University of Nottingham
UK | CHINA | MALAYSIA

Investigating Dynamic Brand Perception in the Airline Industry: A Temporal Analysis of Feature Drift and Interaction in Customer Reviews of British Airways and Emirates Airlines

by

Samiksha Kamath
20703562

2025

A Dissertation presented in part consideration for the degree of MSc. Business Analytics

---

Abstract

Airline brand perception is a critical determinant of customer loyalty, pricing power, and resilience during disruption. The COVID-19 pandemic created unprecedented shocks for the aviation industry, shifting passenger priorities from comfort and service toward safety, refunds, and operational reliability. This study investigates how brand perception evolved across crisis phases for British Airways (BA) and Emirates, using online customer reviews (OCR) as the primary data source. Unlike traditional survey-based or static sentiment analyses, the research applies advanced explainable machine learning techniques—Model Class Reliance (MCR), feature drift analysis, and feature interaction analysis—to move beyond prediction toward explanatory insight.

The methodology integrated structured sub-ratings (e.g., seat comfort, food quality, ground service) with text-derived features (sentiment polarity, topic modelling, emotion scores) to predict overall passenger ratings. MCR was used to establish model-agnostic reliance bounds across the Rashomon set of near-optimal models, with emphasis on the conservative upper bound (MCR^{+}) as a stable measure of feature necessity. Drift analysis then quantified temporal changes in reliance distributions across pre-, during-, and post-COVID periods, validated through non-parametric statistical tests. Interaction analysis, supported by permutation-based inference, assessed how attributes reinforced or offset each other in shaping evaluations.

The findings show that BA exhibited abrupt and crisis-sensitive drift: refund and disruption-related features spiked during COVID before receding, while ground and staff service reasserted importance post-crisis. Emirates, by contrast, displayed incremental reweighting anchored in sentiment-driven attributes such as staff and entertainment quality, reflecting a more stable explanatory structure. Interaction effects revealed compounding penalties for BA when multiple service domains failed, and reinforcing synergies for Emirates across sentiment features.

Overall, the study demonstrates that reliance intervals, drift trajectories, and interaction effects provide a transparent and robust explanatory account of brand perception dynamics. By moving beyond conventional sentiment and predictive modelling, this research contributes both theoretically and practically to understanding how airline reputations are stress-tested and reshaped during crises

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

The aviation industry is one of the most brand-sensitive sectors in the global economy. Airlines compete not only on price and network coverage but also on intangible factors such as service quality, safety, and reputation, which collectively shape customer loyalty and market differentiation (Doganis, 2019). In contrast to commodity markets, where cost tends to dominate consumer choice, air travel decisions often hinge on perceived reliability and experience. This makes brand perception a critical strategic asset that influences repeat bookings, willingness to pay, and resilience in times of disruption.

The outbreak of COVID-19 marked the most severe shock in modern aviation history, with global passenger demand collapsing by more than 60% in 2020 (IATA, 2021). Beyond the financial losses, the pandemic altered consumer expectations: safety, hygiene, and flexibility replaced traditional priorities such as price and in-flight service as dominant evaluative criteria (Suau-Sanchez et al., 2020). Airlines were compelled to manage unprecedented levels of cancellations, refund disputes, and operational uncertainty, all of which directly shaped how passengers evaluated brands. In this context, reputation became not just a competitive differentiator but a determinant of survival and recovery.

This study examines the evolution of brand perception for British Airways and Emirates, two airlines with distinct strategic positions. British Airways, a legacy European flag carrier, has traditionally emphasised heritage and connectivity but has faced reputational challenges related to customer service and crisis handling. Emirates, by contrast, has cultivated an image of luxury, innovation, and premium service, consistently positioning itself at the higher end of the market. Comparing these airlines across the pre-, during-, and post-pandemic phases provides an opportunity to understand how brand equity is stress-tested under crisis conditions and how recovery pathways diverge across carriers.

The analysis draws on online customer reviews (OCR) as a rich, real-time source of consumer insight. Unlike traditional surveys, OCR combines numerical ratings with textual feedback, enabling multidimensional assessment of both *what* passengers rated and *why*. In this study, structured sub-ratings such as seat comfort, food quality, and ground service are integrated with text-derived features, including sentiment polarity, emotional tone, and topic distributions. This combined dataset provides a holistic foundation for evaluating brand perception.

Where most existing research has focused on either descriptive text analysis or predictive modelling of overall ratings, this study advances the field by incorporating

explainable machine learning techniques. Specifically, it employs Model Class Reliance (MCR) to establish model-agnostic feature importance, feature drift analysis to capture temporal shifts in drivers of perception, and feature interaction analysis to uncover how service attributes combine to influence evaluations. These methods move the analysis beyond prediction accuracy toward explanatory robustness, addressing the limitations of widely used approaches such as SHAP or permutation importance, which are often unstable, model-dependent, and static (Molnar, 2019).

By adopting this framework, the study contributes both academically and practically. Academically, it extends research on brand equity and crisis management by showing how service attributes fluctuate in salience across crisis stages. Practically, it provides airlines with diagnostic insights into which aspects of service to prioritise when rebuilding trust and differentiating themselves in a highly competitive, reputation-sensitive market.

1.1 Research Focus and Rationale

The study is motivated by three considerations. First, while the importance of brand perception in aviation is widely acknowledged, most analyses remain limited to static surveys or single timeframes, leaving little understanding of how consumer priorities evolve in response to crisis. Second, existing OCR-based studies in aviation are largely descriptive or predictive, with limited explanatory power. Third, methods that can capture temporal drift and feature interactions remain underexplored in this domain, despite their clear relevance to understanding dynamic consumer decision-making.

This research therefore provides a novel contribution by combining structured and unstructured OCR features with advanced explanatory methods to map the evolution of airline brand perception across time and context.

1.2 Research Objectives

The central aim of this study is to evaluate whether advanced explainable machine learning techniques, specifically Model Class Reliance (MCR), feature drift analysis, and feature interaction analysis, can provide validated explanatory insights into airline brand perception across time. While prior research has relied heavily on static sentiment analysis or model-specific importance measures, this study seeks to test whether these methods can move beyond prediction to explanation.

The study is designed to achieve the following objectives:

Objective 1: Predict overall passenger ratings by integrating structured sub-ratings (e.g., seat comfort, ground service, food quality) with text-derived features such as sentiment polarity, emotion scores, and topic modelling.

Objective 2: Apply Model Class Reliance (MCR) to establish model-agnostic bounds of feature importance across the Rashomon set of equally well-performing models. Unlike SHAP or permutation importance, which are tied to single-model artefacts and prone to instability, MCR provides validated upper and lower bounds that identify features consistently influential across models (Fisher et al., 2019).

Objective 3: Conduct feature drift analysis, grounded in MCR-identified stable features, to examine how the drivers of passenger evaluations shifted across pre-, during-, and post-COVID phases. Drawing on the concept drift literature (Gama et al., 2014), this analysis captures the temporal evolution of consumer priorities under crisis and recovery conditions.

Objective 4: Investigate feature interactions, again using MCR-screened features, to assess how combinations of service attributes (e.g., safety with flexibility, comfort with reliability) jointly shaped brand perception. Since consumer decision-making rarely depends on single attributes in isolation, interaction analysis (Aas et al., 2021) provides a more behaviorally realistic account of how passengers evaluate airlines.

By integrating these objectives, the study addresses the overarching research question:

To what extent can Model Class Reliance (MCR), feature drift analysis, and feature interaction analysis provide explanatory—rather than purely predictive—insights into the evolution of airline brand perception across crisis phases?

2. Literature Review

2.1 Brand Perception and Service Quality in Aviation

Brand perception is a critical determinant of competitive advantage in the airline industry, which is widely regarded as one of the most reputation-sensitive service sectors (Doganis, 2019). Unlike commodity markets where price tends to dominate consumer choice, airline selection reflects a complex interplay of tangible service quality attributes and intangible perceptions such as trust, safety, and heritage. This aligns with established customer-based brand equity frameworks. Aaker (1991) identifies brand loyalty, awareness, associations, and perceived quality as the key drivers of equity, while Keller (1993) conceptualises equity in terms of brand knowledge structures that influence consumer decision-making. These frameworks have been widely applied in aviation to explain how passengers' perceptions translate into loyalty, willingness to pay premiums, and resilience in the face of competition (Hanlon, 2007; O'Connell and Williams, 2011).

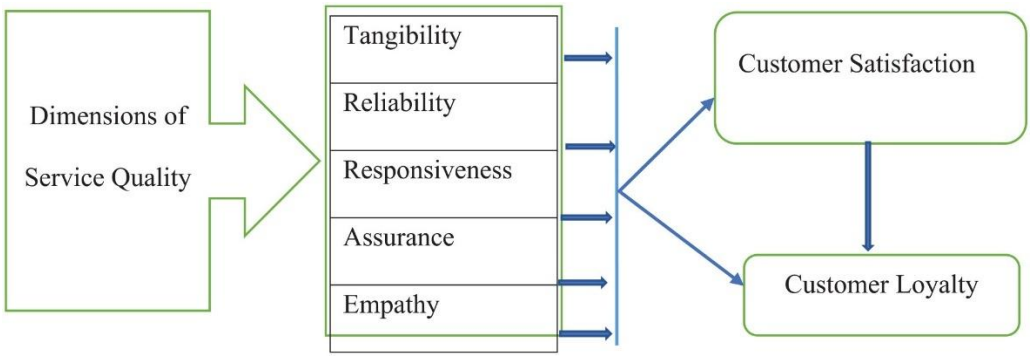


Figure 2.1. SERVQUAL model linking dimensions of service quality to customer satisfaction and loyalty (adapted from Fida, 2020).

Service quality, a cornerstone of brand perception, is frequently operationalised through the SERVQUAL frameworkin, which specifies five dimensions: reliability, responsiveness, empathy, tangibles, and assurance (Parasuraman, Zeithaml and Berry, 1988). Within aviation, empirical studies confirm that reliability (punctuality, schedule adherence), assurance (safety and professionalism), and tangibles (modern aircraft, lounges, comfort) carry disproportionate weight compared to other service industries (Liou and Tzeng, 2007). These findings underscore that airline brands are not constructed primarily through marketing communications but through passengers' lived experiences across multiple service touchpoints.(Figure 2.1)

Crisis contexts amplify the importance of brand perception. The COVID-19 pandemic triggered the largest shock in aviation history, reducing global passenger demand by

more than 60% in 2020 (IATA, 2021). Research demonstrates that hygiene, safety, refund handling, and communication transparency quickly displaced price and comfort as dominant evaluation criteria (Suau-Sanchez, Voltes-Dorta and Cugueró-Escofet, 2020). Budd et al. (2021) highlight how refund disputes and operational disruption significantly damaged consumer trust in some carriers, particularly legacy airlines such as British Airways, while premium carriers like Emirates leveraged reputations for service and safety to mitigate reputational harm. These findings reinforce that brand equity is dynamic, evolving in response to exogenous shocks, and that perception cannot be treated as static.

2.2 Online Customer Reviews (OCR) in Tourism and Aviation

The rise of digital platforms has transformed the ways in which consumers evaluate and share experiences in tourism and aviation. Online customer reviews (OCR) have become central to consumer decision-making, often exerting more influence than traditional marketing communications or advertising (Hu, Pavlou and Zhang, 2009; Filieri, 2015). Unlike conventional surveys, which are structured, infrequent, and sampling-dependent, OCR offers unsolicited, large-scale, and continuous feedback, providing a real-time window into consumer sentiment (Sparks and Browning, 2011).

Research in tourism confirms the strategic value of OCR. Sparks and Browning (2011) demonstrate that review valence and volume significantly influence booking intentions, while Park and Nicolau (2015) show how OCR affects perceptions of brand credibility and trust. In aviation specifically, OCR offers distinct advantages over satisfaction surveys, which often suffer from low response rates and time lags (Budd, Ison and Adrienne, 2021). Longitudinal review data capture how consumer perceptions evolve in response to operational events and crises.

Nevertheless, OCR presents challenges. Reviews tend to be J-shaped, over-representing extreme positive and negative experiences (Hu, Pavlou and Zhang, 2009). Manipulation by fake or incentivised reviews remains a concern (Ott et al., 2011). Furthermore, the unstructured nature of textual content complicates analysis, requiring advanced natural language processing (Zhang, Zhao and Xu, 2016). OCR thus provides rich but imperfect insight, which must be treated with caution.

Applications of OCR in aviation have largely been descriptive or predictive. Early studies focused on sentiment analysis and topic modelling to summarise passenger evaluations (Baker, Donthu and Kumar, 2016). Later research applied machine learning to predict overall ratings from structured and textual features (Archak, Ghose and Ipeirotis, 2011; Zhang, Zhao and Xu, 2016). While such approaches confirm that OCR contains sufficient signal to forecast ratings, they prioritise

prediction over explanation. They reveal what overall ratings are likely to be but not why they take a given form or how the drivers of perception change across contexts.

2.3 Predictive Modelling on OCR

The integration of machine learning into OCR analysis has significantly advanced predictive accuracy. Early econometric approaches such as regression quantified relationships between textual features (e.g., sentiment polarity, review length) and overall satisfaction (Archak, Ghose and Ipeirotis, 2011). These approaches provided interpretable results but were constrained by assumptions of linearity and independence.

With the rise of ensemble methods and deep learning, predictive models such as Random Forests, gradient boosting, and neural networks have become standard in OCR analysis (Zhang, Zhao and Xu, 2016). These methods excel at handling heterogeneous, high-dimensional data, integrating numerical ratings, categorical encodings, and textual embeddings. In aviation, predictive modelling has been applied to forecast satisfaction scores, identify likely detractors, and prioritise service improvements (Baker, Donthu and Kumar, 2016).

However, predictive success comes at the cost of explanatory power. First, these models focus on accuracy rather than causal insight, meaning they do not reveal the mechanisms behind ratings. Second, they often operate as "black boxes," leaving managers without interpretable guidance for decision-making. Third, they assume stability over time, yet evidence shows that models trained on pre-crisis data quickly lose accuracy in crisis contexts, highlighting the presence of concept drift (Widmer and Kubat, 1996; Gama et al., 2014).

Thus, predictive modelling demonstrates the feasibility of mining OCR data but fails to address explanatory and temporal dimensions of brand perception.

2.4 Explainability in OCR Models

The limitations of black-box models have spurred interest in interpretability and explainability. Common methods include permutation importance (Breiman, 2001), Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh and Guestrin, 2016), and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). These methods decompose predictions into contributions from individual features, either globally across datasets or locally for specific instances.

Permutation importance provides intuitive global measures but is unstable when predictors are correlated (Hooker, Mentch and Zhou, 2021). LIME generates

interpretable local approximations but is sensitive to sampling and parameterisation. SHAP, grounded in cooperative game theory, attributes feature contributions additively, offering more robust explanations but still vulnerable to collinearity and model dependency (Molnar, 2019).

While these tools represent progress, their limitations are significant for brand perception research. They typically explain a single fitted model, reflecting one specification rather than the range of equally plausible models. Their static nature also precludes temporal analysis, and they provide little capacity to identify interactions between features. In aviation, where perceptions shift dynamically and attributes operate in combination, such shortcomings are particularly problematic.

2.5 Model Class Reliance (MCR)

Feature importance is a central element in explainable machine learning, but many existing approaches provide unstable or model-dependent results. Widely used methods such as permutation importance (Breiman, 2001), LIME (Ribeiro, Singh and Guestrin, 2016), and SHAP (Lundberg and Lee, 2017) each offer insights but share a critical limitation: they describe importance for a single fitted model. In cases where predictors are correlated, as is common in consumer behaviour datasets, different models may allocate importance differently while maintaining similar accuracy (Molnar, 2019; Hooker, Mentch and Zhou, 2021). This creates uncertainty for interpretation. Model Class Reliance (MCR), introduced by Fisher, Rudin and Dominici (2019), was developed to overcome this issue.(Figure 2.2)

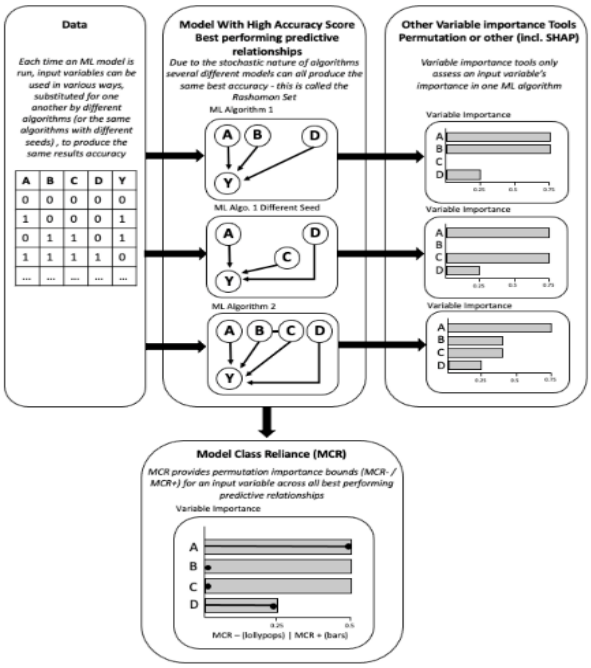


Figure 2.2. Model Class Reliance (MCR) framework with reliance intervals compared to single-model importance measures (NHSX Analytics Unit, 2022).

Rather than focusing on a single model, MCR evaluates the Rashomon set—the group of all models that achieve near-optimal predictive performance. For each feature, MCR calculates a reliance interval. The lower bound (MCR^{-}) reflects the weakest reliance observed when the feature is corrupted across these models, while the upper bound (MCR^{+}) reflects the strongest reliance. This interval-based approach highlights whether a feature is consistently important or only conditionally relevant depending on model specification.

The practical value of MCR lies in its ability to separate indispensable features from those that are substitutable. For instance, in an airline review dataset, service quality measures such as comfort, crew behaviour, and reliability may overlap. MCR shows whether all well-performing models depend on a specific attribute (narrow interval) or whether its role is interchangeable with other predictors (wide interval). This distinction is useful for brand perception studies, where management decisions require stable and defensible insights.

Between the two bounds, MCR^{+} is often considered a more conservative and reliable indicator in domains with correlated variables, since MCR^{-} can underestimate reliance by shifting importance to substitutes (Molnar, 2022). For this reason, the present study adopts MCR^{+} as the basis for further analyses. It ensures that subsequent steps, such as drift detection and interaction analysis, are anchored in features that show robust explanatory significance even in the presence of redundancy.

To date, applications of MCR remain limited. While the method is recognised within the interpretable machine learning community, it has rarely been applied to consumer behaviour or brand perception research (Smith, Mansilla and Goulding, 2020). Furthermore, it has not previously been used as a foundation for sequential analyses such as concept drift or feature interaction. This study therefore extends both methodological and substantive knowledge by positioning MCR as the cornerstone of an integrated explanatory framework.

2.6 Concept Drift and Temporal Dynamics

Predictive models are often built on the assumption that the relationship between features and outcomes remains stable over time. In practice, consumer behaviour and brand perception are dynamic and subject to contextual shifts. In machine learning, this instability is captured under the concept of drift. Drift can occur in two main forms: covariate drift, where the distribution of input variables changes, and concept drift, where the mapping between inputs and outcomes itself shifts (Widmer and Kubat, 1996; Žliobaitė, 2017).

The issue of drift has been studied extensively in domains such as fraud detection, financial forecasting, and recommender systems, where rapid adaptation is essential (Gama et al., 2014; Webb et al., 2016). In these contexts, models trained on earlier data often lose validity when applied to later periods, highlighting the need for monitoring and adaptation. Aviation provides a parallel example: before the COVID-19 pandemic, attributes such as comfort and entertainment were major drivers of satisfaction. During the crisis, however, hygiene, refund policies, and communication transparency became far more important (Suau-Sanchez et al., 2020).

Methodologically, drift can be detected through statistical and model-based approaches. Statistical techniques examine whether feature or residual distributions change across time windows, while model-based methods test whether earlier models retain predictive accuracy (Figure 2.3) when applied to newer data (Gama et al., 2014). Both approaches emphasise that consumer evaluation should not be treated as static but as a dynamic system influenced by both sudden shocks and gradual changes.

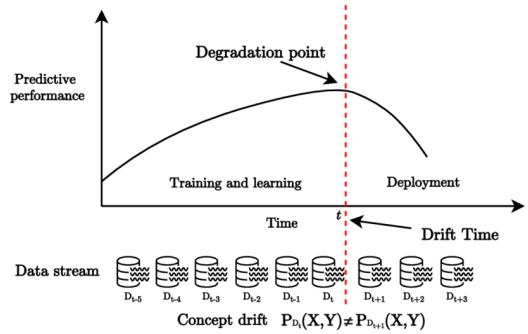


Figure 2.3 Performance-based approach to concept drift (Bayram, Ahmed & Kassler, 2022).

Despite its relevance, drift analysis has rarely been applied to consumer behaviour or brand perception research, particularly within aviation. Most existing studies continue to use static surveys or cross-sectional designs, which fail to capture temporal re-weighting of service attributes. The present study addresses this gap by explicitly applying concept drift analysis to airline review data. By examining how reliance on features changes across pre-crisis, crisis, and recovery phases, the study demonstrates how brand perception evolves dynamically in uncertain contexts. This reframes brand perception from a static construct into a temporal process, where explanatory insights require attention to shifting consumer priorities and adaptive expectations.

