**Predicting Customer’s Buying Intentions when Visiting a Website**

Omkar Modak

Elizabeth Gutierrez

Sagar Kulkarni

Vivek Mankar

University of South Florida

Data Mining (ISM6136)

Dr. [Balaji Padmanabhan](https://usflearn.instructure.com/courses/1356607/users/993403)

Fall 2019

**Abstract**

In our paper we analyzed a dataset selected from Kaggle’s website in which the results of customers use of webpages were recorded throughout a year. The motive for that timeframe was to account for different types of seasons, events, and user specified factors to avoid potential bias and obtain a representative sample of the population. For our analysis we used different statistical and analytical tools such as Python, Tableau, and Azure ML to predict the buying intentions of e-commerce site visitors as well as to identify the factors which mostly affect and define the customers’ buying intentions, which determine the revenue for the businesses. We used three different models to obtain our predictions and compared each of their results to select the one with the highest accuracy and best overall fit. The models implemented were Two Class Boosted Decision Tree, Two Class Decision Tree Forest, and Neural Network. In the preparation of our data we also conducted necessary adjustments to the dataset such as using correlation analysis to select the most correlated factors and obtain more accurate results. Another relevant technique we the use of the Smote algorithm to account for the class imbalance present in our dataset. Through the visualization of our data and the models’ implementation we found out that some of the factors with the highest importance for triggering revenue were the time spent by customers on the page, month of the year, and the page value, with other factors having some importance as well. The model which provided us with the highest accuracy for our results was the Two Class Boosted Decision Tree.

**TABLE OF CONTENTS**

1. Introduction   2

2. Data and Methodology    4

3. Models and Results  9

4. Business Problems and Suggestions 12

5. Conclusion 13

6. References 14

**1.Introduction**

Through our modern years, e-commerce has increasingly become a core feature for the operations and success of most businesses. Using the internet as a platform for advertising, selling, and buying products, more than a trend, is the norm for current business practices. Because of the convenience that e-commerce presents to customers when being able to make online transactions from the commodity of their home, Information Technology is being heavily used to track patterns, analyze information, and create and add value to interactions and relationships between businesses and their clients. Another relevance specifically for business owners is the growth opportunities that e-commerce provides to them since it is less costly for them to sell products and services online than it was while using physical and traditional methods. All these advantages add to the fact that while selling online companies have a much wider reach to potential buyers around the planet, which increases their revenue opportunities. Because of how crucial e-commerce is for our world’s economic growth, specifically for our small businesses’ revenue growth, we chose to conduct this statistical analysis in hopes to predict the buying intentions of customers when visiting e-commerce sites and the factors affecting these intentions the most, which lead to revenue generation for business.

For our analysis we chose a Kaggle’s dataset registering web pages’ customers usage for a year of time. Our data consisted of 12,330 sessions, each session representing a user with 10 numerical and 8 categorical attributes, or factors each. This was the number of attributes to begin with, which was later adjusted through data processing to account for importance to our study. Examples of attributes in our dataset are: Administrative, Administrative duration, Informational, Informational Duration, Product Related, and Product Related Duration, each representing the amount of these types of pages and the time spent in these pages by each user in a session. These values come from the pages’ URL information after users visit them. Also, factors such as Bounce Rate, Exit Rate, and Page Value are present as metrics whose values are obtained from Google Analytics analysis. Other important attributes are region, traffic type, browser and operating system type, returning or new visitor, weather a weekend visit, month of the year, and special day if the visit coincides with some holiday or significant day event. Our target variable was Revenue, which had a binary response of True or False, True meaning the generation of revenue and False meaning no revenue. The main purpose of our analysis was to predict and identify the factors which highly affected the Revenue, which was a direct response from the customers’ buying intentions. We also wanted to discover other significant insights from the data such as specific periods of times which yielded higher or lesser revenue, pages where customers spent the most and least time, and other discoverable information with the potential to help business owners find areas with opportunities for improvement.

**2. Data and Methodology:**

**2.1 Some independent Variables in dataset:**

The business has a website with three different webpages –

Administrative, Informational, Product Related

Bounce Rate – Bounce rate can be defined as the rate at which a user navigates to a different web page without exiting from the current web page after viewing it only one time.

