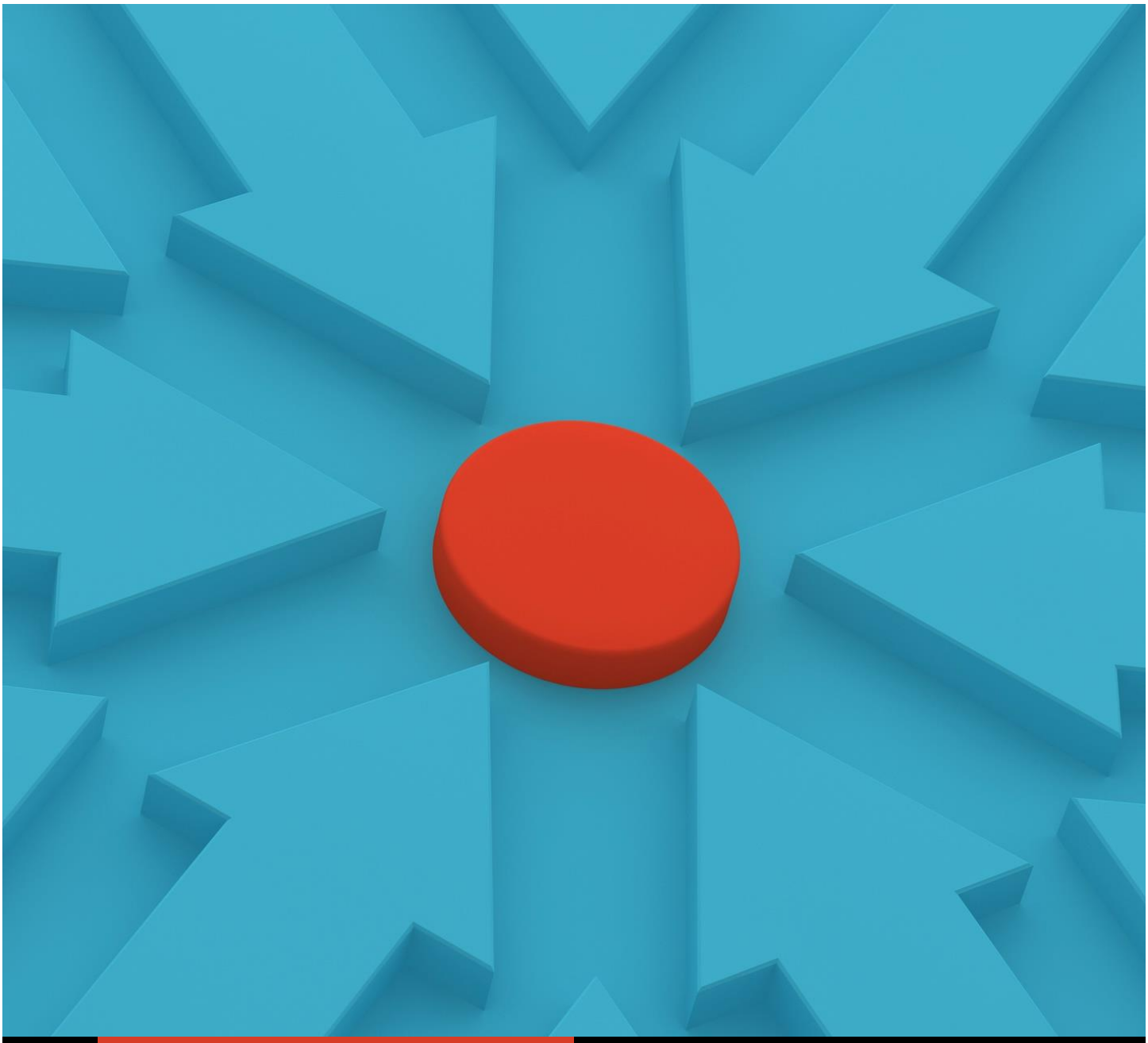
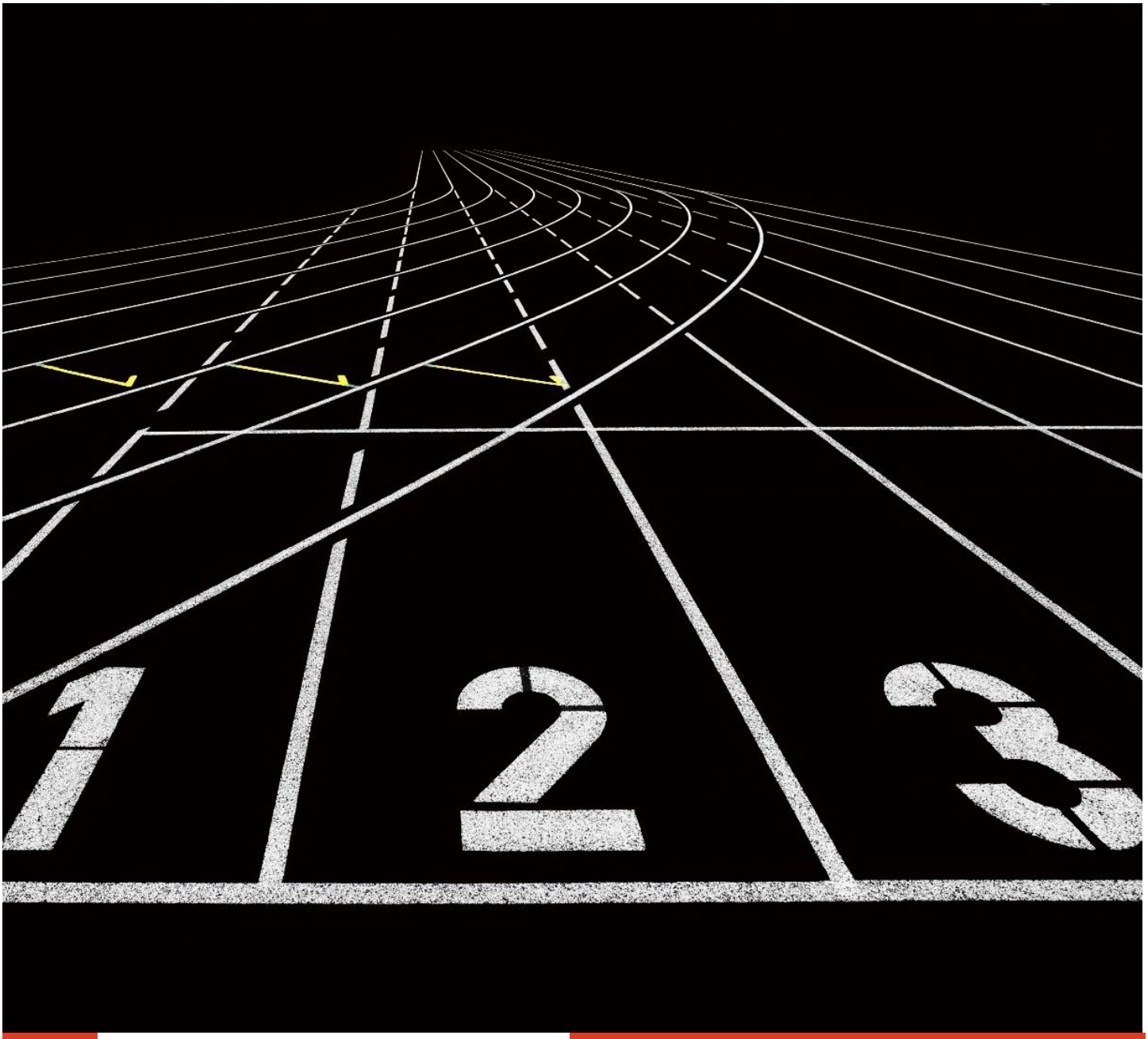


