Lead Scoring Case Study

BY:
ABHISHEK JAIN

NEETHU MARIA TOM

Problem Statement

The X Education currently has a lead conversion rate of around 30% and now they want to improve the lead conversion rate by giving more attention to potential leads through sales team. A logistic regression model need to be built, that can assign a score between 0 to 100 for each leads. Higher the score, higher the chances of conversion.

- The business want to identify top features that increase the probability of lead conversion.
- Business would also like to know how changes will be made to the model pertaining to specific scenarios of cost savings.

Analysis Approach

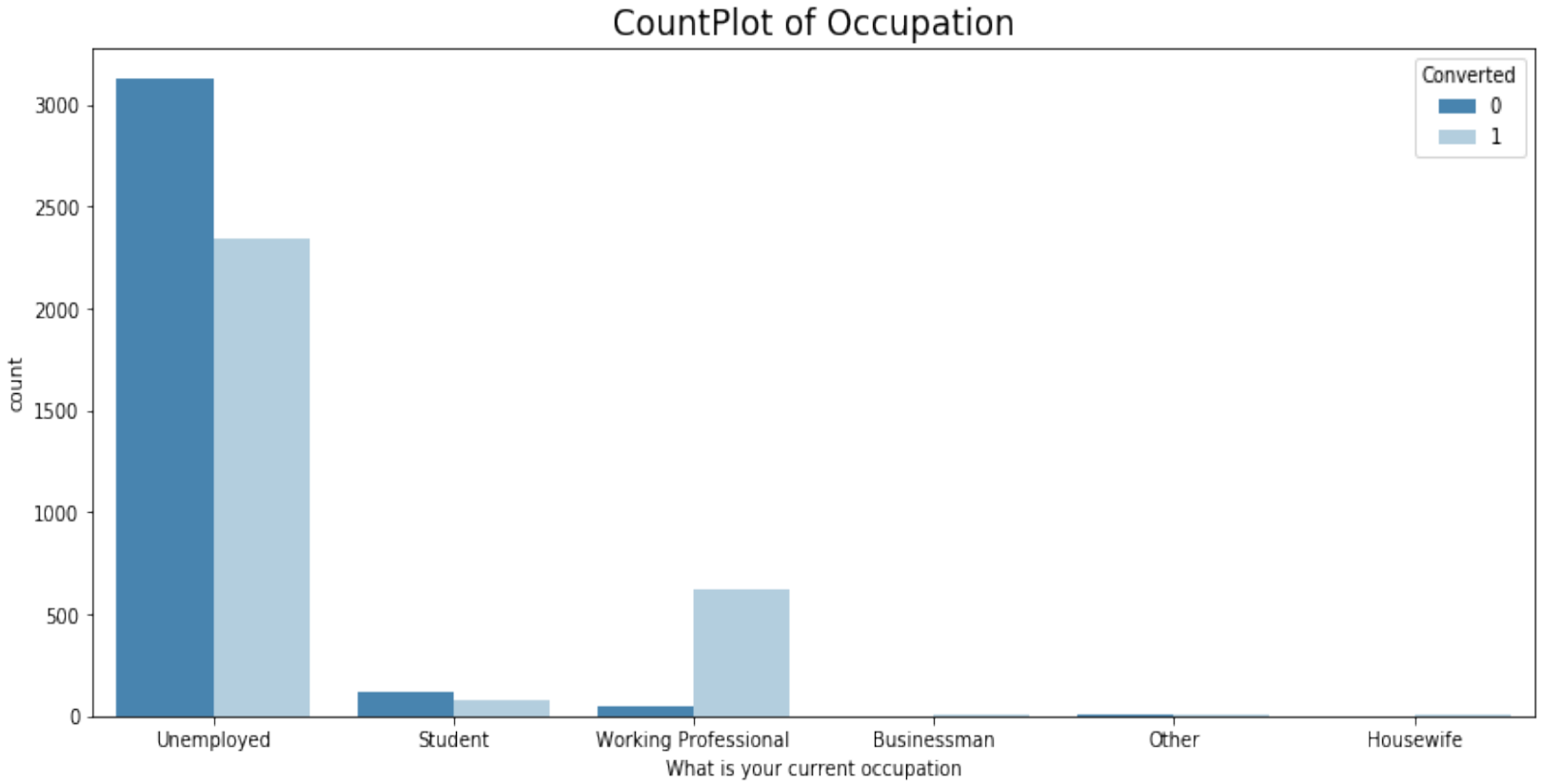
Analysis approach focuses on identifying key factors/variables that could possibly influence lead conversion rate.

