# Logistic Regression

Lead Scoring Case Study

# Problem Statement

- X Education is a company providing online courses directed at professionals
- Their website is promoted at various online avenues. Interested people visit the website and fill up a form. These are classified as leads
- The company contacts these leads to convert them towards completing purchase of the course
- The current strategy is to contact every single lead, which is very resource intensive
- Moreover, there is a very low conversion rate of only 30% even after contacting all the leads
- Hence, the company requires a solution that can help identify leads most likely to convert so they can focus their resources on them

# Data Exploration and Cleaning

- The data file was perused and the data dictionary was used to understand the different columns
- A brief EDA of the dataset was done to identify columns with missing values and irrelevant columns that could be dropped
- A number of categorical cells contained the value ‘select’. This was replaced with null value
- Missing values were dealt with by:
  - Deleting columns with large percentage of missing values
  - Deleting rows for those columns where a very small percentage of the data was missing
  - Creating a value “missing” where neither deletion not imputation made sense

# Data Exploration and Cleaning

- There were a number of columns containing Yes/No values. These were replaced with binary 1/0 numerical values
- Columns where the data was dominated by a single value were dropped as these did not add value to the analysis
- Certain categorical values had a large number of options in the data. Bucketing was done for these to consolidate options with lower percentage representation
- Finally, some column names were modified to enhance readability

# Data Preparation

- The categorical columns were converted into dummy variables with n-1 columns created for n options in the categorical columns. Original columns with categorical data were then dropped
- The dataset was split into train and test sets. 70% was taken as training set and 30% for test set
- The training dataset was scaled to address the large variation in numerical data values. MinMax scaler was used in this case
- The conversion rate of the train dataset was checked to ensure there was no class imbalance. A decent rate of 38.5% was obtained

# Model Building

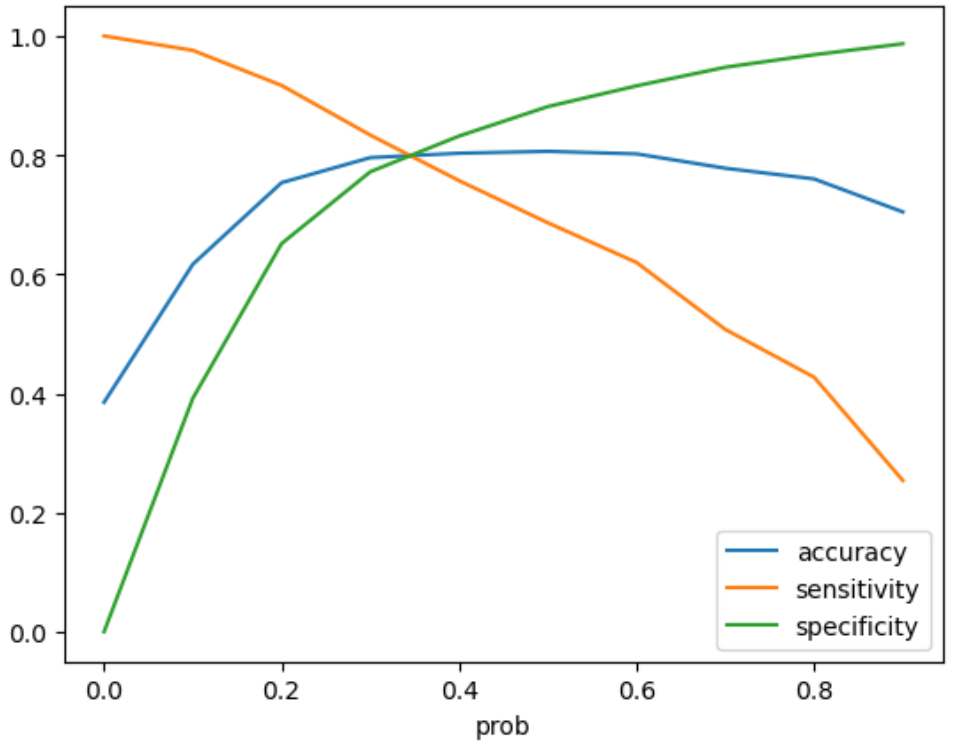
- The training dataset was split into X (independent variables) and y (target variable)
- Recursive feature elimination was used to automate model creation and obtain the best features. Overall 10 features were identified
- These features were then used to create the logistic regression model in StatsModel
- The intention was to fine-tune the independent variables basis the p-value and VIF obtained. However, in the first iteration itself all variables had p-value < 0.05 and VIF < 2 so this model was finalised for further evaluations

# Model Evaluation

- Predicted conversion probability values were obtained using the model
- In order to assign binary values to conversion prediction, a cutoff point had to be estimated
- This was done by obtaining binary prediction value at different cutoff points and calculating the model accuracy, sensitivity and specificity at these values
- The value were plotted to see at what cutoff could the metrics be optimised

