Online Shoppers Purchasing Intention

Apoorv Lunkad
Dept. of Mathematics
Stevens Institute of Technology
Hoboken, NJ
alunkad@stevens.edu

Ayush Ashutosh Panigrahi
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, NJ
apanigr1@stevens.edu

Kasyap Rayalacheruvu
Dept. of Mathematics
Stevens Institute of Technology
Hoboken, NJ
krayalac@stevens.edu

***Abstract*—This project is entailed to observe the activity of the shoppers using their click-stream data to predict whether a customer's visit to the online website leads to a purchase or not depending whether the visit lead to a revenue generation.**

Keywords - Shoppers, Revenue, EDA, Machine Learning

I. Introduction

A. Background

Consumer consumption habits have changed dramatically as a result of rapid advancements in computer technology and e-commerce. This not only includes business to consumer shopping but also business to business shopping and besides using a virtual online website on the internet you also have m-commerce with the increase in app development in the present days due to the ever increasing number of smart devices connected over the internet. This has further increase the percentage of purchase done from foreign websites( ) too Shopping online has opened up a new realm in the world of business which has been expanding at a constant rate. People nowadays are more prone to opening up the various websites or apps to look up the new inventory or the required essentials they might be in need of rather than going up to the stores. The popularity of online purchasing is growing. A accurate analysis system of online purchase patterns allows online shopping platforms to gain a better knowledge of customer psychology and develop better business tactics to enhance sales. At this time, it has become a great ordeal for the various sellers and the online platforms to delve deeper into the shopping habits and behavior of their customers to assess their customers. This pattern in the customers purchase intention can be easily predicted by analyzing the history of the customers. By getting valuable insights from the shoppers behavior the businesses can be benefited as they can understand which factors were more detrimental in resulting in a purchase adding onto their overall revenue. Based on the Online Shoppers Purchasing Intention Data Set, which comprises of aggregated page view data throughout the visit session as well as other user information, this report focuses on factors that may impact visitors' purchase intention.

B. About Data set

We are using the data set "Online Shoppers Purchasing Intention"[1] from the UCI repository a collection of databases that are used by the machine learning community for the empirical analysis of the multitude of machine learning algorithms[2]. Feature vectors from 12,330 separate sessions are included in the data set we're utilizing. To eliminate bias in the data set due to a proclivity for a single campaign or event, limits were placed on the number of users from a certain special day, user profile, or time. It was created such that each session belonged to a different user throughout the year with regard to these limitations. 85 percent (10,422) of the 12,330 sessions in the data set were negative class samples, which did not complete their shopping and generated no income, while the rest (1908) were positive class samples, which completed their shopping and generated money.The data collection has ten numerical and eight categorical features. The revenue feature is the class label, which is either true or false depending on whether the consumer purchased something, resulting in revenue generation for the vendor.
The terms "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" refer to the variety of different types of pages that can be accessed by the viewer during that session, as well as the total spending time on each of these page sections. These characteristics' values are obtained from the URL information of the user's visited pages and updated in real time when the user performs an action, such as switching from one page to another. The statistics measured by "Google Analytics" for each page on the e-commerce site are represented by the "Bounce Rate," "Exit Rate," and "Page Value" features. The proportion of visitors that reach the site through that page and subsequently exit ("bounce") without making any more requests to the analytics server during that session is the value of the "Bounce Rate" feature for that page. For a certain web page, the value of the "Exit Rate" feature is determined as the percentage of all page views to the page that were the last in the session. The "Page Value" function displays the average value of a web page seen by a user prior to making an e-commerce purchase. The "Special Day" feature highlights how near the site visiting time is to a certain special day (for example, Mother's Day or Valentine's Day), when sessions are more likely to be completed with a purchase. The value of this feature is established by taking into account e-commerce factors such as the time taken between an order and its delivery. For example, lets consider the event of valentines day, between February 2 and February 12 this value takes a nonzero value, zero before and after this date unless it is near to another special day and a maximum value of 1 on February 8. The data set also includes information about the operating system, browser, area, traffic type, visitor type (return or new), a Boolean value indicating if the visit was on a weekend, and the month of the year.

II. Related Work

A data-level approach and feature selection approaches are proposed as a solution for the classifying of unbalanced data in [3]. One of the basic challenges in artificial intelligence, notably for classification in machine learning, is imbalance class categorization. Imbalanced data has been shown to degrade the performance of machine learning algorithms, where imbalance data refers to the fact that the total data from each class differs substantially. The suggested approach is tested using a dataset from the UCI repository and the major testing metric is the area under the curve (AUC). In many cases, especially when the target group is a minority, the false positive rate (FPR) is set at an acceptable level, and the goal is to reduce the false negative rate (FNR) while keeping the FPR below a certain level. Various techniques have been offered to overcome this problem, in addition to tampering with the performance statistic. Algorithmic level techniques, data level techniques, ensemble classifications, and cost-sensitive techniques are the four categories of methods (a mixture of algorithmic and data level techniques).In order to overcome the problem of skewed class distribution in the learning phase, data preparation techniques adjust the data distribution. The algorithmic modification techniques change current algorithms to make minorities more significant. To deliver varying misclassification costs for each class in the learning process, cost-sensitive techniques use both algorithm and data modification approaches. Finally, to address the imbalance problem, ensemble of classifier sampling methods modify the ensemble learning algorithm, but they usually do not change the base classifier [4]. [3] claims that combining data preparation approaches with an ensemble classifier outperforms other methods. Before the model training step, the preprocessing data approach involves resampling uneven training data sets. Although no one strategy has been shown to perform effectively for all unbalanced dataset situations, sampling methods have shown tremendous promise in that they aim to enhance the dataset rather than the classifier. By either oversampling the minority samples or undersampling the majority samples, sampling procedures alter the distribution of each class observation. In the case of oversampling, sampling procedures create additional minority instances to balance the dataset, but in the case of undersampling, sampling methods eliminate some majority instances. Because the removal of majority instances may remove essential information from the dataset, undersampling approaches have been found to be less efficient than oversampling methods, especially when the dataset is small. Random sampling is the most basic way of oversampling. It chooses a minority instance at random and replicates it until the minority class has grown to the appropriate size. Over-fitting occurs when random oversampling creates new instances that are highly close to the original instances.To solve this problem, the Synthetic Minority Oversampling Technique (SMOTE) is utilized, in which new synthetic instances are created by generating new synthetic instances from randomly picked minority instances and their NN-nearest neighbors, where NN is a user-defined variable. However, because new instances are created without taking into account the majority instances, this may result in overgeneralization, increasing the overlap between minority and majority classes.When the dataset has a higher imbalance ratio, overgeneralization can be magnified because the minority class

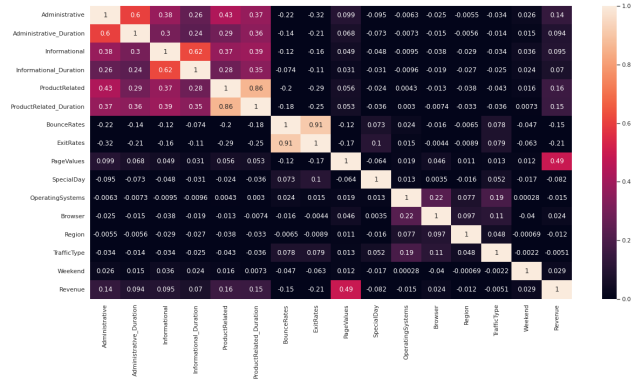

Fig. 1. Correlation Heatmap

instances are very limited and can become contained within the majority class after oversampling. This can make future classification performance much worse.

III. Our Solution

A. Descriptive Analysis of Data set

We make no prior assumptions about the user's purchase intent, do not analyze previous behavior and trends, and learn entirely from the session dynamics whether it will finish with a purchase by considering each session as an anonymous separate session.

B. Machine Learning Algorithms

This subsection describes machine learning algorithms that you plan to use. For each ML algorithm, briefly 1) explain why it might be appropriate for the problem and 2) describe your main design. For example, if it is neural network, provide the network structure and your initial choice of some key parameters (e.g., activation function to use, number of layers, number of hidden nodes of each layer). You may change the parameters during the training process.

C. Implementation Details

This subsection describes details of your implementation. Please focus on how you test and validate the performance, tune the hyperparameters, and select the best-performing models. Elaborate on techniques that you apply to improve the performance and explain why you use these techniques. You include few most important results/figures to illustrate your idea but do not let figures/tables dominate the content of the report. You can include few lines of critical code if needed. But please avoid paste lengthy code in your report. Please make sure the figures/tables/code snapshots are of appropriate size including the font size.

IV. Comparison

This section includes the following: 1) comparing the performance of different machine learning algorithms that you used, and 2) comparing the performance of your algorithms with existing solutions if any. Please provide insights to reason about why this algorithm is better/worse than another one.

V. Future Directions

This section lays out some potential directions for further improving the performance. You can image what you may do if you were given extra 3-6 months.

VI. Conclusion

This section summarizes this project, i.e., by the extensive experiments and analysis, do you think the problem is solved well? which algorithm(s) might be better suitable for this problem? Which technique(s) may help further improve the performance?

Last but not the least, don't forget to include references to any work you mentioned in the report.

References

[1] M. K. . Y. K. C. Okan Sakar, S. Olcay Polat, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks." *Neural Computing and Applications*.

[2] D. Dua and C. Graff, "UCI machine learning repository."

[3] I. Kurniawan, A. Abdussomad, M. Akbar, D. Saepudin, M. Azis, and M. Tabrani, "Improving the effectiveness of classification using the data level approach and feature selection techniques in online shoppers purchasing intention prediction," *Journal of Physics: Conference Series*, vol. 1641, p. 012083, 11 2020.

[4] I. Nekooeimehr, "Oversampling methods for imbalanced dataset classification and their application to gynecological disorder diagnosis," 2016.