Assignment #5

Course: ISA 414

Instructor: Dr. Arthur Carvalho

Points:100

Due date: November 13th, 2022, before 11:59 pm

Submission instructions: this assignment is to be done individually. All your answers should be in a single python script or Jupyter notebook. Your code must be well formulated (*i.e.*, no errors) and sound (*i.e.*, it does what the question asks it to do). In particular, the grader must be able to copy and paste the content of your file into Databricks and run the code without running into errors. Code with errors may receive zero points. Submit the final document on Canvas before the due date.

Question 1 – Case Study: bol.com

A 2015 study sponsored by the Dutch e-commerce company bol.com, led by Arthur Carvalho (previously: Rotterdam School of Management – Erasmus University; currently: Farmer School of Business – Miami University) and Esther Hundepool (PwC), investigated some of the factors that affect consumers' willingness-to-buy in B2C e-commerce environments. The case below is an adaptation of the above study.

Business Understanding:

Over the past 25 years, the Internet has changed the way consumers buy goods and services. From groceries to vacation packages and clothing, more and more people use the Internet to shop online. The online selling of products or services by businesses to consumers is often defined as business-to-consumer (B2C) electronic commerce (e-commerce).

One can argue that *trust perception* is one of the most significant barriers for consumers to engage in electronic commerce. A potential lack of trust will likely discourage consumers from participating in online shopping. Therefore, it is interesting to study how to manage trust in e-commerce environments and the influence of different trust types on consumers' willingness-to-buy online.

In addition to trust perception, *risk perception* can be another challenging factor in e-commerce. Different types of risk perceptions are likely to influence consumers' attitudes towards online transactions.

Finally, consumers' demographic traits might also be of influence when it comes to online shopping behavior.

This study investigates the variables that positively or negatively influence consumers' willingness-to-buy in B2C e-commerce environments. Following the above background sketch, one can formulate the underlying business problem as:

What are the determinants of consumers' willingness-to-buy in B2C e-commerce environments?

In particular, this study aims at measuring the effects of perceived risk and perceived trust on consumers' willingness-to-buy online. As e-commerce sales are expected to continue growing over the years, understanding these factors and how to deal with them effectively will play a crucial role in the online strategies of companies engaging in e-commerce.

Data Understanding:

The data in this study were collected by means of an electronic survey developed in partnership with PwC and bol.com. To illustrate the process of online shopping, we started by showing the respondents a 5-minute video containing actual browsing and shopping behavior on bol.com, the number one online retailer in the Netherlands. Specifically, after exhibiting some website features, the video showed the search for and the purchase of a digital camera. When the video was over, we showed a web page from bol.com containing a detailed description of the purchased camera.

Following the video and product description, a survey measured three dimensions of perceived risk and three dimensions of perceived trust using five question-items per dimension. The six dimensions are *Perceived Product Risk* (PPR), *Perceived Informational Risk* (PIR), *Perceived Economic Risk* (PER), *Perceived Integrity* (PI), *Perceived Safety* (PS), and *Perceived Benevolence* (PB).

Next, the survey measured the main dimension of interest, *Willingness-to-Buy* (WTB), using five question-items. All the question-items used a 0-100 scale. Think about a chosen scale-value as the probability (represented in percentage values) that the respondent agrees with the statement in the question-item. Finally, the survey collected demographic information, such as respondents' age, income, and gender.

We invited participants via social networks and by sending emails to subject pools from Rotterdam School of Management and the office of the company PwC located in Rotterdam (the Netherlands). In total, 360 participants started the survey.

After the data collection phase, we prepared the resulting data set for posterior analysis by removing all incomplete survey responses, which resulted in 199 complete observations in the data set, a completion rate of 55.27%. We show below the structure of the survey we used to collect data (translated from Dutch):

- Perceived Product Risk (PPR)
  - *PPR_1*: I think this product will perform as expected.
  - *PPR_2*: The product purchased will likely not perform as expected.
  - *PPR_3*: I think it is difficult to judge the quality of this product adequately.
  - *PPR_4*: A product purchased on this website will likely fail to work as expected.
  - *PPR_5*: I believe the chances are high that something is wrong with the performance of this product.

- Perceived Informational Risk (PIR)
  - *PIR_1*: It is clear to me whether Bol.com intends to give my personal information to third parties.
  - *PIR_2*: I believe this website will protect my personal information from exposure to third parties.
  - *PIR_3*: I believe Bol.com does not intend to misuse the personal information provided by me.
  - *PIR_4*: I believe Bol.com will protect and store my personal information correctly.
  - *PIR_5*: I believe Bol.com is likely to misuse my personal information.

- Perceived Economic Risk (PER)
  - *PER_1*: Purchasing from this website would involve economic risk (fraud, difficulty to return, *etc.*).
  - *PER_2*: I believe I can return this product and get a refund easily.
  - *PER_3*: I believe there is a high chance of losing money if I purchase this product.
  - *PER_4*: When I purchase this item from Bol.com, I have the chance of financial loss.
  - *PER_5*: I believe there is a great chance I do not receive the intended product.

- Perceived Integrity (PI)
  - *PI_1*: Bol.com acts sincerely in dealing with their customers.
  - *PI_2*: I believe this online shop is honest with its customers.
  - *PI_3*: I believe Bol.com would keep its promise.
  - *PI_4*: I would characterize Bol.com as honest.
  - *PI_5*: Bol.com acts truthfully in dealing with their customers.

