

PREDICTING THE LIKELIHOOD OF A RETAIL PURCHASE

Springboard | Mentor: Dipanjan Sankar

Using Online Shoppers’ Purchasing INtent

Lisa Patel

## INTRODUCTION

Now more than ever, online shopping is one of the primary ways people spend their money and represents a huge portion of B2C revenue. The ecommerce sector has drastically exploded in recent years, but especially more in 2020 due to the COVID pandemic, which completely altered human behavior in regards to the retail sector. Since the vast majority of retail’s customer base had to rely on shopping online, the US ecommerce sector grew by a staggering 44 % in 2020 according to data by the US Department of Commerce.[[1]](#footnote-1) The change can be clearly seen in the graph comparing the growth between ecommerce and total retail sales. Additionally, if you leave out Amazon, the top 100 retailers had a 74.1% of ecommerce growth in this past year, up significantly from a 49.4% share in 2019. These statistics are just scratching the surface since this sector will continue to grow in the upcoming years, so it is crucial that retailers analyze page-to-page clickstream data from their website to create a pleasant experience for the visitor and make effective, data-informed future business decisions to grow revenue.

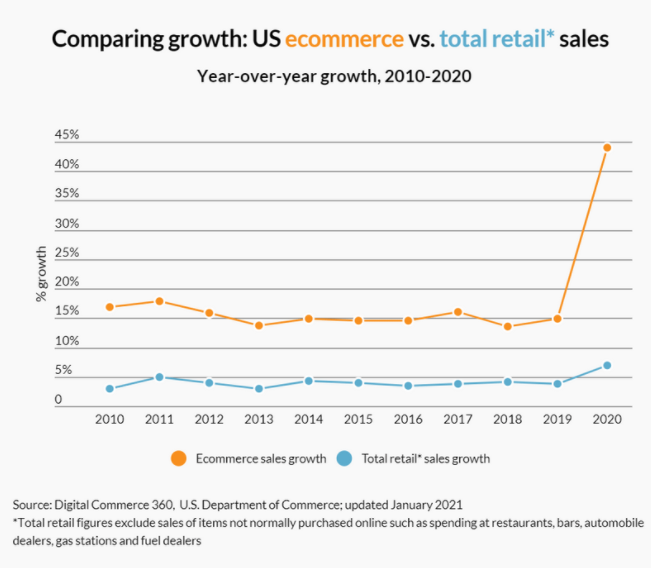


Figure : A visualization of the drastic growth in US ecommerce in 2020 as a result of a global pandemic that affected the majority of brick and mortar shops.

This type of analysis is also crucial because in a typical bricks-and-mortar environment, stores employ sales people that have learned to distinguish between shoppers based on their in-store behavior. In the cases where shoppers appear to be very focused in looking for a specific product, sales people may step in and help the shopper find what they are looking for. In other cases, if the shopper is merely “window shopping”, the experienced sales person can identify these shoppers and either ignore them and let them continue window shopping, or try and stimulate a purchase in the appropriate manner. However, in the virtual shopping environment, there is no sales person to perform that role.

By using page-to-page clickstream data from a given online store, visits can be categorized as a buying, browsing, searching, or knowledge-building visit based on observe navigational patterns in-store, including the general content of the pages viewed. Each type of visit varies in terms of purchasing likelihood. The shoppers, in each case, are also driven by different motivations and therefore would respond differentially to various marketing messages. The ability to categorize visits in such a manner allows the e-commerce marketer to identify likely buyers and design more effective, customized promotional message.

## OBJECTIVE

The main objective revolves around the identification of key metrics which contributes the most towards predicting a shopper's behavior and website abandonment likelihood. Along the way, we also suggest prioritized critical recommendations and performance improvements to increase our target feature, Revenue.