2.7 Feature Interactions in Customer Perception

Consumer evaluations rarely emerge from individual service attributes considered in isolation. Instead, perceptions are shaped by combinations of features across

multiple touchpoints. Marketing research has long acknowledged this multidimensional nature of service experiences. Parasuraman, Zeithaml and Berry (1988) showed that quality dimensions can interact to amplify or mitigate overall satisfaction. For example, poor in-flight comfort may be tolerated if staff service is strong, while service recovery following a disruption can restore trust that might otherwise be lost.

In machine learning, explicit methods have been developed to capture such interaction effects. Traditional global feature importance measures, such as permutation importance or SHAP, largely focus on marginal contributions and can obscure joint influences. To address this limitation, the H-statistic (Friedman and Popescu, 2008) and SHAP interaction values (Aas, Jullum and Løland, 2021) have been introduced to identify cases where two variables have a combined effect greater or smaller than the sum of their individual contributions.

In the context of aviation, the importance of interactions is particularly evident. Passengers typically evaluate their experiences holistically, with impressions shaped by the interplay between comfort, safety, staff behaviour, and flexibility. For instance, generous refund policies during a crisis may only strengthen brand trust when coupled with effective communication, while high entertainment quality may improve perceived value primarily in the presence of attentive cabin crew. Ignoring these combinations risks oversimplifying the mechanisms that underpin satisfaction and loyalty.

Despite their relevance, systematic applications of interaction analysis remain scarce in consumer behavior research using online customer reviews. Most existing studies still rely on additive models, which provide partial but incomplete explanations of brand perception. The present study addresses this limitation by incorporating interaction analysis into its framework. Following the identification of robust features through MCR^+, interaction detection methods are applied to examine how attributes combine across pre-crisis, crisis, and recovery phases. This step moves beyond one-dimensional explanations and produces a behaviorally realistic account of how perceptions evolve through interdependent service factors.

2.8 Synthesis and Research Gap

The reviewed literature highlights important insights but also reveals significant limitations in the study of aviation brand perception, online customer reviews, and explainable machine learning.

From a marketing perspective, established frameworks such as customer-based brand equity (Aaker, 1991; Keller, 1993) and service quality (Parasuraman, Zeithaml

and Berry, 1988) provide robust foundations. However, empirical studies in aviation largely rely on static survey methods or cross-sectional data, which capture general sentiment but overlook the temporal re-weighting of service attributes that occurs during crises or competitive shifts.

Online customer reviews offer a step forward by providing unsolicited, large-scale, and temporally granular data. Yet most applications in aviation have remained descriptive, using sentiment analysis or topic modelling, or predictive, using machine learning optimised for accuracy (Archak, Ghose and Ipeirotis, 2011; Zhang, Zhao and Xu, 2016). These approaches demonstrate that OCR contains meaningful signals but fail to provide stable explanations of why perceptions evolve. Explainability tools such as SHAP, LIME, and permutation importance add some interpretive value but remain tied to single fitted models, leaving them unstable under correlation and unable to capture temporal or interactive effects (Molnar, 2019; Hooker, Mentch and Zhou, 2021).

This review identifies three main gaps. First, robust and model-agnostic measures of explanatory importance are absent in OCR-based brand perception studies. MCR$^+$ directly addresses this by bounding reliance across the Rashomon set, providing a stable foundation for interpretation. Second, the temporal dynamics of brand perception are largely unexplored. Concept drift analysis offers the tools to capture how reliance on attributes evolves across pre-crisis, crisis, and recovery phases, reframing brand perception as dynamic rather than static. Third, the interactional nature of consumer evaluations has been neglected. Brand perception rarely emerges additively; it is shaped by interdependencies between service attributes. Systematic interaction analysis is required to uncover these mechanisms.

The present study responds to these gaps by integrating MCR$^+$, concept drift analysis, and interaction analysis into a sequential explanatory framework. MCR$^+$ establishes which features are consistently important across correlated models, drift analysis traces how their influence changes over time, and interaction analysis reveals how attributes combine to shape evaluations. This sequence ensures both methodological rigour and behavioural realism.

By adopting this framework, the study makes dual contributions. Methodologically, it demonstrates how MCR can serve not only as a measure of importance but as the anchor for temporal and interactional analyses, extending its application beyond technical discussions in interpretable machine learning. Substantively, it advances brand perception research in aviation by providing a transparent, data-driven account of how consumer priorities shift under crisis conditions and how interdependent service features jointly structure evaluations. In doing so, it bridges a key gap between predictive accuracy and explanatory insight, offering contributions with both theoretical and managerial value.

3. Methodology

This chapter outlines the methodology adopted to investigate the determinants of airline brand perception (Figure 3.1). It describes the dataset and preparation steps, the feature engineering process, the predictive modelling framework, and the explanatory analyses of model class reliance, feature drift, and feature interactions. The aim is to provide a rigorous and transparent account of how the research objectives were operationalized into data-driven analysis.

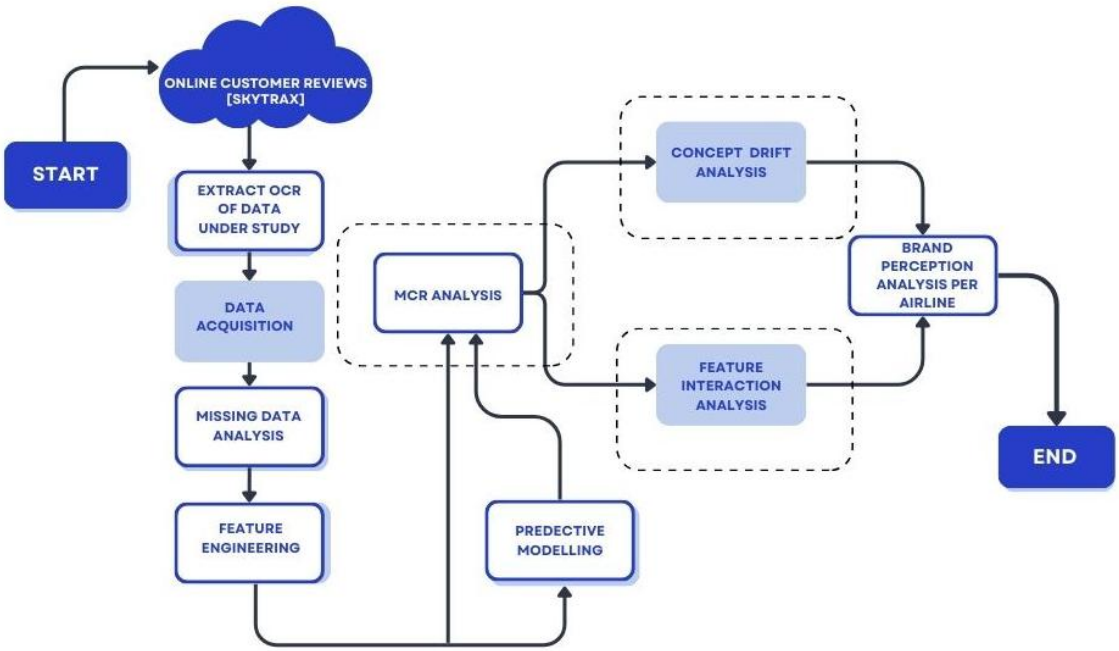


Figure 3.1 Research methodology flowchart

3.1 Data Acquisition

This study uses **Skytrax airline** review data obtained via Kaggle for British Airways (BA) and Emirates Airlines. The BA file contains **3,692 reviews and 18 columns**; the Emirates file contains **1,540 reviews** and **19 column**s. Both datasets include an **overall rating** and rich, **aligned sub-ratings** (seat comfort, staff service, food quality, entertainment, Wi-Fi, ground service, value for money), **free-text review**, and **contextual metadata** (traveller type, travel class, route, aircraft, verified flag, and review dates). Reviews were filtered to a common window, yielding BA coverage **2015 to 2023**. Columns were harmonised to a unified schema, extraneous fields were dropped, and airline labels were added for downstream comparison.

To enrich the route information, each record was parsed into origin, stopover (where applicable), and destination. These locations were geocoded into latitude–longitude coordinates using the **Google Maps API**, and great-circle distances were computed via the **Haversine formula.**The Haversine formula is a mathematical equation used to calculate the great-circle distance between two points on the surface of a sphere (like Earth), given their latitude and longitude. Flights were then categorised as **short-haul** (<1,500 km), **medium-haul** (1,500–3,500 km), or **long-haul** (>3,500 km), with unresolved cases labelled as unknown.

3.2 Missing Data Analysis

Incomplete data represent a pervasive challenge in user-generated review datasets. If not addressed, missingness can diminish statistical power and introduce bias into model estimates. To mitigate these risks, the analysis employed a structured approach integrating descriptive exploration, visualization, statistical diagnostics, and tailored strategy design as illustrated in figure 3.2, in line with established methodological guidance (van Buuren, 2018).

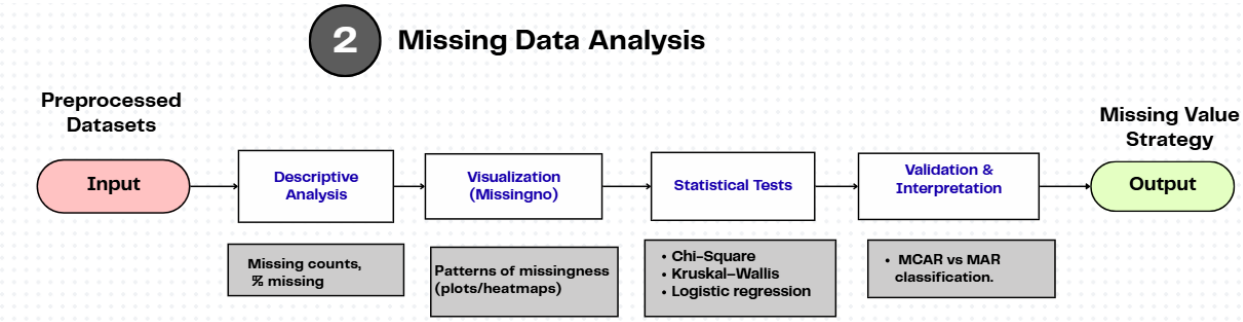


Figure 3.2 Missing Data Analysis flowchart

Descriptive and Visual Exploration. Missingness was first quantified through null counts and percentages, followed by matrix plots, bar charts, and heatmap correlations generated with ***missingno***.This triangulation helped detect whether gaps appeared randomly or followed structural patterns (e.g., clustering within specific travel classes).[Appendix B]

Statistical Diagnostics. Formal tests were then used to classify the missingness mechanism. **Chi-square tests** assessed dependencies with categorical features such as travel class; **Kruskal–Wallis test**s compared distributions of continuous predictors (e.g., distance) across missing vs. observed groups; and **Logistic regressions** modelled missingness directly as a binary outcome. Convergence across these diagnostics provided robust evidence for whether variables followed MCAR (completely at random) or MAR (systematically related to context).

Validation and Interpretation. Results indicated heterogeneity across features. *Travel_class* and *overall_rating* followed **MCAR** patterns, permitting straightforward imputation, whereas sub-ratings (*seat_comfort, food_quality, staff_service*) exhibited **MAR** mechanisms linked to travel context, warranting conditional treatment. By contrast, systematic absences in *wifi* and *entertainment* reflected product availability (e.g., short-haul aircraft), for which missingness indicators were introduced in lieu of imputation.

This distinction is critical for brand perception, as evaluations of Emirates and BA depend on both the quality and availability of services. The adopted strategy thus functioned dually as a safeguard against biased modelling and as a design choice preserving brand-relevant signals.

3.3 Feature Engineering

The feature engineering stage transformed raw review data into structured variables serving as explanatory inputs for predictive and explanatory modelling. Given the mixed nature of online reviews—structured sub-ratings and free-text narratives—this step was critical to align the dataset with the study's objective of uncovering drivers of airline brand perception. The engineered features fall into two groups: (i) text-derived variables capturing sentiment, emotion, and themes, and (ii) structured encodings based on traveller characteristics and missingness patterns. Together, these features provide a multidimensional representation of passenger experiences that goes beyond rating prediction toward insights into how perceptions of brand value form.(Figure 3.3)

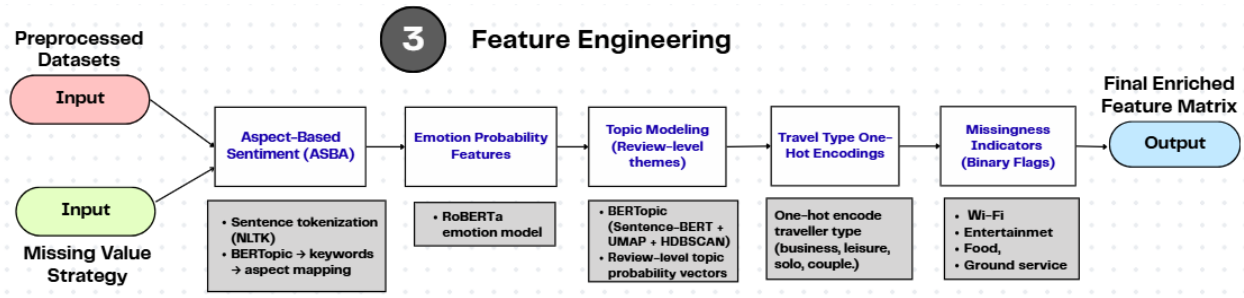


Figure 3.3 Feature Engineering flowchart

3.3.1 Text-Derived Features

Aspect-Based Sentiment Features. Aspect-based sentiment analysis (ABSA) was employed to capture the multidimensional nature of brand perception. Reviews were segmented into sentences, each mapped to core service dimensions—staff, seat comfort, food, and entertainment—and scored using a transformer-based RoBERTa sentiment model

(Barbieri et al., 2020). Aspect-level polarity scores (−1 to +1) were then aggregated per review. This approach isolates how evaluations of specific service elements contribute to overall brand perception, avoiding the oversimplification inherent in single global sentiment measures.

Emotion Features.
Brand perceptions are shaped by both rational evaluations and affective tone. To capture this, the study applied CardiffNLP's RoBERTa-based emotion classifier (Barbieri et al., 2020), trained on social media text annotated for affect. The model generated probability scores for **joy, anger, optimism, and sadness** (Mohammad et al., 2018), enabling differentiation between superficially similar reviews (e.g., relief vs. excitement) and revealing how emotional framing amplifies or moderates service impacts on brand image.

Topic Modelling

Predefined service dimensions capture only part of the customer experience; emergent themes such as refunds, delays, or lounge facilities also play critical roles in shaping brand trust and loyalty. To detect these, **BERTopic** (Grootendorst, 2022) was used to cluster reviews into approximately twenty coherent themes using sentence embeddings (Reimers & Gurevych, 2019), dimensionality reduction (UMAP), and HDBSCAN clustering. Each review was represented by probabilities across these topics, enabling multi-theme attribution. These latent themes capture *how narratives beyond standard ratings contribute to brand perception*, for example, whether operational disruptions or premium services dominate discourse.

3.3.2 Structured Encodings

Passenger background shapes brand evaluations, a business traveller in business class perceives service quality differently from a leisure passenger in economy. To capture this heterogeneity, variables such as ***traveller_type*** were one-hot encoded, enabling models to detect group-specific perceptions. In parallel, missingness in sub-ratings (e.g., Wi-Fi, entertainment) often reflected structural service gaps rather than random omission. **Binary indicators** were therefore introduced to distinguish "not provided" from "poorly rated," ensuring models did not misinterpret absences and preserving the informational link between service provision and brand perception.

3.3.Final Feature Set

Category	Feature(s)	Description
Aspect-Based Sentiment	`food_sentiment`, `seat_sentiment`, `staff_sentiment`, `baggage_sentiment`, `refunds_sentiment`, `lounge_sentiment`, `entertainment_sentiment`, `cleanliness_sentiment`, `airport_service_sentiment`	Sentiment polarity scores (−1 to +1) for specific service aspects.(float64)
Emotions	`anger`, `joy`, `optimism`, `sadness`	Emotion probabilities from RoBERTa model, capturing affective tone of reviews.(float64)
Topics (BERTopic)	`topic_entertainment`, `topic_staff_customer_seating`, `topic_refunds_lounge`, `topic_business_experience`, `topic_meals_food`, `topic_baggage`, `topic_qantas_related`, `topic_drinks_water`	Probability that review content belongs to latent topics (multi-theme distribution).(float64)
Traveller Type (One-Hot Encodings)	`type_Business`, `type_Couple`, `type_Family`, `type_Solo`, `type_Unknown`	Binary indicators for passenger group/travel type. (Binary)
Missingness Flags	`travel_type_missing`, `food_quality_missing`, `wifi_missing`, `entertainment_missing`, `ground_service_missing`	Indicators marking systematically missing ratings, preserving information on service unavailability.(Binary)

Table 3.1 Final Engineered Features

This hybrid feature set combines structured sub-ratings with unstructured textual cues and contextual encodings. It captures not only what passengers rated, but also how they felt, what else they mentioned, and which services were unavailable. This multidimensional representation ensures subsequent modelling can disentangle utilitarian and experiential drivers of airline brand perception across customer groups and temporal contexts.

3.4 Baseline Random Forest Prediction Model

The baseline stage of analysis established a predictive benchmark for how observable service features and textual signals shape overall passenger ratings, used here as a proxy for airline brand perception. Random Forest regressors were selected as the methodological backbone, serving both as (i) a benchmark for prediction accuracy and (ii) a consistent foundation for the subsequent Model Class Reliance (MCR) analysis of feature robustness.

3.4.1 Resources and Data Curation

Baseline models were trained on enriched passenger review datasets from British Airways and Emirates, spanning **2015–2023**. To ensure validity, the feature space was curated to remove leakage and redundancy, so predictors reflected genuine customer evaluations rather than artefacts(Appendix B). Data were partitioned into three temporal windows—**pre-, during-,** and **post-COVID**—to test whether drivers of brand perception remained stable or shifted across crisis and recovery. The design assumed some service attributes remain central, while others vary with situational pressures such as disruption or uncertainty.

3.4.2 Modelling Framework

Random Forests are ensemble learners combining bootstrapped decision trees with random feature subsampling, capturing non-linearities and reducing overfitting (Breiman, 2001; Molnar, 2022). They were selected for two reasons:

- **Methodological alignment:** Model Class Reliance is formally defined for Random Forests, with a tractable implementation in the *mcrforest* package (Smith, Mansilla & Goulding, 2020). Alternative learners (e.g., boosting, SVMs) would break comparability.
- **Empirical suitability:** The heterogeneous structure of airline reviews—numeric ratings, categorical encodings, and text-derived features—requires a flexible model capturing complex, non-linear patterns. Random Forests meet this need while offering interpretable outputs via feature reliance and SHAP decompositions.

3.4.3 Training, Evaluation, and Role in Framework

Training used an 80/20 split with hyperparameters tuned via random search cross-validation (full grids in Appendix [B]). Performance was assessed using **R^2** (variance explained) and **RMSE** (average deviation from observed ratings). Strong results across both metrics confirmed that the enriched feature set captured substantive drivers of passenger evaluations.

The tuned Random Forests served not only as predictive benchmarks but also as scaffolding for reliance and drift analyses. Carrying identical models into the MCR framework allowed reliance intervals to be interpreted alongside predictive accuracy, strengthening internal validity and ensuring observed changes reflected genuine shifts in brand perception.

3.5 Model Class Reliance (MCR) Analysis

The objective of this stage was to determine which features are **robustly necessary** for predicting overall passenger ratings, treated here as a proxy for airline brand

perception. Conventional feature importance methods (e.g., permutation importance, SHAP) describe a single fitted model and are therefore sensitive to sampling variation, correlated predictors, and specification choices. Their conclusions may shift substantially if another equally well-performing model is fitted.

Model Class Reliance (MCR) quantifies how much a predictive model relies on a given feature across the Rashomon set of near-optimal models (Fisher, Rudin & Dominici, 2019). For feature j, the reliance interval is defined as:

$$MCR^{-}_{j} = \inf\{\Delta_j(f) : f \in \mathcal{F}\varepsilon\}$$
$$MCR^{+}_{j} = \sup\{\Delta_j(f) : f \in \mathcal{F}\varepsilon\}$$

where $\Delta_j(f)$ is the performance degradation when feature X_j is corrupted, and $\mathcal{F}\varepsilon$ denotes the set of models whose risk lies within an ε-tolerance of the empirical optimum.

- **Narrow, high intervals** indicate features that are indispensable across all near-optimal models.
- **Wide intervals** suggest that reliance is conditional, either because features are substitutable or their influence emerges only through interactions.

This distinction is particularly important in brand perception research, as it enables the identification of attributes that are not only influential in a single fitted model but consistently necessary across multiple valid specifications to explain variations in customer satisfaction.

3.5.1 Data and Model Alignment

Random Forests were chosen as the sole learner family since MCR is implemented for forests via the ***mcrforest*** package (Smith, Mansilla & Goulding, 2020), which extends RandomForestRegressor to compute reliance intervals while retaining predictive validity. Using the same model class for prediction (Section 3.4) and reliance ensured methodological consistency.

The baseline curated feature space (Section 3.4) were reused: reviews segmented into pre-, during-, and post-COVID windows, per brand. Hyperparameters tuned for the baseline forests were retained, so reliance intervals were estimated under the same validated predictive regime.This alignment ensured observed reliance patterns reflected genuine structural shifts in customer evaluation, not data or model inconsistencies.

