# Lead Scoring Assignment From upGrad


Submitted By :

Maddali Sainadh
Naman Arora
Harshit Aggarwal

# INTRODUCTION

In this assignment, we analyzed a lead scoring dataset to help company prioritize its leads based on the probability of conversion. The dataset contains information about different leads, such as their activities on the website, their demographic information, and their communication history with the company.

**Business Problem :**

Our business problem was to help the company prioritize its leads based on the probability of conversion .We wanted to build a model that could predict the probability of a lead getting converted based on the available data.

**Solution:**

We proposed a solution by building a logistic regression model that could predict the likelihood of conversion. We used different feature selection techniques and preprocessing steps to make the model more accurate.

**Business Impact:**

Our solution can help the company prioritize its leads better and allocate its resources more efficiently. By focusing on the leads that are more likely to get converted, the company can improve its conversion rates and increase its revenue. Our model can also help the company identify the areas where it needs to improve its lead generation process.

# TECHNICAL ANALYSIS

**Data Importing and Data Checking :**

•We started by importing the required libraries and the dataset. We used libraries like pandas, numpy, and seaborn to perform data analysis.

•We also used sklearn libraries to preprocess the data and build a logistic regression model. The dataset was checked for its shape, data types, and basic statistical details.

**Data Cleaning and Preprocessing:**

•We removed the unnecessary columns, handled missing values, and performed data imputation. We also performed feature scaling on the dataset.

**Feature Selection and Model Building:**

•We used feature selection techniques like Recursive Feature Elimination (RFE) and variance inflation factor (VIF) to select the most important features for the model. We then built a logistic regression model to predict the likelihood of conversion.

# Exploratory Data Analysis (EDA)

•We performed Exploratory Data Analysis (EDA) on leads dataset. We started by checking for duplicates based on 'Prospect ID' and 'Lead Number' and removes the first two columns ('Prospect ID' and 'Lead Number') as they are unique ids which would not be required for analysis.

•After that we checked the null values for each columns. Then we dropped the columns which have more than 40% null values.

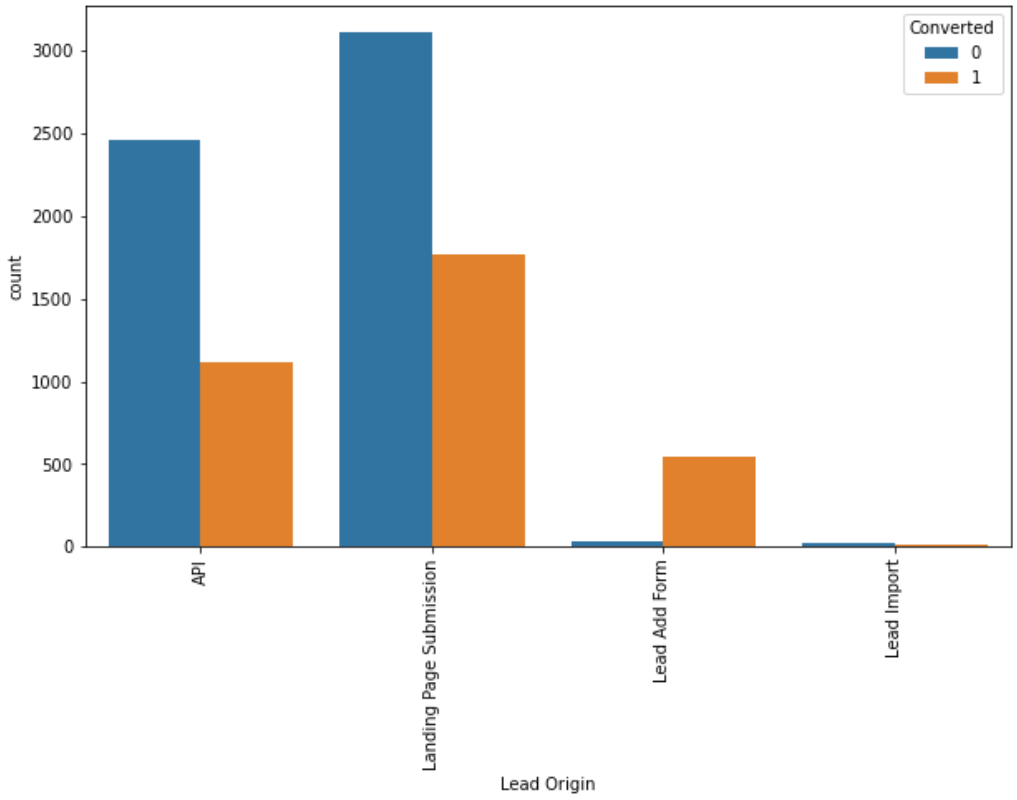
•Then we worked with missing values in the 'Tags', 'Lead Profile', 'What is your current occupation', 'What matters most to you in choosing a course', 'Country', 'How did you hear about X Education', 'Specialization', and 'City' columns.

