Predicting online shoppers' purchasing intention

Manas Sharma[1], Saikat Sengupta[2], and Abdur Rahman[3]

[1]*BS2137*
[2]*BS2140*
[3]*BS2146*

Abstract

We analyze a e-commerce shopper analytics dataset and build a model that can predict whether a customer will generate revenue or not. From the given requirement, it is clear that classification is the suitable solution since we need to bucket the customers into two categories (revenue generated or not generated). Once we build the model, we will discuss strategies that e-commerce sites use to increase the conversion rate.

Keywords Logistic Regression, Online shopper behaviour, Conversion rate, Bounce rate, Exit rate, Page value

1 Introduction

E-commerce is flourishing at a fast rate across India (evident from its' increasing contribution to India's GDP). Like any other business, losses do occur in e-commerce too. Every company tries to cut down their losses, and the one who does it efficiently becomes an e-commerce giant. For example, Amazon is a giant and Snapdeal is a heavy loss making company.

At shopping malls, salesmen provide customers a wide range of customized choices based on his/her experience and feedback from past customers. This "experience" and "customer feedback" has an important influence on saving time, purchase conversion rates and sales figures. E-commerce sites use algorithm that imitate a salesman's role and provide customers with varied choices based on their activity on that site.

The question that we are going to address in this report is "Given clickstream and session data of a user who visits an e-commerce website, we want to predict whether or not that visitor will make a purchase". Answering this question is of utmost importance for companies in order to ensure that they remain profitable. This information can be used to nudge a potential customer in real-time to complete an online purchase, increasing overall purchase conversion rates. Examples of nudges include highlighting popular products through social proof, exit intent overlay on webpages, push the sales of a specific product by putting it next to similar products that aren't perceived to be as good of a deal (Decoy effect ; Apple has been it over the years) etc.

2 Dataset Description

The dataset used in our analysis was obtained from the UC Irvine Machine Learning Repository. The dataset was specifically formed so that each session (period of time spent by a user on an e-commerce website) would belong to a unique user over a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The total number of sessions in the dataset is 12,330.

We have a total of 18 covariates; brief description of these are given below. "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

Table 1. Numerical features used in user behavioural analysis.

Feature	Description
Administrative	# pages visited by the visitor about account management
Administrative duration	total time spent (in seconds) by the visitor on administrative pages
Informational	# pages visited by the visitor about communication, address etc.
Informational duration	total time spent (in seconds) by the visitor on informational pages
Product related	# pages visited by the visitor on product related details
Product related duration	total time spent (in seconds) by the visitor on product related pages
Bounce rate	average bounce rate value of the pages visited by the visitor
Exit rate	average exit rate value of the pages visited by the visitor
Page value	average page value of the pages visited by the visitor
Special day	closeness of the site visiting time to a special day

Table 2. Categorical features used in user behavioural analysis.

Feature	Description	# values
OperatingSystems	operating system used by the visitor	8
Browser	browser used by the visitor	13
Region	geographic region from which the session has been started by the visitor	9
TrafficType	traffic source (e.g., banner, SMS, direct)	20
VisitorType	visitor type as “New Visitor,” “Returning Visitor,” and “Other”	3
Weekend	value indicating whether the date of the visit is weekend	2
Month	month value of the visit date	12
Revenue	whether the visit has been finalised with a transaction	2

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Independence Day, Mother’s Day) in which the sessions are more likely to be finalized with transaction.

The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year. The attribute “Month” includes 10 months except January and April.

The target variable is Revenue which takes "TRUE" or "FALSE" corresponding to whether or not a user purchased or not during a session. Of the 12,205 sessions in the dataset, 84.37 % (10,297) were "FALSE" revenue class (that did not end with shopping) and the rest 15.63 %(1,908) were "TRUE" revenue class (ending with shopping). So this dataset is imbalanced.

3 Data Preparation and Data Analysis

3.1 Data Cleaning

There are duplicate records in the dataset, after removing 125 duplicates, 12,205 records remain and there is no missing value. The dataset has good data quality in general. There are some minor

issues with outliers, but we have decided not to handle them.

