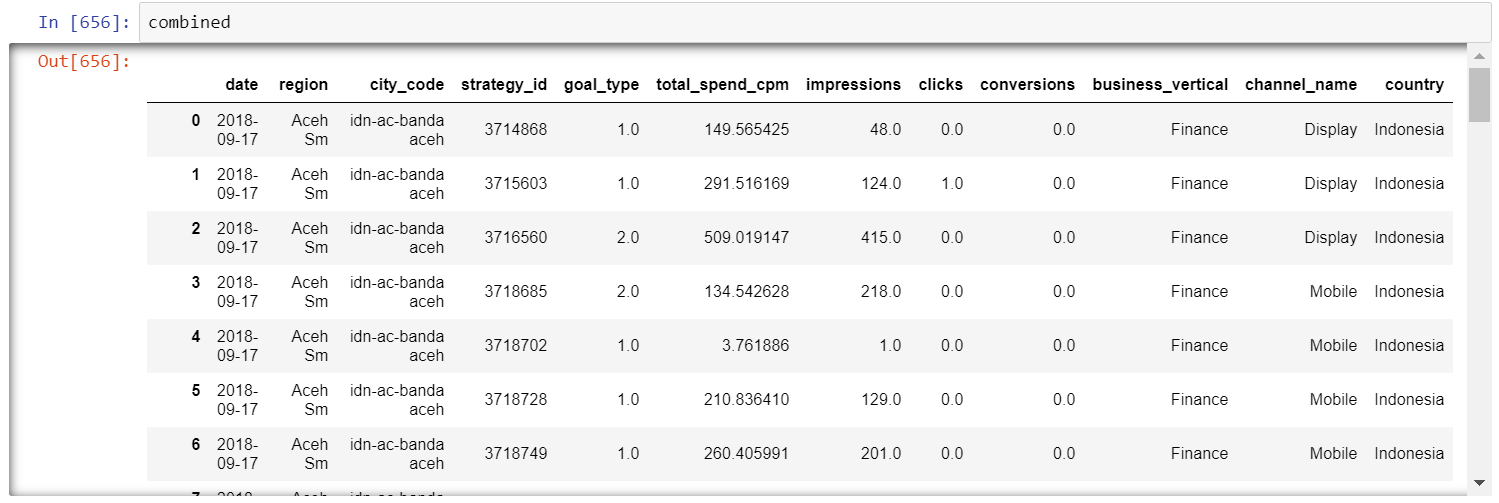
## Data table joining

Here I used join function in Python, join method” left join”. After joining, remove duplicated key and rename the joined column has been implemented.



## Detect Missing Data

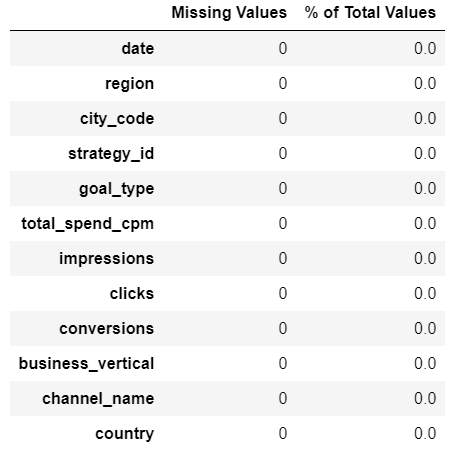
I write a python function (missing\_value\_table) to detect missing data in dataset:

There are 4 columns having missing data: impressions, clicks, conversions,goal\_type