## DATA ACQUISITION

The Online Shoppers Purchasing Intent[[2]](#footnote-2) dataset was found online on UCI Machine Learning Repository and consists of feature vectors belonging to 12,330 individual sessions, making it suitable for machine learning. 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 has 10 numerical features and 8 categorical features. They are:

* **Administrative**: Number of pages of this type (administrative) that the user visited.
* **Administrative**\_**Duration**: Amount of time spent in this category of pages.
* **Informational**: Number of pages of this type (informational) that the user visited.
* **Informational**\_**Duration**: Amount of time spent in this category of pages.
* **ProductRelated**: Number of pages of this type (product related) that the user visited.
* **ProductRelated\_Duratio**n: Time spent in this category of pages.
* **BounceRates**: The percentage of people who leave your website after visiting only a single page. Here are some scenarios that count as a bounce on your website:
  + Someone clicks the back button after viewing a single page.
  + Someone exits their browser after viewing a single page.
  + A user clicks to another website that takes them elsewhere after viewing only a single page on your site
* **ExitRates**: The percentage of pageviews on the website that end at that specific page.
* **PageValues**: The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction. This value is intended to give you an idea of which page in your site contributed more to your site's revenue.
* **SpecialDay**: This value represents the closeness of the browsing date to special days or holidays (eg Mother's Day or Valentine's day) in which the transaction is more likely to be finalized.
* **Month**: Contains the month the pageview occurred, in string form.
* **OperatingSystems**: An integer value representing the operating system that the user was on when viewing the page.
* **Browser**: An integer value representing the browser that the user was using to view the page.
* **Region**: An integer value representing which region the user is located in.
* **TrafficType**: An integer value representing what type of traffic the user is categorized into. Read more about traffic types here.
* **VisitorType**: A string representing whether a visitor is New Visitor, Returning Visitor, or Other.
* **Weekend**: A boolean representing whether the session is on a weekend.
* **Revenue**: A boolean representing whether or not the user completed the purchase.

## DATA CLEANING

After the dataset was imported as pandas data frames, it was inspected for data quality, cleaned and transformed accordingly. The data cleaning steps were relatively simple since none of the columns had any null values or any that fell outside the typical range. There are 10,422 false cases and 1908 true cases. That imbalance in cases will have consequences in the modeling results, so it should not be taken lightly. In the exploration stage of data analysis, certain features will be dropped after considering their relevancy and usefulness for modeling.

## EXPLORATORY DATA ANALYSIS

In the exploratory data analysis stage, we will examine who the visitors were, when they were more likely to make purchases, how they arrived at the site, which region they were in, their operating system preference, what pages they visited and how long they stayed before making a purchase. Along the way, we also detail additional opportunities for growth and improvement in sales. The initial look at the correlation map in Figure 1 indicates the PageValues is strongly correlated with Revenue. Not strongly correlated with the target column are ExitRates and BounceRates, which make sense since it means people are leaving the site before making a purchase. There are strongly correlated with each other however.