- Perceived Safety (PS)
  - *PS_1*: I believe this online shop has sufficient technical capacity to prevent hackers from intercepting my data.
  - *PS_2*: I believe this online shop shows great concern for the security of any of the transactions.
  - *PS_3*: I think this online shop has mechanisms to ensure the safe transmission of my information.
  - *PS_4*: I believe in having a safe transaction when purchasing from Bol.com.
  - *PS_5*: Purchasing from this online shop is safe.

- Perceived Benevolence (PB)
  - *PB_1*: When problems occur, I believe this website will be prepared to solve my problems.
  - *PB_2*: In case of a problem, I believe it will be easy to report a complaint to this website.
  - *PB_3*: I believe, when required, Bol.com would do its best to offer help.
  - *PB_4*: In case of a problem, I believe this website will make all the necessary efforts to solve it.
  - *PB_5*: I believe this online shop keeps the well-being of the consumer needs in mind.

- Willingness to Buy (WTB)
  - *WTB_1*: The likelihood that I would shop at this online shop is high.
  - *WTB_2*: I would consider buying this product at this price.
  - *WTB_3*: I would be willing to recommend this online shop to friends.
  - *WTB_4*: I would be willing to buy at this online shop.
  - *WTB_5*: It is likely that I will purchase at this online shop.

- Demographics:
  - Gender: What is your gender?
    - Male
    - Female
  - Age: What is your age?
    - Below 18 years old
    - Between 18 and 25 years old
    - Between 26 and 35 years old
    - Between 36 and 45 years old
    - Between 46 and 55 years old
    - Above 55 years old
  - Income: What is your current yearly income?
    - Less than $20.000
    - Between $20.000 and $35.000
    - Between $35.000 and $50.000
    - Between $50.000 and $65.000
    - More than $65.000
    - I prefer not to say

Data Preparation:

It is now time to analyze our data. You will be using the Spark technology in conjunction with the Python programming language inside the Databricks platform.

a) Start by downloading the data set *bol.csv* from Canvas. Next, upload the data to Databricks and create a Jupyter notebook there. **[0 points]**

Unless otherwise stated, all the following questions must be answered with code executed on the Spark cluster.

b) Note that the scales of PPR_1, PIR_5, and PER_2 are different from the other items' scales in their dimensions (constructs). For example, the scale of PPR_1 increases positivity, whereas the scales of PPR_2, PPR_3, PPR_4, and PPR_5 decrease in positivity. Therefore, you have to transform the scales for the sake of consistency. The goal of this preprocessing step is to have all risk-related variables using scales in increasing negativity and all trust-related variables using scales in increasing positivity. To do so, transform the variables PPR_1, PIR_1, PIR_2,

PIR_3, PIR_4, and PER_2 by subtracting their original values from 100, *e.g.*, the new values of PPR_1 must be equal to 100 minus the old values. Hint: look at the Spark function withColumn(). **[20 points]**

c) After fixing the scales, it is now time to create our variables. Remember that we measured each risk and trust dimension using five question-items. Since the question-items are highly subjective, one should expect that the respondents' answers contain some "random component." A common approach to eliminating some of this "randomness" is by averaging the question-items' values across each dimension. In practice, one would have to perform reliability analysis and check for internal consistency before doing so (*e.g.*, performing confirmatory factor analysis and calculating Cronbach's alpha), but this is beyond the scope of this assignment. Using the withColumn function (same as above), add the following features to the data set in the Spark cluster: **[20 points]**

PPR = (PPR_1 + PPR_2 + PPR_3 + PPR_4 + PPR_5)/5

PIR = (PIR_1 + PIR_2 + PIR_3 + PIR_4 + PIR_5)/5

PER = (PER_1 + PER_2 + PER_3 + PER_4 + PER_5)/5

PI = (PI_1 + PI_2 + PI_3 + PI_4 + PI_5)/5

PS = (PS_1 + PS_2 + PS_3 + PS_4 + PS_5)/5

PB = (PB_1 + PB_2 + PB_3 + PB_4 + PB_5)/5

WTB = (WTB_1 + WTB_2 + WTB_3 + WTB_4 + WTB_5)/5

Data Modeling:

d) Next, you will build an explanatory model that relates risk and trust dimensions to willingness-to-buy. To simplify the analysis, ignore the demographic variables in the data set. Using the LinearRegression function from the pyspark.ml.regression module, build a linear regression model where the dependent variable is WTB. The independent variables are PPR, PIR, PER, PI, PS, and PB. **[20 points]**

Conclusion:

e) Assume that all the assumptions behind the regression analysis are valid. Recall that in practice, you have to always check the assumptions. But doing so is beyond the scope of this course. Assume that your model is represented by the variable model. You can then retrieve the coefficients and p-values associated with your model by running model.coefficients and model.summary.pValues. Important: ignore the last p-value since it is related to the intercept in your model. Note that the order of the coefficients and p-values in the above lists is the same as the order of the variables in your "predictors" list, which you used to build the model.

Given the coefficients and p-values from above, which actions would you suggest bol.com take to increase consumers' willingness-to-buy? Specifically, carefully suggest at least one feature that bol.com could add to its website to alleviate the underlying risk for each significant coefficient and associated variable. That is, say you find three statistically significant variables, you should then discuss three website features help with them. (*sloppy answers may receive zero points*) **[40 points]**