3.2 Converting Categorical variables to Numeric variables: One Hot Encoding

We have several categorical variables like OperatingSystems, Browser, Region etc. We know that we can't fit a model directly using the categorical variables. So we convert the categorical variables to strings of 0's and 1's. Say a user is using operating system number 3 (out of the 8 OS used in the dataset); then it corresponds to (0,0,1,0,0,0,0,0). This is popularly known as One Hot Encoding. However there is a drawback; there is a drastic increase in number of columns, hence increasing the dimension. Here number of columns increases from 8 to 69.

3.3 Data Analysis

The first 6 numeric features represent number of pages visited of different types and average time spent on these pages, of which the medians of numbers are 1, 0 and 18 and the medians of time are 9, 0, 608.9 respectively. This shows only a small percentage of visitors visiting the product page visits the checkout page. Also this illustrates that very few visitors choose to view only one product rather the probability that they explore more related products is relatively much higher.

3.4 Right Skewed Distributions

We observed that many numeric features are right-skewed with long tails. This is common in e-commerce sites because some users have extremely high usage statistics (they browse a pot comparing products to get the best deal).

3.5 Importance of Page value

Page value is defined as the average page value of the pages visited by the user. In an e-commerce context, values are normally assigned to important pages such as checkout pages, or pages preceding the checkout process. As seen from Figure 1., having a Page value of above 5 increases the likelihood of purchase conversion. Hence, this Page value feature provides a strong signal on whether the user will make a purchase or not.

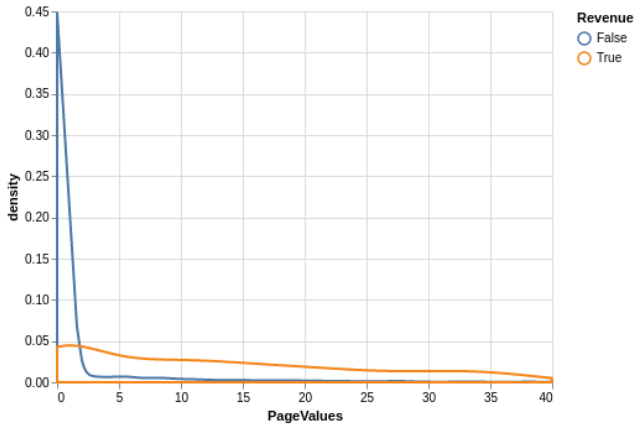

Figure 1. Density Plot of Page value

3.6 Correlation plot of the dataset

This is the Correlation Heatmap Plot. The colour intensity of the square (i,j) represents how good is the correlation between the i-th and the j-th covariates. The colour intensity is high only along the diagonal (obvious since a covariate is highly correlated to itself) and colour intensity is low in the off-diagonal areas. Since the covariates are uncorrelated, its good for our model.

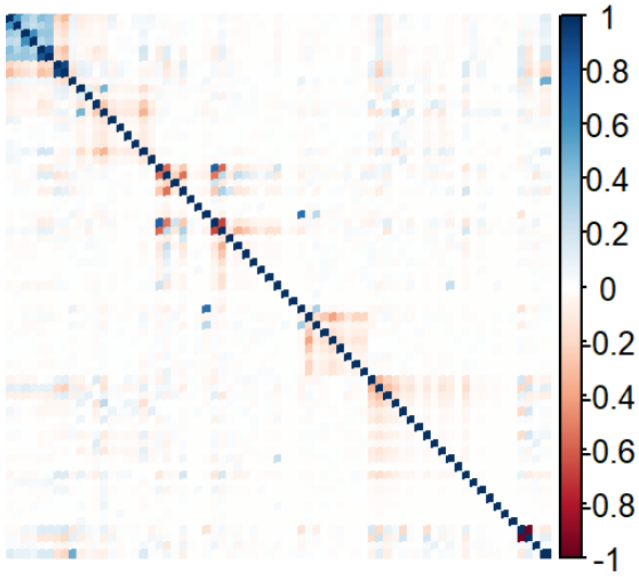

Figure 2. Correlation Plot

3.7 The problem of imbalance

The data as pointed out before, is heavily imbalanced. About 84% of the data is classified as 0's or a non revenue generating individual. This begs the question, how exactly are we supposed to fit a good predicting model to a dataset when even if we just blindly say that no revenue is generated, it will be accurate 84 times out of 100. How do we ensure that the model that we fit is not suffering from overestimation of 0's? And how do we consider a better metric than the accuracy score that may hide the overestimation.Let's first look into a few simple and intuitive methods to counteract class imbalance!

