Lead Scoring Assignment Summary

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google etc. Although X Education gets a lot of leads, its lead conversion rate is very poor.

The company needs a model to derive a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Solution Summary:

Determining the conversion of the leads received can be done using the Linear Regression model using the dataset provided with the leads.

Following are the steps involved in the analysis.

Step 1: Reading and understanding Dataset.

- Import important libraries
- Reading the dataset provided in leads.csv
- Quick review of dataframe
- Shape of Leads dataframe
- Check for conversion rate in dataframe Conversion Rate is 39%

Step2: Data Cleaning: Analyze data and prepare data

- Check for missing values
- Check level of categorical columns
- Identify columns that have default "Select" value
- Check for missing values
- Identify categorical columns with missing values
- Identify quantitative columns with missing values
- Calculate percentage missing values (Re-check)
- Checking distribution of these quantitative variables
- Impute quantitative columns
- Impute missing values for categorical values with less missing values
- Update numcols and nonnumcols since we dropped few columns
- Boxplot for quantitative varibles
- Bivariate Analysis
- Outlier Analysis
- Create dummy variables
- Label encoding for other categorical columns
- Drop columns with no variance
- Checking correlation

Step3: Data Preparation for Modeling

- Creating Dummies
- Train Test split
- Feature Scaling

Step4: Model Building

- Create a function for model building
- Feature Scaling
- Use RFE for feature selection
- Using statsmodel for rfe columns
- Predicting based on latest model
- Create confusion metrics
- Plot ROC Curve
- Find Optimal cutoff value
- Plot accuracy sensitivity and specificity
- Precision - Recall plot

Step5: ***Predicting based on latest model***

- Plot ROC Curve
- Find Optimal cutoff value
- Plot accuracy sensitivity and specificity
- Confusion metrics and scores
- Precision - Recall plot
- Making prediction on test set
- Find Principal Components using PCA
- Logistic Regression on PCA

Step6: Model Conclusion

- Merging train and test prediction
- Merging predictions to original dataframe
- Creating Lead Score column
- Creating a dataframe with cutoff and conversion%