UpGrad Data Science Course

Lead Scoring Case Study for X Education

- Agenda:

  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

- Agenda:
  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

- Problem Statement:

  - X Education sells online courses to industry professionals.
  - X Education is able to generate leads through various sources like online platform, past referrals etc.
  - However, the typical lead conversion rate at X education is around 30% which is very poor.
  - X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- Agenda:

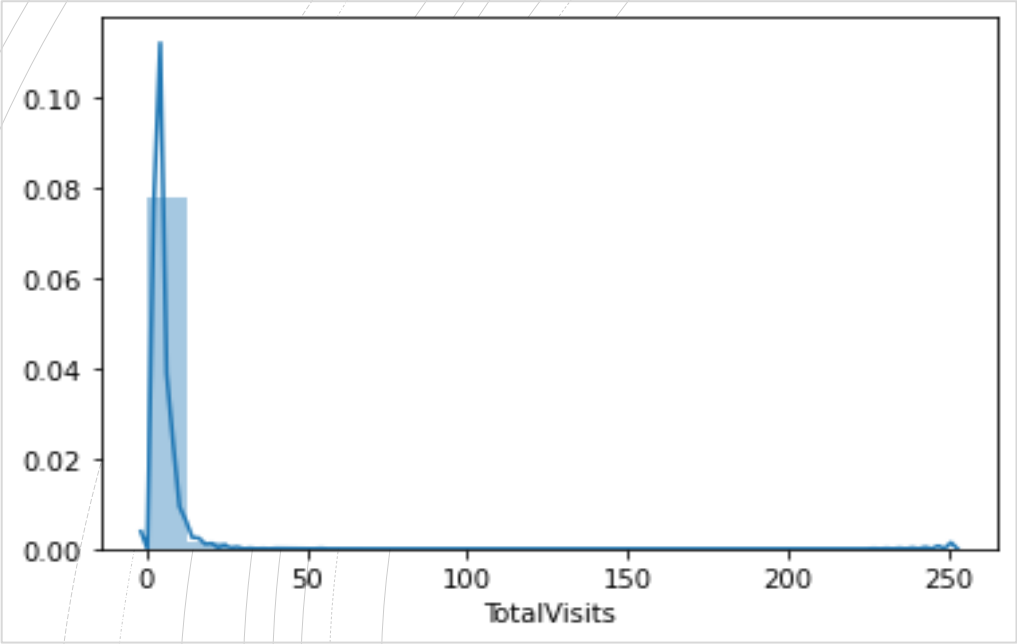
  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

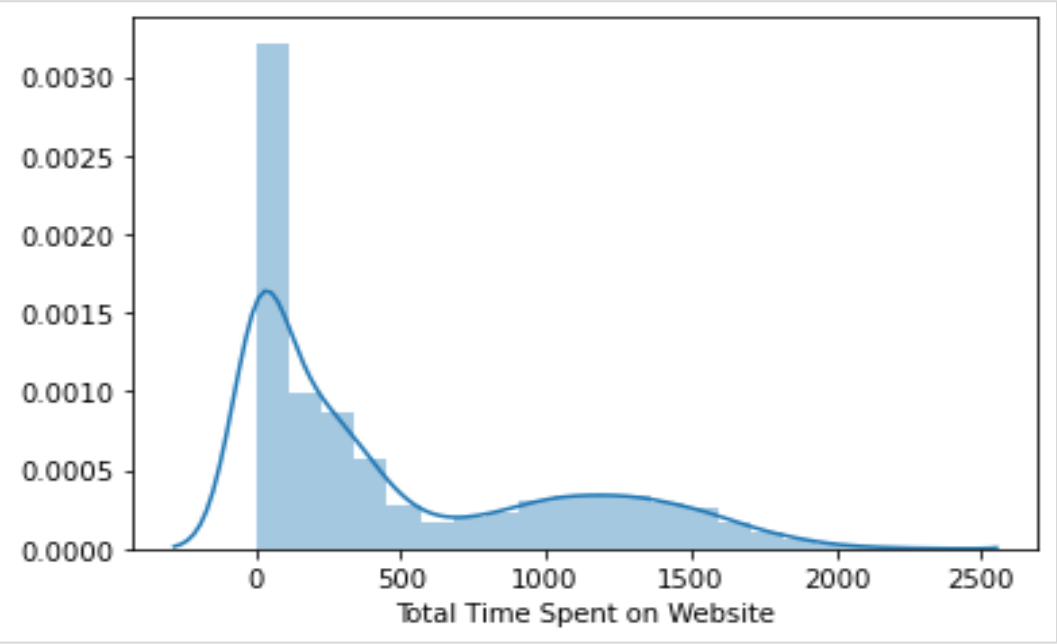
Analysis Approach:

