Lead Scoring Case Study Summary

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- X Education gets a lot of leads, but its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%. In order to increase the lead conversion process more efficiently, the company wishes to identify the potential leads, who can be categorized as 'Hot Leads'.
- Once 'Hot Leads' are acquired, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Goal:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company wants to build a model to identify Hot Leads.
- Improve the lead conversion rate from 30% to 80%

Approach

- Exploratory Data Analysis (EDA)
  - Data Cleaning
  - Categorical Analysis
  - Numerical Analysis
- Data Preparation
- Model Building
- Metrics Validation
- Conclusions & Recommendations

Exploratory Data Analysis (EDA)

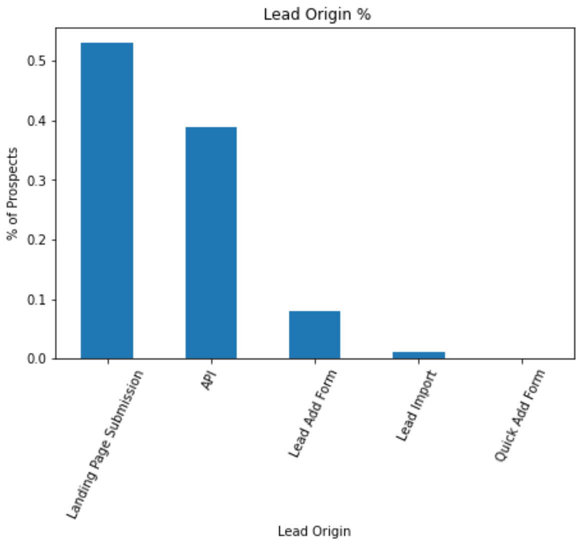
Data Cleaning

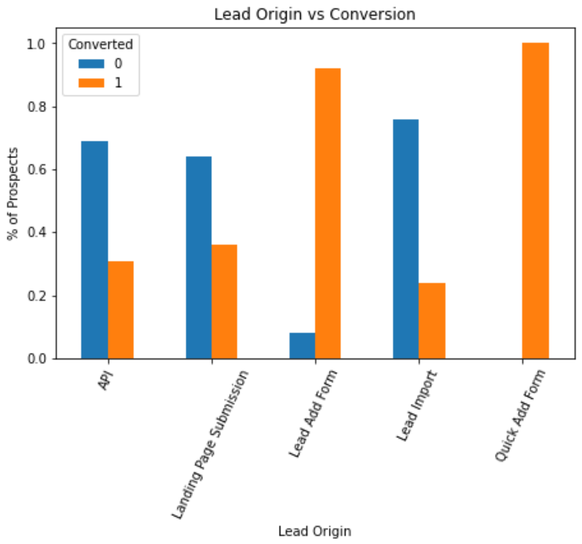
- v 37 columns X 9240 rows
- Data set has a lot of categorical features.
- Data Imbalances for categorical features analyzed and dropped features with high imbalance
- Data set has conversion ratio of 38%
- 17 columns have missing values, dropped columns Missing values greater than 40%.
- Imputing and standardization - Replaced **'Select'** with **'NaN'** based on data analysis as this value is result of not entering values as part of data collection
- Imputing City column Nulls and other low frequency values to 'Others'
- Dropped records where null values are present in numerical columns 'Page Views Per Visit' and 'TotalVisits'.

Categorical & Numerical Analysis:

Univariate and Bivariate analysis of categorical variables done and inferences observed. Numerical analysis is done using correlation diagram, box and distribution plots to identify correlations and perform outlier treatment using capping to retain data.

