**BBS CUSTOMER & MARKETING ANALYTICS: TEAM ASSIGNMENT 1**

**Customer Sentiment/ Rating Scores In Product Reviews: Their Impact On Business Outcomes**

**DESCRIPTION OF THE DATA**

In this assignment, you have a dataset from 600 customer reviews from an online shopping platform selling consumer electronics and PC accessories to individual customers.

*Online Reviews/Ratings*

On this online shopping platform, customers can write online reviews on listed products on the basis different criteria such as (i) design (ii) technical features (iii) price and (iv) delivery terms and conditions. An important information that you need to know is customers pre-inspect products and then eventually write reviews/rate the products (on a 1-5 scale) without buying/before buying the products.

In other words, customers (visitors) who write online reviews on products are not necessarily purchasers no matter what the review/rating is a purchase may or may not occur after the product is reviewed-rated by the visitors.

(a) Rating Scores are already numeric (1-5) and directly given (scored) by customers.

Textual Reviews: Since this was (originally) unstructured text data, our text-analytics team did a great job and they helped us to process/prepare the text data so that we can use the information coming from textual reviews in our modelling exercise. In other words, they quantified (coded) the text data (on reviews) so now that we have them in our data as numeric variables:

As marketing analysts we are curious about three issues (i) sentiment (ii) topics and (iii) length of review

In other words we are curious whether it is the overall sentiment (emotional positivity) of the review, the topics mentioned in the review (which one) and the length of the review (how long the detailed the review is) having an impact on business outcomes – or all of them.

(b) Sentiment Score: This is (pre) calculated by a sentiment-analysis model-algorithm by our text-analytics team and represents the overall positivity of each customers’ (textual) reviews. This is calculated in a 1-5 scale score; and reviews are classified (scored) as 1 being the lowest (very negative) and 5 being the highest (very positive). Now this is also a 1-5 score variable as rating score.

(c) Topics: Customers/Visitors can mention-talk about design, technical aspects, price and delivery conditions in their reviews. We are curious (mentioning of) which topics have an impact and-or relatively more impact on further business outcomes. If there are such topics as marketing team, we can make improvements on that topic or we can improve our communication on those aspects.

*(\*) To initially process-prepare the textual data in numeric format various NLP libraries are used and you see the final-outcome quantified data now in this assignment. You will not see the original (textual) review data in this assignment since we visit this topic later in this course.*

Our text analytics team therefore, (after a topic-analysis) prepared this as a numeric data for us so now we have variables on whether or not each topic is mentioned by each customer (reviewer) in her/his reviews. They have been even more helpful – so that they have also prepared (for each topic) whether each topic is mentioned in a positive or negative context.

(d) Length of the Review: Another important question is whether longer (more detailed) textual reviews are taken as more serious by customers and therefore are more influential on business outcomes. To answer this question, we also have a variable on the length of the review (again a numeric variable) refers to total number of words used in each review.

*Purchase Incidence (after Review)*

Fortunately, we also have data on, whether or not a purchase is occurred (whether or not customer purchased) after pre-reviewing/ pre-rating the products. In other words: we know whether or not each reviewer (visitor) made a purchase after reviewing the product.

*Total Number of Purchases in Relational History*

On top of that, we also have data on how many # purchases each customer (reviewer) made in her/his relational history with our online shopping platform. To be more accurate, this information includes the total (times) of purchases each customer made in the past, starting from the date s/he registered as a customer on our online shopping platform up until today.

*Customer & Product Background Variables*

Last but not the least, we also have some information (variables) on certain customer and product characteristics such as: (i) age, (ii) country and (iii) gender (customer demographics) and (iv) product category of reviews: meaning, whether customer produces reviews on utilitarian products, hedonic products or both (utilitarian & hedonic).