3.5.2 MCR Procedure and Computation

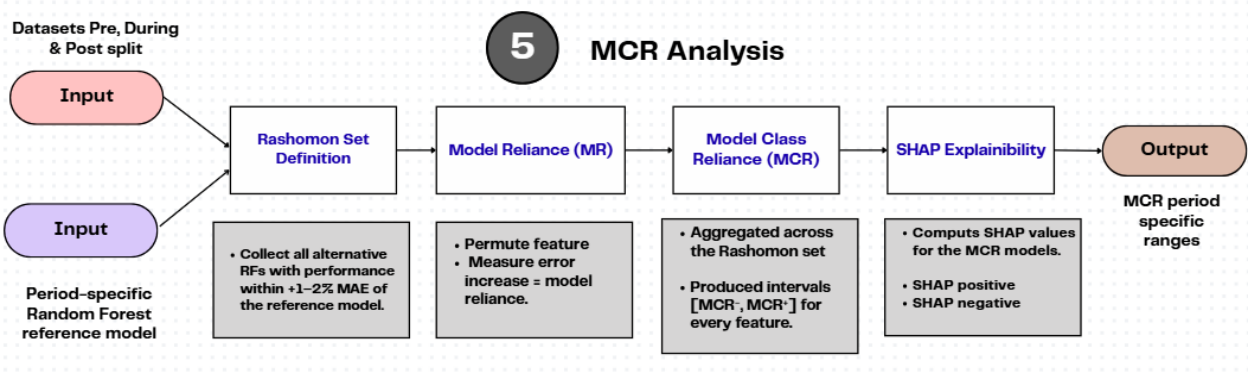


Figure 3.4 Model Reliance flowchart

The MCR procedure consisted of four steps (Figure 3.4):

1. **Baseline fit**: Train the tuned Random Forest on the airline–period subset.
2. **Generate Rashomon set**: Identify alternative forests within an ϵ\epsilonϵ-tolerance of the baseline risk.
3. **Perturbation**: For each feature, apply a corruption operator (permutation conditional on the data distribution) across the Rashomon set.
4. **Aggregation**: Record the degradation in predictive accuracy for each perturbation, and compute the infimum and supremum across forests to obtain MCR^{-}_{j} and MCR^{+}_{j}.

In practice, the **mcrforest** package automates this by repeatedly perturbing features, refitting forests under both "forced-use" and "restricted-use" conditions, and aggregating results into reliance intervals. Full code specifications and diagnostic settings are provided in Appendix B.

3.5.3 Integration with SHAP Analysis

In addition to reliance intervals, SHAP (SHapley Additive exPlanations) was applied to the same airline–period models to decompose feature contributions into negative (SHAP–) and positive ($SHAP^{+}$) effects. While MCR identifies whether a feature is consistently relied upon across the Rashomon set, SHAP clarifies that reliance by quantifying the directional impact of low versus high feature values on predicted ratings (Lundberg & Lee, 2017). Together, MCR and SHAP provide complementary insights: MCR ensures robustness by addressing redundancy and interactions, while SHAP enhances interpretability by distinguishing penalty from reward effects.

3.5.4 Integration with Drift and Interaction Analyses

The reliance intervals formed the foundation for the drift and interaction stages.

Only features with non-negligible reliance were carried forward, ensuring that observed shifts (drift) or synergies/antagonisms (interactions) were grounded in attributes already validated as relevant. This sequencing reduced noise, improved interpretability, and strengthened the validity of conclusions about how airline brand perception evolved across phases.

3.6 Concept Drift Analysis

The aim of this stage was to evaluate whether the reliance of predictive models on service features shifted across time i.e. Pre-COVID, During-COVID, and Post-COVID. In machine learning, *concept drift* refers to changes in the statistical relationship between predictors and the target variable over time (Widmer & Kubat, 1996; Žliobaitė, 2017). In this study, the target is overall star rating, taken as a proxy for airline brand perception. Detecting drift therefore reveals whether the dimensions that passengers use to judge airline quality remain stable or evolve under crisis and recovery conditions. If reliance patterns shift, it suggests that brand perception is context-sensitive rather than fixed.

Concept drift here is operationalised through **Model Class Reliance (MCR)**, which quantifies the degree to which near-optimal models depend on a given feature (Fisher, Rudin & Dominici, 2019). This study emphasised **MCR^+ (upper bound)** because it:

- provides a bounded, interpretable measure for temporal comparison,
- MCR^- can underestimate feature reliance when correlated substitutes exist, because reliance may be shifted to other features. By contrast, MCR^+ reflects the maximum possible reliance of a feature within the Rashomon set, making it less sensitive to this downward bias (Fisher, Rudin & Dominici, 2019).
- directly reflects whether a feature's potential role in shaping predictions has grown or diminished (Molnar, 2022).

MCR^- was computed only as a robustness check. In a brand perception context, this means the analysis captures not just *whether* a feature mattered, but whether its potential influence on customer ratings expanded, contracted, or remained stable across phases.

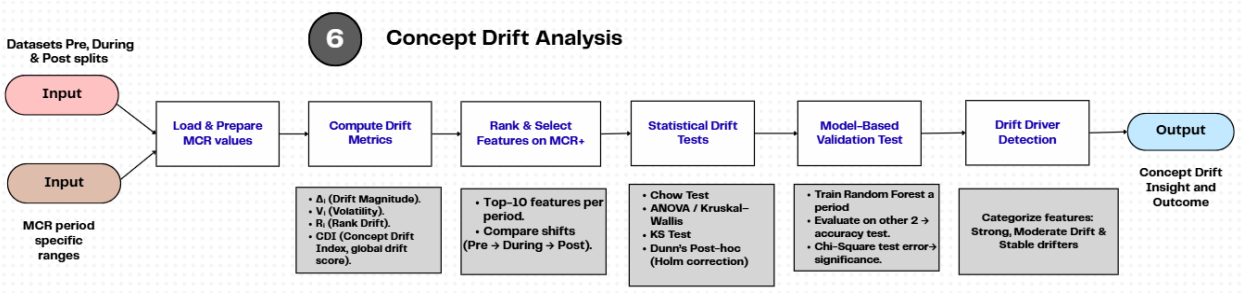


Figure 3.5 Concept Drift Analysis flowchart

3.6.1 Data Preparation and Drift Metrics

Reliance intervals were computed for each airline × period subset using tuned Random Forests (Breiman, 2001). These were structured into a panel dataset and three drift metrics were derived (Gama et al., 2014; Žliobaitė, 2017)(Figure 3.5):

- **Drift Magnitude (Δ_i):** absolute change in MCR^+ of feature *i* across periods.
  $\Delta_i = |MCR^+\_pre(i) - MCR^+\_dur(i)| + |MCR^+\_dur(i) - MCR^+\_post(i)|$

  Interpretation: Measures how much the maximum possible reliance (MCR^+) of feature i changed across all periods.
  - Large Δ_i → feature's importance shifted strongly over time.
  - Small Δ_i → feature's role stayed stable.

- **Rank Drift (R_i):** maximum change in the importance ranking of feature *i* (cf. Nogueira et al., 2017).
  $R_i = \max(|rank\_pre(i) - rank\_dur(i)|, |rank\_dur(i) - rank\_post(i)|)$

  Interpretation: Measures how much the feature's importance rank moved across periods.
  - Large R_i → feature climbed or dropped a lot in relative importance.
  - Small R_i → feature kept roughly the same importance ranking.

- **Concept Drift Index (CDI):** average Δ_i across all features, providing a global measure of system-wide change.

  $CDI = (1/N) * \Sigma\, \Delta_i \quad \text{for } i = 1...N$

  Interpretation: The average drift magnitude across all features (a single system-level number).
  - High CDI → the system overall experienced substantial concept drift.
  - Low CDI → the feature importance structure stayed relatively stable.

- **Volatility Index (V_i)**

$V_i = \max(MCR^+\_pre(i) - MCR^-\_pre(i), MCR^+\_dur(i) - MCR^-\_dur(i), MCR^+\_post(i) - MCR^-\_post(i))$

Interpretation: The maximum width of the importance interval [MCR^-, MCR^+] across time.
- Large V_i → feature's reliance is uncertain/unstable, substitutable with other features.
- Small V_i → feature's reliance is consistent and robust.

Together, these metrics show whether brand perception was consistently anchored in structural service attributes (low drift) or whether passenger priorities shifted toward situational concerns such as refunds, hygiene, or flexibility (high drift).

3.6.2 Statistical Validation

Observed reliance shifts were validated through:

- **Kruskal–Wallis tests** for overall differences across phases.
- **Dunn's post-hoc tests with Holm correction** to identify specific pairwise transitions (Pre → During, During → Post).

This ensured that detected changes represented genuine restructuring of brand perception, not random variation.

3.6.3 Model-Based Validation

As an external check, Random Forest models trained on pre-COVID reviews were evaluated on during- and post-COVID data. Drops in **predictive accuracy** and **F1 scores** indicated that the features driving brand perception before the pandemic no longer explained satisfaction as well in later phases (Gama et al., 2014; Webb et al., 2016). Retraining on Pre + During and testing on Post confirmed that the shifts represented substantive changes in feature–rating relationships rather than stabilisation around a new equilibrium.

3.6.4 Drift Driver Identification

Features were classified into categories using bootstrapped permutation tests (Altmann et al., 2010; Fisher et al., 2019):

Strong Drift Driver: Feature–target relationship changed significantly.
- Significant global test (ANOVA or Kruskal–Wallis, $p < 0.05$)

Moderate Drift / Emerging: Some drift exists, but less robust.

- Not globally significant, but effect-size shift observed.
  Rule of thumb: CI_high_post – CI_low_pre > 0.001

Stable: Feature influence stayed consistent.

- No significant test result.Confidence intervals overlap substantially.

This classification ties drift detection directly to brand perception ie. strong drift drivers highlight emerging priorities shaping airline reputation under crisis (e.g., refund policies during COVID), while stable features signal enduring brand anchors such as staff service.

3.7 Feature Interaction Analysis

This stage extended the analysis beyond individual reliance by examining whether pairs of service attributes exerted **joint effects** on passenger ratings, which serve here as a proxy for airline brand perception. Standard importance measures such as MCR and SHAP provide insight into the marginal contributions of individual predictors, but they cannot identify whether attributes operate **synergistically** (combined influence greater than the sum of individual effects) or **antagonistically** (combined influence weaker than expected). Interactions of this type are central to service evaluation, where passenger impressions often arise from the interplay of dimensions rather than isolated features.

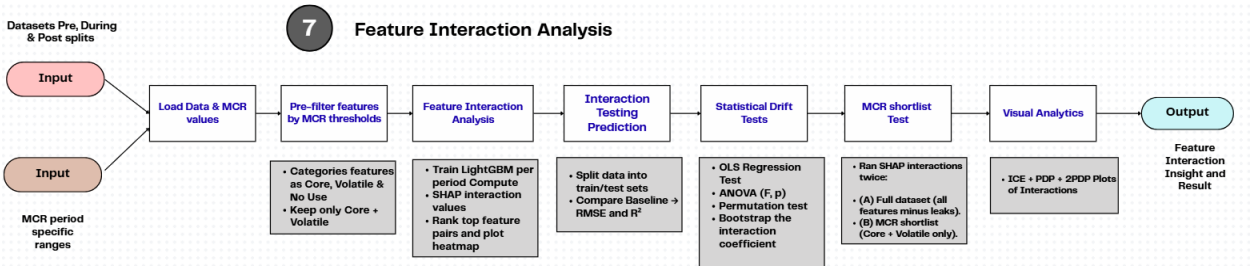


Figure 3.5 Feature Interaction Analysis flowchart

3.7.1 Feature Screening

To ensure tractability and reduce spurious detections, the interaction analysis was restricted to predictors already shown to be substantively relevant in the reliance framework. Features were screened using **Model Class Reliance (MCR) intervals** (Fisher, Rudin & Dominici, 2019), and only those exhibiting consistently high or unstable reliance were retained. This filtering step aligns with recommendations in interpretability research (Molnar, 2022), which caution against searching for higher-order effects in noise variables.(Figure 3.5)

For each feature j at time t, features are classified into Core, Volatile, or No-use according to the following rule:

$$\text{Category}_{j,t} = \begin{cases} \text{Core}, & \text{if } MCR_{j,t}^{+} \geq \text{median}(MCR^{+}) \wedge MCR_{j,t}^{-} \geq P25(MCR^{-}) \\ \text{Volatile}, & \text{if } MCR_{j,t}^{+} \geq \text{median}(MCR^{+}) \wedge MCR_{j,t}^{-} < P25(MCR^{-}) \\ \text{No-use}, & \text{otherwise} \end{cases}$$

Where:

- MCR^{+} = upper reliance bound
- MCR^{-} = lower reliance bound
- $P25(MCR^{-})$ = 25th percentile threshold

3.7.2 Interaction Detection

Within each temporal window (Pre-, During-, Post-COVID), Random Forest models were re-estimated using the shortlisted predictors, and **SHAP interaction values shap.TreeExplainer**(Lundberg & Lee, 2017) were computed. This method extends the SHAP decomposition to attribute variation in predictions not only to main effects but also to pairwise terms, providing a model-consistent measure of interaction strength. Following established practice (Lundberg et al., 2020), mean absolute interaction values were calculated as a summary metric, and the strongest candidates were flagged for formal validation.

3.7.3 Statistical Validation

Interaction candidates were tested using a combination of parametric and resampling-based approaches to ensure robustness:

- **Nested regression tests:** Each interaction was encoded as a multiplicative term (e.g., *seat_comfort* × *staff_sentiment*) and introduced into regression frameworks. **ANOVA F-tests** (Fisher, 1925) were used to compare nested models, assessing whether the interaction significantly improved explanatory power beyond main effects. This test was chosen because it directly evaluates variance explained by added terms in linear settings.

- **Predictive performance tests:** For ensemble models, changes in out-of-sample accuracy were measured using ΔR^2 and $\Delta RMSE$. This approach follows recommendations by Gama et al. (2014) for detecting meaningful concept extensions, ensuring that interactions contribute not only statistically but also practically to predictive validity.

- **Bootstrap confidence intervals:** To account for sampling variability, **bootstrap resampling** (Efron & Tibshirani, 1993) was applied to estimate

distributions of interaction coefficients. This step guards against spurious findings by ensuring that retained interactions remain stable across repeated subsamples, in line with best practices for validating feature effects in nonparametric settings (Altmann et al., 2010).

Only interactions that passed both explanatory and predictive criteria and exhibited stability across bootstrap replicates were retained for interpretation.

3.7.4 Visual Interpretation

Interpretability was enhanced through diagnostic plots that illustrate interaction surfaces. **Partial dependence plots (PDPs)** (Friedman, 2001) provide a global view of the average joint effect of two features, while **individual conditional expectation (ICE) curves** (Goldstein et al., 2015) capture heterogeneity across observations. These tools are widely recommended for diagnosing interactions in applied machine learning (Molnar, 2022), as they clarify whether interactions are consistent across the sample or vary by subgroup.

4. Results and Discussion

4.1 Model Class Reliance Results and Discussion

4.1.1 Model Class Reliance Results

Feature reliance was assessed using MCR⁺ across the Rashomon set, ensuring robustness to redundancy and interactions, while SHAP decomposed contributions into negative (SHAP–) and positive (SHAP+) effects on ratings. Results are reported by period, with British Airways (BA) and Emirates compared side by side

4.1.1.1 Pre-COVID Reliance Dynamics

BRITISH AIRWAYS	EMIRATES AIRLINES

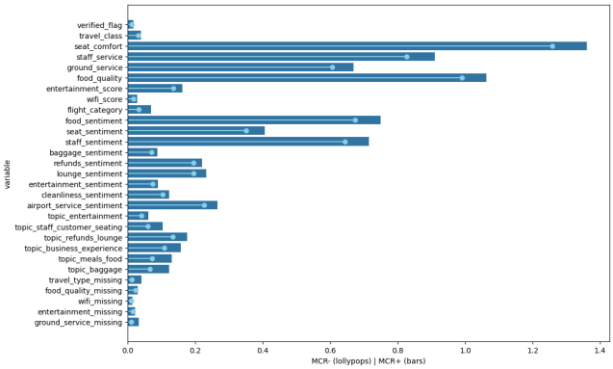
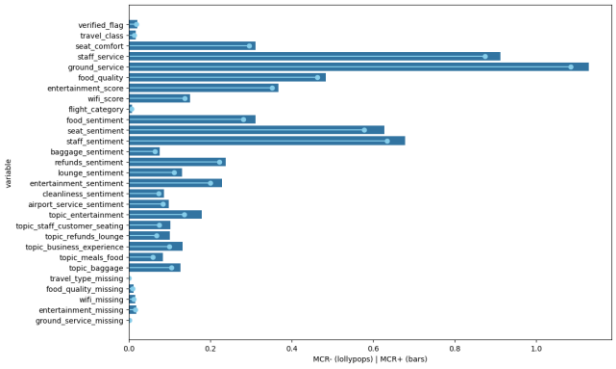
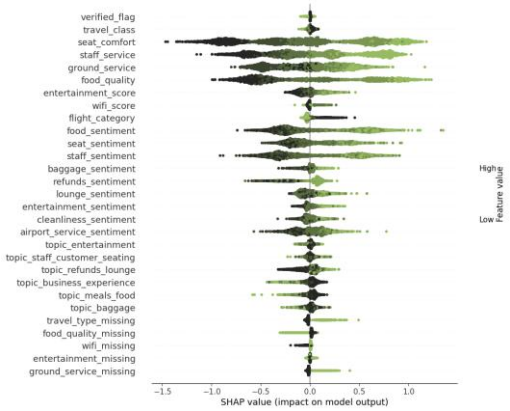
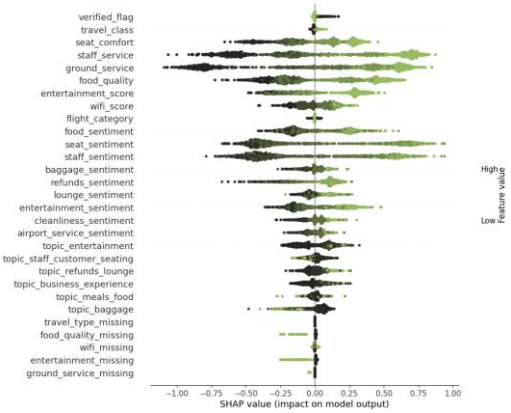

	

Figure 4. MCR Feature Reliance Plots BA Pre	**MCR Feature Reliance Plots Em Pre**
	
SHAP-Based Feature Contributions (Positive) BA Pre	**SHAP-Based Feature Contributions (Positive) Em Pre**

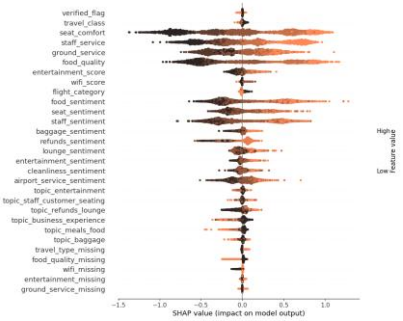
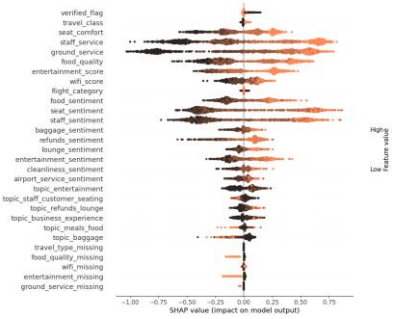
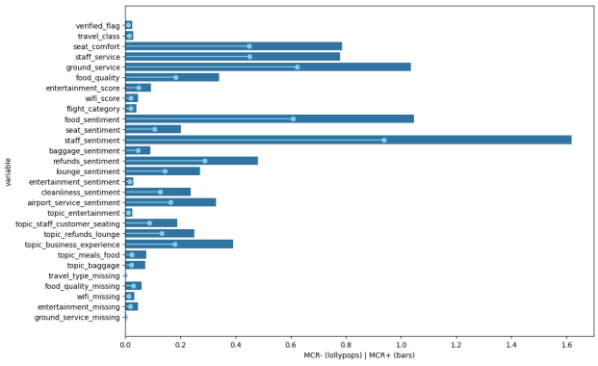
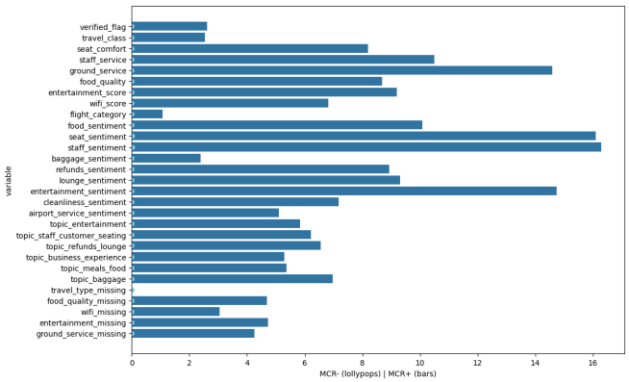
 **SHAP-Based Feature Contributions (Negative) BA Pre**	 **SHAP-Based Feature Contributions (Negative) Em Pre**

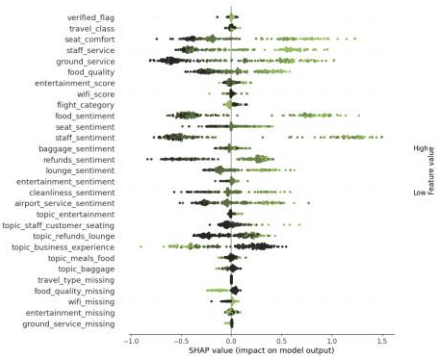
Table 4.1.1 Model Reliance and SHAP plus, minus (Pre-COVID Plots)

British Airways (BA):
MCR⁺ reliance intervals were narrow, identifying **seat_comfort** [0.95–1.35], **food_quality** [0.85–1.05], and **staff_service** [0.70–0.85] as leading features, with **staff_sentiment** [0.55–0.75] moderately influential. SHAP spreads showed **seat_comfor**t had the widest directional range (SHAP–: –1.2, SHAP+: +1.0), while **staff_service** and **food_quality** produced bounded effects (≈ –0.6 to +0.8).