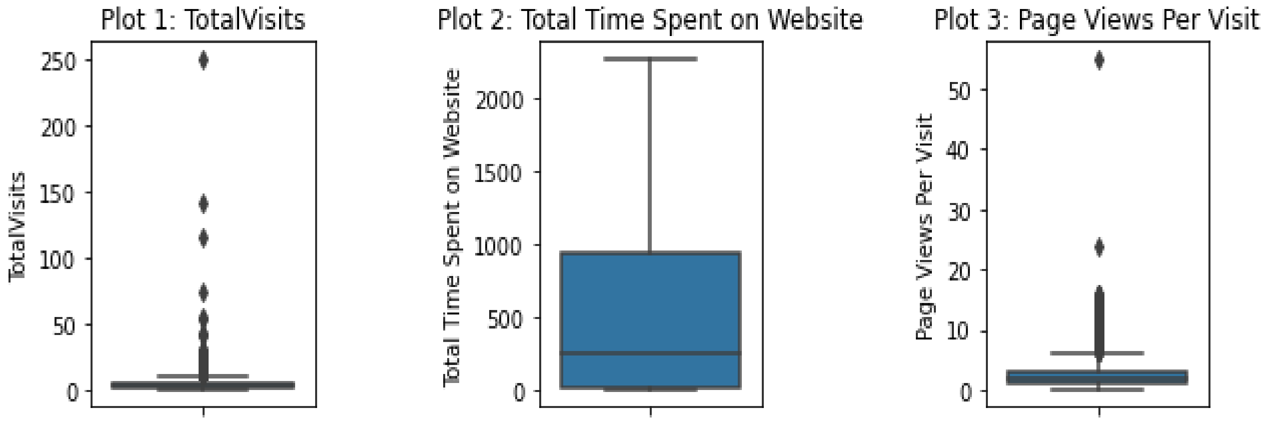
Categorical column ‘Lead Origin’ Analysis visualisation shown below:





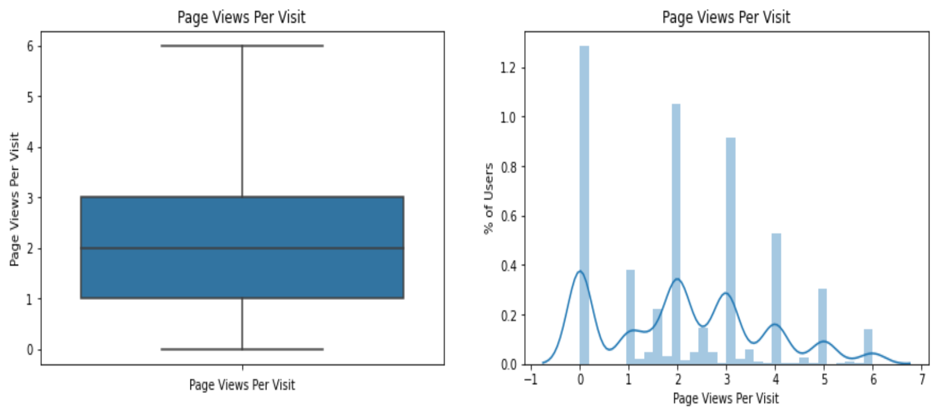
Outlier Analysis Visualisation

Prior to capping



- Capping is performed with IQR*1.5
- Why Capping?

  Capping is done to retain data instead of dropping.
  99% of records are retained for modeling.



‘Page Views Per Visit’ after outlier treatment shown above.

