**Summary**

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. **EDA:**

* We dropped columns which are not required for analysis like Prospect\_Id.
* Quick check was done on % of null values and we dropped columns with more than 45% missing values.
* We replaced the ‘select’ values in some columns with ‘Not specified'.
* Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
* Then we saw the Number of Values for India were quite high (nearly 97% of the Data), so this column was dropped.
* We checked the Imbalance data…..wherever we found that , we just dropped that columns as they can be a factor of biased result.
* We also filled some Null values in those columns which are relevant for analysis.
* To get the only desired columns, we had gone through columns one by one, checked its importance and took the decision.

1. **Train-Test split & Scaling :**

* The split was done at 80% and 20% for train and test data respectively.
* We have done scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

1. **Model Building**

* We have attained top 13 relevant columns out of which 12 are independent and 1 is dependent (‘converted’ column).
* Independent columns contain 9 categorical feature and 3 numerical features.
* Dummy variables of the categorical features were created by using get\_dummies() method.
* As our target variable is categorical, for providing training or fitting the model to the training set, we will import the Logistic Regression class of the sklearn library.
* After importing the class, we have created a classifier object and used it to fit the model to the logistic regression.
* We also observe the summary.

**4. Predicting the Test Result**Our model is well trained on the training set, so we have predicted the result by using test set data.

1. **Model Evaluation**

* We have created the confusion matrix here to check the accuracy of the classification.
* True Positive = 1072
* True Negative = 650
* False Negative = 47
* False Positive = 79

**Accuracy = 93%**

**Precision = 93%**

**Recall = 89%**

**CONCLUSION**

TOP VARIABLE CONTRIBUTING TO CONVERSION:

* + - 1. Lead Origin
      2. Tags
      3. Last\_Notable\_activity

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.