Online Shoppers Purchasing Intention

Emily Zhang, Fuyang Lu, Lesley Yan, Michelle Wang

**2022-12-12**

Introduction

As internet penetration and accessibility increase, the number **of** digital buyers keeps climbing every **year**, **creating** tremendous **opportunities for** developing **more** useful analytics to drive business decisions. This **project explores** *Online Shoppers Purchasing Intention Dataset* which contains e**-**commerce user information. It **consists of 12330 rows where each row represents a session that belongs** 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** consists **of 10 numeric** variables **and** 8 categorical **variables**.

**The numeric variables are:**

•

**Administrative**: Number of pages **visited by** the **visitor about account management**

Administrative **Duration**: Total **amount** of time **(in seconds) spent by** the **visitor on account** management related pages

**Informational** : **Number of pages** visited by **the visitor about Web** site, communication and **address** information **of** the shopping site

**Informational Duration**: Total amount of time (in seconds**)** spent by **the visitor on** informational pages

**Product** Related: Number of **pages visited by visitor about product** related **pages**

**Product-**Related Duration: Total amount of time **(in seconds) spent** by the visitor on product **related**

**pages**

**Bounce Rate** : **Average bounce rate value of the pages visited** by **the visitor**

**Exit** Rate : Average exit rate value of the pages visited by the visitor

Page Value: **Average page value of the pages visited** by **the visitor**

**Special Day:** Closeness **of** the **site** visiting time to **a special** day

**The** categorical variables are:

•

Operating **system**: Operating system **of** the visitor

**Browser: Browser of the visitor**

**Region: Geographic** region **from** which the session has been started **by the** visitor

**Traffic type**: **Traffic source by** which the **visitor has arrived** at the **Web site (e.g.**, **banner, SMS, direct)**

**Visitor** type: **Whether** the **visitor is a New Visitor, a** Returning **Visitor or** Other

• **Weekend**: **Whether the date of the visit is weekend**

**Month** Month **of the visit**

**Revenue: Whether the visit has been finalized with a transaction**

**For** data preparation**, except for** turning binary variables **into** factors and **removing rows that contain** NAs**, we made no** change **to** the **raw data** when building **models** and making **analyses**.

count

Research Questions

Given the data, **we aim** at addressing **features** of **online** shoppers that could potentially **have an impact on** the purchasing **results**. **We will** examine **3** research questions:

1. What **is** the relationship between features related to timing **(**Weekend, Month, and special Day**)** and

**revenue generation?**

2. **Is there an** impact **of** visitor type(**new**, **returning,** other**) on** whether **revenue will** be generated**?**

3. **How do some web metric( BounceRates, Administrative\_Duration**, **PageValues, and**

**productRelated) influence revenue** generation**?**

Research Question 1

**To start** with, **we would like** to look **at the overall relationship between time features** and **revenue. More** specifically, **we would investigate** how Weekend, **Month and Special Day can** impact **Revenue making**.

**We** begin by **showing** a bar **chart** with **density curves on counts** to see how **Revenue is** impacted by Weekend

**and Month**.

**Relationship** between **Revenue**, **Month and** Weekend

Shoppers by Revenue, Month and Weekend

3000 -

2000 **-**

1000-

not Weekend

A

0 **-**

I

I

I

I

I

1

1

I

1

2

3

4

5

6

7

8 9

10

Month

1

2

Weekend

-5

-4

-3

I

I

I

I

4 5 6

7

8 9 10

Revenue

FALSE

TRUE

**From** the graph **above, we** can **first visualize** the distribution of month **by looking at** height of bars on both facets. On March, May**,** November and December**, we** see the **more** online shopper **visits**, and there is no **record for Jan and very few for April. Then**, **looking at height for different** colored **bars**, **we see that** there **are always** more shopper **visits** without **revenue** than **revenue**. **Also, we see** that for **both** facet, **meaning** no matter **it's weekend or** not, **there are more** shopper **visits with revenue** made **on** non**-weekends than weekends**.

**Independence between Month and Weekend**

TRUE

**Also**, **we are** interested in seeing the **amount of visits regardless of revenue, but purely on Month** and **Weekend**. We **use a** mosaic **plot** to **visualize** the **relationship** between **online visits of Month** and Weekend.