Exit Rate – Exit rate can be defined as the rate which a user exits from a page and navigate to a different page.

Page Value - Page value can be explained as the average value of a page which a user visited before landing on the final/goal page. Page value is a dollar value and contributes to your site’s revenue.

Other factors such as the type of the customer and time of the year are also very important factors contributing for the result on the dependent variable:

Operating system: As name itself suggests it is the operating system which the user used while visiting the website

Traffic Type: This factor talks about the amount of traffic on the web page at a particular time during the browsing.

**2.2 Data Cleaning:**

**Dealing with Missing values:**

**For Numerical Columns:**

We are replacing the numerical values in the data with the Mean of the columns.

This works by calculating the mean of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

**For Categorical Columns:**

We are replacing the Categorical values in the data with the mode of the column. This is done by replacing missing data with the most frequent values within each column.

**Dealing with -1 Values in the Duration column.**

We are setting the -1 values in the duration column as NaN and are then imputing them with the Mean of the column.

**Categorical Encoding:**

Machine learning algorithms cannot handle categorical variables unless they are converted to numerical values and many algorithm’s performance varies based on how Categorical variables are encoded.

There are many ways we can encode these categorical variables as numbers and use them in algorithm.

**One Hot Encoding:**

In this method, we map each category to a vector that contains 1 and 0 denoting the presence or absence of the feature. The number of vectors depends on the number of categories for a feature. This method produces a lot of columns that slows down the learning significantly if number of the category is very high for the feature. We have used the get\_dummies function in Pandas which is quite easy to use.

We have used One Hot Encoding on all categorical variables except Revenue and Weekend:

**# one hot encoding**

data\_ohe = pd.get\_dummies(df.iloc[:, :-1])

data\_ohe.columns

Out[63]:

Index(['Administrative', 'Administrative\_Duration', 'Informational',

'Informational\_Duration', 'ProductRelated', 'ProductRelated\_Duration',

'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay',

'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'Weekend',

'Month\_Aug', 'Month\_Dec', 'Month\_Feb', 'Month\_Jul', 'Month\_June',

'Month\_Mar', 'Month\_May', 'Month\_Nov', 'Month\_Oct', 'Month\_Sep',

'VisitorType\_New\_Visitor', 'VisitorType\_Other',

'VisitorType\_Returning\_Visitor'],

dtype='object')

**2.3 Label Encoding:**

In this encoding each category is assigned a value from 1 through N (here N is the number of the category for the feature. One major issue with this approach is there is no relation or order between these classes, but algorithm might consider them as some kind of order or there is some kind of relationship. Hence, we have used this technique only on Revenue and Weekend Columns which has Boolean values and there is no relation between the rows.

**# label encoding of revenue**

from sklearn.preprocessing import LabelEncoder

df1 = pd.DataFrame()

le = LabelEncoder()

df1['Revenue'] = le.fit\_transform(df['Revenue'])

df1['Revenue'].value\_counts()

**# label encoding of Weekend**

data\_ohe['Weekend'] = le.fit\_transform(df['Weekend'])

data\_ohe['Weekend'].value\_counts()

**# Combining 2 dataframes**

df\_encoded = pd.concat([data\_ohe,df1], axis =1)

**2.4 Co-relation Matrix**

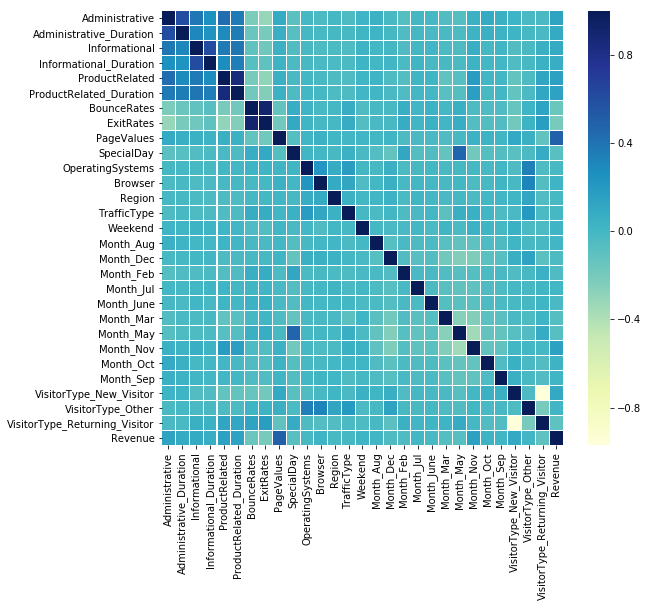
A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

