ACL REPORT MILESTONE 1

This report documents the steps taken, the insights gained from the Data Engineering phase, the development and comparison of the predictive models, and the model explainability analysis.

Done by :

- Aly Serry 55-5284
- Ali Shaheen 55-5288
- Nour Saber 55-12186
- Donia Fathey 55-17419

1. Data Cleaning and Preprocessing

The primary goal of the cleaning phase was to prepare the 4 initial datasets—`AirlineScrappedReview`, `Customer_comment`, `Passanger_Booking_Data` and `Survey_Data_Inflight_Satisfaction_Score`—for analysis and subsequent predictive modeling, adhering to the project's requirement to remove unnecessary columns and handle missing values/duplicates.

1.1 `AirlineScrappedReview` Cleaning

The `AirlineScrappedReview` data initially contained 3575 entries with significant missing values in several columns, notably **'Flying_Date'** (2620 nulls) and **'Layover_Route'** (3091 nulls).

Steps Performed and Justification:

1. **Column Removal:** Dropped **'Flying_Date'** due to the overwhelming number of missing values (over 73% missing). **'Passanger_Name'** and **'Review_title'** were dropped as they were deemed unneeded identifiers or redundant given the presence of **'Review_content'** and were not listed as required features for the model.
2. **Missing Value Handling (Imputation):** Missing values in **'Layover_Route'** were imputed with **"None"** since nulls likely indicate a direct flight.
3. **Handling "Unknown" Values:** Rows where the column contained the placeholder **"Unknown"** were treated as missing data and subsequently dropped, as class is a critical traveler-related feature for the model.
4. **Final Dropping & Deduplication:** Rows with remaining nulls (e.g., in location coordinates/addresses, 'Route') were dropped, followed by the removal of duplicate rows to ensure a clean, unique dataset for analysis.

1.2 `Customer_comment` Cleaning

The `Customer_comment` data primarily involved handling unnecessary columns and encoding categorical features for future integration.

Steps Performed and Justification:

1. **Column Removal:**
    - **'Unnamed: 0'**: A duplicate index column, providing no analytical value.
    - **'ques_verbatim_text'**: Had only one unique value, making it uninformative for differentiation.
    - **'transformed_text'**: This stemmed/lemmatized text column was dropped because it had missing values (1019 nulls) and we have the original **'verbatim_text'** that was kept for potential re-analysis or feature creation.

2. **Missing Value Handling:** Missing values in **'loyalty_program_level'** were filled with **'None'** to retain these records, as this converts the missing value into a **meaningful and explicit category**, allowing the model to learn the impact of a passenger *not* being part of a loyalty program.
3. **Type Conversion and Deduplication:** The **'scheduled_departure_date'** was converted to a datetime object as this conversion is mandatory for any time-based analysis and modeling, and duplicate rows were removed.

1.3 `Passanger_Booking_Data` Cleaning

The `Passanger_Booking_Data` was relatively clean, with no missing values.

Steps Performed and Justification:

1. No major cleaning steps were required beyond the standard checks for nulls and duplicates (the snippet only shows null check and describe; no explicit drop_duplicates is shown but implied by best practice).

1.4 `Survey_Data_Inflight_Satisfaction_Score` Cleaning

The main challenge was parsing the **'score'** column, which contained both numerical satisfaction scores and categorical food item choices.

Steps Performed and Justification:

1. **Score Splitting:** A new binary column **'is_score'** was created to differentiate between numerical scores and food item text. The original **'score'** column was then split into **'satisfaction_score'** (numerical, converted to Int64, 12094 missing) and **'food_item_chosen'** (categorical, 34357 missing). The original **'score'** column was dropped.
2. **Missing Value Handling:**
    - **'cabin_name'** was dropped due to high missing values.
    - **'loyalty_program_level'** nulls were filled with **'None'**.
    - Rows with nulls in **'departure_gate'** and **'arrival_gate'** (critical identifiers for flight context) were dropped.
    - **'media_provider'** nulls were imputed using the mode (**'PANASONIC'**).

