LEAD SCORING CASE STUDY REPORT:

Dataset consists of **9240 rows & 37 columns**. *Converted* is dependent variable to predict

Data Cleaning: No Duplicates Found. Replaced “Select” as missing in entire dataset. Dropped categorical variables having high class imbalance (>90% as same class)

Missing Value Treatment: Dropped columns with >45% values missing. And for rest, imputed missing values with a) Mode or separate class in case of categorical variable b) Median in case of numeric variable

Outlier Treatment: Removed upper and bottom 1% values for numeric features

After cleaning, dataset has 9157 rows; 12 independent features (3 Numeric, 9 Categorical); 1 Target feature

EDA Insights:

- It is observed that **converted leads spend more time on website**; plan website engagement activities like masterclass, quizzes etc.
- **Total visits have high correlation with pages view per visit**; should be taken care during model building
- Lead Source: a) **Welingak website & Reference has high conversions**- plan more leads from these sources b) Direct Traffic & Google have high lead volumes- plan to increase conversions here
- Plan more leads with current occupation **as working professionals** & housewife- show high conversions
- Plan to assign leads ASAP for last notable activity as SMS sent- high conversion
- **Target Imbalance:** 38.5% Class1 and 61.5% Class0. Classes are pretty much balanced to build model

Prepare Data for Modelling:

- Categorical features: a) For binary class encode with 0’s and 1’s b) For multiclass, create n-1 **dummy variables** features (n= number of classes)
- Divide dataset into 70% train and 30% test
- Apply **standardization** for numeric features to have all features on same scale

So, finally after data preparation step, there are 59 predictor variables

Model Building:

- Built Logistic Regression model & applied **RFE** to coarse tune and **obtain 20 significant features**
- Further, used **stats model** to fine tune and **obtain 13 final predictors** having **p-value<5%, VIF<5**

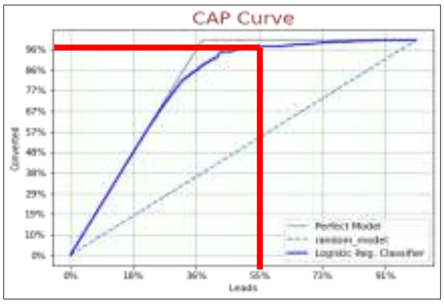
Model Evaluation and Predictions:

Model Perf. on Train data	Model Perf. on Test Data
Accuracy- 90%	Accuracy- 90.3%
Sensitivity- 89.2%	Sensitivity- 90%
Specificity- 90.5%	Specificity- 90.5%
Precision 85.6%	Precision 85.4%

- Made predictions on train data using the built model and constructed **ROC Curve**; resulted in pretty good model with **AUC= 96%**
- Simply multiplied probability with 100 to get a **lead score between 0-100**. **Higher score meaning greater chances of lead conversion**
- Obtained **0.3 as optimal cut off** to balance sensitivity, accuracy, specificity
- Made **predictions on test data**. Resulted in a well generalized model, able to **capture ~90% conversions on both train and test data**

Model Interpretation and Recommendations:

Business Impact driven by Model from CAP Curve: Just by attempting 55% of Leads prioritized by model. We can capture ~96% of Sales. So, Logistic classifier helps optimize sales team accordingly, and increase efficiency on lead churning.



RECOMMENDATIONS:

- Aggressive follow up on leads tagged *Will revert after reading the email*
- Target leads from Welingak website and Reference Lead Sources
- As X-Education offers professional courses- target leads who marked themselves as working professionals
- Target leads that have last notable activity as SMS Sent, and leads that spend more time on website
- Avoid leads with student’s tag- maybe they do not have paying capacity (early stage)
- Avoid leads tagged- *Interested in other courses/ Switched off/ Ringing*