Online Shopper's Intention

How to best optimise the customer's online experience?

CS- 513 A KDD -Final Project

presented by

Prithiv Dev Devendran
10453922

Shreya Vhadadi
10453495

Chran Suresh
10450732

Problem Statement

The customer goes through a series of chronological steps to making a purchase in an E-commerce website. These steps are tracked down and classified to know a purchased is made. From this problem we seek to find the crucial part of the decision making process and improvise the outcome.

Dataset from Kaggle

	Administrati	Administrati	Informationa	Informationa	ProductRela	ProductRelate	BounceRates	ExitRates	PageValue	SpecialDay	Month	Operating	Browser	Region	TrafficTyp	VisitorTyp	Weekend	Revenue
2	0	0	0	0	8	260	0.008333333	0.025	0	0.8	Feb	3	2	1	3	Returning_Vi	TRUE	FALSE
3	0	0	16	1210.397619	5	279.8571429	0.003174603	0.0127642	0	0	Mar	2	2	1	8	Returning_Vi	FALSE	FALSE
4	0	0	0	0	20	927.45	0.011111111	0.0272487	8.0007407	0	Mar	1	1	3	1	Returning_Vi	FALSE	FALSE
5	2	9	0	0	50	836.8	0	0.0063399	0	0	Mar	2	2	3	2	Returning_Vi	FALSE	FALSE
6	10	293.7782051	2	153	96	3283.166739	0.001960784	0.0135094	0	0	Mar	3	2	6	2	Returning_Vi	TRUE	FALSE
7	9	111.5	1	48.5	49	1868.819697	0	0.0207089	1.706015	0	Mar	2	2	7	2	Returning_Vi	FALSE	TRUE
8	3	47	1	51	68	3008.124108	0.007142857	0.0167279	46.530175	0	Mar	2	2	1	1	Returning_Vi	FALSE	FALSE
9	0	0	0	0	2	0	0.2	0.2	0	0	Mar	1	1	1	3	Returning_Vi	FALSE	FALSE
10	3	1226	5	3	24	3230.25	0.036190476	0.096	0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
11	3	52	0	0	9	319	0	0.02	0	0	Mar	1	1	1	1	Returning_Vi	FALSE	FALSE
12	0	0	0	0	3	42	0	0.0666667	0	0	Mar	2	2	3	1	Returning_Vi	FALSE	FALSE
13	0	0	0	0	22	354.3333333	0.009090909	0.0576623	0	0	Mar	1	1	7	2	Returning_Vi	TRUE	FALSE
14	0	0	0	0	98	3556.61241	0.002061856	0.0101735	0	0	Mar	1	1	1	3	Returning_Vi	FALSE	FALSE
15	2	56	1	144	67	2563.783333	0	0.0057971	19.34265	0	Mar	2	2	4	2	New_Visitor	FALSE	TRUE
16	3	112.9607843	0	0	13	3014.018519	0.013068182	0.0614063	0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
17	0	0	0	0	17	840.2333333	0	0.0016667	109.176	0	Mar	2	2	9	2	New_Visitor	FALSE	TRUE
18	3	94	2	125	55	1970.844805	0	0.0017241	96.255116	0	Mar	2	4	1	2	New_Visitor	TRUE	TRUE
19	1	32	0	0	50	2867	0	0.004	153.44325	0	Mar	2	2	7	8	Returning_Vi	TRUE	TRUE
20	0	0	0	0	5	43	0	0.04	0	0	Mar	1	1	1	9	Returning_Vi	TRUE	FALSE
21	5	218	0	0	13	284.5	0	0.0041667	0	0	Mar	1	1	1	2	New_Visitor	FALSE	FALSE
22	0	0	0	0	3	133.5	0	0.0888889	0	0	Mar	1	2	1	8	Returning_Vi	FALSE	FALSE
23	1	119	0	0	12	297.6666667	0	0.0083333	0	0	Mar	3	2	1	10	Returning_Vi	FALSE	FALSE
24	3	281	0	0	16	453.75	0	0.0052632	0	0	Mar	2	2	5	2	New_Visitor	FALSE	FALSE
25	1	18	0	0	16	1331.75	0	0.0125	33.799567	0	Mar	2	5	2	3	New_Visitor	FALSE	TRUE
26	2	40	0	0	5	558.5	0	0.0285714	0	0	Mar	2	2	7	2	New_Visitor	TRUE	FALSE
27	3	107	0	0	4	145.8333333	0	0.0166667	0	0	Mar	3	2	7	1	New_Visitor	FALSE	FALSE
28	0	0	0	0	8	731	0	0.04375	0	0	Mar	2	4	1	3	Returning_Vi	FALSE	FALSE
29	2	49	1	127	4	188	0.033333333	0.0375	0	0	Mar	3	2	3	2	Returning_Vi	FALSE	FALSE
30	0	0	0	0	14	565.3333333	0.037619048	0.0495238	0	0	Mar	2	6	3	1	Returning_Vi	FALSE	FALSE
31	0	0	0	0	2	50	0	0.05	0	0	Mar	3	2	8	11	Returning_Vi	FALSE	FALSE
32	4	57	0	0	7	591	0.006666667	0.0333333	0	0	Mar	1	2	6	6	Returning_Vi	TRUE	FALSE
33	2	2	3	261	10	281	0	0.0428571	0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
34	2	123	2	306.3333333	18	483.8333333	0	0.02	0	0	Mar	2	2	4	1	Returning_Vi	FALSE	FALSE
35	2	118	0	0	4	42	0	0.0666667	0	0	Mar	1	2	7	3	Returning_Vi	FALSE	FALSE
36	0	0	0	0	35	1192.386111	0.005714286	0.0285714	0	0	Mar	2	4	2	2	Returning_Vi	FALSE	FALSE
37	0	0	0	0	12	198	0.016666667	0.075	0	0	Mar	2	2	3	2	Returning_Vi	FALSE	FALSE
38	0	0	0	0	3	15	0	0.0666667	0	0	Mar	2	4	3	12	Returning_Vi	FALSE	FALSE
39	2	38	0	0	14	643	0	0.0133333	35.0928	0	Mar	2	2	5	1	Returning_Vi	FALSE	TRUE
40	0	0	0	0	5	23	0.08	0.12	0	0	Mar	2	2	1	1	Returning_Vi	FALSE	FALSE