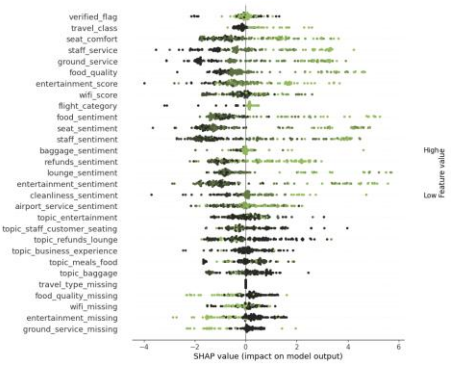
Emirates:
Top reliance fell on **ground_service** [0.95–1.25], **staff_sentiment** [0.65–0.85], and **seat_sentiment** [0.60–0.80], with smaller contributions from **food_quality** and **staff_service**. SHAP confirmed **ground_service** had the broadest bilateral impact (SHAP–: –1.5, SHAP+: +1.0), while sentiment-based features showed narrower ranges (≈ –0.5 to +0.9).

4.1.1.2 During COVID Reliance Dynamics

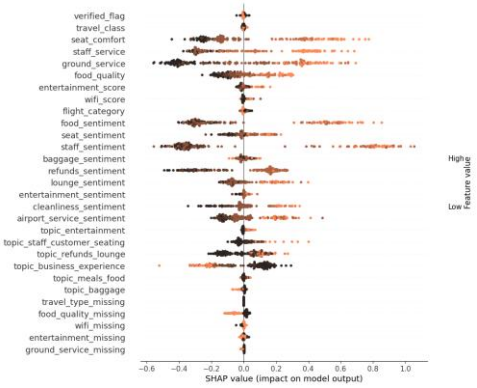
BRITISH AIRWAYS	EMIRATES AIRLINES
 **MCR Feature Reliance Plots BA During**	 **MCR Feature Reliance Plots Em During**



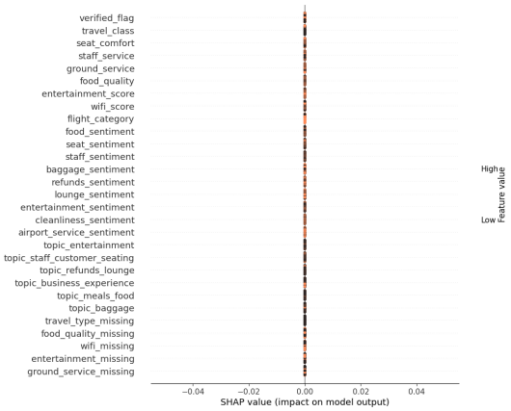
SHAP-Based Feature Contributions (Positive) BA During



SHAP-Based Feature Contributions (Positive) Em During



SHAP-Based Feature Contributions (Negative) BA During



SHAP-Based Feature Contributions (Negative) Em During

Table 4.1.2 Model Reliance and SHAP plus,minus During-COVID Plots

British Airways (BA):

Staff_sentiment [0.94–1.62] and **ground_service** [0.62–1.04] dominated reliance, with **seat_comfor**t [0.45–0.79], **staff_service** [0.45–0.78], and **refunds_sentiment** [0.29–0.48] secondary. SHAP decomposition showed **staff_sentiment** (–1.0 to +1.1) and **ground_service** (–0.9 to +0.8) as the largest contributors, while **seat_comfort** and **refunds_sentimen**t had smaller, more symmetric spreads.

Emirates:

Reliance distributions widened considerably, with **staff_sentiment** [0.00–15.83], **seat_sentiment** [0.00–14.17], and **entertainment_sentiment** [0.00–12.95] displaying high dispersion, alongside ground_service [0.00–11.39] and **staff_service** [0.00–11.35]. SHAP showed bilateral spreads of ≈ –1.0 to +1.3 for **staff and seat sentiment,** and –1.3 to +1.1 fo**r ground_service**, confirming large variability in directional impact.

4.1.1.3 Post-COVID Reliance Dynamics

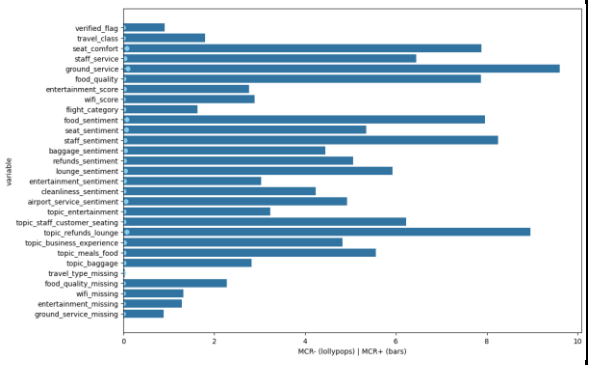
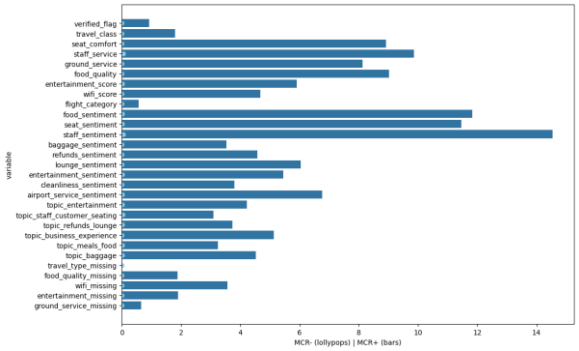
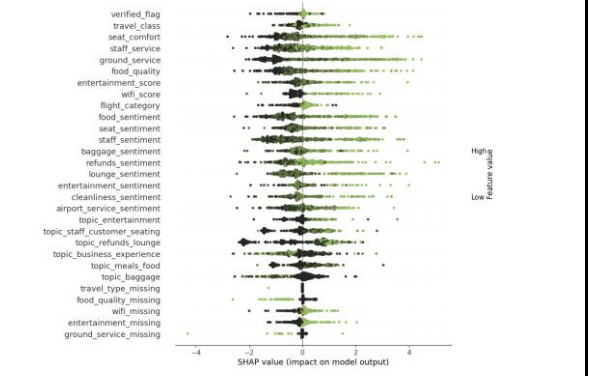
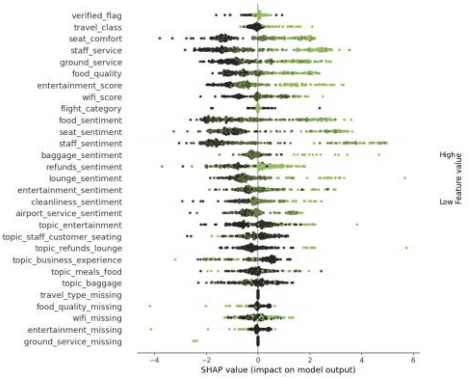
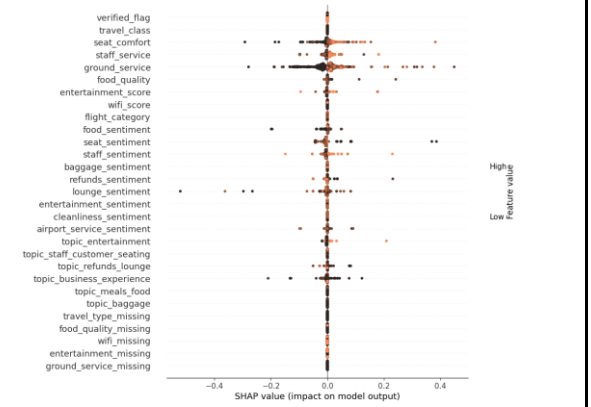
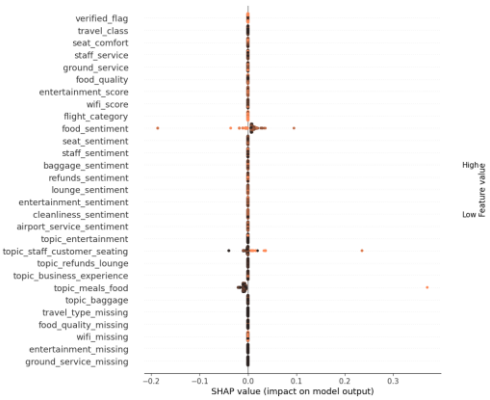
BRITISH AIRWAYS	EMIRATES AIRLINES
 **MCR Feature Reliance Plots BA Prost**	 **MCR Feature Reliance Plots Em Post**
 **SHAP-Based Feature Contributions (Positive)BA Post**	 **SHAP-Based Feature Contributions (Positive) Em Post**
 **SHAP-Based Feature Contributions (Negative) BA Post**	 **SHAP-Based Feature Contributions (Negative) Em Post**

Table 4.1.3 Model Reliance and SHAP plus,minus Post-COVID Plots

British Airways (BA):

Reliance shifted toward **ground_service** [0.44–0.82] and **staff_service** [0.42–0.78], with f**ood_quality** [0.38–0.72], **seat_comfort** [0.35–0.70], and **topic_refunds_lounge** [0.28–0.55] also relevant. SHAP showed moderate

spreads across features: **ground_service** (–0.8 to +0.9), **staff_service** (–0.6 to +0.8), and **food_quality** (–0.5 to +0.7), indicating balanced directional effects.

Emirates:
Reliance intervals narrowed compared to the crisis phase, with **seat_sentiment** [0.65–1.15], **staff_sentiment** [0.60–1.05], and **entertainment_sentiment** [0.55–0.95] leading, followed by **ground_service** [0.50–0.85] and **staff_service** [0.45–0.70]. SHAP spreads remained symmetric: **seat_sentimen**t (–0.9 to +0.9), **staff_sentiment** (–0.8 to +0.9), and **entertainment_sentiment** (–0.7 to +0.8), showing consistent bilateral influence.

4.1.2 Model Class Reliance Discussion

4.1.2.1Explainability of Feature Reliance Over Time

The use of MCR⁺ alongside SHAP decomposition provided a structured account of how feature reliance evolved across the pre-, during-, and post-COVID phases. By estimating reliance over the Rashomon set of equally predictive models, MCR⁺ isolated stable explanatory drivers from those subject to redundancy or model-specific noise (Fisher et al., 2019). In turn, SHAP allowed the directional impact of each feature to be distinguished, separating penalty effects (SHAP–) from reward effects (SHAP+). Together, these methods enhanced the interpretability of overall rating predictions by linking reliance shifts to the context of passenger evaluations over time.

To validate the effectiveness of MCR⁺ as a screening step for Interaction, a comparative interaction analysis was conducted under two setups: (i) the full feature set (excluding target and collinear variables) and (ii) the MCR⁺ shortlist. Results across all periods showed that both approaches identified similar high-impact interactions (e.g., *food_quality* × *staff_sentiment*, *seat_comfort* × *staff_service*), but the MCR⁺ shortlist consistently filtered out weaker or spurious pairs, emphasising brand-relevant dynamics such as *ground_service* × *business_experience* during COVID. This confirms that MCR⁺ provides a robust foundation for interaction analysis by concentrating on stable, behaviourally meaningful features.

For British Airways, explainability was characterised by abrupt shifts in reliance, consistent with crisis-sensitive volatility. In the pre-COVID phase, narrow reliance intervals on seat_comfort and food_quality indicated stable dependence on tangible product attributes. SHAP spreads confirmed asymmetry, with poor comfort exerting disproportionately negative effects compared to the uplift from high comfort. During COVID, reliance migrated towards staff_sentiment and ground_service. SHAP showed stronger SHAP– than SHAP+ contributions,

clarifying that negative staff interactions or poor airport handling sharply reduced predicted ratings, while positive experiences offered only partial recovery. Post-COVID, reliance reweighted again toward ground_service and staff_service. Here SHAP distributions were more balanced, suggesting that BA's recovery period evaluations were explained by a return to symmetric service-based contributions. This sequence demonstrates that BA's explanatory features were not constant but reweighted under external shocks, making overall predictions highly sensitive to context.

In contrast, Emirates displayed a more consistent explanatory structure. Pre-COVID reliance was already anchored in staff_sentiment and seat_sentiment, with SHAP spreads showing balanced bilateral contributions. During COVID, reliance intervals widened, reflecting model uncertainty and multiple explanatory pathways, but SHAP confirmed that staff and seat sentiment, alongside ground_service, consistently explained both positive and negative shifts in ratings. Post-COVID, reliance narrowed again, consolidating around sentiment-oriented features. Unlike BA, Emirates' explainability showed incremental rather than abrupt reweighting, indicating that predictions were consistently structured around service and emotional perception rather than crisis-driven redistributions.

Overall, the joint use of MCR^{+},MCR- and SHAP demonstrates that these models were not black-boxes:reliance intervals distinguished stable features from volatile ones by quantifying maximum and minimum dependence across the Rashomon set, while SHAP clarified directional impacts through penalty (SHAP–) and reward (SHAP+) effects. For BA, reliance shifted abruptly with operational shocks, dominated by negative contributions; for Emirates, patterns remained sentiment-driven and more balanced. Together, these methods explained not just which features mattered, but how reliably and in what direction they shaped ratings over time

The focus is on **MCR^{+}** as the primary indicator of reliance drift and interaction. Following Fisher, Rudin, and Dominici (2019), MCR^{-} values can systematically underestimate a feature's necessity in the presence of correlated or substitutable predictors, since reliance may be displaced to alternative variables within the Rashomon set. By contrast, MCR^{+} captures the **maximum possible reliance** a feature can attain across all near-optimal models, providing a more conservative and interpretable measure for temporal comparison. This upper-bound framing is particularly suited for drift analysis, as it indicates whether the explanatory potential of a feature has expanded, contracted, or remained stable across phases (Molnar, 2022). MCR^{-} was therefore computed only as a robustness check, while reliance shifts are interpreted using MCR^{+} values

4.2 Concept drift Analysis Result & Discussion

This section reports concept drift results, showing how feature reliance shifted across pre-, during-, and post-COVID phases to explain brand perception dynamics.

4.2.1 Concept drift Analysis Result

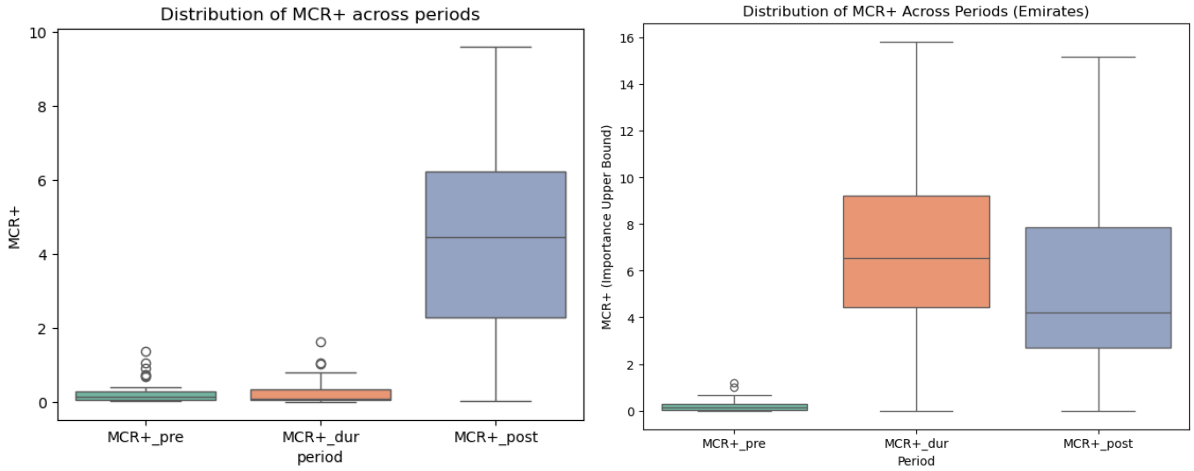


Figure 4.2.1. Distribution of Feature Reliance (MCR) boxplot Across Periods (BA vs Emirates)*

From Figure 4.2.1 British Airways, MCR+ reliance values remained low in Pre- and During-COVID, with a clear upward shift and greater spread Post-COVID.
For Emirates, MCR+ reliance increased sharply During COVID with wide variability, and remained elevated Post-COVID though with slightly reduced spread.

British Airways	Emirates Airlines

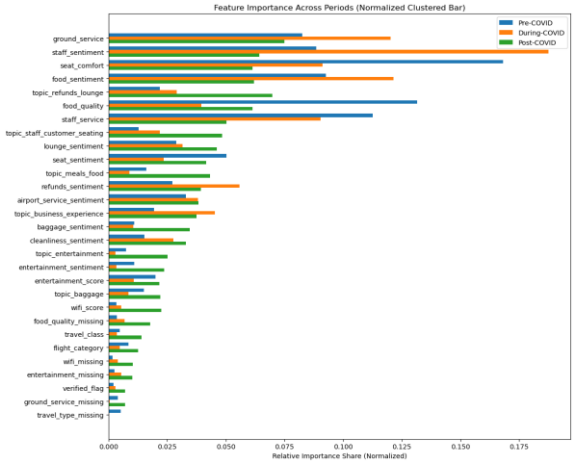
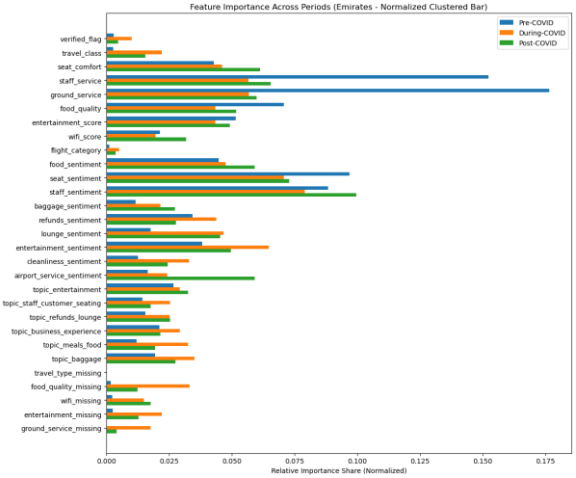

	

Figure 4.2.2 Feature Reliance Across Periods BA	*Figure 4.2.3 Feature Reliance Across Periods Em*

Table 4.2.1 Feature Reliance Across Periods (BA & Em)

For British Airways, reliance was concentrated on ***ground_service, staff_sentiment, seat_comfort,*** **and** ***food_quality*** in all periods. During COVID, ***topic_refunds_lounge*** **and** ***lounge_sentiment*** showed increased importance. In the post-COVID period, ***seat_sentiment*** **rose** in share, while ***travel_class, wifi_missing,*** **and** ***entertainment_missing*** remained consistently low across all periods.

For Emirates, reliance was dominated by ***ground_service*** and ***staff_service*** before COVID. During COVID, ***entertainment_sentiment, refunds_sentiment,*** **and** ***lounge_sentiment*** increased in relative importance. In the post-COVID period, ***staff_sentiment*** and ***seat_sentiment*** became more prominent. Features such as ***travel_class, wifi_missing,*** and ***entertainment_missing*** consistently recorded minimal reliance**.**

4.2.2 Statistical Drift Test

Model Performance Evaluation

British Airways

Period	Accuracy	F1 Score	AUC
Pre	1.000	1.000	1.000
During	0.565	0.502	0.893
Post	0.589	0.550	0.904

Emirates Airways

Period	Accuracy	F1 Score	AUC
Pre	1.000	1.000	1.000
During	0.503	0.506	0.824
Post	0.450	0.393	0.827

Table 4.2.2 Model Performance Across Periods per Brand(Trained on Pre-COVID Only)

Models trained only on Pre-COVID data achieved perfect scores on Pre-COVID sets but showed major declines on During- and Post-COVID periods for both airlines (Table 1). Accuracy and F1 dropped substantially, though AUC remained relatively high (0.82–0.90). (Table 4.2.2)

Chi-Square Test of Error Distributions

Airways	Test	χ^2	p-value	Conclusion
British Airways	Pre vs During vs Post	1170.411	<0.001	Significant difference in error rates
Emirates	Pre vs During vs Post	667.934	0.00000	Significant difference in error rates

Table 4.2.3 – Chi-Square Test of Error Distributions Across Periods per Brand

Chi-square tests confirmed statistically significant differences in error rates across periods for both airlines (BA: $\chi^2=1170.411$, $p<0.001$; Emirates: $\chi^2=667.934$, $p<0.001$) (Table 4.2.3).

