SMU Data Science & Analytics

Maersk Case Competition

Team Data? Tada!

Name	Matriculation No.
Jeffrey Chan Zhong Ping	01391932
Joanna Low Yu Ting	01416899
Heng Jun Yong	01372566
Shaun Ng RuiQuan	01372131

Table of Contents

1. Introduction

1.1 Client Introduction

Maersk is a global leader in logistics and transportation. Since its inception in 1904, Maersk has evolved into an extensive network of container shipping, terminal operations, and logistics services. Enabling customers' businesses in various industry sectors, Maersk move 12 million containers every year via multiple transportation channels including, Ocean Transport, Inland Services, Cross Border Rail Transportation, Maersk Air Freight and Less-than-Container Load.

In this Case Competition, our Team has been tasked to analyze and gain valuable insights from multiple fictional datasets. In doing so, provide innovative recommendations to one of their key business lines, Maersk Air Freight – a line that aims to fly goods for their clients efficiently across the globe.

1.2 Problem Statements

1. Identify factors contributing to Customer Satisfaction and identify any methods to predict a customer's loyalty to the Company.
2. Analyze the Voice of Customer dataset (verbatim comments) to investigate any additional new insights or provide focused recommendations.
3. Recommend new and creative ways of fleshing out service improvement insights and explain how such insights can carry over to Maersk's shipping business.

1.3 Dataset Description

Airline Passenger Satisfaction Data

Category	**Field Name**	**Description**
Passenger Demographic and Flight Details	`Gender`	Male or Female
	`Customer Type`	Customer's loyalty to airline. Loyal or Disloyal.
	`Age`	Actual age of passenger
	`Type of Travel`	Purpose of flight. Business or Personal.
	`Flight Distance`	Distance of Journey
Customer Feedback on each Category on a scale of 1 to 5	`Inflight Wi-fi Service`	Wi-fi provided In-flight.
	`Departure/Arrival time convenient`	Time of flights' provided departure & arrival
	`Ease of Online booking`	Accessibility of online booking of flight
	`Gate Location`	Location of gate of flight
	`Food and Drink`	Food & drinks provided In-flight
	`Online Boarding`	Accessibility of online boarding of flight
	`Seat Comfort`	Comfort of the seats In-flight
	`Inflight Entertainment`	Entertainment provided In-flight
	`On-board Service`	Purchasable service provided In-flight
	`Leg Room Service`	Leg room available In-flight
	`Baggage Handling`	Handling of baggage throughout flight
	`Check-in Service`	Feedback score on the Check-in Service
	`In-flight Service`	Service provided In-flight
	`Cleanliness`	Cleanliness of airline experience
Others	`Departure Delay in Minutes`	Delay in departure of flight
	`Arrival Delay in Minutes`	Delay in arrival of destination of flight
	`Satisfaction`	Either 'neutral or dissatisfied' or 'satisfied'
	`Satisfaction Score`	Final score by customer on a scale of 0 to 10

Voice of Customer Data

Field Name	Description
`Satisfaction`	Only ‘neutral or dissatisfied’ as a value
`Satisfaction Score`	Final score by customer on a scale of 0 to 4
`Voice of Customer`	Verbatim feedback of passengers. No verbal feedback is listed as ‘NIL’

2. Methodology

2.1 Data Cleaning and Pre-processing of Airline Passenger Satisfaction Dataset

The “Airline Passenger Satisfaction” dataset is analyzed in Python on “Cleaned_Maersk_Case_US.ipynb”. The notebook file is optimized to be used in the Google Colab environment.

Binning of Features

These features were one-hot encoded to compare with the rest of the numerical features and target.

Age

Ages of passengers were binned into 4 bins. ‘Child’ with ages ranging 0-17, ‘Young Adult’ with ages ranging from 18-35, ‘Adult’ with ages ranging from 36-60, and ‘Senior’ with ages ranging from 61-100.

Flight Distance

Flights were segmented into three categories, “Short-Haul” – less than 1499 miles, “Medium-Haul” – 1499 to 3500 miles, “Long-Haul” – 3500 to 9000 miles. Flight distance was categorized based on the distance-based definitions provided by the European Union in the aviation sector.