The most straightforward method to counteract class imbalance is undersampling. Undersampling means that you discard a number of data points of the class that is present too often. The disadvantage of undersampling is that you lose a lot of valuable data The advantage of undersampling is that it is a very straightforward technique to reduce class imbalance.

Another simple solution to imbalanced data is oversampling. Oversampling is the opposite of undersampling. Oversampling means making duplicates of the data that is the least present in your data set. You then add those duplicates to your data set. The advantage of this is that you do not have to delete data points, so you do not delete and valuable information. On the other hand, you are creating data that is not real, so you may be introducing false information into your model.

We now introduce SMOTE (Synthetic Minority Oversampling Technique).SMOTE is an algorithm that performs data augmentation by creating synthetic data points based on the original data points. SMOTE can be seen as an advanced version of oversampling, or as a specific algorithm for data augmentation. The advantage of SMOTE is that you are not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points. The SMOTE algorithm works as follows:

- You draw a random sample from the minority class.
- For the observations in this sample, you will identify the k nearest neighbors.
- You will then take one of those neighbors and identify the vector between the current data point and the selected neighbor.
- You multiply the vector by a random number between 0 and 1.
- To obtain the synthetic data point, you add this to the current data point.

4 Fitting a Model

Now we go on to fit a model to our data. Considering we have a binary classification problem at hand, we will aim to do logistic and probit regression to successfully predict whether a customer will generate revenue or not. Our aim being to then choose between the two, the more successful model using our decided metrics.

4.1 Fitting a Logistic Model

The logistic model (or logit model) is a model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.5986   0.0065   0.5902   2.4589

