LEAD SCORING CASE STUDY

Manish Kumar
Mohammad Hussain
Chhavi Mehndiratta

AGENDA

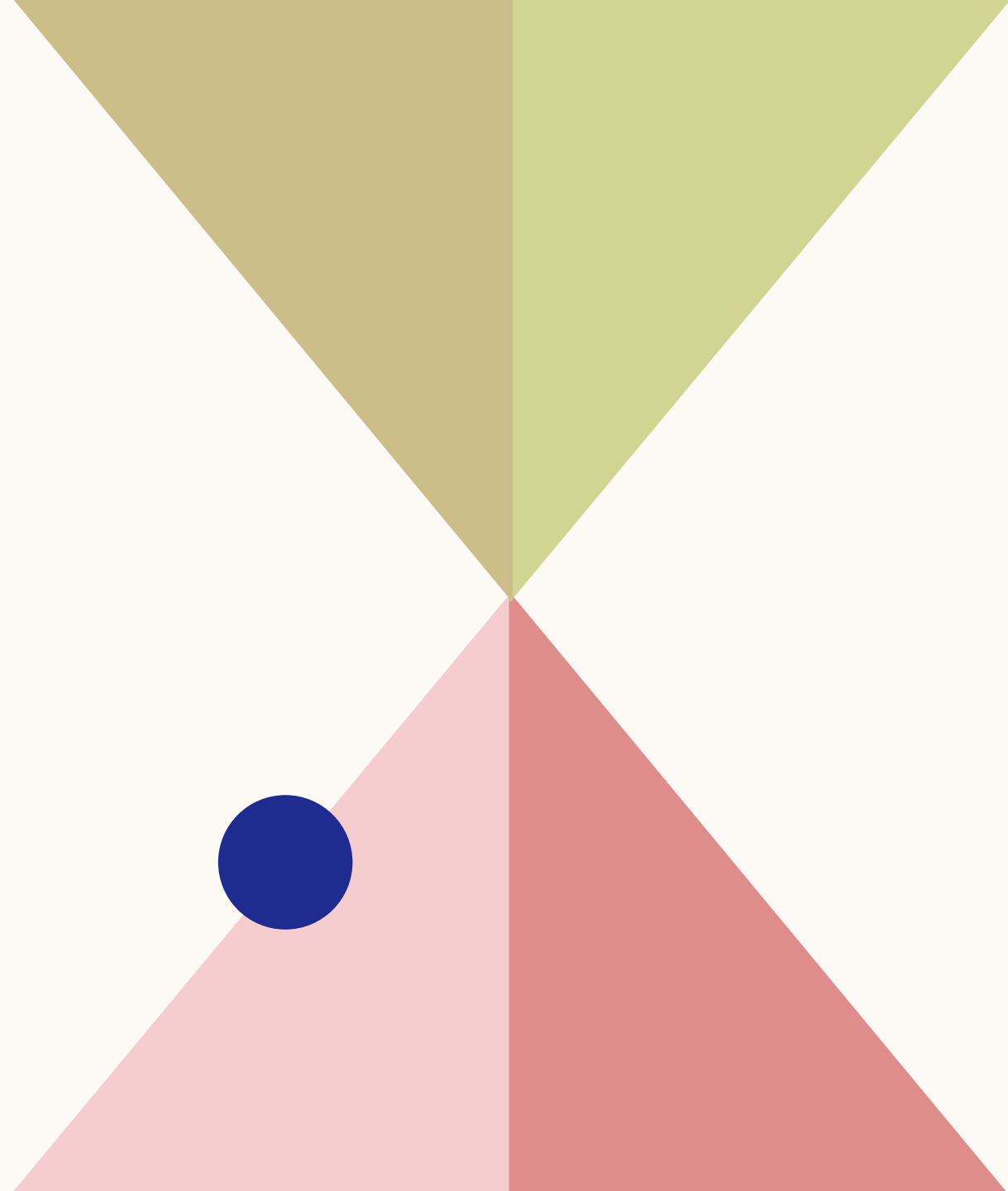
Problem Statement

Methodology

Data Manipulation

Exploratory Data Analysis

Model Building



PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

BUSINESS OBJECTIVE

- X education wants to know most promising leads.
- Building a suitable model which identifies the hot leads.

