0.2 Introduction

The analysis of customer purchase behavior dates back to the beginning of e-commerce. However, in today's world with the advancement in technology it can be utilized in a more profitable way than ever before. It is being used for intent marketing, personalized advertising, prioritize marketing budget to meet sales goals and create future value. Information gained by analyzing customer behavior can be used to target buyers with content they need at the right time in their journey. To predict which prospects are ready to make their purchase, the model evaluates non-transction customer data, such as how much time customer spent on administrative or product related pages or how the customer interacts with the website. The models also takes into account some kind of demographic data like visitor type. The model compares the pre-purchase behavior of the cutomers to the pre-purchase behaviors of millions of previous customers who endend up buying, comparing attributes like what pages they spent time, was the visit close to holidays or weekend. The model tags the customer that behaved the most like the previous buyer is predicted as the customer who is likely to buy. Once the companies have this information they can prioritze their investment on each individual prospective customer. For instance,it can be used to attract customers who are not likely to buy by offering discounts at the write time before they exit the site.

0.2.1 Primary Analysis Objectives

The preliminary goal of the study is to analyze the variables influencing the mindset of the coustomer to predict their purchase intent.

0.2.2 Secondary Analysis Objectives

In businees world, User purchase intent research is very useful. It helps understand what triggers at what point in the visit to the website customers make decision to purchase. Is it the information in the adminstration page, product page or the seasonal offers or the weekend? Once we know the factors that trigger the customer buys, companies can tailor thier messaging across all of their marketing assets to promote activities that are likely to create future value. If we know what kind of visits are going to generate revenue for the company than those kind of visits can be leveraged to meet the sales goals.Hence this study is focused on analyzing the pattern in customer behavior and predicting the intent of the online shoppers.

0.3 Materials and Methods

0.3.1 Data Sources

The data was obtained from the url. The dataset consists of 10 numerical and 8 categorical attributes.The binary categorical variable Revenue is the response.The dataset includes the following variables :

Table 1: Variables in the dataset

Sl.No	Variable Name	Info
1	Administrative	Administrative Value
2	Administrative_Duaration	Duration in Administrative Page
3	Informational	Informational Value
4	Information_Duaration	Duration in Informational page
5	ProductRelated	Product related value
6	ProductRelated_Duration	Duration in product related page
7	BounceRates	Bounce rates of web page
8	ExitRates	Exit rate of a web page
9	PageValues	Page values of each web page
10	SpecialDay	Closeness of site visiting time to a specific special day
11	Month	Month of the year
12	OperatingSystems	Operating system used
13	Browser	Browser used
14	Region	Region of the user
15	TrafficType	Traffic type
16	VisitorType	Types of visitor
17	Weekend	Weekend or not
18	Revevue	Revenue will be generated or not

0.3.2 Statistical Analysis

The data obtained is in csv format.The data analysis is done using the statistical software R version 3.5.2 (2018-12-20) and the study uses binary logistic regression.The data is split into 80% training and 20% test set.An initial logistic regression built with all variables and training set is given by
$logit[P(Y = 1)] = \alpha + \beta_{1Admin} + \beta_2 AdminDuration + ... + \beta_{63WeekendTrue}$.Variable ProductedRelated dropped due to multicollinearity and all other variables are used in stepwise selection to arrive at the final model.The variable Pagevalue is dropped due to complete seperation of data.Multicollinearity, correlation between explanatory variables and stepwise variable selection method is adopted to arrive at the final model. The test set is used in evaluating the performance meterics.

0.3.3 Model Assumptions

All inferences are conducted using $\alpha = 0.05$ unless stated otherwise. The model was built with the train set containing 80% of observations, tested with 20% of the data.No adjustments for multiplicity are made as this is an exploratory analysis. Discrete variables are summarized with proportions and frequencies. The continuous variables are summarized using mean, median, standard deviations, quantiles, variance, maximum and minimum

0.3.4 Preliminary Objective Analysis

Exploratory Analysis: The data set is explored to observe the distribution of the numerical and categorical variables. The tables below shows the 5-summary statistics for the numerical

variables:

Table 2: 5-summary statistics

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
	Min. : 0.0	Min. : -1	Min. : 0.0	Min. : -1	Min. : 0	Min. : -1	Min. :0.0	Min. :0.0	Min. : 0
	1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 7	1st Qu.: 185	1st Qu.:0.0	1st Qu.:0.0	1st Qu.: 0
	Median : 1.0	Median : 8	Median : 0.0	Median : 0	Median : 18	Median : 600	Median :0.0	Median :0.0	Median : 0
	Mean : 2.3	Mean : 81	Mean : 0.5	Mean : 35	Mean : 32	Mean : 1196	Mean :0.0	Mean :0.0	Mean : 6
	3rd Qu.: 4.0	3rd Qu.: 94	3rd Qu.: 0.0	3rd Qu.: 0	3rd Qu.: 38	3rd Qu.: 1466	3rd Qu.:0.0	3rd Qu.:0.0	3rd Qu.: 0
	Max. :27.0	Max. :3399	Max. :24.0	Max. :2549	Max. :705	Max. :63974	Max. :0.2	Max. :0.2	Max. :362
	NA's :14	NA's :14	NA's :14	NA's :14	NA's :14	NA's :14	NA's :14	NA's :14	NA

Only 16% of the visits generated revenue.The response revenue has unbalanced class.This is shown in Figure 1 below.

