Lead Score Assignment Summary

- Logistic Regression model building is implemented to maximize the lead conversion probability
- **Data sanity** Missing values are imputed as mode (in case of few categorical variable), as mean (in case of few numerical variables) and as 98% in case of few outliers.

- **Categorical variables**: Categorical variables are created using'get_dummies' (i.e. k-1 variables for k types of values for a particular categorical variable)
- **Yes/No variables**: Yes- no type of categorical varaibles are replaced with 0 and 1

- **Scaling numerical variables**: Numerical variables are scaled using standard scaler

- **Building mode**: Once, no missing values and outliers are present, dummy variables are created, we build a logistic regression binomial model for predicting the conversion probability

- **RFE method** : RFE method is used with 25 varaibles to start with

- **Statistical significance and Collinearity :** Variables with p values above 0.05 are removed one at a time (statistical significance). VIF score is also checked and variables with VIF above 5 are removed one at a time (for collinearity)

- **Probability threshold for conversion** Model is assessed with conversion threshold of 0.5.
- **ROC curve** : ROC curve is plotted at optimal value comes out to be 0.4

- **Updated Probability threshold for conversion as 0.4** Updated parameters with p value as 0.4

- **Evaluation of model on training set** Accuracy = 92%, Sensitivity =91.8%, Specificity=92% Positive predicte rate = 87%, Negative predictive value =94.8%
- Precision = 90.6%, Recall = 86.28%

- **Evaluation of model on test set :**Values on Test Set
  - Accuracy = 91.59%
  - Sensitivity = 92.6%
  - Specificity = 90.9%

- **Conclusion** : Top 3 variables which should be focussed upon for maximising the probability of conversion
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email