Case Study Summary

Brief overview of the 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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data cleaning and preprocessing techniques

the data cleaning and preprocessing section, we have used various techniques to clean and preprocess the data before feeding it into the lead scoring model. Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data. Data preprocessing is the process of preparing the data for analysis by transforming it into a format that is suitable for machine learning algorithms.

The following are some of the data cleaning and preprocessing techniques that we have used:

1. Missing value imputation: We have identified missing values in the dataset and impute them using appropriate techniques such as mean imputation, mode imputation, or regression imputation.
2. Outlier detection and removal: We have identified outliers in the dataset and remove them using appropriate techniques such as z-score method or interquartile range (IQR) method.
3. Data normalization: We have normalized the data to ensure that all features are on the same scale. This will help to improve the performance of machine learning algorithms.
4. Feature selection: We have selected relevant features that are important for lead scoring and remove irrelevant or redundant features. This will help to reduce the dimensionality of the dataset and improve the performance of machine learning algorithms.
5. Data encoding: We have encoded categorical variables using appropriate techniques such as one-hot encoding or label encoding. This will help to convert categorical variables into numerical variables that can be used by machine learning algorithms.

6. Data balancing: We have balanced the dataset by oversampling or under sampling the minority class to ensure that the model is not biased towards the majority class.

Case Study Summary

Training and testing of the model

After the data cleaning and preprocessing stage, we have split the dataset into training and testing sets. The training set will be used to train the lead scoring model, while the testing set will be used to evaluate its performance.

We have use various machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines to develop the lead scoring model. We have evaluated the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1 score.

We have also use techniques such as cross-validation and hyperparameter tuning to optimize the performance of the model. Cross-validation will help to ensure that the model is not overfitting or underfitting the data, while hyperparameter tuning will help to identify the best set of hyperparameters for the model.

Once we have developed an effective lead scoring model, we have deployed it in a production environment and use it to score leads for X Education. This will help X Education to prioritize leads and focus their efforts on those that are most likely to convert into paying customers.

Model evaluation metrics

The following are the model evaluation metrics that we have use to assess the performance of the lead scoring model:

1. Accuracy: It measures the proportion of correct predictions made by the model.
2. Precision: It measures the proportion of true positive predictions out of all positive predictions made by the model.
3. Recall: It measures the proportion of true positive predictions out of all actual positive cases in the dataset.

We have also use confusion matrix, and ROC curve to evaluate the performance of the model. The confusion matrix will help us to visualize the number of true positives, true negatives, false positives, and false negatives predicted by the model. The ROC curve and AUC score will help us to assess the trade-off between true positive rate and false positive rate for different classification thresholds.