LEAD SCORING CASE STUDY

SUMMARY

Venkatasubramanian
Sundaramahadevan



PROBLEM STATEMENT

An education company ‘X Education’ sells online courses to industry professionals. 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 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 has given a ballpark of the target lead conversion rate to be around 80%.

APPROACH

The following steps were followed in sequence to arrive at the final model

Data Cleaning

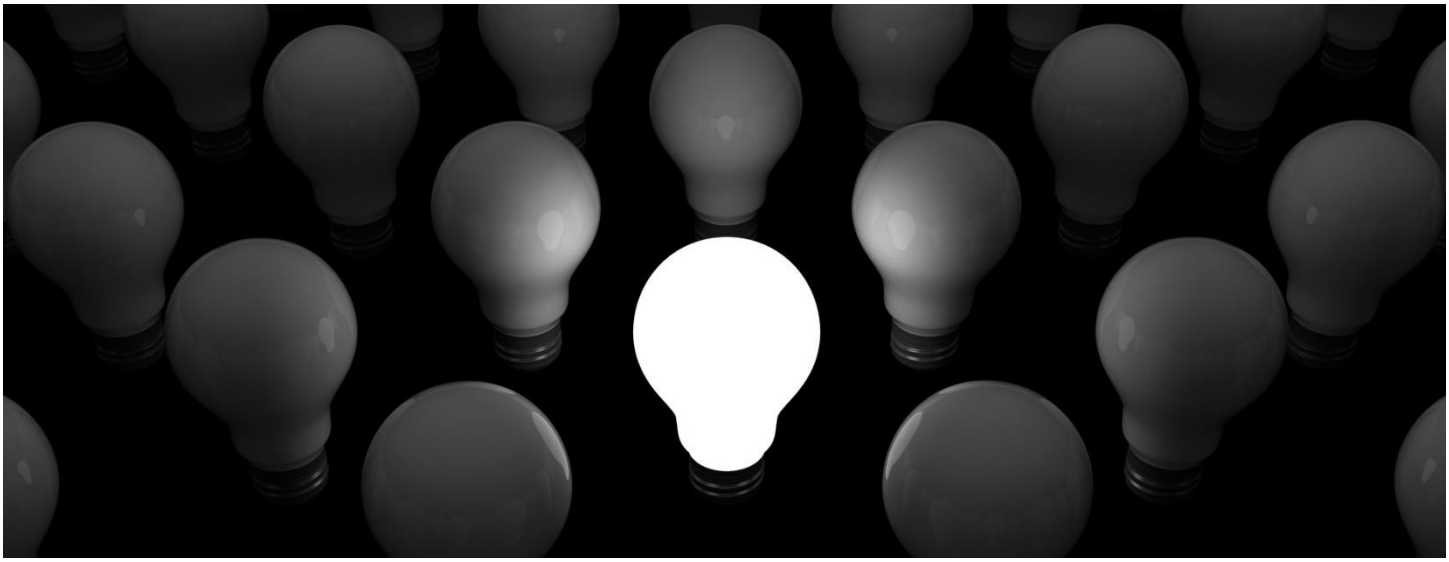
In this step, we imported the leads dataset from the past with around 9000 data points. After that, the following steps were followed as part of cleaning:

