Journal of Smart Tourism Vol. 1 No. 1 (2021) 53-64

DOI: https://doi.org/10.52255/smarttourism.2021.1.1.7

Journal of Smart Tourism

ISSN: 2765-2157 Journal homepage: http://strc.khu.ac.kr/

Will You Buy It Now?: Predicting Passengers that Purchase Premium Promotions Using the PAX Model

Noora Al Emadi[a,b], Saravanan Thirumuruganathan[a], Dianne Ramirez Robillos[c], and Bernard Jim Jansen[a,d,*]

[a] *Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar*
[b] *College of Computer Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar*
[c] *School of Statistics, University of the Philippines, Diliman, Philippines*
[d] *College of Information Sciences and Technology, The Pennsylvania State University, USA*

Abstract

Upselling is often a critical factor in revenue generation for businesses in the tourism and travel industry. Utilizing passenger data from a major international airline company, we develop the PAX (Passenger, Airline, eXternal) model to predict passengers that are most likely to accept an upgrade offer from economy to premium. Formulating the problem as an extremely unbalanced, cost-sensitive, supervised binary classification, we predict if a customer will take an upgrade offer. We use a feature vector created from the historical data of 3 million passenger records from 2017 to 2019, in which passengers received approximately 635,000 upgrade offers worth more than $422,000,000 U.S. dollars. The model has an F1-score of 0.75, outperforming the airline's current rule-based approach. Findings have several practical applications, including identifying promising customers for upselling and minimizing the number of indiscriminate emails sent to customers. Accurately identifying the few customers who will react positively to upgrade offers is of paramount importance given the airline 'industry's razor-thin margins. Research results have significant real-world impacts because there is the potential to improve targeted upselling to customers in the airline and related industries.

Keywords

upselling; price elasticity; aviation company data; prediction; airline travel

1. Introduction

Leading airline companies continuously strive to increase their profits by providing the best competitive services and offers in a highly competitive vertical (Ham, Koo & Chung, 2020). One method of increasing profits in this vertical is upselling, namely convincing a customer to consume more by purchasing an upgraded premium ticket. While certainly upselling of seats is not the only revenue approach, it is one approach to revenue generation. Upselling in the airline industry is a complex decision process. There are competing objectives, such as load factors on flights, cost break-even points for upgrading, and every upgrade means a now vacant seat in the original cabin that now needs to be filled. As the drivers for customer engagement (Chiang et al., 2020) in this area are relatively unknown (Hamari et al., 2020), we investigate how customers of a major international airline company respond to discount offers for upgrading seats (e.g., economy to business, business to first). Like any company in a challenging tight-margin business, this international company is evaluating a variety of pricing and discounting strategies to alter customer behavior.

However, current approaches for upselling leverage rule-based approaches (Sarker & Kayes, 2020) that are not empirically data-driven. As such, there are open questions concerning whether or not more data-driven strategies are worthy of investigation. At least in the travel industry, it is not known what customer attributes are predictive of openness to upselling. *Can a company effectively leverage such information for upselling? Are there external factors, from either the company or the customer, that affect the upselling process?* As such, there are several potentially fruitful areas of investigation. In the research presented here, by utilizing passenger demographics and information about booking and trips, we show that one can perform sophisticated customer targeting that will enhance upselling performance. Specifically, through extensive analysis, we demonstrate that there are indeed distinct and statistically significant customer behaviors that indicate the need for customer segmentation (Oh et al., 2002).

The motivation for investigating the aspect of upselling is clear to companies to increase revenue per customer interaction. However, in particular, there are several unknowns concerning which customer to target with upselling and what attributes to employ to determine these customers. While upselling in specific industries means targeting every customer (e.g., fast food with the "Would you like fries with that?"), this blanketing technique can cause dissatisfaction and lost opportunity cost in many other industries. For example, customers getting bombarded with irrelevant email marketing messages can result in disgruntled customers. With many companies having limitations on how many promotional offers to send within a given period, each irrelevant message means a lost opportunity for a relevant one. Since a company has valuable customer information (i.e., the company knows the customers that purchased, for example), it makes sense to leverage this information for more targeted upselling offers. Therefore, in this research, we investigate the benefits of employing this customer information for upselling.

*Corresponding author:
Bernard J. Jansen, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
E-mail address: jjansen@acm.org
Received 12 January 2021; Received in revised form 21 February 2021; Accepted 10 March 2021

Our research focuses on the central questions of whether an airline company can predict passengers who are the most likely to accept premium upgrade offers, what are the most impactful predictive factors of these customers, and why are these factors are the most predictive. The problem of predicting whether customers will accept an upgrade offer is modeled as a binary (as the decision is binary for accepting the offer or declining the offer) supervised classification from a feature vector that we create from historical passenger data, including demographics and trip information. This data encompasses 3 million trip records from 2017 to 2019.

Our findings reveal that Gradient Boosting Machine outperforms all other classifiers by achieving a 0.75 F1-score (i.e., the harmonic mean of precision and recall of the model), with substantially better performance for some sectors. This promising result opens the door to improving the offer-targeting process using big data and machine learning to more effectively leverage existing customer information, which can greatly impact increasing company revenues. Our insights are in the process of being evaluated by the airline company for actual implementation in real-world A/B testing. We believe this research will be a primary motivator for a rigorous understanding of customer behavior for pricing in this company and will lead to a series of exciting work.

2. Literature Review

The airline industry is an extremely competitive industry with razor-thin profit margins and travel dependence on various factors, such as destinations (Hlee et al., 2020; Song & Lee, 2020). There have been several studies on understanding customer satisfaction based on service quality (Nicolini & Salini, 2006; Zins, 2001; Eti & Mızrak, 2020). However, these works do not explicitly consider the problem of upselling (Mayer et al., 2020), which is a notable gap in the research given the importance of upselling to businesses in the travel sector (Denizci Guillet, 2020). Ostrowski et al. (1993) showed that sufficient investment in excellent service quality influences the raises in customer loyalty in the long-term, positively impacting the revenue; however, the study was based mainly on surveys of U.S. airline travelers in 35 different U.S. airports. Clemes and Choong (2008) examined the travelers' behavior and satisfaction, surveying 428 passengers about how the aviation industry is affected by wars and epidemics in some countries and how these circumstances affect the revenue aligned with competitors.

Prior studies propose theoretical aspects of upselling and how to increase the revenue. Aydin and Ziya (2008) studied the interactions between upselling, customer purchase information usage, and dynamic pricing. They observed that if the promotional product price is dynamically balanced and the regular product is continually available when needed, customers will be offered a discount on the upsell offer. This was for customers who purchase a different product from the promotional product. Further, Johnson and Friend (2015) examined cross-selling and up-selling, which are traditional sale approaches that companies practice to increase revenue. They capture the behavioral trends towards cross-selling and up-selling and set them in a motivation-opportunity-ability (MOA) theoretical framework. Outcomes showed unique factors that are inconsistently associated with cross-selling and up-selling in predicting job satisfaction and achievement. The study performed by (Wiesman, 2006) focused on the fast-food 'business's human factor and how restaurant 'workers' performance can be optimized to increase the revenue by asking customers to ""upsize"" their meal. On the other hand, Norvell et al. (2018) observe that as up-selling made short-term revenues improvement, it affect customer satisfaction and brand loyalty oppositely. Interestingly, down-selling did not reduce short-term revenues and drove a better customer attitudinal response and brand loyalty.

An and Noh (2009) studied the effectiveness of service quality provided in the flights on loyalty and customer satisfaction, using 494 questionnaires input from travelers of business and economy cabin classes. The findings showed that service quality importance factors differ in each cabin class. Hussain et al. (2015) investigated the association among quality of service, perceived value, service provider image, customer satisfaction, loyalty, and expectations in a United Arab Emirates-based airline, again utilizing data from questionnaires. By applying structural equation modeling, the findings implied that the quality of service, recognized excellence, and airline reputation positively influenced client satisfaction that may lead to loyalty.

