Oct 18th, 2022

Lead Scoring Case Study

Summary Report

The first steps towards approaching this case study were data loading, cleaning and EDA. As these are essential steps without which a good ML model can't be built, we went to great lengths to ensure that we miss nothing. Then, we proceeded with data preparation, including dummy variable creation, train-test split and feature scaling. Finally, we trained five logistics regression models, evaluated them and chose the best-performing model.

Data Cleaning

There were a large number of missing values which needed to be handled. We first replaced the value "Select" with NaN. Then, we dropped all features with more than 40% missing values and all rows having features with little missing values. We proceeded to impute most of the remaining missing values using the mode. We had to choose different strategies in some cases. For example, the mode of the "City" feature was "Mumbai", but it couldn't be used with those records that had customers who didn't belong to India.

Besides this, we removed unnecessary features that couldn't contribute anything to the ML model, like features having only one category or huge data imbalances.

Exploratory Data Analysis (EDA)

We checked the data imbalance of the target variable. We also did unsegmented and segmented univariate and bivariate analyses to draw some interesting insights. For some features, we also binned values with very low frequencies into a separate category to avoid the curse of dimensionality. In one case, we also handled outliers to visualise box plots well.

Data Preparation

As data preparation, we first created dummy variables for all categorical features. Then, we proceeded to split the data into train and test sets. We chose 70% data in the train set. Then, we scaled numerical features using a min-max scaler.

Model Training

We finalised five models by carefully removing all insignificant and highly multi-collinear features after each training. We used p-Value and VIF heuristics to select which features were insignificant and highly multi-collinear. The models are different due to initial conditions.

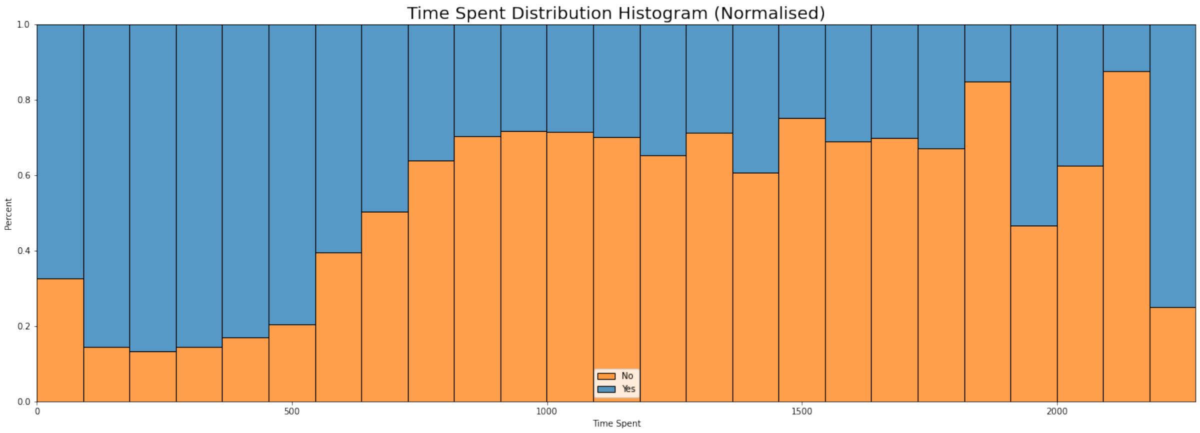
We also found the optimal probability cutoff using sensitivity-specificity trade-off analysis and calculated the optimal accuracy score for each model.