Coefficients: (2 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.653e+00  1.227e+00  -1.348 0.177746
Administrative                 2.910e-02  8.299e-03   3.507 0.000453 ***
Administrative_Duration       -5.566e-04  1.530e-04  -3.639 0.000273 ***
Informational                  4.820e-02  2.101e-02   2.294 0.021814 *
Informational_Duration        -2.444e-04  1.757e-04  -1.391 0.164161
ProductRelated                 2.737e-03  9.553e-04   2.865 0.004164 **
ProductRelated_Duration        5.949e-05  2.162e-05   2.752 0.005926 **
BounceRates                   -4.211e+00  1.929e+00  -2.182 0.029076 *
ExitRates                     -1.041e+01  1.564e+00  -6.658 2.77e-11 ***
PageValues                     1.481e-01  3.010e-03  49.212  < 2e-16 ***
SpecialDay                    -5.523e-02  1.662e-01  -0.332 0.739650
Month_Feb_bool                -8.608e-01  3.376e-01  -2.550 0.010788 *
Month_Mar_bool                -2.223e-02  9.259e-02  -0.240 0.810273
Month_May_bool                -1.537e-01  9.424e-02  -1.631 0.102886
Month_Oct_bool                 5.199e-01  1.197e-01   4.345 1.39e-05 ***
Month_June_bool                4.602e-01  1.691e-01   2.722 0.006498 **
Month_Jul_bool                 1.002e+00  1.312e-01   7.641 2.16e-14 ***
Month_Aug_bool                 7.487e-01  1.323e-01   5.657 1.54e-08 ***
Month_Nov_bool                 1.402e+00  7.743e-02  18.110  < 2e-16 ***
Month_Sep_bool                 9.053e-01  1.271e-01   7.122 1.06e-12 ***
OperatingSystems_1_bool        1.731e-01  1.371e+00   0.126 0.899542
OperatingSystems_2_bool        3.883e-01  1.369e+00   0.284 0.776651
OperatingSystems_4_bool        1.072e-01  1.374e+00   0.078 0.937821
OperatingSystems_3_bool        3.108e-01  1.369e+00   0.227 0.820407
OperatingSystems_7_bool        1.231e+00  1.594e+00   0.773 0.439691
OperatingSystems_6_bool       -1.345e+00  1.545e+00  -0.871 0.383931
OperatingSystems_8_bool        1.262e-02  1.325e+00   0.010 0.992400
Browser_1_bool                 5.926e-01  8.338e-01   0.711 0.477268
Browser_2_bool                 4.052e-01  8.282e-01   0.489 0.624669
Browser_3_bool                -4.252e-01  8.961e-01  -0.475 0.635109
Browser_4_bool                 6.192e-01  8.319e-01   0.744 0.456627
Browser_5_bool                 5.645e-01  8.338e-01   0.677 0.498407
Browser_6_bool                -2.759e-02  8.522e-01  -0.032 0.974171
Browser_7_bool                 1.647e-01  9.000e-01   0.183 0.854819
Browser_10_bool                3.601e-01  8.487e-01   0.424 0.671312
Browser_8_bool                 1.017e+00  8.587e-01   1.184 0.236451
Browser_9_bool                -1.165e+01  8.827e+02  -0.013 0.989470
Browser_12_bool                2.624e+00  1.012e+00   2.593 0.009508 **
Browser_13_bool                       NA         NA      NA       NA
Region_1_bool                 -1.524e-01  1.232e-01  -1.238 0.215807
Region_9_bool                 -4.769e-01  1.693e-01  -2.817 0.004841 **
Region_2_bool                  1.009e-02  1.385e-01   0.073 0.941927
Region_3_bool                 -1.838e-01  1.280e-01  -1.436 0.150993
Region_4_bool                 -4.445e-01  1.411e-01  -3.150 0.001631 **
Region_5_bool                 -4.851e-01  1.932e-01  -2.510 0.012069 *
Region_6_bool                 -8.600e-03  1.478e-01  -0.058 0.953613
Region_7_bool                  5.117e-02  1.459e-01   0.351 0.725852
TrafficType_1_bool            -1.003e+00  2.047e-01  -4.899 9.63e-07 ***
TrafficType_2_bool            -6.894e-01  2.011e-01  -3.428 0.000608 ***
TrafficType_3_bool            -1.170e+00  2.075e-01  -5.636 1.74e-08 ***
TrafficType_4_bool            -8.602e-01  2.156e-01  -3.990 6.60e-05 ***
TrafficType_5_bool            -4.290e-01  2.454e-01  -1.748 0.080383 .
TrafficType_6_bool            -1.129e+00  2.391e-01  -4.720 2.36e-06 ***
TrafficType_7_bool            -8.604e-01  4.304e-01  -1.999 0.045615 *
TrafficType_8_bool            -1.783e-01  2.323e-01  -0.768 0.442756
TrafficType_9_bool            -8.760e-01  5.261e-01  -1.665 0.095931 .
TrafficType_10_bool           -6.346e-01  2.270e-01  -2.796 0.005179 **
TrafficType_11_bool           -6.415e-01  2.487e-01  -2.579 0.009900 **
TrafficType_12_bool                   NA         NA      NA       NA
TrafficType_13_bool           -1.681e+00  2.330e-01  -7.217 5.32e-13 ***
TrafficType_14_bool           -1.757e-01  5.906e-01  -0.298 0.766062
TrafficType_15_bool           -1.336e+01  1.393e+02  -0.096 0.923585
TrafficType_18_bool           -1.332e+01  3.076e+02  -0.043 0.965443
TrafficType_19_bool           -1.347e+01  2.466e+02  -0.055 0.956444
TrafficType_16_bool            1.984e+00  9.407e-01   2.110 0.034890 *
TrafficType_17_bool           -1.275e+01  8.827e+02  -0.014 0.988475
VisitorType_Returning_Visitor_bool  2.814e-01  5.133e-01   0.548 0.583536
VisitorType_New_Visitor_bool   5.176e-01  5.158e-01   1.003 0.315628
Weekend_TRUE_bool              1.293e-01  5.227e-02   2.474 0.013378 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25140  on 18169  degrees of freedom
Residual deviance: 14201  on 18103  degrees of freedom
AIC: 14335