Various methodological aspects have been developed in upselling to increase revenues and maintain customer satisfaction. In (Kubiak & Weichbroth, 1970), the authors discussed different cross and upselling methods, particularly marketing activities automation and integration. They also discussed how to handle suitable advertisements, loyalty campaigns, and promotional offers. They discussed market basket analysis techniques to reveal the relation between offered goods and services in situations where the testing group can verify the 'analysis's outcomes before the promotional 'campaign's actual launching. Their research reports that to improve the single transaction value and sustain loyal customers, the 'offer's personalization and add-on purchases are the primary sources. Some studies have investigated the upselling price models in telecommunication operators (Hu et al., 2016; Manchanayake et al., 2019) to provide to allow customer segmentation in upselling product prices of network 'operators' offers. The results generally showed a promising approach to raise 'customers' willingness to purchase upselling products and increase network 'operators' revenue. Mayer et al. (2020) analyzed the impact of option framing and cognitive load on tourism 'services' customer choices. Irrational behavior has been noticed in that customers consume more when choosing downgrade than when they upgrade. The outcomes rendered robust evidence on downgrade framing strategy effect on upselling tourism services, and the results revealed that cognitive availability did not prevent customers from making non-logical decisions. These outcomes encourage expanding the current discussion on behavioral economics in tourism, especially in airlines.

While there have been several studies on customer satisfaction and loyalty, mainly using survey data, there is not much work on increasing aviation 'companies' revenue. There is little prior work employing actual airline customer behavior information. There is a scarcity of research on building machine learning models to predict whether the customer will accept any 'airline's upgrade offers (Aboelmaged & Mouakket, 2020; Chen et al., 2021; Huang et al., 2016; Ma et al., 2020; Renjith et al., 2020).

There are a handful of prior works that study the impact of pricing on customer behavior. In (Castillo et al., 2017; N. F. Ma et al., 2018), researchers examined the effect of surge pricing (where the base trip fare differs based on demand) on customers and drivers. This has some interesting behavior, as it varies on the cost-benefit of driving (respectively, ride-hailing) at any given time of day for a driver (respectively, a customer). Southern et al. (2017) built a mobile system to estimate and report the total cost of a driving trip that was extended for ride-sharing (Svangren et al., 2018). Understanding pricing and its impact on users is also essential in domains such as crowdsourcing. Wage distribution has been studied in platforms such as Amazon Mechanical Turk (Hanrahan et al., 2015; Hara et al., 2018). Their analysis shows how wage computations were affected by how unpaid tasks are estimated, such as searching for and working on unsubmitted tasks in the end.

Various studies have considered customer demographic information in upselling in different markets, such as (Squires et al., 2007). In (Steffen et al., 2020), the authors examined the impact of default choices on meals and hotels in travel packages. The conclusion of two investigations showed that customers traveling as couples are more likely to purchase a superior hotel room and additional meals if the higher-grade choice is provided as the default instead of the lower-grade option. The investigation reported in (Gupta, 2018) explored the impact of ebanking on the 'consumer's knowledge of cross-selling and up-selling, counting different customer demographics information. The results revealed that customer demographic information, such as age groups, annual income, profession, and gender, have a remarkable impact on the up-selling of banking products and services. In (Dai, 2014), the author modeled customer preferences on viewership behavior using latent Dirichlet analysis (LDA) by analyzing channel viewing behavior as a similar article generation process. The model performed reliable prediction performance.

Several studies also have been conducted on price discrimination in the airline industry. Studies of the most critical airline markets by (Abdella et al., 2019; Borenstein & Rose, 1994; Cui et al., 2018; Stavins, 2001) found that different airlines charged significantly different prices. They compared the carriers' prices against such factors as market structure, price dispersion levels, and the number of competitors. Luttmann (2019) examined price discrimination on a nonstop route on the trip origin and passengers' income. Researchers found that airlines used passengers' income to discriminate in round-trip ticket prices, charging higher prices where demand was low.

The prior works studied various customer satisfaction and service quality factors using data from surveys that tended to rely on convenience customer samples. Other investigations examined price discrimination across different airlines. However, prior work has not addressed any developed model for predicting the acceptance of promotional offers. A review of previous work leaves several open questions, including: *How can companies increase revenue by utilizing passengers' data? Will frequent travelers purchase a premium upgrade offer? Can likely customers for upselling be identified a priori?* These are some of the questions that motivate our research.

3. Research Questions

We focus our investigation on three central questions (R.Q.s):

- RQ1: *Can an airline company predict passengers who are the most likely to accept premium upgrade offers*?
- RQ2: *If so, what are the most impactful predictive factors of these customers?*
- RQ3: *Why are these factors the most predictive?*

These research questions are important to increase the airline company's revenue. While most customers book economy class and few book premium class, it can benefit both the airline companies and the customers if some customers are persuaded to upgrade with a discount offer via email marketing campaigns (Zhu et al., 2019). However, sending messages to all customers booking economy has several disadvantages, including opportunity cost (e.g., sending an email is not free) and customer dissatisfaction (e.g., customers with no intention of upgrading are just irritated). Therefore, it makes practical sense to target those customers only who are likely to upgrade. The research also has theoretical implications for insights into understanding customer behavior.

To address these R.Q.s, we identify the set of demographic and behavioral features that predict the travelers who are most likely to upgrade. We enrich the airline company data with external information that can influence the customer's decision (Fu et al., 2020) to upgrade. We analyze airline company passenger data over three years to investigate factors that impact the customers' decision to avail of a seat upgrade offer to build machine learning models predicting a customer's willingness to upgrade.

Figure 1 shows an overview of the research roadmap. The raw data provided by the aviation company came from multiple teams within the company and hence had several data quality issues.

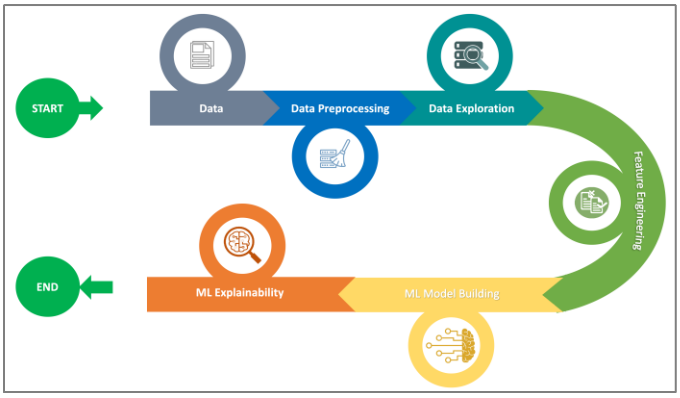

Fig. 1. Research roadmap for predicting passenger propensity to upgrade via an offer to purchase a premium ticket

We developed a complex pipeline to make the data amenable for analysis. We performed extensive data exploration to identify critical factors affecting the willingness to upgrade. Based on extensive discussions with the airline domain experts, we identified many derived attributes that are predictive. We then evaluated several machine learning classifiers using multiple evaluation metrics, an approach with promise in the online price area (Greenstein-Messica & Rokach, 2020). Given that non-experts would use the machine learning model, we also added functionality to explain the predictions.

4. Data Preparation

We describe the data preparation pipeline used for converting the raw data into a form amenable for data exploration and analysis.

4.1 Airline Company Data

The data is from one of the world's largest airline companies and consists of three million trip records of customers traveling from 2017 to 2019, as shown in Table 1.

Table 1. Dataset size from the aviation company and an overview of sent and accepted offer upgrades percentages

Year	Dataset size	Offer sent	Offer accepted
2017	1 million	16.60%	1.69%
2018	1 million	24.26%	1.34%
2019	1 million	22.63%	1.31%
Combined data	3 million	21.17%	1.42%

The dataset contains information concerning which customers were sent upgrade offers and, for those who accepted, at what price they upgraded. This set is valuable data to analyze customer dynamics in a major industry, with implications in many other domains concerned with upselling. From Table 1, we observe that between 2017 and 2019, 634,970 airline customers received upgrade offers, of which 9,042 (1.42%) accepted the upgrade.

The original dataset consists of 23 attributes that include customer demographic information and trip information. We highlight the key ones in Table 2.

Table 2. Aviation company data dictionary

Attribute Name	Data Type	Description
Nationality	Categorical	Customer's nationality code
Gender	Categorical	Customer's gender
Age	Numeric	Age of the customer
Sector_origin	Categorical	Airport code where the flight starts (example IAD)
Sector_destination	Categorical	Airport code where the flight terminates (example BOM)
Flight_date	Date	Date of the departure
Ticket_price	Numeric	Ticket transaction amount in USD
Offer_sent_status	Categorical	Offer sent status to a passenger
Oug_accept_flag	Categorical	Offer accepted flag
Offer_price	Numeric	Upgrade offer price in USD
Cabin_from	Categorical	Original cabin class booked
Cabin_upgrade_to	Categorical	Upgraded cabin class

The overall process of upselling is relatively straightforward. The customer buys a flight ticket for economy cabin class; then, the airline company checks if the customer is eligible to receive the upgrade offer email. Finally, the customer either accepts or rejects the proposed offer.

