GUC
German University in Cairo
الجامعة الألمانية بالقاهرة

ACL Report

Members:

- Amr Maged Hassan 55-21247
- Arwa Saad Shalabi 55-10737
- Shahd Ayman Taha 55-12419
- Osama Hamed Abolata 55-4799

Submission Date: 22 October

1. Data Cleaning

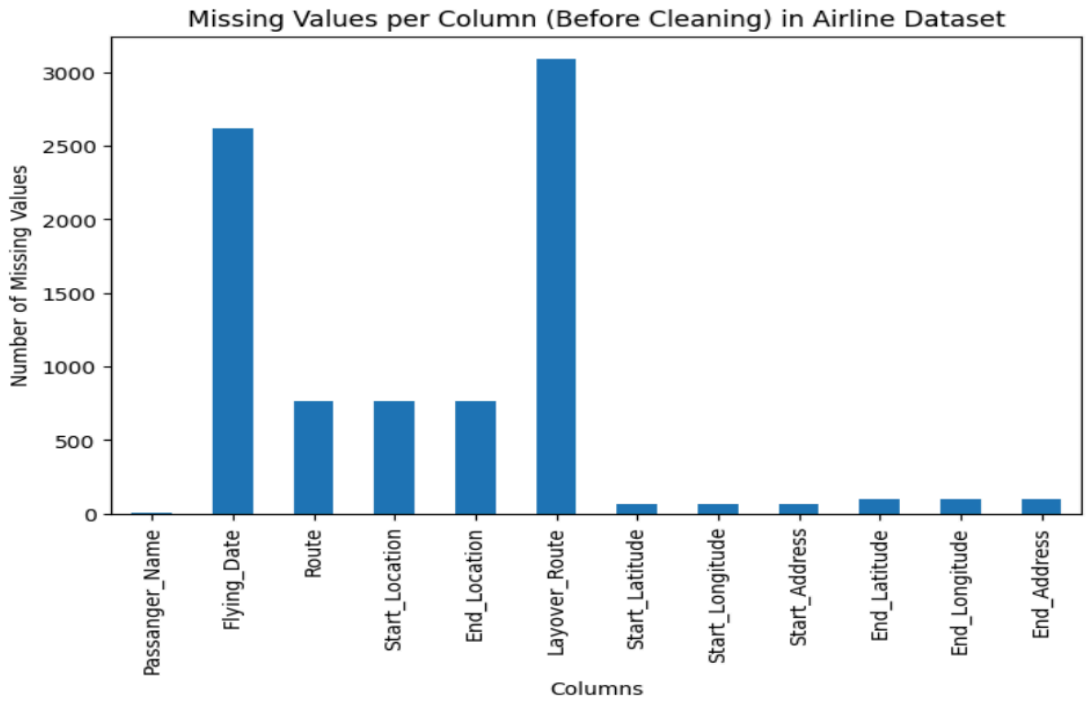
Objective: To ensure data consistency and integrity before conducting analysis or model training.

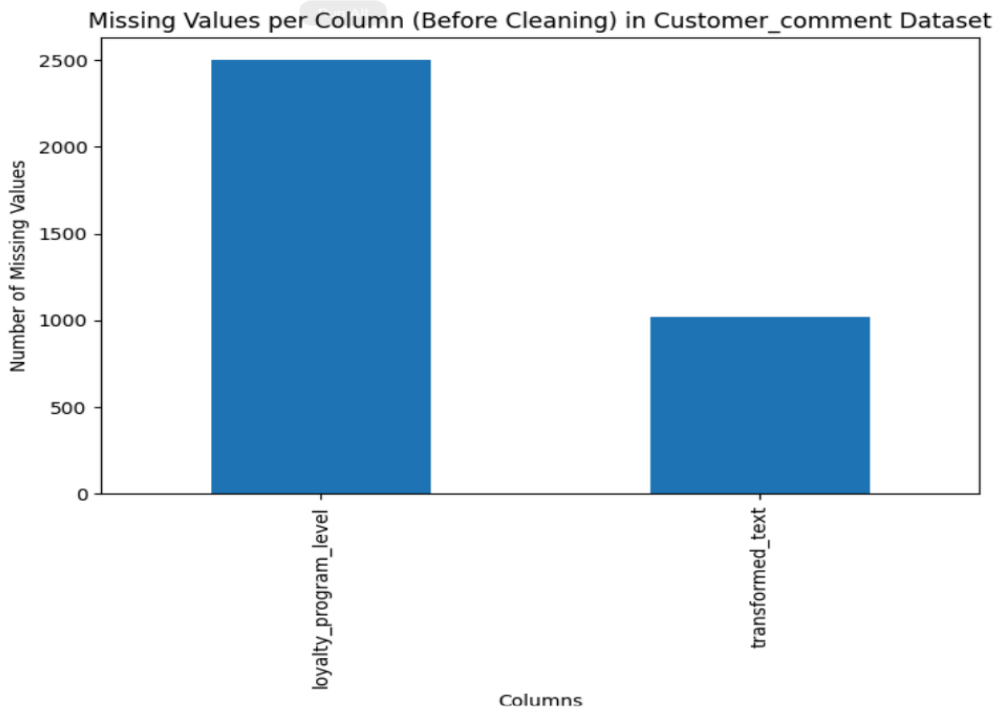
Steps Taken:

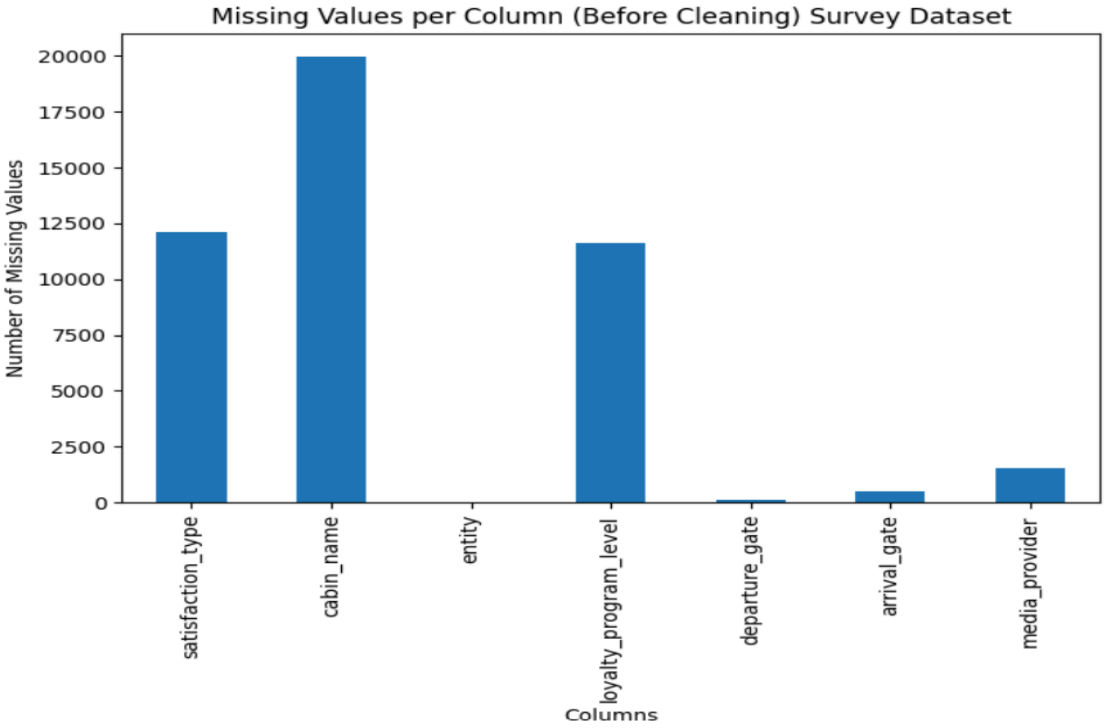
- Removed unnecessary columns.
- Handled null values.
- Removed duplicates.
- Standardized data types.

Justification: Cleaning was essential to ensure accurate sentiment analysis and statistical modeling. Missing or duplicated reviews could distort route popularity and traveler satisfaction results.

