Holland and Barret

Data Science Exercise

The Director of Digital Sales has just had a meeting with you where he pointed out the importance of reducing the incomplete online orders. Currently, approximately 10% of transactions remain incomplete. Holland and Barrett invest a lot in its digital presence in order to boost sales and meet demand, especially during those critical times during the pandemic.

After your meeting, the Director of Digital Sales sent you two csv files with data for 5000 digital transactions (**transaction\_data.csv**) ranging from January to March 2020 along with the customer information (**customer\_data.csv**). They want to use this information in order to predict and prevent incomplete transactions. You are to create a **model** in either R or Python and use the output to write a short presentation to inform the Director on your findings and insights. Your model will show the impact on **Digital Sales**. He also provided an additional csv (**new\_transactions.csv**) file with 1000 transactions that their completeness status was lost during a system failure. He expects you to provide a prediction for each of those transactions as to whether they were completed or not.

Along with your code written in R or python, you should prepare a presentation of no more than 10 slides (including cover slide and any appendix slides) where you talk about the following points:

* Executive Summary for the Director
* Methodology
* The output of the model you used
* How the output can be used to influence digital sales strategy (regarding incomplete baskets)
* Additional data you could ask to enhance your analysis
* Explanation of metrics used to assess the predictive value of your model
* Any Caveats and additional research you would need to do to make the results more accurate

Notes

* Please note that the Director of Digital Sales does not come from a statistical background, so any statistical / data science / machine learning concepts need to be explained in non-statistical terms.
* If successful, you will be expected to present your slides.

Assessment

For this exercise, what we want to see is a good methodology and flow in your code.

What we want to see in your code:

* Mandatory things:
  + Clean code
  + Good flow in the code
  + Meaningful comments describing what you are doing
  + Good Data Science Process in terms of Data Processing, Model Building and Evaluation
* Nice to have:
  + Package the code for easy deployment and usage
  + Include docstrings that provide a concise summary and explanations of the functions
  + Implement unit tests to validate the correctness and reliability of the code

What we want to see in the presentation:

* Storytelling flow
* Actions for the business
* Executive summary containing information that the board would like to see

Explanation of Data Columns (transaction\_data.csv, new\_transactions.csv):

'Transaction ID': Unique transaction Identifier

'Customer ID': Unique customer identifier

'Date': date in YYYY/MM/DD

'Total Items': Total items in basket

'Unique Items': Unique items in basket

'Total Sales': Total expected price

'Discounted Sales': Actual price after discounts and vouchers

'Browsing Duration (minutes)': time spent browsing on website

'Number of Clicks': Actual number of clicks during browsing duration

'Incomplete Transaction': Binary outcome of completeness (0= completed transaction, 1 = incomplete transaction)

Explanation of Data Columns (customer\_data.csv):

'Customer ID': Unique customer identifier

'Age': Age of customer

'Gender': Customer’s gender

'Region': Region characteristic of customer’s household

'Marital Status': Customer marital status

'Education': different levels of maximum education level for customer

'Household Income': Total annual income on household

'Loyalty Card': Customer has loyalty card (0=No, 1=Yes)

'Loyalty Points': Loyalty points per customer at the end of the analysis period