**PROGRESS REPORT :2**

1. **Explain to the class the available data sources for your project. Talk about the various datasets, the number of records, what kind of fields they contain, and anything else about your data that you think is important to mention.**

The data provides valuable insights into when an account opted into the Cross-Sell program, their credit information, and their current account status, including whether an account has been locked or has a security deposit. It shows the financial write-offs and losses for the Cross-Sell accounts across different periods. Which will be helpful in understanding the write-off trends and their impact on the business. The datasets further show records of Customer spending on fuels, and the total spending, form which we can infer the customer's spending on other products.

We are planning to integrate the cohesive analysis that will guide insights into improving account management, Attrition analysis, write off analysis and risk analysis.

Basic )verview of all the sheets is as below

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset Name** | **Description** | **Number of Records** | **Number of Columns** |
| All Acct Segment Score.xlsx | Contains account segmentation scores for various accounts | 3 | 176056 |
| Cross-Sell Acct.xlsx | Contains account information for cross-selling opportunities | 18 | 29410 |
| Non-Cross-Sell Acct Info.xlsx | Contains account info for non-cross-sell accounts | 15 | 180384 |
| Non-Cross-Sell Aging Data.xlsx | Contains aging data for non-cross-sell accounts | 9 | 473626 |
| Non-Cross-Sell DNB Score.xlsx | D&B scores for non-cross-sell accounts | 4 | 480050 |
| Non-Cross-Sell NSF Payment.xlsx | Non-sufficient funds payment data for non-cross-sell accounts | 4 | 11499 |
| Non-Cross-Sell Payment.xlsx | Payment data for non-cross-sell accounts | 4 | 1028038 |
| Non-Cross-Sell Revenue.xlsx | Revenue data for non-cross-sell accounts | 7 | 954354 |
| Non-Cross-Sell Spend.xlsx | Spending data for non-cross-sell accounts | 7 | 1023651 |
| Non-Cross-Sell Vantage Score.xlsx | Vantage scores for non-cross-sell accounts | 3 | 490755 |
| Non-Cross-Sell Write-Off.xlsx | Write-off data for non-cross-sell accounts | 3 | 33347 |

**Data Collection Status:**For now we are using the data that has been shared with us on canvas and has majorly segregated data into 3 segments: Cross\_Sell\_Acct\_Info, All\_Account\_Segment\_Code and Non-cross sell data.

**2.** **How do you plan to prepare, integrate, and clean your data from the available data sources? How do you plan to establish relationships among the various datasets? How do you plan to join the datasets?**

**Data Cleaning**

While reviewing the data, we have found that approximately 63% of the Vanatge score column data in Non-Cross-Sell Vantage Score is blank. Therefore, we are planning to remove that data from the analysis. Furthermore, approximately 25% of the data in column “Days Past due” blank in the non-cross sell aging data sheet. We have checked the corresponding values in their adjacent columns and found that the values those columns have is 0.00. Henceforth, we have decided not to keep the rows with the blank data. Lastly, the 87% of the column “Deposit flag data” is blank in the cross-sell Account sheet which again we will be removing from the dataset.

The Next step is the duplicate removals. After merging the data, we will identify the duplicates and remove them from the datasets.

**Data Integration:**

The key column that will be used to merge the datasets will be Account ID (FAKE\_ACCTCODE), which is common not only across all the non-cross-sell datasets but also in the Cross\_Sell\_Acct\_Info, All\_Account\_Segment\_Code datasets.

**Merging Datasets:**

We will be Merging the PANDAS library in python and will be using the LEFT JOIN to retain all accounts from the primary dataset while integrating the relevant details from the other datasets.

**3.** **After your data synthesis efforts and joins, a lot of rows will be generated. How do you plan to work with your data so that you have one manageable dataset?**

The dataset is indeed large after the merger. However, to ensure that the data remains manageable we are going to check for the redundant columns to ensure the data is clean and ready to use for Analysis. We are going leverage database solutions like filtering the data into inactive and active accounts, high risk account for the risk assessment.