Visualization:




Missing Values per Column (Before Cleaning) in Customer_comment Dataset
Number of Missing Values
2500
2000
1500
1000
500
0
loyalty_program_level
transformed_text
Columns


Missing Values per Column (Before Cleaning) Survey Dataset
Number of Missing Values
20000
17500
15000
12500
10000
7500
5000
2500
0
satisfaction_type
cabin_name
entity
loyalty_program_level
departure_gate
arrival_gate
media_provider
Columns

Survey Dataset Cleaning Summary:

	Before Cleaning	After Cleaning
Rows	47074	47026
Duplicates	0	0
Missing Values	45873	0

Passenger_booking Dataset Cleaning Summary:

	Before Cleaning	After Cleaning
Rows	50002	49283
Duplicates	719	0
Missing Values	0	0

Customer_comment Dataset Cleaning Summary:

	Before Cleaning	After Cleaning
Rows	9424	9415
Duplicates	0	0
Missing Values	3523	0

Airline Dataset Cleaning Summary:

	Before Cleaning	After Cleaning
Rows	3575	3501
Duplicates	74	0
Missing Values	8487	0

2. Data-Engineering Questions

- **Sentiment Analysis:** To better understand customer opinions and experiences, we conducted sentiment analysis on the review content using the VADER (Valence Aware Dictionary for Sentiment Reasoning), a lexicon- and rule-based tool that is particularly effective for analyzing text and short reviews.

1. **Calculate Sentiment Scores:**
   Each review in the dataset was analyzed using VADER's SentimentIntensityAnalyzer() to compute a compound sentiment score, ranging from -1 (very negative) to +1 (very positive). Reviews with missing content were assigned a score of 0.

2. **Classify Sentiment Labels:**
   To simplify interpretation, the numerical scores were converted into categorical labels:
   - **Positive**: compound score ≥ 0.05
   - **Negative**: compound score ≤ -0.05
   - **Neutral**: compound score between -0.05 and 0.05

3. **Resulting Data:**
   The analysis added two new columns to the dataset:
   - **Sentiment_Score:** the numeric score of the review's sentiment
   - **Sentiment_Label:** the categorical classification (positive, neutral, negative)

This allows us to quantify customer sentiment and combine it with other features such as traveler type, flight class, and route, enabling a deeper exploration of trends and patterns in airline reviews.

- **Top 10 Most Popular Flight Routes:**

  To identify the routes most frequently booked by passengers, we analyzed the Route column in the dataset. Routes labeled as 'Unknown' were excluded to ensure accuracy. The frequency of each route was calculated using the .value_counts() method, and the top 10 most booked routes were selected.

  The results are visualized in a bar chart, where:

  - The x-axis represents the flight routes.
  - The y-axis represents the number of bookings for each route.
  - The height of each bar indicates the popularity of that route.