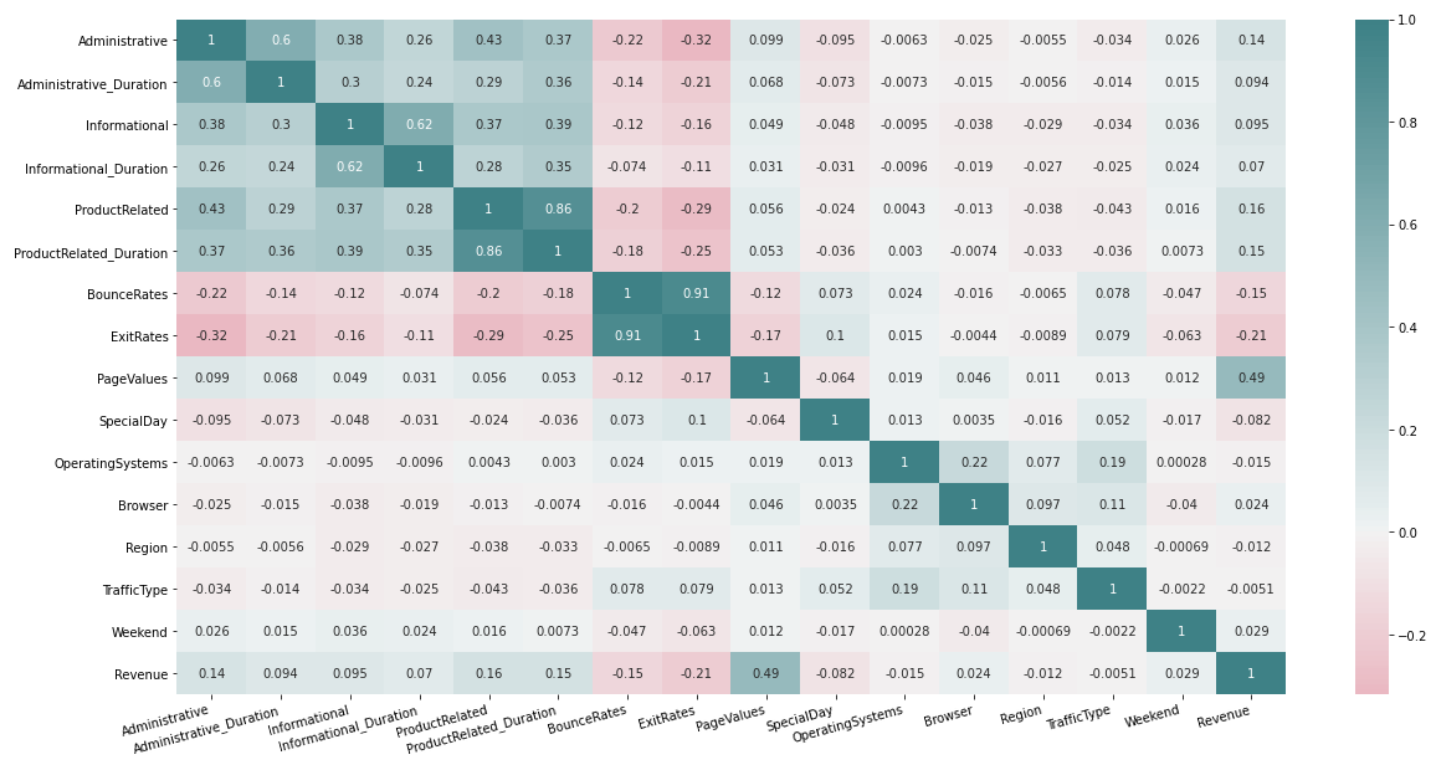


Figure 2: A heat map of all of the features to gain a sense of any strong correlations.

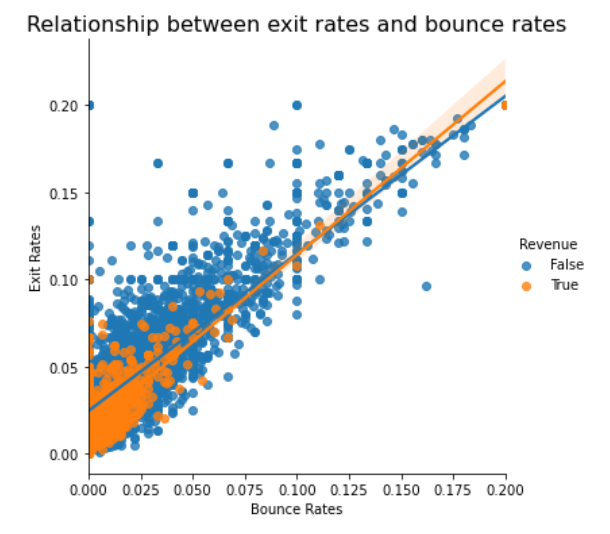
For clarification, the bounce rate is overall percentage of a single engagement session, or in the other words, the percentage of visitors who arrive at a landing page, but bounce off without navigating to a second page. It indicates the performance level of an entire website and is calculated by dividing the total number of single page visits by total numbers of visitors to a website. Exit Rates, on the other hand, the percentage of visitors who leave the site form a particular page after having visited any number of pages on that site. It is only specific to the landing page and is calculated by dividing the aggregation of total exits by the total visits to a page.

Figure : A linear relationship between the exit rates and bounce rates of the retail website. The majority of user who make a purchase have a low bounce rate, unsurprisingly.

One major difference between these is that exit rate is related to the overall percentage of visitors that were within the last session whereas bounce rates account for the percentage of visitors that were part of that one and only session. In other words, prior activity is not considered in the case of bounce rates.

A high bounce rate could indicate issues with user satisfaction from technical reasons such as unfriendly UI of the website, extremely slow throughput or other matters. According to BigCommerce, a bounce rate for retail commerce between 30 - 55% is acceptable. This data shows the bounce rates largely scattered lower than 10% as shown in Figure 3. According to UpSide Business[[3]](#footnote-3), a bounce rate lower than 5% is a cause of concern indicating a possibility of the Google Analytics code was inserted twice. **Hence more investigation is needed on the analytics code and changes needs to be made accordingly**.