2. Data Engineering Questions and Analysis

The following questions were answered using the cleaned `AirlineScrappedReview` and `Passanger_Booking_Data` datasets.

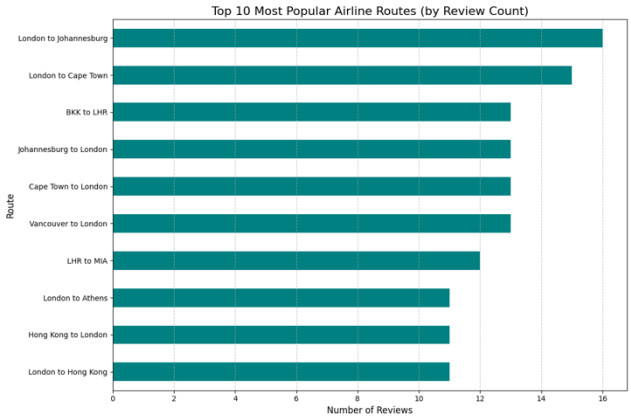
2.1 Sentiment Analysis and Target Variable

1. **Sentiment Analysis:** A **'Sentiment_Score'** column was added to AirlineScrappedReview by running VADER's SentimentIntensityAnalyzer() on the Review_content field. This continuous value (-1 to +1) quantifies the review's emotional tone.
2. **Target Variable:** The binary target variable **'Satisfaction'** was created from the Rating column[5].
    - **Satisfied** (Satisfaction = 1) if Rating > = 5.
    - **Dissatisfied** (Satisfaction = 0) if Rating} < 5.

2.2 Top 10 Most Popular Flight Routes

The routes with the highest number of customer reviews are dominated by long-haul destinations, most originating or terminating in London:

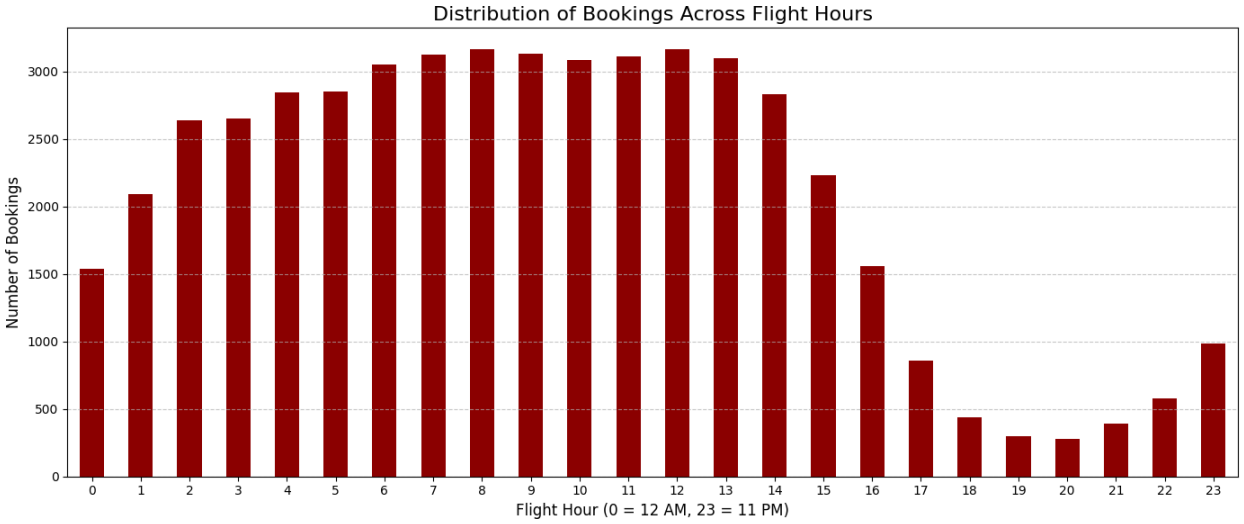
1. **London to Johannesburg** (16)
2. **London to Cape Town** (15)
3. **BKK to LHR** (13)
4. **Johannesburg to London** (13)
5. **Cape Town to London** (13)
6. **Vancouver to London** (13)
7. **LHR to MIA** (12)
8. **London to Hong Kong** (11)
9. **Hong Kong to London** (11)
10. **London to Athens** (11)

