Report Cover

Title: Data Analysis Report
Author: Wenlao Peng
Affiliation: RMIT University
Contact details: s3690691@student.rmit.edu.au(Email), 0403999357(Mobile No)
Date: May 21, 2023

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Content

Abstract

The objective of this study is to investigate the factors that affect online shopping transactions. By analyzing the "Online Shoppers Purchasing Intention Dataset Data Set" dataset containing 12330 data items, we focused on transaction-related characteristics including 'PageValues', 'Weekend', 'Region', ' OperatingSystems', 'TrafficType', 'ProductRelated', and 'ProductRelated_Duration'. By examining the association of these characteristics with transaction outcomes, we explore their impact on transaction intent and purchase behavior. We used two different approaches, namely K-nearest neighbor (KNN) models and decision tree models, combined with cross-validation to build classification models and evaluate performance. By visually analyzing the relationships between attributes and comparing the performance of the two methods, we found that the decision tree model received higher scores, indicating its better performance in predicting online shopping transactions.

These findings have important implications for e-commerce platforms and online retailers to help them optimize their website design, improve user experience and develop more effective marketing strategies to increase transaction conversion rates and sales. However, there are some limitations to this study, including biases and limitations of the data set and the potential impact of other factors not considered on the findings. Future research can explore the factors influencing online shopping transactions in depth by using a larger dataset and more comprehensive feature analysis.

Introduction

The aim of this study is to investigate the factors that influence transactions during online shopping. Specifically, we focus on the following research questions:

1. Which features have a significant impact on users' purchase intention and buying behavior?
2. Are there specific time, region, or operating system types associated with transaction outcomes?
3. Which features have higher predictive power in determining whether a user will make a purchase?

To address these objectives, we will employ various methods and techniques. Firstly, we will utilize the K-Nearest Neighbors (KNN) algorithm for classification prediction, which involves determining the class of an unknown sample based on the labels of its nearest neighbors. Secondly, we will apply the decision tree algorithm, which partitions data based on feature conditions, allowing for classification and regression analysis. In order to evaluate the performance and generalization ability of our models, we will employ cross-validation, a technique that involves splitting the dataset into training and testing sets for model evaluation. Additionally, we will perform data preprocessing, including data cleaning, feature scaling, and feature selection, to enhance data quality and model performance.

Our study will be based on the "Online Shoppers Purchasing Intention Dataset Data Set", where we will analyze features related to transactions such as 'PageValues', 'Weekend', 'Region', 'OperatingSystems', 'TrafficType', 'ProductRelated', and 'ProductRelated_Duration', and investigate their impact on buying behavior. By comparing the performance and effectiveness of different methods, our objective is to uncover the key factors influencing online shopping transactions, providing valuable insights for the e-commerce industry to optimize user shopping experiences and enhance transaction outcomes.

Through this research, we aim to provide targeted recommendations for e-commerce platforms and online retailers, assisting them in formulating more effective marketing strategies to improve conversion rates and sales performance.

Methodology

Throughout my research, I used the python language. python's ease of use and practicality helped me a lot in performing data analysis. When faced with a freshly acquired dataset, I needed to set up a goal for the project based on the characteristics of the dataset. i needed to look at the general picture of the data, so I wrote code to look at the head data, tail data, overall information about the data, and descriptive statistics about the data.

```
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
None
```

Then I browsed the website about the source of the dataset to get the information about the dataset and the specific meaning of the attributes.

The following are the descriptions of the attributes:

1. Administrative: Number of administrative pages visited by the user in the session.
2. Administrative_Duration: Total time spent on administrative pages during the session.
3. Informational: Number of informational pages visited by the user in the session.
4. Informational_Duration: Total time spent on informational pages during the session.
5. ProductRelated: Number of product-related pages visited by the user in the session.
6. ProductRelated_Duration: Total time spent on product-related pages during the session.
7. BounceRates: Percentage of visitors who enter the site and then leave ("bounce") without triggering any other requests to the analytics server during the session.
8. ExitRates: For all pageviews to a specific web page, the percentage that was the last in the session.
9. PageValues: The average value for a web page that a user visited before completing an e-commerce transaction.
10. SpecialDay: Closeness of the site visiting time to a specific special day (e.g., Mother's Day, Valentine's Day) when sessions are more likely to be finalized with a transaction.
11. Month: Month of the year the visit took place.
12. OperatingSystems: The operating system used by the visitor.
13. Browser: The browser used by the visitor.
14. Region: The geographic region from which the visitor accessed the website.
15. TrafficType: The type of traffic that brought the visitor to the website (e.g., direct, referral, organic search, paid search, social media).
16. VisitorType: Type of visitor, either new or returning.
17. Weekend: Boolean value indicating whether the visit date is a weekend.
18. Revenue: Whether the user made a purchase during the session (class label).

