Lead Scoring Case Study - Summary

Sumedh Kurhade
Sai Lokesh
Nishant Agrahari
Mohammed Shebin

The analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. **Data Reading and Understanding:**

    - Number of rows and columns
    - Data types of each columns
    - Checking first few rows how data looks
    - Checking how the data is spread
    - Checking for duplicates, if any

2. **Data Cleaning:**
    - Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
    - Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
    - Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

3. **EDA:**

    - Data imbalance checked- only 38.5% leads converted.
    - Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
    - Time spend on website shows positive impact on lead conversion.

4. **Train-Test split:** The split was done at 70% and 30% for train and test data respectively.

5. **Data Preparation:**
    - Created dummy features (one-hot encoded) for categorical variables
    - Feature Scaling using Standardization
    - Dropped few columns, they were highly correlated with each other

6. Model Building:

- Used RFE to reduce variables from 48 to 15. This will make dataframe more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

7. Model Evaluation:

- Confusion matrix was made and cut off point of 0.45 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 73%.
- As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions

8. Precision – Recall:

- This method was also used to recheck and a cut off of 0.45 was found with Precision around 80% and recall around 73% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   - Google
   - Direct traffic
   - Organic search
   - Welingak website
4. When the last activity was:
   - SMS
   - Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.