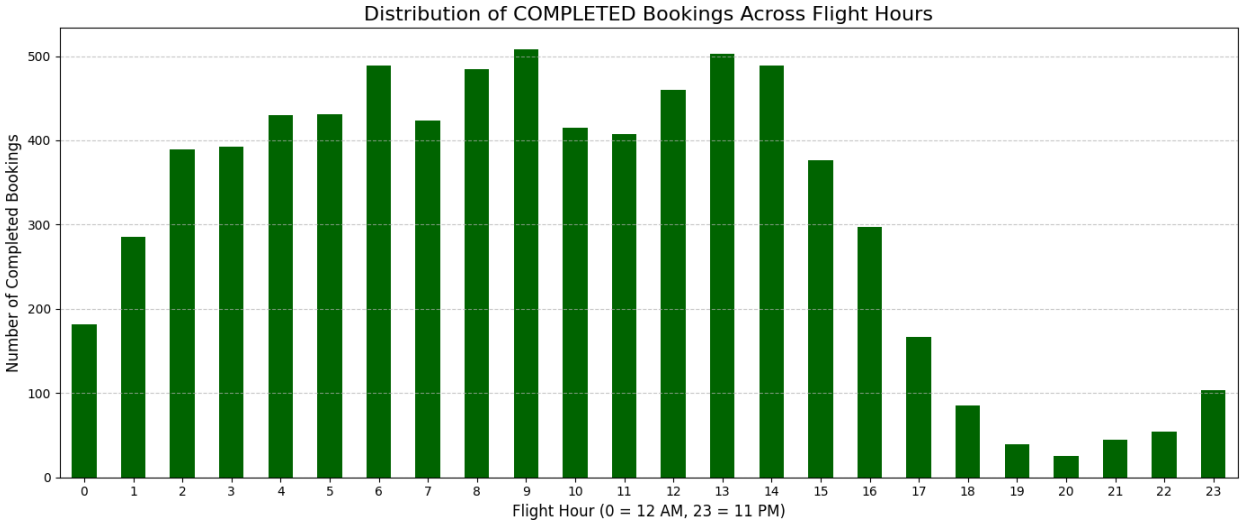


2.3 Distribution of Bookings Across Flight Hours

- **All Bookings:** The distribution shows consistent high volume during the **morning and midday flight hours (6 AM to 2 PM)**.
    - **Peak Activity** occurs broadly around **hour 8 (8 AM)** and **hour 12 (12 PM)** indicating a prolonged period of high demand in the midday window.
    - **Lowest Activity** is observed in the **late night hours**, particularly around **hour 20 (8 PM)**.
- **Completed Bookings:** When focusing only on successfully completed bookings (`booking_complete = 1`), the pattern is similar but features a clearer primary peak.
    - **Absolute Peak Conversion** occurs at **hour 9 (9 AM)**, followed closely by a strong secondary peak at **hour 13 (1 PM)**. This confirms that customer readiness to complete a purchase peaks in the later morning, slightly after the volume peak starts, and again after the typical lunch hour.

- **Lowest Conversion** aligns with the overall lowest activity period, specifically at **hour 20 (8 PM)**.