**Applications of a correlation matrix**

There are three broad reasons for computing a correlation matrix:

* To summarize a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables highly correlate with each other.
* To input into other analyses. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise.
* As a diagnostic when checking other analyses. For example, with linear regression a high amount of correlations suggests that the linear regression’s estimates will be unreliable.

Below is the pictorial representation of the co-relation between the variables in our data:



* From the above graph, we can see that features such as Administrative, Administrative\_Duration, Informational, Informational\_Duration, ProductRelated, ProductRelated\_Duration, PageValues have a positive co-relation with respect to the target variable Revenue.
* Also, Columns such as BounceRate and ExitRate have a negative co-relation with respect to the target variable Revenue.

**2.5 Feature Selection:**

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance.

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

Benefits of Feature Selection:

· **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.

· **Improves Accuracy**: Less misleading data means modeling accuracy improves.

· **Reduces Training Time**: fewer data points reduce algorithm complexity and algorithms train faster.

**Feature Selection Technique Used:**

We have used the SelectKBest method for feature selection.

SelectKBest takes as a parameter a score function, which must be applicable to a pair (X,y). The score function must return an array of scores, one for each feature X[:,i]X[:,i] of X . SelectKBest then simply retains the first k features of X with the highest scores.

So, for example, if you pass chi2 as a score function, SelectKBest will compute the chi2 statistic between each feature of X and y (assumed to be class labels). A small value will mean the feature is independent of y. A large value will mean the feature is non-randomly related to y, and so likely to provide important information. Only k features will be retained.  
We have taken the 15 most co-related variables to the target variable. The features which were finalized were below