METHODOLOGY

- **Data cleaning and data manipulation.**
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- **Exploratory data Analysis**
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- **Feature Scaling & Dummy Variables and encoding of the data**.
- **Classification technique:** logistic regression used for the model making and prediction.
- **Validation of the model.**

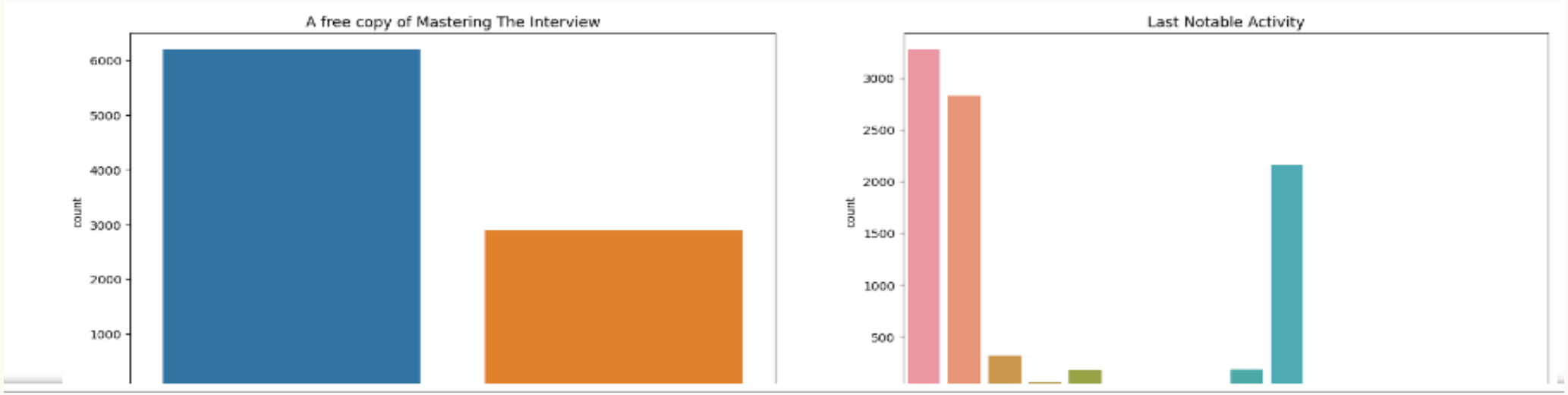
DATA CLEANING AND MANIPULATION

- All the values were converted into lower base for consistency
- Some binary variables (Yes/No) were converted into numeri variables 0/1
- Columns which had only one unique value were dropped off
- Columns having minimum 35% Null values were dropped off
- For, the columns with less than 35% Null values, NaN was converted into ‘Select’ and ‘not provided’ option as NaN means that no option was selected.

PERFORMING EDA (1/2)