A sample of the upgrade offer email that is sent to eligible customers is shown in Figure 2. The email consists of all trip details and the offer price to upgrade from economy to the premium cabin class.

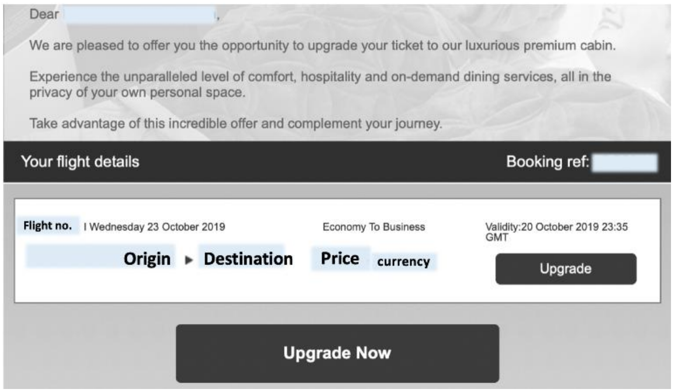

Fig. 2. Example of upgrade offer email sent to a passenger. Note: Branding, including coloring, removed. Message altered to mask the airline company.

There are also supplementary datasets related to the ticketing classes and trip distance. The company represents each cabin class with different single letters within the same cabin class. It maps these letters with actual cabin classes such as Economy, Business, First, Group, and Staff. Another supplementary dataset relates to the destination distance from/to the home-based airport. This includes the three-letter airport code, city and country codes of the airport, and distance to/from the home-based airport in miles.

4.2 Data Pre-Processing

For data cleaning, using data from the airline company, we excluded any records missing demographic information, as state-of-the-art imputation methods did not provide good results (Royston, 2004; Van Buuren, 2018). Passengers' nationalities were also inconsistently presented in two-letters code or three-letters code, which we unified using the IATA format. We also unified the cabin class attribute from a single letter code to the equivalent cabin class such as Economy, Business, and First. We partitioned passengers into age groups using the U.S. Census groupings, which are 13 years and younger, between 13–17, 18–24, 25–34, 35–44, 45–54, 55–64 years, and 65 years and older (Bodrunova, 2018).

We enriched the original dataset with several types of external information. We included a derived attribute whether the flight was on a weekday or a weekend. Another feature was whether the traveler was returning to a home country from abroad. We obtained this information by comparing the passenger's nationality with the flight's destination. We used the destination's distance dataset to enrich the original dataset with the 'flight's distance and duration (Abdollahi et al., 2020). The distance is presented as miles, and the flight duration is calculated using the average speed of (550 miles).

We also collected information about the official and school holidays of the countries of the top ten passenger nationalities and combined them with our original dataset. Given the 'industry's competitive nature, we also collected a list of competitor airlines flying to the same destinations from the same origin. Finally, we created a derived attribute percentage difference between the ticket price and offer price to test whether the difference would affect the customer's decision. Consequently, in addition to the airline company data dictionary shown in Table 2, we used eight more attributes, such as age group, day type, trip distance, trip duration, etc., as shown in Table 3.

5. Research Methodology and Experimental Results

5.1 Data Exploration

We first performed data exploration to study the impact of demographic factors on the acceptance of upgrade offers. The male-to-female breakdown is 55.6% to 44%. From Figure 3, we can see that age groups of 25–34, 35–44, and 45–54 account for almost 62.6% of the trips, whereas the age groups of people under 13, 13–17, 18–24, 55–64, and over 65 represent the remaining 37.4%.

Table 3. Additional attributes to the original data dictionary

Attribute Name	Data Type	Description
Age_group	Categorical	Age group of the passenger
Flight_day_type	Categorical	The flight day is the weekdays or weekends
Distance	Numeric	Miles Distance to/from home-based airport
Duration	Numeric	Duration in hours to/from home-based airport
P_difference	Numeric	The percentage of difference amount between the ticket price and offer price to upgrade the cabin class
To_home	Boolean	Yes, if a passenger is going to his country
Is_holiday	Boolean	Yes, if the flight date is within an official holiday
Is_competitor	Boolean	Yes, if there is a competitor flight to the same destination

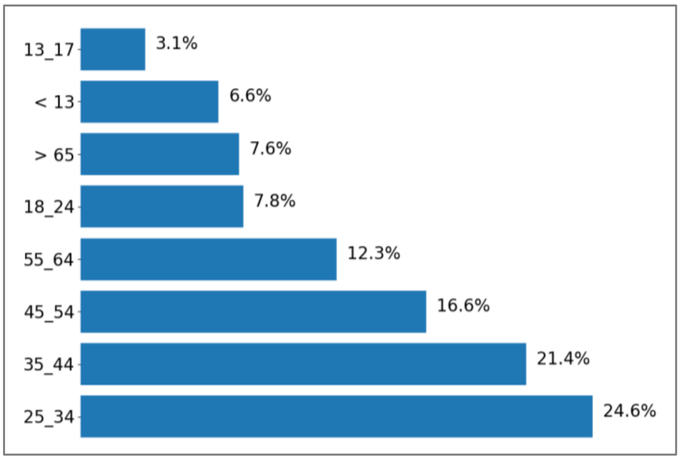

Fig. 3. Statistics of the age group population

Figure 4 shows that 50% of the passengers are flying to the home-based airport (the actual airport is hidden for the confidentiality of the company), while the other 50% is scattered across 165 destinations. This is not surprising due to the 'airline's hub and spoke model (Bryan & O'Kelly, 1999).

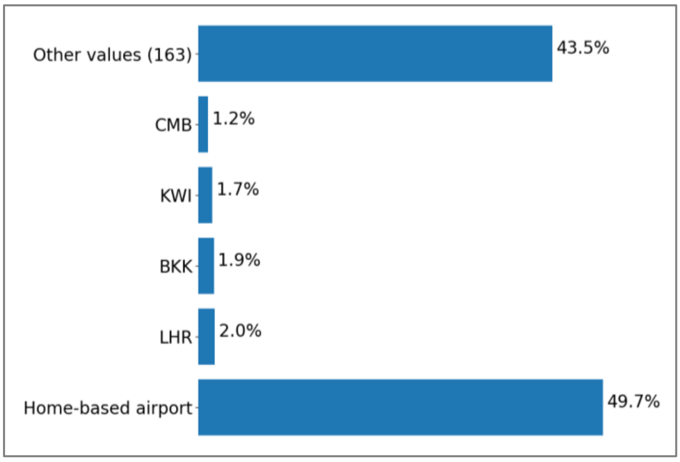

Fig. 4. Destinations overview. Home-based airport is hidden for the confidentiality of the company

Correlations are an effective method for exploring the relationship between two variables. In Figure 5, we applied Pearson's correlation coefficient on the selected attributes to measure the linear correlation between variables (Benesty et al., 2009; Havlicek & Peterson, 1976). We find that there is a strong linear relationship (0.7) between the offer price, trip distance, and trip duration. Although the acceptance ratio for upgrade offers is minimal, the correlation between the offer being sent and being accepted and the amount of difference between the ticket price and the upgrade offer price is 0.67. This means that the higher the difference between the ticket price and the upgrade offer price, the more likely that offer will be accepted.

Of the upgrade offers sent, 43.27% of upgrade offers were sent to females, while 56.73% were sent to males. Of the customers who were contacted, 62.3% of those who accepted the offer were male, and 37.7% were female. An exact binomial test was run to determine whether a greater proportion of male customers were more willing to purchase the upgrade offer compared to females, assuming the percentages were equal; the results indicated a significant difference ($q=0.95$, $p < 0.01$). These results suggest that gender affects the willingness to accept an upgrade offer. Specifically, our results indicate that males are more willing to accept offers than females. This shows a crucial missed opportunity for upselling and could motivate a data-driven approach for more complex models as the rule-based approach obviously missing opportunities.

