REPORT

PROBLEM STATEMENT

Primary objective of this data is to build features and model on these characteristics of users to calculate a **score/rank** for conversion probability of that user. These scores will eventually decide the bidding logic used for each user.

We have worked on the following parts:

a. Feature Engineering (Variable Imputation)

b. Visualisation (Insights)

c. Model Selection Criteria (Basis of choosing the final Technique)

d. Measurement Criteria (Comparison of Various Models)

e. Scope for improvement

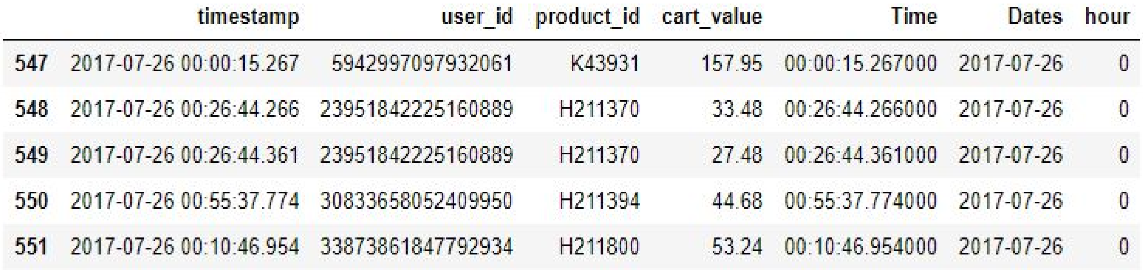
APPROACH TAKEN

The data is divided in to two files. One is the browsing behaviour (containing Timestamp, UserId and website section visited) and the conversion file (containing Timestamp, UserId, Products Purchased in the transaction, Overall Cart Value).

- Firstly, we have tried understanding the data. What we found:
  - o Dealing with missing data.
    The missing values in the data were all focused in the user_id column. We have dropped the user_id where the values are 0 since there is numerous activities in the data attributed to these unknown users, which might act as an outlier. We found it would be unreasonable to

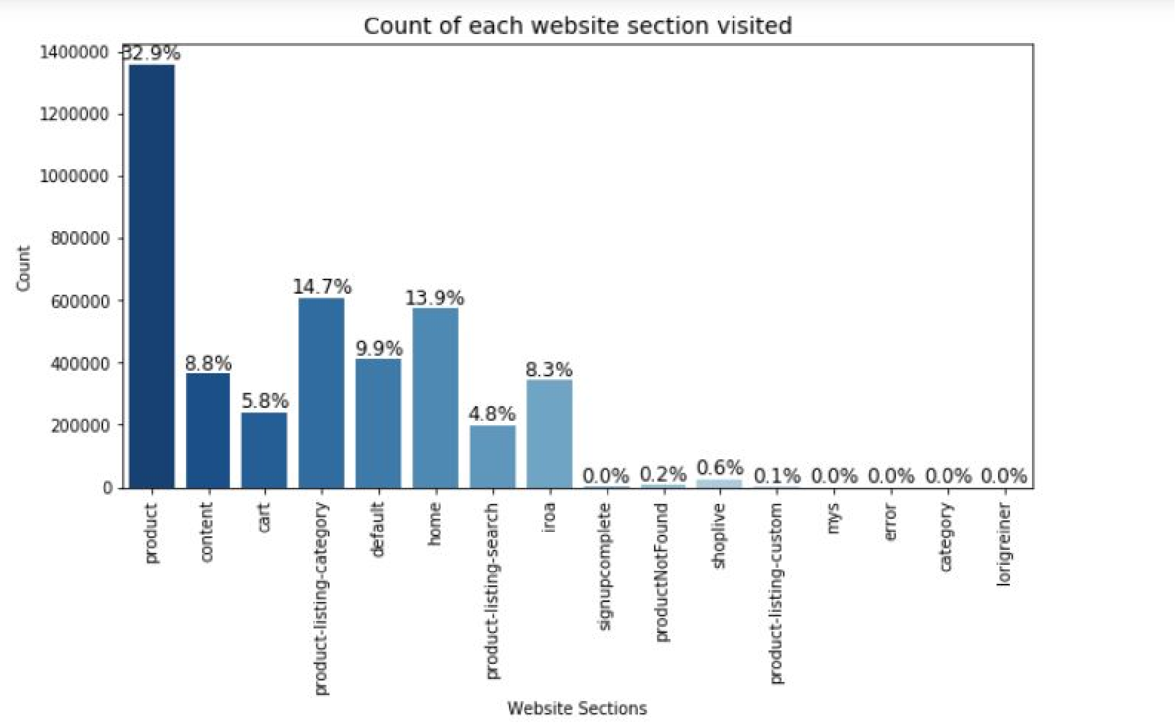
impute it based on the other user_id, as user_id is unique for each customer.

