Lead Scoring Case Study: Summary Report

Process Followed	1. Initial brainstorming & approach formulation: ➢ Involved walkthrough of the original dataset, data dictionary and gaining some domain expertise followed by draft formulation of the approach to be followed. 2. upGrad session for validating the approach: ➢ The session helped in better understanding of the expectations from the case study and the business aspects. 3. Alignment on coding protocols and Data preparation ➢ Team documented the final steps & prepared a common coding skeleton. ➢ Data preparation involved reshaping the data based on business requirements. Some steps involved in that were: Dropping Columns/Rows, Handling Null Values, Outlier Treatment and EDA to gather insights from the data. ➢ Precise coding along with relevant comments helped in the clarity of the steps. 4. Modeling and validations: ➢ First followed RFE approach to pick around 25 variables for the first model followed by dropping 1 variable at a time, based on high p-value/VIF – using statsmodels. ➢ The final model had 14 variables each having p-vale<0.05 and VIF<2.5(indicates almost negligible multicollinearity). ➢ Various metrics to measure the efficiency of model was calculated and then the same were evaluated on test data. ➢ Team reviewed each other's coding work and fine tuned 5. Final consolidation: ➢ Train: Accuracy:0.797, Sensitivity:0.847, Specificity:0.766, Precision:0.696 ➢ Test: Accuracy:0.801, Sensitivity: 0.854, Specificity: 0.769, Precision:0.694 ➢ Lead Score calculated for the complete data and visualizations done to check the relation between Converted and Lead Score (High Lead Score mostly corresponds to Converted value 1)
Learnings	1. Handling categorial variables with numerous categories ➢ Since we do not want our dataset to have huge number of dummy variables, it better to group categories having less weightage into 1 single category as "Others". 2. Handling skewed categorical variables

<table>
<tr>
<td></td>
<td>
Highly skewed categorical columns indicate its less impact to the overall outcome. Hence such columns should be dropped
<ol start="3">
NULL imputation in categorical variables
Null imputation by mode would not always be the best solution for categorical columns.
Process of finding the right cut-off
Checking the Predicted value against multiple cut-off values is a manual approach. Can be done by plotting accuracy, sensitivity and specificity for various probabilities. The intersection of the graphs gives the optimum cut-off. Anything less than that would increase False Positives leading to increase in Sensitivity but decrease in Specificity and anything above that would increase Specificity but decrease Sensitivity.
Adjusting the predictions based on business needed
The scenario-based questions helped us think differently on how to make the business grow in a better way.

</td>
</tr>
</table>