X Education

Lead Identification

Problem Statement:

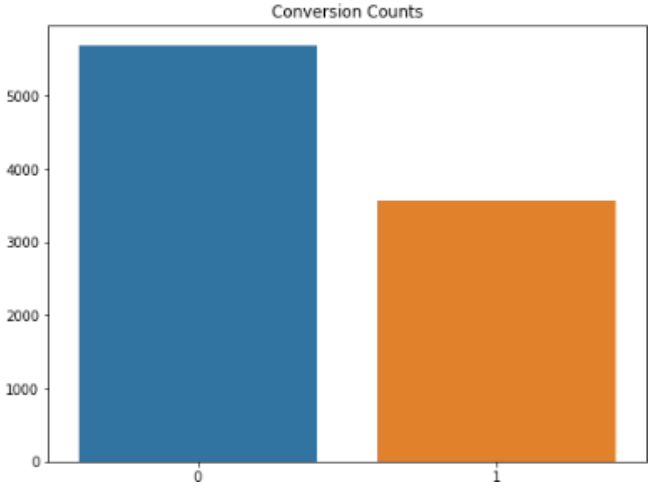
- The objective of this case study was to help the sales team of X education identify & distinguish hot leads from the ones who might not get converted.
- They also would need to be given the top variables and their values to look out for while trying to assess if the lead will get converted or not.

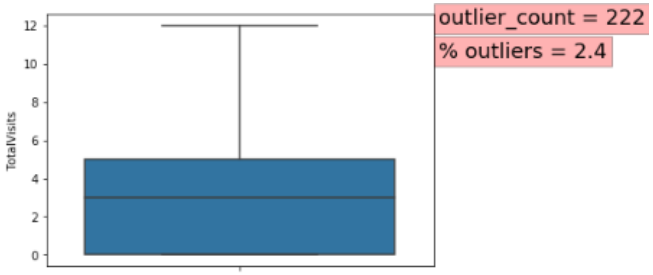
Understanding the data landscape

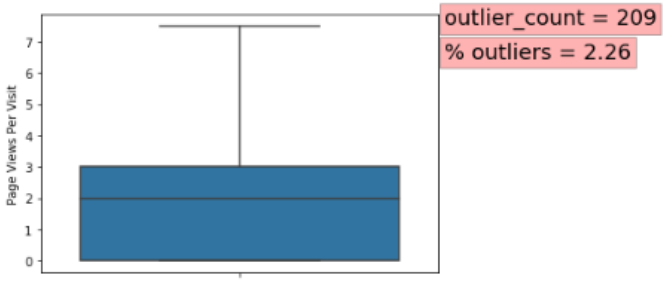
- The data does contain few **extreme** scenarios, in only 2 of the numerical variables which depict the number of visits and page views per visit. We can handle this, by capping the outlier values of these variables to a value slightly higher than the Q3 + 1.5 * IQR
- The data collection for this exercise has resulted in a lot of missing data, hence the **need of a strategic missing value imputation.** We can handle by removing some redundant, imputing few categorical variables with 'NA' category & 0 for numerical indexes.
- A lot of the categorical variables have only one levels or do not depict any difference in the conversion ratio across different levels, and thus can be omitted from the analysis, while some variables can be recoded.