## Fill missing data

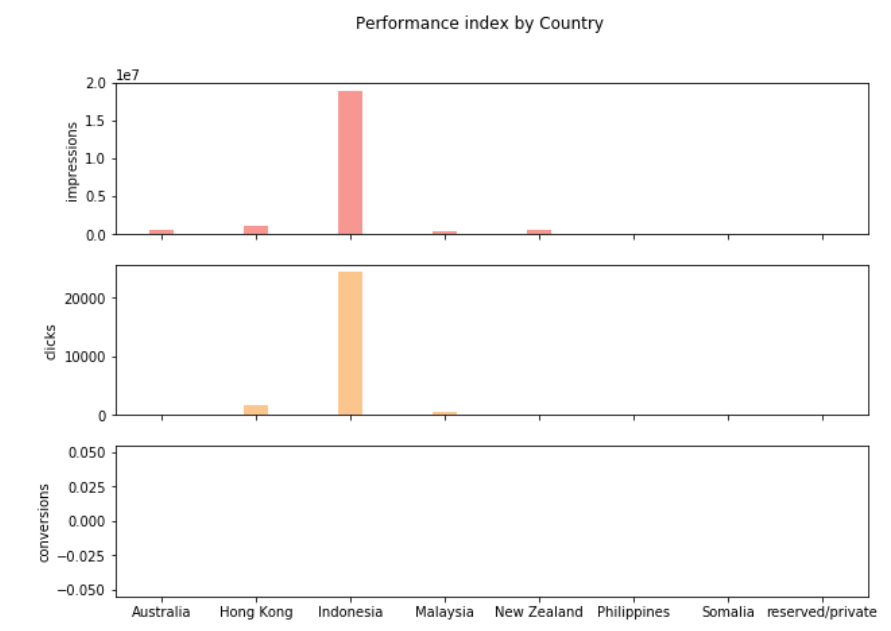
The missing value has been filled with median value of existing data in column.



## EDA

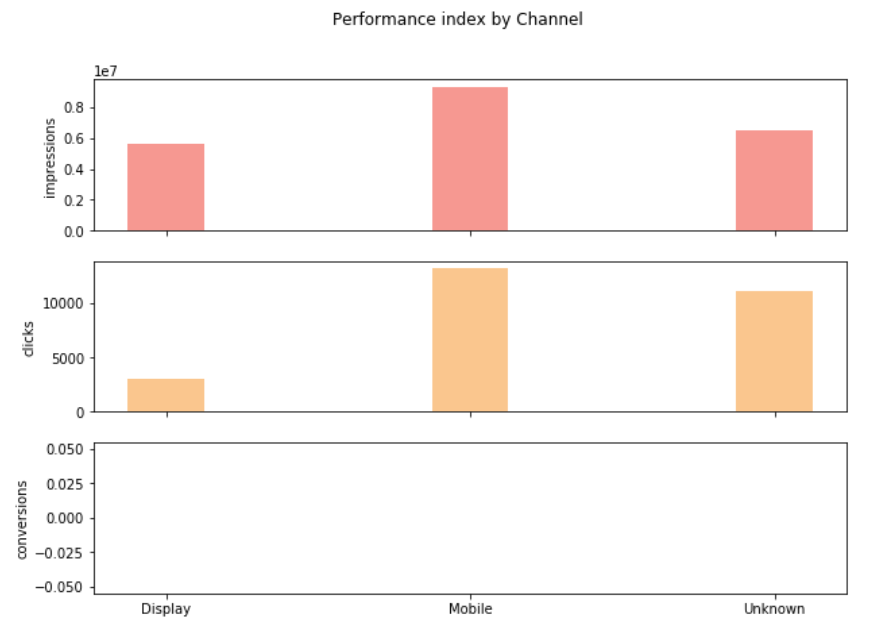
### Performance Index by Country

The plot showing Indonesia is the main market, it contribute majority of total impressions and clicks.



### Performance Index by Channel

The impressions and clicks are showing almost same distribution by channel, Mobile is the top channel for all performance index. There is no conversion at all for all channel