1. Data Importing
2. Dropped the columns
   - With no useful information
   - High number of missing values
3. Missing value imputation
4. In the EDA step we did :
   - Univariate Analysis
   - Bivariate Analysis
   - Multi Variate analysis
5. Dummy Variables Creation (Object Data Type Variables)
6. Train test Split (70/30) & Scaling continuous numeric features(Using MinMaxScaler)
7. Model Building:
   - Using Logistic Regression
   - Combined RFE & manual Selection
   - Chose features which were most significant & had low VIF
8. ROC Curve & Precision Recall tradeoff for Cutoff Value estimate
9. Predicting the results and metrics on the Train - Test data.
10. Adding the Score Column in main data frame

- Agenda:
  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

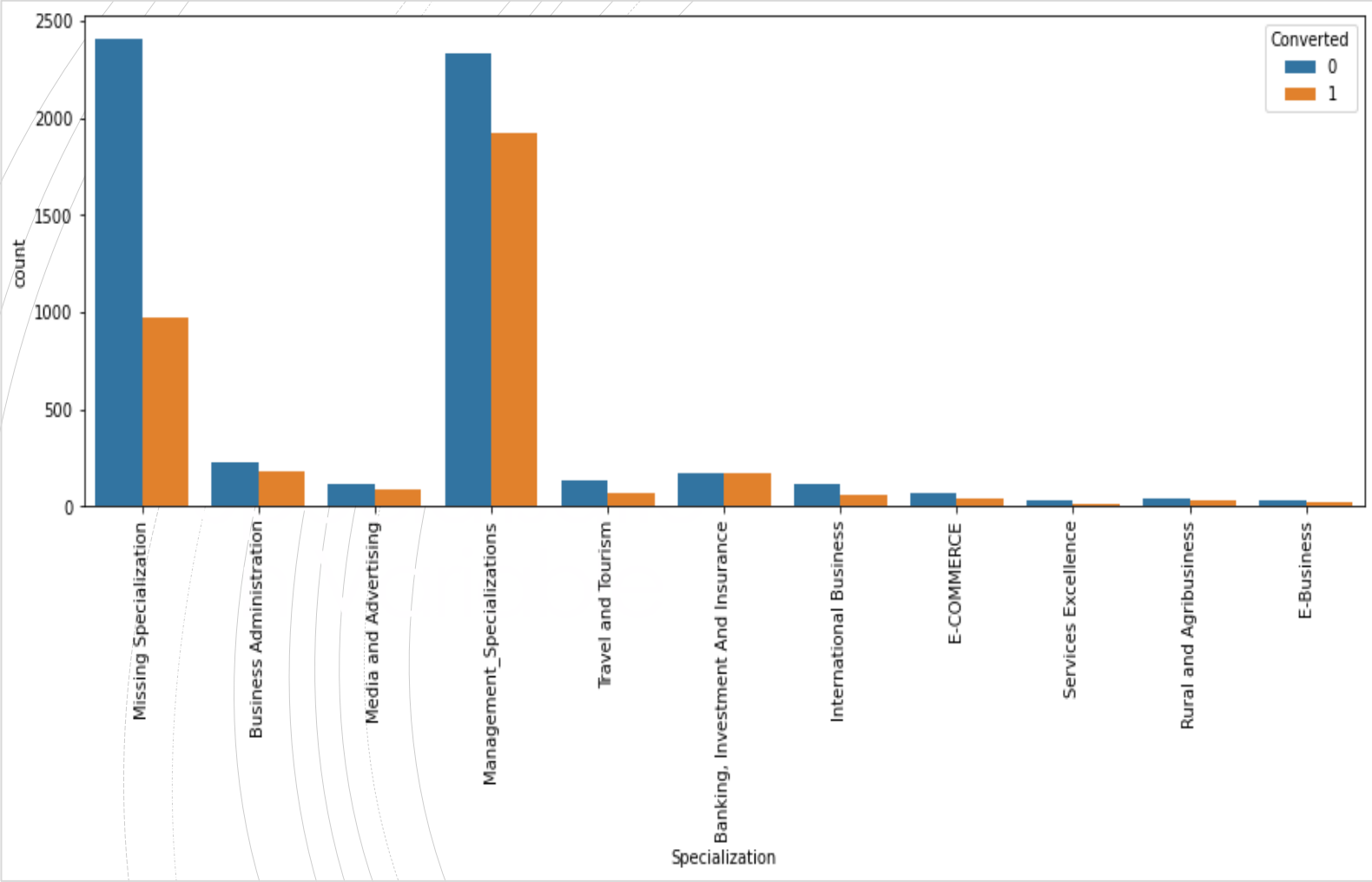
Univariate Analysis:





