LEAD SCORING CASE STUDY

CAPSTONE PROJECT

DATA SCIENCE / ANALYTICS + ML/AI CERTIFICATION PROGRAM AT UPGRAD

YASH JAIN
JULY 2022
COHORT 4332

PROBLEM STATEMENT

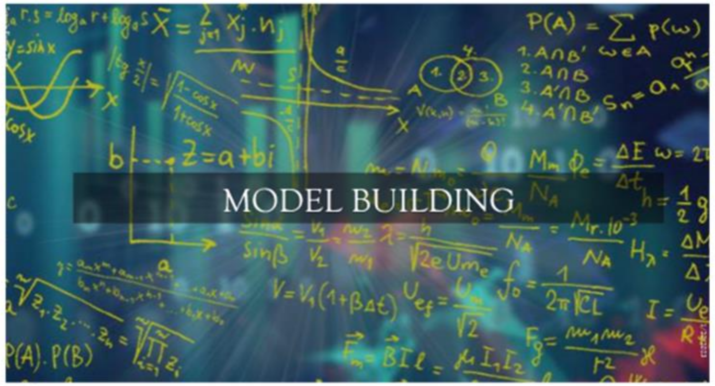
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

(Method to be used – Logistic Regression)

INDEX

DATA UNDERSTANDING

DATA CLEANING AND PREPARATION


MODEL BUILDING

MODEL EVALUATION

PREDICTIONS

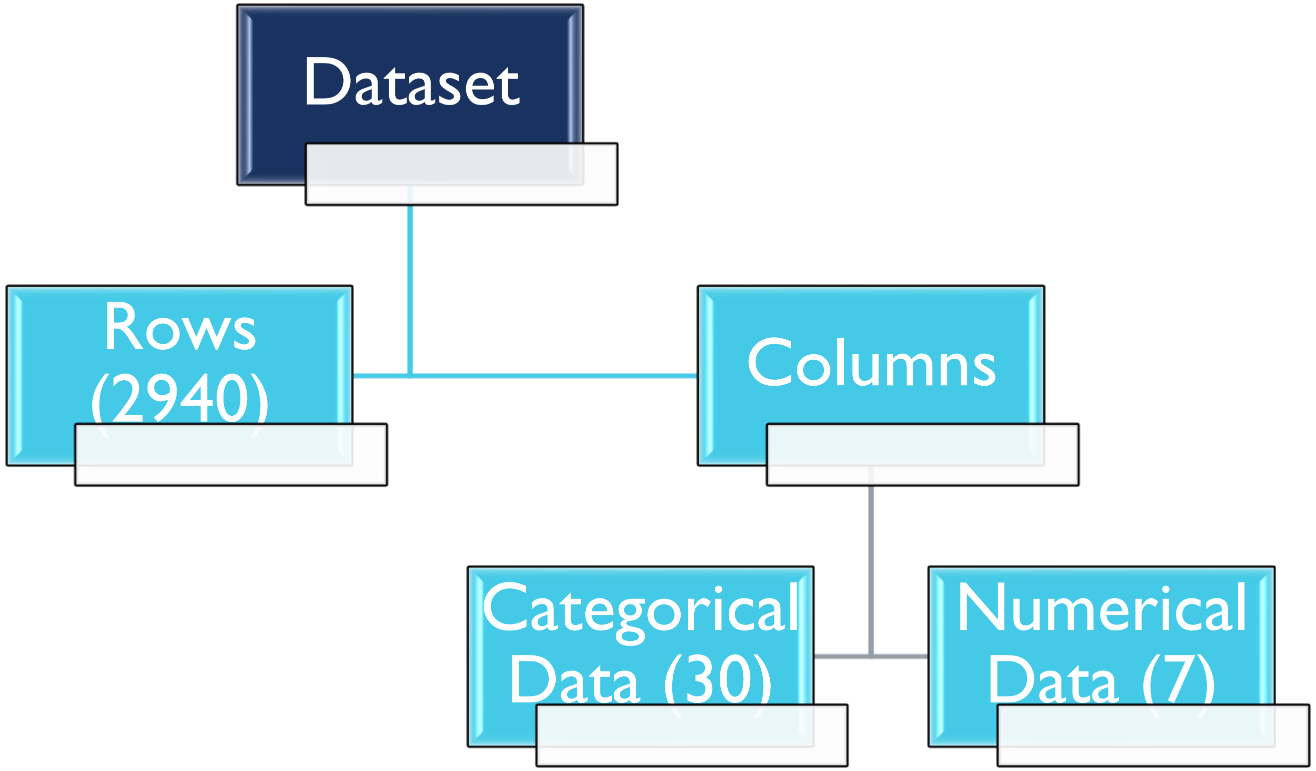
CONCLUSION

DATA UNDERSTANDING

DATA UNDERSTANDING

The dataset contains 9240 entries in total, and 37 attributes, out of which 30 are categorical attributes, and 7 are numerical attributes.