1) Drop columns with NULL values above 32%
2) Remove columns that don't provide any business value such as Country/City/Lead Number etc.,
3) Remove columns that have very less variance such as 'Do Not Call'
4) Remove columns with more than 50% 'select' values (this indicates that user didn't select any value)
5) Remove the rows for all columns with < 32% NULL values

Data Preparation

Once Data Cleaning was complete, we move to Data Preparation. In this step, following activities were performed to ready the data for modeling

1) Convert below binary categorical variables to 1 & 0 values
    a. 'Do Not Email' & 'A free copy of Mastering The Interview'
2) For multi-level categorical variables, we perform One-Hot Encoding
    a. 'Lead Origin', 'Lead Source', 'Last Activity', 'What is your current occupation', 'Last Notable Activity'
    b. Drop the first column for each dummified variable as we need only n-1 dummies to represent unique n values
    c. Drop all the categorical variables for which we performed dummification



Modeling

Once Data Preparation was complete, we move to Modeling. We performed the following operations:

- Split data into train and test with ratio 0.7 to 0.3
- Do Feature Scaling for numerical variables with MinMaxScaler
- Look at Correlations using Heatmap and make a note of correlation above 60% as they indicate multicollinearity
- Create 1st model with all features
- As we have 74 features, we will use RFE (Recursive Feature Elimination) to narrow down to 15 important features. Using this create 2nd model
- Check VIFs (Variance Inflation factor) for all features. This helps us to identify features which have highly collinear (i.e., VIF > 5)
- Check p-value for all features (p-value should be < 0.05)
- From here on, we repeatedly create models (after removing high VIFs and p-value variables one by one)
- The model was finalized at 6th iteration where we all p-values < 0.05 & VIFs < 5

Evaluation

Once Modeling was complete, we move to Model Evaluation. We did the following activities:

- Predict target probability values based on X_train (feature train data set)
- Build base predictions based on 0.5 probability cut-off point and calculate metrics such as accuracy, sensitivity, specificity, False Positive rate, Positive Predictive rate, & Negative Predictive rate
- Plot ROC (Receiver Operating Characteristic) curve to determine trade-off b/w Sensitivity and Specificity. We get AUC (Area under the curve) as 0.86 which indicates a good model
- We will then determine the optimal probability cut-off point based on various probabilities in the range 0, 0.1, 0.2…,0.9 by plotting the accuracy, sensitivity, and specificity curve. The intersection point determines the optimal cut-off. In our case it's 0.43
- We create precision and recall curve as well to validate the optimal cut-off. We observe 0.43 in this case as well
- Predict target probability values based on X_test (feature test data set)
- Calculate accuracy, sensitivity, and specificity for both train and test data based on 0.43 probability cut-off. They are as follows:

Train/Test	Accuracy	Sensitivity	Specificity
Train	0.7895	0.7854	0.7932
Test	0.7850	0.7751	0.7941

- Calculate Lead Score based on the Conversion Probabilities on the test data set by multiplying the probability values by 100

RECOMMENDATIONS

The attributes mentioned below will helps us select the most promising leads that are most likely to convert into paying customers. They are listed in their order of significance.

1) The **total number of visits** made by the customer on the website
2) The **total time spent** by the customer on the website
3) The origin identifier with which the customer was identified to be a lead was through **'Lead Add Form'**
4) Last activity performed by the customer was a **phone conversation**
5) The source of the lead is via **'Welingak Website'**
6) The source of the lead is via '**Olark Chat**'
7) Last activity performed by the customer was **'SMS Sent'**
8) The customer is a **'Student'**
9) Current Occupation of the customer is **'Unemployed'**