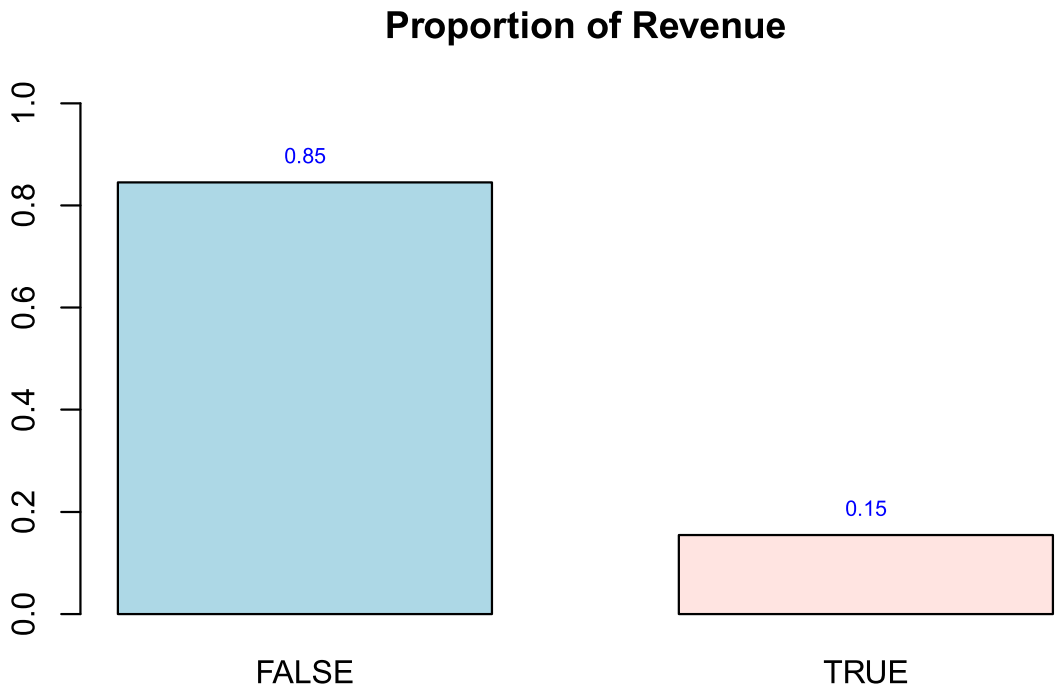

Figure 1: Barplot of Revenue

Data Cleaning

There are 112 NA and -1 values for Admininstarion_duration, Informational_duration and ProductRelated_duration. The rows containing NA's and -1 values are omitted.The Table 3. shows NA's in different columns.

Table 3: Table of variables with NA

	x
Administrative	14
Administrative_Duration	14
Informational	14
Informational_Duration	14
ProductRelated	14
ProductRelated_Duration	14
BounceRates	14
ExitRates	14
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

Univariate Analysis

The distribution of all the numerical explanatory variables are seen in Figure 2. All the explanatory variables are skewed to the right.