Steps followed for analysis:

- Null values were identified and handled. Outliers were identified and removed.
- Derived insights after conducting univariate, bivariate and multivariate analysis.
- Checked for class imbalance
- Modelling was done in iterations and important features were identified.

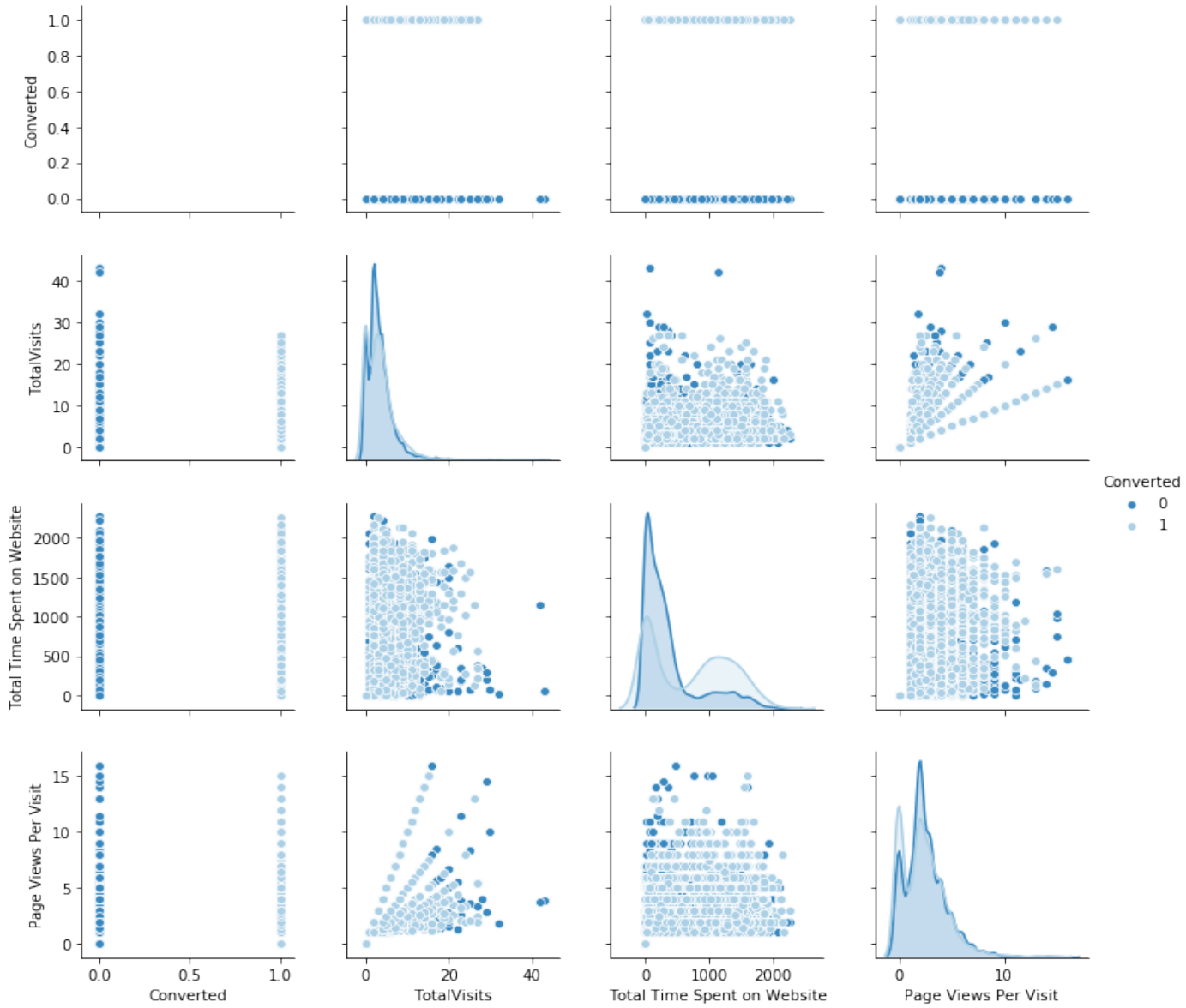
Association of variable Occupation with Converted

- Most of the leads are from Unemployed category and it has high conversion compared to other categories.
- Notably working professionals has the highest conversion rate than any other categories.

