LEAD SCORING CASE STUDY ASSIGNMENT

BY PALLAVI NALAVDE, VELMURUGAN R AND HEENA KANZAR.

PROBLEM STATEMENT

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision

DATA EXPLORATION

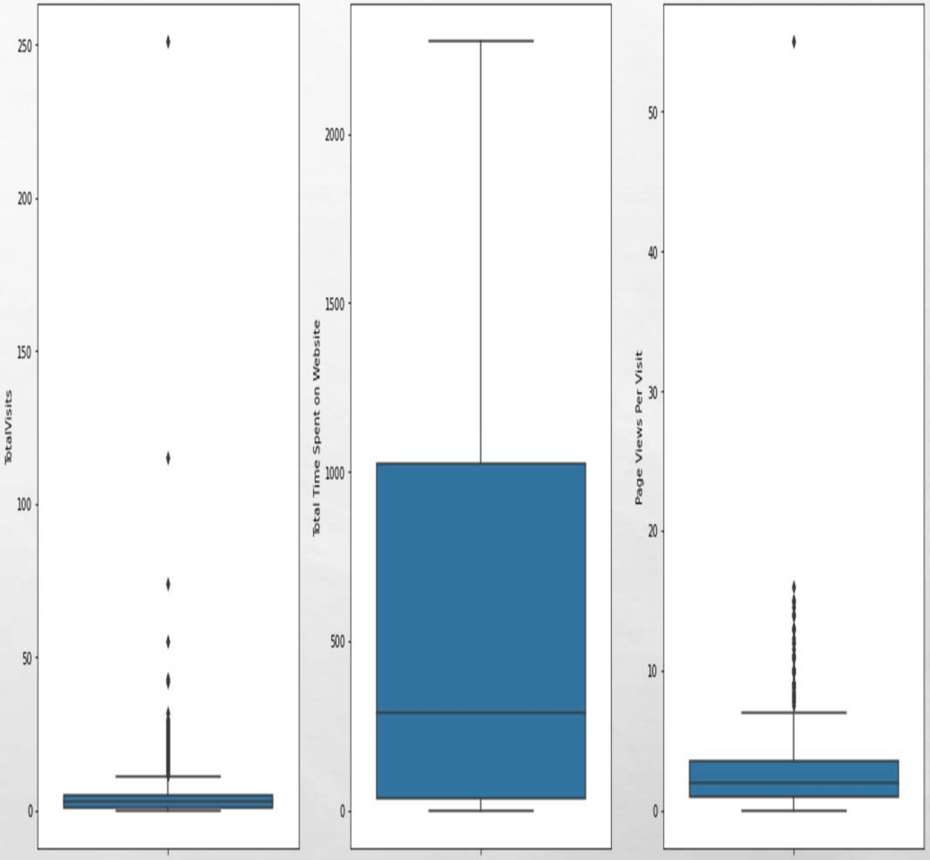
- ‘Leads.Csv’ contains all the information about the leads generated through various sources and their activities.
  - This file contains 9240 rows and 37 columns.
  - Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
- ‘Leads data dictionary.Csv’ is data dictionary which describes the meaning of the variables present in the “leads” dataset.

DATA CLEANING AND PREPARATION

- Following columns contain more than 30% null values initially like lead profile, what is your current occupation , lead quality, etc.
- Following columns have default value of 'select' as a dominating value which is same as null value. So, we have converted 'select' to 'NA' like specialization , how did you hear about X education , city, etc.
- Following columns have been dropped since percentage of missing value is more than 70% like how did you hear about x education, country, city, etc.