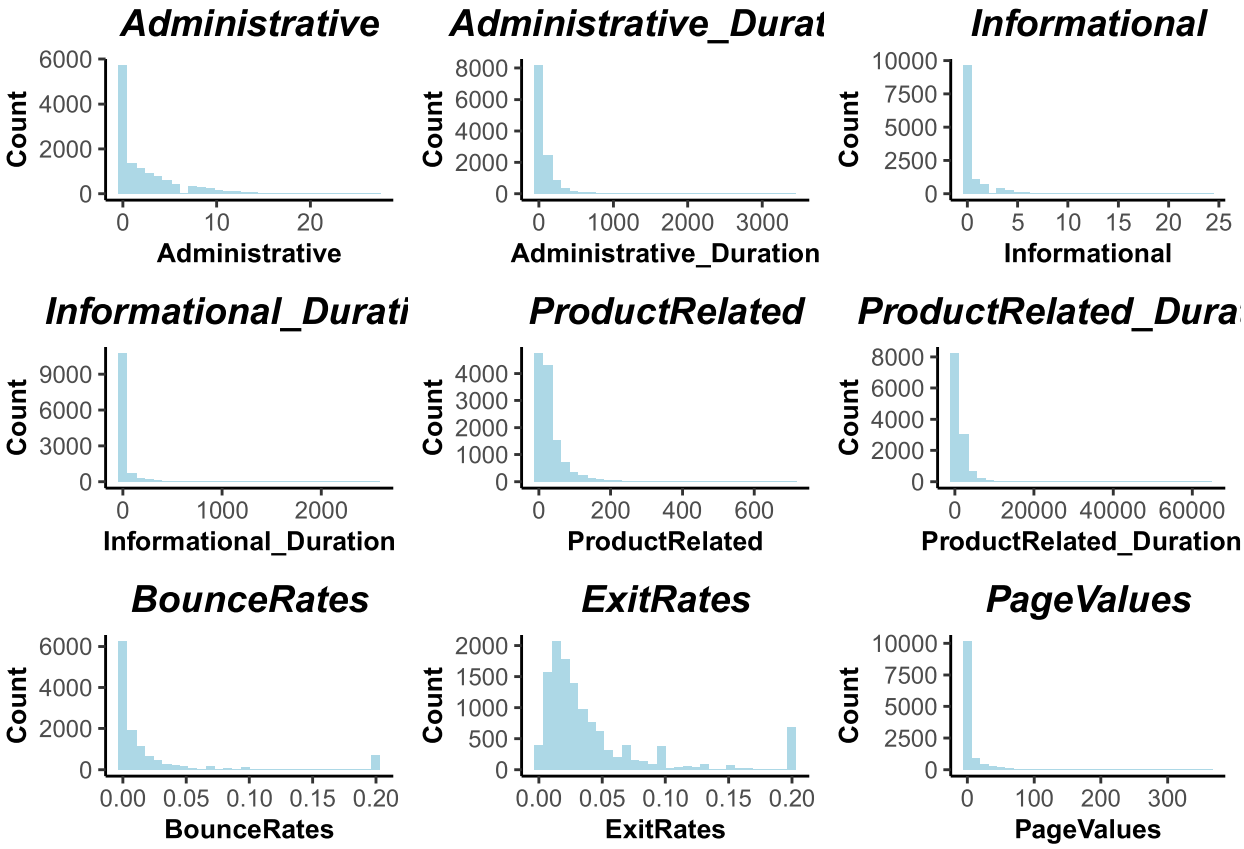

Figure 2: Histogram of numerical variables

The customers visited the website 90% more just before and after the special day.The percenteage of visits were higher for May(27%), Nov(24%) and Dec(14%) than the other months.Most of the visitors to the site used operating system 2, that is about 54% and browser 2 about 57%.People from region1(39%) and region3(19%) visited the site most.The most common traffic type was traffic2(32%) and traffic(17%).85% of the visitors were returning visitors. Most of the people visited the site during weekend(77%).The barplots in Figure 3 and the Table in Figure 4 shows the proportions of categories.Meera Govindan (2002)

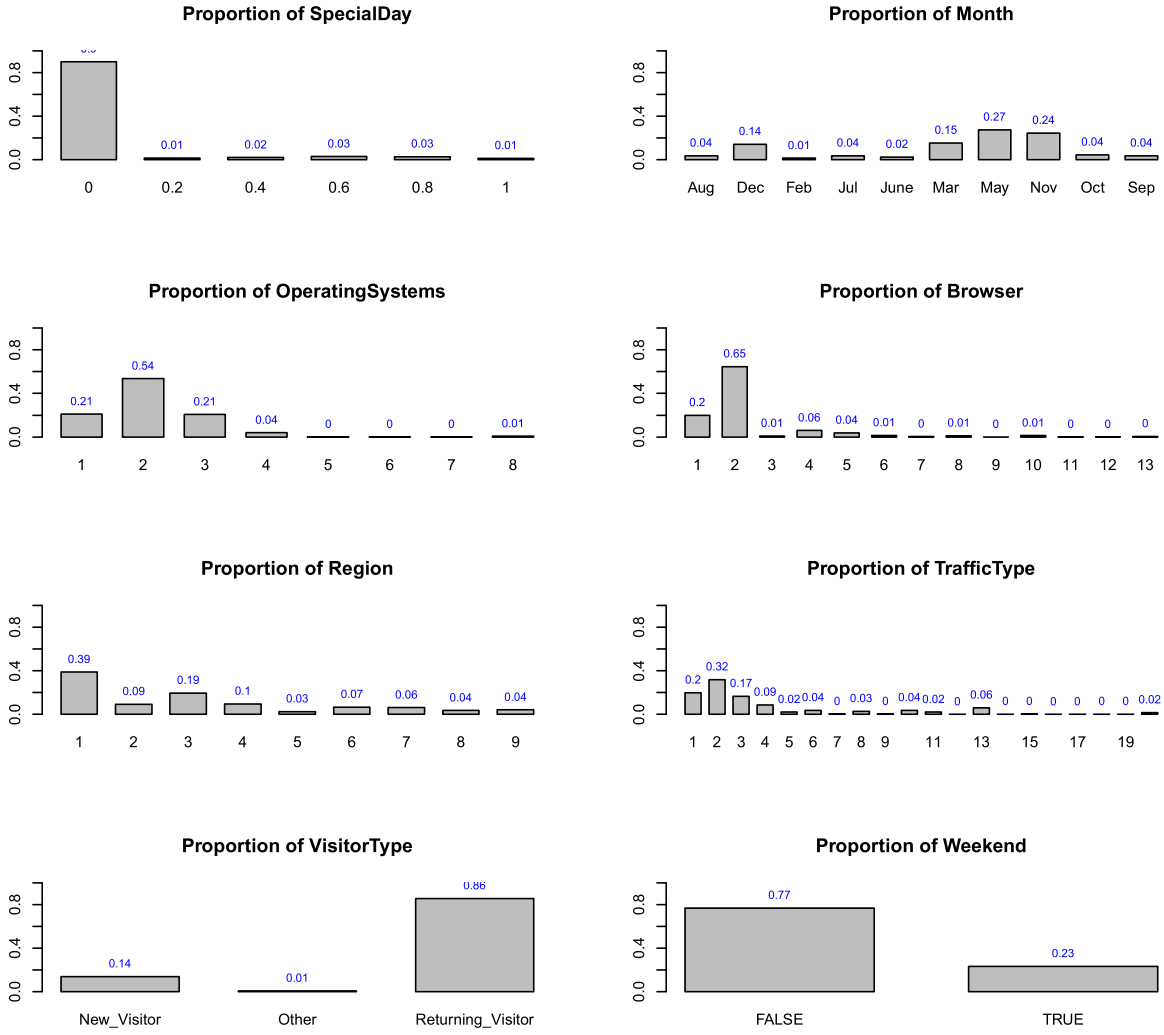

Figure 3: Barplots showing proportion for all categorical variables

Operating System	Proportion
1	0.20948
2	0.53554
3	0.20720
4	0.03875
5	0.00049
6	0.00155
7	0.00057
8	0.00643

SpecialDay	Proportion
0	0.898
0.2	0.014
0.4	0.020
0.6	0.028
0.8	0.026
1	0.013

Region	Proportion
1	0.388
2	0.092
3	0.195
4	0.096
5	0.026
6	0.065
7	0.062
8	0.035
9	0.041