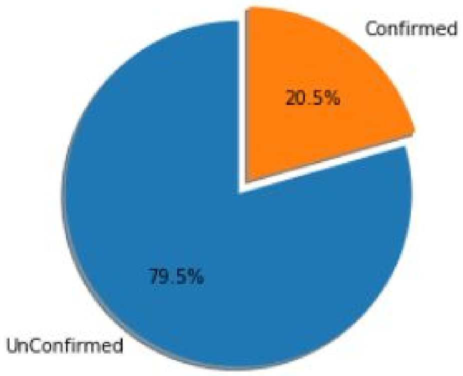
- Cleaned the timestamp column in the both files.
  Firstly, we converted timestamp into a datetime format. After dividing the column into two separate features (date and time), it was noticed the date was the same for all of the data. Hence the primary focus will lie on Time.



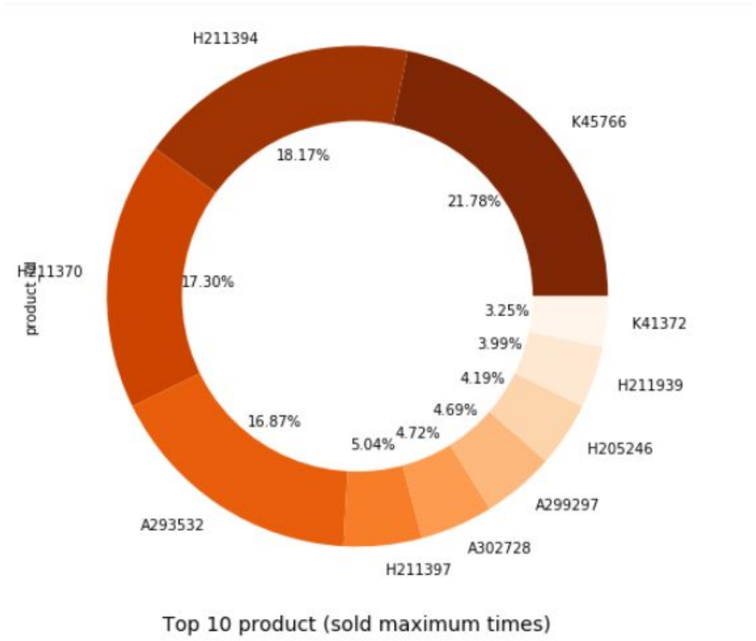
	timestamp	user_id	product_id	cart_value	Time	Dates	hour
547	2017-07-26 00:00:15.267	5942997097932061	K43931	157.95	00:00:15.267000	2017-07-26	0
548	2017-07-26 00:26:44.266	23951842225160889	H211370	33.48	00:26:44.266000	2017-07-26	0
549	2017-07-26 00:26:44.361	23951842225160889	H211370	27.48	00:26:44.361000	2017-07-26	0
550	2017-07-26 00:55:37.774	30833658052409950	H211394	44.68	00:55:37.774000	2017-07-26	0
551	2017-07-26 00:10:46.954	33873861847792934	H211800	53.24	00:10:46.954000	2017-07-26	0