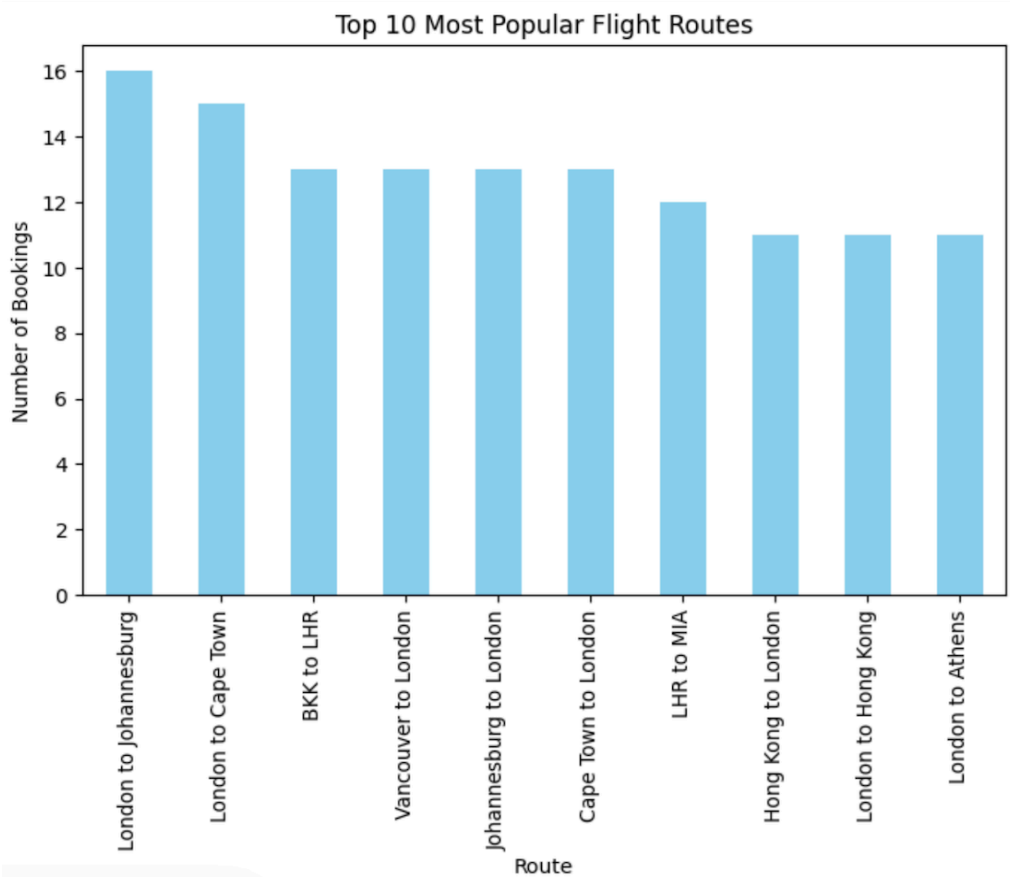


Figure 2.1: The 10 flight routes with the highest number of bookings, with each bar representing a route and its height corresponding to the total number of bookings on that route.

- **Distribution of Bookings Across Flight Hours**

  To understand when passengers most frequently book or take flights, we analyzed the flight_hour column in the dataset. This column represents the hour of the day (0–23) when flights occur. The frequency of bookings for each hour was calculated using the .groupby() and .size() methods.

  The results are visualized in a bar chart, where:

  - The x-axis represents the flight hours (0–23).
  - The y-axis represents the number of bookings during each hour.
  - The height of each bar indicates the volume of bookings at that hour.

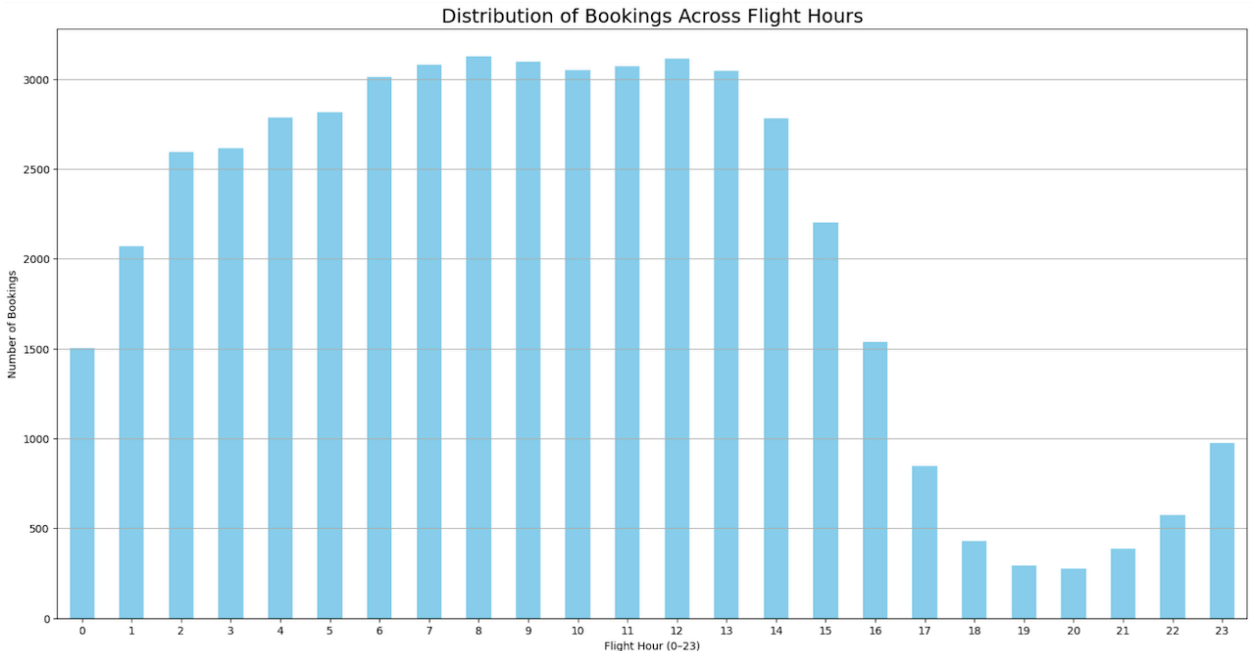


Figure 2.2: The distribution of bookings across different flight hours in a day, with each bar representing a specific hour (0–23) and its height showing how many bookings were made for flights departing at that hour.