Legend

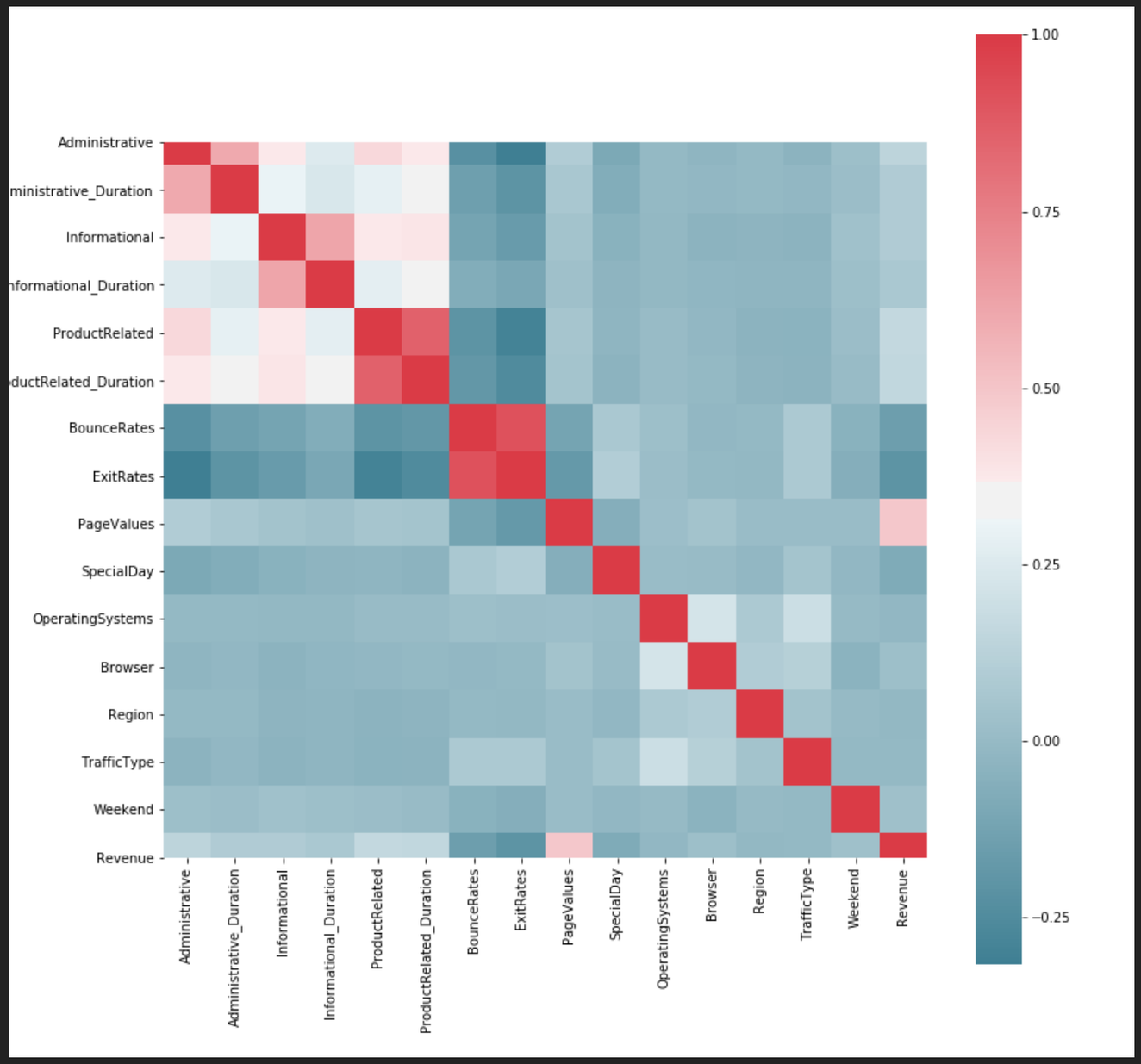
Feature	Description
Administrative	The page relating to Administrative setup.
Administrative_Duration	Time spent viewing the Administrative page.
Informational	The page relating to Profile info.
Informational_Duration	Time spent viewing the Informational page.
Product Related	Cost of the product in dollars viewed by the customer
Product Related_Duration	Time spent viewing the Product_Related page
Bounce Rate	
Exit Rate	Percentage of visitors to a page on the website from which they exit the website to a different site. ifferent website.
Page Values	Average value for a page that a user visited before landing on the goal page or completing an transaction.
Special Day	Holidays, festival days
Month	Month of the year
OS	Operating System used
Browser	Web page terminal
Region	Location of the customer
Traffic Type	How users arrived to that site
Visitor Type	Type of visitor
Weekend	Weekend or not
Revenue	Revenue will be generated or not

Exploratory Data Analysis

- The plotted graph shows the correlation matrix
  between the features in the dataset.

- The “PRODUCT_RELATED” and “PRODUCT RELATED _DURATION” have considerable amount of correlation between them but we felt the cost of the product need not be related to the amount of time spent on the product.

