Lead Scoring Case Study – Summary Report

By – Arindam Bhattacharya & Satyendra Subedi

The objective of the case study is to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The data was first imported. The shape and size of the dataset were inspected along with the data types of the columns. During the data cleaning process, the dataset was checked for missing values. All columns with more than 40% missing values were removed from the dataset. Many variables were found to have a high degree of data imbalance. Some of those were imputed with the most commonly occurring value i.e., the mode. Some columns had only one distinct value, which were removed.

After the cleaning process, the Exploratory Data Analysis (EDA) was performed in which univariate and bivariate analyses of all variables were carried out. It was found that there was a higher conversion rate among leads from "Welingak Websites" and "Reference", who were **working professionals** and those who **spent more time on the websites.** In absolute terms, unemployed leads converted more but the conversion rate was only 30-35%. The top 3 specializations that were selected were: **Finance Management, Human Resource Management and Marketing Management**. Most leads also came from India. In The numerical columns, the data points were capped to the 95th percentile as there were outliers.

While building the model, the **dummy variables** were introduced for all categorical columns, which resulted in a total of 78 variables in the dataset. Next, the dataset was split into a **70:30 ratio** for training and testing. The numerical variables were scaled using a **Standard Scaler**. To select the variables for model building, **Recursive Feature Elimination (RFE)** was done with a parameter to select the best **15 variables** and the p-values were checked using the **Statsmodels library**. Subsequently, more models were made by eliminating variables that had high **p-values and Variance Inflation Factors (VIF)** and by the accuracy with a cut-off of 0.5 which was arbitrarily chosen to test the models. The 5th and final model had all the variables with a p-value of 0 and VIF under 5, which also means that **multicollinearity was minimized**.

The model was evaluated by measuring the **sensitivity, specificity and accuracy**. An **optimal cut-off of 0.34** was selected and the final results were obtained. The difference in the accuracy, sensitivity and specificity was marginal, which shows that the **model has not overfitted**. A sensitivity of 80% was the target of this case study and this was done successfully on the final model.

In conclusion, leads who are working professionals, leads that spend more time on the websites, and those who come from Wellingak Websites and References had a better conversion rate. Lead Origin_Lead Add Form was the strongest feature with the highest coefficient in the model followed by the current occupation of the lead.