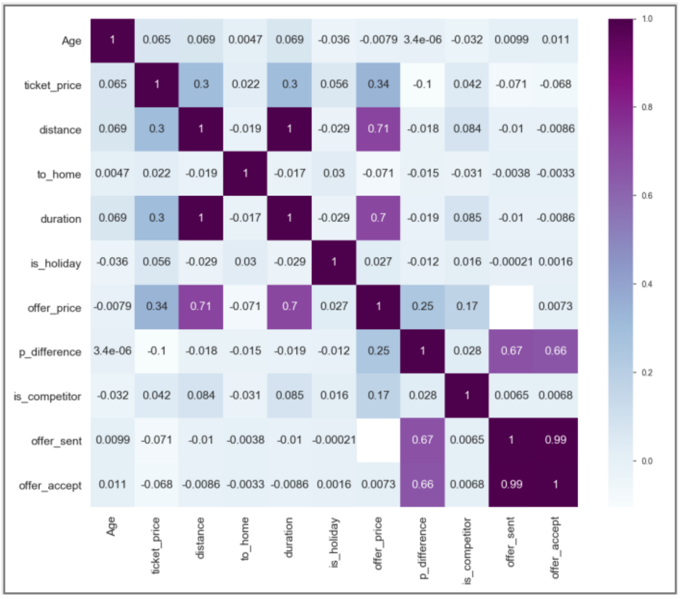

Fig. 5. Pearson's correlation coefficient for attribute pairs

In Figure 6, we show the distribution by age group of those receiving upgrade offers. We find that people in the age groups 25–34 and 35–44 received more than 50% of the upgrade offers, whereas those in the age groups 55–64 and 45–54 received 16.3% and 12%, respectively. We can see that members of some age groups accept offers at a higher rate, which highlights a potential opportunity. A chi-square test of independence was performed to examine the relationship between age group and accepting the upgrade offer, which proved to be significant, with $X2(7, N = 634970) = 363816.45$, $p = 0.00087$. These results suggest that age group has an effect on the willingness to accept an upgrade offer. Specifically, our results indicate that the age groups 25–34, 35–44, and 45–54 are more willing or able to accept offers than the other groups.

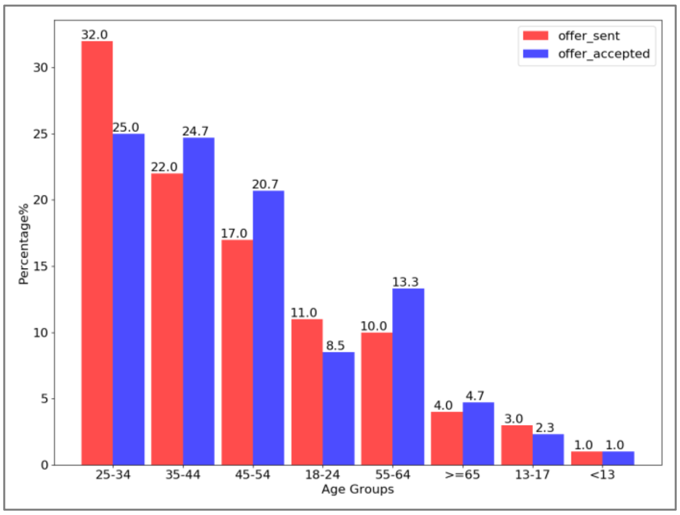

Fig. 6. Percentage of sent and accepted offer upgrades to each age-group

We examine trip destinations also. In Figure 7a, we show the percentage of upgrade offers for the top 50 destinations. The travelers flying to the airline's home-based airport received the highest number of upgrade offers (49.4%), whereas the rest of the upgrade offers were split among the remaining 49 airports. The higher offer rate is to be expected due to the Hub-and-Spoke model used by most major airlines. Surprisingly, as shown in Figure 7b, the travelers to the airline's home-based airport accepted upgrade offers 46% of the time, followed by LHR (London Heathrow) airport at a distant second rate of 3%. A chi-square test of independence examined the relationship between the destination and accepting the upgrade offer, which proved to be significant: $X2(49, N = 634970) = 27569.9$, $p < 0.000017$. These results suggest that the destination affects the willingness to accept an upgrade offer. Specifically, our results indicate that passengers traveling to the home-based airport are more willing to accept offers.

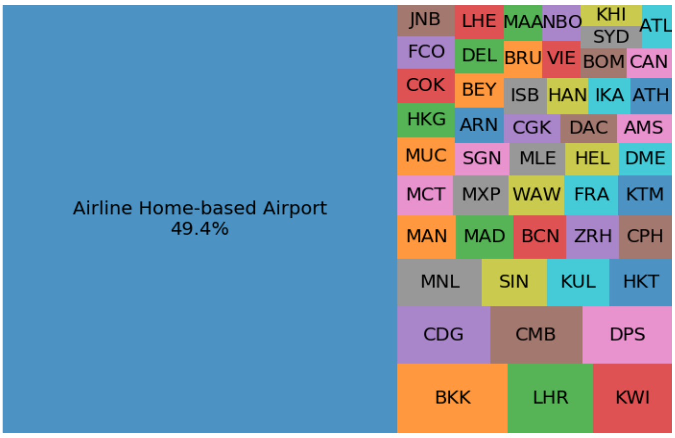

Fig. 7a. Percentage of upgrade offers sent for destinations. Home-based airport is hidden for the confidentiality of the company

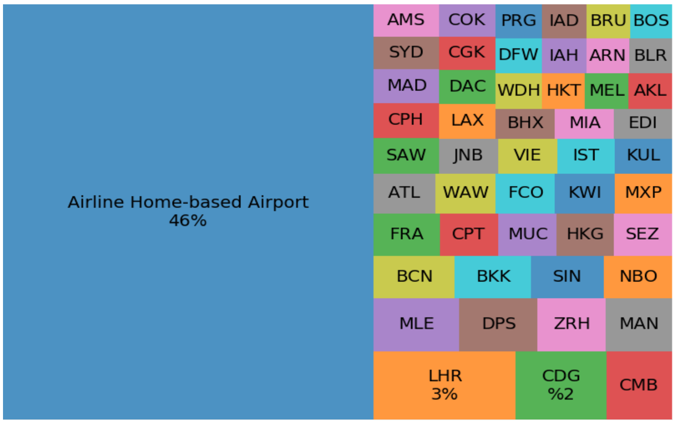

Fig. 7b. Percentage of upgrade offers accepted in destinations. Home-based airport is hidden for the confidentiality of the company

Our analysis shows several intriguing results. For example, men are, perhaps, less price-conscious than women, who are willing to upgrade more. The age groups 25–34, 35–44, and 45–54 exhibit a greater propensity to upgrade than other age groups. We observed that the travelers toward the airline's home-base airport receive and accept upgrade offers more than other airports. This exploratory analysis confirmed our premise that empirical data analysis would be a fruitful avenue for investigation.

5.2 Model Formulation

We formulate the problem as a binary supervised learning classification problem (Nadeau & Turney, 2005). We were given the features of customers and trips; we predict whether the customer would upgrade or not. This is a challenging problem due to the 'data's imbalanced nature (Yang et al., 2020), with approximately 1% of the contacted customers upgrading. We evaluated multiple classification techniques (Duda et al., 2012), such as Logistic Regression (L.R.) (Cheng & Hüllermeier, 2009; Kleinbaum et al., 2002; Murphy, 2012), Gradient Boosting Machine (GBM) (Friedman, 2001), Decision Tree (D.T.) (Safavian & Landgrebe, 1991; Stein et al., 2005; Steinberg, 2009), and Random Forest (R.F.) (Breiman, 2001; Opitz & Maclin, 1999). We evaluated the performance using the F1-score and the AUC-ROC curve. We downsampled the majority of the class, which is the not-upgraded class. We applied cross-validation with 5 K-fold. We tuned the hyperparameters to obtain higher and more precise predictions in terms of accuracy scores for the different classifiers.

We create a feature vector set that combines customer-based information and trip-based information using different approaches that are labeled collectively as the PAX Model:

- **Passenger (P)** – Passenger-based information: age, gender, nationality, age group
- **Airline (A)** – Trip-based information: flight date, flight day type, origin sector, destination sector, is the passenger traveling back to home, trip duration, trip distance, from which cabin the passenger upgrades, ticket price, offer price, percentage of the difference amount between the ticket price and offer price to upgrade.
- **eXternal (X)** – External information: is there any competitor flight; is the flight date within an official holiday?

The process of data transformation is implemented on categorical attributes using the One Hot Encoding technique (Potdar et al., 2017) by converting all categorical attributes into numbers in order to provide it to machine learning models to obtain better performance. One Hot Encoding transfers each categorical variable into a column that indicates the presence of this variable index as 1, and absence as 0. One Hot Encoding is an essential process in data preparation if there are one or more categorical attributes. Few machine learning algorithms can handle categorical attributes, but many other models cannot work on categorical data directly. These algorithms expect all input variables and output variables to be numeric.