The 10 most likely interconnections:

1. Administrative and Administrative_Duration: As the number of administrative pages visited increases, the time spent on those pages should also increase.
2. Informational and Informational_Duration: As the number of informational pages visited increases, the time spent on those pages should also increase.
3. ProductRelated and ProductRelated_Duration: As the number of product-related pages visited increases, the time spent on those pages should also increase.
4. BounceRates and ExitRates: Pages with high Bounce Rates are more likely to have high Exit Rates, as visitors are leaving after viewing just one page.
5. PageValues and Revenue: Higher Page Values should correlate with a higher likelihood of revenue generation, as visitors who view valuable pages are more likely to make purchases.
6. SpecialDay and Revenue: The closer the visit is to a special day, the higher the likelihood of a purchase, as people tend to buy gifts for special occasions.
7.VisitorType and Revenue: Returning visitors have a higher likelihood of making a purchase compared to new visitors, potentially due to their familiarity with the website and its offerings.
8. TrafficType and Revenue: Different traffic types may have different conversion rates, with some sources potentially leading to more purchases.
9. VisitorType and Revenue: Returning visitors may be more likely to make a purchase than new visitors, as they are already familiar with the website and its offerings.
10. Weekend and Revenue: Shopping behavior might differ between weekdays and weekends, with weekends potentially seeing more purchases.

Then, I had the goal of the project and started **"Retrieving and Preparing the Data"**. In this session, I focused on data cleanup and data conversion.

Data cleanup: I looked at missing values to determine the integrity of the data, and I used visual box plots to check for outliers in Administrative, Administrative_Duration, Informational, Informational_Duration, because the number of outliers was greater than 5 percent of the data. percent, so I replaced the outliers with the median, but I set up the prerequisite that only data with non-zero values would be processed and calculated. Since data with a value of 0 means in the study that the user did not view the page, it makes sense for it to exist but there is no need for them to participate in the median calculation. I replaced the outliers of 'ProductRelated', 'ProductRelated_Duration' with the median.

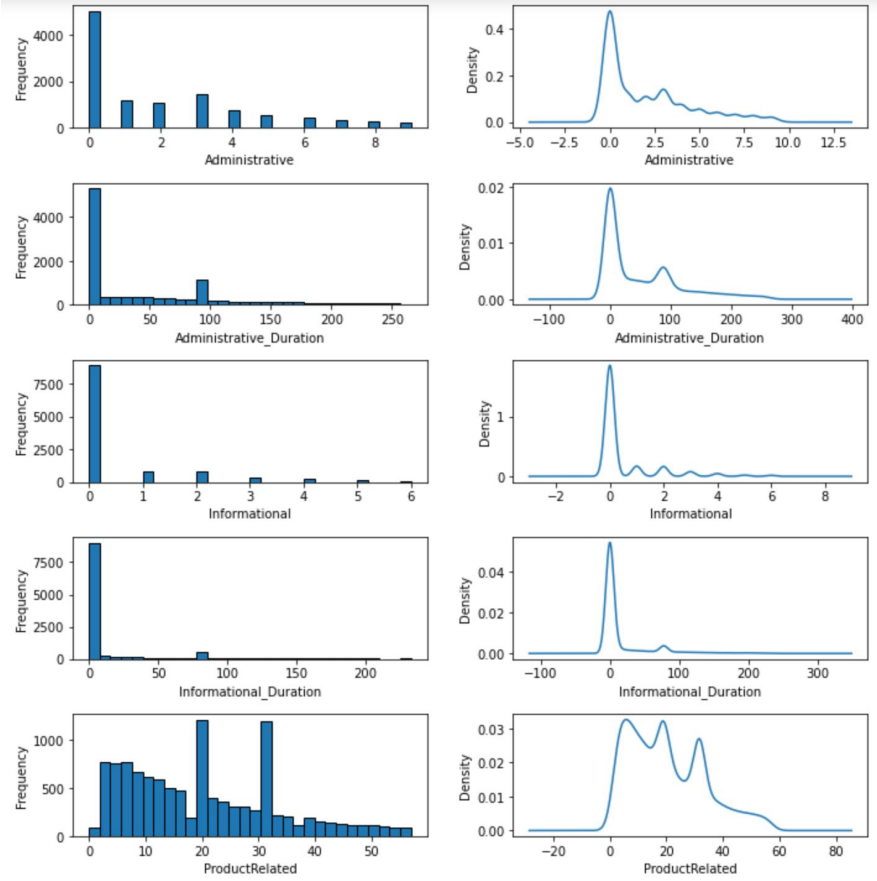
Data conversion: I converted the data types of the "Weekend" and "Revenue" columns in the "data" dataset to integer types to facilitate subsequent data processing and analysis.