- **Average Rating by Traveller Type and Class**

  To analyze passenger satisfaction across different traveler types and flight classes, we calculated the average rating for each combination of Traveller_Type and Class. The dataset was grouped using these two columns, and both the mean rating (avg_rating) and the number of reviews (count) for each combination were computed.

  The data is visualized in a bar chart, where:

  - The x-axis represents the traveler types.
  - The y-axis represents the average rating for each combination.
  - Different classes are shown using color hues.
  - Numerical values on top of each bar indicate the exact average rating.

Figure 2.3: The average ratings for each traveler type across different flight classes, showing how satisfaction varies by both the type of traveler and the class they flew in. Each bar represents the mean rating for a combination of traveler type and class, with the numbers on top indicating the precise average rating.

Features Used In The Predictive Model:

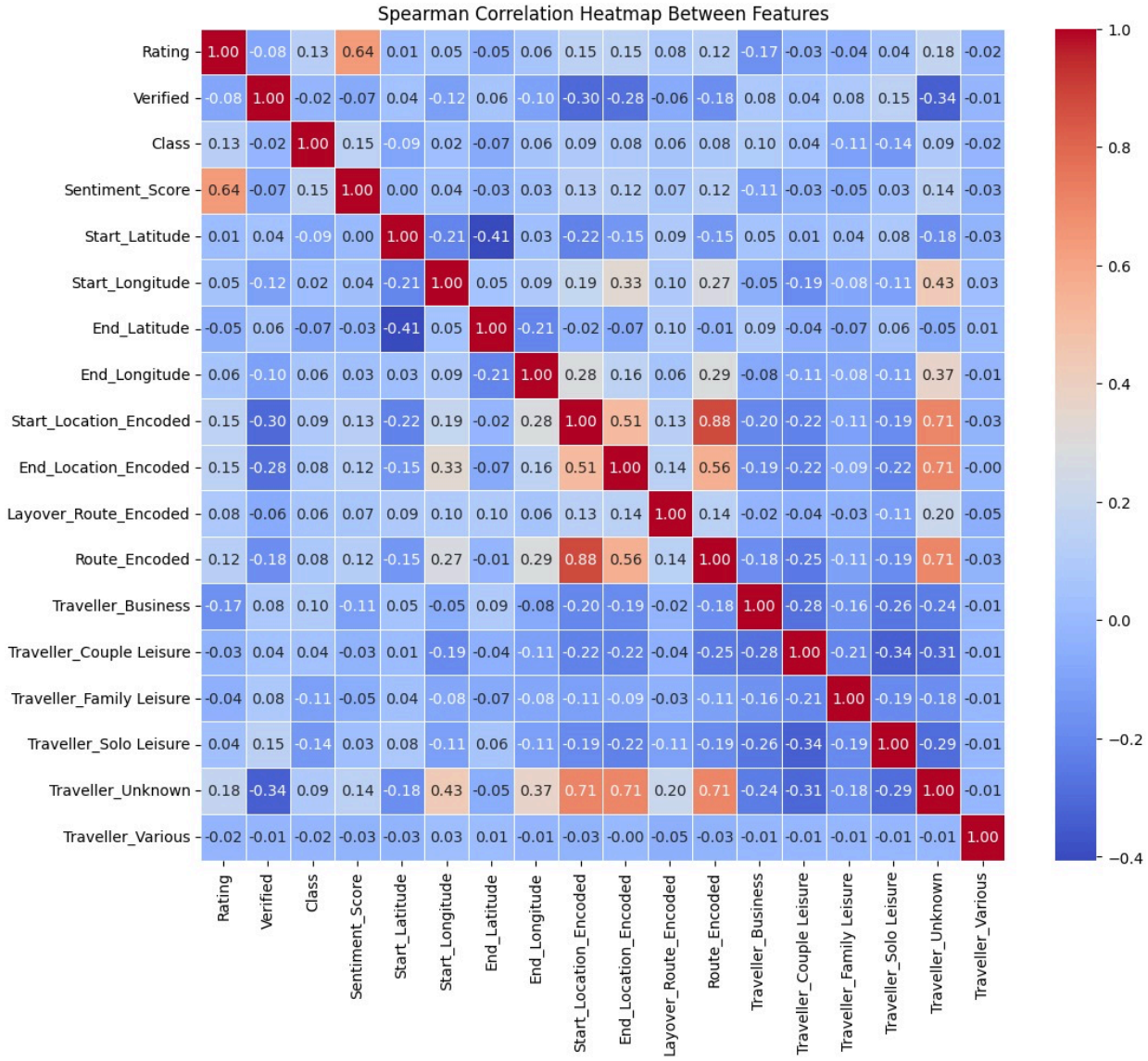
- Verified
- Class
- Sentiment score
- Traveller type
- Start location
- End location