2.4 Traveler Type and Class Combination Ratings

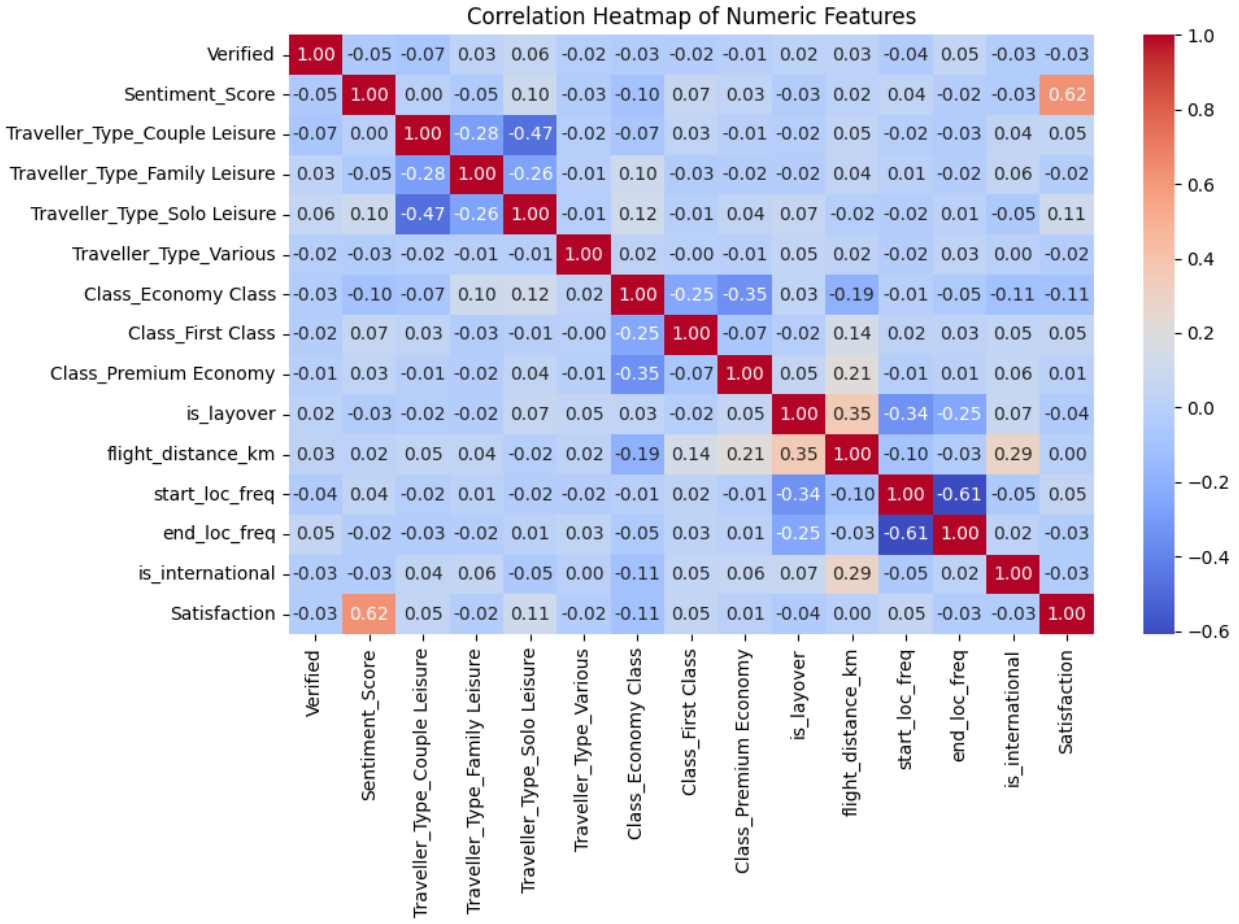
By grouping and averaging the **'Rating'** based on the combination of **'Traveller_Type'** and **'Class'** in the AirlineScrappedReview data (this step was performed to **analyze review patterns** by finding the average customer satisfaction score for every possible pairing of the two most important traveler demographic features), the following extremes were found:

- **Highest Rated Combination: Solo Leisure** in **First Class**, with an average rating of **6.526**.

    - *Justification:* Solo travelers in First Class likely represent a customer segment with high expectations that are generally met by the premium service, leading to the highest average satisfaction.
- **Lowest Rated Combination: Various** in **Economy Class**, with an average rating of **1.000**.
    - *Justification:* The "Various" traveler type is not a distinct segment and may be linked to specific highly negative group or unclassified reviews. Combined with the typically low-frills experience of Economy Class, this combination results in the lowest rating.

2.5 Data Analysis

The heatmap below visualizes the relationships between the numeric features in the dataset. Each cell represents the Pearson correlation coefficient between a pair of variables, where values closer to 1 indicate a strong positive correlation, and values closer to –1 indicate a strong negative correlation.



