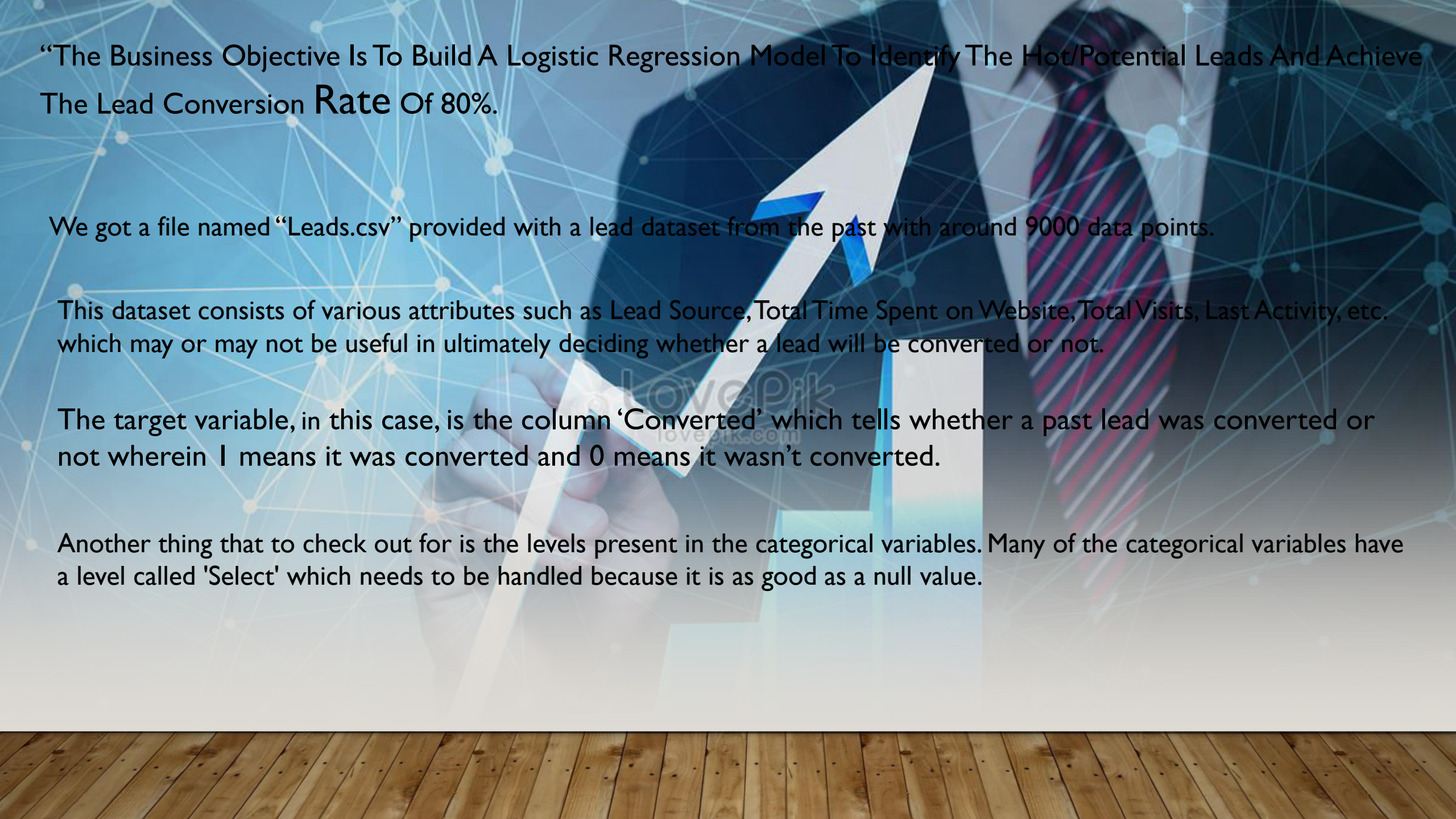
Lead Scoring

An education company named X Education sells online courses to industry professionals.

Although X Education gets a lot of leads, its lead conversion rate is very poor

For example, if say, they acquire 100 leads in a day, only about 30 of them are converted.

The objective is to build a model to identify the hot leads and achieve lead conversion rate of 80%.



"The Business Objective Is To Build A Logistic Regression Model To Identify The Hot/Potential Leads And Achieve The Lead Conversion Rate Of 80%.

We got a file named "Leads.csv" provided with a lead dataset from the past with around 9000 data points.

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Another thing that to check out for is the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

EDA: Numerical Data Total Visits

- The max probability for Total Visits is found to be around 15-20. It increases initially but decreases further.

- The average total visits for both converted and non-converted people is found to be the same. Total Time Spent On The Website

- The probability of time spent is found to be high for the time between 0-300 seconds and decreases further.

- The mean is found to be higher in the case of Converted people rather than non-converted people. Page Views Per Visit

- The max probability for Page Views Per Visit is found to be around to be 3-5. • The average page views for both converted and non-converted is found to be the same

Model – I and II: Basic Model

•We build a basic model using 35 features. Since it is not efficient we perform RFE to obtain a model with Top – 20 features. There are so many variables with high p-values and VIF values, we need to remove them.

Model – III and IV:

Removing variables with p-values > 50%

•Two columns having p –values > 50%: Lead Source_Others and Lead Source Origin.

•Since p–value> 50%, it does not seem to be significant at all.

Model V, VI, and VII: Removing variables having p-value > 10%

•Since we have a cut-off for significance value > 10 % does not improve our model.

•Hence, we remove these variables which are: Current Occupation Student, Specialization International Business, and LastActivityEmail. Model VIII: Removing variables having high VIF

•After model –VII, all p-values < 5%, hence we need to check VIF.

• VIF for Current Occupation_Unemployed = 12.20 which is > 5% . •Hence we drop this variable from our analysis.

Model – VIII: The Final Model

•All p-values < 5% - Hence they are highly significant. •All VIF values are < 5. Hence the dependency of one variable with another is tolerable. •Final model has 14 features in total.

ROC Curve And Optical Cut-Off Probability

ROC Curve represents how much the model is able to distinguish between the classes.

AUC – Area under the curve represents that it is distinguishing the 1's and 0's correctly

-On plotting the ROC curve for our data we see that, AUC is around 0.88 which means at around 88% of the time, the model is able to distinguish the 1's as 1's and 0's as 0's.

- AUC of 0.88 is found to be a very stable model.

- When we plot the sensitivity, accuracy, and specificity of the model together, the optimal cut-off point is found to be at 0.35. This means that at 35% probability, the sensitivity and specificity are found to be balanced.
- With probability = 0.35, we predict y-values with XTrain, in such a way that, any conversion prob > 35% is said to be converted to a lead

Hot Leads

- Hot leads are people who have a high probability to be converted as a Lead and thus need to be identified. They have a higher conversion rate.

- The leads whose lead score is greater than 35% are considered potential leads. The conversion rate is around 73%. When we increase this threshold from 35% to 95% we get Hot Leads.

- Conversion Rate for hot leads is increased from 73% to 96%. This means they have a 96% probability of getting converted to a lead.

- Focusing on Hot Leads will increase the chances of obtaining more value for the business as the number of people we contact are less but the conversion rate is high.

- From our model, we can conclude the following points:

- The customers/leads who fill out the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is an SMS sent or Email opened.
- It's always good to focus on customers, who have spent significant time on our website.
- It's better to focus least on customers to whom the sent mail is bounced back.
- If the lead source is a referral, he/she may not be the potential lead.
- If the lead didn't fill specialization, he/she may not know what to study and are not the right people to target. So, it's better to focus less on such cases.