Prediction of airline customer satisfaction status

Agughalam Davis Munachimso
School of computing
National College of Ireland
Dublin, Ireland
agughalamdavis@yahoo.com

***Abstract*—Passenger satisfaction is a very important concept for airlines as they compete to acquire and retain customers. Predictive analytics using machine learning techniques have been employed to predict customer satisfaction so palliative measures can be taken to appease dissatisfied customers and keep their business. The random forest algorithm is an effective algorithm to apply as it accurately predicts the satisfaction status of the customers and also indicates the importance of each of the considered factors. This work successfully employs this technique to predict customer satisfaction for an airline service with an accuracy score of 96.19%. The results also indicate that different factors drive customer satisfaction across different flight classes.**

***Index Terms*—predictive analytics, machine learning, customer satisfaction**

I. Introduction

Customer satisfaction is a very important part of any business as the profitability of the business generally depends on being able to acquire and retain customers. Owing to this, businesses have largely explored means of improving customer satisfaction. This is more the scenario for airlines as the competition to acquire loyal customers is on the increase hence, customer dissatisfaction has to be avoided or at least remedied through offers and discounts.

In recent years, the retail industry has greatly adopted predictive analytics using machine learning techniques to predict potential customer satisfaction or dissatisfaction. This strategy has proven to increase customer satisfaction and subsequent customer retention as dissatisfaction cases are noticed before time and acted upon. Meinzer et. al. [1] leveraged machine learning in the automobile sales to predict customer satisfaction using a random forest classifier. Though the approach was successful, it failed to quantitatively measure the impact of each of the features on customer satisfaction. In order to resolve this, Liu et. al.[2] proposed a local explanation approach to measure the impact of the factors affecting patient satisfaction from patient survey data. The airline industry can leverage this approach to model and gain insights into the factors affecting customer satisfaction status. This drives a customer-centric approach to business as analysis results can be used to improve aspects of operations, customize and optimize services leading to improved profit margins.

Based on this, this work aims to predict customer satisfaction for an airline by modelling contributing factors such as services and quantifying the impact of each of these factors to indicate the most important factors driving passenger satisfaction. This work also carries out an investigation into the features affecting passenger satisfaction for the different flight classes

II. Related work

Machine learning techniques are becoming increasingly popular for predictive analysis across different aspects of a business. Predicting customer satisfaction is one of such business aspects. Meinzer et. al. [1] employed machine learning techniques such as Adaboost, K-Nearest neighbours (KNN), support vector machines (SVM) and random forest (RF) to predict customer satisfaction for an automobile business. Warranty, dealer system, diagnostics and overall vehicle wellness data were used as features for the analysis. The SVM radial basis kernel performed better than the other algorithms with an accuracy value of 88.8% followed closely by the random forest algorithm with 81.5%. Though successful, This work did not elaborate on feature importance due to the use of 'blackbox' models. Looking to measure the impact of features on patient satisfaction, Liu et. al. [2] used a local explanation algorithm. A random forest classifier was used for prediction of patient satisfaction status from survey data. The research was focused on the unhappy patients to pick out the most important factors causing dissatisfaction. Using this approach, they were able to successfully predict patient satisfaction and measure the impact of each of the considered features. Similar research done by Singh et. al. [3] aimed to identify determinants of passenger satisfaction at an airport. They analysed survey data collected across different passenger demographics on their satisfaction levels for different aspects of the airport services. A random forest algorithm was used for the classification and an accuracy score of 75% was achieved. The features used in the analysis were also rated using a backward selection process and it was found out that the food, artworks, correct signage and cleanliness were the top factors affecting customer satisfaction. Sankaranarayanan et. al.[4] in their research linked the on-time performance of airlines at an airport to passenger satisfaction. They modelled factors such as airport on time performance, on time ranking of the airport and average delays amongst others to predict the level of passenger satisfaction using logistic model trees (LMT). Their model had an accuracy value of 80%. Though successful, this work was limited as they did not consider passenger demographics and other passenger specific information.

Mining customer reviews to ascertain satisfaction or dissatisfaction is another approach researchers have explored. Lucini et. al [5] employed this approach to mine airline customer reviews using latent Dirichlet allocation (LDA) for topic modelling to extract the supposed factors affecting the customer satisfaction or dissatisfaction in the review. From their results, they identified the type of cabin flown to be the biggest impact on customer satisfaction status while the type of passenger had the least feature importance. Using these results, they were able to propose measures targeted at improving customer satisfaction. This research was mostly qualitative as it only identified factors with impact on customer service but failed to quantify this impact. In a similar approach using text mining, Wang et. al. [6] analysed customer satisfaction depending on individual attributes of washing machines in an online store. Positive or negative customer sentiments were mined from the reviews and used to represent customer satisfaction or dissatisfaction. Using a logistic regression model, they were able to understand the impact of each of the product attributes used in the analysis on potential customer satisfaction or dissatisfaction. Amongst the attributes analysed, it was found out that the impact of relevance of design features on customer satisfaction was dependent on the price of the product. This work was insightful, but it did not take into account customer specific data for the analysis as different customers might have different factors affecting their satisfaction status based on personal preferences or attributes. Another work using a text mining approach was done by Sezgen et. al. [7]. They applied latent semantic analysis (LSA) to extract features affecting airline passenger satisfaction from passenger reviews. From the results, it was seen that the class of travel had an impact on the factors affecting satisfaction. For economy flyers, friendliness and helpfulness of the crew was an important factor for satisfaction amongst economy class flyers while product value was important for premium class flyers. Across both class of flyers, legroom space and seat comfort were often reasons for customer dissatisfaction.