Next, we can further examine certain features to see the broader trends as they relate to sales. We can ask questions like: what months were the most active, what day of the week did visitors prefer, how did the sales and types of visitors vary over the year and across region and operating systems, and more. Analyzing the trends will also give us ways to improve to increase revenue.

The large set of figures below gives an overview of how revenue is impacted by these 6 dimensions:

* **Visitor Type**: The majority of sales came from returning visitors. This can be a new opportunity of growth, if a system can be implemented to boost sales. The retailer could engage loyal customers in conversion of other customers by offering discounts for friends/family joining in or for new customers making a purchase in this manner.
* **Month**: The highest transaction volume took place in March, May, November and December. Sales in November account for nearly half of total sales, so what strategy being used should be maintained or improved upon. Zero sales in February are also concerning and needs to be evaluated more to figure out the cause. It should also be noted that certain months have missing data.
* **Region**: Regions 1 and 3 had the highest transaction volume, so more research could be done to find ways to broaden regional reach.
* **Operating Systems**: More visitors were likely to use operating system 2 more than the others. Additional information on the specific operating system would be useful for further analyses.
* **Weekend**: Transactions were more likely to occur on weekdays. This suggests that the firm has dealt with good retention with customers, but there is still room for improvement on conversion rates. Combining these finding could be an opportunity of growth by trying to encourage more returning customers to browse the site on weekends. This could be done by promoting a sale only available on the weekends through various traffic sources.
* **Special Days**: Visitors often bought merchandise during the week of the special day. Special days could also be a returning customer's birthday, so offering birthday discounts has the potential to increase revenue.

## 

Figure 4: Overviews of multiple features to examine how they vary by time, region and operating systems to make business recommendations.

## PRE-PROCESSING | MODELING

Having explored the trends in the data, we can start creating and evaluating the machine learning models. This is a supervised, classification problem, so will be using python’s scikit learn and imblearn to explore ways to improve the performance issues due to the data being imbalanced.

In terms of pre-processing, the categorical features that did not provide any additional useful information to the models were dropped. These include: Browser, OperatingSystem, TrafficType, and multiple page related/duration features. The remaining features that were categorical, like Month and ViistorType, were one-hot encoded. After everything was suitable to run, the data was split into appropriate 75/25 train/test sets.

Before delving into the models, it should be noted that while the prediction accuracy is the most common metric used for classification tasks, but it becomes inappropriate and misleading when it is used on an imbalanced data set. The model overfits to the class that’s represented more in your dataset and become oblivious to the existence of the minority class since the algorithm decides to classify everything in the majority class to get a good accuracy score. The model may naively look good, but in practice has no skills. So, an alternate performance metrics must be used. For imbalanced datasets, the recall score and precision are much better metrics.

The evaluation metrics for the various models that were tried are summarized in the Table 1 below. A simple logistic regression model was used first, giving at 73% precision and 36% recall for the true labels. The recall score significantly improved after modifying the parameter Class to be balanced. Even though the precision declined for this model, it correctly classified more of the true class, which is the metric we’re more concerned about in this scenario.

The other two models that were implemented were the Gaussian Naïve Bayes and an optimized Random Forest Model. Using grid search, the random forest model was optimized with the max depth being 15, minimum samples leaf of 3, minimum samples split of 2 and number of estimators being 100.

Looking at the performance results in Table 1 again, the Gaussian Naïve Bayes had the worst performance in terms of precision, but had a similar recall score to the balanced logistic regression model. While the optimized random forest model performed better in terms of precision compared to the BLG, it has a worse recall score. Considering all of these results, the balanced logistic regression seems to be the most appropriate choice for this data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | Balanced Logistic Regression | Gaussian Naïve Bayes | Optimized Random Forest |
| Precision | 0.73 | 0.54 | 0.37 | 0.73 |
| Recall | 0.36 | 0.68 | 0.69 | 0.51 |
| F1 | 0.48 | 0.61 | 0.48 | 0.60 |
| TP/TN | 170/2544 | 326/2332 | 327/2053 | 244/2514 |
| FP/FN | 307/62 | 151/274 | 150/553 | 233/92 |

Table : Results of various measures for the four machine learning models that were implemented.

There’s room for further analysis by examining the precision-recall curves for the models whenever it is possible for a particular model. Plots from the curves can be created and used to understand the trade-off in performance for different threshold values when interpreting probabilistic predictions. Each plot can also be summarized with an area under the curve score that be used to directly compare classification models. Generally, the higher the AUC-PR score, the better a classifier performs for the given task. A way to calculate AUC-PR is to calculate the AP, which summarizes a precision-recall curve as the weighted mean of precision achieved across all threshold.