The detailed list of all variables, their operationalization and values are given in the table (Appendix)

**QUESTIONS**

Now we ask you to build and estimate few models to investigate the review-rating related and other factors on certain business outcomes. Please feel free use (include) or not to use (exclude) any variables, and, make choice(s) on the model type (regression model type) that you use to model these relationships. Although there are certain choices which are absolutely needed to be done – when it comes to modelling many other things are up to you as a marketing analyst and often times there is no one-only-correct-model (thus multiple ways of doing things are possible). Therefore, while following the certain knowledge-rules, on the other side: feel also free to be creative in your modelling decisions in building your model and while deciding which variables to use. As a matter of fact, making certain choices and building your model is an integral part in this team assignment.

Please only use the knowledge you have learned in Lectures 1-2 in this assignment. You are not responsible (or expected to use) any other methods which will be covered in the later sessions or that might have been captured in other courses. For practical convenience feel free to use 1-5 interval scales as metric/continuous variables. Here are the questions:

**Question 0: Data Preparation**

**Please revisit this question after reading the other questions**

Is your data ready for analysis? Do you need to do any dummy coding, or do you need to recode certain variables in a different way so that they can be used in your models? If yes: Please briefly explain why/how do the required preparations in your data before your models.

* Category, utilitarian hedonic into dummies? 1 column for both categories, 1 column for isUtilitarian

Hedonic | Utilitarian

1. 1- both

0-hedonic

0 1- utilitarian

* Should we divide age into ranges? 18-35, 36-50, 51-65
* Should we removed all the columns that end with \_negative?
* Or should we combine them into a single column?

**Question 1: Rating Scores**

Please build-estimate a model to investigate: What drives higher Rating Scores?

In other words: What are the factors that can explain the higher Rating Scores given by customers. Are certain elements (variables) about Textual Reviews (represented by mentions on certain topics and sentiment score <positivity>) can explain the variation in Rating Scores. Are positive reviews (with high sentiment score) really leading to higher Rating Scores? Or are customers consistent-congruent when they verbally assess/evaluate a product and when it comes to give those products a straight rating score (1-5 scale) Could there be other factors (or variables) that you can consider and that can also have an impact on rating scores or on this relationship?

Please build, estimate your model and briefly explain your results with marketing implications.

**Linear Regression? Because the dependent variable which is the rating score in our point is continuous and take real values from 1 to 5.**

The R-squared value of 0.706 indicates that the model explains 70.6% of the variation in rating scores. The fact that some of the p-values are greater than 0.05 suggests that some of the factors included in the model may not be statistically significant in explaining rating scores. For example, the p-value for "Prod\_Design" is 0.304, which is greater than 0.05. This means that the effect of product design on rating scores may not be statistically significant.

**Question 2: Purchase Incidence**

Please build-estimate a model to investigate:

Can we explain the purchase incidence (purchase occurrence given as 0-1) that may occur after customers pre-reviewing and rating the products, by using rating scores, and/or other variables that might relate to textual reviews (sentiment, content of review: thus mentioning certain topics, and the length of the review). Which elements/variables are more influential in explaining the probability of purchase (purchase propensity) following the (pre) review process? Do better reviews or higher rating scores really lead to increase in purchase probability? Are there any other factors or variables (that could be of help in having a better model) that you can use in your model?

Please build, estimate your model and briefly explain your results with marketing implications.

**Logistic regression? Because of binary outcomes 0s and 1s.**

* **Rating Scores:** The impact of rating scores themselves isn't directly shown here, but the sentiment of reviews on certain aspects of the product (design, technical, service) do appear to influence purchase. Positive sentiment increases the probability of purchase, while negative sentiment decreases it.
* **Textual Reviews:** The number of words in the review isn't shown to have an influence here (coefficient with a p-value greater than 0.05). However, positive sentiment specifically towards service delivery does increase the probability of purchase.

**Question 3: # Number of Purchases in Relational History**

Please build-estimate a model to investigate:

A model explaining the variation in # number of total Purchases in relational history by using the variables (information) that we have. Which variables can we use to predict the total number of purchases in customers’ relational history.

Please build, estimate your model and briefly explain your results with marketing implications.

* Poisson? Number of purchases is a count variable, we can find expected counts.