Kolmogorov–Smirnov Test for Reliance Drift

Airways	Comparison	KS Statistic	p-value	Interpretation
British Airways	**Pre vs During**	0.138	0.95144	No significant drift
British Airways	**During vs Post**	0.862	0.00000	Significant drift (p<0.05)
British Airways	**Pre vs Post**	0.862	0.00000	Significant drift (p<0.05)
Emirates Airlines	**Pre vs During**	0.931	0.000	Significant drift (p<0.05)
Emirates Airlines	**During vs Post**	0.345	0.06301	No significant drift
Emirates Airlines	**Pre vs Post**	0.862	0.000	Significant drift (p<0.05)

Table 4.2.4 –Kolmogorov–Smirnov (KS) Tests for Pairwise Drift in MCR⁺

Kolmogorov–Smirnov tests showed different drift patterns. For BA, no significant drift occurred between **Pre vs During** (p=0.951), while significant drift appeared in During vs Post and **Pre vs Post** (p<0.001). For Emirates, significant drift occurred between **Pre vs During** and **Pre vs Post** (p<0.001), but not **During vs Post** (p=0.063) (Table 4.2.4)

Kruskal–Wallis Test for Distributional Differences

Airways	Test	Statistic	p-value	Interpretation
British Airways	Kruskal–Wallis	49.325	0.00003	Significant difference in medians across periods
Emirates Airlines	Kruskal–Wallis	50.41	0.00022	Significant difference in medians across periods

Table 4.2.5– Kruskal–Wallis Test on MCR⁺ Across Periods

Kruskal–Wallis tests confirmed overall differences in feature reliance distributions across periods for both airlines (BA: χ^2=49.325, p<0.001; Emirates: χ^2=50.41, p<0.001) (Table 4.2.5).

4.2.3 Concept Drift Analysis:

British Airways

Concept Drift Index (CDI) for BA = 4.286

	variable	MCR+_pre	MCR+_dur	MCR+_post	drift_total	volatility	rank_pre	rank_dur	rank_post	r
4	ground_service	0.668522	1.036486	9.606638	8.938116	9.514989	6.0	3.0	1.0	
20	topic_refunds_lounge	0.175955	0.250125	8.962741	8.786785	8.891221	11.0	11.0	2.0	
5	food_quality	1.063328	0.340700	7.866381	8.248308	7.854818	2.0	8.0	6.0	
2	seat_comfort	1.360692	0.786237	7.883940	7.672159	7.815417	1.0	4.0	5.0	
11	staff_sentiment	0.715276	1.618342	8.245396	7.530120	8.217131	5.0	1.0	3.0	
9	food_sentiment	0.749901	1.047059	7.960385	7.210484	7.890150	4.0	2.0	4.0	
19	topic_staff_customer_seating	0.102668	0.189293	6.220343	6.117675	6.216488	17.0	14.0	8.0	
3	staff_service	0.910043	0.778959	6.443255	5.795379	6.413490	3.0	5.0	7.0	
14	lounge_sentiment	0.232595	0.271813	5.929979	5.697383	5.890364	9.0	10.0	9.0	
22	topic_meals_food	0.130174	0.076363	5.555246	5.532694	5.552891	14.0	17.0	10.0	

Table 4.2.6 Concept Drift Index (CDI) for British Airways.

Emirates Airlines

Concept Drift Index (CDI) for Emirates = 8.675

	variable	MCR+_pre	MCR+_dur	MCR+_post	drift_total	volatility	rank_pre	rank_dur	rank_post
15	entertainment_sentiment	0.256469	12.946876	7.554947	18.082336	12.946842	9.0	3.0	9.0
10	seat_sentiment	0.649740	14.174872	11.098056	16.601948	14.170327	3.0	2.0	2.0
11	staff_sentiment	0.592821	15.829949	15.179238	15.887840	15.829927	4.0	1.0	1.0
13	refunds_sentiment	0.230608	8.783448	4.228309	13.107980	8.783446	10.0	9.0	14.0
4	ground_service	1.184108	11.386309	9.128476	12.460033	11.386263	1.0	4.0	5.0
3	staff_service	1.020604	11.351291	9.995768	11.686209	11.351260	2.0	5.0	3.0
14	lounge_sentiment	0.118121	9.347615	6.930943	11.646166	9.347610	15.0	7.0	11.0
25	food_quality_missing	0.011331	6.623272	1.899690	11.335523	6.623272	26.0	13.0	25.0
22	topic_meals_food	0.080882	6.541702	2.948759	10.053763	6.541696	20.0	15.0	20.0
23	topic_baggage	0.130419	7.041197	4.187728	9.764247	7.041192	14.0	12.0	15.0

Table 4.2.7 Concept Drift Index (CDI) for Emirates Airlines.

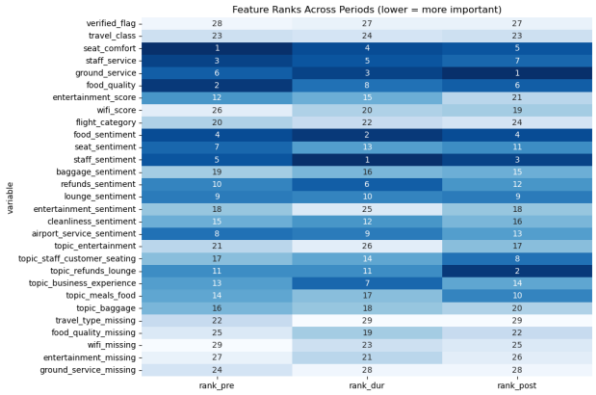


Figure 4.2.4 Feature Rank Dynamics for British Airways Across Periods.

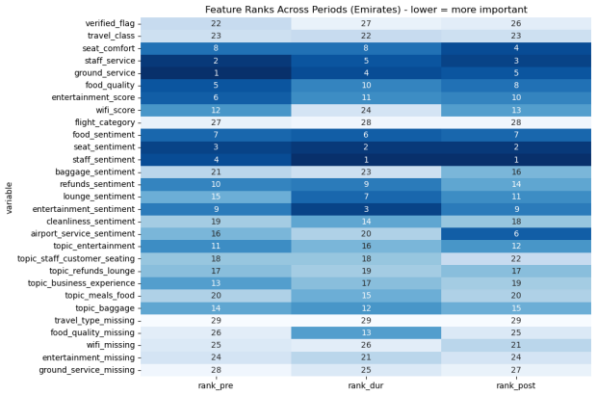


Figure 4.2.5 Feature Rank Dynamics for Emirates Across Periods.

Table 4.2.8– Feature Level Drift and Rank shift across Periods

For British Airways, ***ground_service*** showed the largest drift, increasing from 0.07 (pre) to 8.49 (post), moving from rank 6 to rank 1. ***Topic_refunds_lounge*** and ***seat_comfort*** also exhibited strong shifts, while ***staff_sentiment*** gained importance during COVID before declining. The heatmap confirms these changes, with operational and service-related features climbing to top ranks post-COVID. (Table 4.2.8)

For Emirates, ***entertainment_sentiment*** and ***seat_sentiment*** exhibited the strongest drift, rising from near-zero pre-COVID to top-5 ranks during and post-COVID. ***Staff_sentiment*** and ***refunds_sentiment*** also surged in importance, while ***ground_service*** remained consistently influential but shifted ranks over time. The heatmap highlights this reordering, showing sentiment and service-related features displacing traditional operational variables

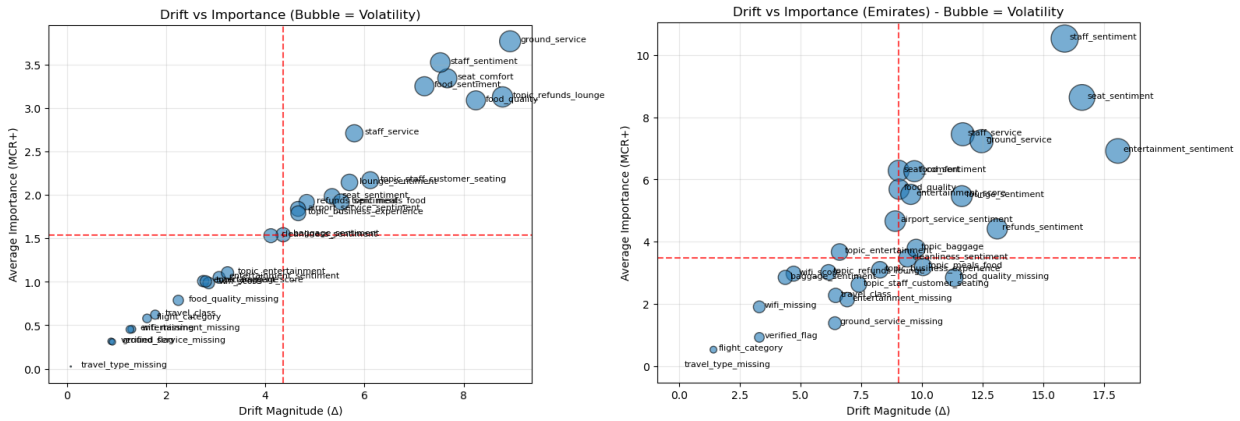


Figure 4.2.6. Drift vs. Reliance per brand (bubble = volatility)

The scatterplots in Figure 4.2.6 compare **drift magnitude (x-axis)** against **average importance (MCR⁺, y-axis)**, with **bubble size indicating volatility**. The red dashed lines mark the median thresholds, dividing features into four quadrants:

- **Top-right:** high drift, high importance.
- **Top-left:** low drift, high importance
- **Bottom-right:** high drift, low importance
- **Bottom-left:** low drift, low importance

For **British Airways**, key drift drivers (top-right) include ***ground_service, staff_sentiment***, ***seat_comfor***t, ***food_sentiment,*** and ***topic_refunds_lounge***, with ***staff_service*** and ***lounge_sentiment*** showing moderate shifts. Low-importance features (*travel class, wifi_missing, entertainment_missing*) remain stable.

For **Emirates**,***staff_sentiment,seat_sentiment,***and ***entertainment_sentiment*** dominate the top-right quadrant, alongside ***ground_service*** and ***staff_service*** with larger volatility. Features like ***baggage_sentiment*** and ***topic_meals_food*** drifted but with limited impact, while ***travel_type_missing, wifi_missing,*** and ***flight_category*** stayed consistently unimportant.

4.2.3 Feature-Level Concept Drift Drivers

British Airways	Emirates Airlines

variable	Pre %	During %	Post %	Δ % (During –Pre)	Δ % (Post– During)	Δ % (Post– Pre)	drift_category
food_sentiment	0.2076	6.7717	0.1893	6.5640	-6.5823	-0.0183	Strong Drift Driver
seat_comfort	5.7171	0.1575	1.2621	-5.5596	1.1047	-4.4549	Strong Drift Driver
refunds_sentiment	0.5026	32.5984	1.4893	32.0959	-31.1091	0.9868	Strong Drift Driver
topic_refunds_lounge	0.0708	0.7874	8.9234	0.7166	8.1360	8.8526	Strong Drift Driver

variable	Pre %	During %	Post %	Δ % (During –Pre)	Δ % (Post– During)	Δ % (Post– Pre)	drift_category
ground_service	10.7192	10.2362	12.7856	-0.4830	2.5493	2.0664	Moderate Drift / Emerging
staff_sentiment	0.1534	2.3622	0.1010	2.2088	-2.2612	-0.0524	Moderate Drift / Emerging
food_quality	1.1231	2.3622	0.2146	1.2391	-2.1476	-0.9086	Moderate Drift / Emerging
topic_business_experience	0.2124	1.2598	2.7767	1.0475	1.5169	2.5644	Moderate Drift / Emerging
seat_sentiment	0.0661	3.1496	0.4796	3.0835	-2.6700	0.4136	Moderate Drift / Emerging

variable	Pre %	During %	Post %	Δ % (During –Pre)	Δ % (Post– During)	Δ % (Post– Pre)	drift_category
ground_service	70.11	0.44	3.78	-69.68	3.34	-66.34	Strong Drift Driver
entertainment_score	1.51	6.98	22.57	5.47	15.59	21.06	Strong Drift Driver
staff_sentiment	3.23	10.12	11.43	6.89	1.30	8.19	Strong Drift Driver
topic_refunds_lounge	0.26	37.78	8.12	37.52	-29.66	7.86	Strong Drift Driver
staff_service	5.44	0.70	18.60	-4.74	17.90	13.16	Strong Drift Driver

variable	Pre %	During %	Post %	Δ % (During –Pre)	Δ % (Post– During)	Δ % (Post– Pre)	drift_category
airport_service_sentiment	0.241497	1.657941	0.377715	1.416443	-1.280226	0.136218	Moderate Drift / Emerging
type_Solo	0.000000	1.396161	0.000000	1.396161	-1.396161	0.000000	Moderate Drift / Emerging
type_Business	1.227611	0.959860	0.000000	-0.267751	-0.959860	-1.227611	Moderate Drift / Emerging
topic_drinks_water	0.167706	1.047120	0.000000	0.879414	-1.047120	-0.167706	Moderate Drift / Emerging
topic_business_experience	1.482525	0.610820	0.472144	-0.871705	-0.138677	-1.010381	Moderate Drift / Emerging

Table 4.2.9– Feature level Concept Drift Drivers

These features are classified as **Strong Drift Drivers** due to the scale and volatility of their reliance shifts.

British Airways:

- **food_sentiment** rose sharply during COVID (6.77%) but dropped back down post-COVID (0.19%), creating sharp swings (Δ% +6.56 → −6.58).
- **seat_comfort** declined from a pre-COVID peak (5.72%) to minimal influence during (0.16%), then partially rebounded post-COVID (1.26%), resulting in a net decline.
- **refunds_sentiment** surged during COVID (32.60%), before falling back to 1.49% post-COVID.
- **topic_refunds_lounge** increased gradually, with the most pronounced growth post-COVID (8.92%).

Emirates Airlines:

- **ground_service** remained consistently relevant(≈10–13%) but gained slightly more influence post-COVID.
- **staff_sentiment** and **food_quality** both spiked during COVID (≈2.36%) but declined again post-COVID.
- **topic_business_experience** increased steadily across all periods (0.21% → 2.78%).
- **seat_sentiment** grew during COVID (3.15%) but declined post-COVID (0.48%).

- **staff_service** was consistently relevant pre- and during-COVID (≈3–4%) but diminished sharply post-COVID (0.18%).

These are classified as **Moderate Drift / Emerging Drivers**, reflecting smaller or transitional shifts in reliance compared to the strong drift features.

British Airways:

- **ground_service** accounted for 70.11% pre-COVID, collapsed to 0.44% during COVID, and partially rebounded to 3.78% post-COVID. This represents the most dramatic overall decline (Δ% −69.68 → +3.34).
- **entertainment_score** rose steadily from 1.51% (pre) to 6.98% (during) and then jumped to 22.57% post-COVID, reflecting a substantial sustained increase.
- **staff_sentiment** grew from 3.23% (pre) to 10.12% (during), followed by a smaller increase to 11.43% post-COVID, yielding a net gain of +8.19%.

Emirates Airways

- **airport_service_sentiment** rose from 0.24% (pre) to 1.66% (during), before falling back to 0.38% post-COVID, leaving only a slight net change.
- **type_Solo** appeared only during COVID (1.40%) and disappeared post-COVID, marking it as an emerging but transient driver.
- **type_Business** decreased consistently from 1.23% (pre) to 0.96% (during) and dropped to 0% post-COVID.

These are **moderate drift drivers**, where shifts exist but are smaller in scale or less sustained than the strong drift group.

4.2.2 Concept drift Discussion

Drift was operationalised through **MCR⁺ reliance distributions** across pre-, during-, and post-COVID periods. Unlike conventional feature importance metrics, MCR⁺ evaluates reliance by resampling across the Rashomon set of equally well-performing models, thereby accounting for feature redundancy and interaction effects (Fisher et al., 2019). As a result, shifts in MCR⁺ values reflect genuine changes in the strength of feature–outcome dependence rather than artefacts of multicollinearity or single-model instability. This framing allows reliance drift to be interpreted in the context of **concept drift**, defined as systematic change in the conditional distribution P(y|x) over time (Widmer & Kubat, 1996; Gama et al., 2014), rather than random model noise.

4.2.2.1 Statistical Evidence and Feature-Level Drift

The observed shifts in reliance were subjected to rigorous statistical validation to ensure they represented genuine **concept drift** rather than stochastic variation.

- **Chi-square tests** confirmed significant differences in error distributions across periods for both airlines (BA: $\chi^2=1170.41$, $p<0.001$; Emirates: $\chi^2=667.93$, $p<0.001$), consistent with methodology for detecting categorical drift in predictive models (Lu et al., 2019).

- **Kolmogorov–Smirnov (KS) tests** identified the *temporal localisation* of drift: for BA, drift emerged primarily in the **post-COVID period** (Pre vs During: ns; During vs Post: $p<0.001$), whereas for Emirates, drift was already significant **during COVID** (Pre vs During: $p<0.001$) but stabilised in the recovery phase.

- **Kruskal–Wallis tests** demonstrated global divergence in reliance medians across all three periods (**BA:** $\chi^2=49.325$, $p<0.001$; **Emirates**: $\chi^2=50.41$, $p<0.001$), aligning with Gama et al. (2014) that drift can manifest as distributional reweighting rather than shifts in mean prediction alone.

Together, these tests reinforce that MCR^+ reliance trajectories are statistically robust indicators of concept drift.These results substantiate that reliance drift was **systematic and structural**, in line with Widmer and Kubat's (1996) foundational definition of concept drift as a change in the conditional distribution $P(y|x)$ over time.

Feature-Level Dynamics of Drift

Analysis at the feature level revealed **contrasting trajectories** between the two carriers.

British Airways (BA): Drift patterns followed a volatile cycle. Features such as *refunds_sentiment* spiked during COVID (32.6%) but collapsed post-COVID (1.49%), reflecting crisis-specific salience that quickly dissipated once operational conditions changed. *Ground_service* displayed the reverse: it collapsed during COVID (0.44%) under conditions of disruption but partially recovered post-COVID (3.78%). *Seat_comfort* and *food_sentiment* also oscillated, suggesting unstable reliance across phases. This volatility exemplifies what Žliobaitė (2016) terms *abrupt drift*, where sharp shifts in importance are tied to external shocks.

Emirates: In contrast, feature drift followed a more **gradual reordering**. *Entertainment_sentiment* and *seat_sentiment* rose from near-zero pre-COVID to consistently high reliance post-COVID. *Staff_sentiment* also gained prominence and sustained its role longer. Unlike BA, Emirates' reliance shifts reflected what Gama et

al. (2014) classify as *incremental drift*, with consumer priorities steadily reweighted rather than radically disrupted.

The **volatility plots** corroborate these patterns. For BA, volatility clustered in high-drift/high-reliance features (e.g., ground_service, refunds_sentiment, seat_comfort), indicating concentrated instability in core brand drivers. For Emirates, volatility was more evenly spread across sentiment-related features, suggesting a distributed and smoother adjustment process. This aligns with evidence from marketing literature that **service perception** evolves more incrementally compared to operational crises, which trigger sharper shifts (Nguyen et al., 2020; Rust & Huang, 2014).

4.2.2.2 Observed Feature-Level Drift Drivers

The detection of drift drivers highlighted that changes in passenger evaluations were not uniform across features but concentrated on a subset of attributes that shifted markedly across the three periods. This aligns with the literature on concept drift (Gama et al., 2014), where reliance distributions often reweight disproportionately toward certain variables during shocks or structural breaks.

British Airways

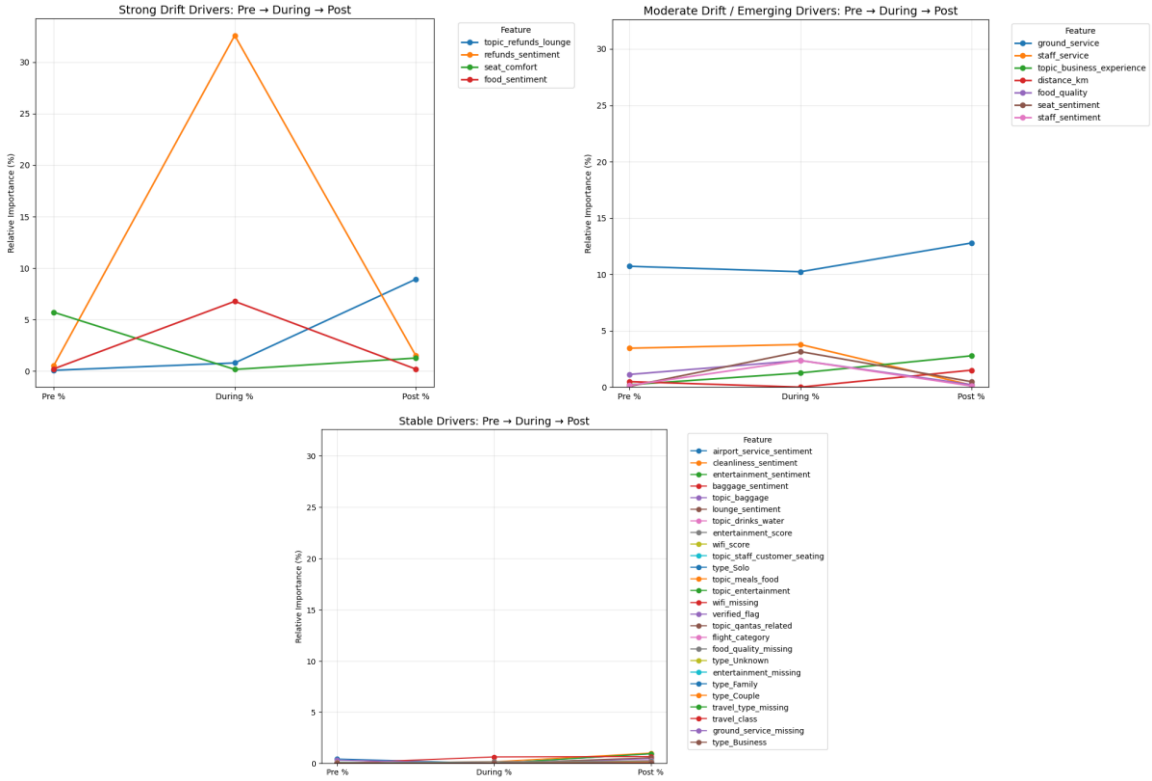


Figure 4.2.7– Feature Drift Drivers Across Phases (Pre–During–Post COVID) British Airways

For British Airways from Figure 4.2.7, the most pronounced drift was observed in **ground_service**, which increased from a negligible reliance pre-COVID (0.07) to the leading driver post-COVID (8.49), moving from rank 6 to rank 1. This structural reordering is consistent with the sharp operational disruptions recorded in the airline industry during the pandemic.