Month	Proportion
Aug	0.035
Dec	0.141
Feb	0.014
Jul	0.035
June	0.023
Mar	0.153
May	0.273
Nov	0.244
Oct	0.045
Sep	0.036

ser	Proportion
	2.0e−01
	6.5e−01
	8.5e−03
	6.0e−02
	3.8e−02
	1.4e−02
	4.0e−03
	1.1e−02
	8.1e−05
)	1.3e−02
	4.9e−04
2	8.1e−04
3	5.0e−03

VisitorType	Proportion
New_Visitor	0.1379
Other	0.0069
Returning_Visitor	0.8552

Weekend	Proportion
FALSE	0.77
TRUE	0.23

	TrafficType	Proportion
47	1	0.1982
48	2	0.3181
49	3	0.1658
50	4	0.0869
51	5	0.0211
52	6	0.0360
53	7	0.0033
54	8	0.0279
55	9	0.0034
56	10	0.0366
57	11	0.0201
	TrafficType	**Proportion**
58	12	8.1e−05
59	13	6.0e−02
60	14	1.1e−03
61	15	3.0e−03
62	16	2.4e−04
63	17	8.1e−05
64	18	8.1e−04
65	19	1.4e−03
66	20	1.6e−02

Figure 4: Table of Proportion

Bivariate Analysis:

In frequency plot in Figure 5, we can see the frequency of buys in numerical variables. The proportion of people who bought is very less for all the variables except for the variable Pagevalues, which is the average value for a web page that a user visited before exiting.The percentage of buys in categorical variables is given in Figure 6. The percentage of buys is the highest(96%) just before and after the special day(special day=0). In the month of november the percentage of buys is the highest 39.8% compared to the rest of the year.Customers that browsed with operating system 2 are the ones that bought most of the time(60.5%).Among the browsers used to navigate the site browser 2 amounts for the largest amount of purchase (64.1%). Region 1 accounts to 40.4% of the purchase. Traffic Type 2 accounts for 44.4% of the purchase. Returning visitor accounts for the largest 77% of the purchase.More purchases are made during weekdays about 73.8% than weekends(26.2%).

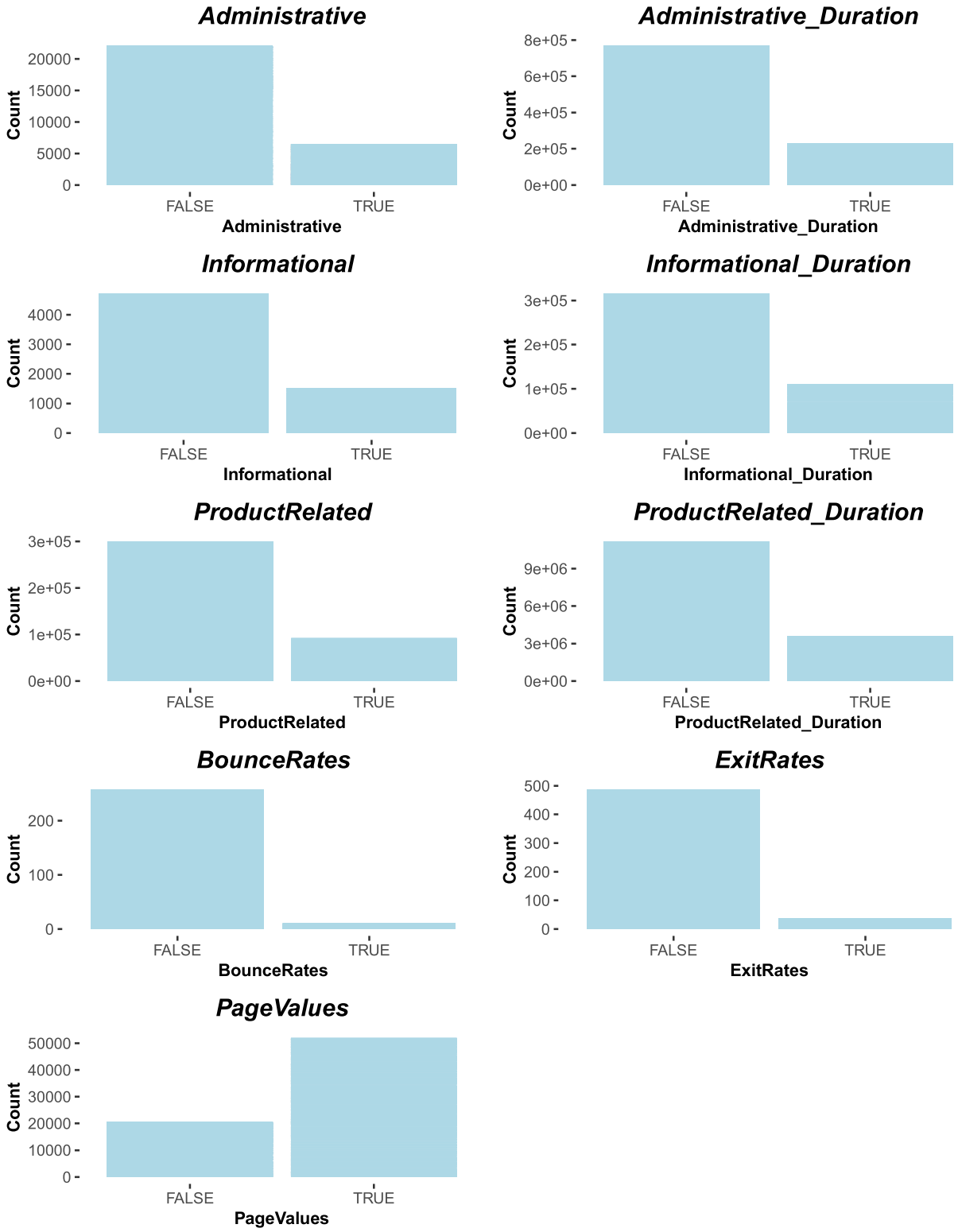

Figure 5: Frequency Plot of Revenue for numerical variables

Figure 6: Percentage of buys in each category

Screening of Covariates:

Data are unbalanced on the response variable,y = 1 are relatively few compared to y=0. This limits the number of explanatory variables p for which effects can be estimated precisely. Guidelines in Agresti (December 3, 2012) suggests that the data set should contain at least 10 outcomes of each type for every explanatory variable. explanatory variables have atleast 10 outcomes, they satisy the guideline.