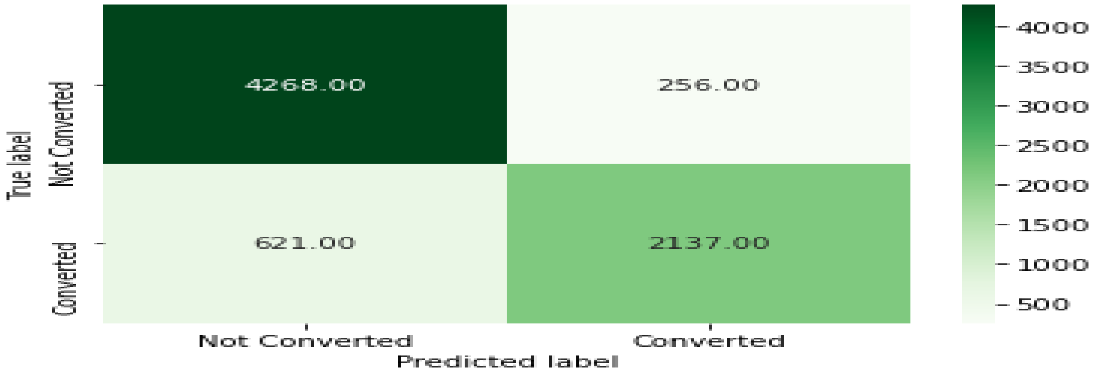
Data preparation

- Dummy creation: Categorical features processed for dummy variables conversion using binary mapping get_dummies function
- Correlation matrix shows there are features with significant correlation. Dropped features with correlation above 0.4
- Train Test Split:_Data set is training and test data sets with test size of 20%
- Scaling:_Scaling of features 'TotalVisits', 'Total Time Spent on Website' is done using StandardScaler for training data sets.

Model Building

- Model 1 - Logistic regression model built using all the features and observed multiple features having high p value which indicates model is not statistically significant and indicates strong evidence for null hypothesis.
- Model 2 - Using RFE (Recursive Feature Elimination) - 15 features are selected for model building.
- Model 3, Model 4, Model 5 and Model 6 are created iteratively by dropping one feature at a time while analyzing p-values and VIF (Variance Inflation Factor) scores
- **Model 6** is resulting with acceptable p-values (<0.05) and VIF scores less than 2. This model is considered for evaluation and metrics analysis.
- Converted probability calculated using Model 6 for training data set and Predicted values for with cutoff of 0.5.

Confusion Matrix for cutoff 0.5 shown below

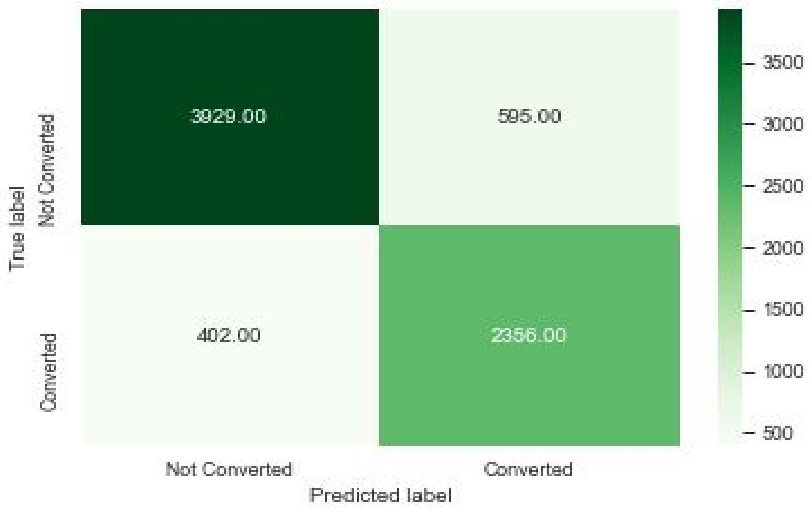


Metrics Validation

ROC Curve, Optimal Cut Off

- The ROC Curve plotted. Good value of Area of ROC Curve 0.93 indicates a good predictive model
- To find Optimal cut off point - Plotted line graphs as below with accuracy, sensitivity and specificity
- Intersecting point is 0.27, so cut off probability is 0.27 and analyzed metrics.

