upGrad

LIVERPOOL JOHN MOORES UNIVERSITY

International Institute of Information Technology Bangalore

Lead Scoring Case Study

Le Minh Khoa – DS C12

INTRODUCTION TO THE ASSIGNMENT AND ITS OBJECTIVES

- This assignment centers on enhancing X Education's current lead conversion rate of 30%, with a target to reach roughly 80%. Using historical data from around 9000 leads, my job is to create a logistic regression model to score each lead's likelihood of conversion. This "Lead Score" will help the sales team focus on the most promising leads. The model should also be adaptable for future business needs. Summarize my insights and suggestions in a concluding PowerPoint slide deck.
- **Analysis Approach:**
  - Data Handling;
  - Exploratory Data Analysis (EDA)
  - Data Preparation;
  - Model Building;
  - Model Evaluation;
  - Test Data Predictions;
  - Recommendations.

Data Handling

- For numerical and categorical data, the mode for imputation has been used. Columns with only a single unique value were discarded. In addition, the following operations were conducted, including handling outlier management, correcting invalid entries, grouping low-frequency values, and mapping binary categories.
- Columns containing more than 40% null values were removed. Value counts within categorical columns were reviewed to determine appropriate actions: remove the column with imputation-related skew, add a new category (others), impute high frequency values, and remove irrelevant columns.
- Numerical categorical data were imputed with ‘mode’ value and columns with only one unique response from customer were dropped.
- Additional activities, including handling outliers, correcting invalid data, grouping low frequency values, mapping binary categorical values, were carried out.

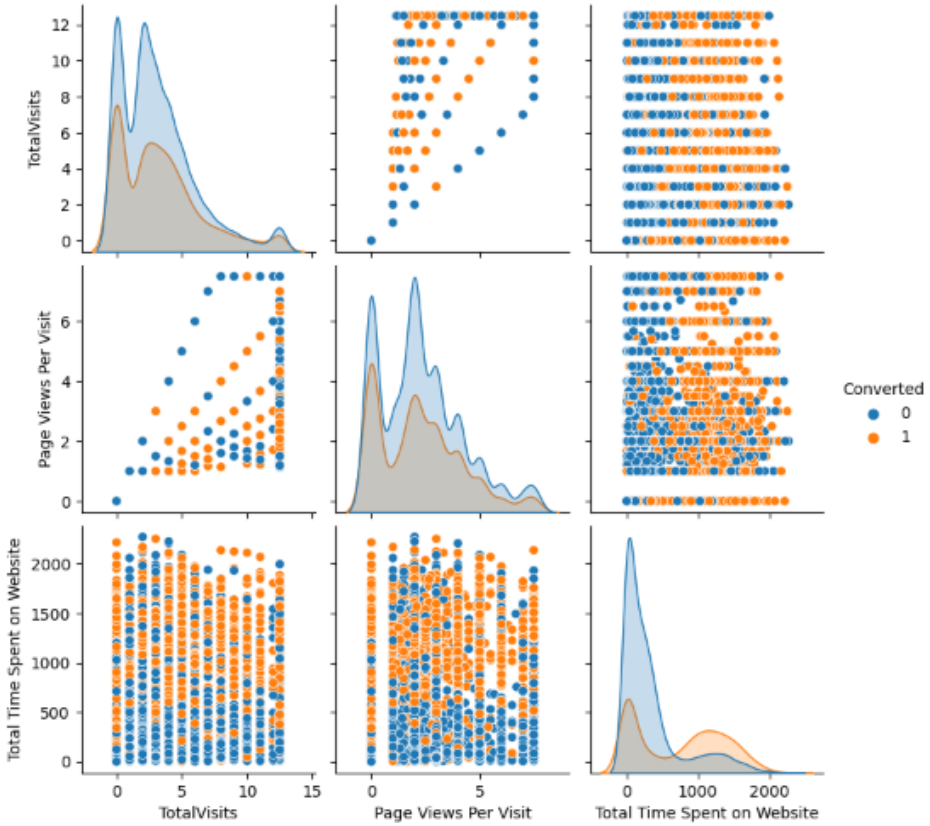
```
def dropNullColumns(data, percentage=40):

    missing_perc_dict = {}

    # Calculate the missing percentage for each column
    for col in data.columns:
        missing_perc_dict[col] = 100 * data[col].isna().mean()

    # Convert the dictionary to a sorted list of tuples
    sorted_missing_perc = sorted(missing_perc_dict.items(), key=lambda x: x[1], reverse=True)

    # Identify columns to be dropped
    col_to_drop = [col for col, perc in sorted_missing_perc if perc >= percentage]

    # Print relevant information
    print("Total columns dropped: ", len(col_to_drop), "\n")
    print("List of columns dropped : ", col_to_drop, "\n")
    print("Shape before dropping columns: ", data.shape)

    # Drop identified columns
    data.drop(labels=col_to_drop, axis=1, inplace=True)

    # Print shape after dropping columns
    print("Shape after dropping columns: ", data.shape)
```