- The “BOUNCE RATE” and “EXIT RATE” also have high correlation but weren’t removed because they were similar aspects of the dataset but represented two important factors of the dataset.



Pre-processing

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Administrati	Administrati	Informationa	Informationa	ProductRela	ProductRelate	BounceRates	ExitRates	PageValue	SpecialDay	Month	Operating	Browser	Region	TrafficTyp	VisitorTyp	Weekend	Revenue
1067									0	0	Mar	2	2	2	1	Returning_Vi	FALSE	FALSE
1134									0	0	Mar	1	1	1	2	Returning_Vi	FALSE	FALSE
1135									0	0	Mar	2	4	5	1	Returning_Vi	FALSE	FALSE
1136									0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
1137									0	0	Mar	3	2	1	1	Returning_Vi	FALSE	FALSE
1138									0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
1475									0	0	Mar	2	2	1	1	Returning_Vi	TRUE	FALSE
1476									0	0	Mar	1	1	6	1	Returning_Vi	TRUE	FALSE
1477									0	0	Mar	2	2	3	1	Returning_Vi	FALSE	FALSE
1478									0	0	Mar	1	1	2	3	Returning_Vi	FALSE	FALSE
2039									0	0	Mar	3	2	4	1	Returning_Vi	FALSE	FALSE
2040									0	0	Mar	2	2	1	2	Returning_Vi	FALSE	FALSE
2041									0	0	Mar	3	2	4	15	Returning_Vi	TRUE	FALSE
2755									0	0	May	2	2	4	13	Returning_Vi	FALSE	FALSE
12332																		