Confusion Matrix for cutoff 0.27 shown below



Precision Recall Curve

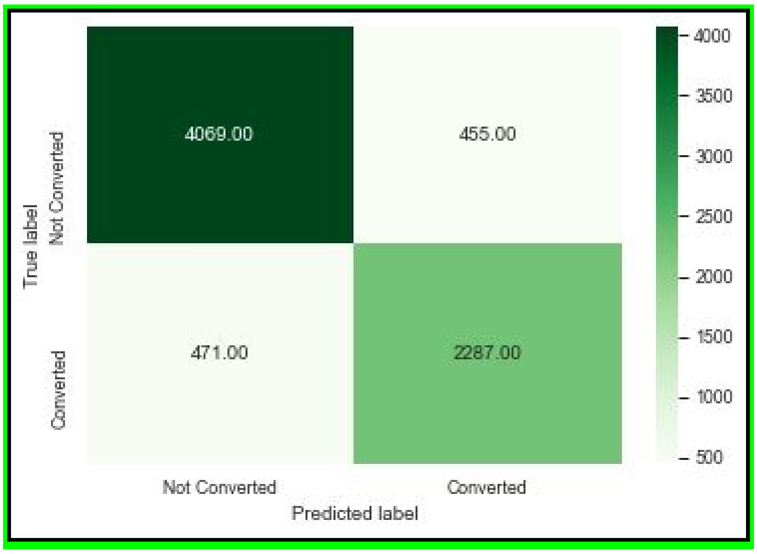
- Tradeoff between Precision and Recall is clearly visible with precision recall curve
- Optimal Cut off is 0.34 based on precision recall curve

Predictions on Test Set

- After scaling Test set and predicting, here are metrics for the final model with 0.34 cut off.

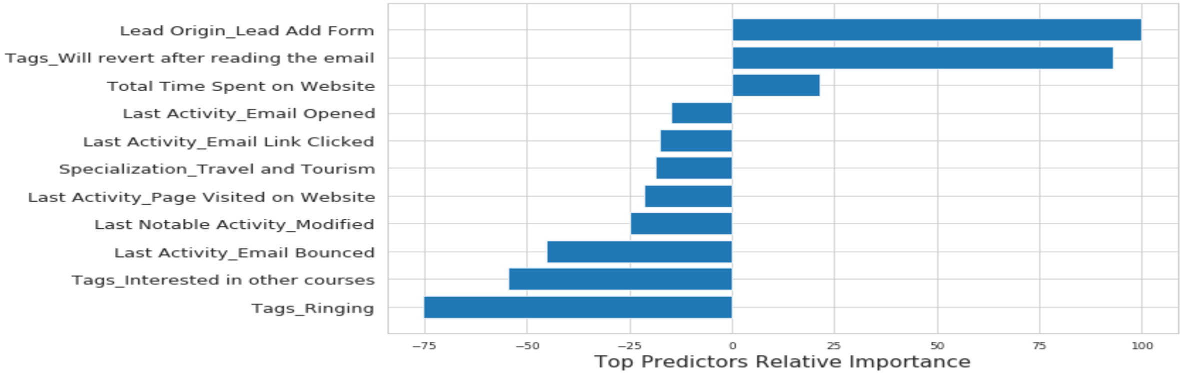
▪ Accuracy	= 87.80 %
▪ Sensitivity	= 83.64 %
▪ Specificity	= 90.43 %
▪ Precision	= 84.60 %
▪ Recall	= 83.64 %

Confusion Matrix for optimal cutoff 0.34 shown below



Model Conclusions

- ✔ ~84% of Recall value means that model is able to predict 84% of actual conversion cases correctly.
- ✔ ~85% of Precision value indicates that 85% of the conversions that our model predicted are actually converted.
- ✔ Top 3 features/dummy variables that contribute towards lead conversion(Hot leads) in this model are:
  - Lead Origin_Lead Add Form
  - Tags_Will revert after reading the email
  - Total time Spent on Website
- ✔ Top Predictors of relative importance from the model are shown below.



Recommendations

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- ❖ **Increase user engagement on their website since this helps in higher conversion**

- **Increase on sending email notifications since this helps in higher conversion**

- **Get Total Visits increased by advertising etc., as this helps in higher conversion**

- **Improve the effectiveness of phone service since this is affecting the conversion negatively**