Lead Scoring Case Study Summary

The company requires a model to be built for selecting the most promising leads. Lead score to be given to each leads such that it indicates how promising the lead could be. Higher the lead score, more promising the lead to get converted.

Approach followed for model building:

- Splitting into train and test sets
- Scale variables in train set
- Build first model
- Use RFE to eliminate less relevant variables
- Build next model
- Eliminate variables by checking the high p-values
- Check VIF values for all variables
- Evaluate accuracy and other parameters
- Precision and Recall Analysis on test predictions

After trying out several models, our final model has following characteristics:

- All p-values are very close to zero.
- VIFs for all features are very low. There is hardly any multicollinearity present.
- The overall testing accuracy of 80.77% at a probability threshold of 0.05 is also very good.

The optimal threshold for the model is 0.35 which is calculated based on tradeoff between sensitivity, specificity and accuracy. According to business needs, this threshold can be changed to increase or decrease a specific metric.

High sensitivity ensures that most of the leads who are likely to convert are correctly predicted, while high specificity ensures that most of the leads who are not likely to convert are correctly predicted.

Twelve features were selected as the most significant in predicting the conversion:-

Features having positive impact on conversion probability in decreasing order of impact:

- Total Time Spent on Website
- Lead Source_Olark Chat
- Lead Source_Reference

- Lead Source_Welingak Website
- Last Activity_Email Opened
- Last Activity_Others
- Last Activity_SMS Sent
- Current_occupation_Working Professional

Features having negative impact on conversion probability in decreasing order of impact:

- Last Activity_Olark Chat Conversation
- Specialization_Hospitality Management
- Specialization_Others
- Lead Origin_Landing Page Submission