### Performance index by Business\_vertical

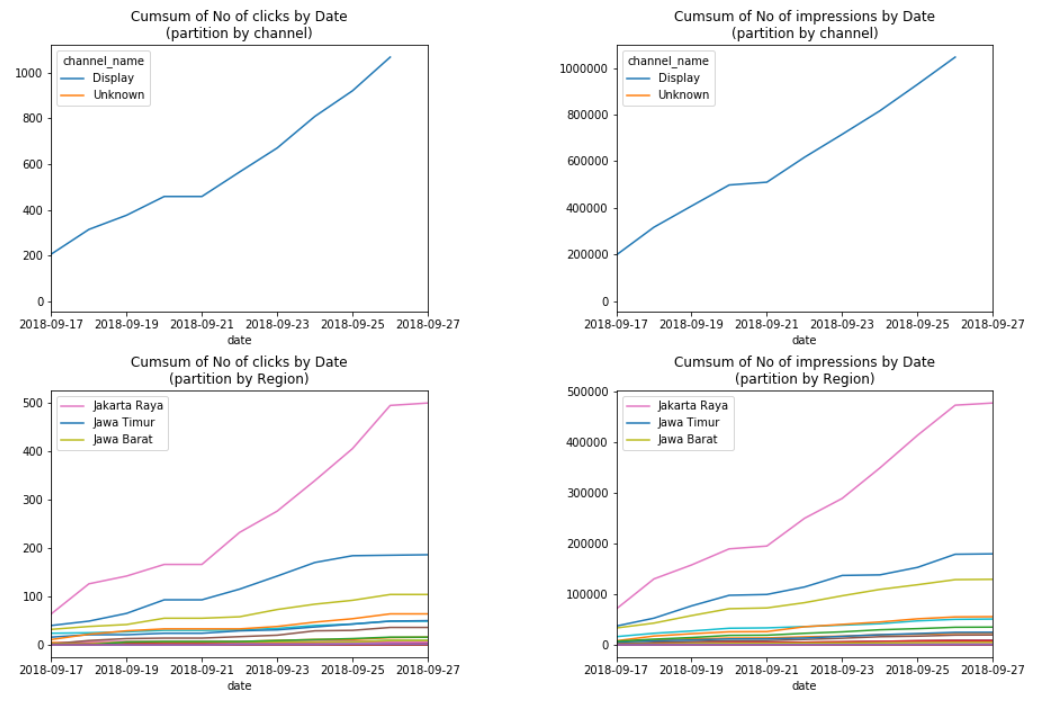
Finance is more than other unknown business\_vertical



### Specific strategy analysis

The plot below showing: Cumulative sum of impressions and click by Channel and Region

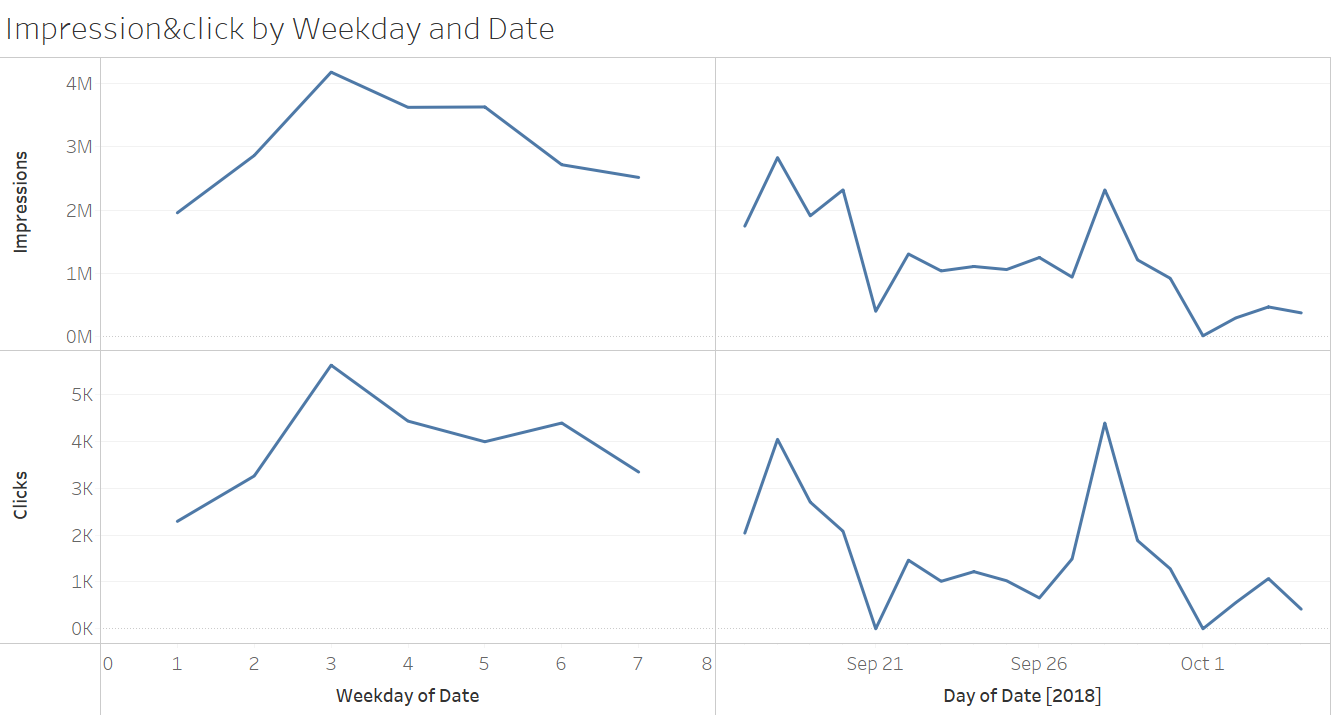
We can see clearly there is a break point at 21-09-2018, impression and click increased sharply in display channel and some Indonesia region, top region is Jakarta Raya, Jawa Timur and Jawa Barat.