The target variable is Converted which is a binary data (0 for False and 1 for True).



Dataset

Rows (2940)

Columns

Categorical Data (30)

Numerical Data (7)

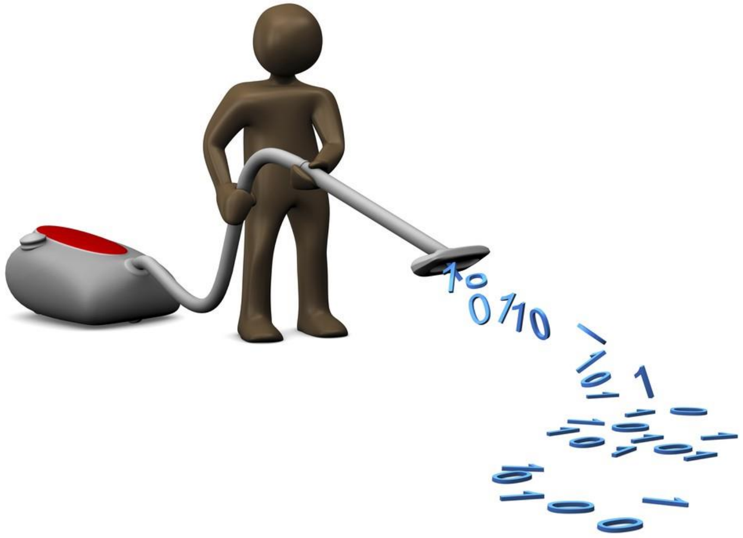
DATA CLEANING AND PREPARATION

DATA CLEANING

We looked for missing values in the dataset, and dropped the attributes containing more than 3000 missing values (~30%).

We also dropped the empty rows from the remaining features, till we had no missing values.

After checking the percentage of data retained, it came out to be around 69%.