- There are 14 missing values in each and every column where all the values are from the same record.
- Due to this fact we removed them, since didn’t want to make imputations that would drastically make modifications.
- We changed the text data into discrete numeric values.
- We changed the Class labels - Revenue-False to be 0 and Revenue-True to be 1.

Baseline Model

Baseline Model

```
The Accuracy for all Zeros Prediction:  0.8555194805194806
```

Y Test	Random
0	0
0	1
0	0
1	0

```
The Accuracy for Random Prediction:  0.5040584415584416
```

Y Test	Zeros
0	0
0	0
0	0
1	0

Y Test	Ones
0	1
0	1
0	1
1	1

```
The Accuracy for all Ones Prediction:  0.1444805194805195
```

- The first tabulation shows random prediction for the output.
- The second tabulation shows all zeros prediction for the output.
- The third tabulation shows all ones prediction for the output

Intuition from the baseline

- It is evident that the dataset is imbalanced.
- Finding accuracy can actually mislead the prediction of this classification problem.
- Thus from this understanding we can use precision as an evaluation metric to find how good each of our class is classified.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Precision has two factors : true positive and false positive.
- True positive - correctly predicts the positive class.
- False positive - mis-predicts the positive class.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Model

Naive Bayes

KNN

Decision Tree

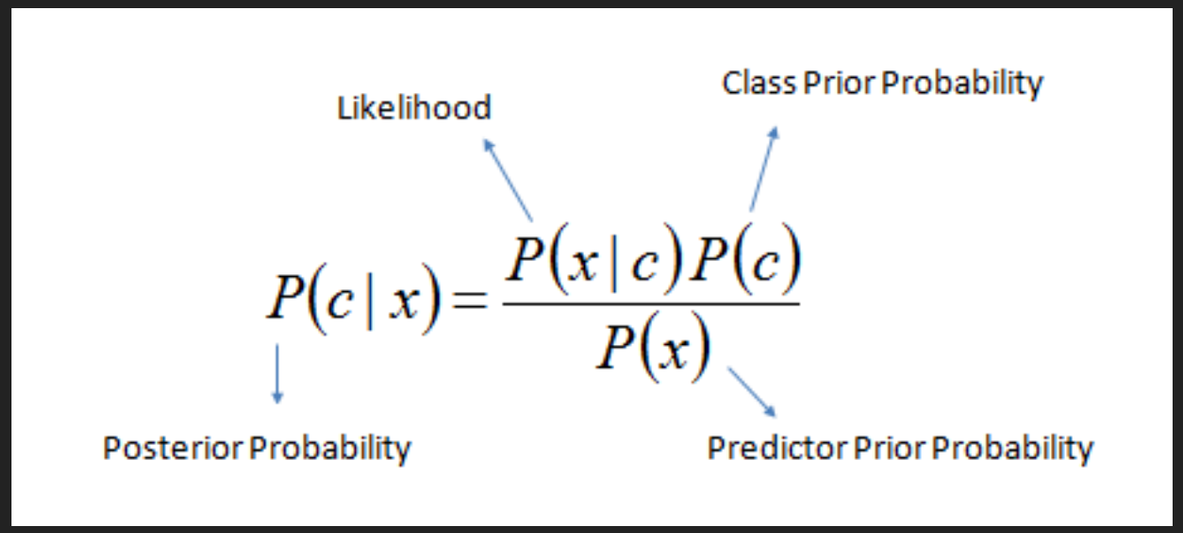
K Means

ANN

Random Forest

Naive Bayes

- We have a naive assumption that each column is independent of each other and the conditional probability with respect to the classes are calculated.



- The precision for the Revenue - False(0) has considerably bet the baseline model(random).