Feature Selection:

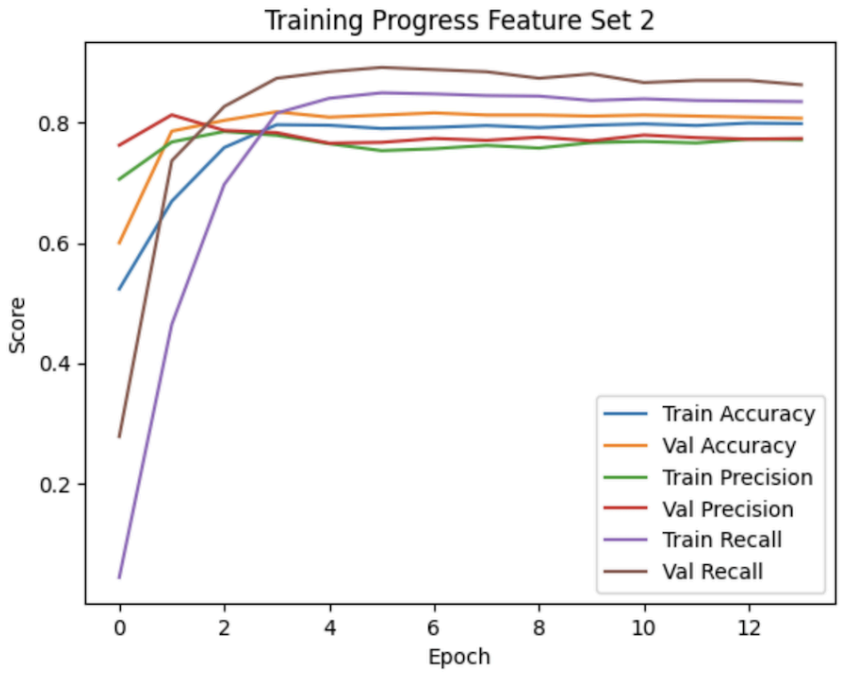
The development of the prediction model involved selecting input features through a blend of exploratory data analysis and iterative experimentation. Initially, a correlation matrix was used to pinpoint features with the strongest relationships to the target variable(Rate), offering insights into their predictive potential. Additionally, a trial-and-error method was employed, testing various feature combinations to assess their individual and combined effects on model performance. This iterative approach helped identify the most relevant and effective features for the final model.

