# **Approach : Analytics Vidya - Jobathan Aug - 22**

By - Aman Panchal ( [panchalaman2000@gmail.com](mailto:panchalaman2000@gmail.com))

LinkedIn - <https://www.linkedin.com/in/aman-panchal/>

Link of the code -Link

Steps:

\* **\_\_Step 1: Importing the Relevant Libraries\_\_**

\* **\_\_Step 2: Data Inspection\_\_**

\* **\_\_Step 3: Data Cleaning\_\_**

\* **\_\_Step 4: Exploratory Data Analysis\_\_**

\* **\_\_Step 5: Building Model\_\_**

\* **\_\_Create a Submission file\_\_**

Step 1: Importing the Relevant Libraries

1. Import all libraries :

Matplotlib, Pandas, Numpy, Seaborn,LabelEncoder,OneHotEncoder,

Metrics, train\_test\_split

Step 2: Data Inspection

2. Read the csv

* Get shape : ((1888, 22), (762, 21))
* Read column names
* Check null values exist or not : Not exist
* Check using describe() and info() :
* As subject\_len, body\_len & mean\_paragraph\_len has higher std number than others so we has to handle this
* is\_timer column has all value zero so we can drop this column.
* Check categorical feature exist : times\_of\_day
* differentiate categorical and numerical columns

Step 3: Data Cleaning

3. Create useful function

* drop\_columns():
* drop the unnecessary columns  
   ['campaign\_id','is\_timer','times\_of\_day']
* replace\_1():
* out of 1888 is ‘price’ column has 1875 value is 0, so changing value that if it is >0 than it is 1

If value = 0 than replace it with 0 else with 1.

* clean\_df():
* Call drop\_columns() function
* Apply replace\_1 on ‘is\_price’ column
* Standardise [‘subject\_len’,’body\_len’,’mean\_paragraph\_len’] to decrease values range to another column.
* dummy():
* One hot Encode ‘times\_of\_day‘ column and combine with the original dataframe.

Step 4: Exploratory Data Analysis

4. Use value\_counts() for [‘times\_of\_day’,’sender‘] to know that data is

balanced or imbalanced.

* check that ‘sender’ column has the same value in train and test csv.
* check ‘times of day’ column that is there any value which is repeating.

5. Create a heatmap of correlation of all columns.

To check which column affects the result most.

6. Draw countplot for ‘times\_of\_day’

Most for ‘evening’ time

7. Draw barplot for ‘click\_rate’ vs ‘times\_of\_day’

So during the morning the click rate is higher.

Step 5: Building Model

* Separate Features and Target for X and Y
* Train test split : 20% data as validation set
* Use ['Linear Regression', 'Ridge Regression', 'Lasso Regression',

'K Neighbors Regressor', 'Decision Tree Regressor','RandomForestRegressor','SVR','XGB','gb regressor']

* Algorithm and create list of rmse error and mae error on test data.
* In which Random Forest, KNeighborsRegressor, xgb, GradientBoostingRegressor perform well
* During model selection use cross\_val\_score to understand that the model generalizes well or not.
* Use GridSearchCV to find best hyper parameter
* Approach:

1. Use single Ml algo which was performing best.
2. Use multiple algos as an ensemble learning to generate better results.
3. Doing StandardScaler on all columns than none.