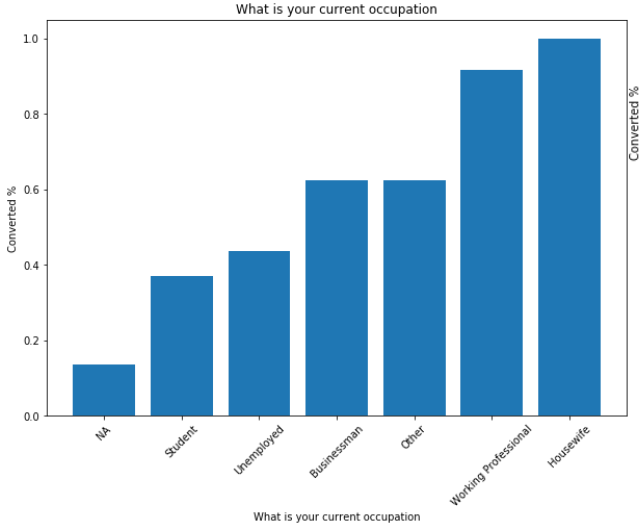
Apoorv Gupta

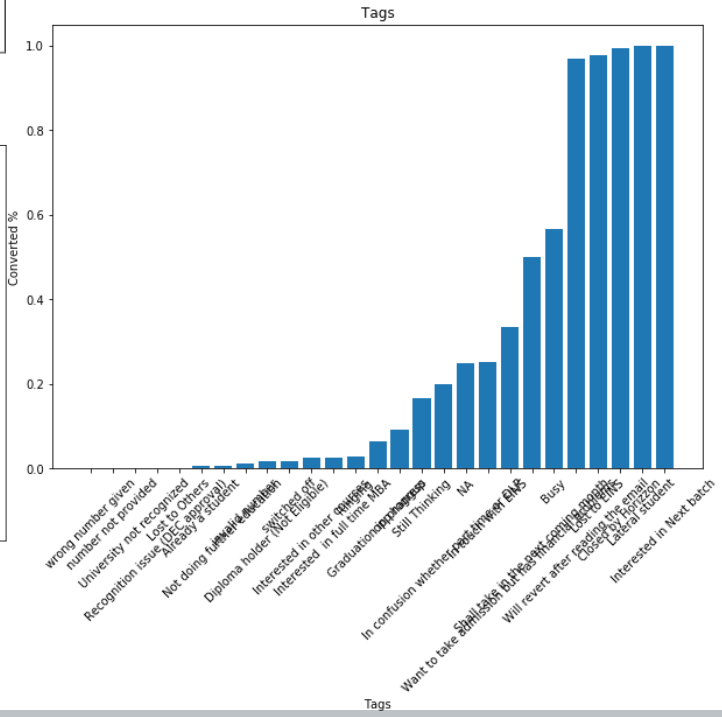
How do the key parameters look?

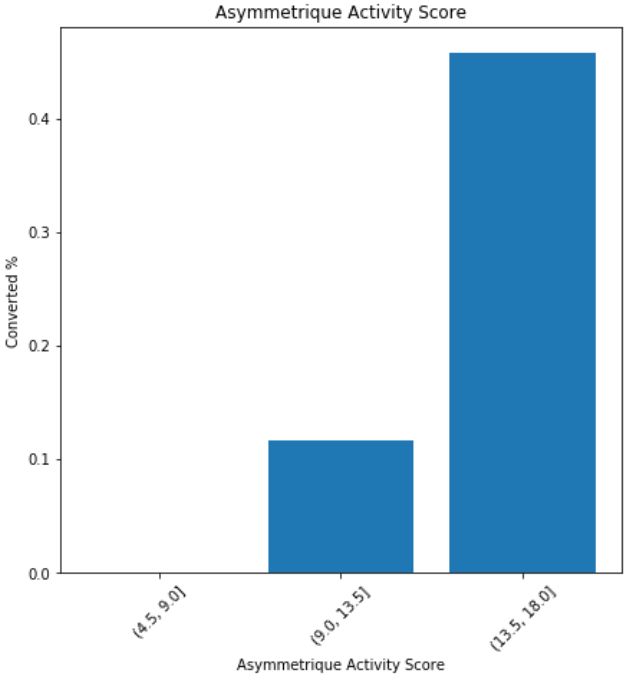












Apoorv Gupta

Lets get technical. What to do next?

- **Treat the variables** – Recode categorical variables, cap the outliers in the numerical variables, drop the insignificant ones.
- Dummy **encode** the categorical variables.
- **Standardize/scale** these parameters of study to a comparable, equal range.

Do the parameters represent similar information?

The data has a very high number of features, and we will need to use VIF/ Correlations to reduce the dimensions.

Variable Pairs With High Correlation

index	variable	value
Asymmetrique Profile Score	Asymmetrique Activity Score	0.973984
Asymmetrique Profile Index_NA	Asymmetrique Activity Score	-0.989867
What matters most to you in choosing a course_NA	What matters most to you in choosing a course_...	-0.999217
Lead Origin_Landing Page Submission	City_NA	-0.808285
Asymmetrique Activity Score	Asymmetrique Activity Index_NA	-0.989867
Asymmetrique Profile Score	Asymmetrique Activity Index_NA	-0.986816
Asymmetrique Profile Index_NA	Asymmetrique Activity Index_NA	1.000000
Lead Origin_Landing Page Submission	Lead Origin_API	-0.842492
Lead Source_Reference	Lead Origin_Lead Add Form	0.853237
Lead Source_Facebook	Lead Origin_Lead Import	0.981709
What matters most to you in choosing a course_...	What is your current occupation_NA	-0.994262
What matters most to you in choosing a course_NA	What is your current occupation_NA	0.995041
Asymmetrique Profile Score	Asymmetrique Profile Index_NA	-0.986816
What matters most to you in choosing a course_...	Tags_NA	-0.812534
What matters most to you in choosing a course_NA	Tags_NA	0.813335
What is your current occupation_NA	Tags_NA	0.809020