Satisfaction

As the “Satisfaction” feature displayed 2 unique values, “neutral or dissatisfied” for satisfaction scores between 0-8 and “highly satisfied” for scores between 9-10, we recategorized the Satisfaction into four groups in order to zero in, more accurately, on the factors that contribute to customer satisfaction. “Highly Dissatisfied” scores between 0 and 4, “Dissatisfied” scores between 5 and 6, “Neutral” scores between 7 and 8, “Highly Satisfied” scores between 8 and 10.

Combining & Rescaling of Features

Departure/Arrival Delay

Assuming that a passenger’s experience in delays is independent of whether the delay is from arrival or departure of the flight, we created a new feature (“Departure/Arrival delay”) that sums up the “Arrival Delay in Minutes” and “Departure Delay in Minutes” and removed the two initial features. We also rescaled this feature to float values between 0 to 5, aligning with the other features.

2.2 Voice of Customer Dataset

The Voice of Customer Dataset are analysed in the files “Sentiment_Analysis_Maersk_Case_Comp.ipynb” and “Voice of Customer.xlsm”. In preparation for the sentiment analysis on text strings, the data was preprocessed on Python, before conducting the analysis on Excel. To reduce noise and ensure consistent text representation throughout the feature “Voice of Customer”, we incorporated (NLP) techniques and topic modeling to achieve this:

Stop word Removal – Common English words that do not contribute to topic identification were removed using NLTK's list of stop words. Punctuation Removal – Eliminate all punctuation marks. Stemming – Words were reduced to their root form using Porter Stemming algorithm.

3. Model Finding

3.1 Analysis on Customer Satisfaction Scores

Subset Data Frame into Flight Length

We decided to segment the data based on relevance to the industry and task. With the improvement of customer service in mind, aligning with major names like Apple, Tesla and Amazon, we broke down the "Airline Passenger Satisfaction" data-frame into 3 subsets based on "Class" – Business, Eco Plus, Eco.

After which, in order to narrow down the 23 independent features for modelling, we defined a function that identifies any variables that had extremely high collinearity (Correlation Coefficient > 0.9). However, the segmented data-frames did not have any variables with high collinearity.

Hence, we furthered the analysis with a High Correlation Filter, where we observe the correlation coefficient of these features with the dependent variable (Satisfaction score) and removed any features that had low correlation to the satisfaction score (Correlation Coefficient < 0.1).

Features Removed from High Correlation Filter

Collinearity and Correlation Filter set at 0.1.

Business Class	Eco Plus Class	Eco Class
1. Departure/arrival delay	1. Departure/arrival delay	1. Departure/arrival delay
2. Flight length	2. Flight length	2. Flight length
3. Gender	3. Age group	3. Age group
4. Departure/arrival time convenient	4. Gender	4. Type of travel
5. Age group (Senior)	5. Gate location	5. Gate location
6. Gate location		

Model Selection and Hyperparameter Tuning

With the final 3 data-frames, we proceeded to do model selection in order to find the best model that can predict the "Satisfaction score" of airline passengers in the 3 categories. The function `run_regression_models()` was defined to run the dataset through 7 regression models:

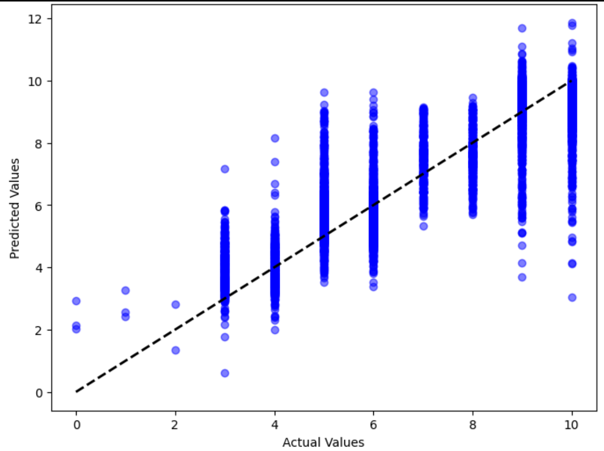
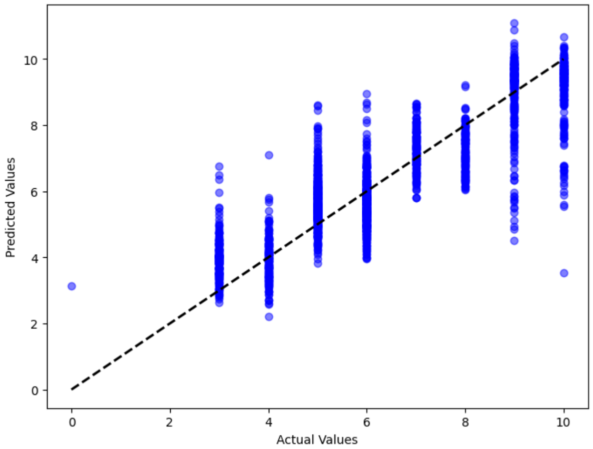
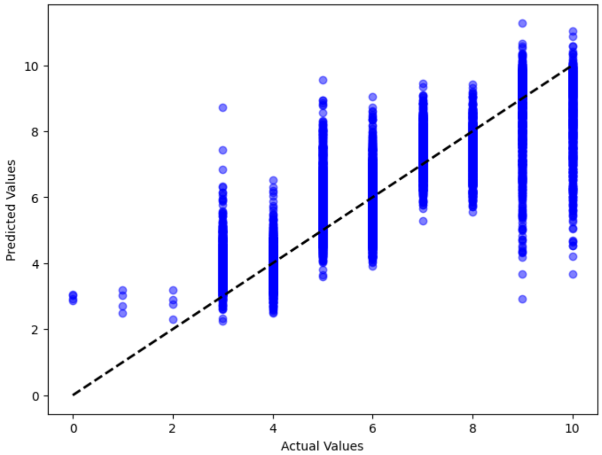
1. Linear Regression – a simple model that can help identify linear relationships.
2. Random Forest Regressor – Ensemble model with decision trees that can handle non-linearity.
3. Gradient Boosting Regressor – Ensemble model boosted to handle complex relationships.
4. Support Vector Regressor – Handles high-dimensional data and regularizes to prevent overfitting.
5. K-Neighbors Regressor – Non-parametric model that performs better for small datasets.
6. LightGBM Regressor – Gradient boosting framework that can handle categorical features.
7. XGBoost Regressor – Gradient boosting framework that handles missing data and outliers.

The function will print out the "Mean Squared Error (MSE)", "Root Mean Squared Error", and "R-squared" of these models. These measurements were chosen to compare regression models as we are facing a regression problem where the target contains continuous values from 1 to 10. A lower MSE or RMSE represents that the models' predictions are closer to the true values, and a higher R-squared represents the goodness of fit and variations in the model.

To improve on the selected model for each dataset, we also performed hyperparameter tuning for the models to improve the robustness and reliability of the final model performance.

Final Model Evaluation and Feature Importance

Having the final models for each class and the hyperparameters tuned for these models, we evaluated the final model and results by printing the actual and predicted values in a graph, along with the RMSE, R-squared value, and the rankings of the feature importance in each category to the "Satisfaction score".

	Feature Importance *Top 10 Features*	**Actual vs. Predicted Values**
Business *XGBoost Regressor* RMSE: 0.80914 R-Squared: 0.85253	1.Online Boarding 2. Inflight Entertainment 3. Type of Travel 4. Inflight Wi-fi Service 5. Customer Type 6. Cleanliness 7. Check-in Service 8. Leg Room Service 9. Ease of online booking 10. On-board Service	
Eco Plus *LGBM Regressor* RMSE: 1.04786 R-squared: 0.75614	1. Inflight Service 2. Baggage Handling 3. Inflight Wi-fi Service 4. Check-in Service 5. Seat Comfort 6. Departure/arrival time convenient 7. On-board Service 8. Ease of online booking 9. Online boarding 10. Inflight Entertainment	
Eco *XGBoost Regressor* RMSE: 1.01341 R-squared: 0.73847	1. Inflight Wi-fi Service 2. Inflight Entertainment 3. Online Boarding 4. Ease of Online Booking 5. Food and Drink 6. Departure/arrival time convenient 7. Cleanliness 8. On-board service 9. Inflight service 10. Seat comfort	

3.2 Analysis on Customer Loyalty

The analysis was done on the “Maersk_Case_Comp_Loyalty.ipynb” file. To identify what makes a passenger loyal or disloyal, we rescaled the column ‘Departure/Arrival delay’ to float values between 0 to 1. The target variable is also changed to the classification of what makes a ‘Loyal’ customer, whereas the original satisfaction score is now classified as an independent variable under numerical features. The categorical features are then encoded using one-hot encoding to allow for the combination of both numerical and categorical features.