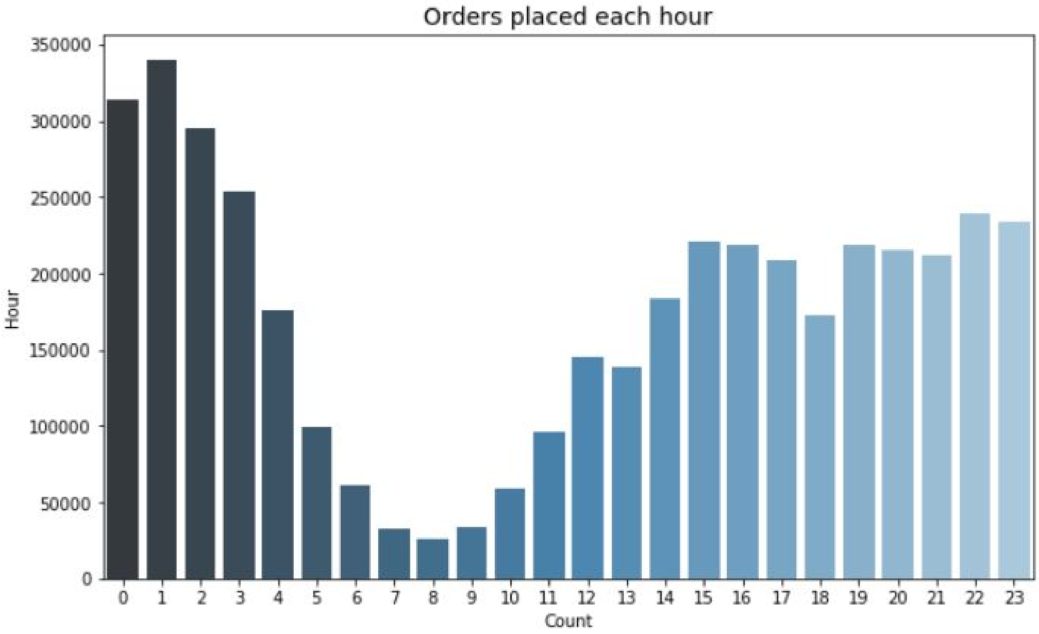
- After cleaning the data, we try to get insights by using visualization techniques with the help of seaborn and matplot library. Some of the insights are as follows:



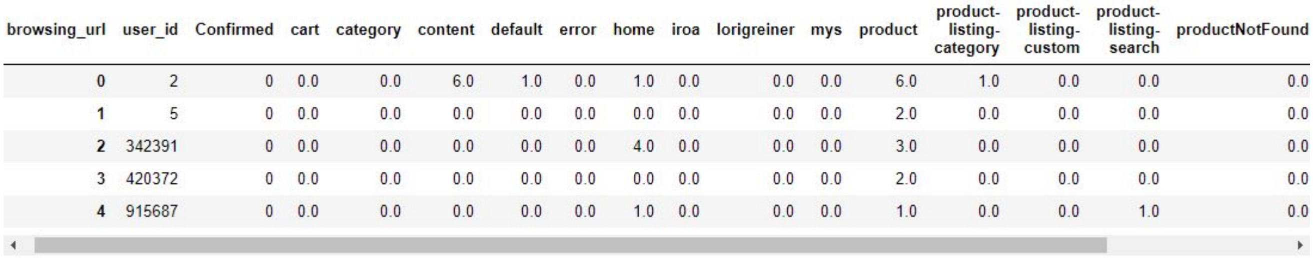


In a particular day almost 20% users are buying any product from the site





- Lastly, for feature selection, we have tried to create features for each browsing_url using pivot_table (panda function) and filled the values with the number of times the user have visited that url.

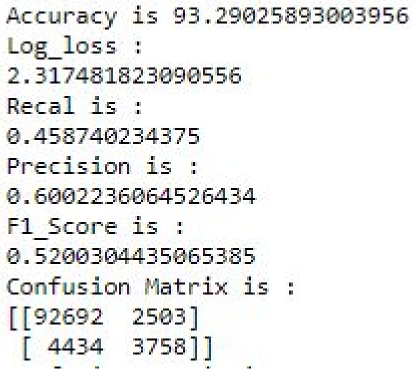


browsing_url	user_id	Confirmed	cart	category	content	default	error	home	iroa	lorigreiner	mys	product	product-listing-category	product-listing-custom	product-listing-search	productNotFound
0	2	0	0.0	0.0	6.0	1.0	0.0	1.0	0.0	0.0	0.0	6.0	1.0	0.0	0.0	0.0
1	5	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
2	342391	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0
3	420372	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
4	915687	0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

We have implemented various machine learning algorithms such as Logistic Regression, RandomForest, AdaBoost and Gradient Boosting in order to identify the most optimal model. For each of these models the Confusion Matrix, Accuracy, Recall and F1score were calculated. Further we found the feature importance and their ranking based on the models we have used.