Variable Retained After RFE & Elimination Through Correlations:

Most levels of the following variables were retained:

Tags

- Asymmetrique Activity Index
- Asymmetrique Profile Index
- Lead Source
- What matters most to you in choosing a course

Along with the numerical variable:

- Asymmetrique Activity Score

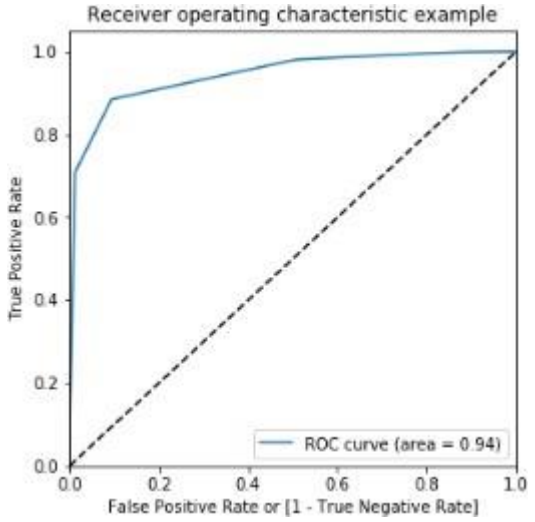
Build - Check Features, Repeat!

- Build the model with the current set of variables
- Check the VIF /p values & eliminate insignificant variables
- Repeat above steps until all the features are significant!

- How did the model perform now?

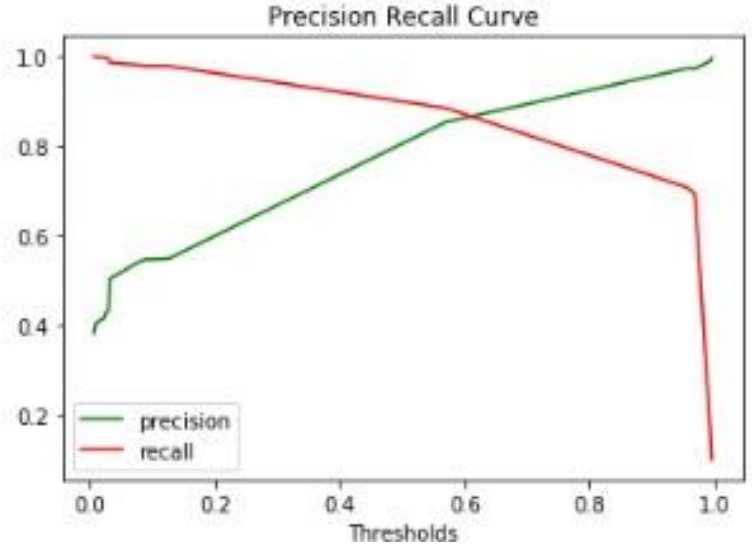
```
Accuracy =  0.8979591836734694
Sensitivity =  0.8848337388483374
Specificity =  0.9060469765117442
False Positive Rate =  0.09395302348825588

Precision =  0.853010164190774
Recall =  0.8848337388483374
```

ind	predicted_no	predicted_yes
actual_no	3626	376
actual_yes	284	2182



Can we improve the performance further?

- The performance of the model can further be improved by tuning the threshold value.
- We can derive from the graph below that 0.6 would be a more optimised threshold.