Model Selection

Logistic Regression

Logistic Regression is a form of supervised learning algorithm and a parametric model, where the model requires labelled data. In this dataset, the customers are assigned binary values, with loyal as 1 and disloyal as 0. This regression model is used for binary classifications, with the goal of predicting the probability of an input sample belonging to one of the two classes. The variables in the dataset are fitted into the model with the loyalty of the customers already labelled, and the algorithm determines the coefficient of the independent variables from the available dataset. Upon determining the coefficients of the variables, the model can be used to predict the class on new, unseen data.

Random Forest

Random Forest is an ensemble learning algorithm used for classification and regression tasks. It combines multiple decision trees, each trained on random subsets of data and features. By using this model as one of our predictive models to predict customer loyalty on airline services, we would be able to leverage on the terminal nodes to determine the interactions between customer loyalty and the rest of the variables in the dataset. By measuring the impurity reduction achieved by each feature split, we could also determine the significance of each variable in making a more accurate prediction.

Model Evaluation and Results

A summary of the various evaluation metrics for each model are stated below:

Classification Report	Logistic Regression	Random Forest
Accuracy	0.90060	0.98722
Precision	0.92863	0.98098
Recall	0.95132	0.94898
F1-Score	0.93984	0.96472

Based on the coefficients which were found using feature importance, the top 5 variables by feature importance for both models and their importance are as follows:

Logistic Regression	Coefficient	Random Forest	Coefficient
Type of travel_personal	6.164003	Type of travel_personal	0.189063
Flight length_Long-Haul	4.350740	Satisfaction score	0.086527
Satisfaction score	3.632568	Cleanliness	0.064826
Inflight entertainment	2.756662	Age group_Young adult	0.064806
Online boarding	2.511966	Inflight wifi service	0.062575

3.3 Incorporation of Voices

In this analysis, the objective was to identify underlying topics from the qualitative negative comments provided by airline passengers. The findings will be used to solidify the conclusions to our analysis done on Customer Loyalty and Customer Satisfaction.

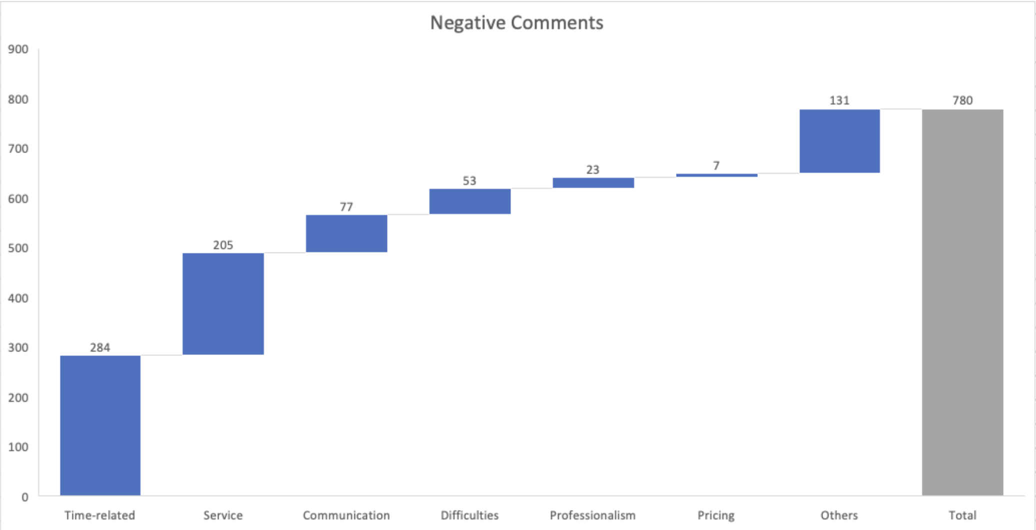
Topic Identification with Python

We started off by trying to find the relevant categories of the data using NLP - a unique index was assigned to every term in a dictionary created using the Gensim library. This data then formed a matrix representing the frequency of terms in documents. Using this matrix, a Latent Dirichlet Allocation (LDA) model was trained to uncover hidden themes within the documents. After 50 iterations, the model identified five key topics (Time-related, Service, Communication, Difficulties, and Pricing). The main words for each topic were listed, and the topics were allocated to the respective comments.