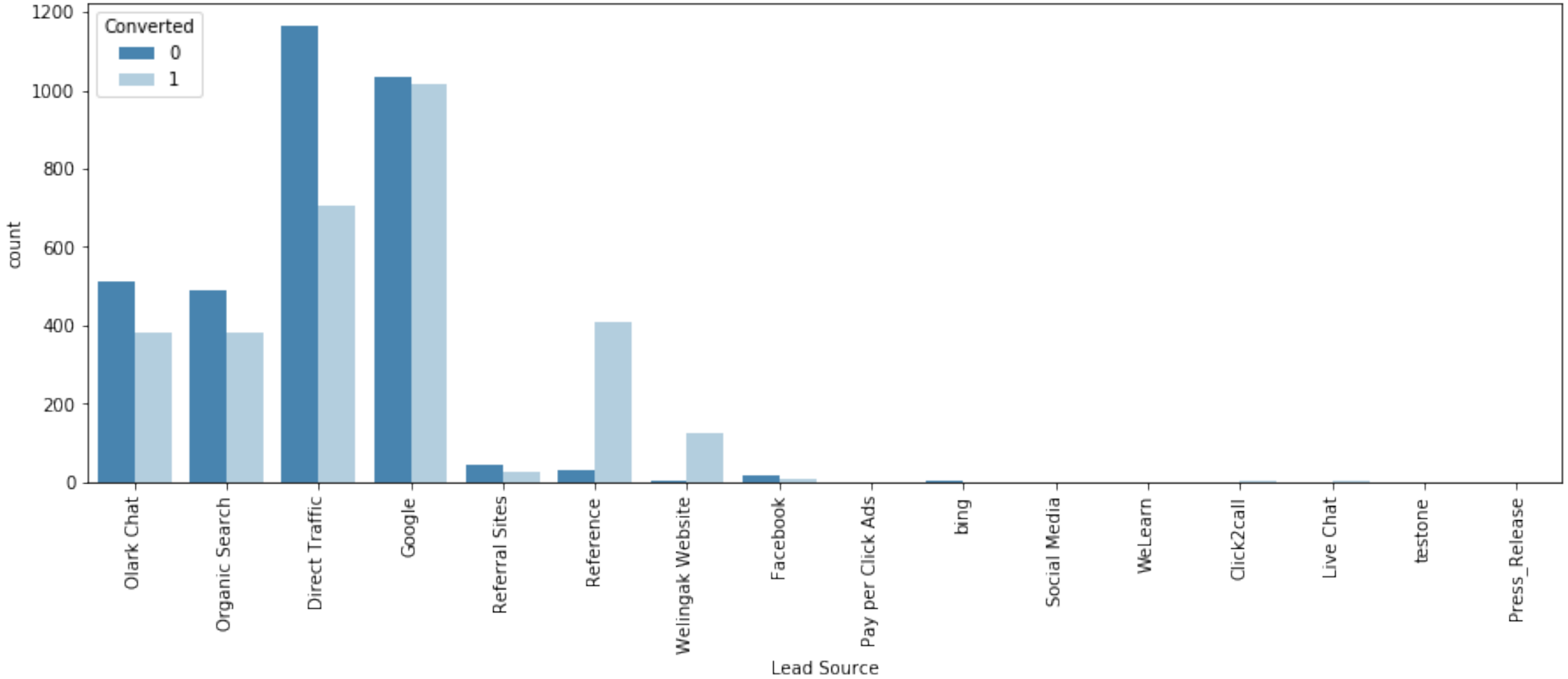


Relation between numeric variables and Converted

Clearly Total Time Spent has a high relation to conversion. Time spent by converted leads are almost half of the time spent by non-converted leads. Whereas other numeric variables doesn't provide significant separation between classes.



Association of feature Lead Source with Converted.



1. Conversion Rate of reference leads and leads through welingak website is high.
2. Maximum number of leads are generated by Google and Direct traffic.

