Summary

- This case study consist of an education company named X Education who sells online courses to industry professionals. Interested people fills the form, to watch some course related videos. They get turned in to a lead when they provides their email, address or phone number. Some leads are also from referrals too.
- Problem faced by X Education company is that these leads may or may not get converted. The rate of converting lead till now is only 30%. Hence X Education company don't want to waste time in focusing Hot Leads that have potential to get convert.
- The X Education company require our help to select the most promising leads. And we should achieve this with 80% of correct accuracy of lead conversion rate.
- We need to clean the data before building the model. While cleaning we came across various columns that are having null values or columns that are irrelevant for the analysis hence we removed them.
- Categorical columns with more than two unique values were made concise by assigning one single category to such values.
- We dropped the categorical columns that are having only one value throughout.
- We dropped columns that are having entries skewed to one value as data will be biased.
- There are certain columns that are representing **views and pages visited, the datatype of which should not be decimal number,** so such values are rounded and stored as integers.
- Removing the columns that are irrelevant for analysis.
- Checking for outliers, as they may affect the data and imputing with appropriate values.
- Here we have checked for data imbalance in the columns as if it is the case then the results will be biased.
- We have used graphs to visualize relation of features with the target and to figure out important ones.
- Created a Dummy variables for model building for all the remaining categorical columns.
- We have used logistic regression model as this is classification-based problem.
- We scaled the values of continuous columns used for training the model.
- Heatmap correlation identifies any underlying multi-collinearity to drop such columns.
- RFE is used to select best features for building model.
- Building Logistic Regression model to check multi-collinearity with the p-values and VIF values that can dropp the columns if necessary.
- Repeating model building and dropping columns based on p-value and VIF until p-value < 0.05 and VIF < 5.
- Adding the additional columns to increase performance of model.
- Deploying model on the test dataset to evaluate it, where conversion ratio is 80% as mentioned.
- Optimize communication channels based on lead engagement impact.
- Need to spend more budget on Welingak Website as it is the top performer, etc.
- Incentives/discounts need to be provided that encourage providing more references.
- Needs to develop a strategies to attract high-quality leads from top-performing lead sources.
- Needs to target working professionals as they have high conversion rate and also they are ready to pay as they have better financial situation to pay higher fees.
- Review landing page submission process for areas of improvement.