Improvement seen in the model:

High improvement in precision, which is crucial in this case

```
Accuracy =  0.8811069882498453
Sensitivity =  0.7076236820762368
Specificity =  0.9880059970014993
False Positive Rate =  0.01199400299850075

Precision =  0.9732292247629671
Recall =  0.7076236820762368
```

Appendix: Model Performance

- Performance on train data:

```
Accuracy =  0.8811069882498453
Sensitivity =  0.7076236820762368
Specificity =  0.9880059970014993
False Positive Rate =  0.01199400299850075

Precision =  0.9732292247629671
Recall =  0.7076236820762368
```

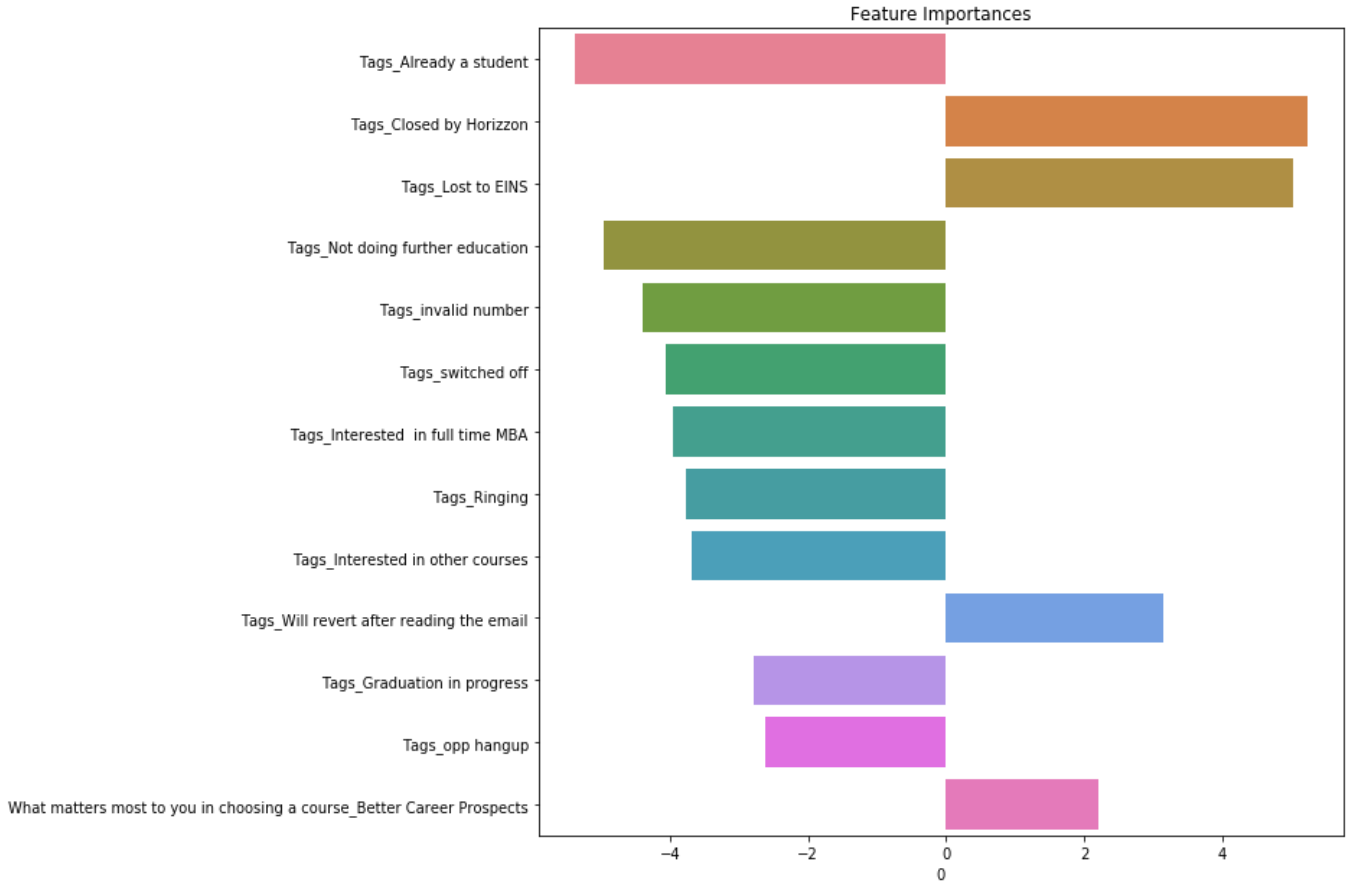
- Performance on test data:

```
Accuracy =  0.8813131313131313
Sensitivity =  0.7205479452054795
Specificity =  0.9862850327966607
False Positive Rate =  0.013714967203339297

Precision =  0.9716748768472906
Recall =  0.7205479452054795
```

- In this problem it is more important to have a higher precision than recall as it is more important to reduce number of false positives than false negatives as if you have a high number of false positives, that is, the model predicted leads who actually would not be potentially converted as a hot lead, then the counselors are actually wasting their time in concentrating on leads who are not beneficial and leaving out the actual hot leads. This might cause a huge loss in the overall conversion.

