INFORMATION:

The given data has 25 columns and 7594 records with **Target Feature (Categorical):** 'IsConverted' and other columns.

Numeric features:

['NumberOfEmployees', 'NumberofOpenings', 'LeadAge', 'VisitWebpageCount', 'FillOutFormCount', 'ClickLinkCount', 'OpenEmailCount', 'ClickEmailCount', 'FillOutFacebookLeadAdsFormCount', 'FillOutLinkedInLeadGenFormCount', 'EmailBouncedCount', 'UnsubscribeEmailCount']

Categorical features:

['CompanyID', 'ConvertedDate', 'Country', 'CreatedDate', 'DomainType', 'EstimatedAnnualRevenue', 'SeniorityLevel', 'Industry', 'LeadSourceType', 'Campaign', 'State', 'StatusReason']

The data i.e., all the features have majority of missing values in them.

MISSING VALUES:

To handle missing values, columns having more than 45% NA values were dropped because imputation in columns having severe missing data could introduce a bias. Hence, it was better decision to drop it.The column which were dropped were ConvertedDate, SeniorityLevel , EstimatedAnnualRevenue .

Later , after careful consideration, Features like  FillOutFacebookLeadAdsFormCount  , FillOutFormCount  ,VisitWebpageCount  , ClickEmailCount , UnsubscribeEmailCount , OpenEmailCount FillOutLinkedInLeadGenFormCount, EmailBouncedCount , ClickLinkCount show similarity in missing records  i.e all these columns have values missing for the  same records, therefore such records were dropped as it was better to drop such consistent records with missing values than imputing them which may cause bias in data.

Then, all categorical columns were checked, and missing values were imputed.

Also, some data cleaning and feature engineering is done to maintain data consistency and by placing the sub types of categories of column into main type.

The numerical data was checked for outlier and imputed using median or mode as most of the features had outliers because mean value is affected by the presence of those outliers.

DATA INSIGHTS:

The Exploratory data analysis is then performed:

**Target Feature:**

The target data is highly imbalanced with Percentage of non- conversion leads: 80.54% and

Percentage of conversion leads: 19.46%

**Categorical Feature:**

The below are some insights:

* Only the leads having domain type‘business' is likely to become an opportunity with 23% conversion rate . So, focus should be given more to customers having domain as business.
* Education and anonymous domain type show no chances (equal to 1%) of lead conversion.
* Website(Call/Chat/Form) and Google(Search/Ad) generates maximum number of leads (equal to 20% conversion rate).
* Facebook (less than 5%) has no conversion rate at all.
* Although, we have less Bing Ad search entries, this lead source type has good conversion rate as good as website and google ad/search.This proves that Advertisement contributes to more conversion rate and should be focused more.
* We also have less records for landing pages, email, referral but this kind of source type have good conversion rate which can be taken into account.
* PaidSearch and OrganicWeb, Other(Imputed missing value) generates maximum number of leads.
* LinkedIn, and Facebook campaign type does not help in lead conversion as we see that can lead non-conversion rate is higher than conversion rate.
* Unknown reasons (Imputed value) show successful conversion rate than any other Status Reason. This shows many people does'nt disclose why the lead converted into opportunity. OR we missed to track this reason down. It is very necessary to take surveys from happy customers and trace down the status of conversion.
* Although, we have less records for 'current customers' status type in data it seems that current customer's lead is likely to convert into opportunity. Thus, current customers satisfaction can be used to generate more opportunities.
* Non-responsive customers show highest unsucessful conversion rate.
* Unqualified, Not now, maybe later, international, price,browsing shows less conversion rate . Infact, all the reason type shows high non-conversion rate than conversion rate.
* US and canada shows highest lead conversion. (Above 20% conversion rate of all records)
* Uk and Germany show relatively good lead coversion than any other countries.
* Also, US shows highest no conversion rate too.
* The users are majorly from US, Canada. Other international countries do not use the service or have very less conversion rate as evident from status type 'International' having higher non conversion rate. Hence, it is very important to concentrate on US, Canada customers for opportunity conversion.Also, international customers from Germany , India , Uk can be targetted if expansion of business is planned.
* California, Texas, Illinous, Florida and NYC has good conversion rate but also high non conversion rate . These cities should be focused more. Only US states have some percent of conversion but less than non-conversion rate.

**Numeric variables:**

A correlation matrix is generated to follow the correlation between contiguous variables which shows:

* Click linkcount and visitwebpagecount have very high correlation so, for modeling purpose , one of these variables must be drop.FillOutFormcOunt and VisitWepbpageCount also show some positive corelation.Followed by FillOutFormcOunt and UnsubscribedEmailCount.
* IsConverted and Visitwebpagecount/ClickLinkCount show positive corelation that means the user who have high clicklinkcount has high possiblity of converting into opportunity. People who click more time on the website are more likely to convert.

The outliers in the numerical variables were checked using boxplot and all of them has presence of outliers. So , it is necessary that All these outliers should be taken into account while creating a lead prediction model by dropping most of them.

Bivariate analysis Is also done.

DATA MODELLING :

As the missing values in data are imputed, outliers are accounted and EDA is done , next step is data modeling.

Considering this is supervised problem of classification which will predict the lead conversion of user into opportunity.

For this, We can use logistic regression,Random Forest classifier,Weighted XGBoost,SVM with  data balancing techniques for predicting the lead conversion to the opportunity.As the dataset is imbalanced Weighted XGBoost can offer better performance on binary classification problems with a severe class imbalance. Also, Random Forest classifier will perform better handling imbalance and outliers. As the dataset is small and we see presence of outliers in most of continous variables, SVM classifier can be used as it is very robust to outliers. We can experiment these mentioned classifier and choose the best performing out of it.

**Approach ⁉**

**1.   Data Preparation for modeling**

* Check Correlation
* Drop high corelated features learned from correlation matrix and keep either one.
* Convert categorical columns into dummy variables or one-hot encoding.
* Divide dataset into X and Y (IsConverted) for model building.
* Train - Test Split (80%-20%) for separating data into train and test.
* To handle the highly imbalanced target data, Apply SMOTE for oversampling where the synthetic samples are generated for the minority class. By random oversampling. This will overcome the overfitting problem which can be caused due to imbalanced data.
* Feature Scaling to normalize the features in the dataset into a finite range regardless the unit of value so that ML model tend to perform well.

**2.   Data Modeling**

* Apply Logistic regression using sklearn package
* Apply SVM using sklearn package
* Apply Random Forest Classifier using sklearn package
* Apply weighted XGBoost.
* Check the performance of all models  and compare the results to choose best performing on dataset. Cross validation techniques can also be used along.

**3. Model Evaluation.**

* As it is classification problem, we will use accuracy as the evaluation metrics. Also, we will Derive and check classification report which would give the precision, recall, F1 score and and classification confusion matrix
* Derive Area under ROC curve which gives the measure of the ability of a classifier to distinguish between classes.
* Using this metrics , we can compare all the above classifiers and choose the best performing one .