- The precision for the Revenue- True(1) is at par with baseline model.

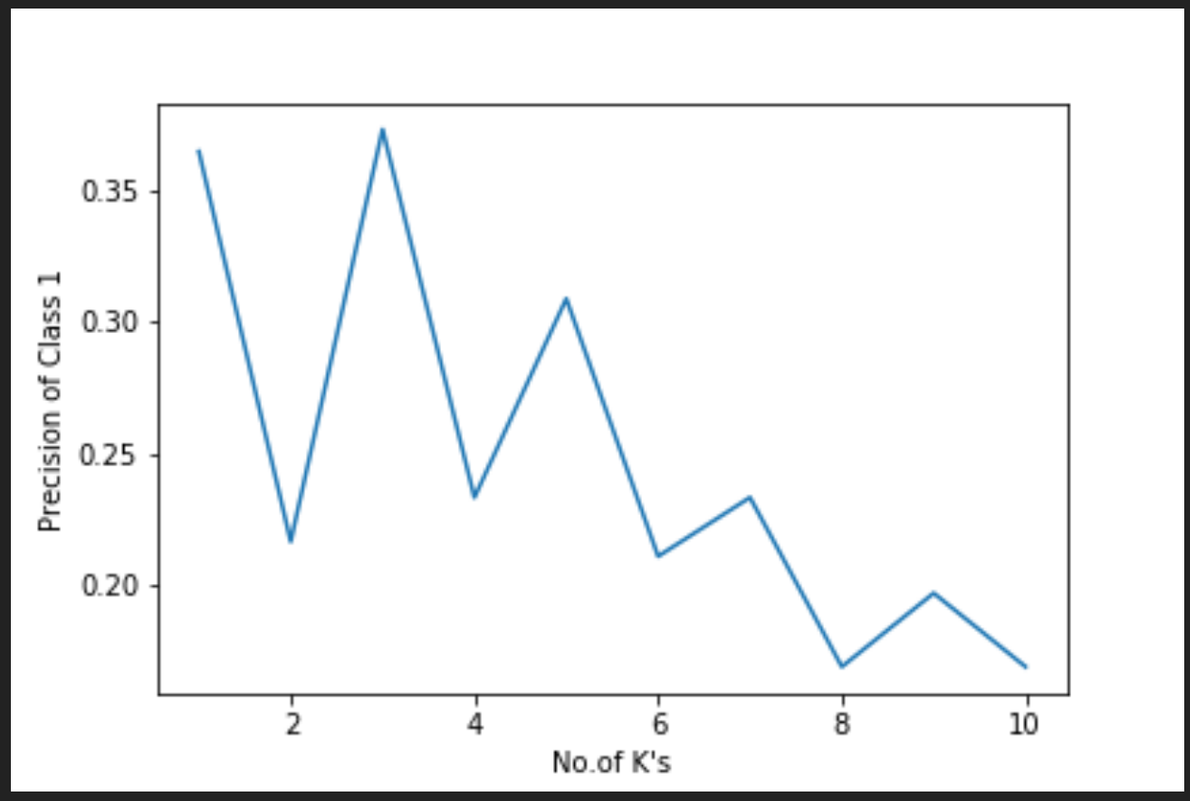
	precision	recall	f1-score	support
0	0.91	0.91	0.91	2092
1	0.50	0.48	0.49	372
accuracy			0.85	2464
macro avg	0.70	0.70	0.70	2464
weighted avg	0.85	0.85	0.85	2464

K Nearest Neighbour

	Revenue - True	Revenue - False
1	0.91	0.37
2	0.98	0.22
3	0.95	0.37
4	0.98	0.23
5	0.97	0.31
6	0.98	0.21
7	0.98	0.23
8	0.99	0.17
9	0.98	0.20
10	0.99	0.17

- Model was iterated for 1 to 10 neighbours
- The precision for Class 0 was relatively high than Class 1
- In the below graph every odd K values tend to shoe more precision because of the fact that the problem is a binary classification problem.

	precision	recall	f1-score	support
0	0.95	0.90	0.93	2236
1	0.37	0.58	0.46	228
accuracy			0.87	2464
macro avg	0.66	0.74	0.69	2464
weighted avg	0.90	0.87	0.88	2464



