**Missing data handling** -

for categorical data missing data is replaced by a new category called ‘absent’ .

If any new category entered by user in future, that will be considered as ‘Others’ as well.

To bring linear relationship with input feature and output target feature did some transformation -

Minmax scaling on numerical attribute like budget and then

On minmax scaled data, did inverse exponential transformation.

Inverse exponential transformation on attributes  - target impression, campaign duration, segment size

Logarithmic transformation on attributes – copy length,

Target encoding on categorical attributes – like

 industry , there are

computers, cosmetics & toiletries, drinks, entertainment & the media, finance, foods so, more than 30 industry there.

So, we can’t do one hot encoding , because that will increase the feature /column number. That’s not good from curse of dimensionality perspective.

Ordinal encoding we can’t do , as there is no specific order in industry, we can’t encode like computer industry as 1, media & entertainment as 2 , that doesn’t make sense.

Label encoding also problem there, label encoding is done based on alphabet in ascending order, like category starting with A will get label as 1, starting with next alphabet ‘like’ C will get label as 2 , like this.

In this way who will get 30 as label, that will get more weightage during model training. That should not happen.

So in target encoding …………………………………………….

    Industry

affluence\_band

youngest\_adult\_vb

barb\_tv\_regions

Inverse exponential transformation –

(1-np.exp(-1\*B\*pow(data\_val, C)))

B and C value taken after lot of trial and error method, finalizing curve coming saturated after a point of data.