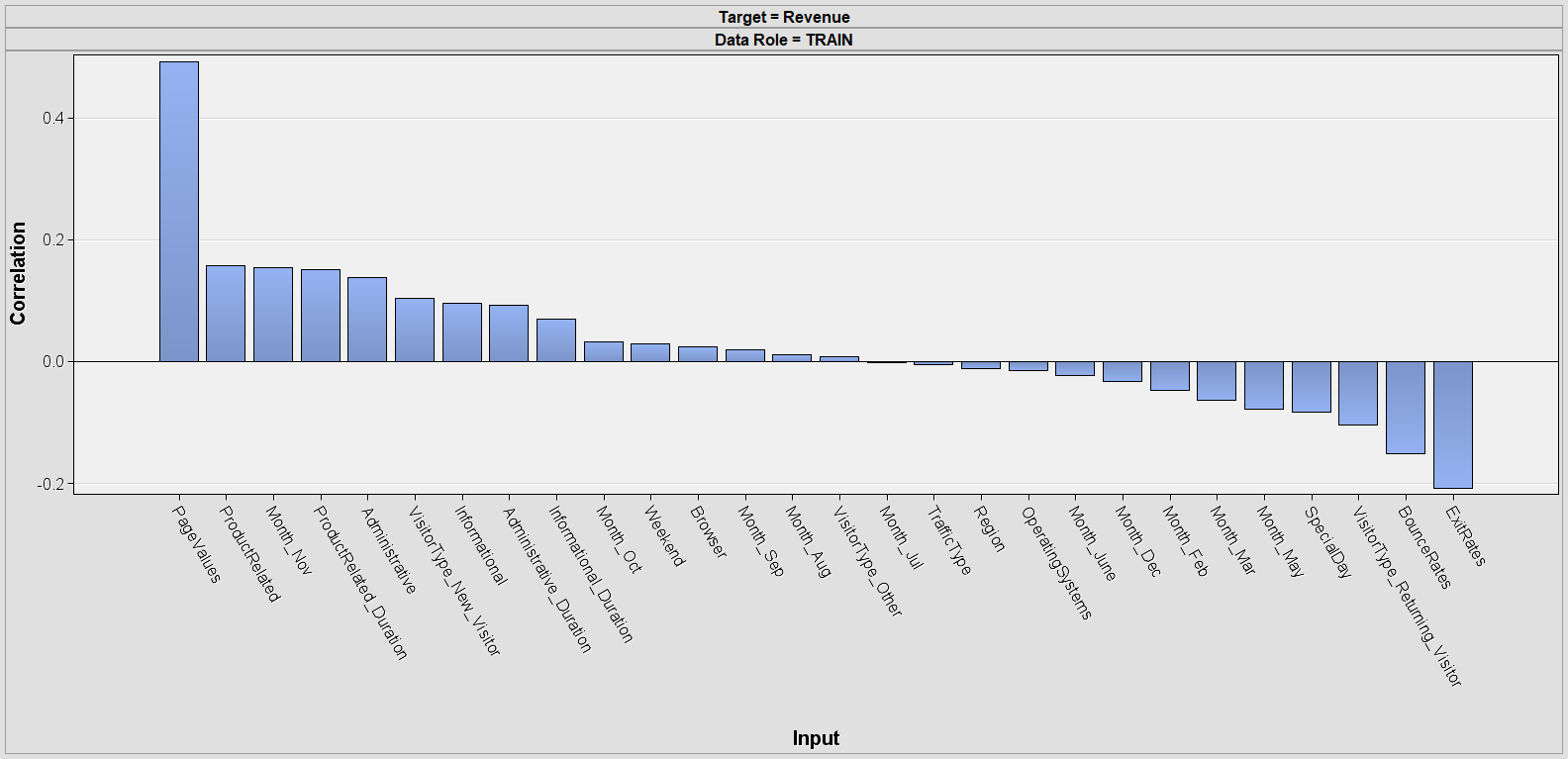
features\_df\_new.columns

Index(['Administrative', 'Administrative\_Duration', 'Informational', 'Informational\_Duration', 'ProductRelated', 'ProductRelated\_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month\_Feb', 'Month\_Mar', 'Month\_May', 'Month\_Nov', 'VisitorType\_New\_Visitor'], dtype='object')

**3. Models and Results**

**3.1** **Selecting Final independent variables for model building**: -

Here we are describing how the study was conducted, and why certain parameter choices were made. Below plots shows the correlation between independent variable and dependent variable (Revenue) in our selected dataset.



As we can see there are certain independent variables which are not related with our target variable. so, we will exclude these variables and start building our model by selecting rest of the variables.

List of variables used:

|  |  |
| --- | --- |
| **Independent Variables** | **Dependent (Target) Variable** |
| Page Values, Bounce Rate, Productrelated, Month\_nov, Product Related Duration, Visitor type\_New Visitor, Informational Duration, Administrative Duration, Month\_feb, Month\_march, Month\_may, VisitorType\_Returning Visitor, SpecialDay, ExitRates | Revenue |

**3.2 Learning Algorithms (Different Models used on datasets):**

We have selected the Azure ML for model building. As our target variable is binary, we have applied these algorithms

🡪 Two class Decision Forest

🡪Two class boosted decision tree

🡪 Neural Network.

**3.3 Model Evaluations and results: -**

**3.3.1 - Two class Decision Forest: -** First we have applied the two-class boosted decision tree on finalized dataset with the help of Azure ML. Random forests are bagged decision tree models that split on a subset of features on each split.

**Results: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **0.89** | **0.716** | **0.53** | **0.6** |

**Accuracy**: - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. we have got 0.89 which means our model is approx. 89% accurate.

**Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. We have got 0.716 precision which is pretty good.

**Recall** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. We have got recall of 0.53 which is good for this model as it’s above 0.5.

**F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is 0.6

**Experiments: -**

Here we have got good results but to improve model performance we have checked other models with applying some techniques to improve accuracy.