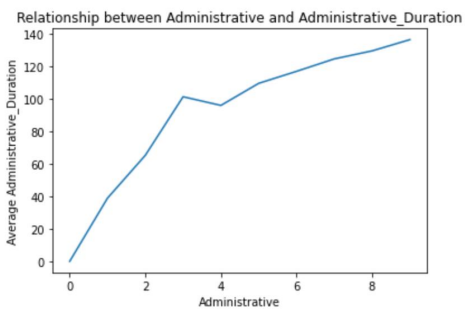
```
Dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
None
```

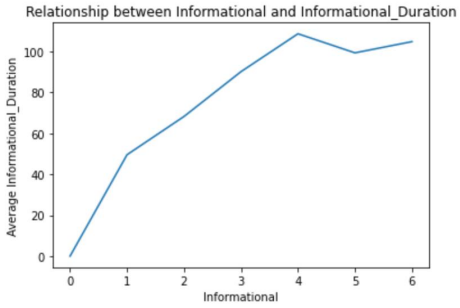
Once I had a clean data set I was able to do **Data Exploration:**

I used box plots, histograms, kernel density estimates, line plots and histograms to visualize the data for a better view.

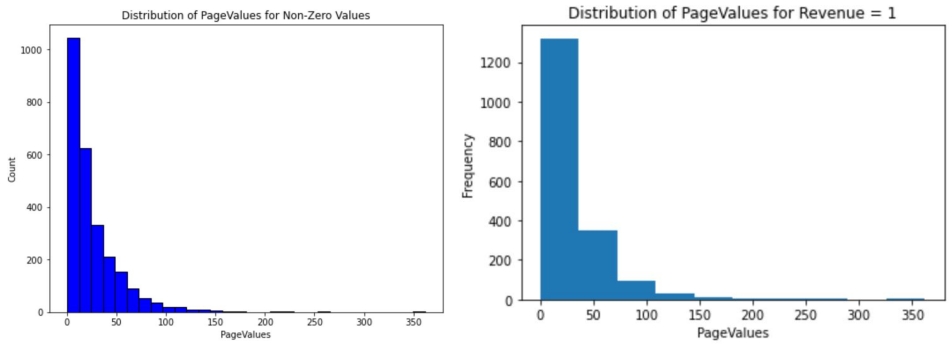
As you can see from the graphs below, most users do not view many specific pages, and the number of pages viewed is positively correlated with the amount of time spent. However, most users view 20-32 pages about a product.



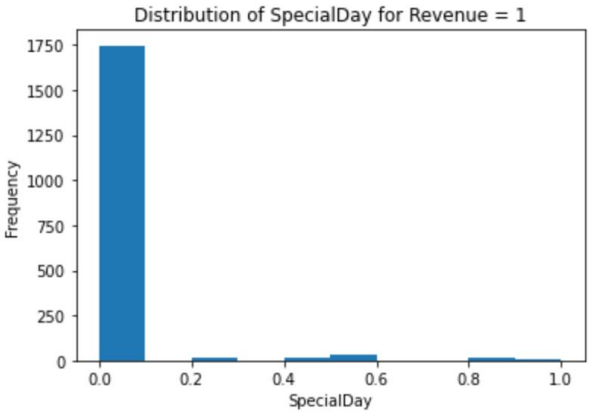




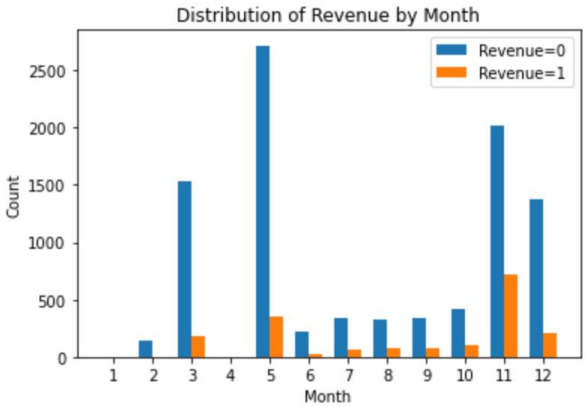
The graph below shows that the average number of pages visited before completing a transaction is concentrated between 0-50 pages, with the maximum number of pages being 10. This indicates that the page value of the site is average and that users are likely to need to view 10-50 pages on this site before making a purchase.