Focusing on the Satisfaction feature, we observe a moderate positive correlation (0.62) with the Sentiment_Score, suggesting that passengers expressing more positive sentiment are also more likely to report higher satisfaction. Other features show very weak correlations, indicating

minimal direct influence on satisfaction individually. This implies that while sentiment strongly reflects passenger satisfaction, other travel-related factors such as class, traveller type, or flight distance may influence satisfaction indirectly or in combination with non-numeric factors like service quality or experience.

3. Feature Engineering and Pre-processing

The predictive task is a binary classification to predict passenger satisfaction (Satisfaction 0 or 1).

3.1 Feature Selection and Reasoning

The features for **Model 1** and **Model 3** were primarily based on the project requirements.

Feature Engineering and Pre-processing of AirlineScrappedReview

Code Block	Description of Step	Need for the Step
1. Binary Encoding (Verified)	Converts the binary categorical feature Verified into numerical integers: 'Trip Verified' to 1 and 'Not Verified' to 0.	Machine learning models require **numerical inputs**. This efficiently encodes a required binary feature, where verification status is assumed to impact rating credibility.
2. One-Hot Encoding (Traveller_Type, Class)	Converts nominal categorical features into multiple binary columns using pd.get_dummies with dtype=int (0s and 1s) and drop_first=True.	**Required Traveler Features.** OHE ensures the model does not misinterpret these nominal categories as having an unintended ordinal relationship (e.g., $1 < 2 < 3$).
3. Engineered Binary Feature (is_layover)	Creates a new binary column by checking if Layover_Route is present (value 1) or explicitly None (value 0).	**Hypothesis-driven feature:** Layover presence is a crucial logistical factor that often correlates with passenger inconvenience, delays, and a potential drop in satisfaction.
4. Engineered Numerical Feature (flight_distance_km)	Calculates the great-circle distance (in km) between the Start and End coordinates using the Haversine formula.	**Required Flight Feature.** Distance is a vital numerical proxy for flight duration, aircraft type, and the corresponding level of

		expected service (e.g., short-haul vs. long-haul).
5. Engineered Frequency Features (start_loc_freq, end_loc_freq)	Calculates the **frequency (count)** of each unique starting and ending location in the dataset and maps these numerical counts back to the DataFrame.	**Hypothesis-driven feature:** Route volume (frequency) acts as a proxy for operational factors, potentially indicating either streamlined, efficient hubs or congested bottlenecks leading to varied satisfaction levels.
6. Engineered Binary Feature (is_international)	Compares the extracted country names from the full address fields: 1 if countries are different (international) and 0 if they are the same (domestic).	**Required Flight Feature.** Explicitly categorizes the flight type, acknowledging that international and domestic flights involve different service standards and customer expectations.
7. Column Dropping (Cleanup)	Removes all raw source columns (Start_Address, End_Address, coordinates, Layover_Route, Review_content, Route) that were used to derive the new features.	**Data Hygiene and Minimization.** Eliminates redundant, highly correlated, or raw text columns, reducing model complexity and preventing potential feature leakage from the input data.

3.2 Data Pre-processing Steps

1. **Categorical Encoding:**
    - **'Verified'** was mapped to 1 (Trip Verified) or 0 (Not Verified).
    - **'Traveller_Type'** and **'Class'** were converted using **pd.get_dummies with dtype=int** for one-hot encoding. This uses numerical 0s and 1s, which is suitable input for a neural network model, especially when comparing against other numerical features.
2. **Location Feature Engineering (Model 2 Only):**
    - For an experimental approach, the high-cardinality nominal features **'Start_Location'** and **'End_Location'** were dropped and replaced with **Target Encoded** versions: Start_Loc_Avg_Rating and End_Loc_Avg_Rating. This implicitly encodes the location's historical satisfaction trend, acting as a hypothesis-driven feature.
3. **Normalization (Need and Effect):**

- **Numerical features** (Sentiment_Score, flight_distance_km, start_loc_freq, end_loc_freq) were scaled using **StandardScaler**.
- **Need:** Normalization is essential for Neural Networks to prevent features with larger ranges (e.g., distance) from disproportionately influencing the weight updates during training (backpropagation), ensuring faster and stable convergence.
- **Effect:** The scaled features have a mean of approx 0 and a standard deviation of approx 1, standardizing their contribution to the model.