5.3 Model and Analysis Results

In the predictive model building process, we set the baseline model as Logistic Regression in this trial as it is a widely used technique for statistical modeling. Table 4 shows all the results from the manual feature vector set of the different classifiers.

Table 4. Results of different ML classifiers of a manual feature vector set. Logistic regression is the baseline.

Model Name	F1-score
Logistic Regression	0.68
Decision Tree	0.73
Random Forest	0.72
Gradient Boosting Machine	**0.75**

Logistic Regression (Cheng & Hüllermeier, 2009; Kleinbaum et al., 2002; Murphy, 2012) is a supervised machine learning algorithm and a predictive analysis algorithm based on a probability concept.

Decision Tree forms the classification as a tree structure and divides a large dataset into smaller subsets to get easier and efficient classification results in the form of decision and leaf nodes (Murphy, 2012; Safavian & Landgrebe, 1991; Stein et al., 2005; Steinberg, 2009).

Random Forest attempts to de-correlate the bottom learners by learning trees based on a randomly determined group of input variables and based on a group of data samples that is chosen randomly. Random forest often obtains good prediction accuracy, is a popular ensemble learning method (Opitz & Maclin, 1999), and is widely used in several domains (Breiman, 2001; Murphy, 2012).

Gradient Boosting Machine (Friedman, 2001) is also known as MART (Multiple Additive Regression Trees). Gradient Boosting Machine develops an onward stage-wise additive model by applying gradient descent in function space that minimizes the overall prediction error.

Applying the Gradient Boosting Machine classifier achieved the highest performance in terms of accuracy score, precision, recall, and F1-score, which obtained 0.75 for the mentioned evaluation metrics followed by Decision Tree classifier that obtains 0.73 accuracy score, precision, recall, and F1-score. On the other hand, the Random Forest classifier achieved 0.72 in accuracy, precision, recall, and F1-scores, whereas the logistic regression classifier obtained 0.68 accuracy, precision, recall, and F1-scores. Obtaining an 0.75 F1-score by Gradient Boosting Machine classifier is an outstanding achievement to be used in the airline company to predict the 'traveler's willingness to upgrade. As mentioned earlier, the prior work has not addressed any developed model for predicting the acceptance of promotional offers. Our model can be used as a filter allowing a more targeted approach in sending offers to travelers.

To show the misclassification of the predictions that occur in each label, we also report the confusion matrices of each classifier in Figure 8. Each row of the confusion matrix corresponds to a not-upgraded class and an upgraded class, representing the prediction results of each class; each column corresponds to the real class.

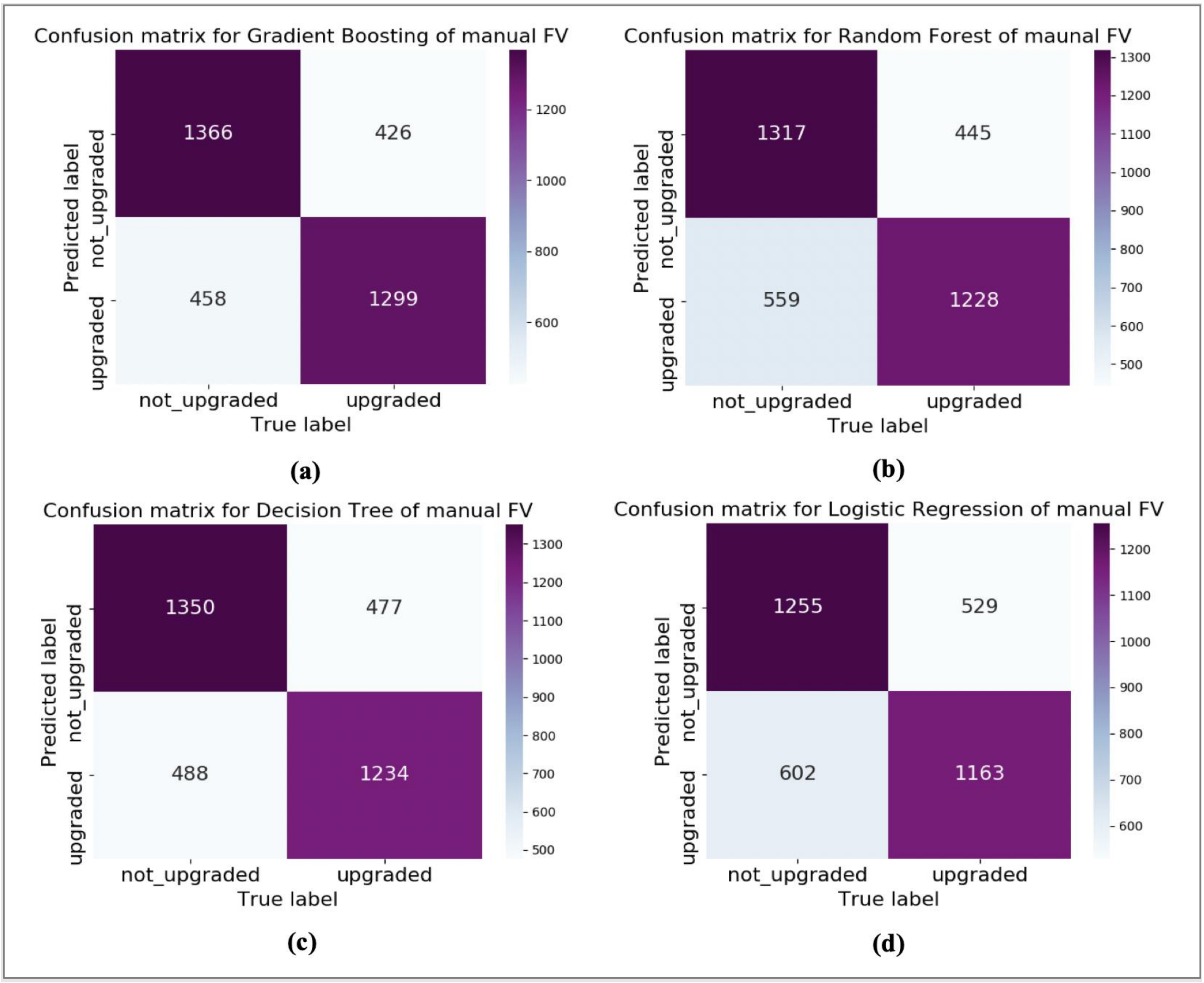

Fig. 8. Confusion matrices for different classifiers built over the manually constructed feature vector

Area Under the Curve - Receiver Operating Characteristics Curve (AUC ROC Curve) is one of the primary performance measurements for binary classification problems (Bradley, 1997; Hanley & McNeil, 1982). In Figure 9, we observe the AUC of the Gradient Boosting Machine classifier performance, which obtained a 0.83. The corresponding scores for Random Forest, Decision Tree, and Logistic Regression are 0.79, 0.78, and 0.76, respectively.

5.4 Machine Learning Explainability

Machine learning explainability and interpretability are effective approaches for extracting human-understandable insights from machine learning models and reporting the prediction performance in more straightforward and logical terms. We utilize InterpretML (Nori et al., n.d.), which can assist in explaining BlackBox methods. To explain individual predictions, we use LIME (Ribeiro et al., 2016). In the images below, the negative features are presented in blue bars, which indicate a not-upgraded class. The positive features are shown in orange bars, which indicate upgraded class.

In Figure 10, we observe an example of an individual prediction of the not-upgraded class by Explainable Boosting Machine (Caruana et al., 2015; Lengerich et al., 2020; Lou et al., 2013, 2012; Tan et al., 2018), with nationality and price data mainly impacting the decision.

Figure 11 also shows that demographic and price information are the key factors impacting the correct prediction decision.

