module : linear models)
2. Read data to python using pandas.read_csv
3. For column 'page_category', get a list of categories where frequency is higher than 200
4. For the categories found in (3), create dummies in the data
5. Remove column 'page_category' from the data
6. For columns 'Post Published Weekday' and 'Base Date Time Weekday' replace ['Sunday','Monday'.....] with [1,2, ...]
7. Instead of creating dummies for date time type 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 (6) as follows :

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

8. Remove columns 'Post Published Weekday' and 'Base Date Time Weekday' from the data
9. Break data into two parts (80/20) randomly
10. Build a simple linear regression model for ''Comments_in_next_H_hrs''. Use cross validation to check its performance. check its performance on the 20% data [Mean Absolute Error]
11. Build Linear Regression model with L2 penalty . Use Gridsearch to find best value of penalty and its cross validated performance. check its performance on the 20% data [Mean Absolute Error]
12. Build Linear Regression model with L1 penalty . Use Gridsearch to find best value of penalty and its cross validated performance.

check its performance on the 20% data [Mean Absolute Error]

13. How many features were removed from the model when you used L1 penalty

Classification Problem

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. use option skiprows to ignore first row
2. Create dummies using pd.get_dummies for following columns : Gender, Education, Marital status . Add them to data and drop the original columns
3. Break the data into two parts (80/20)
4. Use Randomised Grid Search to build a classification model using logistic regression with best parameters . Check its cross validated performance . Check its performance on 20% data [AUC score]. [Parameters to tune : C , penalty]
5. How is the performance different when you use class_weight='balanced'