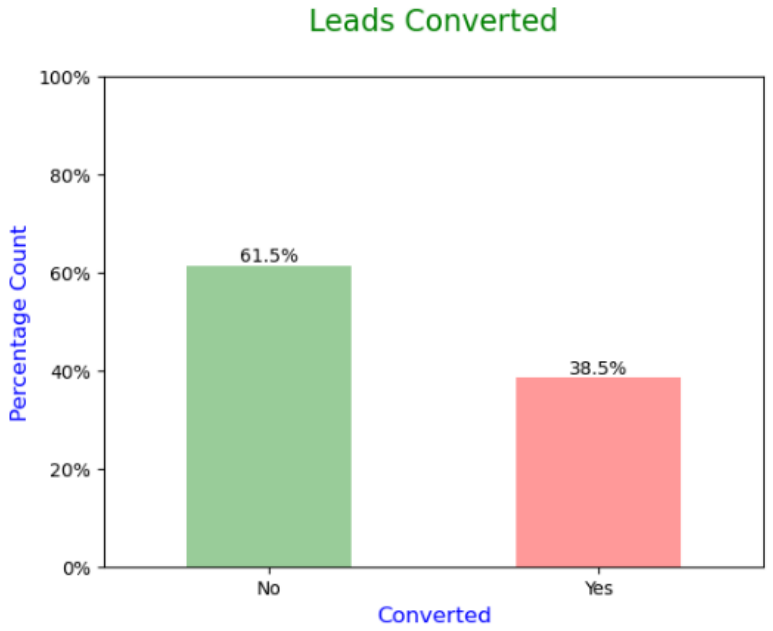
```
missing_values={'Specialization':'Others','Lead Source':'Google','Last Activity':'Email Opened',
                'What is your current occupation':'Unemployed'}
df_leads=df_leads.fillna(value=missing_values)
```

```
def Check_Outliers(data,columnList):

    plt.figure(figsize=[22,11])
    plt.subplots_adjust(wspace=0.4,hspace=0.5)

    for i,j in enumerate(columnList):
        plt.subplot(2,2,i+1)

        sns.boxplot(y=data[j])     # y = df_leads[j] to make plot verticle

        plt.suptitle("\nChecking Outliers using Boxplot",fontsize=20,color="green")
        plt.ylabel(None)
        plt.title(j,fontsize=15,color='red')
```

Exploratory Data Analysis (EDA)