Correlation The correlation plot in Figure 7 shows high correlation for variables (ProductRelated, ProductedRelated_Duration) and (BounceRates,ExitRates), since they have similar kind of information.We could drop one of the variables in the pair.

Assumptions of the model:

First, binary logistic regression requires the dependent variable to be binary.Figure 1. Second, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.The observations in the data are not from repeated measurements. Third, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.In Figure 7 we can see the corelation between variables and the Table 3 we see that the variables ProductRelated and ProductRelated_Duration have vif greater than 5. These variables are dropped in the final model to meet the assumption

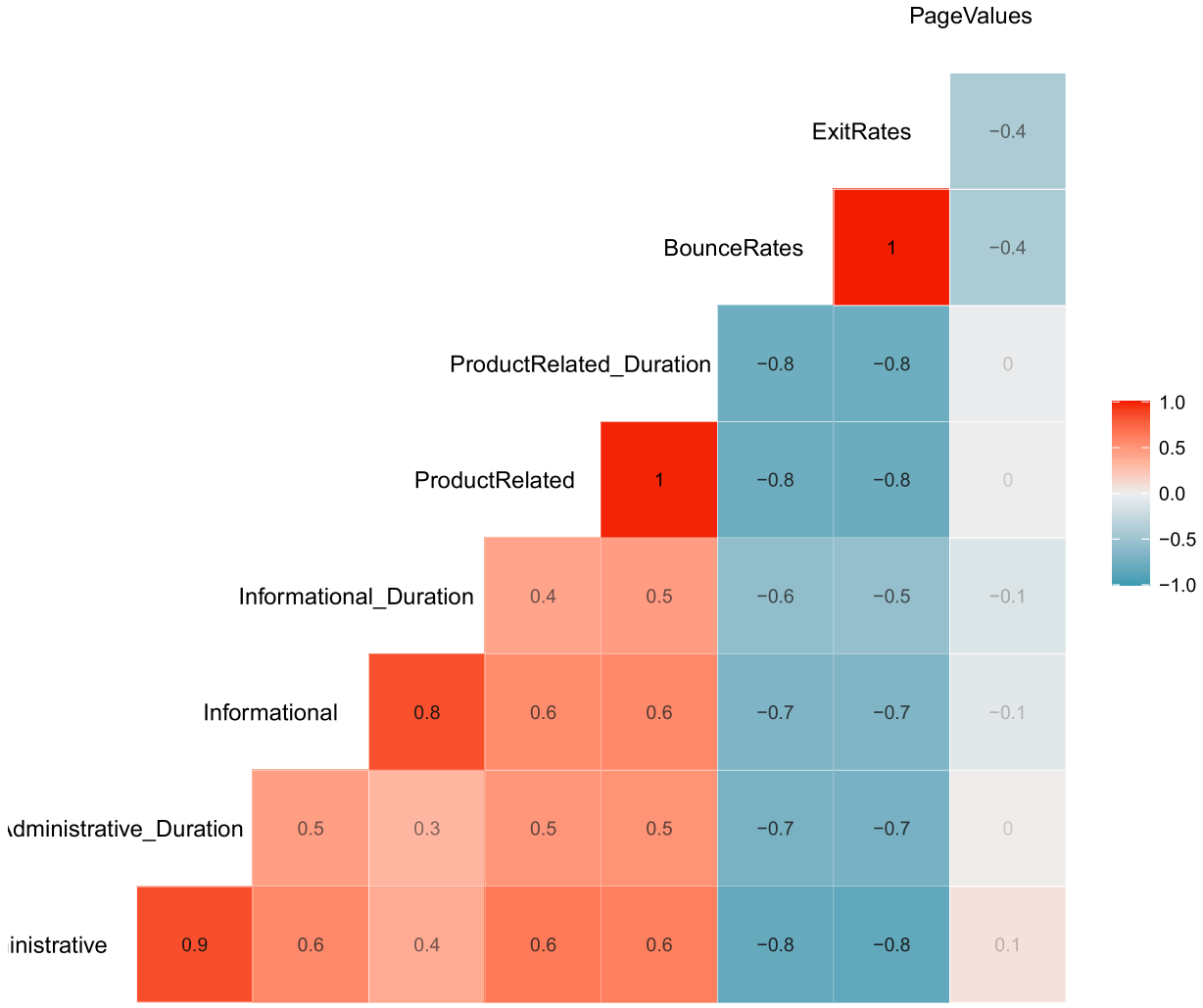

Figure 7: Correlation Plot

Multicollinearity:

The variables ProductRelated and Producted Related_Duaration are higly correlated as shown in the correlation plot in Figure 7, Hence they have vif factor more than 5. Since they both convey the same information, one of them ProductRelated will be dropped.

