Summary:

- We have an education company X Education who sells online courses to industry professionals. People fill forms, to watch some course related videos and are turned to lead when they provide email, address or phone number. Also leads are from referrals too.
- Problem is that these leads may or may not convert. Also, typical conversion of lead form company till now is only 30%. Company does not want to waste time and focus on only the Hot Leads which have potential to convert.
- They require our help to help them with selecting the most promising leads. And requirement of company is that our performance should be around 80% lead conversion rate.
- Data is cleaned before building the model. Here, we came across columns with high number of null values hence we removed them.
- Categorical columns with more than two unique values were made concise by assigning one single category to such values.
- Dropping categorical columns with only one value throughout.
- Dropped columns which have entries skewed to one value as data will be biased.
- There were columns representing **views and pages visited, which should not be decimal number,** so such values are rounded and stored as integers.
- Identifying such columns that have no use in analysis.
- Checking for outliers, as they may affect the data and imputing with appropriate values or other measures.
- 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.
- Dummy variables were created for model building for all the remaining categorical columns.
- We have used logistic regression model as this is classification-based problem
- Scaling values of continuous columns for training the model for training purpose.
- Using correlation heatmap to identify any underlying multi-collinearity and dropping such columns.
- RFE is used to select best features suitable for building model with the help of RFE support.
- Building logistic Regression model and checking multi-collinearity with the p-values and VIF calculated and dropping if necessary.
- Repeating model building and dropping columns based on p-value and VIF until p-value < 0.05 and VIF < 5 for all columns.
- Adding additional columns which were thought to increase performance of model
- Deploying model on the test dataframe and evaluating it where conversion ratio is 80% as per requirement.
- It is observed not all columns might be necessary for model building
- Addition of new columns can improve performance
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- spend more budget on Welingak Website as it is the top performer, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- target working professionals as they have high conversion rate as well as they are ready to pay as they have better financial situation to pay higher fees too.
- Review landing page submission process for areas of improvement