Reflections and Learning Outcomes:

As a student trying to complete this case study, the journey was undoubtedly challenging yet. The following reflections encapsulate the key learnings, the encountered challenges, and the strategies employed to overcome them in a learning context.

Key Learning

I can sum up my key learning in one phrase: Trial and Error. This doesn't mean I started doing this case study unprepared. However, it's the realization that data mining is often a mixture of planning plus experimenting and adjusting the plan according to the experimentation. It is the realization that SEMMA is truly an iterative process.It would be inefficient to stick rigidly to exploring, then strictly move to modifying without going back to sampling. The process should be flexible. In this case study, I nailed it by moving through exploration, experimentation, and modification as the situation demanded.

Challenges

Overall, the challenges can be summed up in one word—-Uncertainty. There was uncertainty in picking data sources. There was uncertainty in trying to decide whether the data size can be reduced for the sake of balancing. I am a big believer that data quality determines the performance of the model. Therefore, I spend a lot of time ensuring that the dataset I have is of good quality. With a lot of the time spent in preparing the data in consideration of issues that might happen in modeling, such as overfitting or underfitting, the time that can be given for trial and error might not be enough.With there being a time constraint, it can be very difficult to go through trial and error and try different things within 24 hours.

For the specific challenges, they were overcome by the tools provided by the software.

The Lacking of Domain Knowledge.

Challenge: Without domain knowledge, it can be very difficult to truly understand how these variables affect the operation of an ecommerce business. Understanding which variables were crucial for predicting customer churn without domain knowledge was like navigating a maze. With the SAS variable charts and Chi-Square Analysis, I was able to pinpoint the more important ones, but there was still uncertainty as I didn't know if the modeling result would turn out well.

Feature Engineering for Membership Level:

Challenge: The creation of the "Membership Level" variable based on specified conditions required careful implementation. I wanted to ensure that there was correlation between this variable with the other variables, but I didn't want to ruin the dataset. This challenge was eventually overcome by brainstorming about the definition of membership and what it means to a business. The conclusion was that membership meant people who are rich enough to buy items or register for a high level membership, indicating that they might live ina higher city tier. Also, they should be loyal and should be frequent customers. And so, I was able

to decide a combination of city tiers, tenure, and app usage conditions as the base for generating this variable.

Imbalanced Data:
Challenge: The dataset showed an imbalance in the target variable "Churn," with significantly more instances of non-churn (Churn = 0) than churn (Churn = 1). This challenge was somewhat overcome by using the Equal Size Sampling Node in Knime to create a more balanced distribution. Even though the sample was still big, I was unsure if this might affect the modeling results. Thankfully, the model performance turned out well.

Strategies to Overcome Challenges.
In a learning context, several strategies were employed to navigate and overcome the challenges encountered during this comprehensive case study:

Strategy 1: Using visualization tools like Talend Data Prep and Knime for data exploration. Recognized the importance of these tools in unraveling patterns, inconsistencies, and potential features within the dataset.

Strategy 2: I relied on the Variable Worth Charts and Chi-Square analysis to handle the challenge of feature selection as I lacked the domain knowledge.

Strategy 3: Acknowledged the need for effective time management. Emphasized the importance of discerning when to delve deeper into exploration and when to make decisions based on available information.

Strategy 4: When faced with uncertainties in deciding imputation methods, I developed a pragmatic approach. The strategy was to compare the imputation methods and choose the one that maintained the overall integrity of the dataset and embraced the iterative nature of data preparation.

Strategy 5: I used software tools like SAS variable charts, Chi-Square Analysis, and the Equal Size Sampling Node in Knime to overcome specific challenges, demonstrating a practical application of available resources.

These strategies collectively contributed to my learning experience, emphasizing not only technical skills but also critical thinking, adaptability, and a resilient approach to problem-solving.