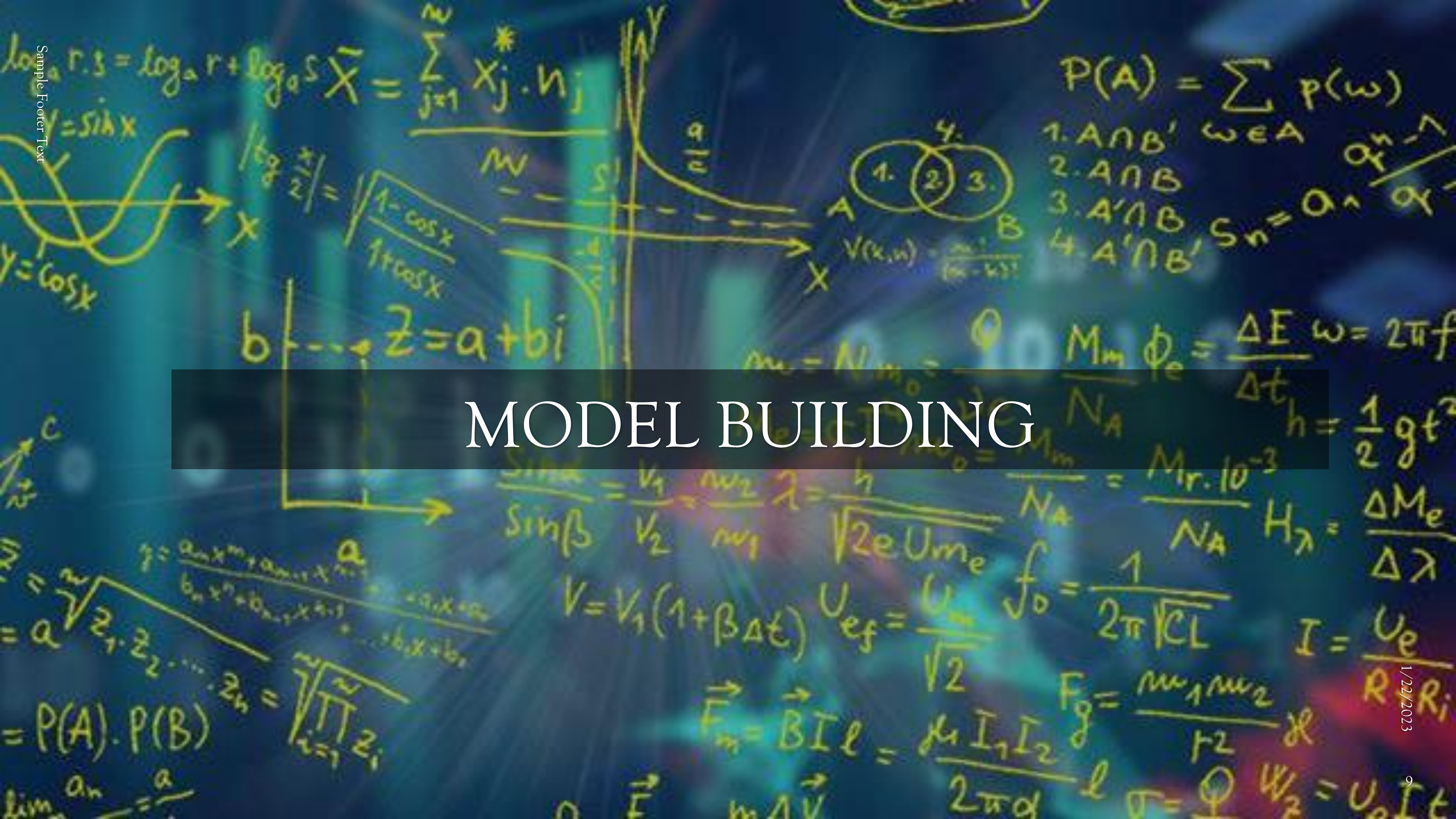
DATA PREPARATION

Divided our dataset first into the variables X (all features except target variable) and Y (target variable – ‘Converted’).

Split out X and Y variables into training and testing sets of 70:30 ratio, using train-test-split and then –

1. Used standardized scaling using StandardScaler to scale numerical variables
2. Used dummy variables for categorical features to prepare them for model building.

After preparation, we came out to have 75 features in total.



$\log_a r.s = \log_a r + \log_a s$
$\tilde{X} = \sum_{j=1}^{\tilde{n}} x_j^* . n_j$
$y = \sin x$
$y = \cos x$
$\left|\tg \frac{x}{2}\right| = \sqrt{\frac{1-\cos x}{1+\cos x}}$
$P(A) = \sum_{\omega \in A} p(\omega)$
1. $A \cap B'$
2. $A \cap B$
3. $A' \cap B$
4. $A' \cap B'$
$z = a + bi$
$\omega = 2\pi f$
$\Phi_e = \frac{\Delta E}{\Delta t}$
$h = \frac{1}{2} g t^2$
$\frac{\sin\alpha}{\sin\beta} = \frac{v_1}{v_2} = \frac{m_2}{m_1}$
$\lambda = \frac{h}{\sqrt{2eUm_e}}$
$H_\lambda = \frac{\Delta M_e}{\Delta \lambda}$
$V = V_1(1 + \beta\Delta t)$
$U_{ef} = \frac{U_m}{\sqrt{2}}$
$f_0 = \frac{1}{2\pi\sqrt{CL}}$
$I = \frac{U_e}{R + R_i}$
$F_g = \frac{m_1 m_2}{r^2} \varkappa$
$\vec{F}_m = BIl = \frac{\mu I_1 I_2}{2\pi d} l$
$= P(A) . P(B)$