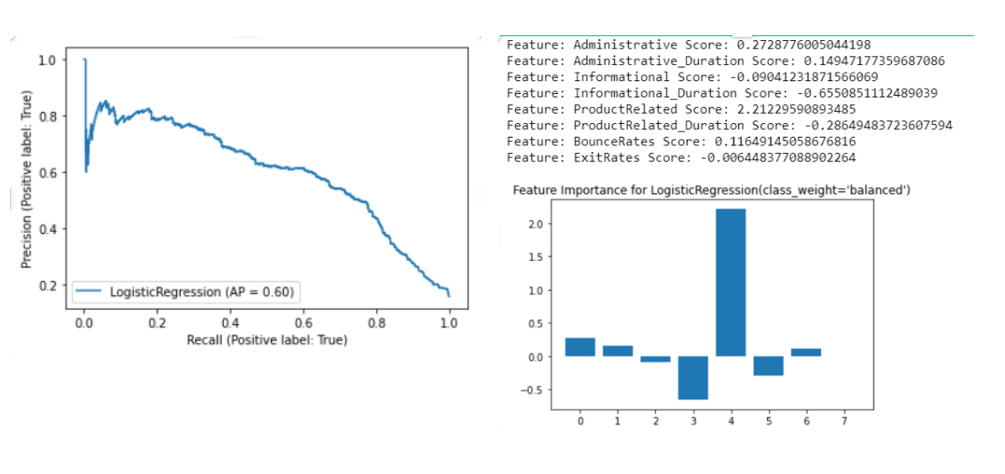
Since the best model was chosen to be the balanced logistic model and PR curves are not possible for certain others, only that was created. Figure 5 shows the PR curve and the resulting AP score of 60%, which indicates that this is a decent model. The figure also has the graph of feature importance, and it clearly shows the most important feature out of the 8 is Product Related Score. This makes sense since the customer is more likely to make a purchase, the longer they spend on the product page. So, it is critical for retail website to effectively design those pages and address any errors quickly.

Figure 5: Graph of PR curve for the chosen model as well as feature importances.

There was also an attempt to find alternate methods of fixing the imbalance issue by using a popular resampling technique called SMOTE, or Synthetic Minority Oversampling Technique. Oversampling usually involves duplicating the minority data from the minority class population. SMOTE, however, works by utilizing a k-nearest neighbor algorithm to create synthetic data by first choosing random data from the minority class. With the k-nearest neighbors from the data set, synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. This procedure is repeated enough times until the minority class has the same proportion as the majority class.

Unfortunately, when the oversampling method was implemented and the models were rerun, they were significantly worse than the models without the oversampling. They performed badly at categorizing both the positive and negative classes, so the results are not included in here. There are also other techniques, such as undersampling or a combination of both, that can be experimented with to further improve the performance of the model, but that will be reserved for future work.

## FUTURE WORK | CONCLUSION

Additional work with this data set should consist of getting additional information about the specifics of the features, so that a more in-depth analysis and business recommendations can be made. Any issues with the abnormally low bounce rate also need to be examined and fixed to get a more accurate interpretation. There is also room for improvement through a further optimization with additional hyperparameter tuning, such as trying a random grid search for all of the models. Lastly, interesting results may come from trying other resampling method or combining oversampling with undersampling.

Considering that there is much room for improvement in the models, this initial analysis did accomplish the main objective of finding the key feature(s) that contributes the most towards predicting a retail purchase as well making business related recommendations to increase revenue. The analysis showed that the number of pages of a product that a user visited has a major impact on their likelihood of making a purchase and that with a balanced logistic regression model, we can categorize users with a 54% precision and a 68% recall. With this knowledge, a retailer can start better designing their website and business practices to retain their existing customers and attract new visitors to their sites to survive these tumultuous times.

1. https://www.digitalcommerce360.com/article/us-ecommerce-sales/ [↑](#footnote-ref-1)
2. Data can be found here: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset [↑](#footnote-ref-2)
3. https://upsidebusiness.com/blog/my-website-bounce-rate-is-5-is-that-too-low-is-it-too-good-to-be-true/ [↑](#footnote-ref-3)