Table 4: Table of VIF factors

	GVIF	Df	GVIF^(1/(2*Df))
Administrative	2	1	1
Administrative_Duration	2	1	1
Informational	2	1	1
Informational_Duration	2	1	1
ProductRelated	6	1	2
ProductRelated_Duration	6	1	2
BounceRates	2	1	1
ExitRates	2	1	1
PageValues	1	1	1
SpecialDay	2	5	1
Month	3	9	1
OperatingSystems	1	1	1
Browser	34	11	1
Region	1	8	1
TrafficType	4	17	1
VisitorType	24	2	2
Weekend	1	1	1

Stepwise selection: The variables selected by both the stepwise forward and backward method are same.They are in Table 6

Table 5: Variables selected by stepwise method

	x
(Intercept)	-1.37
Administrative	-0.04
Administrative_Duration	0.00
ProductRelated_Duration	0.00
BounceRates	-16.10
ExitRates	-10.14
PageValues	0.09
MonthDec	-1.67
MonthFeb	-14.04
MonthJul	-0.03
MonthJune	-0.30
MonthMar	-0.52
MonthMay	-0.89
MonthNov	0.24
MonthOct	0.18
MonthSep	0.17
VisitorTypeOther	-1.82
VisitorTypeReturning_Visitor	-0.45
WeekendTRUE	0.30

Influential Points

The plot of cook's distance in Figure 8 shows that observation 6166, 1573 and 1210 have the largest distance.The standardized residuals plot in Figure 9 is observed to verify if any of the observations are above or below three standard deviation. In this study no observations are found to be above or below three standard deviation.Zhang (2016)