Number of Fisher Scoring iterations: 13
```

Figure 3. Summary of Logistic Model

4.2 Fitting a Probit Model

Probit regression, also called a probit model, is used to model dichotomous or binary outcome variables. In the probit model, the inverse standard normal distribution of the probability is modeled

as a linear combination of the predictors.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.6680   0.0030   0.7347   2.5256

Coefficients: (2 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.889e-01  6.454e-01  -0.758 0.448734
Administrative                   1.657e-02  4.638e-03   3.573 0.000353 ***
Administrative_Duration         -2.790e-04  8.375e-05  -3.331 0.000866 ***
Informational                    1.874e-02  1.191e-02   1.574 0.115494
Informational_Duration          -7.667e-05  1.008e-04  -0.761 0.446679
ProductRelated                   1.143e-03  5.472e-04   2.089 0.036667 *
ProductRelated_Duration          4.309e-05  1.260e-05   3.420 0.000626 ***
BounceRates                     -7.326e-01  1.026e+00  -0.714 0.475080
ExitRates                       -8.297e+00  8.603e-01  -9.645  < 2e-16 ***
PageValues                       6.139e-02  1.226e-03  50.061  < 2e-16 ***
SpecialDay                      -7.105e-02  8.629e-02  -0.823 0.410298
Month_Feb_bool                  -4.392e-01  1.709e-01  -2.570 0.010174 *
Month_Mar_bool                   1.836e-02  4.985e-02   0.368 0.712693
Month_May_bool                   2.342e-02  5.012e-02   0.467 0.640377
Month_Oct_bool                   3.660e-01  6.626e-02   5.523 3.34e-08 ***
Month_June_bool                  3.140e-01  9.332e-02   3.364 0.000767 ***
Month_Jul_bool                   6.093e-01  7.384e-02   8.251  < 2e-16 ***
Month_Aug_bool                   4.617e-01  7.444e-02   6.202 5.57e-10 ***
Month_Nov_bool                   8.121e-01  4.273e-02  19.006  < 2e-16 ***
Month_Sep_bool                   5.320e-01  7.148e-02   7.443 9.87e-14 ***
OperatingSystems_1_bool         -3.021e-01  7.104e-01  -0.425 0.670686
OperatingSystems_2_bool         -1.605e-01  7.089e-01  -0.226 0.820913
OperatingSystems_4_bool         -3.504e-01  7.121e-01  -0.492 0.622639
OperatingSystems_3_bool         -2.391e-01  7.091e-01  -0.337 0.735950
OperatingSystems_7_bool          1.714e-01  8.629e-01   0.199 0.842568
OperatingSystems_6_bool         -9.986e-01  7.965e-01  -1.254 0.209945
OperatingSystems_8_bool         -5.366e-01  6.866e-01  -0.782 0.434465
Browser_1_bool                   3.705e-01  4.331e-01   0.855 0.392307
Browser_2_bool                   2.606e-01  4.296e-01   0.606 0.544190
Browser_3_bool                  -2.018e-01  4.645e-01  -0.435 0.663902
Browser_4_bool                   3.465e-01  4.317e-01   0.803 0.422161
Browser_5_bool                   3.946e-01  4.331e-01   0.911 0.362184
Browser_6_bool                   7.234e-02  4.427e-01   0.163 0.870211
Browser_7_bool                   1.482e-01  4.716e-01   0.314 0.753326
Browser_10_bool                  2.710e-01  4.413e-01   0.614 0.539133
Browser_8_bool                   6.237e-01  4.481e-01   1.392 0.164011
Browser_9_bool                  -3.759e+00  6.051e+02  -0.006 0.995043
Browser_12_bool                  1.608e+00  5.451e-01   2.950 0.003183 **
Browser_13_bool                         NA         NA      NA       NA
Region_1_bool                   -8.462e-02  6.877e-02  -1.230 0.218523
Region_9_bool                   -2.411e-01  9.322e-02  -2.587 0.009689 **
Region_2_bool                    1.473e-02  7.723e-02   0.191 0.848794
Region_3_bool                   -8.311e-02  7.147e-02  -1.163 0.244888
Region_4_bool                   -2.253e-01  7.804e-02  -2.887 0.003893 **
Region_5_bool                   -2.955e-01  1.060e-01  -2.787 0.005317 **
Region_6_bool                   -7.680e-03  8.245e-02  -0.093 0.925784
Region_7_bool                    1.881e-02  8.168e-02   0.230 0.817889
TrafficType_1_bool              -5.936e-01  1.164e-01  -5.099 3.41e-07 ***
TrafficType_2_bool              -4.306e-01  1.146e-01  -3.757 0.000172 ***
TrafficType_3_bool              -7.419e-01  1.178e-01  -6.297 3.04e-10 ***
TrafficType_4_bool              -5.535e-01  1.220e-01  -4.537 5.69e-06 ***
TrafficType_5_bool              -3.300e-01  1.394e-01  -2.367 0.017931 *
TrafficType_6_bool              -6.855e-01  1.338e-01  -5.122 3.02e-07 ***
TrafficType_7_bool              -4.224e-01  2.309e-01  -1.829 0.067341 .
TrafficType_8_bool              -1.178e-01  1.323e-01  -0.890 0.373572
TrafficType_9_bool              -5.344e-01  2.843e-01  -1.880 0.060092 .
TrafficType_10_bool             -3.609e-01  1.289e-01  -2.800 0.005105 **
TrafficType_11_bool             -3.819e-01  1.411e-01  -2.707 0.006796 **
TrafficType_12_bool                     NA         NA      NA       NA
TrafficType_13_bool             -1.032e+00  1.306e-01  -7.898 2.83e-15 ***
TrafficType_14_bool             -1.684e-01  3.324e-01  -0.507 0.612281
TrafficType_15_bool             -5.230e+00  9.272e+01  -0.056 0.955018
TrafficType_18_bool             -5.132e+00  2.072e+02  -0.025 0.980237
TrafficType_19_bool             -5.458e+00  1.610e+02  -0.034 0.972960
TrafficType_16_bool              1.078e+00  5.312e-01   2.029 0.042446 *
TrafficType_17_bool             -4.467e+00  6.051e+02  -0.007 0.994110
VisitorType_Returning_Visitor_bool  1.874e-01  2.831e-01   0.662 0.507981
VisitorType_New_Visitor_bool     3.460e-01  2.843e-01   1.217 0.223520
Weekend_TRUE_bool                8.790e-02  2.918e-02   3.012 0.002595 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25140  on 18169  degrees of freedom
Residual deviance: 15072  on 18103  degrees of freedom
AIC: 15206

Number of Fisher Scoring iterations: 15
```