Other strong drift drivers included **seat_comfort** and **refunds_sentiment**. Seat comfort declined from a pre-COVID peak (5.72%) to minimal reliance during COVID (0.16%) before partially rebounding post-COVID (1.26%), indicating volatility in how passengers weighted cabin quality. Refunds sentiment surged during COVID (32.60%) but collapsed to 1.49% post-COVID, underscoring its transitory importance tied to refund disputes.

The **topic_refunds_lounge** feature showed gradual but consistent growth, reaching 8.92% post-COVID, reflecting the embedding of refund and lounge concerns into evaluations beyond the immediate crisis. **Food_sentiment** also displayed strong drift, peaking at 6.77% during COVID before falling to 0.19% post-COVID.

Moderate drift drivers included **staff_service** (0.70% during → 18.60% post-COVID, Δ +17.9%) and **entertainment_score** (1.51% pre → 22.57% post-COVID), both showing marked rank increases but less instability compared to refund-related attributes.

Emirates Airlines

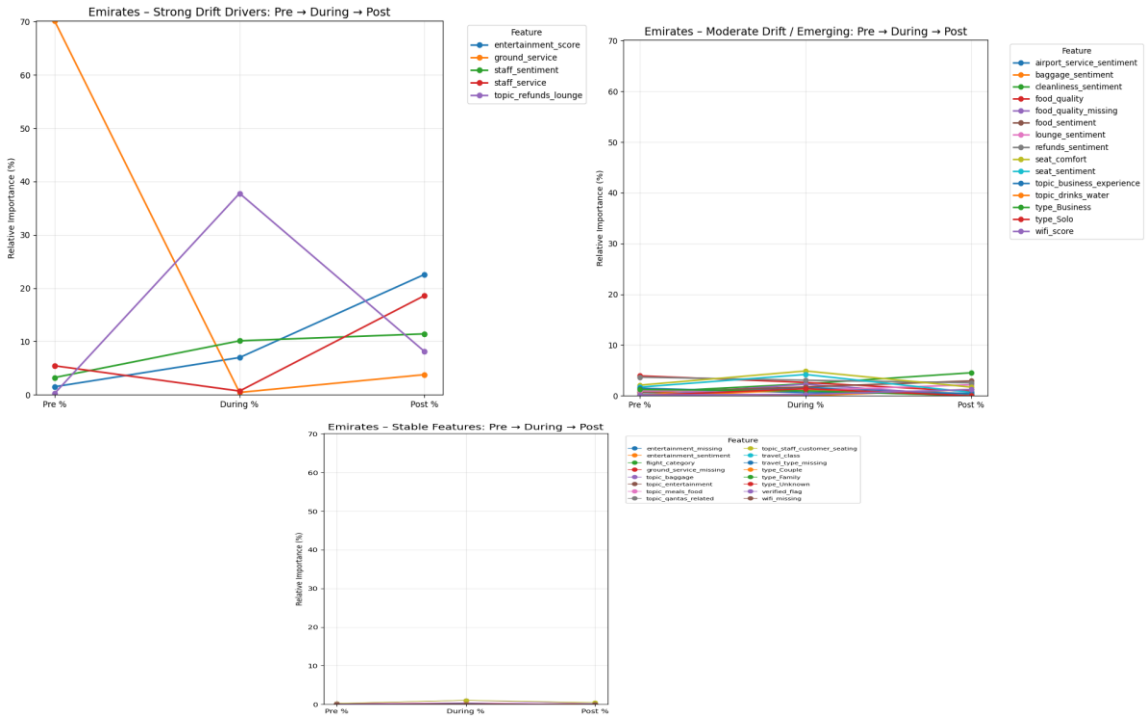


Figure 4.2.8– Feature Drift Drivers Across Phases (Pre–During–Post COVID) Emirates

For Emirates, drift was driven primarily by **sentiment-based features** rather than operational ones. **Entertainment_sentiment** and **seat_sentiment** rose from negligible pre-COVID levels to top-five positions during and post-COVID, displacing traditionally stable attributes. **Staff_sentiment** also increased sharply during COVID (≈2.36%) before declining post-COVID, while **refunds_sentiment** followed a similar trajectory.(Figure 4.2.8)

In contrast to British Airways, **ground_service** remained consistently important (≈10–13%) across all three periods, although its rank shifted relative to emergent sentiment-based features. **Staff_service**, while influential pre- and during-COVID (≈3–4%), declined post-COVID (0.18%), showing reduced weight in the recovery phase. Moderate drift drivers included **airport_service_sentiment** (0.24% pre → 1.66% during → 0.38% post), **topic_business_experience** (0.21% → 2.78% → 0.47%), and transient categorical factors such as **type_Solo** (1.40% during only) and **type_Business** (declining to 0% post-COVID).

4.2.2.3 Complementary Evidence on Feature Drift Dynamics

The comparative plots (*Figure 4.2.6. Drift vs. Reliance*) placed drift magnitude on the x-axis, reliance on the y-axis, and volatility as bubble size. For British Airways, the top-right quadrant (high drift, high importance) contained ground_service, staff_sentiment, seat_comfort, refunds_sentiment, and food_sentiment, with entertainment and lounge topics emerging as additional drivers. For Emirates, the quadrant was dominated by staff_sentiment, seat_sentiment, entertainment_sentiment, ground_service, and staff_service, with higher volatility in sentiment-based features than BA.

Overall, the results confirm statistically significant drift, but the nature of drift diverged ie. BA's reliance shifted toward operational and refund-related factors during the crisis and rebalanced to tangible services post-COVID, while Emirates' drift was concentrated on sentiment-driven evaluations, with operational attributes retaining stable weight across periods.

It is important to note that the features highlighted in the drift–reliance plots do not always perfectly overlap with those identified as drift drivers in the feature-level analysis. This discrepancy reflects methodological differences. The scatterplots emphasise *relative positioning* across importance and drift simultaneously, highlighting features that were both volatile and central in average reliance terms. In contrast, the drift-driver detection isolates features with the largest ***absolute shifts*** in reliance values or ranking across periods, regardless of their baseline magnitude, as how BA's **refunds_sentiment** shows up as a drift driver because of a sharp surge-and-collapse cycle, but its average reliance across periods is too low to dominate the top-right quadrant. Conversely, entertainment_sentiment for

Emirates gained importance steadily enough to rank in the scatterplot, even if its period-to-period jumps were less extreme than refund-related features.

Thus, the divergence between the two perspectives reflects different dimensions of drift: scatterplots capture ***salience-adjusted volatility***, while drift-driver detection captures ***absolute reweighting shocks***. Together, they provide complementary evidence on how brand perception attributes evolved.

4.2.2.4 Linking Concept Drift to Brand Perception Dynamics

The findings address the central research question: whether concept drift analysis can provide explanatory insight into the evolution of brand perception. The reliance shifts identified through MCR^+ mirrored documented changes in airline narratives across the pandemic. For British Airways, the surge of refund- and ground service–related reliance during COVID is consistent with widespread disputes over cancellations and complaints in the UK market (IATA, 2021). For Emirates, the rise of sentiment-based features, particularly staff and entertainment sentiment, aligns with its established positioning around premium service and passenger experience (O'Connell, 2011).

These trajectories highlight structurally different brand evolutions: BA exhibited crisis-sensitivity and reactive volatility, while Emirates showed greater stability, with incremental reweighting of sentiment-driven attributes. This pattern is consistent with Gama et al. (2014), who emphasise that concept drift manifests either through abrupt shocks or gradual reordering. Importantly, drift analysis adds explanatory value alongside interaction effects: while interaction analysis revealed how features combined to influence ratings at a given time, drift analysis demonstrated when and why certain features surged or declined in salience across phases of disruption and recovery. Together, they provide a complementary account of brand perception dynamics, showing both the structural timing of shifts and the mechanisms by which attributes interacted to shape evaluations.

4.3 Interaction Analysis Results & Discussion

This chapter reports the outcomes of the interaction analysis conducted across the three COVID-19 phases. The results are structured in three parts: (i) SHAP interaction rankings, (ii) baseline OLS model fit, and (iii) validation of top interactions through OLS with significance testing.

4.3.1 Interaction Analysis Results

4.3.1.1 British Airways Per Period Interaction Analysis:

Pre-COVID

=== Top 10 Interactions (pre-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
42	food_quality	staff_sentiment	0.103044
0	seat_comfort	staff_service	0.102450
2	seat_comfort	food_quality	0.089049
4	seat_comfort	food_sentiment	0.088691
40	food_quality	food_sentiment	0.076448
17	staff_service	food_sentiment	0.065339
6	seat_comfort	staff_sentiment	0.054244
1	seat_comfort	ground_service	0.047849
19	staff_service	staff_sentiment	0.047005
61	food_sentiment	staff_sentiment	0.042292

Table 4.3.1. Top 10 SHAP-Based Feature Interactions (Pre-COVID).

=== Pre-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	628.2
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	0.00
Time:	02:23:12	Log-Likelihood:	-4548.0
No. Observations:	2559	AIC:	9128.
Df Residuals:	2543	BIC:	9222.
Df Model:	15		
Covariance Type:	nonrobust		

=== Pre-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
1	seat_comfort × staff_service	63.244459	2.721528e-15	0.792632
4	food_quality × food_sentiment	47.679633	6.316829e-12	0.791386
3	seat_comfort × food_sentiment	45.693076	1.709269e-11	0.791226
2	seat_comfort × food_quality	41.674594	1.286759e-10	0.790901
0	food_quality × staff_sentiment	38.619301	6.000263e-10	0.790654

Table 4.3.2. OLS Regression Results and Interaction Tests (Pre-COVID).

For Pre, SHAP identified **food_quality × staff_sentiment** and **seat_comfort × staff_service** as key interactions. OLS confirmed **seat_comfort × staff_service** as strongest ($\Delta R^2 = +0.017$, $F = 63.24$, $p < 0.001$). Model fit was strong ($R^2 = 0.676$, $F = 1778$, $p < 0.001$), with both features significant ($\beta = 0.203$; $\beta = 0.240$) and **their interaction positive** ($\beta = 0.259$, $p < 0.001$).

During-COVID

=== Top 10 Interactions (During-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
42	food_quality	staff_sentiment	0.120879
38	ground_service	topic_business_experience	0.114934
31	ground_service	staff_sentiment	0.077118
4	seat_comfort	food_sentiment	0.072371
83	staff_sentiment	topic_business_experience	0.069454
37	ground_service	topic_refunds_lounge	0.049501
43	food_quality	refunds_sentiment	0.047688
1	seat_comfort	ground_service	0.043168
70	seat_sentiment	refunds_sentiment	0.041723
19	staff_service	staff_sentiment	0.041584

Table 4.3.3. Top 10 SHAP-Based Feature Interactions (during -COVID).

=== During-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.821
Method:	Least Squares	F-statistic:	59.04
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	3.12e-60
Time:	02:25:54	Log-Likelihood:	-330.38
No. Observations:	191	AIC:	692.8
Df Residuals:	175	BIC:	744.8
Df Model:	15		
Covariance Type:	nonrobust		

=== During-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
1	ground_service × topic_business_experience	7.359018	0.007343	0.841705
0	food_quality × staff_sentiment	6.580558	0.011153	0.841022
4	staff_sentiment × topic_business_experience	4.941944	0.027500	0.839567
3	seat_comfort × food_sentiment	3.279048	0.071895	0.838062
2	ground_service × staff_sentiment	2.048669	0.154133	0.836930

Table 4.3.4. OLS Regression Results and Interaction Tests

For During, SHAP highlighted **food_quality × staff_sentiment** and g**round_service × topic_business_experience**.OLS confirmed **ground_service × topic_business_experience** as strongest (ΔR² = +0.007, F = 7.36, p = 0.007), improving R² to 0.842. Ground service was significant (β = 1.675, p < 0.001), business experience not (β = −0.162, p = 0.829), while their interaction was **significant and negative (β = −1.339, p < 0.001)**.

Post-COVID

=== Top 10 Interactions (Post-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
28	ground_service	food_sentiment	0.122288
1	seat_comfort	ground_service	0.096917
36	ground_service	topic_refunds_lounge	0.073047
5	seat_comfort	staff_sentiment	0.056527
27	ground_service	food_quality	0.055680
71	staff_sentiment	lounge_sentiment	0.054335
47	food_quality	topic_refunds_lounge	0.044032
41	food_quality	staff_sentiment	0.042335
3	seat_comfort	food_sentiment	0.040130
50	food_sentiment	seat_sentiment	0.036543

Table 4.3.5. Top 10 SHAP-Based Feature Interactions (post -COVID).

=== Post-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.781
Method:	Least Squares	F-statistic:	111.6
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	3.92e-141
Time:	02:27:15	Log-Likelihood:	-828.19
No. Observations:	467	AIC:	1688.
Df Residuals:	451	BIC:	1755.
Df Model:	15		
Covariance Type:	nonrobust		

=== Post-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
4	ground_service × food_quality	37.845122	1.693724e-09	0.804237
1	seat_comfort × ground_service	30.784042	4.930580e-08	0.801362
0	ground_service × food_sentiment	28.141642	1.772484e-07	0.800264
3	seat_comfort × staff_sentiment	28.007715	1.891770e-07	0.800208
2	ground_service × topic_refunds_lounge	0.823112	3.647573e-01	0.788161

Table 4.3.6. OLS Regression Results and Interaction Tests (post -COVID)

While for Post, SHAP identified **ground_service × food_sentiment** and **seat_comfort × ground_service** as top pairs. OLS confirmed **ground_service × food_quality** as strongest (ΔR² = +0.016, F = 37.85, p < 0.001), raising R² to 0.804. Ground service was significant (β = 0.446, p = 0.001), food quality not (β = 0.083, p = 0.515), but their interaction was **significant and positive** (β = 0.284, p < 0.001).

4.3.1.2 Emirates Airlines Per Period Interaction Analysis:

Pre-COVID

=== Top 10 Interactions (pre-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
31	ground_service	seat_sentiment	0.188691
19	staff_service	seat_sentiment	0.121649
42	food_quality	seat_sentiment	0.086660
14	staff_service	ground_service	0.080598
20	staff_service	staff_sentiment	0.060436
69	food_sentiment	seat_sentiment	0.056067
77	seat_sentiment	staff_sentiment	0.055384
32	ground_service	staff_sentiment	0.051896
33	ground_service	refunds_sentiment	0.045868
28	ground_service	entertainment_score	0.043597

Table 4.3.7. Top 10 SHAP-Based Feature Interactions (Pre-COVID).

=== Pre-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.828
Method:	Least Squares	F-statistic:	346.9
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	0.00
Time:	03:01:33	Log-Likelihood:	-1858.6
No. Observations:	1080	AIC:	3749.
Df Residuals:	1064	BIC:	3829.
Df Model:	15		
Covariance Type:	nonrobust		

=== Pre-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
4	staff_service × staff_sentiment	22.474120	0.000002	0.833769
3	staff_service × ground_service	22.453861	0.000002	0.833766
1	staff_service × seat_sentiment	19.549665	0.000011	0.833320
2	food_quality × seat_sentiment	11.953255	0.000567	0.832142
0	ground_service × seat_sentiment	7.260244	0.007161	0.831406

Table 4.3.8. OLS Regression Results and Interaction Tests (Pre- COVID)

Now in Pre, SHAP identified **ground_service × seat_sentiment** and **staff_service × seat_sentiment** as top interactions, with **food_quality × seat_sentiment** and **staff_service × ground_service** also relevant. OLS confirmed **staff_service × staff_sentiment** as strongest ($\Delta R^2 = +0.004$, $p < 0.001$), raising R^2 to 0.834. Both staff_service and ground_service were significant, and sentiment-based interactions consistently **improved model** explanatory power.

During-COVID

=== Top 10 Interactions (During-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
82	staff_sentiment	topic_baggage	0.110768
14	staff_service	ground_service	0.087939
97	entertainment_sentiment	topic_baggage	0.072182
77	staff_sentiment	refunds_sentiment	0.058474
69	seat_sentiment	staff_sentiment	0.050885
88	refunds_sentiment	topic_baggage	0.050855
84	refunds_sentiment	lounge_sentiment	0.048996
43	food_quality	refunds_sentiment	0.046159
52	entertainment_score	staff_sentiment	0.045965
5	seat_comfort	seat_sentiment	0.044175

Table 4.3.9. Top 10 SHAP-Based Feature Interactions (during -COVID).

=== During-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.784
Model:	OLS	Adj. R-squared:	0.760
Method:	Least Squares	F-statistic:	33.08
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	3.93e-38
Time:	03:02:21	Log-Likelihood:	-281.87
No. Observations:	153	AIC:	595.7
Df Residuals:	137	BIC:	644.2
Df Model:	15		
Covariance Type:	nonrobust		

=== During-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
1	staff_service × ground_service	11.506357	0.000909	0.800502
3	staff_sentiment × refunds_sentiment	2.358465	0.126928	0.787312
4	seat_sentiment × staff_sentiment	2.128726	0.146867	0.786958
2	entertainment_sentiment × topic_baggage	1.422126	0.235130	0.785863
0	staff_sentiment × topic_baggage	0.612978	0.435030	0.784594

Table 4.3.10. OLS Regression Results and Interaction Tests (During-COVID)

In During, SHAP identified **staff_sentiment × topic_baggage** and **staff_service × ground_service** as top interactions, with secondary effects from **entertainment_sentiment × topic_baggage** and **staff_sentiment × refunds_sentiment.** OLS confirmed **staff_service × ground_service** as

strongest (ΔR^2 = +0.017, F = 11.51, p < 0.001), raising R^2 to 0.801. Ground- and staff-related features were **significant,** while baggage and refunds sentiment were weak and **non-significant.**

Post-COVID

=== Top 10 Interactions (Post-COVID) ===

	Feature1	Feature2	MeanAbsSHAP
14	staff_service	ground_service	0.124711
103	airport_service_sentiment	topic_baggage	0.072905
32	ground_service	staff_sentiment	0.069453
20	staff_service	staff_sentiment	0.062254
30	ground_service	food_sentiment	0.054479
21	staff_service	refunds_sentiment	0.048123
98	lounge_sentiment	topic_baggage	0.047975
89	staff_sentiment	topic_baggage	0.045547
0	seat_comfort	staff_service	0.044209
84	staff_sentiment	refunds_sentiment	0.043470

Table 4.3.11. Top 10 SHAP-Based Feature Interactions (Post -COVID).

=== Post-COVID Baseline OLS ===

OLS Regression Results

Dep. Variable:	y	R-squared:	0.826
Model:	OLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	71.52
Date:	Mon, 01 Sep 2025	Prob (F-statistic):	6.54e-77
Time:	03:03:09	Log-Likelihood:	-414.78
No. Observations:	242	AIC:	861.6
Df Residuals:	226	BIC:	917.4
Df Model:	15		
Covariance Type:	nonrobust		

=== Post-COVID Interaction Tests (OLS) ===

	Interaction	F	p_value	R2
0	staff_service × ground_service	32.636344	3.498891e-08	0.848031
3	staff_service × staff_sentiment	11.362939	8.817129e-04	0.834354
4	ground_service × food_sentiment	5.823908	1.661043e-02	0.830379
2	ground_service × staff_sentiment	4.598568	3.307065e-02	0.829473
1	airport_service_sentiment × topic_baggage	0.564373	4.532886e-01	0.826424

Table 4.3.12 OLS Regression Results and Interaction Tests.(Post -COVID)

As for Post, SHAP identified **staff_service × ground_service** as the top pair, with **airport_service_sentiment × topic_baggage** also emerging. OLS confirmed **staff_service × ground_service** as strongest (ΔR^2 = +0.022, F = 32.63, p < 0.001), raising R^2 to 0.848. Staff service was **significant** (β = 0.246, p < 0.001), ground service also **significant**, and their interaction reinforced **explanatory power**, while airport–baggage remained non-significant (p = 0.827).

Period	R^2	F-statistic (p)	Significant Predictors	Interaction Effect (β, CI, p)
Pre-COVID	0.676	1778, p < 0.001	Seat comfort (β=0.203, p=0.006); Staff service (β=0.240, p<0.001)	Seat_comfort × Staff_service: β=0.259, CI [0.221–0.297], p<0.001
During-COVID	0.67	126.3, p < 0.001	Ground service (β=1.675, p<0.001)	Ground_service × Business_experience: β=-1.339, CI [-1.886–-0.793], p<0.001
Post-COVID	0.695	351.3, p < 0.001	Ground service (β=0.446, p=0.001)	Ground_service × Food_quality: β=0.284, CI [0.203–0.365], p<0.001

Table 4.3.13 Validated OLS Models for British Airways Across COVID Phases.