# Model Evaluation

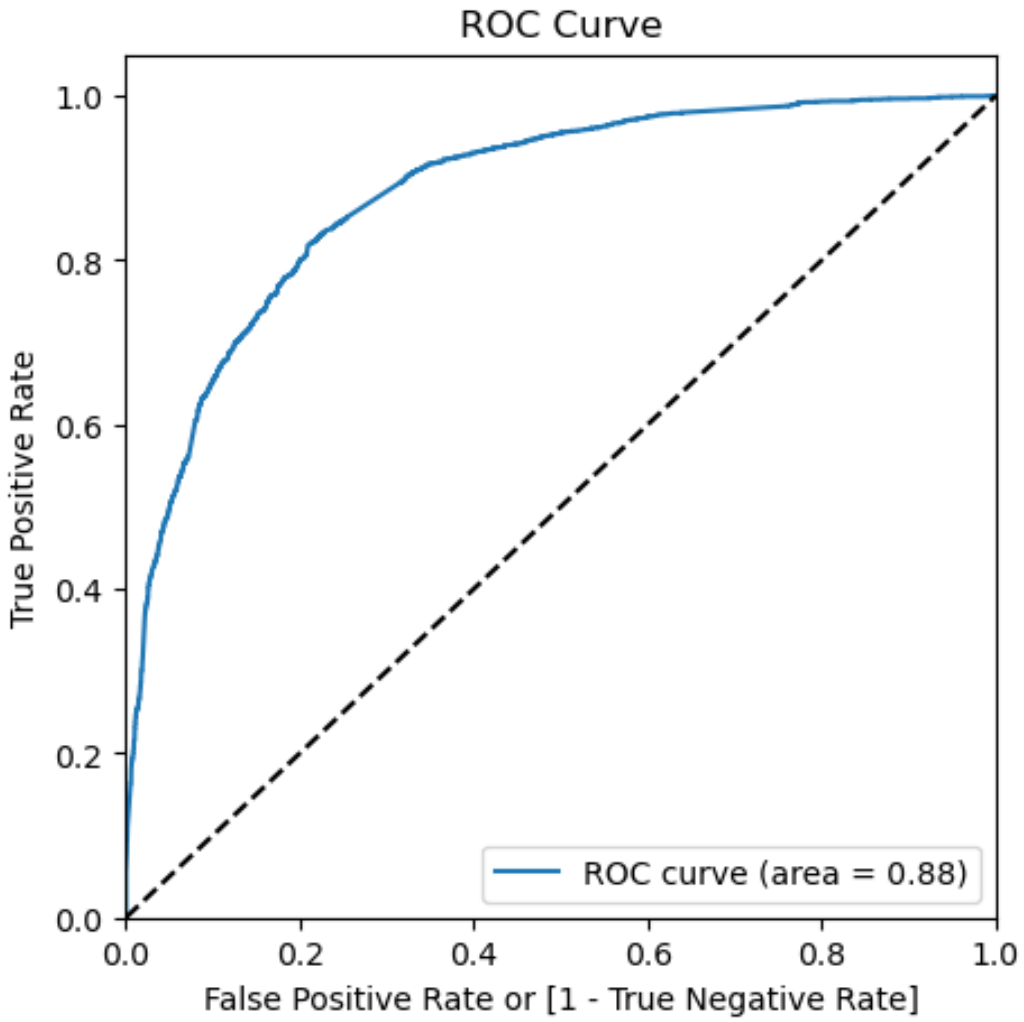
| prob | accuracy | sensitivity | specificity |
|---|---|---|---|
| 0.0 | 0.385136 | 1.000000 | 0.000000 |
| 0.1 | 0.616753 | 0.975879 | 0.391805 |
| 0.2 | 0.753740 | 0.917007 | 0.651472 |
| 0.3 | 0.795780 | 0.833606 | 0.772087 |
| 0.4 | 0.803181 | 0.757155 | 0.832010 |
| 0.5 | 0.806330 | 0.686427 | 0.881434 |
| 0.6 | 0.802078 | 0.619787 | 0.916261 |
| 0.7 | 0.777988 | 0.507359 | 0.947503 |
| 0.8 | 0.760038 | 0.427228 | 0.968502 |
| 0.9 | 0.704613 | 0.253884 | 0.986940 |




- Of the three metrics, sensitivity was focussed on the most as the goal of the model is to correctly identify conversions (i.e. True Positives)
- Using this, 0.3 was selected as the cutoff point

# Model Evaluation

- With this cutoff point, the model achieved an **accuracy of 79.5%** and a **sensitivity of 83.3%,** which is in line with the CEO's requirement of over 80% conversions
- An ROC curve was also plotted. It provided the desired left-hand curve shape with sufficient area under the curve, indicating a good model
- As a final step, the lead scores were assigned by using the formula conversion_probability * 100



# Making Predictions on Test Data

- Transformations were done on the test data set to mirror those on train set. This included:
  - Splitting into X (independent variables) and y (target variable)
  - Scaling the numerical variables
  - Selecting the relevant columns used in the model
- Predicted probability values were generated using the model and predicted conversion was assigned based on the earlier selected cutoff point
- The metrics were checked and we obtained an **accuracy of 79.1%** and **sensitivity of 82%**, similar to the training set. This indicates the model is able to generalise sufficiently
- Finally, the lead scores were assigned to each prospect using the formula predicted_probability * 100

# Model Results – Business Insights

While the model was created using 10 independent variables, we can see that the variables have come from 5 notable columns:

- Total visits to the site – Prospects most likely to convert would visit the site multiple times, indicating a continued interest in the course
- The time spent on the website – longer time spent by the prospect indicates a higher chance of conversion
- Lead origin – leads originating from the leads add form and the leads import data are likeliest to convert
- Additionally, leads coming from Olark chat also show a higher tendency to convert
- Within occupations, working professionals had a higher weightage towards conversion, indicating the course could be used to further one's career path.