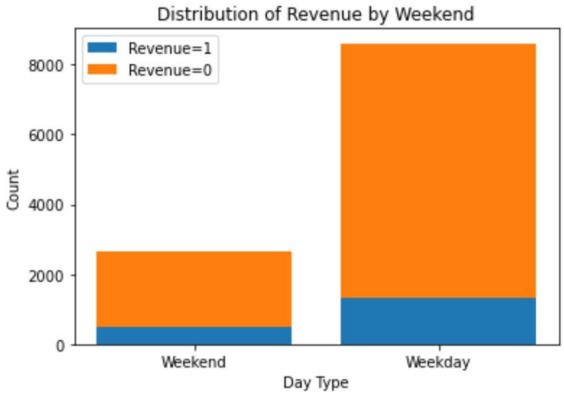
From the following graph, we observe that very few users shop on this site on special days, so there is no correlation between users' transactions and special days.



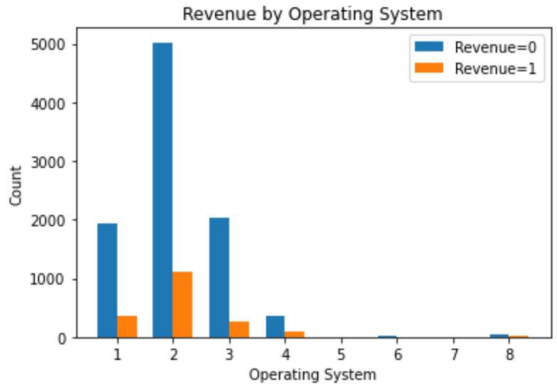
As we can see from the graph below, the number of visits to the site each month is proportional to the number of transactions, with the majority of transactions occurring in March, May, November and December.



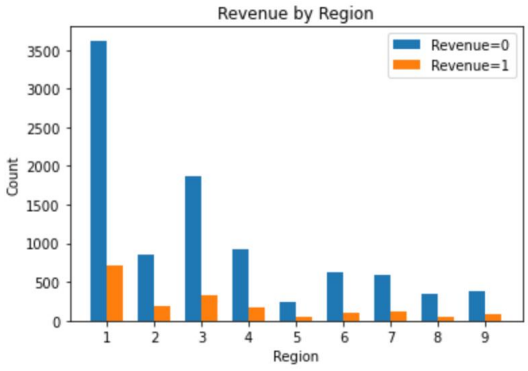
In this figure below we can see that a large number of customers visit the site and complete transactions on weekdays, so the conclusion is that users usually visit the site and complete their purchases on weekdays.



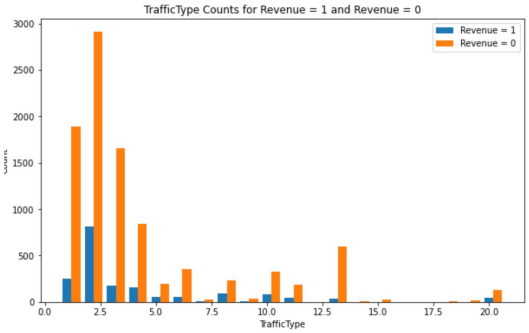
As you can see from the figure below, most of the users will be using 1, 2, 3, and 4 operating systems to browse the site and most of the transactions will be conducted on these systems, which allows the site to explore deeper business value.



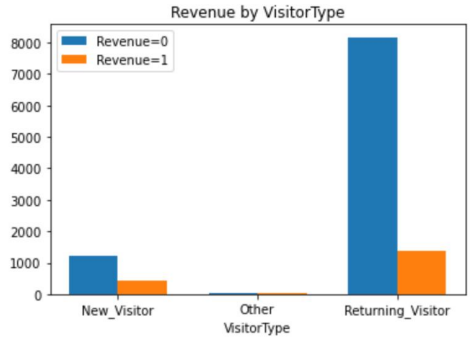
As you can see from the graph below, the users browsing this site are mainly located in areas 1, 2, 3 and 4. Most of the transactions also take place in these areas, so if the company wants to increase the number of transactions, it can try marketing campaigns for areas 1, 2, 3 and 4. However, we do not rule out the possibility of logistical factors.