- Highlights:
  - Most of the people paid not more than 25 visits
  - The time spent on the website was also showing dropping trend
  - Univariate analysis was not able to add much of insight to the study
  - However we did not remove any outliers as they were adding important information to the model.

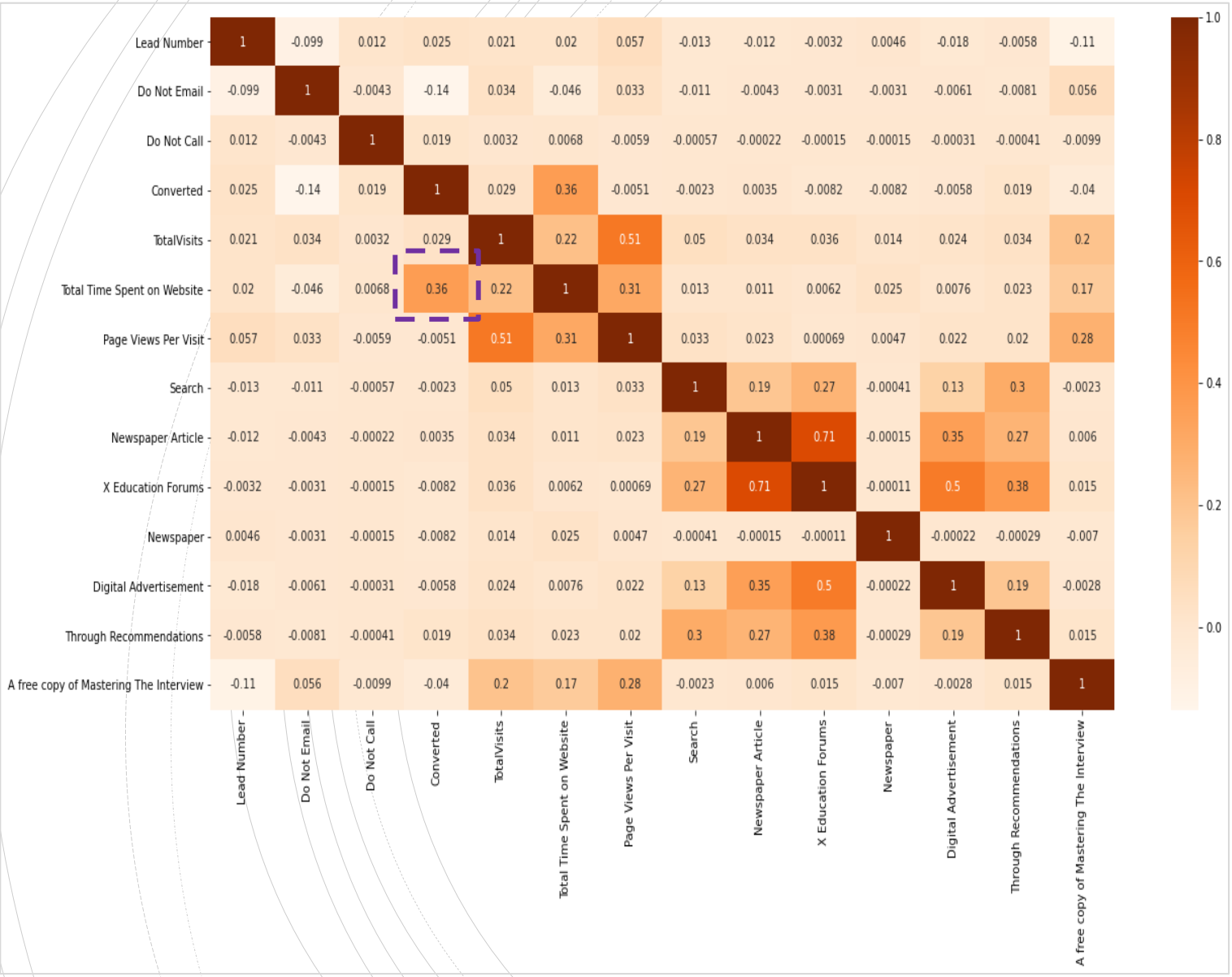
Bi-Variate Analysis:



Highlights:

- Maximum number of leads did not disclose their Specialisation
- The Management specialisation is which the maximum leads have shown interest towards the courses.

Multivariate Analysis:



Highlights:

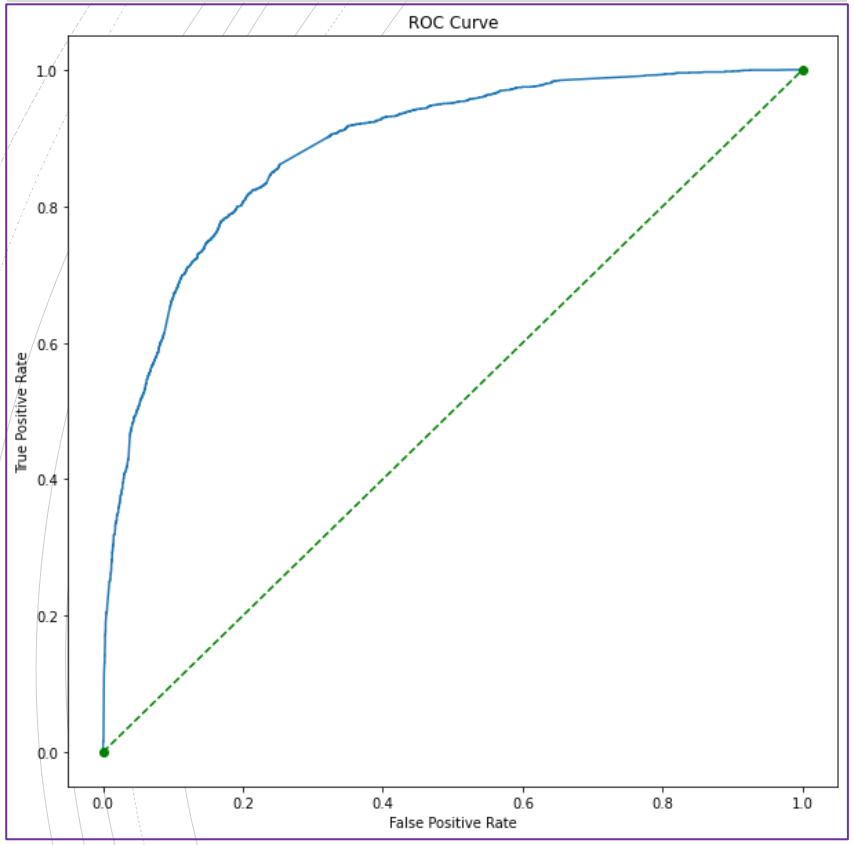
- There is a significant positive relationship of Total time spent on website and the conversion rate
- Hence it is clear & logical indicator that the interested candidate will share more amount of time on the website.