Figure 4. Summary of Probit Model

We usually have contracted the set of co-variates to the set of only the significant ones in previous classes and then proceeded to check if it improves the prediction but here, there is a slight problem. Since we wished to avoid the dummy variable trap, we already removed one category when making our dummy variables. Our insignificant predictors lie in this set of dummy variables. But when we remove these predictors, we actually lose prediction power. This is because the nth variable that we removed, for e.g. December in months, are significant. Say we remove insignificant March. Now if we remove a second category, all 0s in the category variables implies it either belongs to March or December. We average out the significance of these predictors essentially, clearly losing significance. Hence we can’t remove insignificant dummy variables with abandon.

5 Choosing between two models

We will now select between the two models through some parameters that allow us to estimate model efficiency above particular data considerations.

5.1 Balanced Accuracy

We start with the simplest of them all, plotting the curve for balanced accuracy with respect to cutoffs to see if we can tell if one curve lies below the other and hence directly comment on which model is better. But as we plot these below, we see it is very hard to tell as such. Thus we try to look at a better way to judge.

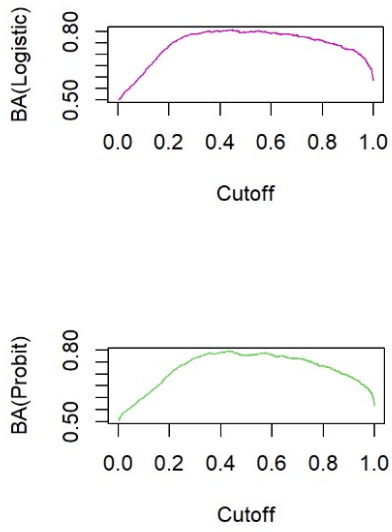

Figure 5. Balanced Accuracy Plots

5.2 AUCROC

Another approach is one going through the Receiver Operating Characteristic Curve (ROC). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Just like the ones we have plotted. Ours are mirrored since we plot Specificity on the x axis which is 1-FPR. The graph tells us how Sensitivity(TPR) and Specificity(1-FPR) vary together. The best point obviously being (1,1). The curve in itself is very hard to draw inference from and hence we consider a more concise statistic derived from the curve. AUC or Area Under Curve is exactly what it says, the area under the ROC curve. A higher AUC states a higher classification rate and hence a better model.

