IS517 - Methods of Data Science

Project : Online Shoppers Intention

Rohan, Asmitha, Sahit

Introduction

- E-commerce is a 5 trillion dollars industry.
- Customer tracking metrics are key for success of E Commerce business
- customer dynamics is of paramount importance for the firm, is widely used, and is less known to the general masses
- Example: Google Analytics, Heap, Hubspot

Dataset Description

- Rows : 12330 , Columns : 18
- The dataset consists of feature vectors belonging to 12,330 sessions.
- The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user,profile, or period.
- The dataset consists of 10 numerical and 8 categorical attributes.
- The 'Revenue' attribute used as the class label.

Research Question

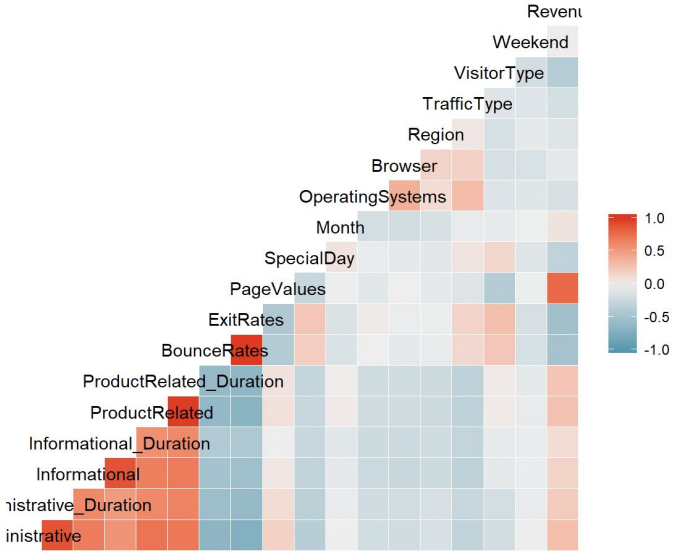
1) Given the information about an online shopping session can we predict if revenue will be generated in that session.

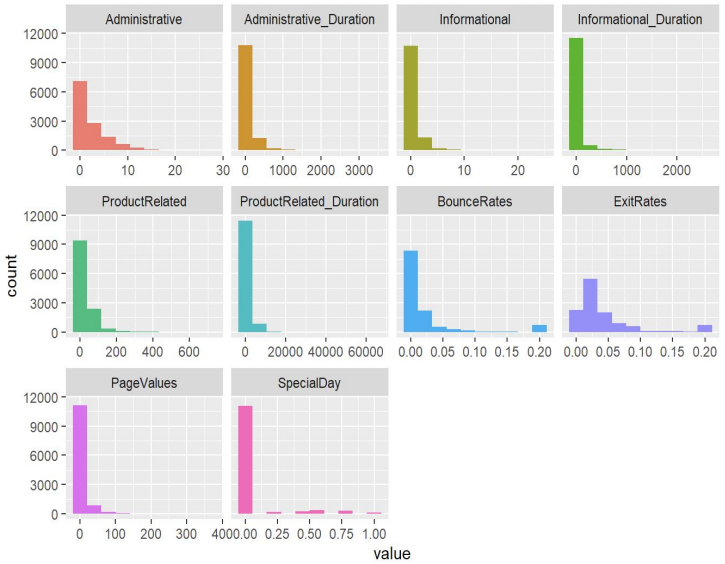
This will be a binary classification problem and we will attempt to employ different feature engineering techniques as well as numerous machine learning models to determine which approach provides the best classification results.

