# Summary

**Introduction:**

The analysis is done for X Education to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The data given to us provided the information about how the potential customers visit the site, the time they spend there, how they reached the site, the conversion rate, etc.

**Background:**

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

**Objective:**

1. X education wants to select the most promising leads
2. X education wants to build a model to identify the hot leads.
3. X educations wants to deploy the model for future use as well in case requirement changes

Now that we have already defined the objective of the case study and data has been provided by X education, the following are the steps used:

1. **Cleaning data:**

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information.

**2. EDA:**

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good. Potential outliers are identified, particularly within the class labeled as "No Conversion". However, at the aggregate level, the impact of outliers appears to be neutralized, as the distribution of the "Conversion" class is less skewed compared to the "No Conversion" class

3. **Train-Test split:**

Dataset was split into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance

The split was done at 75% and 25% for train and test data respectively.

4. **Model Building:**

Based on the results, we chose the RandomForest Classifier, GradientBoosting, LightGBM & Catboost on which we tested the other metrics to see in depth performance of these 4 models based on several different metrics to choose the best model for our analysis.

5. **Model Evaluation:**

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

Analysis Summary:

A) Model Accuracy:

Random Forest:

- Train Accuracy: 98.47%
- Test Accuracy: 91.69%

Gradient Boosting:

- Train Accuracy: 91.74%
- Test Accuracy: 91.66%

LightGBM:

- Train Accuracy: 94.58%
- Test Accuracy: 91.55%

CatBoost:

- Train Accuracy: 94.05%
- Test Accuracy: 92.02%

B) Model Precision:

Random Forest:

- Train Precision (Class 0): 97.95%
- Train Precision (Class 1): 99.30%
- Test Precision (Class 0): 91.84%
- Test Precision (Class 1): 91.42%

Gradient Boosting:

- Train Precision (Class 0): 90.71%
- Train Precision (Class 1): 93.63%
- Test Precision (Class 0): 91.37%
- Test Precision (Class 1): 92.18%

LightGBM:

- Train Precision (Class 0): 94.12%
- Train Precision (Class 1): 95.37%
- Test Precision (Class 0): 92.11%
- Test Precision (Class 1): 90.56%

CatBoost:

- Train Precision (Class 0): 93.43%
- Train Precision (Class 1): 95.15%
- Test Precision (Class 0): 92.07%
- Test Precision (Class 1): 91.91%

C) F1-Score:

Random Forest:

- Train F1-Score (Class 0): 98.75%
- Train F1-Score (Class 1): 97.99%
- Test F1-Score (Class 0): 93.42%
- Test F1-Score (Class 1): 88.73%

Gradient Boosting:

- Train F1-Score (Class 0): 93.45%
- Train F1-Score (Class 1): 88.80%
- Test F1-Score (Class 0): 93.42%
- Test F1-Score (Class 1): 88.58%

LightGBM:

- Train F1-Score (Class 0): 95.64%

- Train F1-Score (Class 1): 92.83%
- Test F1-Score (Class 0): 93.27%
- Test F1-Score (Class 1): 88.62%

CatBoost:

- Train F1-Score (Class 0): 95.23%
- Train F1-Score (Class 1): 92.09%
- Test F1-Score (Class 0): 93.68%
- Test F1-Score (Class 1): 89.17%

The analysis emphasizes the trade-offs between training and testing performance for each model. Further optimization and hyperparameter tuning are recommended, with a focus on Random Forest and CatBoost models for enhanced results.

6. **Tune Hyperparameters:**

Tuning the hyperparameters of the model helps to optimize its performance.

7. **Prediction:**

Prediction was done on the test data frame with accuracy, sensitivity and specificity of ~90%.

8. **Precision – Recall:**

This method was also used to find the precision, which is approximately 91.8% and the recall is approximately 86.4%. These metrics provide insights into the model's ability to make accurate positive predictions and capture all the positive instances, respectively

**Conclusion:**

It was found that the variables that mattered the most in the potential buyers:

1. The total time spend on the Website.

2. Total number of visits.

3. When the lead source was:

   a. Google

   b. Direct traffic

   c. Organic search

   d. Welingak website

4. When the last activity was:

   a. SMS

   b. Olark chat conversation

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.