Period	R^2	F-statistic (p)	Significant Predictors	Interaction Effect (β, CI, p)
Pre-COVID	0.694	812.6, p<0.001	Staff sentiment (β=0.990, p=0.006); Staff service (β=1.068, p<0.001)	staff_sentiment × staff_service: β=0.461, CI [0.313–0.637], p<0.001
During-COVID	0.653	93.6, p<0.001	Ground service (β = −0.191, p = 0.484); Staff service (β = −0.431, p = 0.067)	ground_service × staff_service: β=0.445, CI [0.342–0.551], p<0.001
Post-COVID	0.804	325.8, p<0.001	Staff service (β=0.424, p=0.001)	ground_service × staff_service: β=0.246, CI [0.148–0.323], p<0.001

Table 4.3.14 Validated OLS Models for Emirates Airlines Across COVID Phases.

Table 4.3.13 and 4.3.14 shows the validated OLS models for British Airways and Emirates across the three COVID-19 phases. For British Airways, seat comfort and staff service were significant pre-COVID, while ground service became the main predictor during and post-COVID. For Emirates, staff sentiment and staff service were key pre-COVID, while ground–staff interactions emerged during COVID, and staff–ground interactions remained significant post-COVID. Across both airlines, R^2 values were stable between 0.65 and 0.80.

4.3.2 Interaction Analysis Discussion

4.3.2.1 Observations of the Interactions

The interaction analysis revealed that passengers did not evaluate airline services on isolated attributes alone; rather, their overall ratings were shaped by combinations of features that either reinforced or undermined one another. This aligns with the argument in Aas et al. (2021) that interaction effects provide a behaviourally realistic account of decision-making beyond additive models. Importantly, the observed interactions were **statistically validated through OLS regression with F-tests, permutation checks, and bootstrap confidence intervals**, confirming that the effects were genuine and not artefacts of a single model specification.

4.3.2.2 Comparative findings of the Interactions

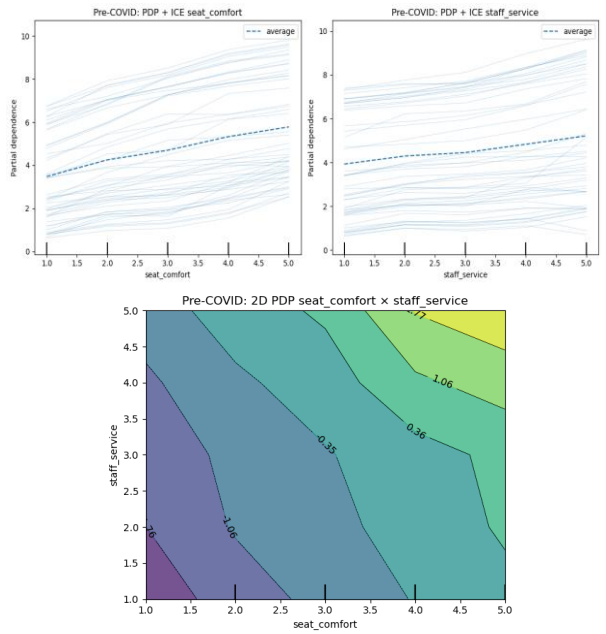
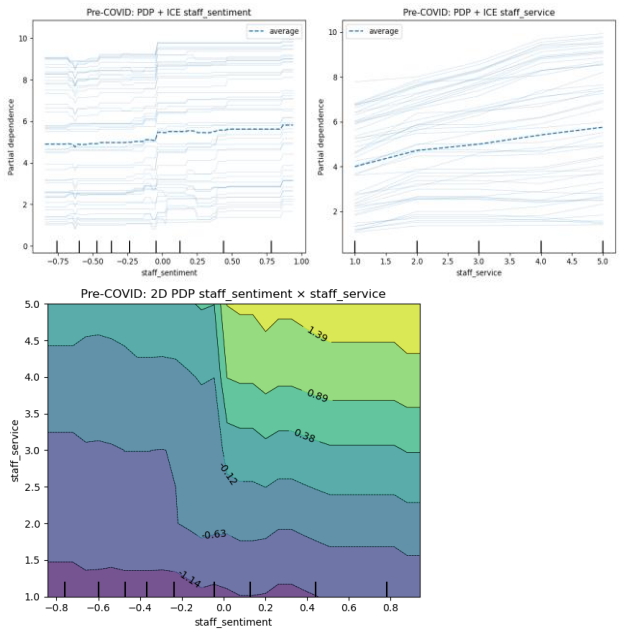
British Airways	Emirates Airlines
	

Table 4.3.15 Pre-COVID PDP and ICE Plots with Interaction Surfaces for British Airways and Emirates

Pre-COVID

In Table 4.3.15,**British Airways** the dominant interaction was ***seat comfort* × *staff service*.** OLS confirmed its significance (β = 0.259, p < 0.001), with a marked increase in R^2 (+0.017). PDP/ICE plots showed that ratings increased only when both were **high**, indicating BA's evaluations were driven by product–service consistency. Comfort alone or service alone did not sustain strong ratings.For Emirates, the strongest driver was ***staff sentiment* × *staff service*.** SHAP ranked this interaction highest. OLS validation showed a stable, significant effect ($R^2 \approx 0.793$).The plots revealed asymmetry, negative sentiment sharply lowered ratings even when service was strong, whereas positive sentiment reinforced service to yield the highest scores.

Therefore, BA's perception was anchored in **synergistic product–service consistency**, while Emirates' relied on **reinforcing emotional–service dynamics marked by asymmetry**,BA's perception was anchored in **cabin and staff consistency**, whereas Emirates' relied on **emotional–service reinforcement**. Both highlight service, but BA leaned toward tangible cabin factors, while Emirates leaned toward human interaction.

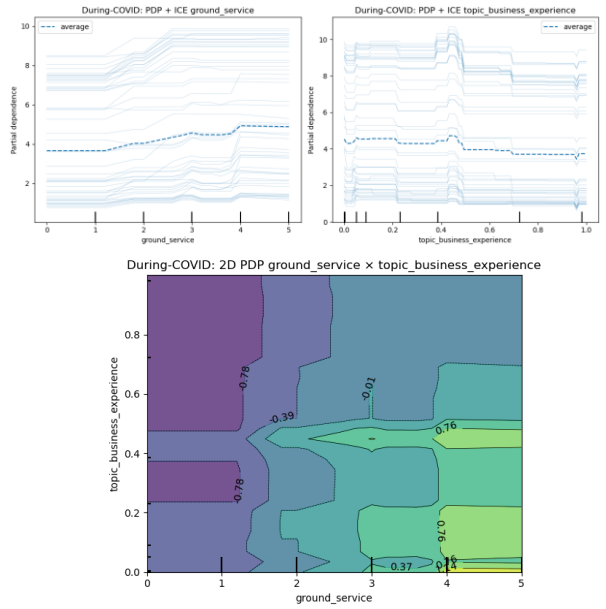
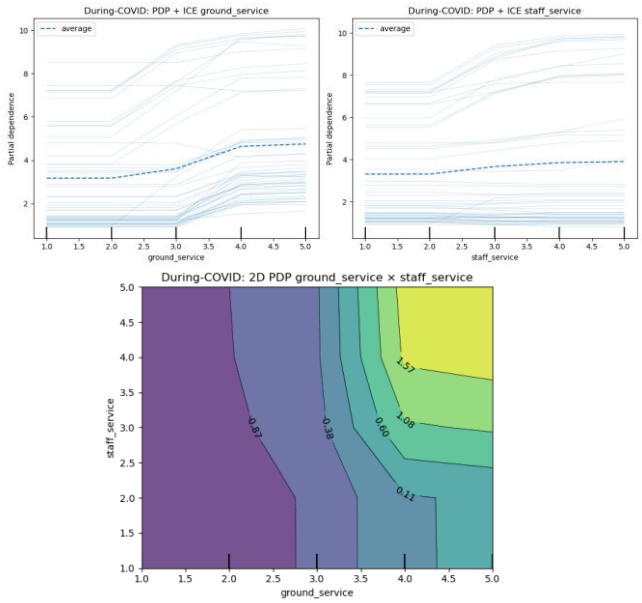
British Airways	Emirates Airlines
	

Table 4.3.16 During COVID PDP and ICE Plots with Interaction Surfaces for British Airways and Emirates

During-COVID

British Airways's standout interaction was ***ground service × business experience***. OLS results confirmed significance, the effect was **negative** i.e. ratings dropped sharply when ground handling conflicted with business travel needs. PDP/ICE plots illustrated that no compensating uplift from other features occurred under poor ground handling.For EmiratesThe key driver was ***ground service × staff service***. This interaction significantly boosted explanatory power ($\beta = 0.445$, $p < 0.001$;$R^2 = 0.653$). Ratings climbed when both elements were **strong**, but **declined** if either lagged.(Table 4.3.16)

In both airlines, ground service dominated brand perception. However, BA's interaction was **conflictual** (linked to disruptions and refund issues), while Emirates' was **synergistic** (smooth coordination between ground and staff services).

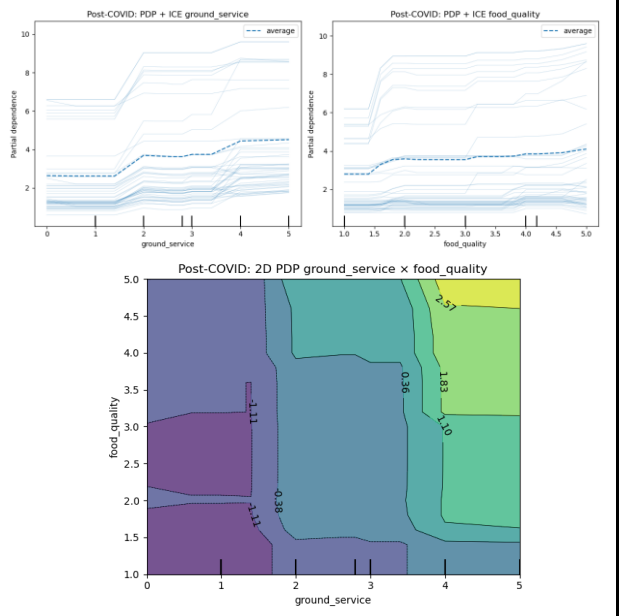
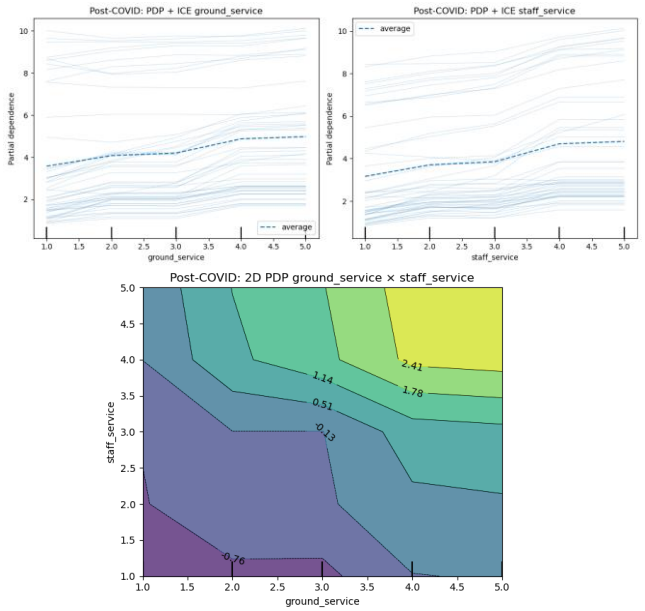
British Airways	Emirates Airlines
	

Table 4.3.17 Post-COVID PDP and ICE Plots with Interaction Surfaces for British Airways and Emirates

Post-COVID

From table 4.3.17,British Airways's leading interaction was ***ground service × food quality***. OLS validation confirmed its significance (β = 0.284, p < 0.001, F = 37.85, R^2 = 0.804). **Ratings rose** when strong ground service was combined with improved food quality, reflecting a shift toward tangible recovery elements. In Emirates, the main driver remained ***ground service × staff service***. OLS confirmed it as strongly significant (β = 0.246, p < 0.001, True R^2 = 0.804). PDP/ICE plots showed ratings **increased steadily** when both dimensions aligned.

Therefore, BA's perception reflected a **synergistic recovery dynamic**, where ground operations reinforced food quality to restore brand value, while Emirates maintained a **stable synergy** between ground and staff service. In comparison, BA's perception shifted to **ground + food recovery**, while Emirates retained a **ground + staff service foundation**. This reflects BA's reactive adjustment after reputational damage versus Emirates' continuity in service delivery.

4.3.2.3 Synthesis of findings

Across all three periods:

- **British Airways** showed **volatility**: cabin–staff consistency pre-COVID, ground–business disruption during COVID, and ground–food recovery post-COVID.
- **Emirates** showed **stability**: consistently service-focused, anchored in staff and operational performance across all phases.

The robustness of these results was reinforced by consistency across SHAP rankings, OLS regression, F-tests, and bootstrap validation, strengthening confidence that the observed effects were not artefacts of a single model specification but reflected genuine behavioural mechanisms in passenger evaluation.

The prominence of ground service in both carriers during and after COVID is consistent with industry reports (IATA, 2021) that identified airport handling, cancellations, and refund disputes as key determinants of passenger satisfaction in the pandemic and recovery period.

The analyses confirmed that brand perception was statistically captured through feature interactions. SHAP consistently highlighted key interacting features across phases, and OLS regression verified their significance through formal tests. For British Airways, interactions such as seat comfort with staff service pre-COVID, and ground service with food quality post-COVID, were validated as significant explanatory mechanisms. For Emirates, staff-related and ground–staff interactions were repeatedly confirmed as significant across all phases. These results establish that the explanatory structure of brand perception operated through validated feature interactions rather than isolated attributes.

5. Conclusions

This study set out to answer the research question:

To what extent can Model Class Reliance (MCR), feature drift analysis, and feature interaction analysis provide explanatory—rather than purely predictive—insights into the evolution of airline brand perception across crisis phases?

The findings confirm that explanatory insight is possible within a predictive modelling context. MCR^{+} quantified feature necessity with conservative reliance bounds across the Rashomon set, drift analysis identified when and how reliance shifted under crisis conditions, and interaction analysis revealed how service attributes reinforced or offset one another in shaping customer evaluations. Together, these methods offered a transparent account of why predicted ratings rose or fell over time, and why British Airways and Emirates followed distinct brand trajectories during and after COVID-19. Importantly, the study moved beyond conventional sentiment analysis and single-model prediction, which often provide static or unstable insights, by introducing validated explanatory methods that captured both the stability and evolution of brand drivers across contexts

Objective 1 focused on predicting overall passenger ratings with a combined feature space. Random Forests provided strong baseline performance, confirming their suitability for heterogeneous airline review data. Performance declines when pre-COVID models were applied to later phases highlighted the presence of concept drift, reinforcing the need for explanatory approaches rather than static predictions.

Objective 2 applied MCR to establish model-agnostic reliance bounds. Here, the study explicitly emphasised **MCR^{+}** over MCR^{-}, because MCR^{-} can understate feature necessity in the presence of correlated substitutes, shifting reliance artificially across redundant predictors. By contrast, **MCR^{+} captures the maximum possible reliance of each feature across the Rashomon set**, providing a conservative and interpretable upper bound for temporal comparisons (Fisher et al., 2019; Molnar, 2022). The findings showed that for British Airways, reliance was unstable, with features such as staff sentiment and ground service gaining weight abruptly during the crisis. For Emirates, reliance remained anchored in sentiment features with narrower reweighting. This demonstrated both the robustness of MCR^{+} as a reliance measure and its explanatory value for perception research.

Objective 3 evaluated temporal changes in reliance through concept drift analysis. Reliance distributions were statistically tested using Kruskal–Wallis for global differences, Kolmogorov–Smirnov for pairwise shifts, and Chi-square tests for error distribution changes. These validated that reliance shifts were not random

fluctuations but systematic structural changes in feature–rating relationships. Substantively, British Airways displayed abrupt drift: refund- and disruption-related attributes spiked during COVID before receding, while ground service and staff service reasserted importance post-crisis. Emirates, in contrast, exhibited incremental drift, with sentiment-based features steadily gaining weight over time. Together, these results showed how statistical validation of drift aligns with perceptual dynamics in brand evaluation—abrupt volatility for BA versus smoother reordering for Emirates.

Objective 4 addressed feature interactions. Statistical validation was conducted using **bootstrapped permutation tests**, which confirmed whether observed synergies or antagonisms were stronger than expected under randomised feature distributions. The tests showed that interaction effects were not artefacts of sampling but represented genuine dependencies within the data. For British Airways, significant negative interactions emerged between staff service and ground operations during COVID, illustrating compounding penalties when multiple service touchpoints failed simultaneously. For Emirates, significant positive reinforcement was detected between staff sentiment and entertainment sentiment, producing a more resilient explanatory structure. These results underscored both the statistical validity and substantive importance of interactions: passengers evaluate airlines through interdependent attributes, and failures or strengths in one domain can amplify effects in another.

Overall, the combined use of MCR^+, drift, and interaction analyses confirmed that explanatory insight is possible in a predictive modelling context. MCR^+ quantified feature necessity with conservative stability bounds, drift analysis revealed when and how reliance shifted under crisis conditions, and interaction analysis clarified how attributes reinforced or offset one another in shaping evaluations. This triangulation provided a transparent account of why overall rating predictions rose or fell, and how brand perception evolved differently for carriers with distinct strategic positions.

5.1 Limitations

Several limitations should be acknowledged when interpreting the findings of this study. First, the reliance on online customer reviews (OCR) introduces potential **data bias**, as such reviews are often written by passengers with particularly positive or negative experiences, thereby underrepresenting the "average" customer voice. Second, the dataset itself was **limited in size** and contained **missing values** across structured sub-ratings (e.g., food or entertainment). Although imputation strategies were applied, both the small sample sizes within certain periods and the presence of missing data may have constrained the reliability of feature reliance estimates.

Third, the study's **temporal framing** divided reviews into discrete pre-, during-, and post-COVID windows. This facilitated comparability but oversimplifies the continuous evolution of brand perception, potentially masking gradual shifts. Fourth, the restriction to **Random Forests** as the sole learner family, necessitated by the current operationalisation of Model Class Reliance (MCR) for tree ensembles, limits generalisability; reliance patterns may differ if MCR were extended to boosting or neural models.

Fifth, **MCR itself carries methodological constraints**. The approach is computationally expensive, particularly when generating Rashomon sets across multiple carriers and time periods. Reliance intervals are also sensitive to **sample size**, meaning that in relatively sparse review subsets, intervals may appear artificially wide. The study emphasised **MCR$^+$** (upper bounds) for drift analysis because it provides a conservative, interpretable measure of maximum reliance. However, this choice may risk overstating feature importance in cases of strong collinearity. Finally, MCR remains a **relatively new framework**, meaning that its interpretive conventions are not yet as established as those of more widely used importance measures.

Together, these limitations underscore that while the results provide valuable explanatory insights, they should be interpreted as **indicative rather than definitive** patterns of brand perception dynamics.

5.2 Future Research

Several promising directions emerge from this study.Firstly, future work could explore **real-time drift detection methods**, such as streaming algorithms, to capture shifts in brand perception as they occur rather than retrospectively. This would increase the practical relevance of reliance-based monitoring for airline management.

Second, the dataset could be broadened to include **multiple carriers across diverse markets**, enabling cross-sectional comparisons between legacy, low-cost, and premium airlines. Such expansion would test the external validity of findings beyond British Airways and Emirates.

Third, future studies should aim to **triangulate OCR-based results** with alternative data sources such as structured surveys (e.g., Net Promoter Score) or operational metrics (e.g., financial performance, complaint volumes). This would help validate reliance-driven insights against managerial benchmarks.

Fourth, more work is needed on **feature interaction and drift within the MCR framework**. While this study applied drift and interaction analyses sequentially,

future research could integrate them more tightly, examining how reliance intervals themselves reflect interaction synergies or antagonisms over time. This would allow for a more dynamic explanatory model of brand perception.

Finally, future research could investigate whether **MCR can be adapted as a direct tool for drift and interaction detection**, rather than only serving as an input for subsequent analyses. Developing MCR-based drift indices or interaction-aware reliance measures would extend its explanatory power and offer a unified framework for tracking how feature relevance evolves under changing conditions.

6. Appendix

Appendix A

Instructions for Using the Jupyter Notebooks

1. **Environment Setup**
    - Install Python 3.9+ and Jupyter Notebook or JupyterLab.
    - Install required libraries listed in the project (e.g., `pandas`, `numpy`, `scikit-learn`, `matplotlib`, `seaborn`, `shap`, `statsmodels`).

2. **Data Files**

    - Place the datasets (*BA_AirlineReviews.csv* and *Emirates Airways Reviews.csv*) in the working directory.
    - Ensure filenames are unchanged to match notebook references.

3. **Execution Order**

    - Run notebooks in sequence:

        1. *1.Data_Input_Processing.ipynb*
        2. *2.Data_Missing_Analysis.ipynb*
        3. *3.Data_Sentimental_Analysis.ipynb*
        4. *4.Feature_Engineering_Cleaning_BA.ipynb* / *5.Feature_Engineering_Cleaning_Em.ipynb*
        5. *6.Prediction_Star_BA_EM.ipynb*
        6. *7.MCR_BA_EM_Analysis_F.ipynb*
        7. *8.BA_Feature_Drift.ipynb* / *9.Em_Feature_Drift.ipynb*
        8. *10.BA_Feature_Interaction.ipynb* / *11.Em_Feature_Interaction.ipynb*
4. **Outputs**

    - Each notebook produces plots, tables, and metrics referenced in the dissertation chapters.
    - Outputs include feature reliance plots, drift heatmaps, SHAP interaction rankings, OLS validation tables, and PDP/ICE surfaces.