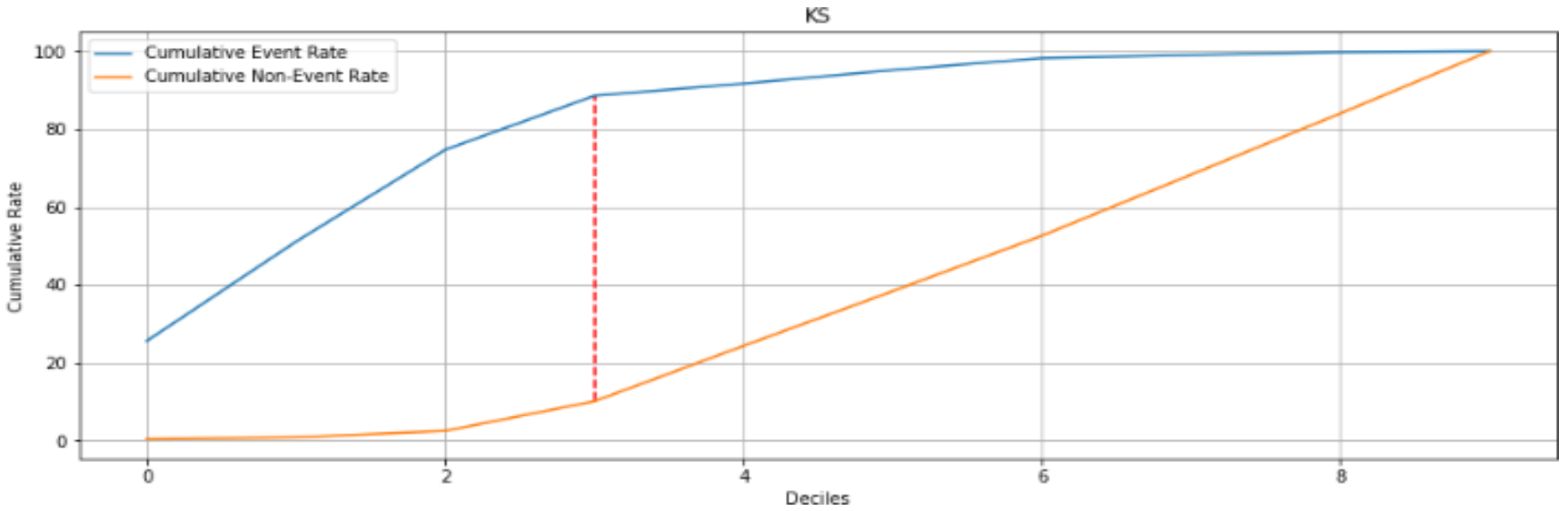
Feature Importance

- Below are the importance of the variables along with the direction of their impact:



Can the sales team achieve more?

- We can see that as per the probability cut-off that we have taken as 0.6, we get the best precision. This section of the data is covered in the top 3 deciles itself. However, if there is time left out for the team, it is possible to consider the next best set of the leads, that is the 4th decile, only if it is worth it.
- In the KS Chart above, we can clearly see that the 4th decile also gives a good or atleast a decent conversion rate, post which, from the 5th decile, the conversion rate drops drastically. So even if the team completes following up on the leads as per the threshold of 0.6, we can push the team to only pursue the next set of 10% of leads only and not further than that as the conversion rate drops as low as 3% post the 4th decile.



Conclusion

- The top important variables in understanding if the lead will get converted or not are below:
  - Tag
    - Tags_Already a student
    - Tags_Closed by Horizzon
    - Tags_Lost to EINS
    - Tags_Not doing further education
  - What matters most to you in choosing a course
    - Better Career Prospects
- The model can very well assist the sales team in identifying the targets as it has a precision of 97% in both the train and test data. It also has a very good accuracy of 88%. This means that the team can precisely identify and convert 97% of the hot leads by just trusting the model.
- However, if the sales team does run out of the existing leads and even completed the sales target for the quarter, it is still beneficial if the sales team can call the next set of leads as the model has a good conversion rate until the 4th decile. So even if the sales team calls the next best set, they will get a good portion of them converted and not waste their time!