**Mosaic Plot: Shoppers by Month and Weekend**

FALSE

Aug

**Dec** FebJulJune Mar

May

Nov

Oct Sep

Standardized

Residuals:

In **the mosaic** plot above**, the width of** the **rectangles is proportional** to a marginal **distribution**, and **the** height **is proportional** to **a conditional distribution**. **Thus**, **from the above plots, we** can easily **assess the marginal** distribution of **Month** by looking at **widths**, and conditional distribution given Weekend by looking at heights.

**Now we analyze the mosaic** plot, it appears that there **are significantly more online** shoppers **on November's**

Weekend **than we would** expect under the null **hypothesis of** independence, and significantly less online **shoppers on** Feburuary, June, May's **Weekend** than **we would expect** under **the** null **hypothesis of** independence. Thereby **suggesting that we** should reject the null **hypothesis**, **i.e.**, **we** would conclude that

**Month** and Weekend are not independent for online shoppers data.

**Revenue by Special Day**

-4:-2 -2:0

**<**-4

2:4

0:2

>4

Shoppers

**Meanwhile, we would also like to** see **whether** there **is a relationship** between Revenue **and Special Day. Thus, we used a line** plot **to visualize the Conditional distribution of Special** Day **given Revenue**.

Revenue by Special Day

4000

3500

3000

2500

2000

1500

1000

500

0

0

0.2

0.4

0.6

0.8

1

Month

Revenue

FALSE

TRUE

**From** the **above** graph**, we** can **see that most online shoppers visit** the **website** on **Non-special day occasions**. And **there is a slight increase** in **visits around special** day **index = from 0.2 to 0.6,** which **corresponds to from** 4 **days** to **2 days prior to a special day. This** phenomenon might be **caused by needs for special day** gifts **and** there **is a** decrease from **special** day **index = from 0.6** to 1**,** which might be caused by shipping **and** handling time for the **product**.

Research Question 2

**We would like** to **better understand** whether **the** type **of shoppers could** have **an** impact on **revenue generation**. **By** doing **so**, **businesses** would be **able** to **identify** and **target potential buyers more** accurately **using** different approaches **such** as pop-up **advertisements.** Specifically, **we** looked **at** the **variables Revenue and**

**VisitorType**.

**First**, **we constructed** a bar chart **of VisitorType** to **assess** the marginal distribution of **VisitorType** and **have a better understanding of our variable** in **interest**. **From the graph**, **we** can **see** that **returning visitors have** the highest **count**, **followed** by **new visitors,** and **visitors from** Other category have the **lowest** count.

Count

10000 -

7500-

5000-

2500-

Distribution of Visitor Type

0-

New Visitor

Other

Returning Visitor

Visitor Type

VisitorType

New Visitor

Other

Returning Visitor

**Then**, **we** moved **on** to **investigate** the relationship **between** Revenue and **VisitorType**. Here **we** use a **mosaic plot to assess the independence** of the **variables**. **From the mosaic plot**, **we see** the cell **of new visitors who have** made a **purchase is colored** blue, which **means this** cell **has observed counts** that **are significantly** higher than **what** we **would** expect under independence. **Besides**, **we** noticed **that** the cell of **new visitors** who **did** not **make** a purchase and **the** cell **of returning visitors** who have made **a** purchase **are colored red**, which **means this cell has observed counts that are significantly** lower than **what we would expect under** independence**. Thus, we say** that **VisitorType** and **Revenue are** dependent **on** each other and new visitors **are more** likely **to move forward with a transaction**.

**Mosaicplot of Revenue and Visitor Type, colored by Pearson residuals**

New\_Visitor

Other

Returning Visitor

FALSE

**=**

TRUE

1

I

000

Standardized

Residuals:

<-4

-4:-2

-2:0

2:4

0:2

>4

Although it seemed that **visitor** type and **revenue** are not independent, **we couldn't make** any statistical **conclusions** just based on observing the **above** graph. **We** need **to** conduct **additional** statistical **tests** to have enough evidence. Since **we** are dealing **with two** categorical variables**, we** would **use the** chi**-**square test to test **whether** Revenue and **VisitorType are** truly dependent. Since **the** p-**value is** less than **0.05**, **we** conclude that

**VisitorType and** Revenue **are** not independent **which further validates our previous interpretation of** the mosaic plot.

