LEAD SCORE – CASE STUDY

HIMANSHI THAKUR & LAWANG MISHRA

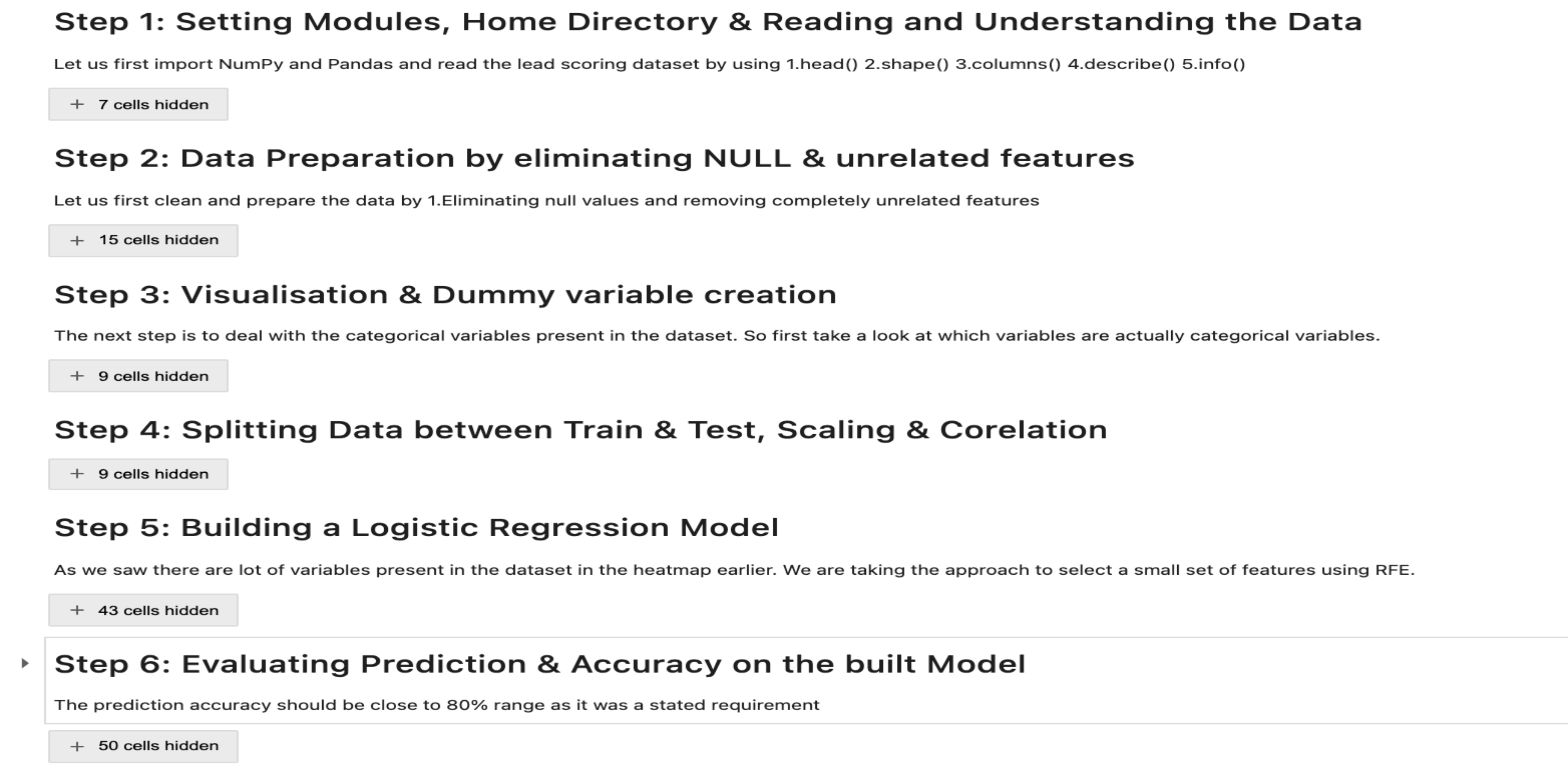
PROBLEM STATEMENT

Our client X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires us to build a model wherein we 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.

Note: The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

APPROACH



Step 1: Setting Modules, Home Directory & Reading and Understanding the Data

Let us first import NumPy and Pandas and read the lead scoring dataset by using 1.head() 2.shape() 3.columns() 4.describe() 5.info()

Step 2: Data Preparation by eliminating NULL & unrelated features

Let us first clean and prepare the data by 1.Eliminating null values and removing completely unrelated features

Step 3: Visualisation & Dummy variable creation

The next step is to deal with the categorical variables present in the dataset. So first take a look at which variables are actually categorical variables.

Step 4: Splitting Data between Train & Test, Scaling & Corelation

Step 5: Building a Logistic Regression Model

As we saw there are lot of variables present in the dataset in the heatmap earlier. We are taking the approach to select a small set of features using RFE.

Step 6: Evaluating Prediction & Accuracy on the built Model

The prediction accuracy should be close to 80% range as it was a stated requirement

KEY POINTS

- Removed the Prospect_Id and Lead_Number, City, Country columns as they were not related to the context based on the data dictionary study.
- Created Dummy Variables for the categorical variables and ensured that the Null and missing data rows were removed before model building exercise.
- During model building Kept 15 columns as the number of features to be selected and dropped column based on the Variance Inflation Factor and P value analysis.
- Calculated the Confusion Matrix and in the final model it was baselined when Precision/Recall was 78%

SUMMARY

Based on the above model we see that the factors like Total time spent on Website, Unemployment , Lead Source_Olark Chat are significant factors for converting a Lead.

When these leads expressed interest to be contacted on a preferred channel we should pursue them to convert them

Furthermore we should evaluate each lead carefully and understand their personal requirements for courses so that we can customize the details while reaching out.