4. Predictive Modeling

4.1 Model Choice and Architecture

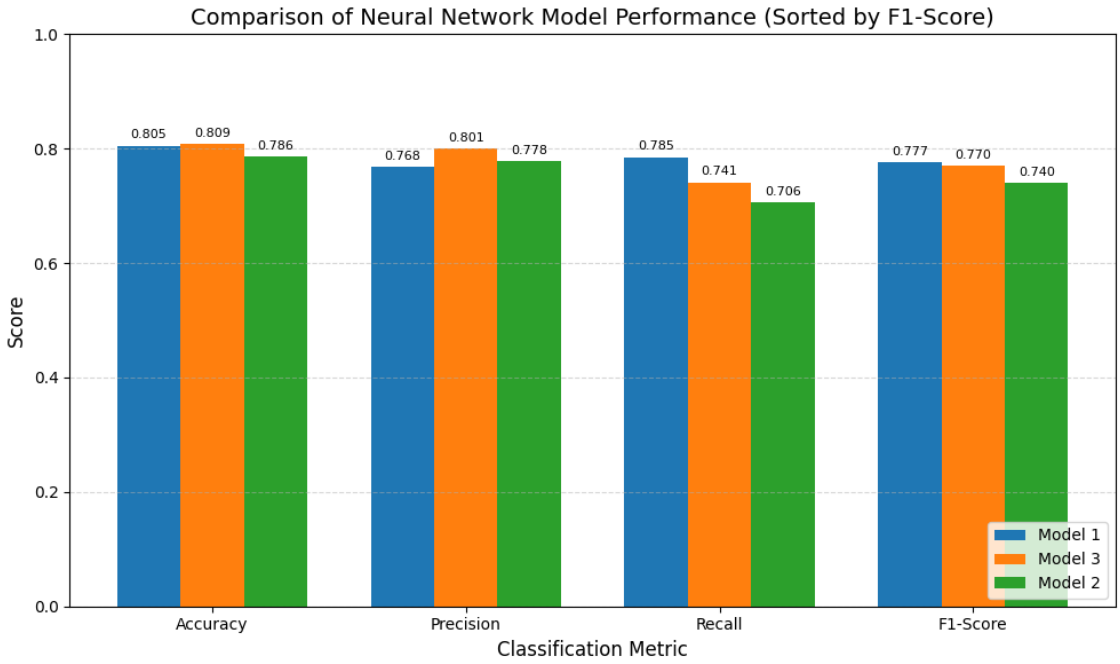
The predictive task is a **binary classification** of passenger satisfaction. Three shallow Feed-Forward Neural Networks (FFNNs) were developed to fulfill the baseline requirement and investigate feature/architecture variations. An **EarlyStopping** callback was used to prevent overfitting.

Model	Input Features	Architecture	Rationale/Limitation
Model 1 (Baseline)	OHE/Binary/Numerical	16-8-4-1 nodes	**Selected Baseline Architecture** (shallow FFNN) with core engineered features.
Model 2 (Feature Test)	Target Encoded Features	16-8-4-1 nodes	**Trial-and-error:** Tests the effect of replacing location features with Target Encoding.
Model 3 (Architecture Test)	OHE/Binary/Numerical	32-16-8-4-1 nodes	**Comparison:** Deeper/Wider network using Model 1 features to justify Model 1's sufficient complexity.

4.2 Model Performance and Comparison

The models were evaluated on the unseen test data using Accuracy, Precision, Recall, and F1-Score[13]. **F1-Score** was prioritized as the key metric for model selection due to its balance between Precision and Recall.