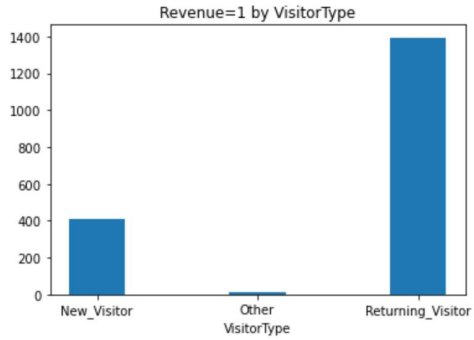


As you can see from the chart below, customer visits and transactions are more likely to be influenced by traffic types 1 to 4.



As you can see from the chart below, there are more visits and transactions from repeat customers than new users, so the main transactions and visits to this site come from repeat customers.





Once I had a clear view and understanding of the attributes, I started to work on **Data Modelling:**

I chose to use classification methods to model the data. I have two combinations, the first one is: K Nearest Neighbor (KNN) algorithm for classification prediction and cross-validation. It is used to solve the classification problem and classify the attributes as whether they have an impact on the transaction or not. The second one is: using decision tree algorithm and cross-validation. Prediction of revenue categories (with and without) based on known factors.

Results

```
Cross-Validation Scores: [0.88326676 0.86856128 0.86722913 0.84902309 0.85257549]
Average Score: 0.8641311486007603
Cross-Validation Scores: [0.8996893  0.88055062 0.86634103 0.85213144 0.86012433]
Average Score: 0.8717673455325798
Cross-Validation Scores: [0.88948069 0.87655417 0.86634103 0.84236234 0.86634103]
Average Score: 0.8682158542901945
Cross-Validation Scores: [0.90501553 0.88454707 0.86856128 0.85479574 0.86989343]
Average Score: 0.876562609632785
Cross-Validation Scores: [0.89480692 0.87566607 0.87033748 0.85257549 0.87078153]
Average Score: 0.8728334984969714
Cross-Validation Scores: [0.90501553 0.88543517 0.87078153 0.85657194 0.87122558]
Average Score: 0.8778059488867813
Cross-Validation Scores: [0.89702619 0.88365897 0.87033748 0.85346359 0.87033748]
Average Score: 0.8749647401254613
Cross-Validation Scores: [0.90368398 0.89076377 0.86989343 0.85612789 0.86811723]
Average Score: 0.8777172571956555
Cross-Validation Scores: [0.90013316 0.88676732 0.87211368 0.85035524 0.86811723]
Average Score: 0.8754973238760397
Cross-Validation Scores: [0.90235242 0.88854352 0.87166963 0.85301954 0.86944938]
Average Score: 0.8770068958775313
Cross-Validation Scores: [0.89525078 0.88587922 0.87211368 0.85346359 0.86500888]
Average Score: 0.8743432281725806
Cross-Validation Scores: [0.8996893  0.88587922 0.87522202 0.85834813 0.86589698]
Average Score: 0.8770071323887076
Cross-Validation Scores: [0.89436307 0.88587922 0.87166963 0.85657194 0.86545293]
Average Score: 0.8747873567432096
Cross-Validation Scores: [0.89791389 0.88365897 0.86989343 0.85568384 0.86412078]
Average Score: 0.8742541817146903
Cross-Validation Scores: [0.89569463 0.88587922 0.87211368 0.85479574 0.86589698]
Average Score: 0.8748760484343354
Cross-Validation Scores: [0.89791389 0.88543517 0.87122558 0.85435169 0.86545293]
Average Score: 0.8748758513416884
Cross-Validation Scores: [0.89525078 0.88454707 0.87211368 0.85657194 0.86545293]
Average Score: 0.8747872779061507
Cross-Validation Scores: [0.89613848 0.88543517 0.87122558 0.85790409 0.86234458]
Average Score: 0.874609579175664
Highest score (0.8778059488867813, 7)
```