- Agenda:

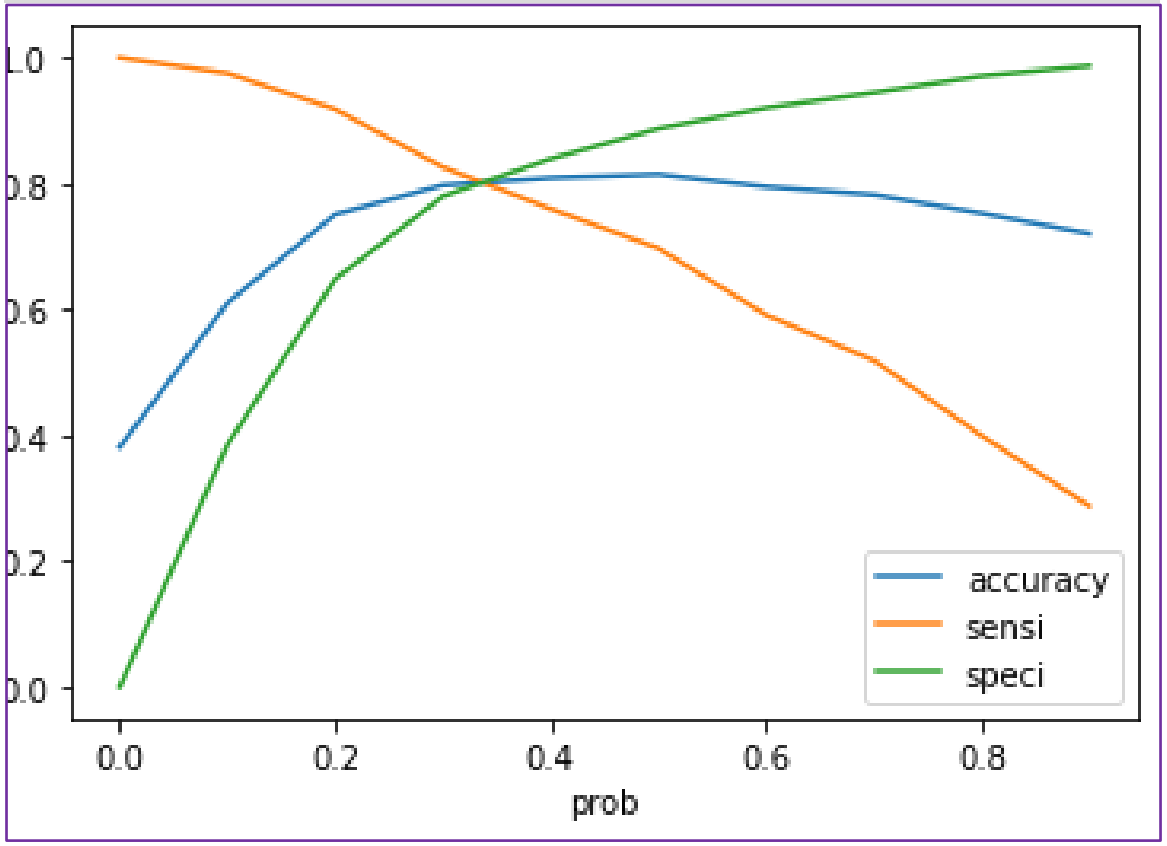
  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

Conclusion

ROC Curve



Accuracy , Sensitivity , Specificity v/s probaility



Highlights:

- Area under the ROC Curve came to be ~ 0. 89 which is a fair score.
- Accuracy , sensitivity and specificity seem to cross each other 30% approx.
- However we chose 25% as the cut off as we wanted to have high recall/sensitivity.

Conclusion

- VIF is maintained below 2 which helps to reduce the multicollinearity for the model.
- All the features that are preserved in model are significant i.e. they have p-value below 0.05
- The evaluation metrics are proving that the model is stable and predicting results decently.
  - Accuracy – 0.79
  - Recall – 0.86
  - Precision – 0.68
  - F1 Score – 0.76

- Agenda:

  - Problem Statement
  - Analysis Approach
  - Exploratory Data Analysis
  - Conclusion
  - Recommendations

Recommendation

- After looking at the Final Model we understand that the Company can focus on following variables to convert a Lead into customers:
  - Should focus on:
    - Number of visits to the Platform
    - Duration spent on Platform
    - Regular phone conversation with the Lead
  - Should not focus on:
    - People who do not give Email Id’s
    - People who do not intend to give their occupational Details.

Thank you