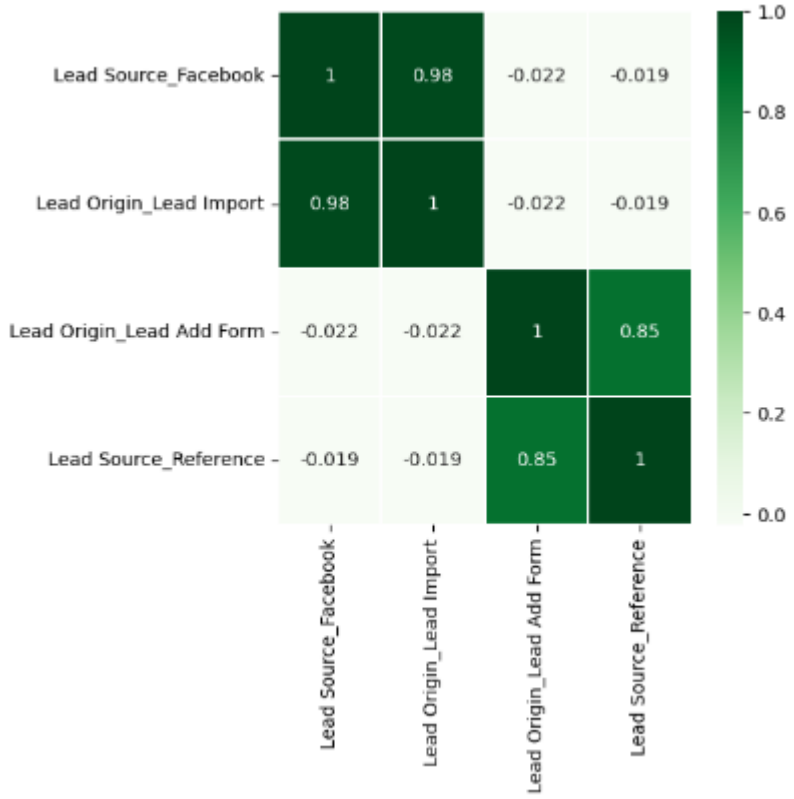
- The data was checked for imbalance, only 38.5% of leads was found converted.





- Univariate and bivariate analyses on categorical and numerical variables ‘Lead Origin’, ‘Current occupation’, ‘Lead Source’, etc. provided valuable insights into the effect on the target variable.
- It was observed that spending more time on the website has a positive impact on lead conversion.

Data Preparation



- Divided the dataset into training and testing sets in a 70:30 ratio, standardized the data for feature scaling, and removed highly correlated columns to avoid redundancy.

```
# Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