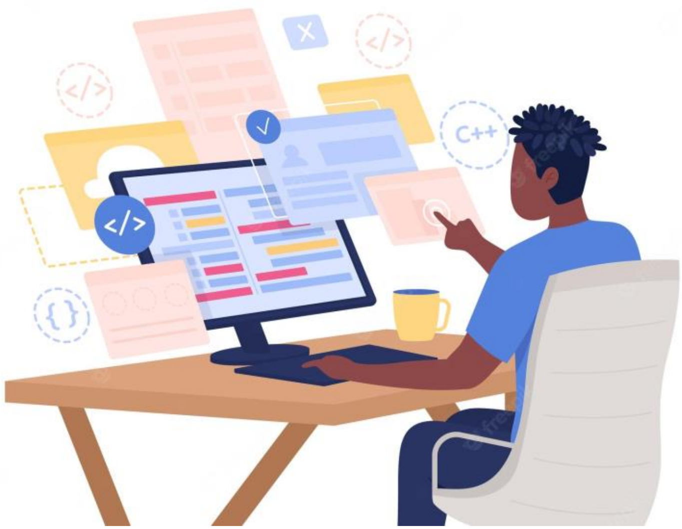
MODEL BUILDING

MODEL BUILDING

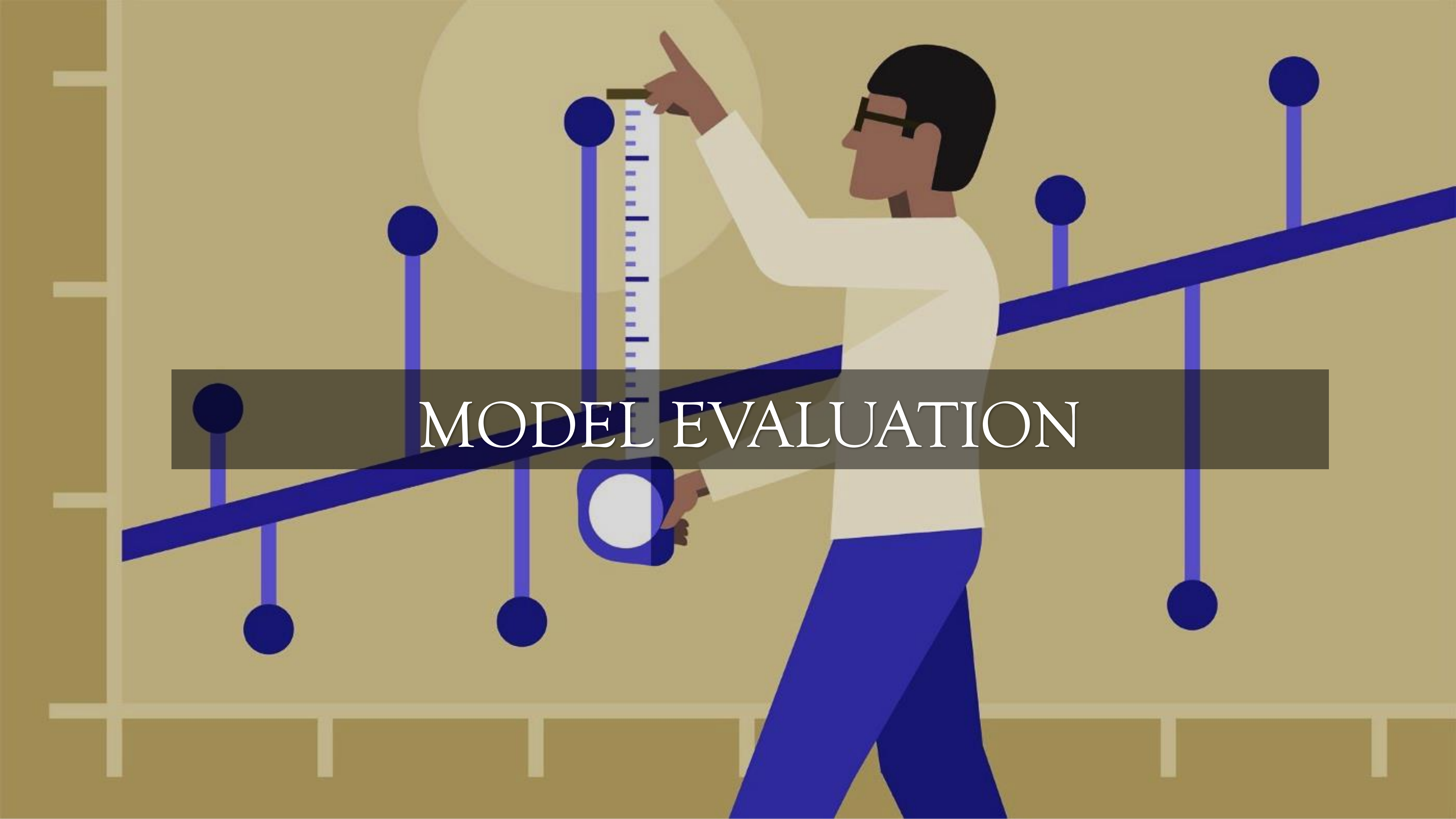
Used RFE (Recursive Feature Elimination) for selecting 15 features out of those 75 features, using the RFE Module in Python.