Excel VBA for Categories

We created a VBA function named "CategorizeComment" to categorize customer comments into seven predefined areas of dissatisfaction, such as time, service, pricing, and more. Each category had specific keywords representing unique issues. When a comment is inputted, the function initially labels it as "Other”. It then scans the comment for any of the defined keywords, and if found, assigns the corresponding category. The process ends once a category is matched, and the function returns this category, ensuring that each comment is categorized according to the criteria.

Model Evaluation and Results



Categorization of Negative Comments from Customers

Analyzing the collection of 780 negative comments, insights reveal a distinct pattern in customer dissatisfaction. Time-related issues are the most prevalent, accounting for 284 comments, indicating significant concerns with delays or time management. Service-related complaints follow, with 205 mentions, reflecting potential gaps in quality or responsiveness of service. Communication and difficulties are also notable areas of dissatisfaction, though they are less frequent with 77 and 53 comments respectively. Professionalism, although mentioned in 23 comments, and pricing, noted in just 6 comments, seem to be less concerning for customers.

4. Overall Inference of Analysis

Satisfaction Score

Drawing from the analysis of Customer Satisfaction, there were some common grounds that all three classes of customers factored heavily. The online services such as online boarding, online check-ins, online bookings and even in-flight Wi-fi availability were important to all three classes. However, there were also differences in factors that customers prioritized.

Passengers in Eco Class were more concerned with in-flight services & facilities such as Wi-fi and entertainment. They considered the food and drinks available in-flight heavily, and physical factors like the comfort of the seats.

Those in Eco Plus Class factored in the handling of their baggage & comfort of their seats more significantly than the factors that Eco Class did in theirs.

The Business class passengers prioritized conveniences and efficiency like online boarding. The quality received during the flight, such as cleanliness and legroom experienced was a key factor in their satisfaction as well. In this segment, the loyalty of customers to the airline was an important factor too.

Customer Type

Complementing these findings, the analysis of Customer Loyalty allowed us to understand that passengers taking the flights for “Personal” reasons were more likely to be loyal. One possible explanation is because for personal travelers, they have the freedom of choice to choose between different flight providers, whereas business travelers tend to take whichever airline was provided to them by their company. This makes business travelers less likely to consistently pick one airline over the other, whereas personal travelers do so to stick with familiarity.

Voice of Customer

Among the passengers who were highly dissatisfied, the common trend found was that most of these dissatisfactions were linked to time-related issues with the flight. Despite departure & arrival delays being a feature of low correlation with customer type and satisfaction score, we can infer that it is a deciding factor for customers to swing between a high dissatisfaction and the opposite.

5. Recommendations

In our ideation process, we looked at the similarities between major companies known for their exemplary customer service – Apple, Tesla, Amazon. We then aligned them to the improvement of key features identified to link to a customer’s satisfaction in the fictional airline dataset before bridging these ideas over to Maersk’s Air Freight business line.

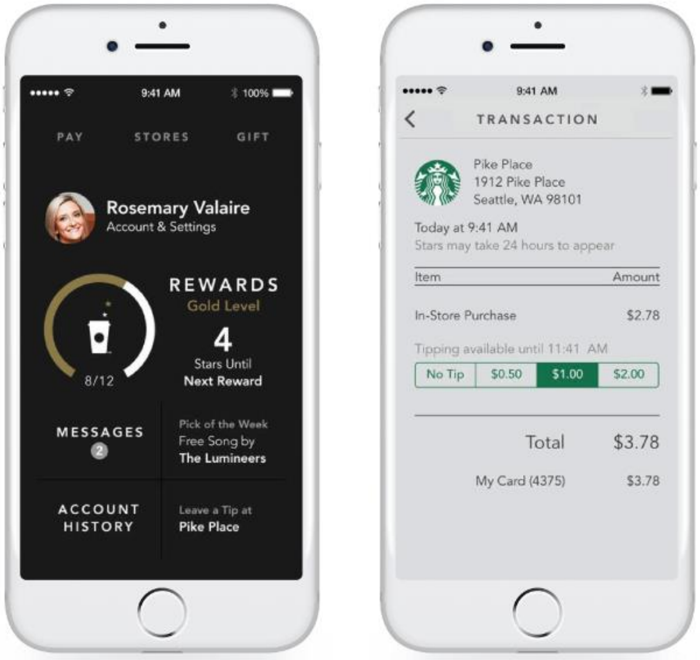
A common ground that these companies boast is a memorable customer experience. Tesla has the most loyal customers of any car company with a satisfaction rating of 90%, of which 80% buy or lease another Tesla as their next vehicle. Apple, likewise, boasts these five steps of service that goes down to each of their staff and customer. With the acronym A-P-P-L-E, every employee is trained to walk a customer through an experience in their stores. The scale of Amazon has allowed them to venture into several innovations to offer clients digital service experiences.

