Summary report

Business Objective

The objective of this project was to develop a logistic regression model for X Education in order to improve lead conversion rate.

Data Wrangling

- Inconsistencies in column names were treated/shortened to ensure better readability of the code and analysis.
- Missing values was investigated and treated/flagged. ‘Select’ category was converted to NaN.
- Features with high percentage of NaN (such as AP/AA Score, Index) or have a Missing At Random relationship with other features (such as Lead_Quality, Lead_Profile,…) were also removed.
- Constant/quasi-constant features such as the surveys where the leads learned about the company from were also dropped.
- Duplicated observations were also removed. There were 1281 duplicated entries from 196 actual unique observations. They did not add more information and could lead to overestimated performance of the model
- Outliers that did not make sense in Total_Visits and Page_Views_Per_Visit were capped at a reasonable limit to ensure they would not affect the coefficients of the model.
- Category names were spellchecked/shortened. Infrequent classes were grouped into “Others” as there may not have enough observations to provide reliable estimates of their impact.

EDA

- Potential features in predicting probability and insights regards to Total_Visits, Total_Time_Spent, Tags, Lead_Source, Occupation, etc… were discovered, which greatly helped in feature selection and coefficients interpretation.

Data Preparation

- Categorical features were converted to WOE values and dummies variables.
- Drop highly correlated features based on the information values (for eg Lead_Source/Lead_Origin). Having highly correlated features can lead to unreliable p-values and coefficients signs.
- Min-max scaling was done to ensure all features stay in a similar range.

Feature selection

- Top features were selected using RFE and their VIF values (<5)

Model building and tuning

- Iterate model building and eliminate insignificant features (p-values >0.05)
- Different models were attempted with different initial features.
    - First model: only WOE values of categorical variables were used
    - Second and third models: the WOE variables with the highest coefficient were replaced with their dummies to better understand the impact of individual category.
- This approach made sure the first model was kept relatively simple and subsequent models could be scaled up in complexity where needed.

Model evaluation

- Due to business objective and data imbalance, F1 score was the primary metric instead of accuracy. We want to limit both the false positives (waste of resources) and false negatives (loss of revenue).
- Model performance are evaluated based on the confusion matrix on both train and test data to ensure the model is not overfitted.
- Threshold optimization using precision/recall curve was performed.
- K-fold cross validation was done to ensure the model can generalize well.

Final model selection

- Model 3 was selected due to its performance and features explainability. AUC = 0.975, F1 = 0.915, Accuracy = 0.932.
- Features that positively and negatively affect the probability of conversion were identified and presented.
- Other subjective questions were also solved with model 3.

Kfold cross validation result

```
cv_result
```

	Model	Accuracy	Precision	Recall	F1	AUC
0	model1	0.924237	0.907712	0.902293	0.904994	0.973579
1	model2	0.930645	0.918283	0.907320	0.912769	0.974302
2	model3	0.932278	0.918354	0.911719	0.915024	0.974777