LEAD SCORING CASE STUDY

PREPARED BY ANISH LAKHOTIYA, KUMAR SHUBHAM AND ANKITA KAREKAR

PROBLEM STATEMENT:

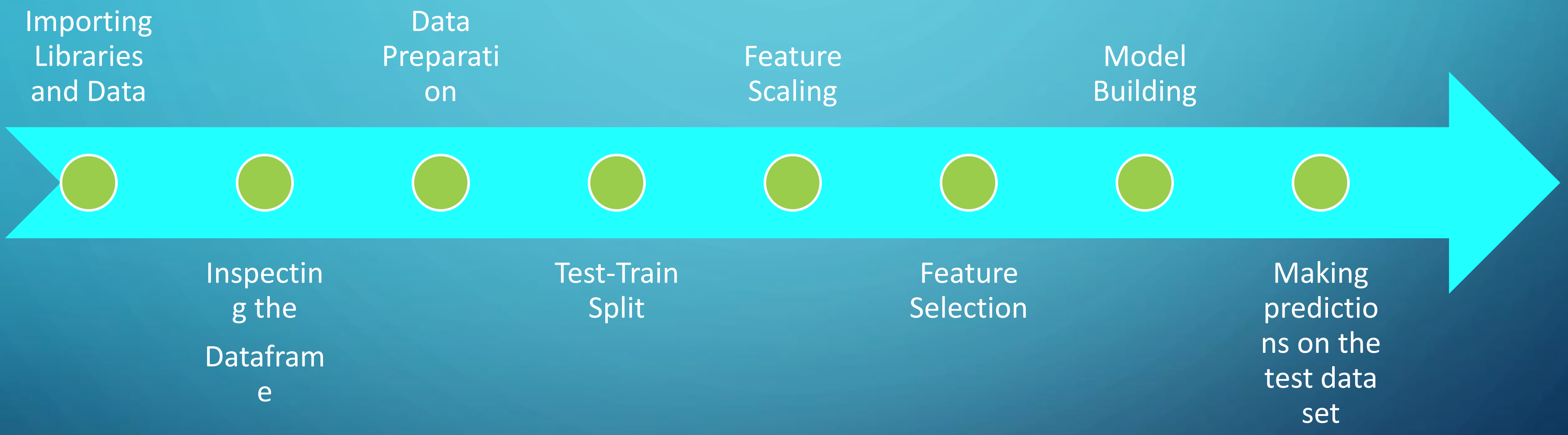
- X Education engages in the sale of online courses to industry professionals, employing various marketing channels such as websites and search engines. Upon users expressing interest by filling out a course form, they are categorized as leads. The sales team then initiates contact to convert these leads into students, with a standard conversion rate of approximately 30%.
- To enhance the lead conversion rate, the company is eager to identify high-potential leads, referred to as 'Hot Leads.' Pinpointing this subset enables the sales team to concentrate efforts on communicating with these promising leads.
- The company's objective is to construct a model that assigns a lead score to each prospect.
- The targeted lead conversion rate is set at 80%.

A typical lead conversion process

GOAL OF THE CASE STUDY: 

- The primary goal is to develop a logistic regression model that allocates a lead score ranging from 0 to 100 to each lead. Higher scores indicate a greater likelihood of conversion, while lower scores suggest leads that are less likely to convert.
- Business Problems to Address:
  - a) Identify the top 3 variables contributing the most to the probability of lead conversion.
  - b) Identify the top 3 categorical variables that should be prioritized to increase the probability of lead conversion.
  - c) Develop a strategy to convert potential leads predicted by the model effectively.
  - d) Propose a strategy to minimize the occurrence of unnecessary phone calls in the conversion process, except when absolutely necessary.

MODEL BUILDING WORKFLOW



STEP 1: IMPORTING LIBRARIES AND DATA

- The following Python libraries were imported:

  - Data analysis– 1) Numpy 2) Pandas
  - Data visualization– 1) Matplotlib 2) Seaborn
  - Machine Learning– 1) Statsmodels 2) Scikit Learn

- The datafile “Leads.csv” was uploaded to start the EDA process. Upon uploading the dataframe was stored as “lead_df”. The dataframe was inspected using head() function. The data frame contains 37 columns.