Model	Test Loss	Accuracy	Precision	Recall	F1-Score
Model 1	0.434	0.832	0.814	0.789	**0.802**
Model 3	0.508	0.819	0.766	0.833	**0.798**
Model 2	0.461	0.796	0.770	0.750	**0.760**



Conclusion: Model 1 is the best performing model. It achieved the highest F1-Score (0.802) and accuracy, demonstrating that its architecture is sufficiently complex and its feature set (relying on engineered numerical features over Target Encoding) is the most predictive.

5. Model Explainability (XAI)

Model 1 was selected for explainable AI technique to interpret feature contributions.

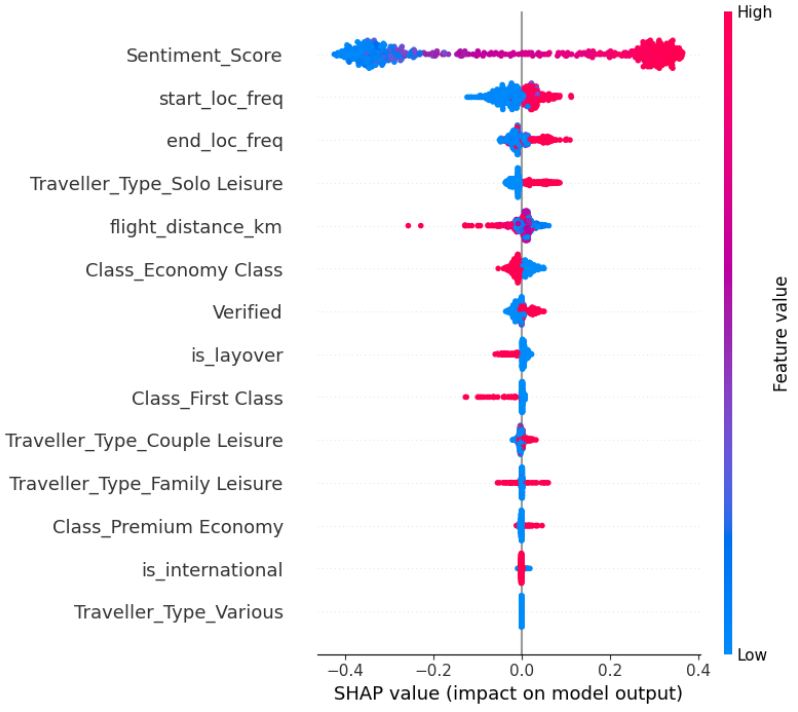
5.1 Global Explanation (SHAP Summary Plot)

SHAP values quantify how much each feature contributes to pushing the model's prediction from the average output (baseline) to its final predicted value. The following bar plot shows the **global impact** (mean absolute SHAP value) of each feature across the training set.

Key Feature Contributions (Based on Magnitude)

1. **Sentiment_Score (Dominant Driver):** This engineered feature, which quantifies the emotional tone of the review text, is the **single greatest predictor** in the model. Its impact is significantly higher (magnitude of **0.2789**) than all other features combined, validating the use of **NLP (VADER) as a core feature engineering step**. A higher (positive) sentiment score will strongly push the prediction toward **Satisfaction (1)**.
2. **Route Volume/Hub Activity (Features at Rank 2 & 3):** The engineered frequency features, **`start_loc_freq` (Rank 2)** and **`end_loc_freq` (Rank 3)**, emerge as the second and third most important global features. This shows that the **operational context of the route (volume)**, determined by how often a specific location is reviewed, is a critical factor in predicting satisfaction, **surpassing all direct traveler or class indicators** in importance magnitude.
3. **Traveller Type and Class (Rank 4 & 6):**
    - **`Traveller_Type_Solo Leisure` (Rank 4)** is the most impactful traveler type, suggesting that **solo travelers are a key segment** influencing the model's output probability.
    - **`Class_Economy Class` (Rank 6)**, although an expected negative driver of satisfaction, has a smaller global impact magnitude than the route frequency features.
4. **Least Important Features: `Traveller_Type_Various` (Rank 14)** has a **zero impact**, meaning the model learned to completely disregard this traveler type for prediction. The **`is_international` (Rank 13)** feature also has a near-zero impact.