**## Pearson's Chi-**squared **test**

**##**

**##** data**: table(online$**Revenue**, online$VisitorType)**

**## X-squared = 135.25,** df = **2, p-**value **< 2.2e-16**

**In conclusion**, **we say** that **there is** an **impact of Visitor type on whether there will** be **revenue** generated **or not**. **We also** found that the **new visitors are more** likely to **make** a purchase than the **other two** types **of** visitors. **Businesses** should **try** to utilize **different** approaches to target each type **of** visitor in order **to** generate more

**revenue.**

**Research** Question 3

**In this section, we are interested** how **other web metrics of** individual **visitors affect revenue** generation**.** To be **more concise**, **we will investigate** in the **relationship of some quantitative variables**, **such as**

**Administrative\_Duration,** PageValues**, BounceRates, and ProductRelated**, **with** Revenue**.** When **we look at the dataset**, **there are 10422 observations with no revenue and** 1908 **observations with revenue**. **To deal with such an** unbalanced dataset, **we** adopted a **wide**-accepted technique called **resampling**. **It** consists **of removing samples from** the **majority** class.

**## # A tibble: 2** × **2**

**##**

**Revenue**

**n**

**##**

**<**fct>

**<int**>

**## 1 FALSE**

**10422**

**## 2 TRUE**

**1908**

**Data Resampling**

To begin **with** the **analysis, we will** do **resampling** to better **visualize** whether the **four** variables **are** distributed **differently** depending on **revenue** generation. Notice that in **the** current dataset there are far **more** data **without revenue** generated than **with** revenue**.** With this **end**, **we randomly** undersample the original data **without revenue so that** the **size of data** with and without **revenue are roughly** equal.

Next, **we want to make sure that** undersampling the **majority class doesn't** lead to underfitting**,** i.e. the **new** sample fails **to** capture the **general** pattern **in** the **data**. We create side-**by-**side boxplots for **all** the four **quantitative variables** to **compare** their **distributions between** the **original and** the **new dataset**.

Boxplots of 4 Variables: Old vs New Dataset

2000-

3000-

2000-

Admin. Duration (old)

1000-

0-

-0.4 -0.2 0.0 0.2 0.4

0.2

0.20-

Bounce Rates (old)

0.05-

0.10-

boo.

0.15-

Bounce Rates (new)

Admin. Duration (new)

1500-

1000-

500-

300-

200-

Page Values (old)

100 **-**

0-

0 -

1

-0.4 -0.2 0.0 0.2 0.4

0.20-

0.15-

0.10-

0.05 -

600-

400-

Product Related (old)

200-

100 100

-0.4 -0.2 0.0 0.2 0.4

-0.2

Product Related (new)

300

200-

Page Values (new)

400-

200-

100-

0-

-0.4 -0.2 0.0 0.2 0.4

0.00 **-**

0.00-

-0.4 -0.2 0.0 0.2 0.4

-0.4 -0.2 0.0 0.2 0.4

0-

-0.4 -0.2 0.0 0.2 0.4

0-

-0.4 -0.2 0.0 0.2 0.4

**From** the graph **above, we** can **see that the distributions of Administrative\_Duration**, **BounceRates,** and

**ProductRelated are** roughly **equal between the two datasets**. **In terms of PageValues, we observe that although the distributions** between **two datasets** are **different**, **this** can be explained by the **extreme** right- **skewness** in the **original dataset**.

**Clusters and Distances**

**We will** then **investigate** how the **overall distances in** the **four variables among the data differ** depending on **how whether** revenue **is generated for a certain visit**. **To** do **this**, **we** create a dendrogram **using** complete**-linkage** as **clustering method**, **with the leaves colored** by **revenue generated**. **In this case**, **the red leaves represented** the data **with revenue**, **while** blue **leaves represent the data without revenue**.