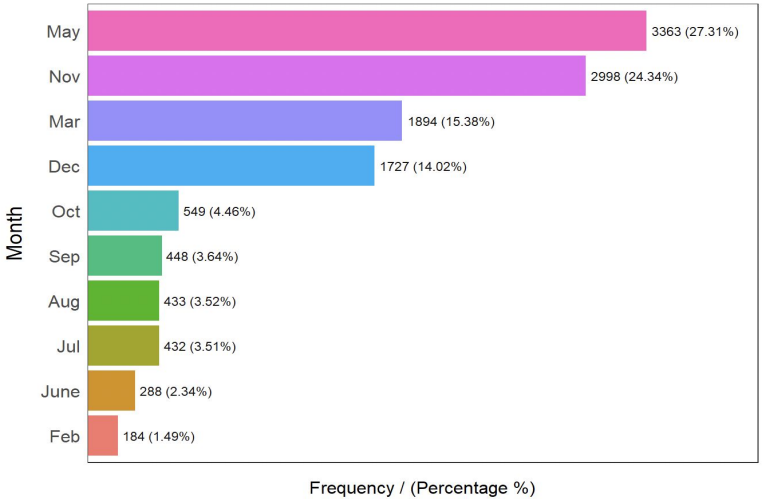
2) What are the important attributes that indicate the customer is most likely to purchase (create revenue) during an online session?

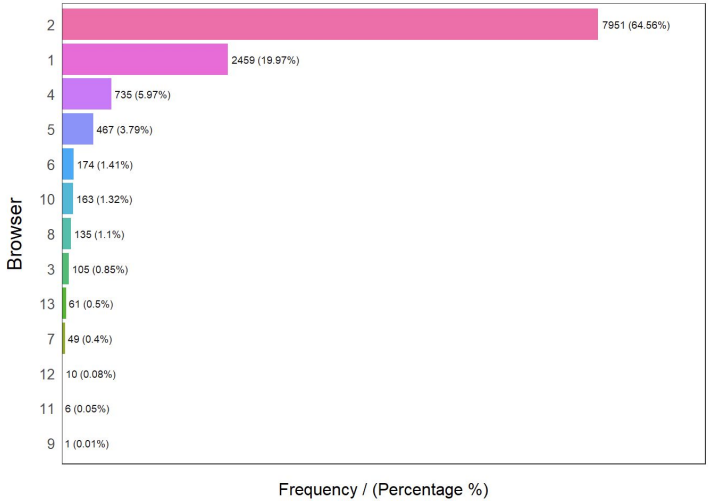
Any E-commerce firm would enjoy having highly accurate models that can predict if there is going to be a sale during a given session. But, at the same point, it is important for them to understand what factors are playing a role in generating revenue.