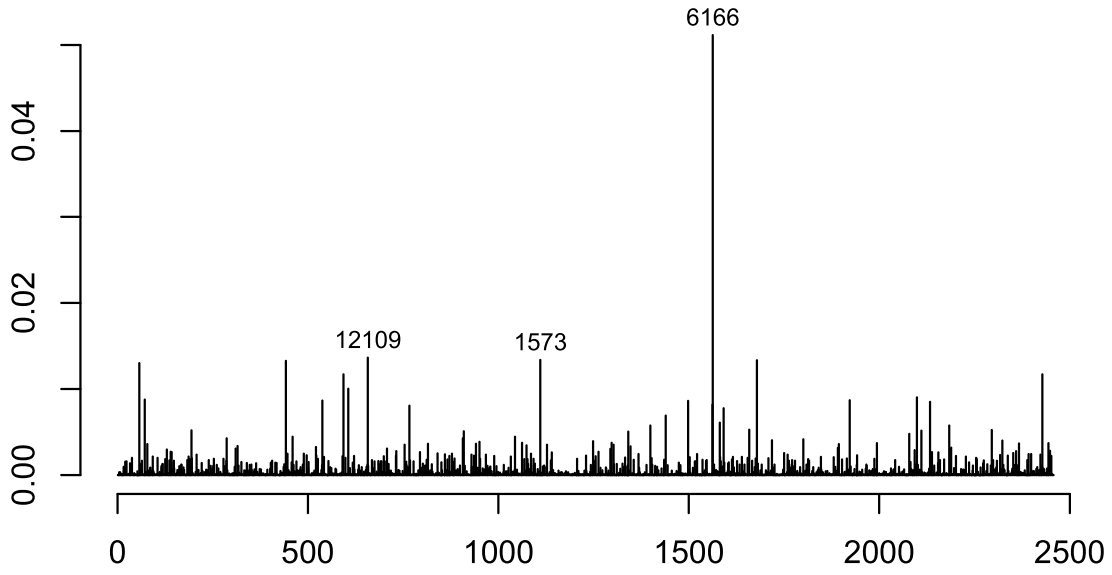

Figure 8: Plot of cooks distance

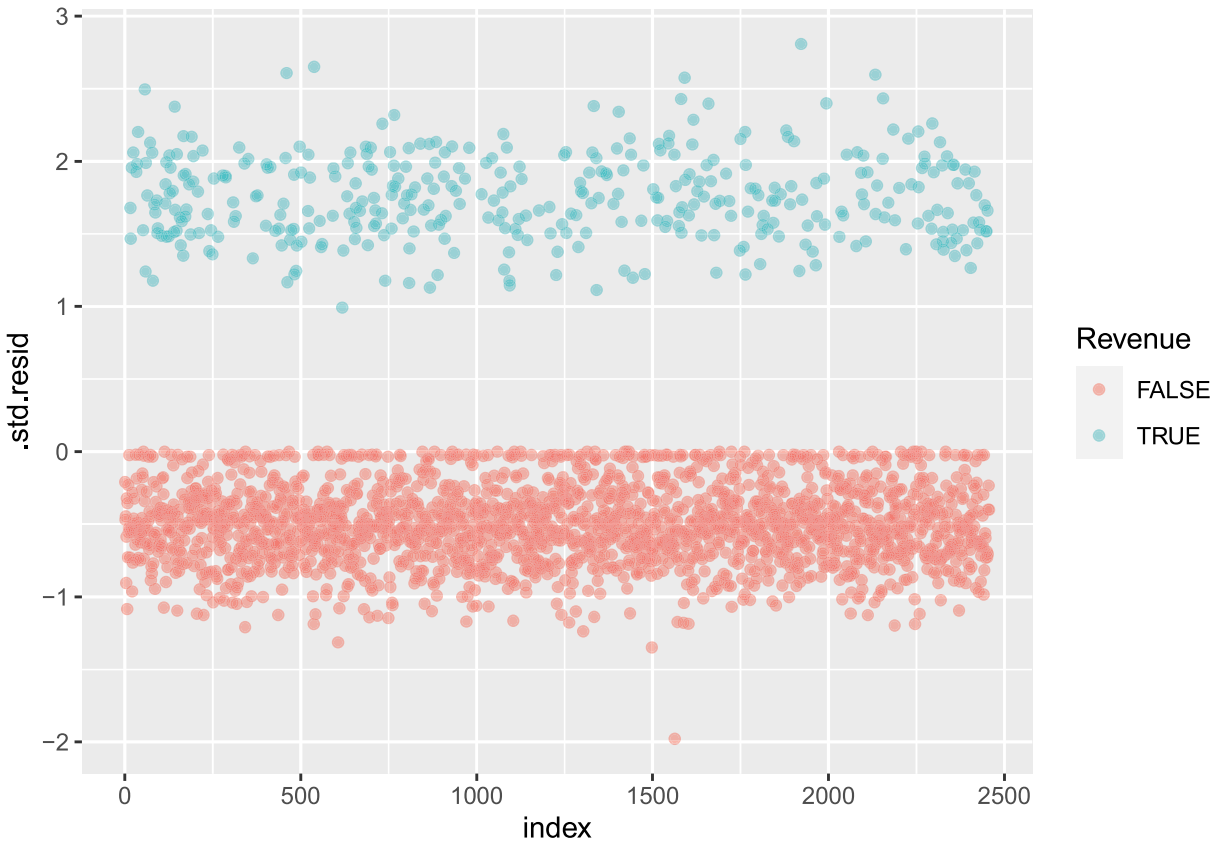

Figure 9: Standardized Residuals

0.3.4.1 Goodness of Fit Test

The likelihood ratio statistic comparing the model deviances for the full model and model obtained from stepwise selection is 1329.3 − `rL1` = -521, with a df = 45.The evidence of effect is not strong (P=1) for th more complex model.The test suggests that simpler model is adequate. Table 6 shows results of the wald test for the variables. T

Table 6: Wald test results for the variables

	Std.Error	z value	Pr(>|z|)
(Intercept)	3e-01	-1.25	2e-01
Administrative	2e-02	-1.88	6e-02
Administrative_Duration	4e-04	1.83	7e-02
ProductRelated_Duration	0e+00	1.87	6e-02
BounceRates	8e+00	-0.89	4e-01
ExitRates	5e+00	-5.74	0e+00
MonthDec	4e-01	-1.64	1e-01
MonthFeb	4e+02	-0.03	1e+00
MonthJul	4e-01	0.18	9e-01
MonthJune	5e-01	-0.05	1e+00
MonthMar	3e-01	-1.24	2e-01
MonthMay	3e-01	-1.12	3e-01
MonthNov	3e-01	1.30	2e-01
MonthOct	4e-01	0.97	3e-01
MonthSep	4e-01	1.18	2e-01
VisitorTypeOther	6e-01	0.49	6e-01
VisitorTypeReturning_Visitor	2e-01	-3.73	2e-04
WeekendTRUE	1e-01	0.78	4e-01

0.4 Results

The final model is given by the regression equation: log([P(y=1)]) = 0.69 + 0.9593Administration + 1.0007Admin_duration + 1.00006ProdRelated_duration + 0.00113BounceRates + 1.14e-12ExitRates +0.56197MonDec + 6.1601e-07MonFeb + 1.07MonJul + 0.97MonJun + 0.65MonMar +0.69MonMay+ 1.51MonNov+1.44MonOct + 1.59MonSep + 1.36VisitorOther + 0.56VisitorReturning + 1.11WeekendTrue. The significance of individual predictors is assessed using wald's test.It is found that the predictors administration, admininstration_duration, productrelated_duaration, BounceRates, ExitRates, Month, VisitorType and Weekend can predict if customer will buy or not.The regression coefficients of the predictors can be interpreted as: $ln(odds_{Revenue|Administrative=x+1}) - ln(odds_{Revenue|Administrative=x}) = -0.0416 = exp(-0.0416) = 0.96$. This means holding all other variables constant, for one unit increase in administrative values we expect to see the odds of generating revenue decrease by 4%. We can expect to see a 0.07% increase in the odds of generating revenue for one unit increase in administrative duaration,holding all other variables constant.We can expect to see a 0.006% increase in the odds of generating revenue for one unit increase in product related duaration, holding all other variables constant.We can expect to see a 0.11% increase in the odds of generating revenue for one unit increase in bounce rate, holding all other variables constant.We can expect to see a 1.14e-10% increase in the odds of generating revenue for one unit increase in exit rate, holding all other variables constant. The odds of buying in Dec over the odds in buying in Aug is 0.56 with all variables constant.In terms of percentage

change, the odds of buying in Aug are 56% higher than the odds of buying in Dec. Similarly the odds of buying in Aug are 6.1601e-05% higher than the odds of buying in Feb with all other variables constant.The odds of buying in July is 7% higher than than the odds of buying in Aug with all other variables constant. The odds of buying in June is 97% lower tan the odds of buying in Aug with all other variables constant. The odds of buying in March is 65% lower than the odds of buying in Aug with all other variables constant. The odds of buying in May is 69% lower than the odds of buying in May with all other variables constant. The odds of buying in Nov is 51% higher than the odds of buying in Aug with all other variables constant. The odds of buying in Oct is 43% higher than the odds of buying in Aug with all other variables constant. The odds of buying in Sep is 58% higher than the odds of buying in Aug with all other variables constant. The odds of buying by Other visitor is 36% more than the odds of buying by new visitor with all other variables constant. The odds of buying by Returning visitor is 56% lower than the odds of buying by new visitor with all other variables constant. The odds of buying during weekend is 11% more than the odds of buying during weekdays with all other variables constant.