Model Evaluation

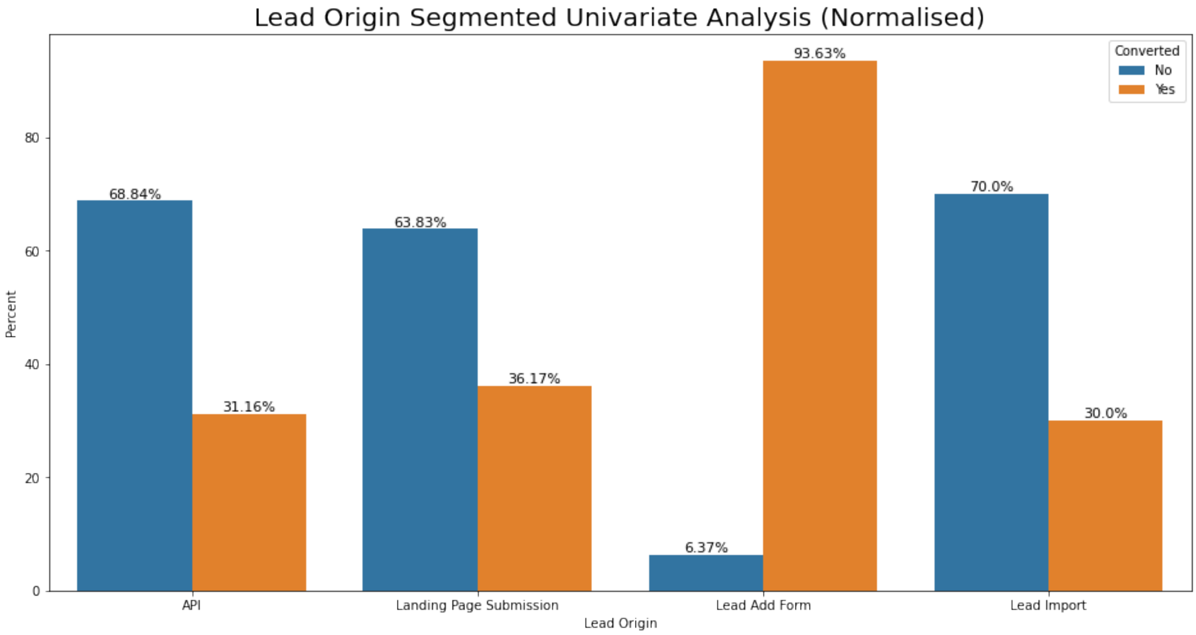
We used the models on test data and then calculated the accuracy score of the predictions. As a final model, we chose a model that used only nine predictors but gave more than an 80% accuracy score on both train and test data.

Findings

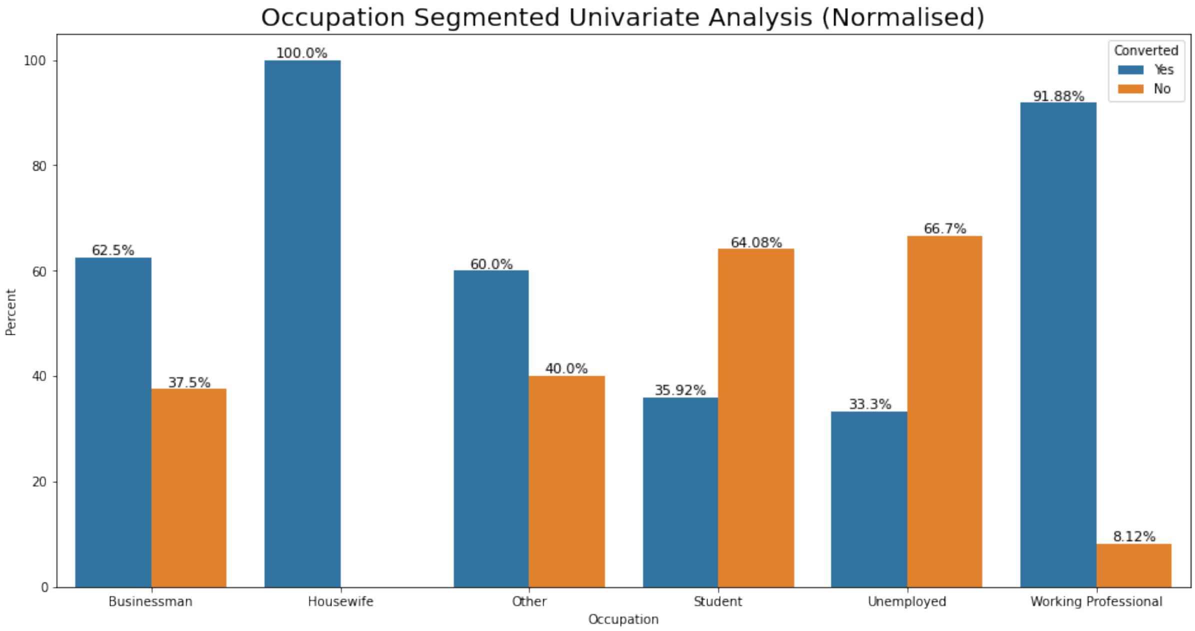
The top three predictors having the highest coefficients in the final model are “Total Time Spent on Website”, “Lead Origin_Lead Add Form”, and “Occupation_Working Professional”. This makes sense if we look at the EDA’s univariate segmented analysis.



The above segmented normalised visualisation clearly shows how higher time spent on the website leads to a higher conversion rate.



The above segmented normalised visualisation clearly shows that the “Lead Add Form” lead origin dictates a higher conversion rate.



The above segmented normalised visualisation clearly shows that the “Working Professional” occupation dictates more conversion rate.