STEP 2: INSPECTING THE DATAFRAME

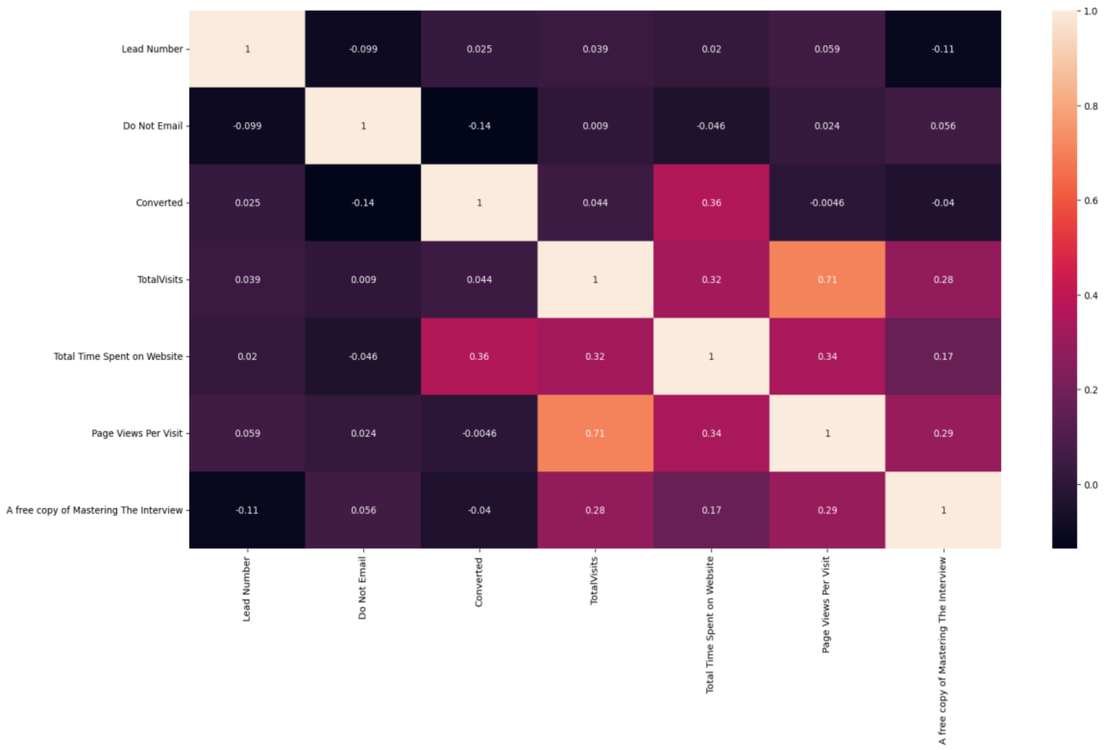
- The shape of the DataFrame (9240 rows, 37 columns) was examined using the shape attribute.
- To understand the data type of each column and number of missing values in each column, info() function was used The data type of 3 columns is “ type, 4 columns are “ type and rest of the 30 columns are of “ type
- The info() function revealed column data types and identified missing values in 17 columns.
- Descriptive statistics of numerical columns were obtained using the describe() functions.

STEP 3: DATA PREPARATION

- Columns with values like "Select" were considered as not selected; these were replaced with NaN.
- Columns with 35 missing values were dropped.
- Highly skewed columns were dropped as they contribute less to model preparation.
- A separate category was created for columns with numerous categorical values.
- Missing values in categorical columns were imputed with mode values, and numerical columns with median values.
- Outliers in numerical columns were identified and replaced with values capped at the 99th percentile.
- Columns acquired after lead identification were dropped

STEP 4: DATA PREPARATION (CONTINUED)

- Correlation relation between numerical variables was visualized using a Correlation Heatmap.
- A of 0.71 was observed between Total Visits and "Page Views Per Visit."



Correlation Heatmap of numerical columns

- Dummy columns were created for categorical variables.
- Dummy dataframes were concatenated, and redundant columns were dropped.
- The target variable "Converted" was removed and assigned as "y"; the rest of the dataset was assigned as "X.”

STEP 4: TEST-TRAIN SPLIT OF DATASET

- The dataset was split into train and test sets using train_test_split.
- 70% of the data was allocated for training, and 30% for testing the model performance.
- random_state was et to 100 for result consistency across runs.

STEP 5: FEATURE SCALING

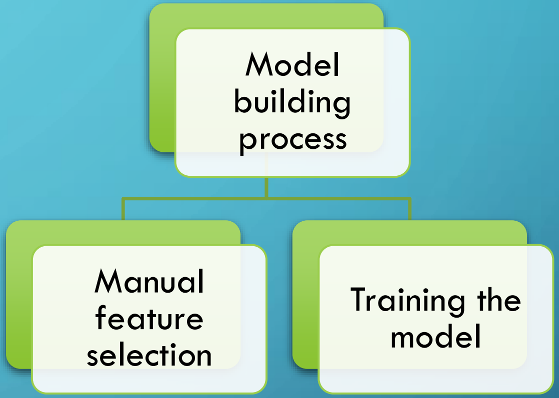
- Numerical columns (TotalVisits, Total Time Spent on Website, Page Views Per Visit) were scaled using Standard Scaler.
- The existing lead conversion rate was determined as 38.5% for sanity check

STEP 6: FEATURE SELECTION USING RFE

- The automated feature selection technique from Scikit Learn library was used
- Recursive Feature Elimination (RFE) selected the most relevant 15 variables from X_train.
- Manual elimination based on p-values and Variance Inflation Factor (VIF) was integrated into the model building process.

STEP 7: MODEL BUILDING

- Stats models API was used for model building.



- Feature selection/elimination based on p-values and VIF values.

	p value > 0.05	*p value < 0.05*
VIF value > 4	Drop First	Drop Last
VIF value < 4	Drop Second	Retain

STEP 7: MODEL BUILDING CONTD…

- The manual feature elimination process was continued till “p values” and “ values of all features become less than 0.05 and 4 respectively
- The model was considered to be satisfactory
- The final model was used to predict the lead conversion
- To get the optimum probability cutoff, different probabilities are used (from 0.0 to 0.9)
- The optimum probability cutoff was found to be 0.25
- The accuracy of the model is 77.55% whereas the sensitivity or recall value is 78.22%
- The values of 9 features in the final model are as under

Feature	p value	VIF
Do Not Email	0.000	1.11
Total Visits	0.000	1.42
Total Time Spent on Website	0.000	1.25
Lead_Origin_API	0.000	3.03
Lead_Origin_Landing Page Submission	0.000	1.70
Lead_Source_Google	0.000	1.90
Lead_Source_Olark Chat	0.000	2.94
Lead_Source_Reference	0.012	1.17
CO_Working Professional	0.000	1.18

STEP 8: MAKING PREDICTIONS ON THE TEST DATASET

- The final model predicted lead conversion cases.
- Lead scores were determined based on conversion probability.
- Model accuracy and sensitivity were found to be 76.62% and 77.80%, respectively.
- β values were determined to understand feature contribution, high beta value denotes the stronger contribution to the target variable
- With the top three contributing variables being "Lead_Origin_Landing Page Submission" (3.46), "Lead_Origin_API" (3.44), and "CO_Working Professional" (2.86