INTERPRETATION OF RESULTS

The results that were inferred from the data were:



```
Accuracy is 93.29025893003956
Log_loss :
2.317481823090556
Recal is :
0.458740234375
Precision is :
0.6002236064526434
F1_Score is :
0.5200304435065385
Confusion Matrix is :
[[92692  2503]
 [ 4434  3758]]
```

Confusion Matrix:

The matrix represents 92692 true positive values which means that all of these user transactions were correctly predicted as leading to a Confirmation.

3758 are true negative values which means that all of these user transactions were correctly predicted as not leading to a Confirmation.

2503 are false positive values which means that all of these user transactions were predicted as leading to a Confirmation but did not actually result in a Confirmation.

4434 are false negative values which means that all of these user transactions were predicted as not leading to a Confirmation but actually result in a Confirmation.

Accuracy:

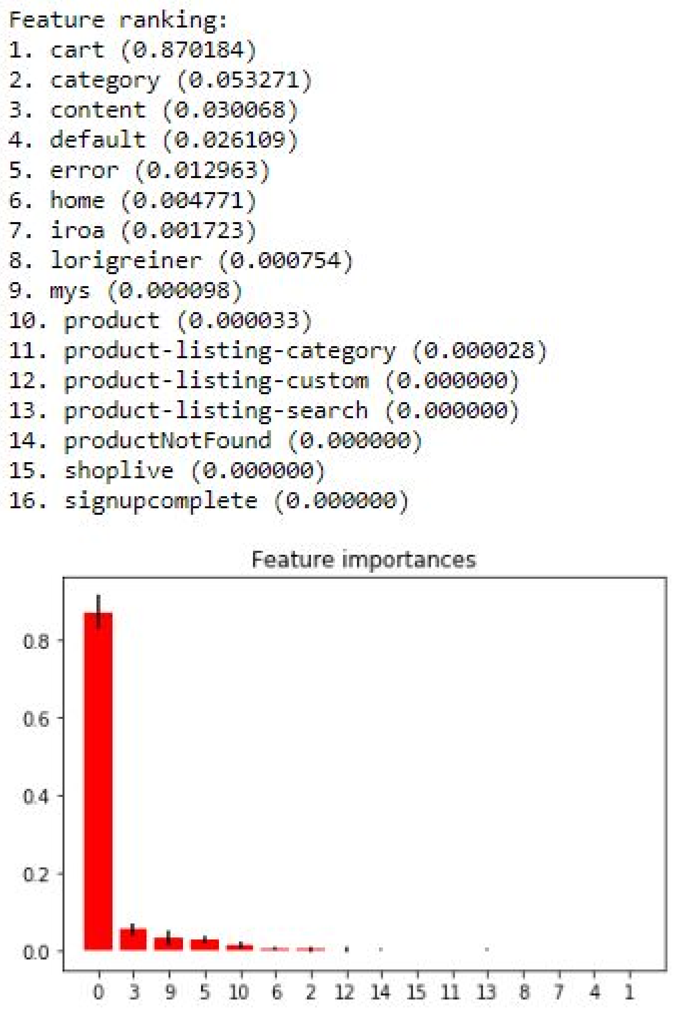
The accuracy 93.29 is not used as an appropriate representation of the model because classes are imbalanced.

Precision:

The score 0.60 implies that if a model predicts a class for the user as confirmed there is a 60% likelihood that the model is correct.

Recall:

The score 0.45 implies that the model can recall 45% of instances of a particular class.



```
Feature ranking:
1. cart (0.870184)
2. category (0.053271)
3. content (0.030068)
4. default (0.026109)
5. error (0.012963)
6. home (0.004771)
7. iroa (0.001723)
8. lorigreiner (0.000754)
9. mys (0.000098)
10. product (0.000033)
11. product-listing-category (0.000028)
12. product-listing-custom (0.000000)
13. product-listing-search (0.000000)
14. productNotFound (0.000000)
15. shoplive (0.000000)
16. signupcomplete (0.000000)
```

SCOPE FOR IMPROVEMENT

In future, we will improve our model by

- Finding how much total time is spent by each user in each section and using that feature improve the model and results.
- Using various optimization techniques and hyper parameter tuning: Optimization formulations contain scalar parameters that balance data-fitting with desired structure. Need to tune these parameters