### Abnormal Date detection

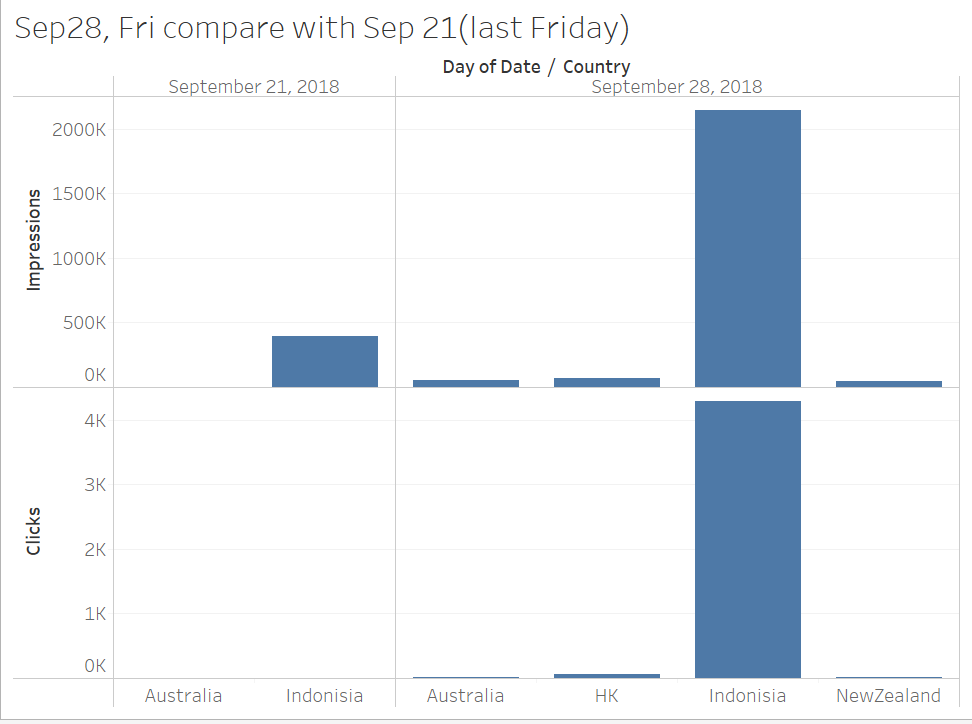
In Sep 28, there is heavy Earthquake and Tsunami, which may impact impressions and clicks.

Compare with last Friday, it is extremely high.