Used the Statsmodels.api library to import logistic regression model object, and used it to create a logistic regression model, after adding a constant to our training set (X_train).

Checked the p-values (<=0.05) and VIF values (>=5) for all features in the model, and kept modifying it till the optimal conditions are met.



MODEL EVALUATION

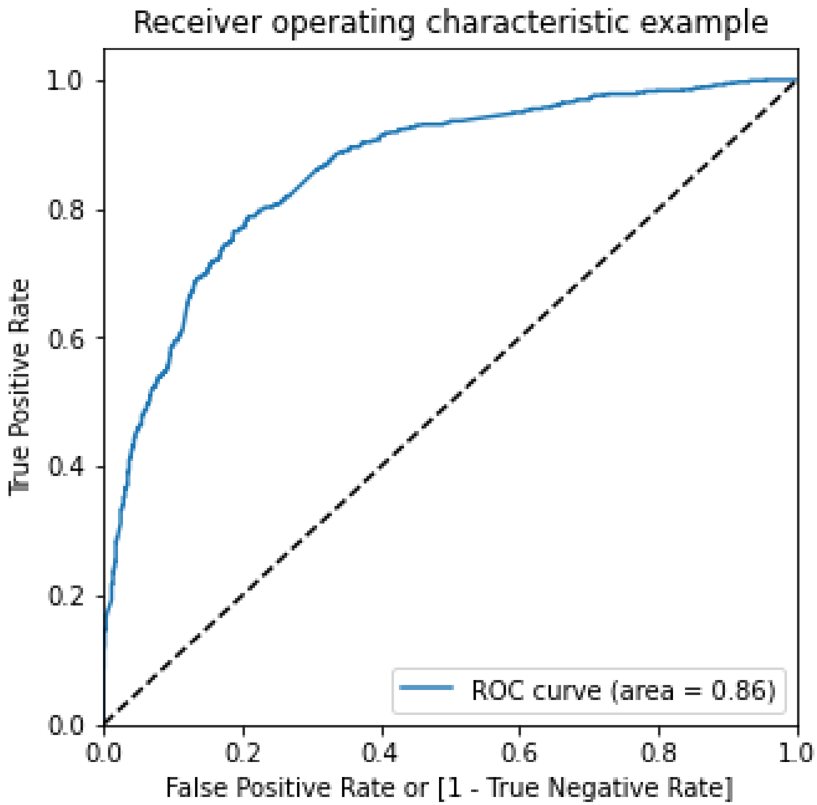


MODEL EVALUATION

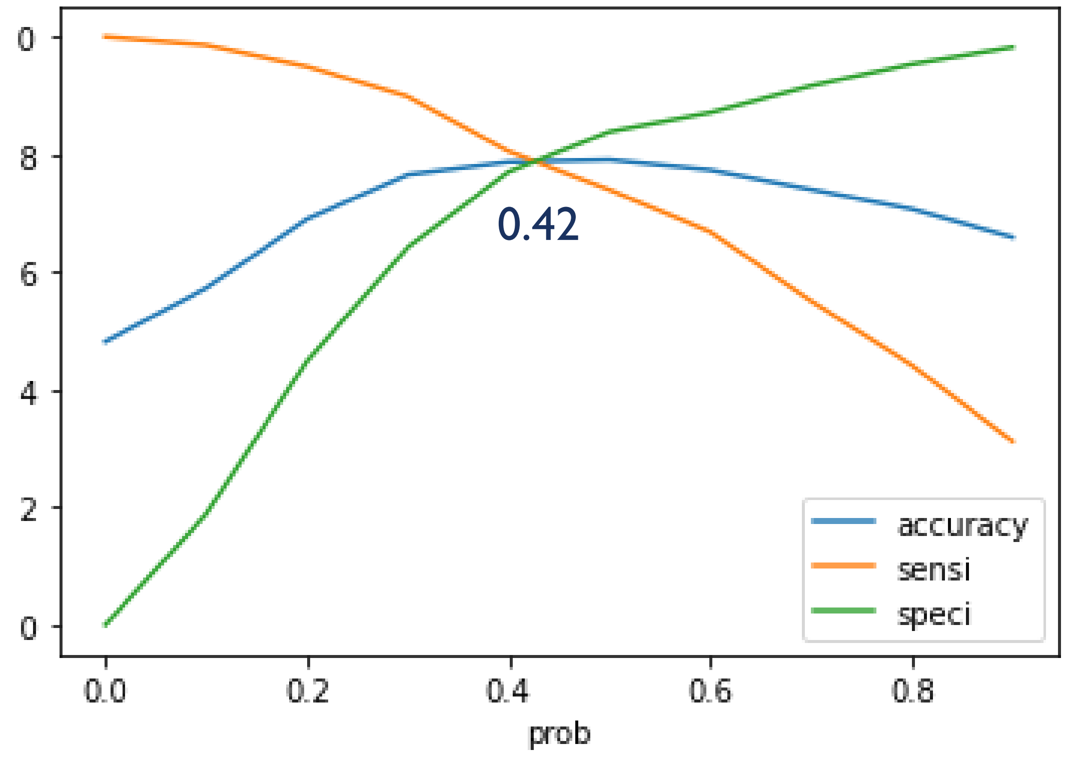
Used our model to predict the values for training set, and compared them to the training values of the target variable 'Y' to evaluate the model. This was done on 3 fronts using the Confusion Matrix –

1. Accuracy
2. Sensitivity
3. Specificity

We used an arbitrary cut-off of 0.5 to predict the expected values, from the probabilities that were predicted via the model. If prob <=0.05, value will be 0 otherwise it will be 1.



CUTOFF POINT



Used the ROC Curve (Receiver Operating Characteristic) and the specificity – sensitivity trade-off to find the optimal cut-off point for our model, which came out to be 0.42.

The ROC Curve returned an area of 0.86, which showed that the model was working efficiently.

PREDICTIONS

MAKING PREDICTIONS

Used the cut-off point to make predictions on our training and our test sets, and compared the model performance for both -

Property	Training Set	Test Set
Accuracy	78.995 %	78.922 %
Sensitivity	78.734 %	78.820 %
Specificity	79.238 %	79.016 %

CONCLUSION

CONCLUSION

1. Handled data (cleaning and preparation) as per requirement for modelling
2. Used dummy variables and scaling after train-test-split to prepare the data for modelling.
3. Used logistic regression analysis technique to create a model for predicting the lead conversion probability.
4. Evaluated the model on the basis of accuracy, sensitivity, and specificity for both training and test sets.

YASH JAIN
JULY 2022
COHORT 4332