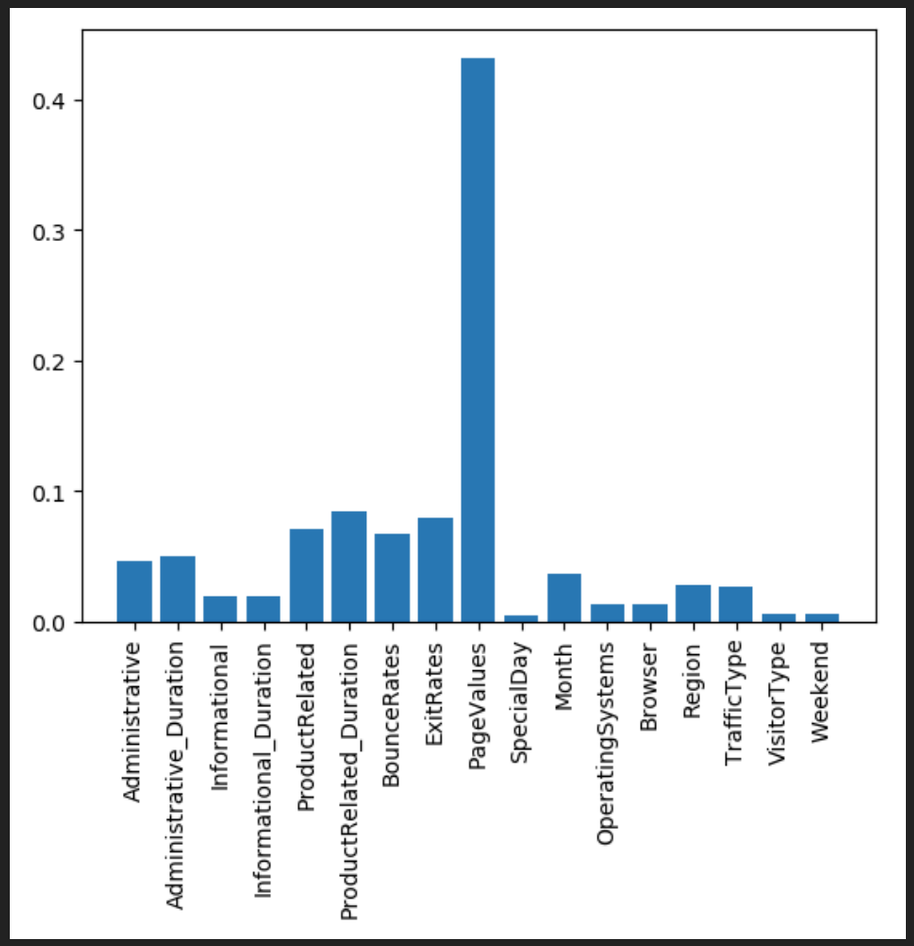
Decision Tree

- Similarly like Naive Bayes, Decision Tree has surpassed
  the baseline model for Revenue-False(0), but on par with the
  baseline model for Revenue-True(1).

- This model is on par with NB.

- From this model the most important features used to build the model
  is extracted and plotted.

- In which “PAGE VALUE” seems to be the most important attribute
  for classifying.



	precision	recall	f1-score	support
0	0.91	0.92	0.91	2086
1	0.51	0.48	0.50	378
accuracy			0.85	2464
macro avg	0.71	0.70	0.71	2464
weighted avg	0.85	0.85	0.85	2464