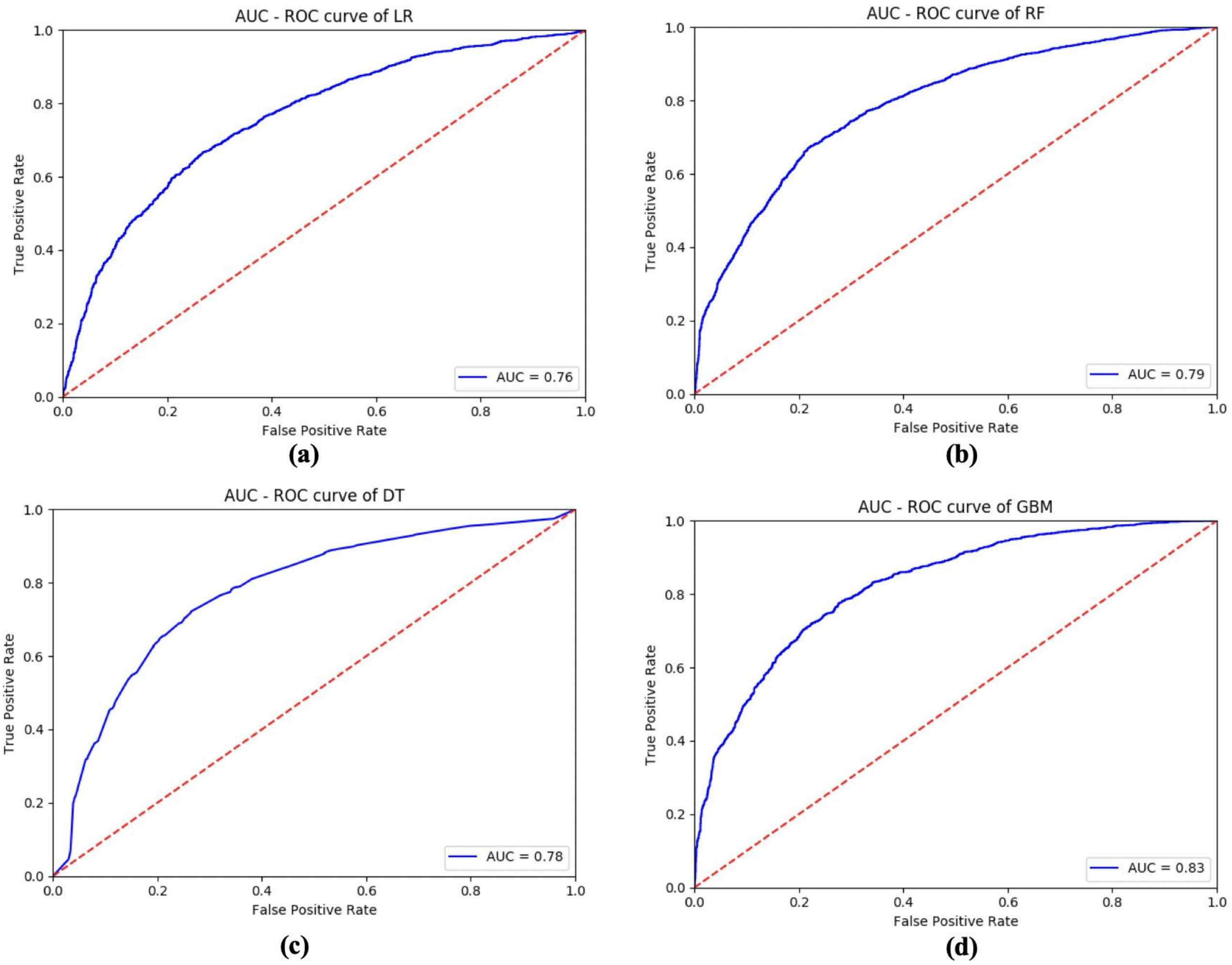

Fig. 9. AUC - ROC curve of the employed classifiers

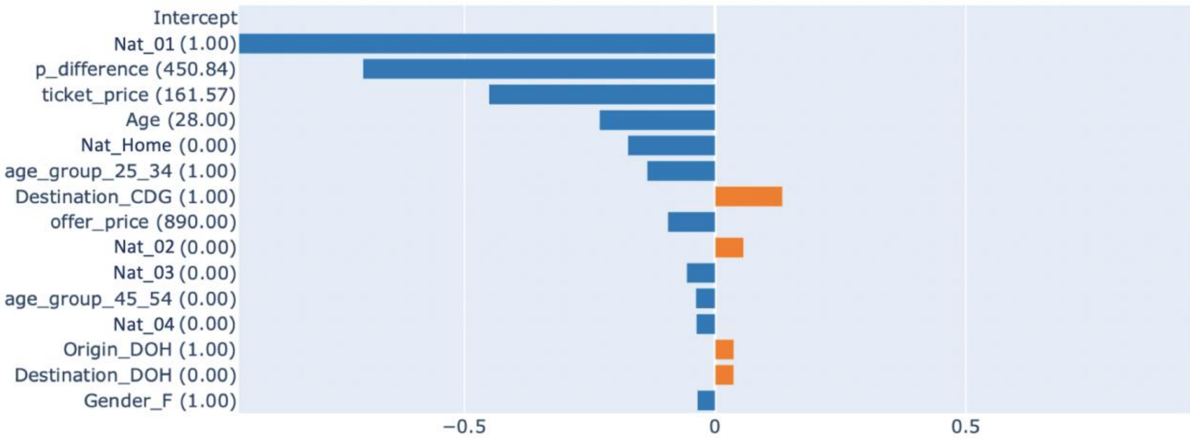

Fig. 10. Example of individual prediction of the not-upgraded class by Explainable Boosting Machine.
(Nationalities are hidden for the confidentiality of the company.)

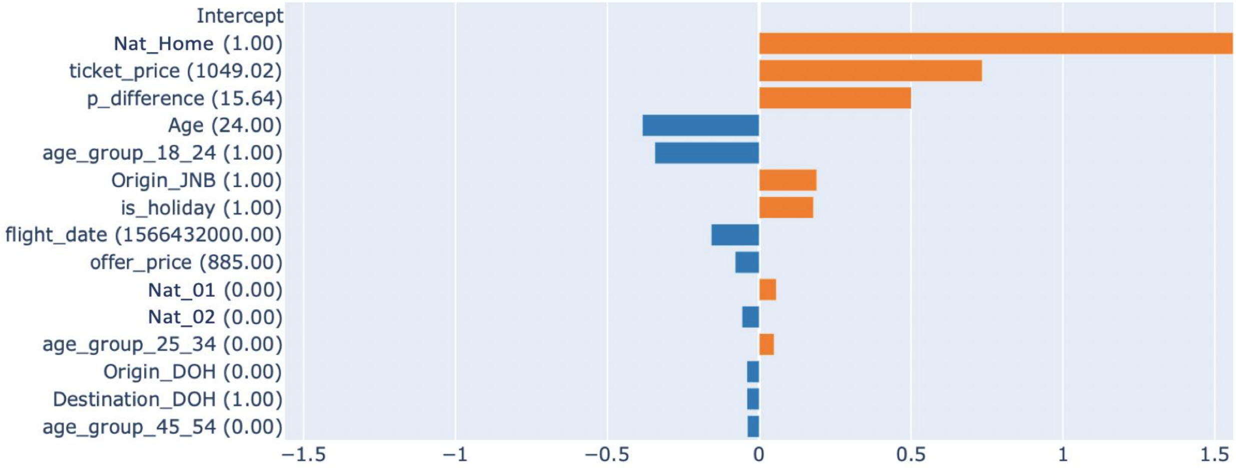

Fig. 11. Example of individual prediction of the upgraded class by Explainable Boosting Machine. (Nationalities are hidden for the confidentiality of the company.)

Next, we analyze the most important factors at a global level by analyzing the model sensitivity. We identify the essential attributes by sorting them in decreasing order of Mean Absolute Score. This score is computed to get the features having the most significant impact on the prediction by getting the sorted average of each 'feature's absolute predicted value in the training dataset. Figure 12 shows that *price_difference* and *ticket_price* are the essential features in the predictions; these are followed by nationality, customer age, flight date, and age group. We conclude that both demographics and price information are valuable data features in predicting the willingness to accept the upgrade offers.

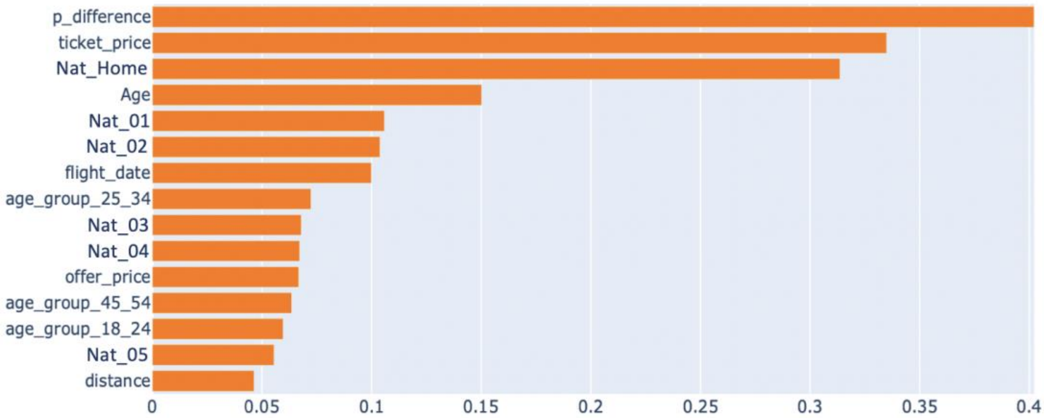

Fig. 12. Overall importance of features in the prediction. (Nationalities are hidden for the confidentiality of the company.)