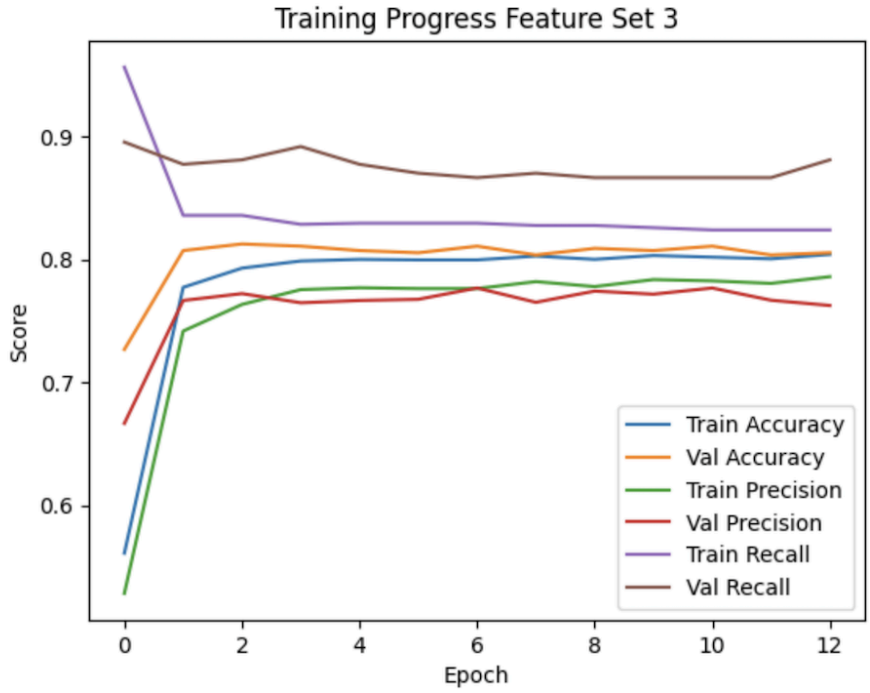
For the model's architecture, multiple configurations were tested by adjusting the number of hidden layers, neurons per layer, and incorporating dropout layers to prevent overfitting. These modifications significantly enhanced the

model’s performance, improving prediction accuracy and minimizing loss.