Univariate Analysis

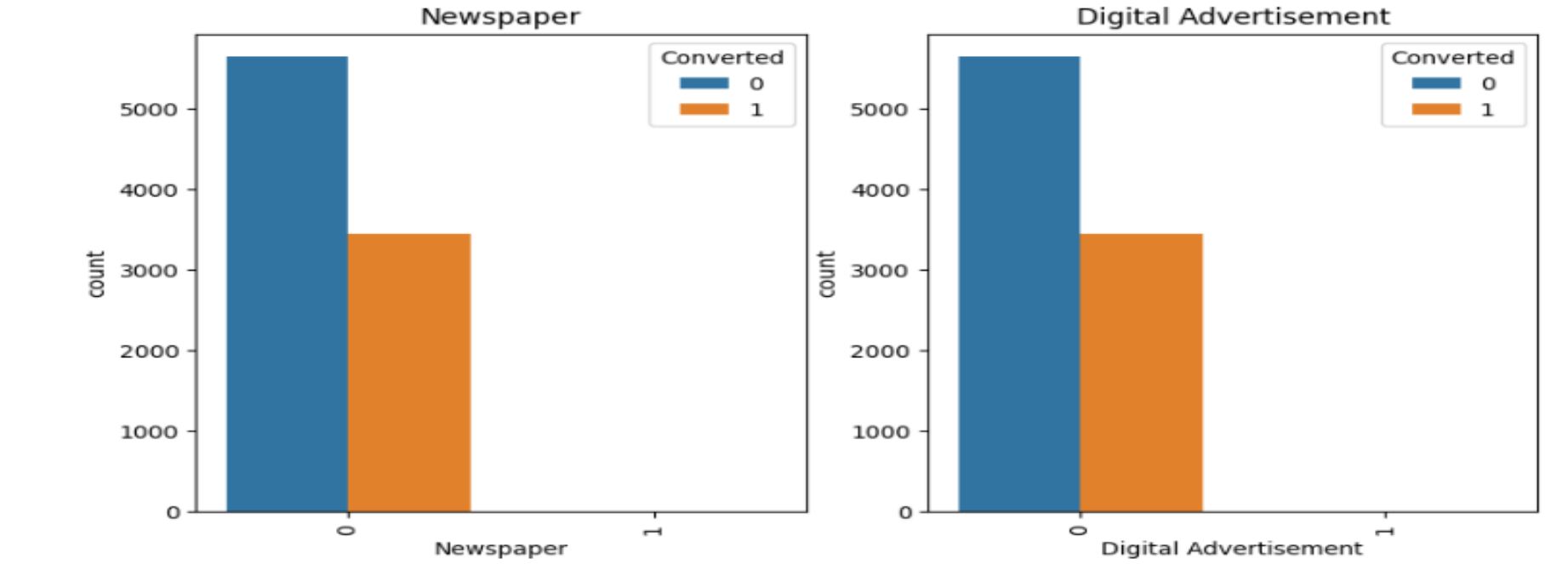
In Univariate analysis, mainly count plots and histograms were plotted to ge the understanding of the counts and frequencies of the different variables present in the data



PERFORMING EDA (2/2)

Bivariate Analysis

In Bivariate analysis, count plots were plotted keeping ‘Converted’ as the target variable and other independent variables



DATA CONVERSION

- **Before initiating the model building, the dummy variables were created using ‘get dummies’**
- **The data type of Boolean columns was also converted into integer type**

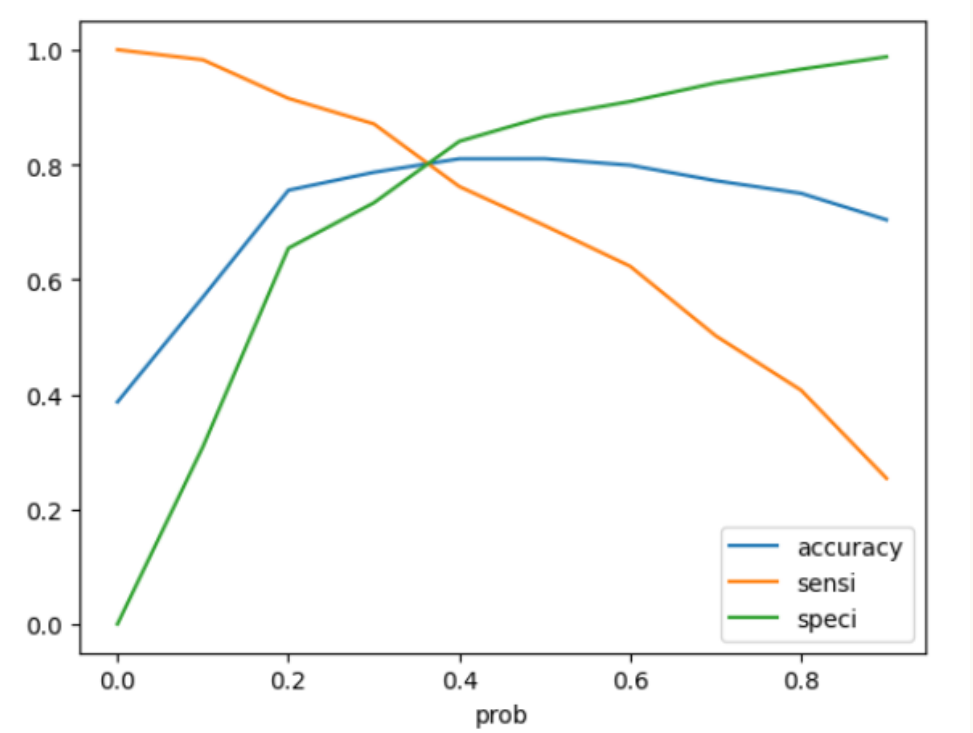
MODEL BUILDING (1/3)