6. Discussion and Implications

Our analysis shows that there are several non-obvious relationships among PAX factors impacting a customer's upgrade behavior, which has implications in a range of airline-industry-related domains and situations.

As discussed in the literature review, there is limited prior work employing actual airline customer behavior information. Specifically, there is a scarcity of research on building machine learning models to predict whether the customer will accept any upgrade offers provided by the airline. We found only a handful of prior works that study the impact of pricing on customer behavior and demographic aspects. There are modest attempts to build pricing models from previous studies, but there is no such attempt made based on machine learning to predict the willingness to upgrade in the airline-industry-related domains.

The practical implication of these findings is beneficial in two critical ways:

- **Identifying Non-Responders and Responders**: Several customer segments have a marked behavior of under-responding. This implies the need for more refined targeting of these customer segments lest the organization spam them with too many unwanted upgrade emails. Conversely, there are customer segments that eagerly embrace the upgrade offers.

- **Balancing Offers with Accepts:** When the customer segment ratio does not match the upgrade offers, we conclude that the airline company is missing an opportunity in targeted upgrade offers to increase their revenue. For example, the airline company sends a disproportionate amount of upgrade offers to segments that historically do not upgrade often. However, there are customer segments that have a history of upgrading with a higher proportion. By sending offers to customer segments that mostly will not accept, the company is losing opportunity costs, as the messages could be used to upsell other services possibly. Conversely, the company is losing upgrading opportunities by not sending offers to the customer segments most likely to accept.

Our proposed model could be used as a filter to prevent sending offers to customers who are likely not to upgrade or identify customers who are likely to upgrade, allowing a more targeted approach. Such an approach would ensure that the otherwise vacant premium-class seats are filled, thereby increasing revenue and increasing customer satisfaction. However, these changes must be performed with more care. Naively sending more upgrade offers to certain customer segments could result in potential cannibalization, where the customers learn to game the system. Specifically, if they figure out that they will get an upgrade offer, they might wait instead of directly buying the business class. This results in an intriguing new dynamic that we plan to investigate more in future research.

We also observed that tackling a real-world data science problem such as this is quite challenging. The data is noisy, needing an extensive pipeline with many data-processing steps of data cleaning, unifying attributes, and enriching the data. A naïve application of a classifier on the raw data provided suboptimal results. There is also the issue of acceptance of the machine learning approach. The use of model explainability was essential to get buy-in from airline domain experts. As such, the airline company is in the process of evaluating and integrating our project into their workflow.

7. Conclusion

In this research, we tackle the problem of predicting the willingness to accept upgrade offers in the aviation company to increase revenue and profit. We leveraged aviation company data from 2017 to 2019 to create different feature vector sets, and we developed the PAX model using binary classification. The Gradient Boosting Machine achieved the highest accuracy score of 0.75. Our proposed model could be operated as a filter to avoid sending offers to customers who are likely not to upgrade or identify customers who are likely to upgrade, allowing a more targeted approach. Such an approach would guarantee that the otherwise vacant premium-class seats are filled, thereby increasing revenue and increasing customer satisfaction.

In this study, we define some limitations that we encountered. These limitations revolve around the information we obtain from the airline company and the related information. For example, the marital status and average income of the traveler, official holidays in their home country, flights heading to the same destinations from competing companies in the same period, and ticket prices of premium class seats at those competitor companies. We also focused on this study only on particular segments of travelers who receive and accept the upgrade offers to study their behavior and neglected the other segments, which might show exciting more findings.

For future work, we propose investigating more derived features related to the demographic and trip information and external and hidden factors that might play a significant role in enhancing the prediction model. We want to investigate more machine explainability algorithms regarding individual predictions, such as SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to study the 'attributes' prioritization. A fruitful research avenue would be finding the optimal offer price, since the pricing attributes are essential predicting features. Finally, we could explore other customer data, such as panel data, for additional upselling insights.

Declaration of competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

This publication was made possible by GSRA grant, I.D. # GSRA5-2-0602-18131, from the Qatar National Research Fund (a member of Qatar Foundation).

References

Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F. (2019). Airline ticket price and demand prediction: A survey. *Journal of King Saud University - Computer and Information Sciences*.

Abdollahi, M., Khaleghi, T., & Yang, K. (2020). An integrated feature learning approach using deep learning for travel time prediction. *Expert Systems with Applications*, *139*, 112864.

Aboelmaged, M., & Mouakket, S. (2020). Influencing models and determinants in big data analytics research: A bibliometric analysis. *Information Processing & Management*, *57*(4), 102234.

An, M., & Noh, Y. (2009). Airline customer satisfaction and loyalty: Impact of in-flight service quality. *Service Business*, *3*(3), 293–307.

Aydin, G., & Ziya, S. (2008). Pricing promotional products under upselling. *Manufacturing & Service Operations Management*, *10*(3), 360–376.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In I. Cohen, Y. Huang, J. Chen, & J. Benesty (Eds.), *Noise reduction in speech processing* (pp. 1–4). Berlin: Springer.

Bodrunova, S. S. (2018). *Internet science.* Proceedings of the 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018. Springer.

Borenstein, S., & Rose, N. L. (1994). Competition and price dispersion in the U.S. airline industry. *Journal of Political Economy*, *102*(4), 653–683.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bryan, D. L., & O'Kelly, M. E. (1999). Hub-and-spoke networks in air transportation: An analytical review. *Journal of Regional Science*, *39*(2), 275–295.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission*. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730.

Castillo, J. C., Knoepfle, D., & Weyl, G. (2017). *Surge pricing solves the wild goose chase*. Proceedings of the 2017 ACM Conference on Economics and Computation, 241–242.

Chen, Y. L., Yeh, Y. H., & Ma, M. R. (2021). A movie recommendation method based on users' positive and negative profiles. *Information Processing & Management*, *58*(3), 102531.

Cheng, W., & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, *76*(2), 211–225.

Chiang, C. T., Yang, M. H., Koo, T. L., & Liao, C. H. (2020). What drives customer engagement behavior? The impact of user participation from a sociotechnical perspective. *Journal of Electronic Commerce Research*, *21*(3), 197–214.

Clemes, M. D., & Choong, M. (2008). An empirical analysis of customer satisfaction in international air travel. *Innovative Marketing*, *4*(2), 15.

Cui, Y., Orhun, A. Y., & Duenyas, I. (2018). How price dispersion changes when upgrades are introduced: Theory and empirical evidence from the airline industry. *Management Science*, *65*(8), 3835–3852.

Dai, B. T. (2014). *How can consumer preferences be leveraged for targeted upselling in cable T.V. services? Pacific Telecommunications Council (PTC'14)*. Singapore: Research Collection School of Information Systems.

Denizci Guillet, B. (2020). Online upselling: Moving beyond offline upselling in the hotel industry. *International Journal of Hospitality Management*, *84*, 102322.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. Hoboken: John Wiley & Sons.

Eti, S., & Mızrak, F. (2020). Analysing customer satisfaction of civil aviation companies of Turkey with text mining. In H. Dincer & S. Yüksel (Eds.), *Strategic outlook for innovative work behaviours: Interdisciplinary and multidimensional perspectives* (pp. 21–41). New York: Springer International Publishing.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Fu, H., Manogaran, G., Wu, K., Cao, M., Jiang, S., & Yang, A. (2020). Intelligent decision-making of online shopping behavior based on internet of things. *International Journal of Information Management*, *50*, 515–525.

Greenstein-Messica, A., & Rokach, L. (2020). Machine learning and operation research-based method for promotion optimization of products with no price elasticity history. *Electronic Commerce Research and Applications*, *40*, 100914.