- •After observations in the 'Tags' column, we replaces the null values with 'Unknown'. In the 'Lead Profile' column, since there are 74% missing values, the column is dropped entirely. In the 'What is your current occupation' column, null values are replaced with 'Unemployed', and in the 'What matters most to you in choosing a course' column, null values are replaced with 'Better Career Prospects'.

- •In the 'City' column, since the value of 'Mumbai' is much higher compared to other values, we replaced null values with 'Mumbai'.

# EDA – Column Wise Analysis

We checked the value counts of the "Lead Origin" column and visualizes it using a plot. It observes that "API" and "Landing Page Submission" bring a higher number of leads, and it needs to focus on improving the conversion rate of these two. Also, it needs to generate more leads from the "Lead Add Form."



# EDA – Column Wise Analysis

- We replaced NaN values and combines low-frequency values in the "Lead Source" column.

- We visualizes the distribution of the "TotalVisits" variable using a boxplot and droped the top and bottom 1% of outlier values.

- Then, we checked the percentile values for the "Total Time Spent on Website" variable and visualizes it using a boxplot.

- We checked the distribution of the "Page Views Per Visit" variable using a boxplot and removes the top and bottom 1% of outlier values. We also checks the distribution of the "Total Visits" variable with the "Converted" variable using a boxplot.

- We then worked with categorical variables by getting dummy variables and dropping the first column.

# Logistic Regression – Model Building

•After EDA analysis and working with categorical variables first we splited the data into a training set and a testing set.

•After that we scaled the numeric columns of the training set using StandardScaler. Next, we used RFE from sklearn to select the important features for the model.

•We then builds the logistic regression model using statsmodels. Then, we droped the column with the highest p-value, rebuilds the model, and checks the VIF. This process is repeated until all features have a p-value less than 0.05 and VIF is low.

•The final model is built, where it derives the probabilities, lead score, and predictions on the training data.

•We used confusion_matrix to evaluate the performance of the model on the training data. The we ploted the ROC curve.

# Final Observation:

**On Train Data-**

- Accuracy: 92.29%
- Sensitivity: 91.70%
- Specificity: 92.66%

**On Test Data-**

- Accuracy: 91.34%
- Sensitivity: 92.15%
- Specificity: 90.89%

**The model appears to accurately predict the conversion rate, which should give the CEO confidence in making informed decisions based on this model's predictions.**

# INSIGHTS

**The following are the categorical variables in the model which should be focus by the sales most on in order to increase the probability of lead conversion**

•**Lead Source_Reference:** This variable indicates that leads that come from a reference or referral have a higher probability of conversion. Therefore, the sales team should focus on nurturing and prioritizing such leads.

•**Lead Source_Social Media:** This variable suggests that leads that come from social media platforms have a higher probability of conversion. Hence, the sales team should focus on utilizing social media channels to generate and nurture leads.

•**Lead Source_Olark Chat:** This variable indicates that leads that initiate a chat on the website have a higher probability of conversion. Therefore, the sales team should prioritize and focus on leads that engage in chats on the website.

# INSIGHTS

**The following are the variables in our model which contribute most towards the probability of a lead getting converted**

**Total Time Spent on Website:** This variable has a positive contribution towards lead conversion. The longer a lead spends on the website, the higher the probability of conversion. Therefore, the sales team should focus on nurturing such leads.

**Lead Source_Reference:** This variable also has a positive contribution towards lead conversion. If the lead source is from a reference or referral, the probability of conversion is higher, as referrals provide cashback incentives and the assurance of trusted sources. The sales team should prioritize such leads.

**What is your current occupation_Student:** This variable has a negative contribution towards lead conversion. If the lead is already a student, the probability of them taking up another course designed for working professionals is low. Therefore, the sales team should not focus on such leads.