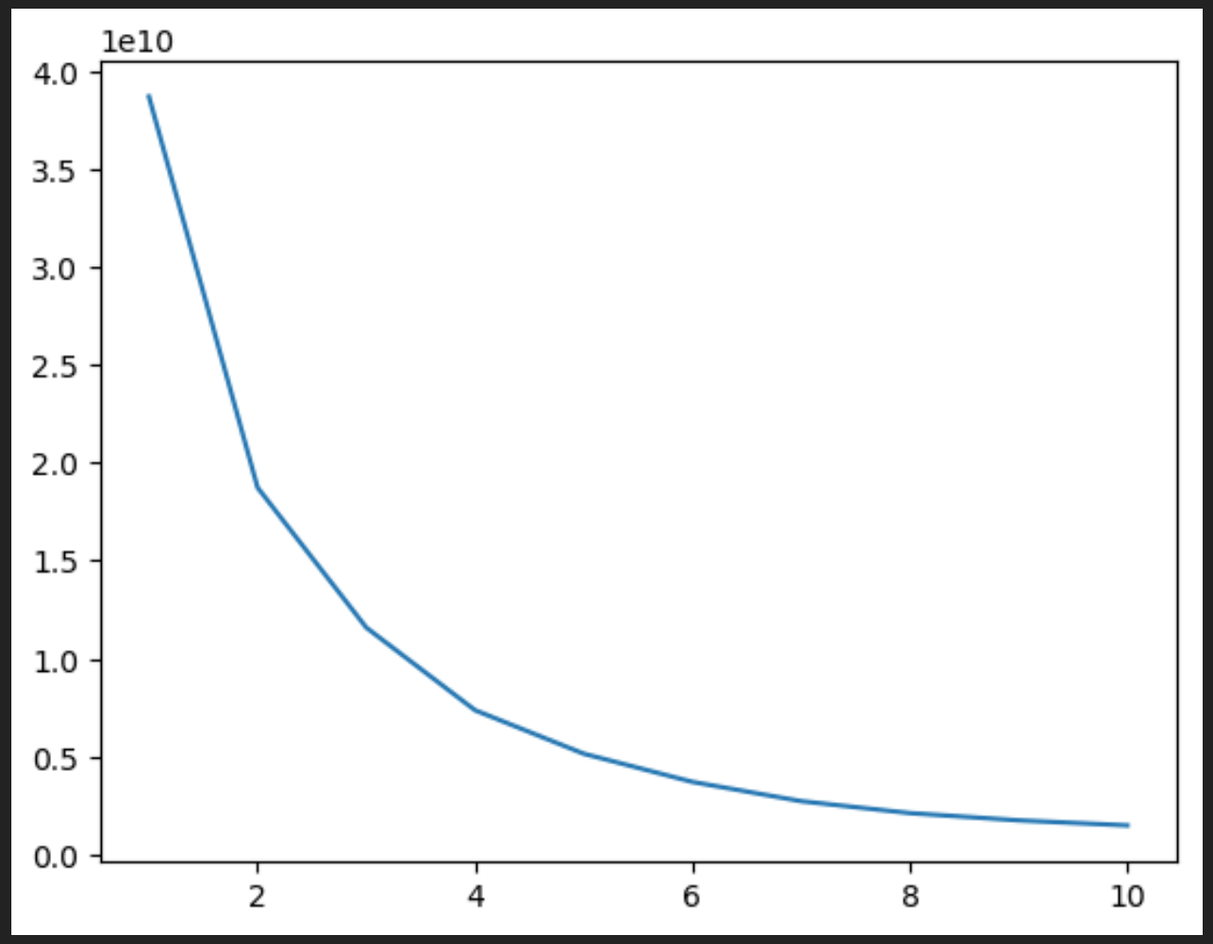
K Means

- K Means is an unsupervised model that groups the
features into clusters.

- To find the optimal number of clusters we draw the elbow
curve using the WCSS-Within Cluster Sum of Squares.

- WCSS is a statistical method to compute the difference in the square of distances between the centroid and the points in cluster.

- The goal is to have minimal number of clusters with minimal WCSS.



$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

Artificial Neural Network

- Comparing with the other models and baseline, ANN has performed well for Revenue-False(0), but has failed to beat the baseline in Revenue-True(1).
- This model was built with 8 hidden layers, with learning rate 0.0001 and optimiser to be Adam.
- All the parameters are used with hyper tuning, i.e., the parameters were tried with trial and error.
- Models with higher or lower hidden layers tend to fail with perfect precision for the Revenue-False(0) and no precision on Revenue-True(1)

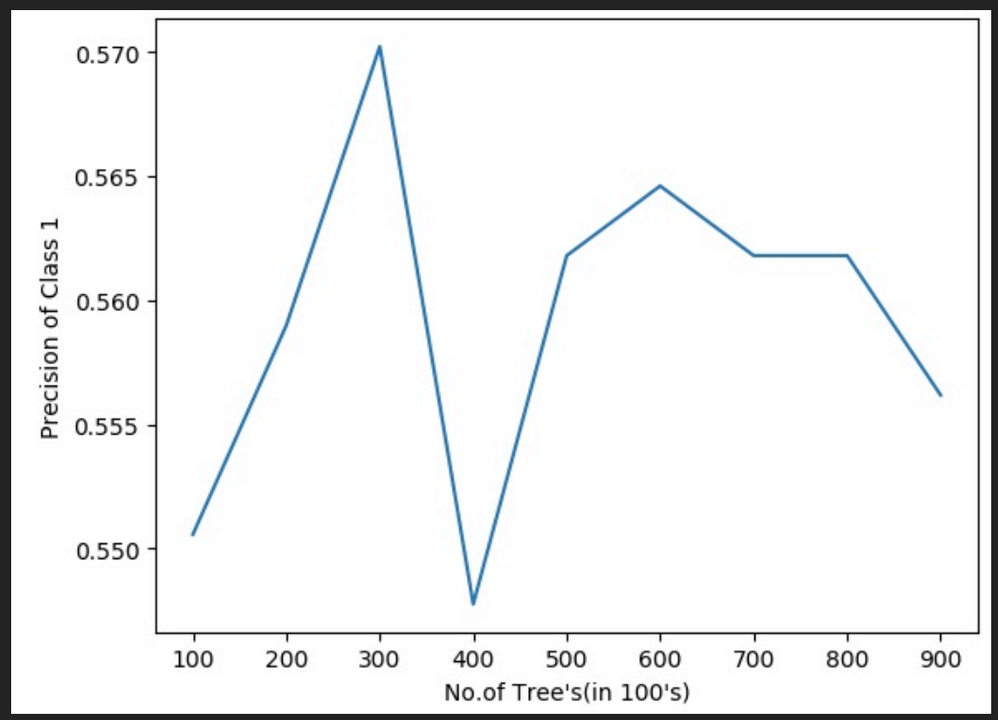
	precision	recall	f1-score	support
0	0.96	0.91	0.93	2239
1	0.41	0.65	0.51	225
accuracy			0.88	2464
macro avg	0.69	0.78	0.72	2464
weighted avg	0.91	0.88	0.89	2464

No of Trees	Revenue - False	Revenue - True
100	0.96	0.55
200	0.96	0.56
300	0.96	0.57
400	0.96	0.55
500	0.96	0.56
600	0.96	0.56
700	0.96	0.56
800	0.96	0.56
900	0.96	0.56
1000	0.96	0.56

	precision	recall	f1-score	support
0	0.96	0.93	0.94	2176
1	0.57	0.70	0.63	288
accuracy			0.90	2464
macro avg	0.76	0.82	0.79	2464
weighted avg	0.91	0.90	0.91	2464

Random Forest

- Till now the best model had been the DT and NB, where RF is and extension(ensemble) of many DTs.
- The tabular column shows that the Revenue-False(0) to be constant irrespective of the number of trees and the Revenue-True(1) tends to be increasing till 300 trees and then stays constant.
- So ultimately 300 trees seem to be optimal solution for this ensemble method.
- The precisions seems to have considerably increased with respect to the DT and NB.



Conclusion

- From the tabulation it is evidently seen that Random Forest is the best classification model for our dataset.
- So we decided to train our RF model with only the most important features which was “PAGE VALUES” and “EXIT RATES” and we found out the precision for the Revenue-True class decreased by .03
- Instead of having 17 features to actually predict the model, predicting merely the same accuracy with two features found to be more effective for the business.

Model	Revenue-False	Revenue-True
Baseline-Random	0.50	0.50
Naive Bayes	0.91	0.50
KNN	0.95	0.37
Decision Tree	0.91	0.54
ANN	0.96	0.41
Random Forest	0.93	0.58

	precision	recall	f1-score	support
0	0.93	0.93	0.93	2101
1	0.58	0.57	0.57	363
accuracy			0.88	2464
macro avg	0.75	0.75	0.75	2464
weighted avg	0.87	0.88	0.87	2464