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Feature scaling
- Checking correlation of variables using heat maps
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predicting the probabilities on train set and evaluating the model using accuracy, sensitivity specificity
- Predictions on test data

MODEL BUILDING (2/3)

```
# Plotting the same


cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```




```
# Calculating the sensitivity
TP/(TP+FN)
```

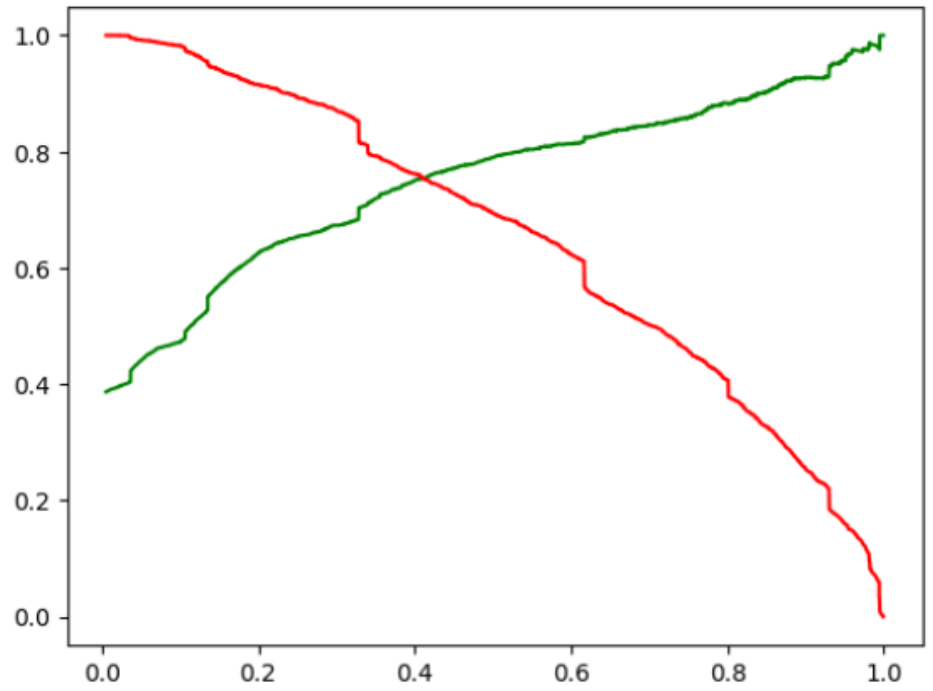
```
0.7927524429967426
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.8048780487804879
```

```
## with present cut off set as 0.35 we have sensitivity at 79% and specificity at 80%
```

MODEL BUILDING (3/3)




```
# Precision = TP / TP + FP
TP / (TP + FP)
```

```
0.7313725490196078
```

```
#Recall = TP / TP + FN
TP / (TP + FN)
```

```
0.7620020429009193
```

```
## With the  cut off as 0.41 we have Precision around 73% and Recall around 76%
```

KEY INSIGHTS

The variables identified as most influential in predicting potential buyers, ranked in descending order of importance, are:

- Total time spent on the website.
- Total number of visits.
- Lead source, with emphasis on Google, Direct traffic, Organic search, and Welingak website.
- Last activity, particularly SMS and Olark chat conversation.
- Lead origin as Lead add format.
- Current occupation as a working professional.

In conclusion, focusing efforts on these key variables presents an opportunity for X Education to significantly enhance its conversion rates by effectively engaging and persuading potential buyers.

THANK YOU