But we have AUC values of 0.894 and 0.883 respectively for logistic and probit respectively. Values so close, it is hard to say one is better or worse. But we get an inkling from both above methods that logistic seems to hold an edge over the probit model.

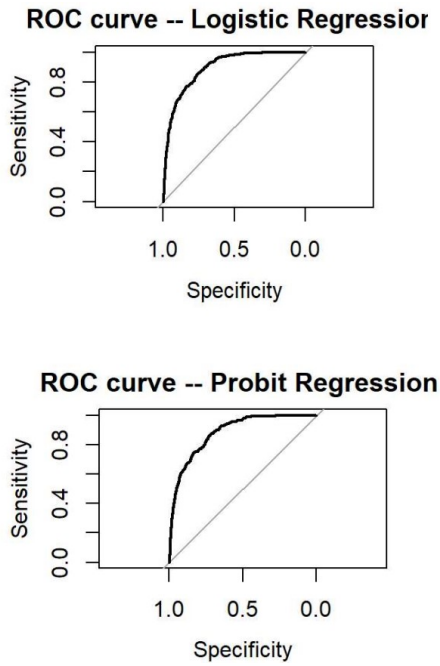

Figure 6. ROC Plots

5.3 AIC

To confirm our suspicions at least at the level of our specific dataset at least, we use the last statistic that we have already encountered above. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.
We can clearly see in the above pictures, that the AIC of the logistic model is clearly lower than the one for probit. We assume this to be enough working evidence to bias ourselves towards the logistic regression model.

6 Choosing the optimal cutoff

Any firm that is wishing to rise above its loss making days will wish to know a way to correctly classify whether a customer will give it revenue. Hence we maximise balanced accuracy which is the arithmetic mean of specificity and sensitivity which allows a firm to understand which customer it has to sell to. As we have previously created the graph, we just choose the maximum point on the graph as our threshold.

7 Insights from our Model

- We see that data for the month of April and January is missing because one marks the start of the year and the other marks the start of the financial year and hence is used for maintenance.
- Another observation is that estimate for the month of February is negative. This may seem absurd at first glance because we except heavy sales due to Valentine's Day. We hypothesized that; most of the buyers buy something expensive for Valentine's Day, and a buyer will always prefer physical shopping over online shopping when they are buying something expensive.
- Higher page value implies that chance of revenue generated is high, while a higher exit rate implies lower revenue which seems highly logical.
- The estimate for special day is negative. special day feature indicates the closeness of the site

visiting time to a specific special day in which the sessions are more likely to be finalized with transaction. Visitors usually wait for offers during special days; so most of them keep products added to their cart and purchase it as the special day comes closer than a threshold. So there is a disruption in regular sales which are not very far away from the special days.

- Browser 13 is NA because of the fact that it is a chromium based browser. Chromium browser data gets intermingled causing the co-variate in that area to become redundant due to over counting of the site data.

8 Conclusion

8.1 Recommendations to increase conversion rate based on our model

- Recommendations from personal sources generate more revenue. For example, an endorsement by an celebrity you like or a recommendation by a close one is expected to generate more revenue compared to an ad on a random website. So companies must invest more in personal recommendation traffic sources.
- Sites must ramp up their advertising campaign and give special deals during weekends and special days. But these special campaigns should not be too long or heavy; buyers would wait for these special deals and usual revenue balance would be affected.
- Sites must provide a variety of choices to the visitors, but the number of alternatives should be minimally sufficient. Otherwise exit rates would increase and there would be a loss of interest among buyers.
- User interface of administrative and informational pages must be optimized for speed. According to our model, less time spent on administrative and informational pages would generate more revenue.

8.2 Final Comments

Looking at the fit parameters, while the logistic regression model holds good predictive power, there seems to be a very clear chance of improvement. But as a model for behavioural analysis, it has the benefit of having meaning to its results. We can infer about an individual's behaviour through the values of the coefficients of the co-variates. This allows us to check whether what the model predicts holds true in our own logical framework.

9 Citation

1. Logistic Regression
2. Probit Model
3.Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks