**Online Shoppers Intention**

Isaac Cortes

Ryan Lertpichitkul

Samyak Jain

Abraham David Ungos

Jack Hannum

California State University Long Beach

CECS 456 Machine Learning Sec 02

Mahshid Fardad

December 3, 2022

**Table of Contents**

**1 Introduction**………. …………………………………………………………………………...**2**

**2 Data**………………..……………………………………………………………………………**3**

**2.1 Dataset Description**……..……………………………………………………………**3**

**2.2 Variable Analysis**…....……………………..…………………………………………**4**

**3 Methodology**……………………………………………………………………………………**9**

**3.1 Principal Component Analysis.**……………………………………………………**11**

**3.2 K-nearest Neighbor Clustering Algorithm**.……………………………………….**14**

**4 Results**…………………………………………………………………………………………**14**

**4.1 Predicting with Principal Component Analysis**…….…………………………….**14**

**4.2 Predicting without Principal Component Analysis**.…..………………………….**16**

**4.3 Resultant Model**.…..………………………………………………………………..**17**

**5 Conclusion**…………………………………………………………………………………….**18**

**References**……………………………………………………………………………………….**19**

**1 Introduction**

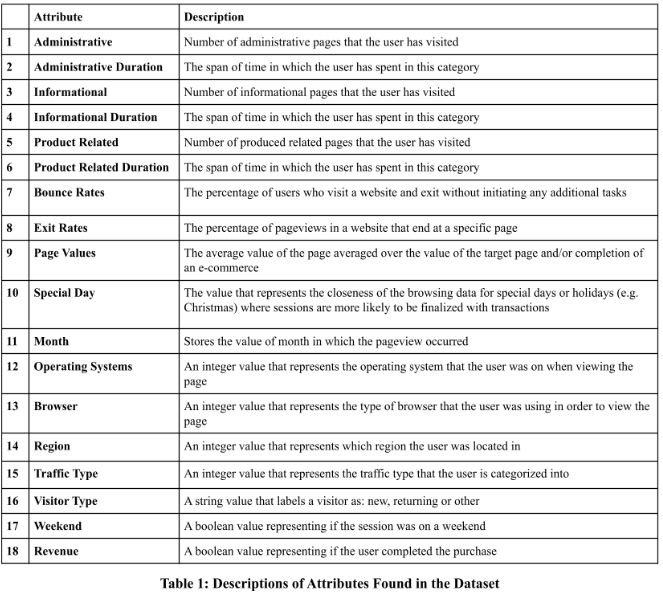
In the last couple of years, online e-commerce shopping has increased as more and more shoppers browse online to purchase things that they want. This is evident especially during the holidays such as Cyber Monday online shopping. Oftentimes, the intent of the shopper who is browsing websites such as Amazon or eBay will choose to shop on these websites instead of physical retail stores due to the very quick and convenient nature of buying items online. The nature of e-commerce makes shopping easier since shoppers have the ability to compare prices and bargain for the cheapest items when surfing multiple e-commerce websites. There are various types of online shoppers that range from low to high intent. The intent of a shopper is their behavior in making a purchase on a website. A shopper has two main behaviors: stay on the website and make a purchase or abandon the site. Based on these behaviors, e-commerce stores can collect and analyze shopper’s intentions to improve their online shopping user experience.

It is very important for e-commerce websites to study their shopper’s behaviors since they want more consumers to visit and purchase items from their website. If these online businesses are able to gather this data and analyze it, then they would be more aware of how they can market their website to cater towards their users (Baati, K. et al, 2020). In this research, we will take a look at a dataset provided by the UCI Machine Learning Repository. Using the dataset, we will use clustering and classification algorithms to make a predictive model around the shoppers’ intentions thus leading to a better understanding of customers visiting online stores. Lastly, we will conclude the paper by observing the results and comparing it to the predictions of the data.

**2 Data**

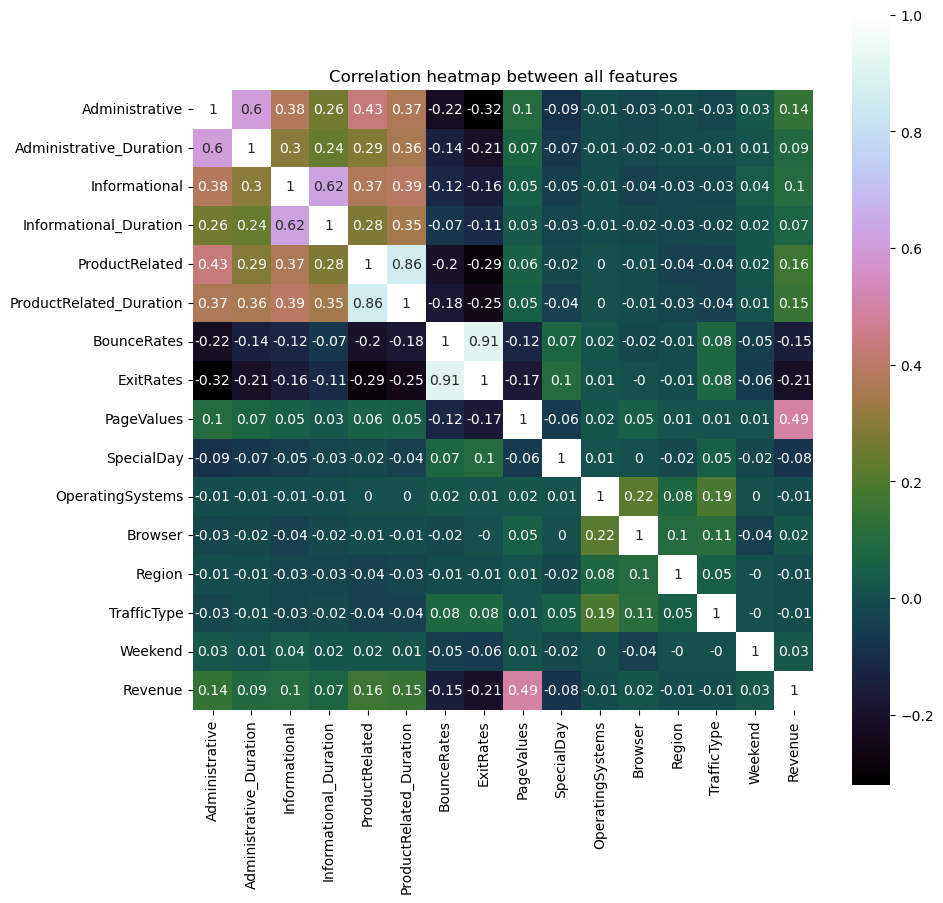
**2.1 Dataset Description**

In our study, we utilized the Online Shoppers Purchasing Intention dataset retrieved from the UCI database which contains data that recorded 12,330 sessions of users who shopped online (Sakar et al). In our analysis, we first cleaned the dataset. First, we checked if the data set contained any values that were undefined or missing across all columns. We found that there was no missing data from any of the columns. In addition, we uncovered 18 attributes as shown below in Table 1.



**2.2 Variable Analysis**

In the next step, we analyzed the relationships between the attributes or features by using a correlation heatmap. With this heatmap, we can visualize the association of relationships between all attributes. In other words, it will help us understand which variables are related to each other and how strong their relationship is. Figure 1 shows the correlation heatmap between all features of this dataset.



**Figure 1: Correlation Heatmap Between All Features**

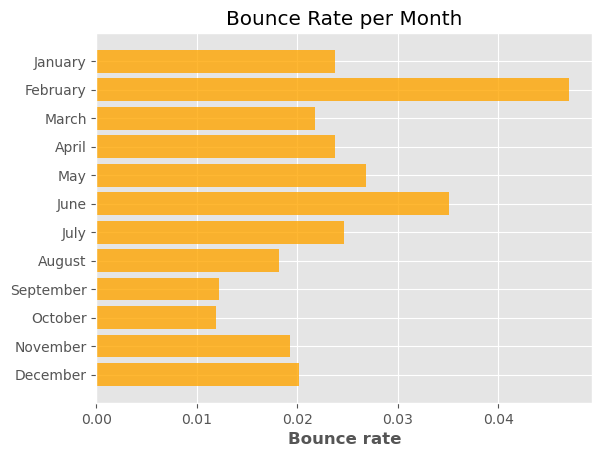
The heat map shows that attributes that have a higher relationship with another attribute are shown in a lighter color while attributes that have a darker color will result in a lower relationship.

Next, the dataset was analyzed by looking into the bounce rates of each month. Table 2 shows the numerical value of bounce rates that corresponds to each month. Figure 2 displays a visualization of the bounce rate per month. By observing the result of this data, we can conclude that February had the highest bounce rate. We assume that this is due to the fact that Valentine’s Day takes place in February. A lot of online shoppers are probably scrambling to find a perfect gift for their loved ones. We predict that this is what caused February to have such a high bounce rate since online shoppers would hop from one website to another trying to find the right item to purchase.



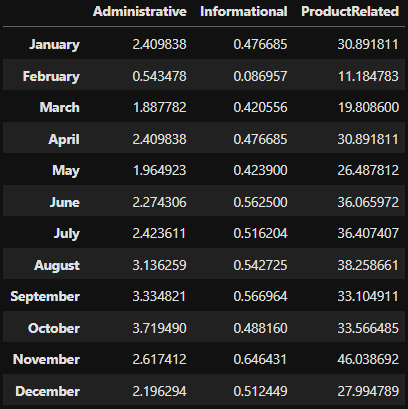
**Table 2: Numerical Value of Bounce Rate Per Month**

**Figure 2: Bounce Rate Per Month**



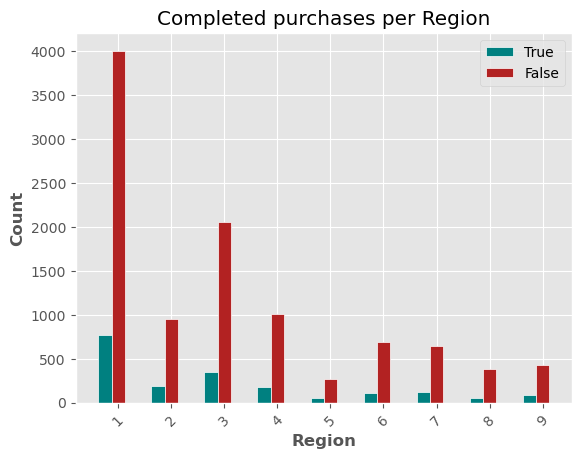
In the next step in our data analysis, we analyzed the number of visits for the type of each session: administrative, informational, or product related when grouping the data by each month.

Table 3 shows a numerical representation of the type of visits by month and type while Figure 3 visually displays the values separated by each month. According to our observations, we concluded that in this dataset the highest type of visits of the sessions are Product Related. The second highest recorded value is administrative and the lowest value counted was informational. This makes sense since a vast majority of online shoppers' intention is to buy a product that they are looking for. As you can see in Figure 3, the Product Related visits are more apparent in the second half of the year (July to November). We predict this is due to the fact that many online shoppers spend their time looking for early Christmas gifts before everything gets sold out.

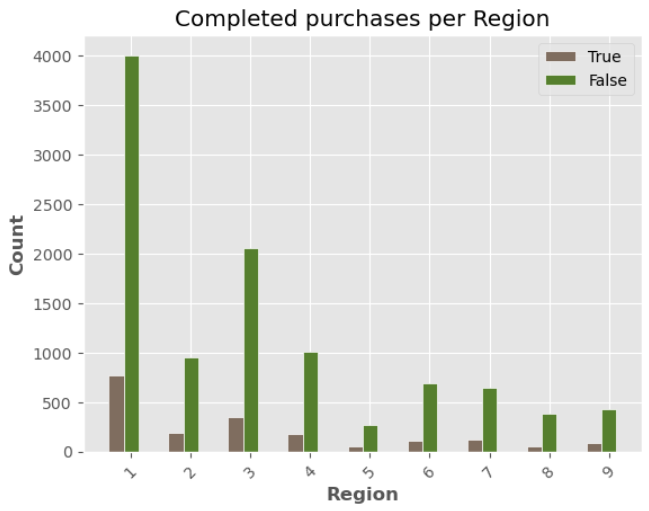
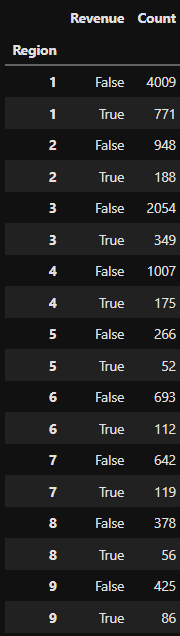


**Table 3: Numerical Values of Visits by Month and Type**

**Figure 3: Visits by Month and Type**



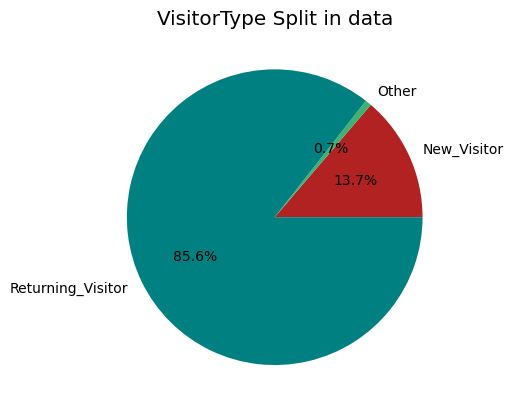
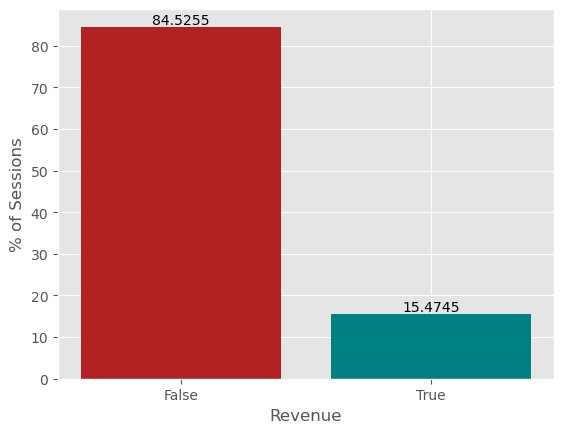
In addition to analyzing the dataset, we wanted to inspect the revenue per region. In this dataset, there are a total of 9 different regions. Table 4 displays the numerical representation of the revenue by region. Figure 4 visually displays the revenue by region. By observing the results, we can determine which region has a higher rate of not buying an item in comparison to buying an item from an online store. Region 1 has the highest count in both true/false values in revenue. We assume that this region is more densely populated in comparison to the other regions. In addition, it shows that at most of the visits end in leaving the website without the user actually buying something.



**Table 4: Numerical Values of Revenue by Region**

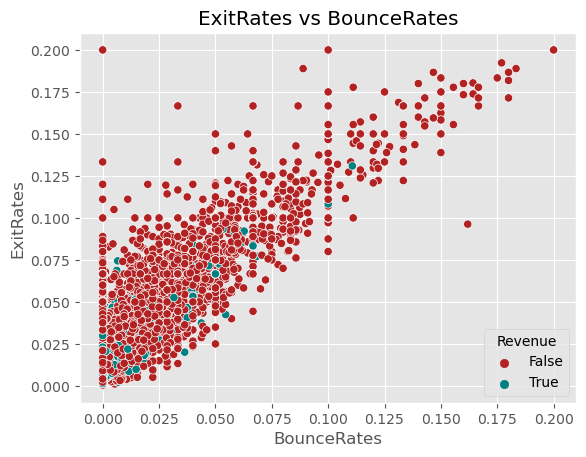
**Figure 4: Revenue by Region**

In the final step, we have further analyzed the data by looking at the revenue, visitor, and exit and bounce rates shown in Figures 5-7. There exists about 85% false revenue and 15% of the revenue is true. We also found that 85% of the data are returning visitors. In addition, we can see by the graph below that the exit rates versus bounce rates have a lot of false revenue data in comparison to revenue data that is true. As a result, we can see that this data set is very imbalanced. A majority of the data points did not actually create revenue and was false.



**Figure 5: Revenue Analysis**

**Figure 6: Visitor Type Data**

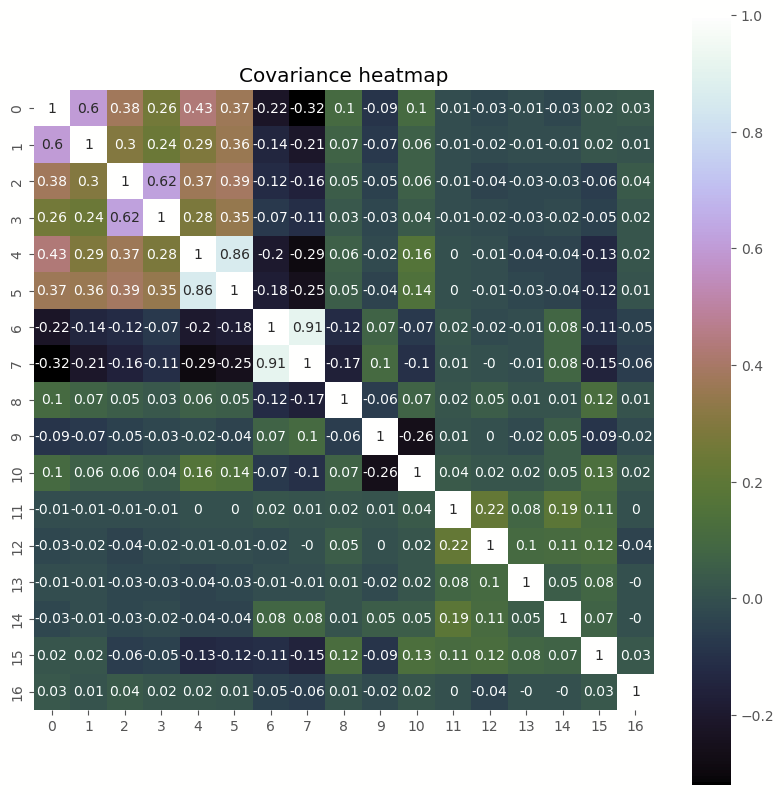


**Figure 7: Exit Rates Versus Bounce Rates**

In our data analysis, we discovered various attributes or features that exist within the dataset. The importance of this analysis is to get an idea of the composition of the data. We found out that some attributes have a close correlation with another while some do not. In addition, we observed the behavior of online shoppers by grouping several attributes together. We found that certain months had a higher bounce rate in which a user would visit a website or an online store without actually doing anything such as making a purchase. In addition, we found out that a majority of the sessions are classified as Product Related visits. Many users tend to shop online due to product related matters possibly due to holiday shopping or ordering necessities. Finally, we inspected the revenue by region which observes user data on their region location where they either do or do not make a purchase. Our data analysis serves an important role in helping e-commerce websites. If we analyze an online shopper’s intentions then we can predict how a user will interact when visiting a specific website. For example, Amazon can put up Christmas gifts such as toys during the months with higher Product Related visits. In addition, online stores such as Flower shops can put up flowers and chocolates to attract users to their websites for Valentine’s Day gift shopping.

**3 Methodology**

One of the first things we did to eliminate bias and make the graphs and charts more accurate was to standardize the data. This is taking all the data and attributes and resizing them so they have an average of 0 with a variance of 1. This would eliminate improper importance on certain features that are normally in a range of the 100s and another that is in the range of 10s. An example shown is that there are 9 regions, thousands of bounces, and only 2 options for if a sale was made or not. Without data standardization, the algorithm would not be able to work properly and it would give inaccurate results. After the data standardization we did another analysis of the relationship between them to make sure that it would still be closely related as before. Figure 5 is shown below that shows a covariance heat map that represents the relationship after the data standardizations.



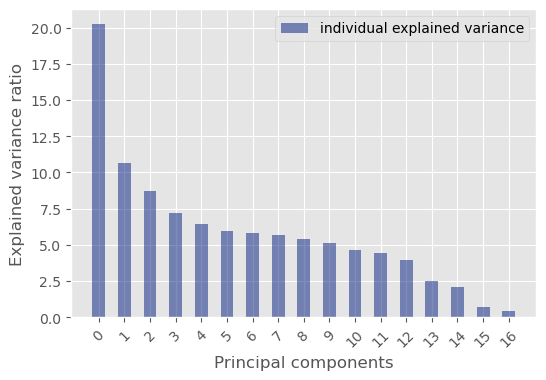
**Figure 5: Data Standardization**

As you can see the data is very similar to the heatmap from above and this shows that even after the standardization, the relationship between the data is still similar. This helps keep the calculations simple and avoid biases towards certain numbers than others, because of their varying range or size.

For this project we had to look into the 18 different variables to analyze if a purchase was going to be made on an ecommerce website or not. One of the biggest reasons we did a lot of variable analysis is that there are 12,330 different instances each with18 variables so we need to find some way to minimize the complexity of the calculations without losing significant data. One of the ways this is accomplished is through a method called Principal Component Analysis. For this we were able to eliminate some of the features and conduct faster calculations while staying accurate. Just to be sure, we then compared the data cluster calculation using both the original data and the newly minimized data. Now we discuss in further details about what the exact machine learning algorithms we implemented.

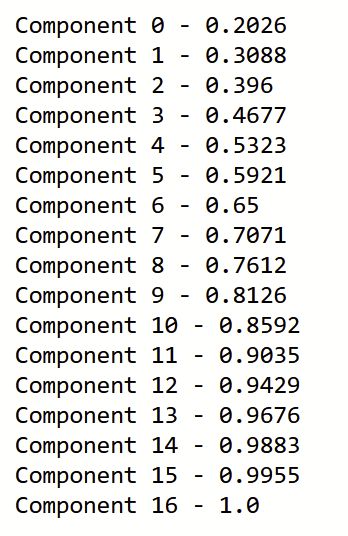
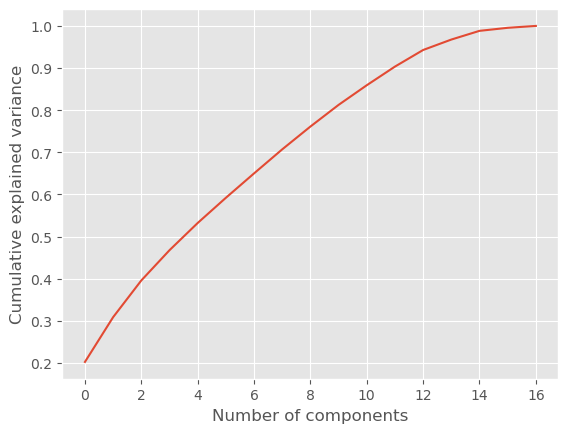
**3.1 Principal Component Analysis**

For Principal component analysis we first had to order the data and look at which values had the strongest insight on the data in comparison to others. To do this we took the individual explained variance of each component to understand how it impacts the data and the quality of information it offered. Figure 6 is shown below as a graph that shows the ratios for these component numbers.



**Figure 6: Explained Variance and Principal Components**

Seeing as the graph slowly curves down in descending order, this would make it easier to determine the number of features we would need to keep. Now in order to know when to draw the line and know when we have enough variables, we need the cumulative explained variance. This would give us insight as to how much of the picture we get as we add more and more components. Figure 7 displays a graph and a chart demonstrating this.



**Figure 7: Reducing the Number of Components**

Over here we see that as the components are added we get closer to the whole picture. The curve of this graph also shows us that the most important components are being added first and the least one at the very end. After having the first 10 components we cover about 85% of the data representation. We then conducted the principal component analysis with only the first 10 components. Now our calculations will cover 85% of the data while eliminating almost half of the variable needed.

**3.2 K-nearest Neighbor Clustering Algorithm**

The main algorithm we will run is called K-Nearest Neighbor. This is the algorithm that determines K clusters and based on any new points, it will determine based on how close its surrounding points are, which cluster it would be a part of. To run KNN we imported a python library called “sklearn” and fed our data into some of their methods. We used their k-neighbor classifier to identify the clusters that exist within the data set and then we fit it using the training and labeling data. We then predicted the output from the testing data.

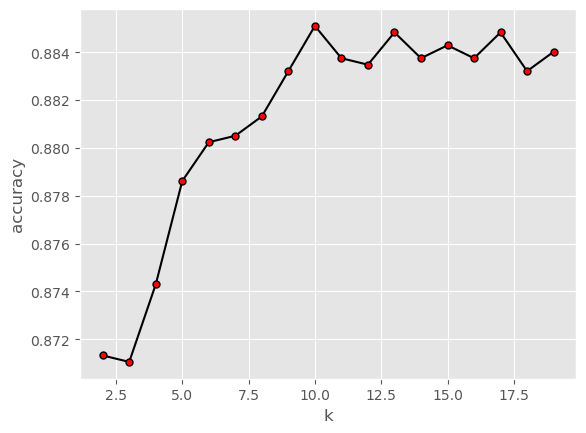
We ran the data for both the standardized 18 features and the minimized data from PCA with the number of k-clusters ranging from 2 to 20. After running the algorithms, we tested their accuracy by taking the average of the number of correct labels over total tests for each k cluster amount.

**4 Results**

**4.1 Predicting with Principal Component Analysis**

For the results of this algorithm, we were able to graph the accuracy of the data set based on the number of K clusters we had determined and whether or not the model has PCA implemented. We compared both models by looking at their overall accuracies as the number of k clusters increases. Figure 8 below shows the accuracy and k-cluster chart for the PCA dataset.

**Figure 8: Predicting with PCA**

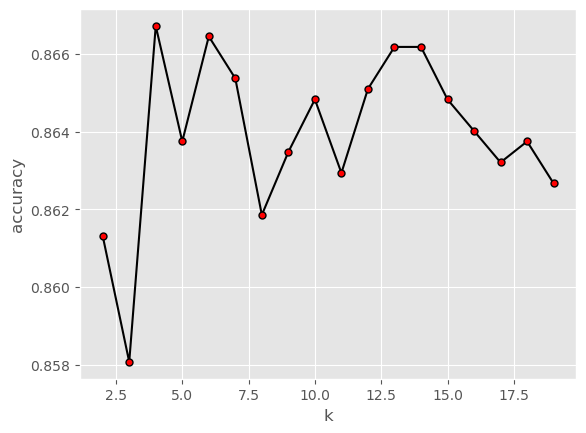


This graph shows that the greatest accuracy was achieved with about 9 clusters and it goes down, plateaus a little, then decreases in its accuracy. A bellhop shape is sort of expected because at one point you would define more clusters than there should be and the accuracy would go down. This graph seems to overall maintain a higher level of accuracy compared to the one from the standardized data.