The highest score obtained for the KNN model is (k = 7): 0.8778

```
Cross-Validation Scores: [0.91788726 0.89653641 0.86278863 0.85479574 0.86944938]
Average Score: 0.880291484257422
Cross-Validation Scores: [0.91344874 0.89653641 0.86545293 0.85923623 0.87033748]
Average Score: 0.8810023580164282
Cross-Validation Scores: [0.91078562 0.89786856 0.86722913 0.85790409 0.86589698]
Average Score: 0.8799368751670359
Cross-Validation Scores: [0.91433644 0.89476021 0.86545293 0.86012433 0.86989343]
Average Score: 0.8809134692326552
Cross-Validation Scores: [0.90368398 0.88898757 0.86145648 0.85790409 0.86634103]
Average Score: 0.8756746284212327
Cross-Validation Scores: [0.90412783 0.89387211 0.86367673 0.85701599 0.87078153]
Average Score: 0.8778948376705541
Cross-Validation Scores: [0.90190857 0.88188277 0.85657194 0.85390764 0.86900533]
Average Score: 0.8726552479070732
Cross-Validation Scores: [0.89835775 0.88188277 0.85479574 0.84991119 0.86722913]
Average Score: 0.8704353145874576
Cross-Validation Scores: [0.89480692 0.87833037 0.85479574 0.84991119 0.86190053]
Average Score: 0.8679489514276998
Cross-Validation Scores: [0.90057701 0.86989343 0.85257549 0.84902309 0.85923623]
Average Score: 0.866261049999251
Cross-Validation Scores: [0.90412783 0.86145648 0.85035524 0.84902309 0.85879218]
Average Score: 0.8647509655568774
Cross-Validation Scores: [0.89835775 0.86234458 0.84724689 0.84813499 0.86234458]
Average Score: 0.8636857586371912
Cross-Validation Scores: [0.88948069 0.85612789 0.84591474 0.84369449 0.85612789]
Average Score: 0.858269140258223
Cross-Validation Scores: [0.88681758 0.85257549 0.8401421  0.8383659  0.84857904]
Average Score: 0.8532960197534134
Cross-Validation Scores: [0.87616511 0.85479574 0.84058615 0.83170515 0.84591474]
Average Score: 0.8498333778762716
Cross-Validation Scores: [0.87660897 0.85124334 0.83614565 0.82992895 0.83969805]
Average Score: 0.846724990322751
Cross-Validation Scores: [0.87083888 0.84946714 0.8250444  0.82637655 0.83436945]
Average Score: 0.8412192860673633
Highest score: (0.8810023580164282, 4)
```

The highest score obtained for the decision tree model is (max_depth=4): 0.8810

Discussion

Decision tree has higher average and highest scores than KNN. The higher scores of the decision tree model may indicate that it has better performance and accuracy in predicting transaction outcomes. This means that the decision tree model is able to identify the transaction outcomes more accurately when using the given features for classification prediction and may have better generalization ability. So I think choosing the decision tree model is the better choice.

Conclusion

This study aims to investigate the factors influencing online shopping transactions and explores transaction-related features by analyzing the "Online Shoppers Purchasing Intention Dataset Data Set" dataset. We used K-nearest neighbor (KNN) algorithm and decision tree algorithm for classification prediction and evaluated the model performance through cross-validation. The following are our main findings:

1. The decision tree model performs better in predicting online shopping transactions, with higher accuracy and performance scores. This indicates that the decision tree model is able to identify transaction outcomes more accurately and has better generalization capabilities.
2. The features 'PageValues', 'Weekend', 'Region', 'OperatingSystems', 'TrafficType', 'ProductRelated', and 'ProductRelated_Duration' have significant effects on users' purchase intention and purchase behavior .
3. Through visual analysis and relationships between attributes, we gain insight into the impact of these features on transactions.

These findings have important implications for e-commerce platforms and online retailers. By optimizing website design, improving user experience, and developing more effective marketing strategies, they can improve transaction conversion rates and sales performance.

However, there are some limitations to this study, including biases and limitations of the data set and other factors not considered that may have an impact on the results. Future research could use larger datasets and more comprehensive feature analysis to further deepen the understanding of online shopping transactions.

In summary, this study reveals key factors that influence online shopping transactions and demonstrates the strength of decision tree models in predicting transaction outcomes. These findings provide useful insights for the e-commerce industry and provide guidance and references for optimizing user shopping experience and improving transaction outcomes.

References

Decision Trees (Tu7.ipynb)

KNN (Week5-NumberClassification)

Outlier processing (in 1-3 weeks)

Comparing weeks of interest for different models (weeks 6 and 8)

Cross-validation https://scikit-learn.org/stable/modules/cross_validation.html

Set x, y labels https://stackoverflow.com/questions/21487329/add-x-and-y-labels-to-a-pandas-plot

Plot.bar https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.bar.html