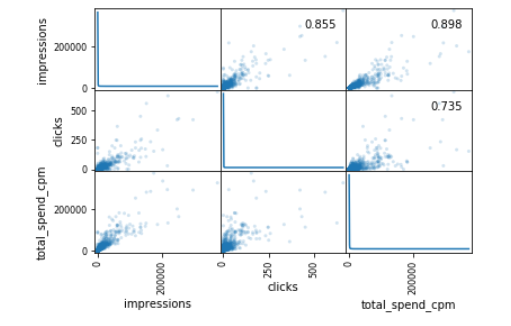
Sep 28, Fri

Indonesia Earthquake



### Numerical data correlation study

There are 3 numerical data in the dataset, from the scatter plot below, we can see there is strong correlation among total\_spend\_cpm, clicks and impression. We can use this relationship for prediction in building model.



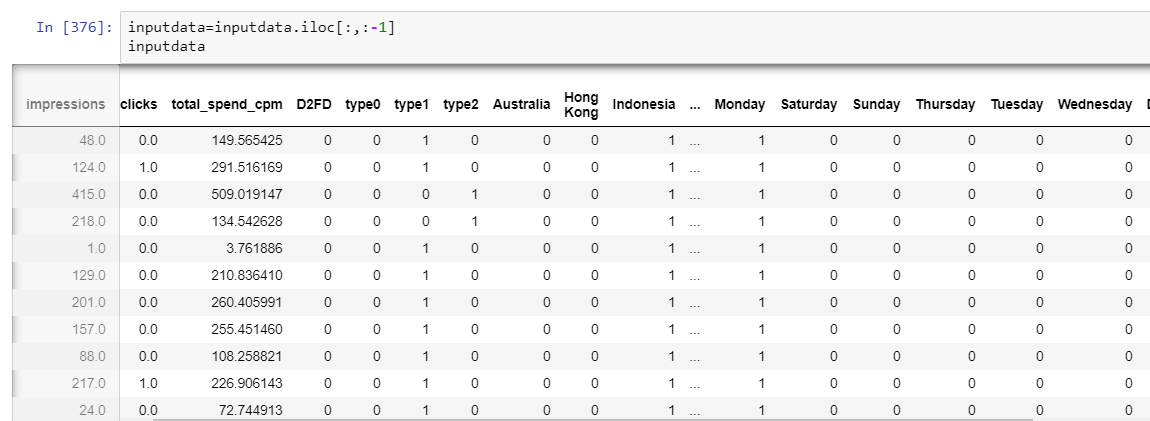
## Feature Engineering

In order to get more influencing features,I have conducted three Feature Engineering actions:

1. Get weekday by using Date variable
2. Get 'D2FD'variable which is the Number of the days from first day of the campaign started
3. Perfom one-hot encoding to categorical variables such as: country,weekday,business\_vertical,goal\_type

The Feature Engineering actions have been write into a function,which can be called anytime

Picture below showing how the dataset looks like after feature engineering.



## Impression prediction

### Split dataset to training and testing:

70% random selected data will be used as training set, 30% will be used as testing set.

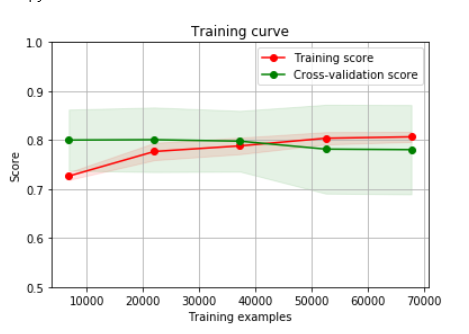
### Liner Regression with one predictor

Here I am using Lasso regression, the result will be same as Liner Regression.

The input variable is only ‘total\_spend\_cpm’, since we already found there is strong correlation between impressions and total\_spend\_cpm in EDA phase.

The learning curve showing that the model performance is keep increasing as more sample added in model, in the first 40000 samples, the model looks like a bit under fitting. After further training with more data, Training score and cross-validation score intersected at the 40000th sample, after that the model trend to over fitting.

