**Model Validation:**

In this section of the project, the test data underwent a preprocessing phase to prepare it for evaluation with the trained models. The test data was manually encoded to match the format of the training data used to train the models. Specifically, certain values in the test data were transformed or encoded to ensure compatibility with the models' input requirements.

To create the test dataset, certain entries needed to be excluded. The first 6 entries of each loan were removed using the 'groupby' function and other alpha functions. This step ensured that the test data reflected the loan's repayment history, starting from the point when the first payment occurred. By removing the initial entries where no payment took place, the test data better aligned with the models' expectations.

Once the initial entries were removed, the remaining data was further processed to select specific entries for evaluation. Depending on the specific model being tested, the first 1, 6, or 9 entries were chosen for each loan in the test dataset. The selection of these entries was based on the requirements of the respective models. For instance, the LSTM model required data from the window where the loan is first paid, along with a specific number of subsequent entries, depending on the defined window size.

After the necessary entries were selected, the test data was divided into different subsets based on the models to be evaluated. The data was split into pools, ensuring that loans with the same Identifiers were grouped together in each pool. This step allowed for a more comprehensive evaluation of the models' performance across different loan types.

Sample Script :

import pandas as pd

input\_file = 'ds2.csv' # Path to the input file

output\_file = 'test\_pool.csv' # Path to the output file

entries\_per\_loan\_id = 9 # Number of entries to select per Loan Identifier based on the model for which I/P is to be given

df = pd.read\_csv(input\_file)

# Create a new DataFrame without the first 6 rows for each Loan Identifier

filtered\_data = df.groupby('Loan Identifier').apply(lambda x: x.iloc[6:])

# Group the DataFrame by 'Loan Identifier' and select the top entries for each group

top\_entries\_df = filtered\_data.groupby('Loan Identifier').head(entries\_per\_loan\_id)

# Save the top entries DataFrame to the output CSV file

top\_entries\_df.to\_csv(output\_file, index=False)

**Characteristic based Pool Creation :**

This section of the project report focused on the creation of test pools based on four different characteristics: Loan-to-Value (LTV) ratio, Debt-to-Income (DTI) ratio, Loan Age, and Credit Score. These characteristics were used to categorize the loans and evaluate the performance of the models across different segments of the data.

To create the test pools, specific threshold values were determined for each characteristic. For LTV and DTI, the median values of 68 and 29, respectively, were used as the thresholds. Loans were categorized as either "Low LTV" or "High LTV" based on whether their LTV ratios were below or above the median value. Similarly, loans were categorized as "Low DTI" or "High DTI" based on their DTI ratios relative to the median.

For Credit Score, industry-standard categorizations were utilized. Loans with credit scores below 660 were classified as "Near Prime," while those with credit scores between 660 and 720 were classified as "Prime." This allowed for a further segmentation of the test data based on creditworthiness.

Additionally, the Loan Term was used as a characteristic to divide the data into two categories: 180 months and 360 months. This division provided insights into how the models' predictions performed for different loan durations.

Once the test pools were created based on these characteristics, the trained Deep Neural Network (DNN) models were utilized to predict the Single Monthly Mortality (SMM) and Cash Inflows. The predictions from the DNN models were taken as the output for these calculations. Additionally, a Random Forest model was tested on the data to predict the Weighted Average Life (WAL).

The results obtained from these predictions were analyzed and compared to the actual values. The predictions were plotted, allowing for a visual examination of the performance of the models across different test pools. Inferences were drawn based on the observations, highlighting any outliers or discrepancies between the predicted and actual values.

Overall, the evaluation of the models' predictions demonstrated their effectiveness in estimating SMM, Cash Inflows, and WAL for different loan segments based on LTV, DTI, Loan Age, and Credit Score. The results indicated that the models' predictions were generally close to the actual values, providing valuable insights into loan performance and risk assessment within specific segments of the data.