**1. Coefficients:**

* This section shows the estimated effect of each predictor variable on the expected number of purchases.
  + A positive coefficient indicates that an increase in that variable is associated with an increase in the expected number of purchases.
  + Conversely, a negative coefficient suggests that an increase in the variable leads to a decrease in the expected number of purchases.
* Pay attention to the significance level (p-value) associated with each coefficient. A low p-value (typically less than 0.05) indicates that the effect of the variable is statistically significant, meaning it's likely not due to chance.

**2. Null Deviance and Residual Deviance:**

* **Null Deviance:** This represents the overall variation in the number of purchases before fitting the model.
* **Residual Deviance:** This indicates the remaining variation in the number of purchases after accounting for the effects of the predictor variables in the model.
* A lower residual deviance compared to the null deviance suggests that the model explains a good portion of the variation in purchase counts.

**Question 4: Make 1-2 simple Interactions for one of your models.**

Another valuable (and fun) aspect of modelling is to explore/spot certain interactions – telling us which variables could have combined-effects (interaction effects) on our outcome variable. Interactions also tell us: the impact of which drivers (or variables) on outcome variables are observed stronger – when combined (used by firm, done/mentioned together by customers, or customers are being exposed to together; or a variable only has a significant for a certain customer/or product group) This is also a nice exercise to see which variables produce a stronger impact (or maybe not) when they are observed together. Now please look at your models in Question 1-2-3. CHOOSE one of them. NOT all.

For the model you chose (not for all models):

Please try to explore and test 1-2 interactions/potential interactions that you find potentially interesting, likely or valuable for managers.

*Note: As there is unlimited number of interactions (and simplicity is an important priority in modelling) and we have limited time please do not try to interact everything-with-everything and try to include max 1-2 interactions which you find managerially and analytically interesting. The main idea behind this part of assignment is to give you a hands-on practice with creating-estimating and interpreting 1-2 interaction effects in your analysis. We are not trying to find the best interactions in the data.*

Please show how you include interaction effects: which variables you choose as potential interaction variables (why) and compare your model with main effects only model. Are your interactions significant? What are the marketing implications? Do you think you now have a better model compared to main effects only models? Briefly explain your results with marketing implications.

* **Significance of interaction terms:** The p-value of each interaction term will tell us if the interaction effect is statistically significant (p-value < 0.05). Significant interaction terms indicate that the effect of one variable on purchase frequency depends on the level of another variable.

**Instructions & Practical Matters**

*Feel free to use any analytical software or environment in your assignment (R, Python, SPSS, STATA or SAS) We are not interested in (and it is not a grading element) which software you use. What matters is properly building/estimating the model and the conclusions-implications you draw from your results.*

*Please share your assignment with me ([u.konus@uva.nl](mailto:u.konus@uva.nl)) latest by* ***Thursday 21 March 09:00*** *before the lecture where you present (and we discuss) team assignments in class.*

*Assignment will be in PPT (Powerpoint) format max 8-9 slides including cover page and appendices. Your assignment = your presentation. You don’t need to prepare a separate report. If you think you need detailed explanations: use “notes” tab beneath the ppt slides. Sure I will read them in detail.*

*Your assignment (ppt) will include (i) a short executive summary in few bullet points (ii) 1-2 slides per each question: what are the modelling choices/variables you use and why <with brief explanations> ; your model (with results/estimates) and your conclusions on each model/question. You don’t have to re-run your data practice during presentations: you can use screenshots or output tables from your model results in presenting your results. Please be brief-concise and clear in showing your reasoning and results: You don’t have to describe (show) in detail or data processing and preparation steps or other preliminary checks-controls you do in your analytical process. You can briefly mention them either in notes or verbally during your presentation. In your presentation (report) please mainly focus on (building and presenting) your model and your results.*

**Appendix: Variables in the Data Set**



**Additional Note:** In case you wish to share your R/Python codes-scripts or STATA-SPSS syntax together with your assignment please do not show them on your slides. You can share them as a separate appendix (separate file). Your slides (and explanations) should only include brief reasoning-explanations on your modelling decision, brief description or equations on your models and finally your (main) output tables (results) and your conclusions and implications on those results.