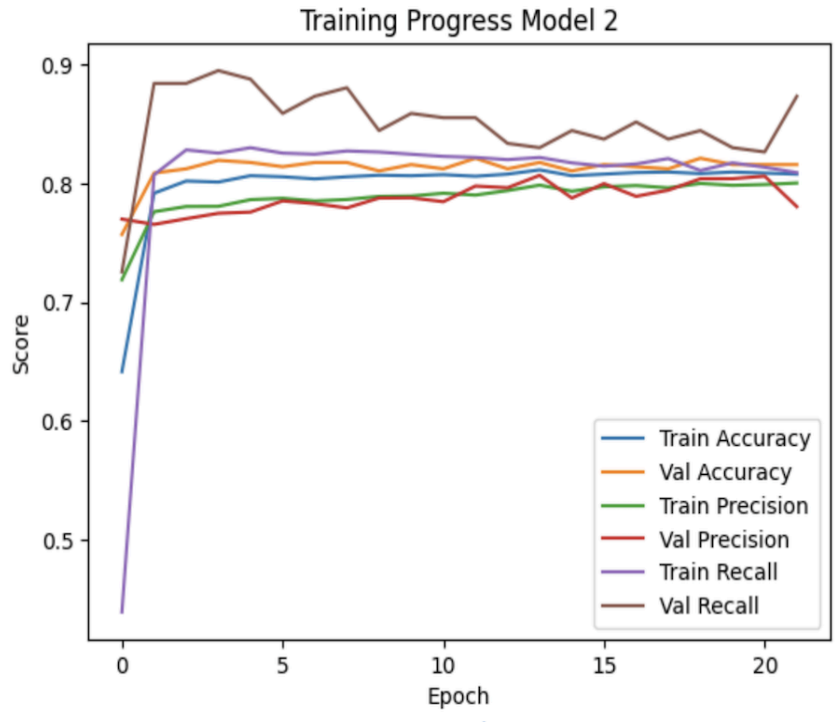


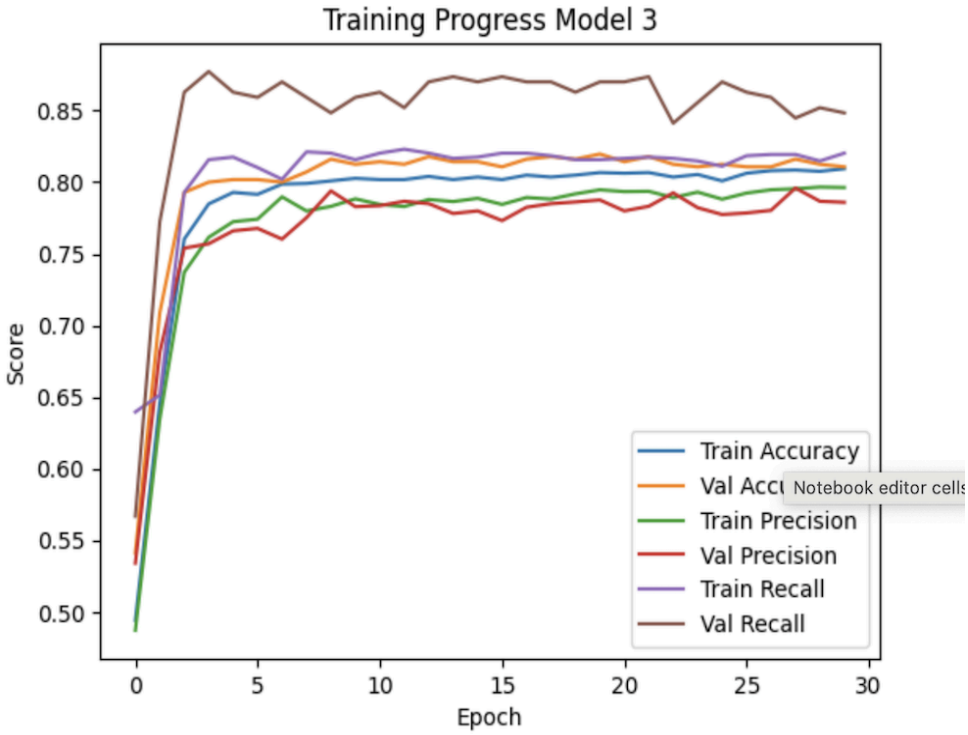
- Here’s our testing for different datasets:


Training Progress Feature Set 2
Score
Epoch
0.2
0.4
0.6
0.8
0
2
4
6
8
10
12
Train Accuracy
Val Accuracy
Train Precision
Val Precision
Train Recall
Val Recall

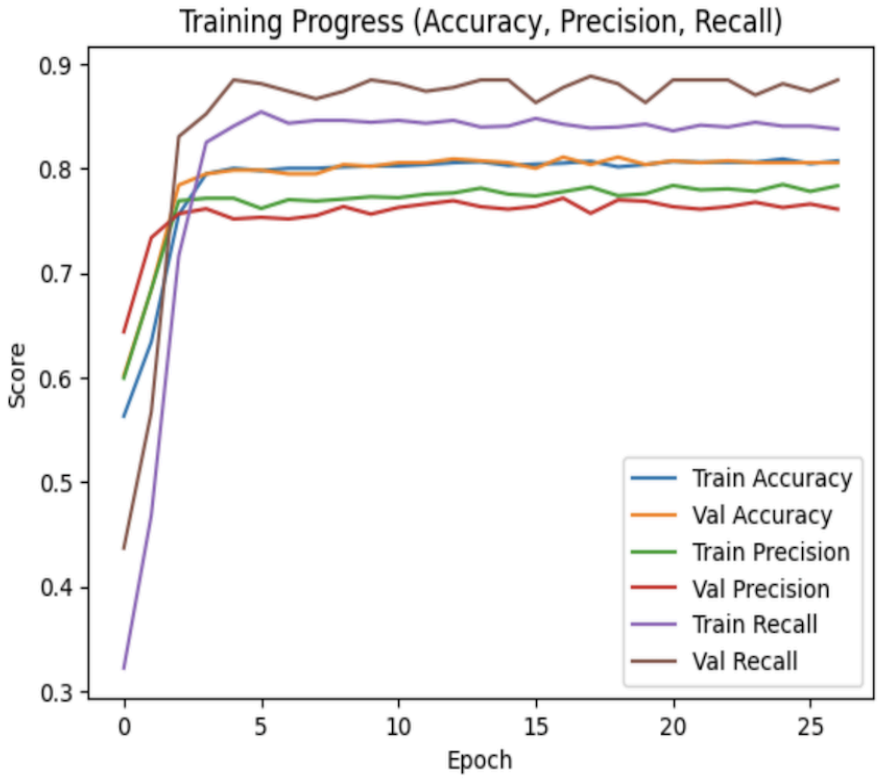

Training Progress Feature Set 3
Score
Epoch
0.6
0.7
0.8
0.9
0
2
4
6
8
10
12
Train Accuracy
Val Accuracy
Train Precision
Val Precision
Train Recall
Val Recall

- And Also we have tried change in in the architecture and observe if it will improve the performance :





- And That is the graph for the model we stick to it:



- After conducting several tests using different feature combination architectural modifications,we found that the model's performance remained relatively stable ,the accuracy shows fluctuation of approximately 1-2%.