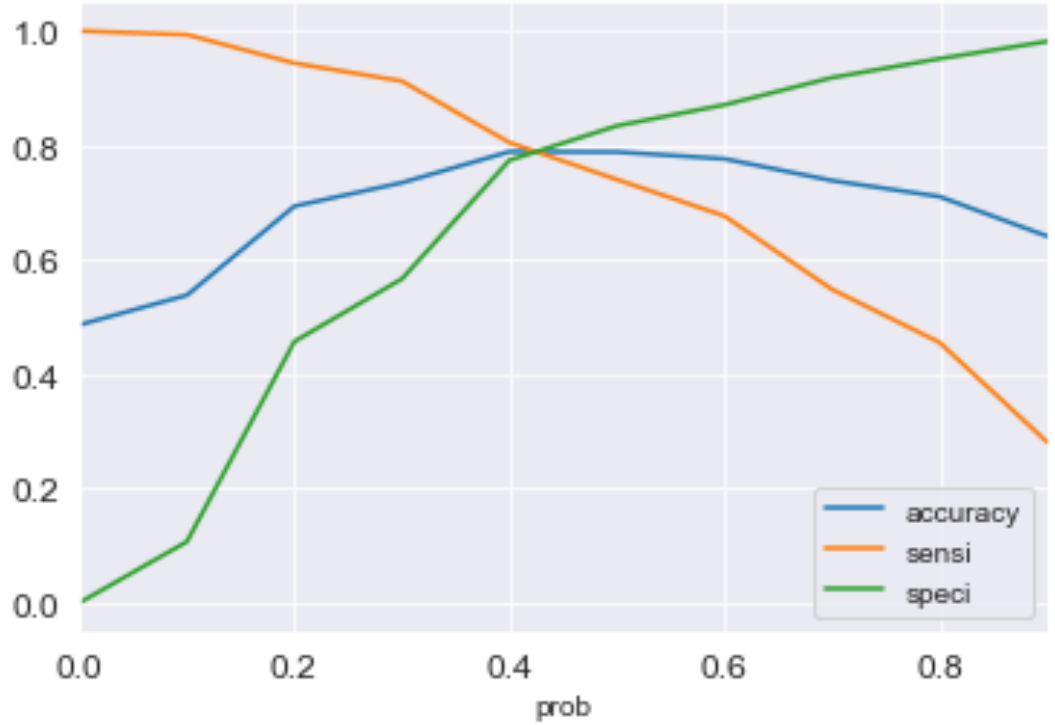
Results from model evaluation

- There was no need to deal with class imbalance as almost 48% of data had converted value as 1.

- Metrics from test data evaluation:

  - Accuracy: 77.6%
  - Sensitivity/Recall: 77.8%
  - Specificity: 77.4%
  - Precision: 75.2%

- Cut off used - 0.42



Results for business queries

Q. Top three variables in your model which contribute most towards the probability of a lead getting converted.

1. Lead Source
2. Total Time Spent on Website
3. TotalVisits

Q. top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion.

1. Lead Source_Welingak Website
2. Lead Source_Reference
3. Last Notable Activity_Unreachable

Business Queries

- To make lead conversion more aggressive

Here all of the potential leads need to be identified as they have enough employees to make calls to leads. In this case the company is not concerned about the false positive rate and our aim is to increase Sensitivity/Recall. So, our objective here will be to reduce the number of false negative observations. That means, we will have to decrease the cut off for the predicted probability. This cut off has to be confirmed with the help of business, considering achievable number of daily calls per employee.

- When company want to minimize the rate of useless phone calls.

Here the aim is to reduce useless phone calls, which means they are not concerned about false negative values but are interested in high Precision of the model. By increasing the cut off of the model, the number of observations falsely identified as positives can be reduced. Thus, only those leads having high probability will be predicted as 1. Again this cut off will have to be confirmed with the business based on the number of phone calls they are planning to make per day.

Summary

- A logistic regression model that can predict lead score with around 77% of accuracy and sensitivity on test data was built to achieve the business goal to increase lead conversion rate. Here business does not want to miss any potential customers.
- As per the model, important features that contribute most towards the probability of a lead getting converted are Lead Source, Total Time Spent on Website and TotalVisits.
- Based on the business requirement the cut off for the model that differentiate between Converted or not-Converted can be changed which ultimately affect model metrices like Recall or Precision.