5. **Reproducibility**
    - Random seeds were fixed where applicable.
    - Results should replicate within minor numerical tolerances due to stochastic resampling.

Python Notebooks and Functions

1. **1.Data_Input_Processing.ipynb** – Imported, cleaned, and structured raw airline review data.
2. **2.Data_Missing_Analysis.ipynb** – Diagnosed missingness (MCAR/MAR/MNAR) and applied imputation with flags.
3. **3.Data_Sentimental_Analysis.ipynb** – Extracted sentiment polarity, emotion scores, and topic models from text.

4. **4.Feature_Engineering_Cleaning_BA.ipynb** – Processed British Airways sub-ratings, categorical encodings, and text-derived features.
5. **5.Feature_Engineering_Cleaning_Em.ipynb** – Processed Emirates Airlines sub-ratings, categorical encodings, and text-derived features.

6. **6.Prediction_Star_BA_EM.ipynb** – Trained predictive models for overall star ratings using BA and Emirates datasets.
7. **7.MCR_BA_EM_Analysis_F.ipynb** – Computed MCR^+ reliance bounds for stable feature identification across both airlines.
8. **8.BA_Feature_Drift.ipynb** & **9.Em_Feature_Drift.ipynb** – Conducted drift analysis (Chi-square, KS, Kruskal–Wallis) and feature rank shifts across COVID phases.
9. **10.BA_Feature_Interaction.ipynb** & **11.Em_Feature_Interaction.ipynb** – Identified and validated feature interactions via SHAP rankings, OLS tests, and PDP/ICE plots.

Appendix B

Additional Information

Table 6.1: Original Passenger Review Features and Descriptions

Features Originally	Description
overall_rating	Target variable; star rating given by the passenger [1–10]
review_title	Short headline of the passenger's review (string).
review_date	Date when the review was posted (datetime).
verified_flag	Boolean flag indicating if the review was verified (1 = verified).
review_text	Full written passenger review (string, free-form).
travel_type	Category of trip (e.g., leisure, business, couple, solo, family).
travel_class	Cabin class (Economy, Premium Economy, Business, First)
flight_category	Flight haul (Short,Medium,Long)
seat_comfort	Passenger's numeric rating for seat comfort (1–5).
staff_service	Passenger's numeric rating for staff service (1–5).
ground_service	Passenger's numeric rating for ground/airport service (1–5).
value_for_money	Numeric rating for value-for-money (1–5).
recommended	Binary indicator if passenger would recommend airline (Yes/No).
food_quality	Numeric rating for food/beverage quality (1–5).
entertainment	Numeric rating for inflight entertainment (1–5).
wifi	Numeric rating for inflight Wi-Fi connectivity (1–5).

Figure 6.1 Missingness Analysis for British Airways Features

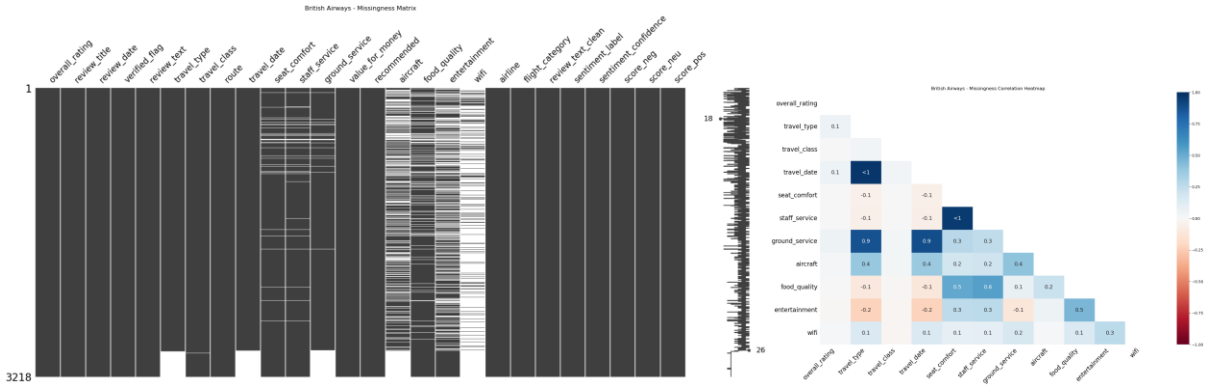


Figure 6.2 Missingness Analysis for Emirates Features

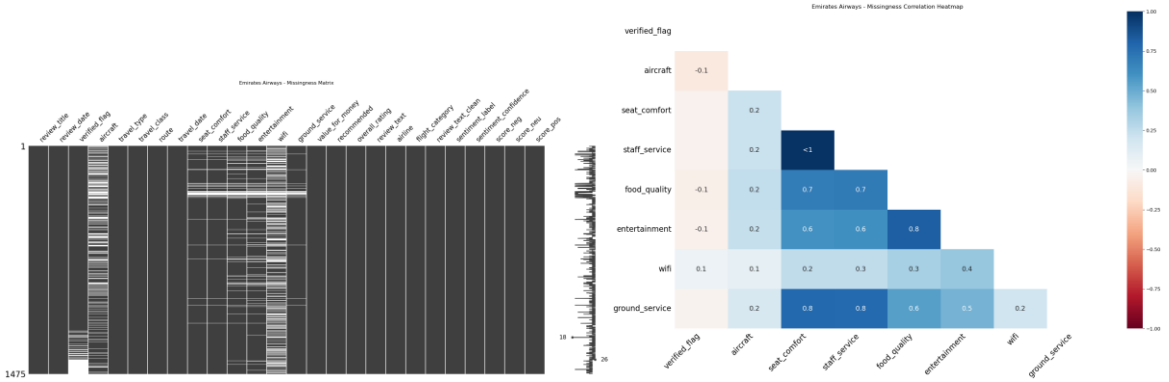


Table 6.2:Feature engineered descriptions

Feature Engineered	Description Datatype	
food_sentiment	Sentiment score for food/meals mentions	float64
seat_sentiment	Sentiment score for seat comfort mentions	float64
staff_sentiment	Sentiment score for staff/crew mentions	float64
baggage_sentiment	Sentiment score for baggage/luggage handling	float64
refunds_sentiment	Sentiment score for refund-related experiences	float64
lounge_sentiment	Sentiment score for lounge mentions	float64

entertainment_sentiment	Sentiment score for inflight entertainment	float64
cleanliness_sentiment	Sentiment score for cleanliness/cabin hygiene	float64
airport_service_sentiment	Sentiment score for ground/airport service	float64
anger	Emotion probability for anger in text	float64
joy	Emotion probability for joy	float64
optimism	Emotion probability for optimism	float64
sadness	Emotion probability for sadness	float64
topic_entertainment	Probability review is about entertainment	float64
topic_staff_customer_seating	Probability review is about staff/service/seating	float64
topic_refunds_lounge	Probability review is about refunds/lounges	float64
topic_business_experience	Probability review is about business travel	float64
topic_meals_food	Probability review is about meals/food	float64
topic_baggage	Probability review is about baggage/luggage	float64
topic_qantas_related	Probability review is about Qantas (competitor)	float64
topic_drinks_water	Probability review is about drinks/water	float64
type_Business	Dummy: Business travel	bool (0/1)
type_Couple	Dummy: Couple travel	bool (0/1)
type_Family	Dummy: Family travel	bool (0/1)
type_Solo	Dummy: Solo travel	bool (0/1)
type_Unknown	Dummy: Unknown travel type	bool (0/1)
travel_type_missing	Indicator if travel type missing	int64
food_quality_missing	Flag: food_quality rating missing	int64

wifi_missing	Flag: wifi rating missing	int64
entertainment_missing	Flag: entertainment rating missing	int64
ground_service_missing	Flag: ground_service rating missing	int64

Table 6.3 Missing Logistic Regression test

Variable	Overall Missingness	Significant Factors (Tests)	Logistic Regression Key Results	Interpretation
Wifi Score	**39.25%** (very high)	**Travel Class** (Chi² p=0.0032) → strong dependenceFlight Category: ns (p=0.99)Distance: ns (p=0.85)	Travel Class: borderline (**p=0.051**)Flight Category & Distance: ns	**Severe MAR issue**. Wifi score missingness is systematic across classes but not by distance or category. Needs careful imputation.
Food Quality	**7.80%**	**Travel Class** (Chi² p=0.0040) → higher in lower classesDistance (KW p=0.0228) → longer flights more missingFlight Category: ns	Travel Class: **negative coef, p=0.001** → higher class = less missingDistance: borderline (p=0.084)	**Moderate MAR**. Missingness depends mainly on travel class and slightly on distance.
Entertainment Score	**10.24%**	**Travel Class** (p=0.0052) significantDistance (p=0.023) significantFlight Category: ns	Travel Class: **p=0.002** (less missing in higher classes)Distance: **p=0.014** (longer flights = more missing)	**Moderate MAR**. Strong class + distance dependence.
Ground Service	**2.51%** (low)	**Flight Category** (p=0.0002) strongTravel Class borderline (p=0.0567)Distance: ns	Travel Class: **p=0.016**Flight Category: nsDistance: ns	**Low MAR**. Mostly systematic by category & class, but overall low missingness.

Table 6.4 Missing tests and observations

Variable	Overall Missingness	Significant Factors (Tests)	Logistic Regression Key Results	Interpretation
Wifi Score	**39%** (very high)	Travel Class: Chi² not shown, but regression suggests no strong effectFlight Category: ns (p=0.99)Distance: ns (p=0.27)	Flight Category: **negative coef, p<0.001** → certain categories less missingTravel Class: ns	**Severe MAR issue**, but unlike earlier airline, here **flight category** drives missingness..

Entertainment Score	**35.43%** (very high)	**Travel Class** (p=0.0000) → lower classes have more missing**Flight Category** (p=0.0000) → strongest effect**Distance** (p=0.0000) → shorter flights have more missing	Flight Category: **coef = -1.35, p<0.001** → certain categories far less missingDistance: **negative coef, p<0.001** → missingness higher on shorter flightsTravel Class: ns	**Strong MAR**, driven by **flight category + flight distance**. Systematic bias if dropped.
Food Quality	**11.96%**	**Travel Class** (p=0.0000) → lower classes more missing**Flight Category** (p=0.0000)**Distance** (p=0.0000) → shorter flights more missing	Travel Class: **coef = -0.57, p<0.001**Flight Category: **coef = -0.68, p<0.001**Distance: ns	**Moderate MAR**, primarily by **class + category**, distance effect visible in univariate but not regression.
Ground Service	**11.28%**	**Travel Class** (p=0.0061)**Flight Category** (p=0.0000)Distance: ns	Travel Class: borderline (p=0.096)Flight Category: nsDistance: ns	**Moderate MAR**, systematic by category and slightly by class, but logistic regression suggests weak effects.

Table 6.5: Summary of Missing Data Types and Suggested Treatments (BA vs. Emirates)

Airline	Column	Missing %	Likely Missingness Type	Suggested Treatment	Rationale
British Airways	**wifi**	81.08%	MNAR-ish	Add missingness flag + knn Impute	Extremely high missingness; reflects lack of WiFi on many flights. Imputing would distort availability vs quality.
	entertainment	35.43%	MNAR-ish	Add missingness flag only	Often missing because no IFE on short-haul flights. Do not fabricate quality ratings where no system existed.
	food_quality	10.43%	MAR	Impute	Correlated with other service ratings; recoverable via MICE or regression imputation.
	ground_service	11.28%	MAR	Add missingness flag + Knn impute	Tied to travel type & flight category; sometimes "not applicable" (e.g., transit).
	travel_type	~9%	MAR	Impute	Categorical; linked with route/class.
	staff_service	3.43%	MAR	Impute	Strongly predictable from other service scores.

	seat_comfort	3.13%	MAR	Impute	Similar to staff_service; reliable to recover.
	travel_class	0.05%	MCAR	Simple impute (mode)	Negligible; no flag needed.
	overall_rating	0.14%	MCAR	Simple impute (median)	Very low missingness; safe to impute directly.
Emirates Airways	**wifi**	39.35%	MAR	Add flag + knn impute	Missingness tied to aircraft/route; can impute, unlike BA where it signals absence.
	entertainment	10.39%	MAR	Add flag + knn impute	Still linked to route/class, but less severe; flag preserves availability differences.
	verified_flag	9.22%	Possibly MCAR	Simple impute ("Unknown")	Low missingness; not strongly linked to other features.
	food_quality	7.92%	MAR	Impute	Strong correlations with other scores.
	staff_service	4.16%	MAR	Impute	Strong predictor links.
	seat_comfort	4.09%	MAR	Impute	Strong predictor links.
	ground_service	2.53%	MAR	Add flag + knn impute	Informative for drift; low but non-random.F

Figure X.Y: Correlation Matrix of Passenger Review Features (British Airways)

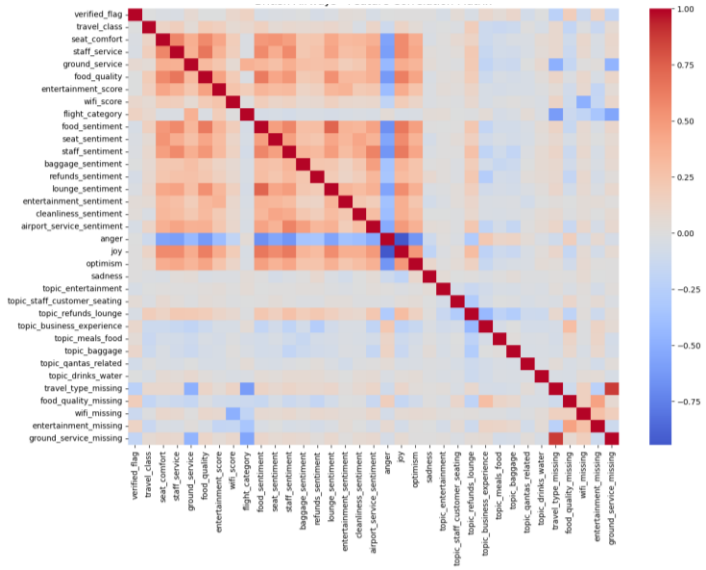


Table 6.6: Random Forest Model Performance Across COVID Phases – British Airways

Airline Period	Best CV R^2	Test R^2	Test RMSE	Best Params
Pre-COVID	0.811	0.820	1.288	n_estimators=200, min_samples_split=2, min_samples_leaf=2, max_features="sqrt", max_depth=None
During-COVID	0.798	0.798	1.558	n_estimators=200, min_samples_split=5, min_samples_leaf=2, max_features="log2", max_depth=30
Post-COVID	0.819	0.750	1.547	n_estimators=200, min_samples_split=5, min_samples_leaf=1, max_features="sqrt", max_depth=20

Table 6.7: Random Forest Model Performance Across COVID Phases – Emirates

Airline Period	Best CV R^2	Test R^2	Test RMSE	Best Params
Pre-COVID	0.836	0.831	1.310	n_estimators=500, min_samples_split=5, min_samples_leaf=2, max_features="sqrt", max_depth=30
During-COVID	0.751	0.802	1.549	n_estimators=300, min_samples_split=2, min_samples_leaf=1, max_features="log2", max_depth=30
Post-COVID	0.766	0.804	1.419	n_estimators=200, min_samples_split=5, min_samples_leaf=2, max_features="log2", max_depth=30

References

Aaker, D.A. (1991) *Managing brand equity: Capitalizing on the value of a brand name*. New York: Free Press.

Aas, K., Jullum, M. and Løland, A. (2021) 'Explaining individual predictions when features are dependent: More accurate approximations to Shapley values', *Artificial Intelligence*, 298, 103502.

Anderson, P., Fernando, B., Johnson, M. and Gould, S. (2016) 'SPICE: Semantic propositional image caption evaluation', *European Conference on Computer Vision (ECCV)*, pp. 382–398.

Archak, N., Ghose, A. and Ipeirotis, P.G. (2011) 'Deriving the pricing power of product features by mining consumer reviews', *Management Science*, 57(8), pp. 1485–1509.

Baker, T.L., Donthu, N. and Kumar, V. (2016) 'Investigating how word-of-mouth conversations about brands influence purchase and retransmission intentions', *Journal of Marketing Research*, 53(2), pp. 225–239.

Budd, L., Ison, S. and Adrienne, N. (2021) 'European airline response to the COVID-19 pandemic – Contraction, consolidation and future considerations', *Research in Transportation Business & Management*, 37, 100578.

CAPA – Centre for Aviation (2021) *Airline strategies in the COVID-19 recovery*. Available at: https://centreforaviation.com/ (Accessed: 4 September 2025).

Doganis, R. (2019) *Flying off course: Airline economics and marketing*. 5th edn. London: Routledge.

Filieri, R. (2015) 'What makes an online consumer review trustworthy?', *Journal of Business Research*, 68(6), pp. 1261–1270.

Fisher, A., Rudin, C. and Dominici, F. (2019) 'All models are wrong, but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance', *Journal of Machine Learning Research*, 20(177), pp. 1–81.

Friedman, J.H. and Popescu, B.E. (2008) 'Predictive learning via rule ensembles', *The Annals of Applied Statistics*, 2(3), pp. 916–954.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) 'A survey on concept drift adaptation', *ACM Computing Surveys*, 46(4), pp. 1–37.

Han, H., Yu, J. and Kim, W. (2019) 'Environmental corporate social responsibility and the strategy to boost the airline's image and customer loyalty intentions', *Journal of Travel & Tourism Marketing*, 36(3), pp. 371–383.

Hanlon, P. (2007) *Global airlines: Competition in a transnational industry*. 3rd edn. Oxford: Butterworth-Heinemann.

Hooker, G., Mentch, L. and Zhou, S. (2021) 'Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance', *Statistical Science*, 36(2), pp. 264–278.

Hu, N., Pavlou, P.A. and Zhang, J. (2009) 'Why do online product reviews have a J-shaped distribution? Overcoming bias in online word-of-mouth communication', *Communications of the ACM*, 52(10), pp. 144–147.

International Air Transport Association (IATA) (2021) *Annual review 2021*. Available at: https://www.iata.org/en/publications/annual-review/ (Accessed: 4 September 2025).

International Civil Aviation Organization (ICAO) (2020) *Effects of novel coronavirus (COVID-19) on civil aviation*. Available at: https://www.icao.int/ (Accessed: 4 September 2025).

Keller, K.L. (1993) 'Conceptualizing, measuring, and managing customer-based brand equity', *Journal of Marketing*, 57(1), pp. 1–22.

Liou, J.J.H. and Tzeng, G.H. (2007) 'A non-additive model for evaluating airline service quality', *Journal of Air Transport Management*, 13(3), pp. 131–138.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp. 4765–4774.

Mathews, T., Xie, L. and He, X. (2016) 'SentiCap: Generating captions with sentiment', *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

McKinsey & Company (2021) *Back to the future? Airline sector after COVID-19*. Available at: https://www.mckinsey.com/ (Accessed: 4 September 2025).

Molnar, C. (2019) *Interpretable machine learning*. Munich: Leanpub.

Molnar, C. (2022) *Interpretable machine learning: A guide for making black box models explainable*. 2nd edn. Munich: Leanpub.

O'Connell, J.F. and Williams, G. (2011) *Air transport in the 21st century: Key strategic developments*. Farnham: Ashgate.

Ott, M., Choi, Y., Cardie, C. and Hancock, J.T. (2011) 'Finding deceptive opinion spam by any stretch of the imagination', *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 309–319.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1988) 'SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality', *Journal of Retailing*, 64(1), pp. 12–40.

Park, S. and Nicolau, J.L. (2015) 'Asymmetric effects of online consumer reviews', *Annals of Tourism Research*, 50, pp. 67–83.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?": Explaining the predictions of any classifier', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144.

Smith, J., Mansilla, D. and Goulding, J. (2020) 'Model class reliance in explainable AI: A comparative evaluation', *arXiv preprint*. Available at: https://arxiv.org/abs/2005.03219 (Accessed: 4 September 2025).

Sparks, B.A. and Browning, V. (2011) 'The impact of online reviews on hotel booking intentions and perception of trust', *Tourism Management*, 32(6), pp. 1310–1323.

Suau-Sanchez, P., Voltes-Dorta, A. and Cugueró-Escofet, N. (2020) 'An early assessment of the impact of COVID-19 on air transport: Just another crisis or the end of aviation as we know it?', *Journal of Transport Geography*, 86, 102749.

Vedantam, R., Lawrence Zitnick, C. and Parikh, D. (2015) 'CIDEr: Consensus-based image description evaluation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575.

Webb, G.I., Hyde, R., Cao, H., Nguyen, H.L. and Petitjean, F. (2016) 'Characterizing concept drift', *Data Mining and Knowledge Discovery*, 30(4), pp. 964–994.

Widmer, G. and Kubat, M. (1996) 'Learning in the presence of concept drift and hidden contexts', *Machine Learning*, 23(1), pp. 69–101.

Žliobaitė, I. (2017) 'Learning under concept drift: An overview', *arXiv preprint*. Available at: https://arxiv.org/abs/1010.4784 (Accessed: 4 September 2025).

Zhang, Z., Zhao, K. and Xu, S. (2016) 'Who will leave online reviews? The role of experience attributes and consumer characteristics', *Information & Management*, 53(5), pp. 715–727.