```
print("X_train:", X_train.shape,"\ny_train:",y_train.shape)
```

```
X_train: (6468, 48)
y_train: (6468,)
```

```
print("X_test:", X_test.shape,"\ny_test:",y_test.shape)
```

```
X_test: (2772, 48)
y_test: (2772,)
```

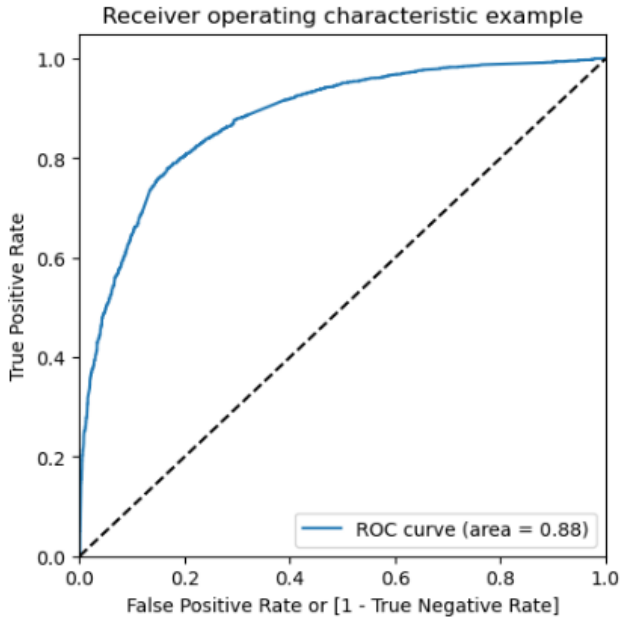
Model Building

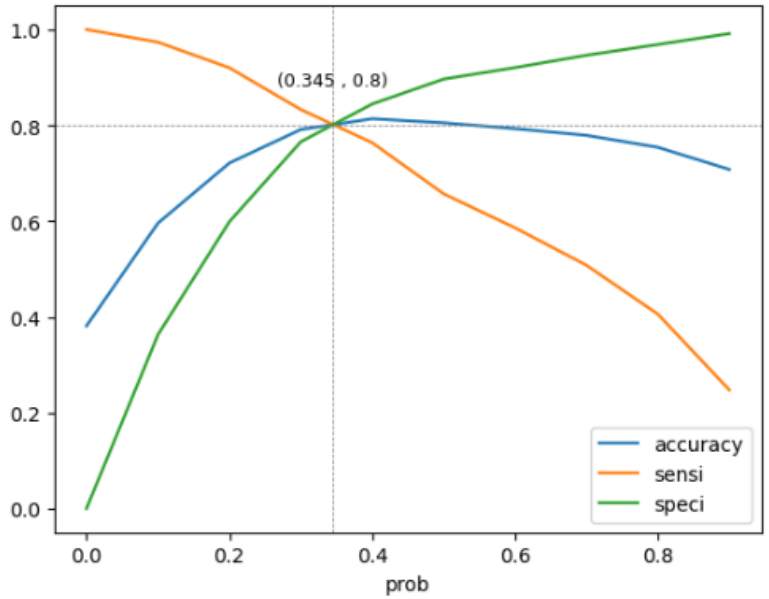
- Used RFE to reduce the number of variables from 48 to 15, making the dataset more manageable.
- Manually excluded variables with p-values of greater than 0.05 to build models.
- A total of 3 preliminary models were built before reaching the final Model 4 which was stable with p-values of less than 0.05 and a VIF of less than 5.
- logm4 was selected for the final model with 12 variables used for training set, testing set, and making predictions.

	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

Model Evaluation

- A confusion matrix was created, and a cut-off point of 0.345 was selected based on the accuracy, sensitivity and specificity plot. This cut-off point resulted in accuracy, specificity and precision values of roughly 80%, whereas the precision-recall view showed lower performance metrics of roughly 75%.
- To address the business problem of boosting the conversion rate to 80%, the sensitivity-specificity view will be selected for the optimal cut-off point used for final predictions due to the drop of metrics when taking the precision-recall view. The lead score assigned to the training data using a cut-off point is 0.345.
-





Test Data Predictions

- The final model was used to make predictions on the test data with scaling and predicting techniques.
- The evaluation metrics for both the train and test data were roughly 80%.
- A lead score was assigned to the data, and the top three features were found to be ‘Lead Source_Welingak Website’, ‘Lead Source_Reference’, and ‘Lead Source_Working Professional’.

Lead Source_Welingak Website	5.388662
Lead Source_Reference	2.925326
Current_occupation_Working Professional	2.669665
Last Activity_SMS Sent	2.051879
Last Activity_Others	1.253061
Total Time Spent on Website	1.049789
Last Activity_Email Opened	0.942099
Lead Source_Olark Chat	0.907184
Last Activity_Olark Chat Conversation	-0.555605
const	-1.023594
Specialization_Hospitality Management	-1.094445
Specialization_Others	-1.203333
Lead Origin_Landing Page Submission	-1.258954

Recommendations

- Target communications through channels, such as SMS and email.
- Follow up with leads who have interacted with X Education through a variety of touchpoints.
- Tailor follow-ups (tracking what leads are doing on the on the Welingak Website) and outreach to leads who have spent a significant amount of time on the website (shooting them a message that indicates directly what they've been browsing and offers some exclusive, limited-time discounts).
- Implement some initiatives related to referrals, for instance, rewards (gift cards, discount coupons, etc.) for referrals, or contests or competitions for existing customers who refer the most customers, etc.
- Use LinkedIn Ads to target working professionals while they are scrolling through LinkedIn, create helpful content that solve real problems for working professionals, and share them on platforms where working professionals hang out online.