Table 7: Table of model parameters

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.41	3e-01	-1.25	0.21
Administrative	-0.04	2e-02	-1.88	0.06
Administrative_Duration	0.00	0e+00	1.83	0.07
ProductRelated_Duration	0.00	0e+00	1.87	0.06
BounceRates	-6.79	8e+00	-0.89	0.37
ExitRates	-27.46	5e+00	-5.74	0.00
MonthDec	-0.58	3e-01	-1.64	0.10
MonthFeb	-14.33	4e+02	-0.03	0.97
MonthJul	0.08	4e-01	0.18	0.86
MonthJune	-0.02	5e-01	-0.05	0.96
MonthMar	-0.43	3e-01	-1.24	0.22
MonthMay	-0.36	3e-01	-1.12	0.26
MonthNov	0.41	3e-01	1.30	0.19
MonthOct	0.36	4e-01	0.97	0.33
MonthSep	0.46	4e-01	1.18	0.24
VisitorTypeOther	0.31	6e-01	0.49	0.63
VisitorTypeReturning_Visitor	-0.58	2e-01	-3.73	0.00
WeekendTRUE	0.11	1e-01	0.78	0.44

Table 8: Table of OR and Confidence Interval

	OR	2.5 %	97.5 %
(Intercept)	0.664	0.3	1e+00
Administrative	0.959	0.9	1e+00
Administrative_Duration	1.001	1.0	1e+00
ProductRelated_Duration	1.000	1.0	1e+00
BounceRates	0.001	0.0	2e+03
ExitRates	0.000	0.0	0e+00
MonthDec	0.562	0.3	1e+00
MonthFeb	0.000	0.0	0e+00
MonthJul	1.080	0.5	2e+00
MonthJune	0.976	0.3	3e+00
MonthMar	0.652	0.3	1e+00
MonthMay	0.694	0.4	1e+00
MonthNov	1.511	0.8	3e+00
MonthOct	1.438	0.7	3e+00
MonthSep	1.586	0.7	3e+00
VisitorTypeOther	1.361	0.3	4e+00
VisitorTypeReturning_Visitor	0.561	0.4	8e-01
WeekendTRUE	1.111	0.8	1e+00

0.5 Discussion and conclusion

The final model includes the predictors administration, The model is fitted with 80% of the train data and predictions are made with the 20% of the data.The model is evaluated with confusion matrix, the metrics precision, F1 score , sensitivity and specificity shown in Table 9 and Table 10.The Figure 10 shows the ROC curve. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. The model has an AUC of 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class. The model had very low sensitvity= 0.004 and specificity=0.84 due to class imbalance of the response.The wald test used to test the significance of the variables indicates all predictors are statistically significant.The assumptions of logistic regression except the assumptions of linearity are met.The VIF for the predictors is less than 5.There is no multicollinearity among the predictors in the model.Table 4. Althoug there are few extreme observations they are below three standard deviations. The model does not include influential observations.

Table 9: Confusion Matrix

	FALSE	TRUE
FALSE	8254	45
TRUE	1504	24

Table 10: Performance Metrics

	x
Precision	0.35
Recall	0.02
F1Score	0.01
AUC	0.70

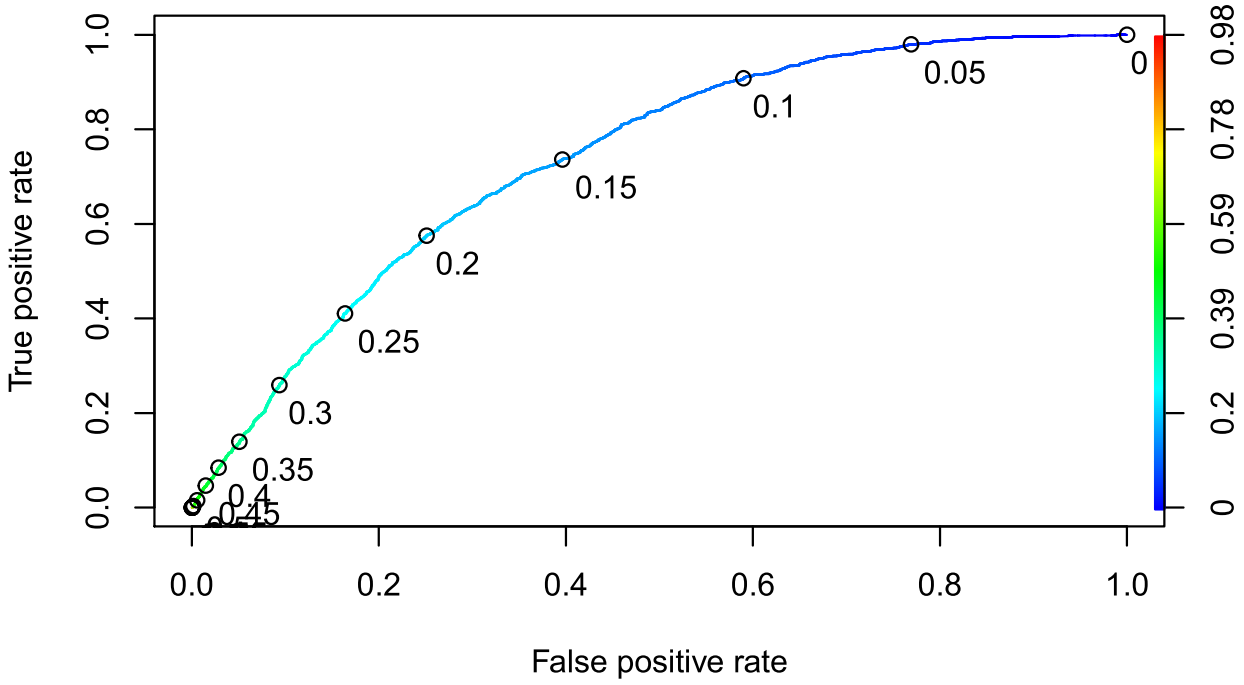

Figure 10: ROC curve