There is a class imbalance in our target variable means if we compare 1 and 0 values then we have 30% and 70% distribution so to balance it we used smote function.

**What is Smote?**

The SMOTE function oversamples your rare event by using bootstrapping and k-nearest neighbor to synthetically create additional observations of that event. The definition of rare event is usually attributed to any outcome/dependent/target/response variable that happens less than 15% of the time.

Now, we have again built the model to check whether we have improved our results or not.

**Results with Smote function: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **0.925** | **0.931** | **0.92** | **0.93** |

Now, we can see that our accuracy, precision, recall and F1 score all have increased so we have got better results when we used Smote function.

**3.3.2 - Two class boosted decision tree: -** After that we have used two class boosted decision tree with smote function. Decision tree learning uses a [decision tree](https://en.wikipedia.org/wiki/Decision_tree) (as a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling)) to go from observations about an item (represented in the branches) to conclusions about the item's target value.

**Results: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **0.932** | **0.933** | **0.940** | **0.936** |

Here we can see, we are getting better results than with using Two class decision forest.

**3.3.3 Two-class Neural Network: -** We have also built this model with the help of neural network algorithm. Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. Neural networks help us cluster and classify.

**Results: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **0.90** | **0.711** | **0.594** | **0.647** |

Here, we are getting our results which are not better than Two-class decision forest and Two-class boosted decision tree.

**3.3.4 Summary of applied models and result: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr.no** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **1** | **Two class Decision Forest** | **0.89** | **0.71** | **0.53** | **0.6** |
| **2** | **Two class Decision Forest with Smote function** | **0.92** | **0.93** | **0.92** | **0.93** |
| **3** | **Two class Boosted Decision tree** | **0.93** | **0.93** | **0.94** | **0.936** |
| **4** | **Two-class Neural Network** | **0.9** | **0.71** | **0.59** | **0.64** |

We can see with this table comparison Two-class Boosted Decision tree is producing better results as compared with any other algorithm hence we have finalized Two-class Boosted Decision tree as our final algorithm.

**4. Business Problems and Suggestions**

**4.1** **The business problems which we were able to identify were:**

* As a business owner, how can I increase my revenue of ecommerce website.
* Identify the pages where customers are spending less time.
* On which month do we get the least revenue.

**4.2 Suggestions:**

**4.2.1** The main concern for the owner would be to increase their revenue. While visualizing the data we identified that factors like "amount of time spent by a customer on a page" and " month of the year" affects the revenue. However, we were able to identify some hidden pattern in the data which suggested that factors like "page value", "Bounce Rate", "Exit Rate “also affects the overall revenue for the business. Thus, we would recommend the business to work on these factors to improve their revenues.

**4.2.2** Factors like – Weekend, Months describe the time of the year when the user visited the website and helps business to introspect which month showcased their majority revenue and which month had the least revenue.

**4.2.3** Visitor type: New visitor, returning user or any other user also affects the dependent variables and helps business to keep a track of the customer’s transactional history.

**4.2.4** We also identified that customers visiting the website were spending more time on “Product\_Related” pages and lesser time on pages like “Informational” and “Administrative”. Thus it would be a good practice for business to look upon these pages and try to increase the by average time spent by the customers.

**4.2.5** One important pattern which was learned from the data was month “July” saw the least amount of revenue as compared to all other months. Business could focus on this month to structure a better foundation and ultimately improve their revenue.

**5. Conclusion:**

Thus, to conclude we identified factors impacting the business revenues from data visualization and the predictive model helped us in finding some hidden pattern in the data.

Using three different models gave us a wide prospective to compare our results and come up with a better explanation and prediction for the business

**6. References**

**1.**

<https://www.kaggle.com/roshansharma/online-shoppers-intention>

**2.**

<https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest>

**3**

<https://www.displayr.com/what-is-a-correlation-matrix/>

**4**

<https://pandas.pydata.org/pandasdocs/stable/reference/api/pandas.get_dummies.html>

**5** <https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

**6**

<https://en.wikipedia.org/wiki/Decision_tree>

**7**

<https://skymind.ai/wiki/neural-network>