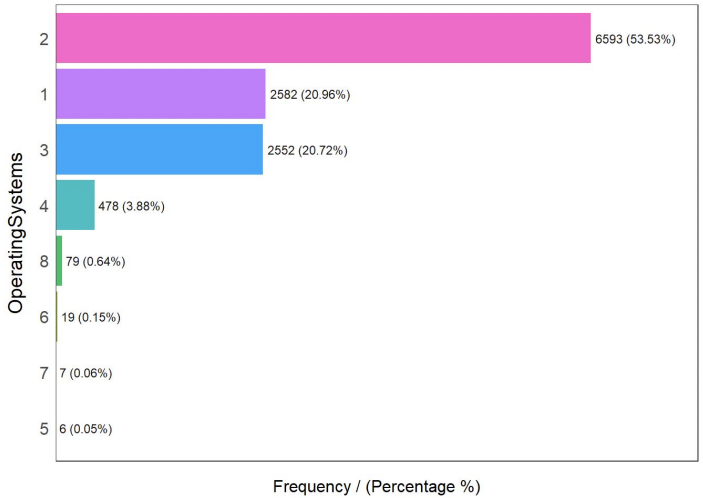
Exploratory Data Analysis

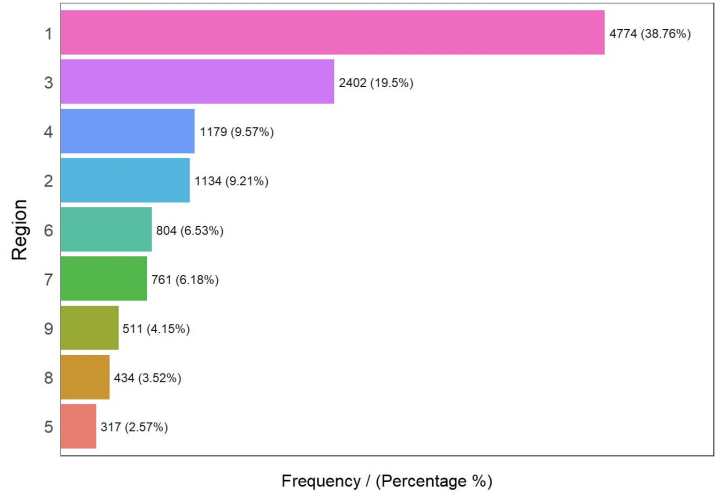


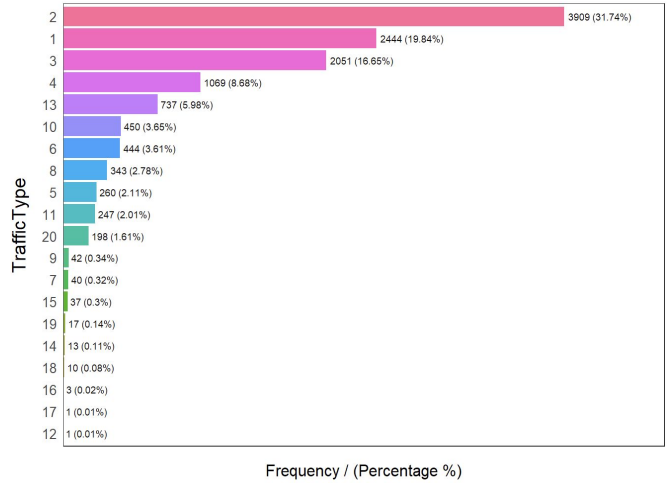


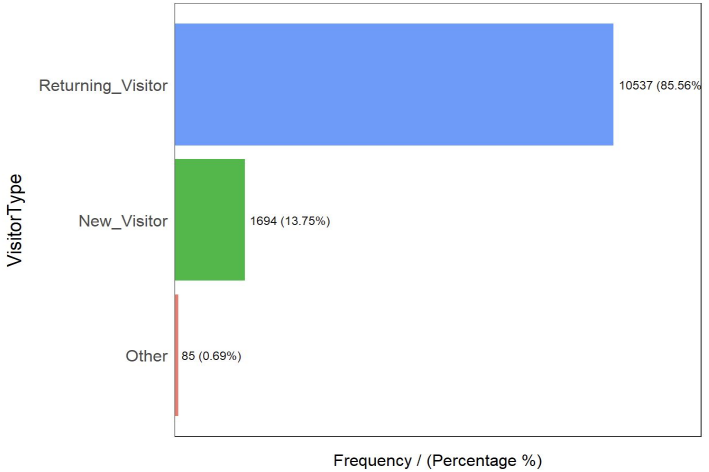

Month
May
Nov
Mar
Dec
Oct
Sep
Aug
Jul
June
Feb
3363 (27.31%)
2998 (24.34%)
1894 (15.38%)
1727 (14.02%)
549 (4.46%)
448 (3.64%)
433 (3.52%)
432 (3.51%)
288 (2.34%)
184 (1.49%)
Frequency / (Percentage %)

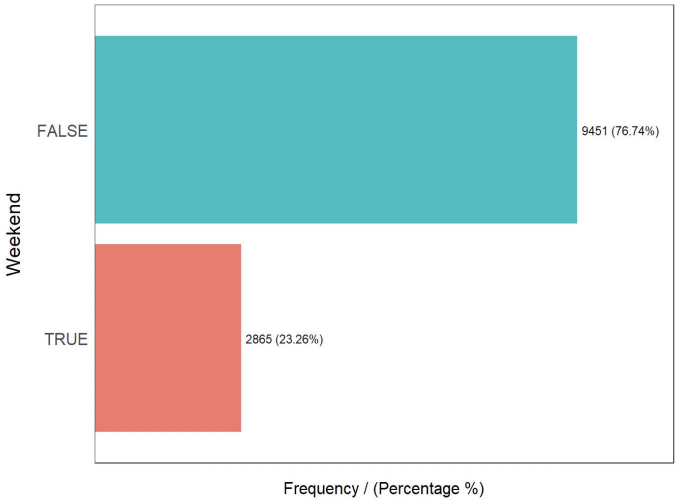

Browser
2
1
4
5
6
10
8
3
13
7
12
11
9
7951 (64.56%)
2459 (19.97%)
735 (5.97%)
467 (3.79%)
174 (1.41%)
163 (1.32%)
135 (1.1%)
105 (0.85%)
61 (0.5%)
49 (0.4%)
10 (0.08%)
6 (0.05%)
1 (0.01%)
Frequency / (Percentage %)

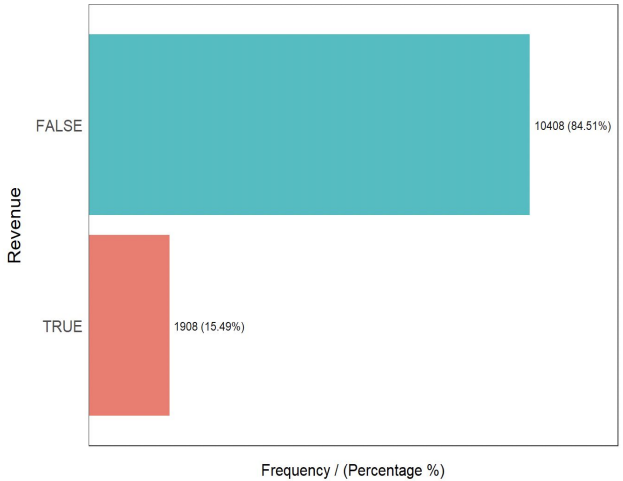

OperatingSystems
2
1
3
4
8
6
7
5
6593 (53.53%)
2582 (20.96%)
2552 (20.72%)
478 (3.88%)
79 (0.64%)
19 (0.15%)
7 (0.06%)
6 (0.05%)
Frequency / (Percentage %)


Region
1
3
4
2
6
7
9
8
5
4774 (38.76%)
2402 (19.5%)
1179 (9.57%)
1134 (9.21%)
804 (6.53%)
761 (6.18%)
511 (4.15%)
434 (3.52%)
317 (2.57%)
Frequency / (Percentage %)


TrafficType
2
1
3
4
13
10
6
8
5
11
20
9
7
15
19
14
18
16
17
12
3909 (31.74%)
2444 (19.84%)
2051 (16.65%)
1069 (8.68%)
737 (5.98%)
450 (3.65%)
444 (3.61%)
343 (2.78%)
260 (2.11%)
247 (2.01%)
198 (1.61%)
42 (0.34%)
40 (0.32%)
37 (0.3%)
17 (0.14%)
13 (0.11%)
10 (0.08%)
3 (0.02%)
1 (0.01%)
1 (0.01%)
Frequency / (Percentage %)


VisitorType
Returning_Visitor
New_Visitor
Other
10537 (85.56%)
1694 (13.75%)
85 (0.69%)
Frequency / (Percentage %)


Weekend
FALSE
TRUE
9451 (76.74%)
2865 (23.26%)
Frequency / (Percentage %)

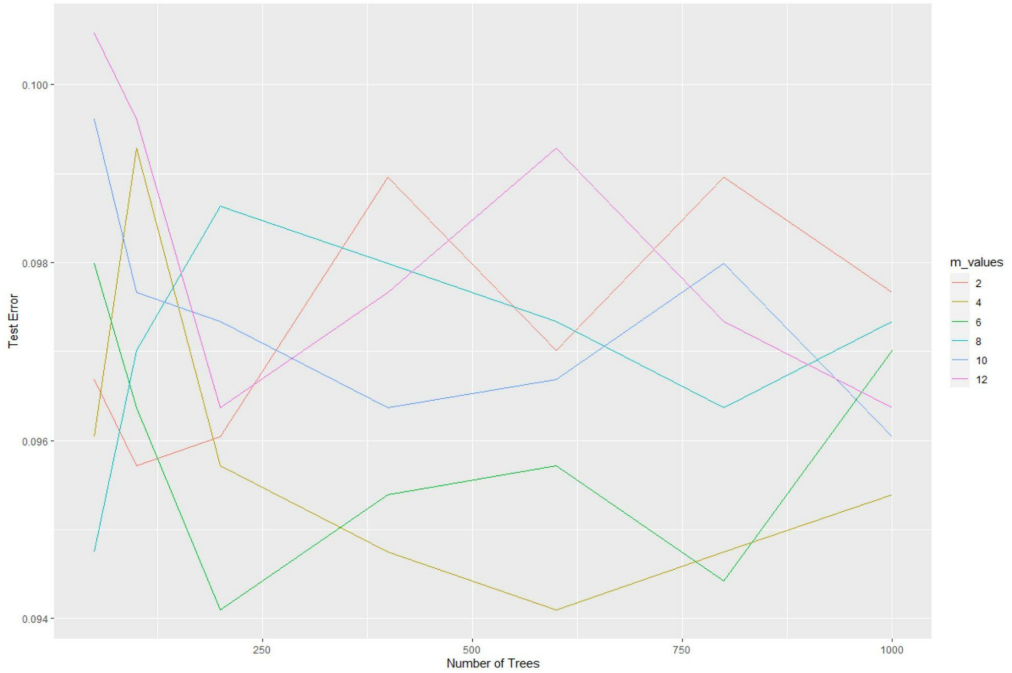

Revenue
FALSE
TRUE
10408 (84.51%)
1908 (15.49%)
Frequency / (Percentage %)

RQ 1

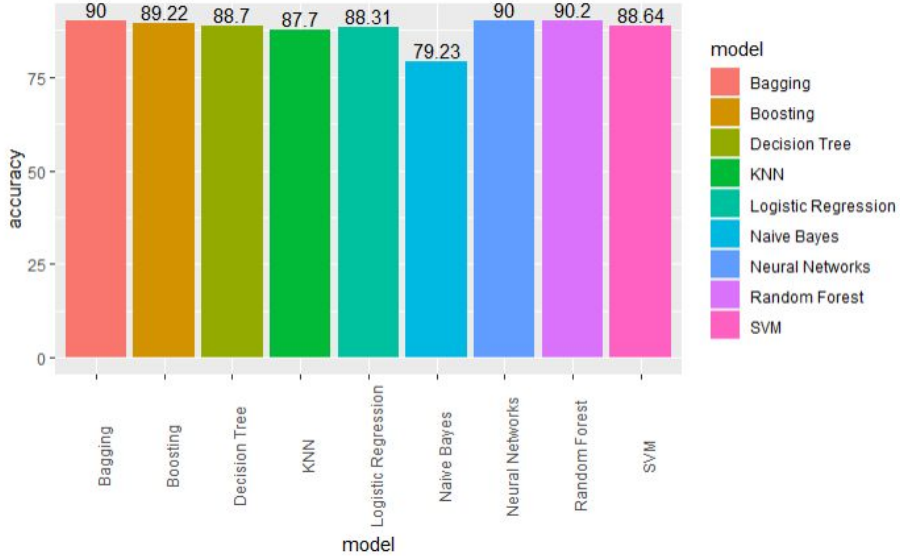
Given the information about an online shopping session can we predict if revenue will be generated in that session.

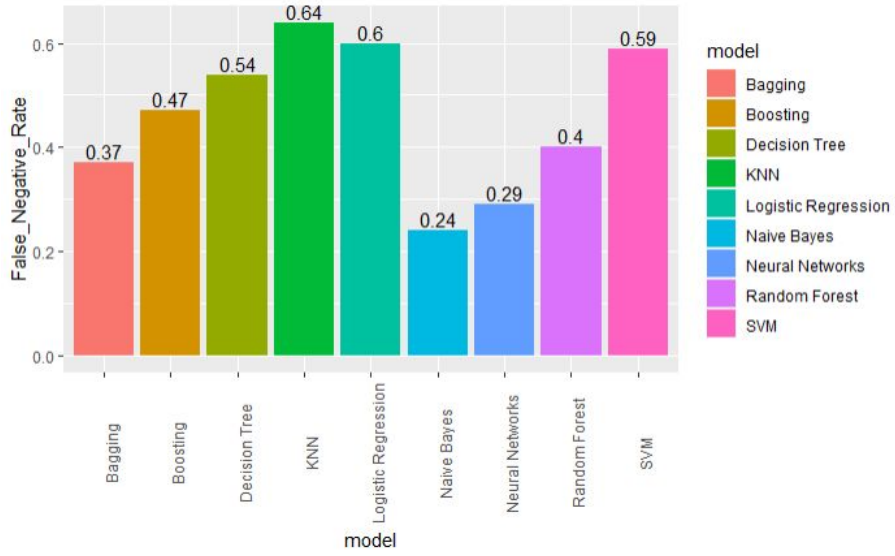
- Logistic Regression
- KNN
- Naive Bayes Classifier
- Decision Tree
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Networks

Fine Tuning


0.100
0.098
0.096
0.094
Test Error
250
500
750
1000
Number of Trees
m_values
2
4
6
8
10
12

Model vs Accuracy vs False Negative Rate





RQ 2:

What are the important attributes that indicate the customer is most likely to purchase (create revenue) during an online session?

- Logistic Regression
- Bagging

RQ 2: Logistic Regression

Summary uses Wald-test.

Also tells us about how each feature affects odds.

ANOVA tests the explanatory power of the predictor.

Important predictors are :

1. PageValues
2. BounceRates
3. ProductRelated_Duration
4. ExitRates
5. Month
6. TrafficType
7. SpecialDay

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.720e+00	2.442e-01	-7.044	1.86e-12	***
Informational	3.926e-02	3.157e-02	1.244	0.213666	
Informational_Duration	-7.799e-05	2.658e-04	-0.293	0.769232	
ProductRelated_Duration	1.053e-04	1.798e-05	5.857	4.72e-09	***
BounceRates	-1.322e+00	3.596e+00	-0.368	0.713167	
ExitRates	-1.439e+01	2.746e+00	-5.241	1.60e-07	***
PageValues	8.209e-02	2.825e-03	29.057	< 2e-16	***
SpecialDay	-8.285e-02	2.776e-01	-0.298	0.765374	
MonthDec	-8.770e-01	2.136e-01	-4.107	4.02e-05	***
MonthFeb	-2.825e+00	1.073e+00	-2.634	0.008441	**
MonthJul	5.302e-02	2.505e-01	0.212	0.832361	
MonthJune	-4.430e-01	3.147e-01	-1.408	0.159257	
MonthMar	-7.562e-01	2.109e-01	-3.586	0.000336	***
MonthMay	-6.728e-01	1.997e-01	-3.369	0.000755	***
MonthNov	3.932e-01	1.893e-01	2.077	0.037824	*
MonthOct	-1.450e-01	2.318e-01	-0.626	0.531540	
MonthSep	4.652e-02	2.406e-01	0.193	0.846705	
OperatingSystems2	2.806e-01	1.901e-01	1.476	0.139870	
OperatingSystems3	5.132e-02	2.045e-01	0.251	0.801829	
OperatingSystems4	1.083e-01	2.096e-01	0.516	0.605576	
OperatingSystems5	3.660e-01	1.363e+00	0.269	0.788237	
OperatingSystems6	-5.273e-01	1.047e+00	-0.504	0.614538	
OperatingSystems7	1.349e+00	1.168e+00	1.155	0.248022	
OperatingSystems8	-2.784e-01	9.326e-01	-0.298	0.765352	
Browser2	-1.805e-01	1.918e-01	-0.941	0.346666	
Browser3	-1.523e+00	7.676e-01	-1.985	0.047172	*
Browser4	-1.216e-01	2.406e-01	-0.506	0.613161	
Browser5	1.303e-02	2.620e-01	0.050	0.960352	
Browser6	-6.026e-01	4.025e-01	-1.497	0.134340	
Browser7	-1.299e-01	5.660e-01	-0.229	0.818529	
Browser8	3.208e-01	3.560e-01	0.901	0.367544	
Browser9	-1.210e+01	1.455e+03	-0.008	0.993367	

: Model Summary

Analysis of Deviance Table

Model: binomial, link: logit

Response: Revenue

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			9247	7968.8		
Informational	1	72.70	9246	7896.1	< 2.2e-16	***
Informational_Duration	1	0.39	9245	7895.7	0.5343	
ProductRelated_Duration	1	146.01	9244	7749.7	< 2.2e-16	***
BounceRates	1	322.05	9243	7427.6	< 2.2e-16	***
ExitRates	1	233.74	9242	7193.9	< 2.2e-16	***
PageValues	1	1565.36	9241	5628.5	< 2.2e-16	***
SpecialDay	1	17.47	9240	5611.0	2.925e-05	***
Month	9	215.80	9231	5395.2	< 2.2e-16	***
OperatingSystems	7	10.77	9224	5384.5	0.1489	
Browser	11	15.76	9213	5368.7	0.1504	
Region	8	4.62	9205	5364.1	0.7969	
TrafficType	19	53.06	9186	5311.0	4.586e-05	***
VisitorType	2	3.71	9184	5307.3	0.1567	
Weekend	1	0.15	9183	5307.2	0.6981	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

: Anova

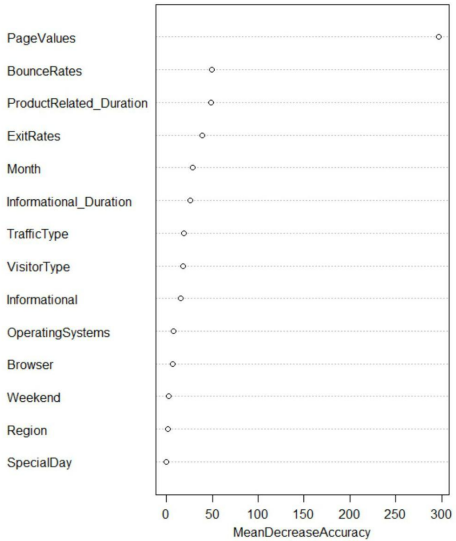
RQ 2: Bagging

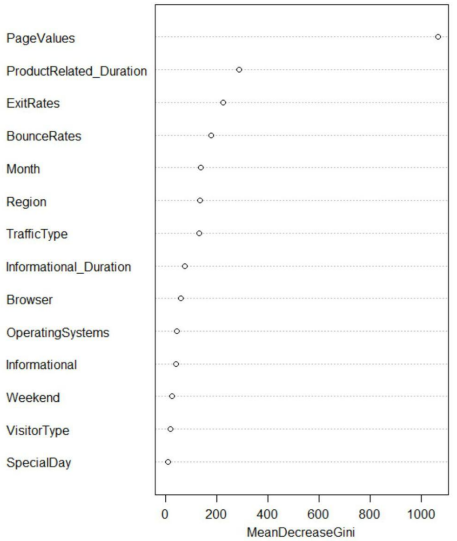
Two measures MeanDecreaseAccuracy and MeanDecreaseGini

Quantify Impact of the predictors

Top 5 predictors:

1. PageValues
2. BounceRates
3. ProductRelated_Duration
4. ExitRates
5. Month





Conclusion & Future Direction

Conclusion

1. Non-Linear Models seem to perform better. Fine tuning increased the accuracy significantly for some models.
2. Model selection should not be done solely on accuracy.
3. Firms can focus research more into the important predictors to better understand the consumer purchase dynamics.

Future Work

1. Months and special days impact revenue. Thus using different temporal aspect can be interesting.
2. Trying similar analysis on different dataset as a regression problem may lead to new insights.
3. Larger Datasets and various other Ensemble/ Neural Network techniques can be tried for higher accuracy.

Thank You