0

Pairwise Distance

**LO**

10

15

**From** the dendrogram **above, we observe that** the left **cluster** tend to **have visits with revenue generated**. **This** indicate **that some visits within** 17 **dendrogram** distance generating **revenues** tend to **have** similar **web metrics in terms of Administrative\_Duration, Pagevalues, BounceRates**, and **ProductRelated. From** the right cluster**, we** cannot **draw any** clear **conclusion as because of the strong noises**.

**Dimension Reduction**

**Given** that **we** are analyzing a **4-**dimensional dataset, **we will do** principal component analysis **to** reduce the **data** into lower **dimensions. To know how many components explain the majority of** the total **variations**, we **create an elbow plot as below.**

**Cluster Dendrogram**

Proportions of Variations

0.35-

0.30-

0.25-

0.20 -

1

1

elbow plot

2

component

3

4

**As** the elbow plot **suggests**, **the first two** principal components explain **over 60% of the** total variations. **Thus**, **for** the **next part**, **we will investigate how the four** variables **are associated with** the **first** two principal components. To **visualize** the **relationship**, **we** create the **following** biplot with data points colored by whether there is **revenue** generated:

standardized PC2 (26.9% explained var.)

-5-

5-

10-

PCA biplots: PC1 vs PC2

Bounce Rates

0

PageValues

Adordous Rallye Duration

I

3

6

groups

FALSE

TRUE

standardized PC1 (34.7% explained var.)

**From** the biplot above**, we observe** that **although** there **isn't** a clear relationship between the **first** principal component and Revenue, higher **values** of the **second** principal component **are** associated with higher likelihood **of revenue. Also**, **we** find **that** BounceRates tend to be **negatively associated with** the **first two principal** components, while **PageValues is positively associated** withe the first **two** principal components. In **terms** of **ProductRelated and Administrative\_Duration, they are positively associated with** the **first principal** component, but are both negatively associated **with the** second principal component. To infer **from this,** web **visits with** revenue generated **may have lower BounceRates** and higher **PageValues.** The **relationship between Revenue and Administrative\_Duration and ProductRelated** tend **to be unclear from**

**the two dimensions.**

**In** conclusion, from clustering based on **distances**, we find that web **metrics for visits with** revenue generated **tend to be similar in terms of Administrative\_Duration**, **ProductRelated**, **PageValues, and BounceRates**, while such similarity **seems** unclear **for** web **visits** without revenue**. Moreover**, **by** dimension**-reduction** for the **four variables, we** find that Revenue tends to be positively associated **with** the **second** principal component. **Also, we** observe **that BounceRates and** PageValues **are positively and** negatively associated **with Revenue**, **respectively**.

Conclusion

In **this project**, **we have shown that, first, time** variables **have** some impact **on** visits **and** revenue **making**. People tend to make online **visits** on some **months'** weekends **more** frequently than others**,** indicated by **dependency** between Weekend and **Month** variables. We **also** found that more visits and revenue **are made on weekdays** than on weekends. And people **do not seem** to **make more visits and** purchases near special **days**. **Furthermore**, **we discovered** that the types **of visitors and purchasing** behavior depend on each other. **More specifically**, **new visitors are more** likely to **make** a purchase **than returning visitors.** Finally**, we** find **that** web

metrics for visits **with** revenue generated tend to be similar in terms of Administrative\_Duration,

ProductRelated, PageValues, and Bounce Rates. At the same **time**, **such similarity** seems unclear for web **visits without revenue**.

Besides **our three** research questions, **there were** additional questions that this project **has** not answered. **First**, **future research** should take a look at the geographical **feature** Region since the shopping **styles** could **vary**

in

**different areas**. Another **topic of interest would be Traffic** Type **to see** the **source** by **which the visitor** has **arrived at the Web** site. It **can** be **beneficial for** the **business** to choose the **best way of advertising**. **Last but not** least, classification **models using** machine learning **techniques such** as K **nearest neighbors**, **Logistic regression,** Support Vector **Machine**, and decision trees **can** be **used to identify** purchasing and non-purchasing **customers** more accurately. **The** capability of **segmenting** people into **'**likely to **purchase**' and 'unlikely to **purchase' groups would** facilitate **optimization** in **all aspects of the business from product recommendation,** and **sales promotion strategies** to **operations management**.