ADDITIONAL RESEARCH

Just getting things wrong

To ensure the model stability (i.e. moving from the training to the validation dataset). it is necessary to ensure the correct understanding of current problem being solved. In this case, the key area of interest is the prediction of the target variable TargetBuy to evaluate if a customer likely to buy the products, and this objective should be kept in mind when identifying the relevant predictor variables (e.g. gender, regions, amount spent, etc.), selecting the right units of measurement for all variables, and interpreting the results. If any of these steps is not well aligned and relevant to the identification of the likelihood of buying by customers, the purpose of building the model is defeated in real-life applications where the model becomes useless.

Overfitting

The issue of overfitting happens when the model simply memorizes the patterns of the training dataset and hence, performs poorly in the validation ones because it does not learn the general patterns very well. In this case study, one example could be that the built models might emphasize and rely on spurious relationships between TargetBuy and Affl or Gender, leading it to pick up the patterns and believe that customers with certain level of affluence or from a gender group are more likely to buy clothing. This means that the overfitting issue occurs as the model fails to capture the more generalized trends that actually influence purchasing decisions. For a predictive model, this also implies model instability as it performs well in training dataset but would struggle with new data when this spurious relationship does not exist. Therefore, after using the validation dataset or new dataset, the model is likely to perform poorly in prediction reliability because it remembers the coincidental relationships, which is not the case when the new data is inputted.

Sample bias

The issue of sample bias occurs when the model's training dataset underrepresents the real-world population, implying a poorer performance in prediction with larger dataset. In this case, for example, if the training dataset mainly consists of customers from a specific gender, or loyalty status, the model's prediction capability might be biased towards these particular groups. Specifically, if the training dataset use a disproportionate number of 'M', 'F' and 'U' as per the result in question 1c where the 'F' gender (i.e. 'female' ones) accounts for over 55% of the training dataset, it casts some doubts that the result might be biased toward the female group of buyers. Another example could be the TV region where the majority of it is from London, implying that the predicted outcome might not be well accurate for different regions, given that customers with this TV region might have more interest in fashion than others. Therefore, the in the case that the sample collected is not representative for the wide population, the prediction accuracy and capability are affected and hence, the model would perform poorly when being applied to different segments of the customer base.

Future not being like the past

Predictive models are built on the basis of using historical data for the model to learn, extract key patterns and apply to predict the future outcomes. However, the issue of "Future not being like the past" might occur because under the assumption that the past information is the best reflection of the future, the model might struggle to predict future changes, especially when the changes happen unexpectedly quickly. In this case, some of the sudden changes could be the changes in customer preferences, fashion trends, economic factors, etc. that the customer behaviors have been shifted. Also, in such a clothing retail industry, customer preferences are expected to change rapidly and hence, the model could not capture those on a timely manner. For example, customers at certain age shows more likelihood (i.e. TargetBuy) to purchase clothing in the past, but future customers at the same age might not inherit the similar purchasing behaviors due to changes in economic conditions for example. Under this circumstance, the model would result in poor prediction accuracy because the learned pattern from historical dataset can no longer be applicable to newly uncaptured changes in behaviors.