**4.2 Predicting without Principal Component Analysis**

Figure 9 below shows that the graph has a lower accuracy overall and is more inconsistent. It contains far more local minimums and maximums and its overall shape is more jagged.

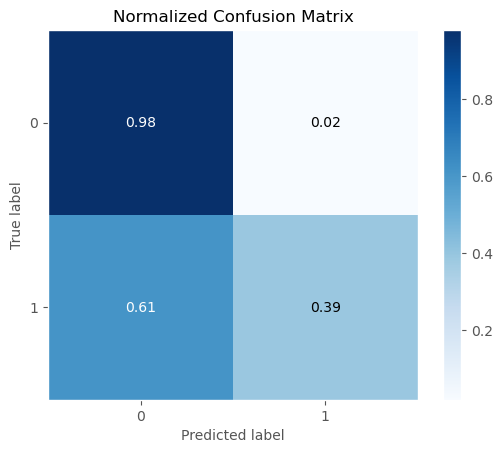
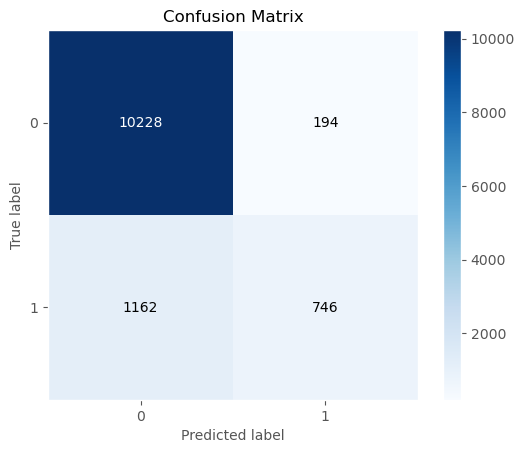
**Figure 9: Predicting without PCA**



The PCA data set allows KNN to be more accurately identified while being less complex and covering 85% of the data representation. It identified the best number of clusters to be about 9 and the accuracy for properly labeling the correct clusters is about 0.879.

**4.3 Resultant Model**

As shown in Figure 10 below, our model may result in high accuracy and this shows after seeing all the predictions and how most of the results were false. This is due to the imbalance of data present in this dataset where most of the revenue was false in comparison to true data.



**Figure 10: Confusion Matrix and Normalized Confusion Matrix**

**5 Conclusion**

For this project, we analyzed the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository. We examined the data in many different ways to study the relationship between the attributes/variables and the impact they have on the data set as a whole. This is important as it allows us to determine and set up the proper explained variance individually and collectively to determine Principal Component Analysis (PCA). PCA allows us to have fewer variables while maintaining a high representation of the data. In our study, we have constructed two models for the k-nearest neighbors algorithm. One had PCA implemented while the other did not. After observing the differences in the graphs, we can determine that the model with PCA implemented had a higher accuracy in comparison to the one that did not. It was even more accurate with identifying KNN labels and its accuracy graph was overall more consistent and accurate. In addition, our resultant model has shown that it had great accuracy, this is apparent since a lot of the data had false revenue. The imbalanced data between true and false revenue was shown in our confusion matrix. This can play a role if we were to deploy the model on new data and how accurate it will be when categorizing data. For future studies, we hope to try and explore more data and perform more analysis and other methods in order to get more accurate results. Overall, we hope we can discover more about user’s and their decisions in buying an item. As a result, this research as well as similar research conducted can help better understand the behaviors of users and can help gain further knowledge of how to improve user experiences when visiting online websites and stores.

# 

# **References**

1. Baati, K., Mohsil, M. (2020). Real-Time Prediction of Online Shoppers’ Purchasing Intention Using Random Forest. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1\_4
2. C. Okan Sakar, Yomi Kastro. Online Shoppers Intention Dataset Data Set. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
3. Sakar, Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing & Applications, 31(10), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0