In overall, the model is showing stable performance at R2: 0.79 and MAE:136, which means with this model, the prediction will have in average around 136 error



### RandomForest with 25 predictors

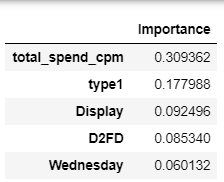
I put all 25 predictors which generated by feature engineering process into RandomForest model. After using grid search method tuning parameter, I decide to use 100 decision tree to construct RandomForest.

The learning curve showing the model is bit overfitting, the training score is higher than cross validation score, but cross validation score curve is trend to stable around 0.79.

The final prediction result with testset : R2:0.82 and MAE:114



The top 5 important feature:

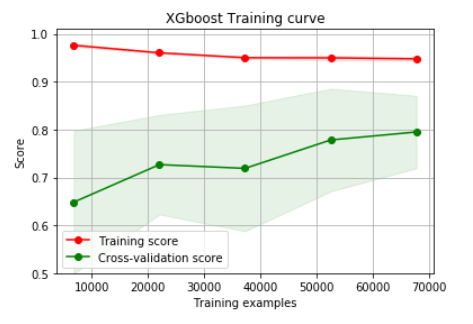


### Xgboost with 25 predictors

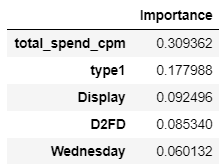
In order to get more better fitted model, the Xgboost has been applied. Boosting method will solve bias problem. The regularizer feature in Xgboost will overcome high variance issue.

The learning curve is showing that, as more sample trained by model, the overfitting issue is getting weak. With training 70000 sample, the model already can get 0.8 cross validation score.

The testing result for xgboost : R2:0.84, MAE:122

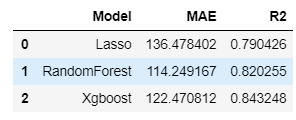


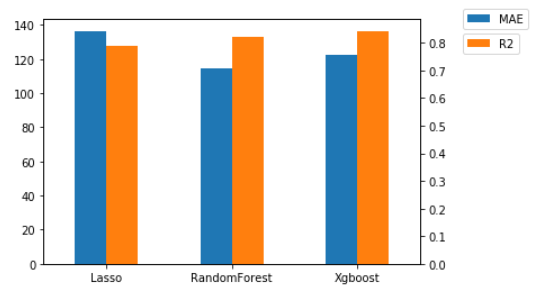
The top 5 important feature: same as RandomForest.



### Model comparison:

I used two Matric to evaluate model, the summary result is below:





The Xgboost showing best fitness and relatively lower MAE in the three model, and learning curve showing it has great potential to get better fitness. So the Xgboost will be choose as final model to predict Impressions.

## Click Prediction

### 1. Click classification

By looking at the data, only 3% of data have more than 1 click, the rest sample have no click at all. Directly use regression technique will poor fitness and high error.

My solution is to divide the problem into two question: 1. Would this observation have click. 2. If yes, how many clicks will have.

Train

Set

Regression

Model

Classification

Model

Model

Integration

Model Evaluation

Integration

Test

Set

To answer the first question, the RandomForest classification model is used. The following step has been used for construct training dataset:

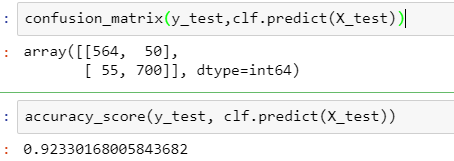
1. Randomly sample 2000 no-click observations from original dataset

2. Combined with 2000 clicked observations

3. Create new target column to indicate if the observation has clicks

4. Use RandomForest to predict

The RandomForest classifier showing very high accuracy in test dataset. Around 92% classification accuracy is achieved.



### 2. Model integration

With the classification result, if the observation is classified as ‘yes, it will have click’, it will be put into regression model for predicting how many clicks will have. If the observation is classified as ‘no, it will have no click’, the 0 will be the number of clicks in final result.

The regression model achieved R2:0.68 and MAE:4.0

The final result will be compared with test set using MAE metric, the MAE result is: 4.07.