\



5.2 Local Explanation (SHAP Force Plot)

This section examines how **Model 1** arrived at its prediction for a specific individual review (Instance 10). The final prediction, expressed as the probability of being Satisfied, is compared to the "SHAP Base Value", which represents the average Satisfaction probability across the training data.

Interpretation for Instance 10

The model starts at the base probability of 0.4392 and must apply feature contributions to reach the final probability of 0.1292.

Features Pushing Prediction Towards DISSATISFIED (Negative SHAP Values):

The model was primarily pushed toward the **Dissatisfied** class by the following features:

1. **Sentiment_Score (-0.2484):** This was the single most powerful negative factor. The feature's low value (negative compound sentiment) immediately signaled a poor review outcome, causing the largest drop in the predicted probability of satisfaction.
2. **start_loc_freq (-0.0255):** The frequency of the starting location had a noticeable negative impact, suggesting this specific starting location may be an unusually congested or problematic hub.
3. **end_loc_freq (-0.0201):** Similarly, the end location's frequency contributed negatively, reinforcing that the route's operational volume negatively affected this passenger's satisfaction.

Features Pushing Prediction Towards SATISFIED (Positive SHAP Values):

The model received some positive pushes, but they were significantly overshadowed by the negative sentiment:

1. **Traveller_Type_Couple Leisure (+0.0375):** For this instance, the traveler being a Couple Leisure exerted the largest positive push, suggesting that this traveler type is generally associated with higher satisfaction for this instance's feature set.
2. **Class_Economy Class (+0.0256):** Although Economy Class is globally a negative driver, for this specific instance, its *presence* pushed the prediction *up*. This is a **local counter-intuitive insight**, meaning that while the average Economy Class experience is poor, this passenger's overall feature combination made their Economy Class flight better than the model expected for that class.

Conclusion

The final prediction of **Dissatisfied (12.92% probability for Satisfied)** is the result of the extremely powerful negative sentiment associated with the review. The collective positive contributions from the traveler type and class were not nearly enough to counteract the significant negative impact from the **low Sentiment Score** and the **high-traffic/problematic location frequencies**.

5.3 Local Explanation (LIME)

Local explanations reveal how an individual instance's feature values combine to produce a specific prediction, often contrasting with the model's average (Base Value). The predicted class for Instance 10 is **Dissatisfied**.

The LIME weights below show the contribution of each feature to the predicted class (**Dissatisfied**). **Positive weights push toward Dissatisfied**, while **negative weights push toward Satisfied**.

Key Factors Driving Prediction for Instance 10

Feature Condition	LIME Weight	Interpretation (Push)
Sentiment_Score (Low)	+0.2670	**Strongest Push to Dissatisfied**. The Normalized Sentiment Score being in the low range <= 0.28 is the single biggest factor confirming the negative prediction.
Traveler Type (Non-Solo/Non-Various)	+0.0738, +0.0677	The conditions Traveller_Type_Solo Leisure} = 0.00 and Traveller_Type_Various = 0.00 contribute significantly to **Dissatisfied**. This implies that belonging to the *unspecified* default group (or being an unclassified non-solo/non-various traveler) is a negative indicator.
Class (Economy Class Not Present)	-0.0653	**Strongest Push to Satisfied**. The condition Class_Economy Class = 0.00 (meaning the customer was in a premium cabin) pushes the prediction **away from Dissatisfied**, acting as a strong mitigating factor.
Traveler Type (Couple Leisure)	-0.0594	Being a **Couple Leisure** traveler (where the feature is 1.00) also strongly pushes the prediction **away from Dissatisfied** (towards Satisfied).

Conclusion

The model is highly confident in predicting **Dissatisfied** for Instance 10 because the extremely negative contribution from the **Sentiment Score** (pushing by +0.2670) vastly overwhelmed the combined positive factors (like the presence of a non-Economy/Couple Leisure traveler, which pushed back by about -0.12). This local analysis aligns with the global insight that the **Sentiment Score is the dominant feature** for Model 1.