Another resemblance between these three companies is their commitment to transparency. Many of Tesla’s customers were discouraged by the push-back in the delivery of their Model 3, and this was responded with honest, some even personal, updates by Elon Musk. Amazon offers access to unfiltered reviews and ratings to their products, accompanied by a recommender system that helps customers identify which products might suit them based on the terms searched.

5.1 Gamification and Loyalty Programs

Insofar, customers interested in and/or currently using Maersk's Air Freight's business line can conduct most of the online functions that aligned with the passengers of the airline in the fictional dataset. They can connect to a helpdesk, view and purchase packages and track their cargo in real-time. Additionally, Maersk Air Freight offers multiple tiers of service to meet differing global demands – Priority Air, Premium Air, Economy Air.

The key difference that can make Maersk stand out in the Air Freight industry would be to appeal to the customers' loyalty with their brand imaging and unique marketing/positioning. The Starbucks Star Rewards system and application features a gamification of a loyalty system that rewards their loyal customers with free drinks and food during special occasions or upon reaching a certain loyalty level.



Pike Place
1912 Pike Place
Seattle, WA 98101

Today at 9:41 AM
Stars may take 24 hours to appear

Item	Amount
In-Store Purchase	$2.78

Tipping available until 11:41 AM

Total	$3.78
My Card (4375)	$3.78

Maersk Air Freight can replicate this model into their application users to not only boost the number of their customers using the application, but to also attract new leads into using their business line. This collection of data can also help the company enhance the tracking and tracing functionality of each of their customers, to provide personalized recommendations and suggestions for services and shipping options.

5.2 Integration of Complementary Services

As some cause for concern among the passengers of the fictional airline included baggage handling, and for the highly dissatisfied customers, delays in arrival or departure, we suggest that the Maersk Application also integrate complementary services as a strategic move.

On top of providing real-time updates on their cargo tracking, Maersk Air Freight can work with the various stakeholders in the shipment process, external platforms like weather forecasts and port information. By doing so, the company can not only receive notifications of any delays, but also anticipate potential delays or understand the causes for the delays.

The business can also partner with insurers to provide cargo insurance options to protect their shipments and offer peace of mind to customers.

5.3 Create a Strong Online Presence and Community

Maersk Air Freight can also elevate service standards by introducing initiatives for their valued customers together through shared experiences, valuable insights, and meaningful interactions. By capturing the key increases in connectivity and cultivation of influence, information and ethics will spread to assist the business line in gaining more ground in the market.

With the motivation of improving the digital service experience of clients, Maersk can engage in user-generated content contests with attractive rewards, driving greater engagement within their community. Having forums or channels to facilitate knowledge sharing allows customers to foster a sense of belonging and camaraderie.

Embracing these customer-centric initiatives is a cost-effective way to poise Maersk Air Freight into being more than just a logistics service provider.

6. Conclusion

With data-driven insights and innovative recommendations, Maersk Air Freight is poised to elevate the air freight experience. In this project, we have unearthed valuable insights from the provided fictional datasets on the passengers of an airline.

We studied and learned from the pioneers and exemplars of customer service such as Apple, Tesla and Amazon, and delved into key drivers of customer satisfaction and loyalty to gain critical insights that we can bring over to Maersk Air Freight.

While the analysis serves as a solid foundation, it is essential to acknowledge the limitations. As the dataset was used to represent a fictional group of passengers of an airline, it does not fully encompass the complexities of an air freight operation. Implementing recommendations requires perfect knowledge of both industries and understanding the unique challenges of an air freight company.

Along with the analysis, we ideated three creative ways for Maersk to embrace a memorable customer journey and transparency. By introducing gamification and loyalty programs, integrating complementary services into their applications, and creating a strong online presence & community.

Challenges may arise in integrating complementary services & collaborating with external stakeholders. Ensuring customer adoption of gamification and loyalty programs requires effective communication and incentivization. Data privacy and cyber security must be addressed when building an online community. The key is to allocate resources effectively and ensure continuous innovation to maintain a competitive edge.