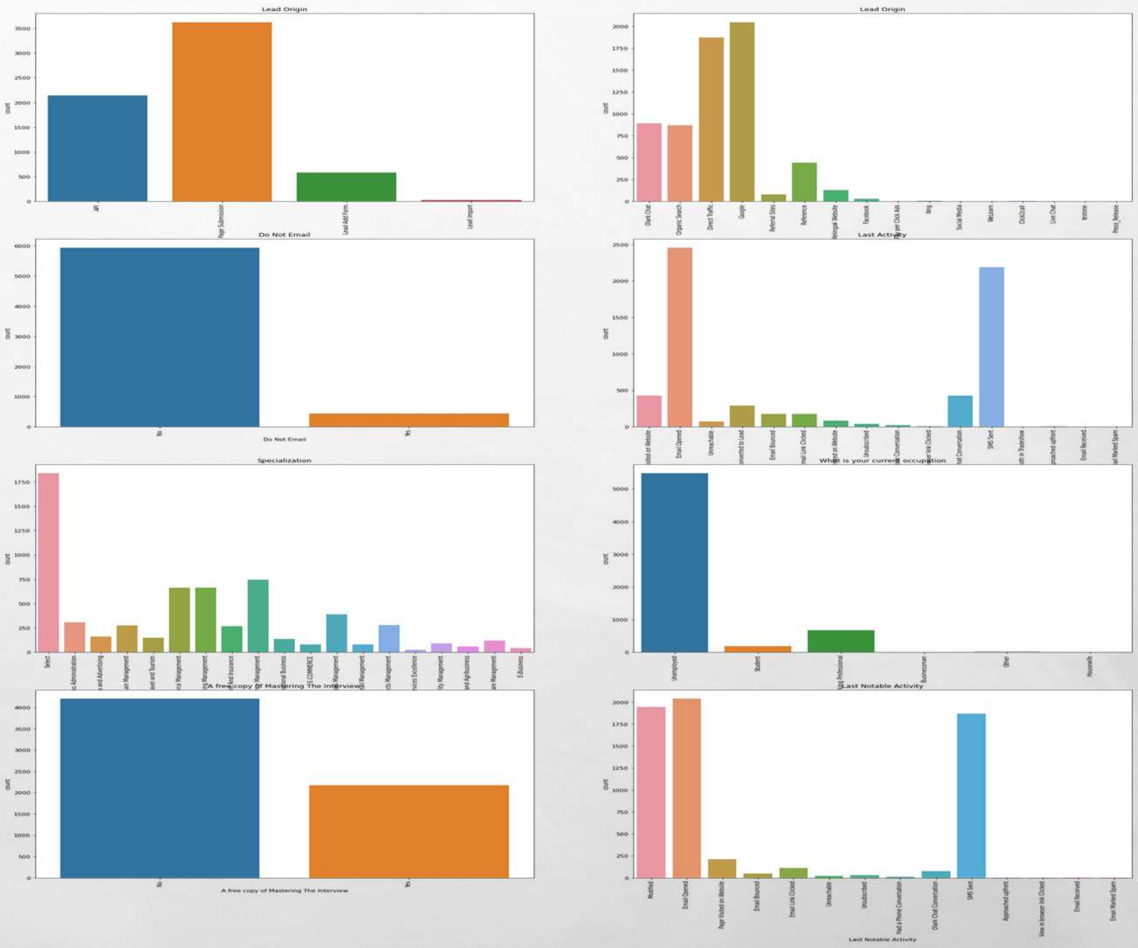
UNIVARIATE ANALYSIS

- Univariate analysis revealed data distribution and outliers in ‘leads’ data. Key columns where outliers were identified are:- a. Totalvisits b. Page views per visit c. Asymmetrique activity score d. Asymmetrique profile score.
- Inter quantile range (iqr) method has been used to treat outliers in the data.
- Decision has been taken to not remove any outliers since the % is high (9%).
- We will review the final model to ensure this does not impact the score



BIVARIATE ANALYSIS

- Converted' column has been chosen as target variable. So, bivariate analysis of important variables has been performed with respect to the target variable.
- Lateral students and the visitors showing interest on next batch have higher chances of getting converted.
- Lead quality tagged with "high in relevance" has high conversion rate history.
- Lead originated through "lead add form" and "quick add form" has high possibility of getting converted.
- Lead belongs to welingak website, welearn, live chat and nc_edm converts more than any other sources.



BUILDING A MODEL

- We build a model with all the features included and found there were many insignificant variables present in our model.
- We need to drop them, but we can't do it one by one as it is time consuming and not an efficient way to do so.
- Hence, we started with rfe method to deduct those insignificant variables. We choose with RFE count 19 and 15.
- We did two rfe count because we want to find out our final model stability.
- We started creating our model with rfe count 19 and went dropping variables one by one until we reach the point where the model is having all significant variables and low vif values.
- Now we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

PRECISION AND RECALL

- We used this cutoff point to create a new column in our final dataset for predicting the outcomes.
- After this we did another type of evaluation which is by checking precision and recall
- As we all know, precision and recall plays very important role in build our model more business oriented and it also tells how our model behaves.
- Hence, we evaluated the precision and recall for this model and found the score as 0.73 for precision and 0.79 for recall.
- Now, recall our business objective - the recall percentage i will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on recall than precision.
- I.E we get more relevant results - as many as hot lead customers from our model .

PREDICTION ON TEST SET

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.E. Finding the accuracy, precision and recall.
- The accuracy score we found was 0.82, precision 0.76 and recall 0.79 approximately.
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

CONCLUSION

- Valuable insights –
- The accuracy, precision and recall/sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- In business terms, this model has an ability to adjust with the company’s requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
- A) last notable activity had a phone conversation
- B) lead origin lead add form and
- C) what is your current occupation working professional.

Thank You