Gupta, N. (2018). Influence of demographic variables: Customers' perception about cross-selling and up-selling for eBanking. *International Journal of Electronic Customer Relationship Management*, *11*(2), 126–141.

Ham, J., Koo, C., & Chung, N. (2020). Configurational patterns of competitive advantage factors for smart tourism: An equifinality perspective. *Current Issues in Tourism*, *23*(9), 1066–1072.

Hamari, J., Hanner, N., & Koivisto, J. (2020). "Why pay premium in freemium services?" A study on perceived value, continued use and purchase intentions in free-to-play games. *International Journal of Information Management*, *51*, 102040.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hanrahan, B. V., Willamowski, J. K., Swaminathan, S., & Martin, D. B. (2015). *TurkBench: Rendering the market for Turkers*. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1613–1616.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). *A data-driven analysis of workers' earnings on Amazon Mechanical Turk*. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14.

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. *Perceptual and Motor Skills*, *43*(3_suppl), 1319–1334.

Hlee, S., Lee, H., Koo, C., & Chung, N. (2020). Will the relevance of review language and destination attractions be helpful? A data-driven approach. *Journal of Vacation Marketing*, *27*(1), 61–81.

Hu, W. H., Lin, B. T., Lu, F. S., & Jeng, J. Y. (2016). *An upselling pricing model using SDP-based rating mechanism with dynamic weight*. Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016, 1–6.

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). *Learning deep representation for imbalanced classification*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5375–5384.

Hussain, R., Al Nasser, A., & Hussain, Y. K. (2015). Service quality and customer satisfaction of a UAE-based airline: An empirical investigation. *Journal of Air Transport Management*, *42*, 167–175.

Johnson, J. S., & Friend, S. B. (2015). Contingent cross-selling and up-selling relationships with performance and job satisfaction: An MOA-theoretic examination. *Journal of Personal Selling & Sales Management*, *35*(1), 51–71.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Berlin: Springer.

Kubiak, B. F., & Weichbroth, P. (1970). Cross and up-selling techniques in e-commerce activities. *The Journal of Internet Banking and Commerce*, *15*(3), 1–7.

Lengerich, B., Tan, S., Chang, C. H., Hooker, G., & Caruana, R. (2020). *Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models*. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 2020, Palermo, Italy, 2402–2412.

Lou, Y., Caruana, R., & Gehrke, J. (2012). *Intelligible models for classification and regression*. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 150–158.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). *Accurate intelligible models with pairwise interactions*. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 623–631.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Red Hook, NY: Curran Associates, Inc.

Luttmann, A. (2019). Evidence of directional price discrimination in the U.S. airline industry. *International Journal of Industrial Organization*, *62*, 291–329.

Ma, N. F., Yuan, C. W., Ghafurian, M., & Hanrahan, B. V. (2018). *Using stakeholder theory to examine drivers' stake in Uber*. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–12.

Ma, Y., Mao, J., Ba, Z., & Li, G. (2020). Location recommendation by combining geographical, categorical, and social preferences with location popularity. *Information Processing & Management*, *57*(4), 102251.

Manchanayake, S. M. A. M., Samarasinghe, D. P., Perera, L. P. J., Bandara, H. M. M. T., Kumaradasa, K. C., Premadasa, N., & Samarasinghe, A. P. (2019). *Potential upselling customer prediction through user behavior analysis based on CDR data*. 2019 14th Conference on Industrial and Information Systems (ICIIS), 46–51.

Mayer, V. F., Santos, G. E. de O., & Marques, O. R. B. (2020). Option framing for upselling tourism services: Does cognitive availability prevent irrational choices? *Tourism Economics*.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Nadeau, D., & Turney, P. D. (2005). A supervised learning approach to acronym identification. In B. Kégl & G. Lapalme (Eds.), *Advances in artificial intelligence* (pp. 319–329). Berlin: Springer.

Nicolini, G., & Salini, S. (2006). Customer satisfaction in the airline industry: The case of British Airways. *Quality and Reliability Engineering International*, *22*(5), 581–589.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (n.d.). *InterpretML: A unified framework for machine learning interpretability*. Ithaca, NY: Cornell University.

Norvell, T., Kumar, P., & Contractor, S. (2018). Assessing the customer-based impact of up-selling versus down-selling. *Cornell Hospitality Quarterly*, *59*(3), 215–227.

Oh, H., Parks, S. C., & Demicco, F. J. (2002). Age- and gender-based market segmentation. *International Journal of Hospitality & Tourism Administration*, *3*(1), 1–20.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Ostrowski, P. L., O'Brien, T. V., & Gordon, G. L. (1993). Service quality and customer loyalty in the commercial airline industry. *Journal of Travel Research*, *32*(2), 16–24.

Potdar, K., S., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, *175*(4), 7–9.

Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing & Management*, *57*(1), 102078.

Ribeiro, M. T., Singh, S., & 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, 1135–1144.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227–241.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674.

Sarker, I. H., & Kayes, A. S. M. (2020). ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, *168*, 102762.

Song, W. K., & Lee, H. C. (2020). An analysis of traveler need for and willingness to purchase airline dynamic packaging: A Korean case study. *Journal of Air Transport Management*, *82*, 101735.

Southern, C., Cheng, Y., Zhang, C., & Abowd, G. D. (2017). *Understanding the cost of driving trips*. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 430–434.

Squires, J., Wilder, D. A., Fixsen, A., Hess, E., Rost, K., Curran, R., & Zonneveld, K. (2007). The effects of task clarification, visual prompts, and graphic feedback on customer greeting and up-selling in a restaurant. *Journal of Organizational Behavior Management*, *27*(3), 1–13.

Stavins, J. (2001). Price discrimination in the airline market: The effect of market concentration. *Review of Economics and Statistics*, *83*(1), 200–202.

Steffen, A., Weibel, C., Stämpfli, A. E., & von Arx, W. (2020). Upselling by default: The effect of default options on travelers' board and lodging choices. *Journal of Travel Research*, *59*(7), 1253–1267.

Stein, G., Chen, B., Wu, A. S., & Hua, K. A. (2005). *Decision tree classifier for network intrusion detection with GA-based feature selection*. Proceedings of the 43rd Annual Southeast Regional Conference - Volume 2, 136–141.

Steinberg, D. (2009). CART: Classification and regression trees. In *The top ten algorithms in data mining* (pp. 193–216). Boca Raton, FL: Chapman and Hall/CRC.

Svangren, M. K., Skov, M. B., & Kjeldskov, J. (2018). *Passenger trip planning using ride-sharing services*. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). *Distill-and-compare: Auditing black-box models using transparent model distillation*. Proceedings of the 2018 AAAI/ACM Conference on A.I., Ethics, and Society, 303–310.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

Wiesman, D. W. (2006). The effects of performance feedback and social reinforcement on up-selling at fast-food restaurants. *Journal of Organizational Behavior Management*, *26*(4), 1–18.

Yang, S., Wang, H., Zhang, C., & Gao, Y. (2020). Contextual bandits with hidden features to online recommendation via sparse interactions. *IEEE Intelligent Systems*, *35*(5), 62–72.
Zhu, G., Wu, Z., Wang, Y., Cao, S., & Cao, J. (2019). Online purchase decisions for tourism e-commerce. *Electronic Commerce Research and Applications*, *38*, 100887.
Zins, A. H. (2001). Relative attitudes and commitment in customer loyalty models: Some experiences in the commercial airline industry. *International Journal of Service Industry Management*, *12*(3), 269–294.

Author Biographies

Noora Al Emadi is a Senior Software Engineer in the social computing group at the Qatar Computing Research Institute. She received her bachelor in computer science from Qatar University and a Master of Science in Data Science and Engineering from Hamad Bin Khalifa University. Her research interests span social and urban computing, applied machine learning, and spatial-temporal data analysis.

Saravanan Thirumuruganathan is a scientist in the Data Analytics group of QCRI, HBKU. He earned his PhD from the University of Texas at Arlington. He is broadly interested in data integration and machine learning for data management.

Dianne Ramirez Robillos is a travel industry professional with a M.S. degree in statistics from the School of Statistics, University of the Philippines, Diliman, Philippines.

Bernard Jim Jansen is a Principal Scientist in the social computing group of the Qatar Computing Research Institute, and Professor at the College of Science and Engineering, Hamad bin Khalifa University. He is the Editor-in-Chief of the Information Processing & Management (Elsevier), and the former Editor-in-Chief of Internet Research (Emerald). He is also an adjunct professor with the College of Information Sciences and Technology at The Pennsylvania State University.