As is seen from the reviewed papers herein, the random forest model is an effective model for customer satisfaction prediction as it can be used to identify feature importance and has a high degree of accuracy as seen in [1], [2] and [3]. The logistic regression algorithm is interpretable and can measure feature importance but offers lower accuracy values than the random forest algorithm. The other approaches are blackbox models that can not be interpreted to explain feature importance. As feature importance is an essential part of this analysis, a random forest algorithm is adopted. The important features are investigated to indicate which features can be improved upon to see an immediate increase in passenger satisfaction levels.

III. Methodology

A. Dataset

The dataset used in this analysis was procured from the Kaggle data repository submitted by [8]. The data is a customer satisfaction survey carried out by an airline to gauge the level of customer satisfaction on the each of the offered services. The customers rate each aspect of the service on a scale of 0 through 5 where 0 represent no satisfaction and 5 represent maximum satisfaction. Some of the aspects of service rated by the customers include inflight Wi-Fi service, baggage handling, ease of online booking, seat comfort and food and drink. The data contained 129,880 observations across 24 variables.

B. Data Pre-processing

As is the case in any analysis with data, data pre-processing into the correct format for an insightful analysis is required. Though this dataset was fairly clean, it had missing values. The observations with missing values were negligible as only 393 out of the 129,880 observations had missing entries. A decision was made to drop them from the data as it would not have any negative impact on the results.
The categorical variables in the data were also one hot encoded as most machine learning models cannot analyse categorical data directly.

C. Data Exploration

The data is first explored to expose preliminary insights. The average ratings for each of the services is shown in Table 1. It is shown that the customers were moderately satisfied with each of the services though there is still a lot of room for improvement of these services. Further exploration was done to check if the flight class affected the satisfaction levels of the customer. From Fig. 1, it was seen that across the 3 flight classes eco, eco plus and business, only business class customers were more satisfied than dissatisfied. Though not surprising as business class customers are often offered better services by airlines, it is further investigated in the analysis.

TABLE I
AVERAGE RATINGS FOR DIFFERENT SERVICES

Service	Average rating
Inflight WiFi	2.72
Departure/Arrival Time convenience	3.06
Ease of online booking	2.76
Gate Location	2.97
Food and Drink	3.2
Online boarding	3.25
Seat comfort	3.44
Inflight entertainment	3.36
Onboard service	3.38
Leg room service	3.35
Baggage handling	3.63
Checkin service	3.31
Inflight service	3.64
Cleanliness	3.28

D. Applied technique

From the works reviewed, it is seen that classification algorithms such as logistic regression, random forest, support vector machine (SVM), Adaboost, K-nearest neighbours (KNN) and the Naïve Bayes approach can all be used for

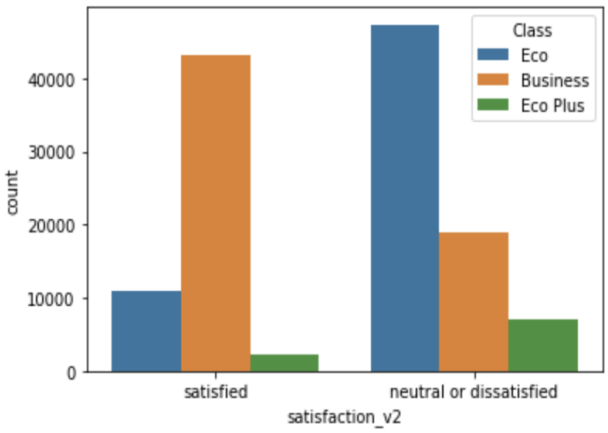

Fig. 1. Satisfaction levels across different flight classes

predicting customer satisfaction status. Some of these approaches though satisfactory have some drawbacks as their results are not interpretable. That is, the impact of each of the features on customer satisfaction or dissatisfaction cannot be measured. Algorithms such as SVM, KNN, and Naive Bayes are blackbox algorithms as they produce no interpretable model at the end of the analysis [9]. They also do not indicate the level of importance of each of the features used in the analysis on the outcome.

Logistic regression offers full interpretability as it produces a mathematical model at the end of the analysis. It is able to quantify the impact of each of the features on the outcome. However, it is a linear model [9] and may produce sub optimal results if the data is not linearly separable. The data to be used is also highly dimensional as over 24 features are to be considered for the analysis. This limits the performance of the logistic regression algorithm.

The random forest algorithm though a blackbox model, can be used to indicate the impact of constituent features on the outcome as well as perform predictions to a high degree of accuracy. It also handles higher dimensional and non linearly separable data as it is an ensemble of decision trees. As feature importance is an essential part of this work, as well as highly accurate predictions, the random forest model is the chosen technique to be applied. The random forest algorithm is an ensemble algorithm made up of individual decision trees. They can be used for both classification and regression purposes. Each tree is built with a different set of data sampled from the main data. Sampling the data with replacement is a form of sampling called bootstrap aggregation (bagging) [9]. When splitting the tree at each node, a subset of the features are used. The final result is obtained by averaging the results from each of the constituent decision trees.

IV. Results and Discussion

A. Model Evaluation

The random forest classifier had an accuracy value of 96.19% with precision and recall values of 0.94 and 0.98. Denoting satisfied customers as positives and dissatisfied customers as negatives, the model makes a false positive error when it classifies a dissatisfied customer as satisfied and makes a false negative error when it classifies a satisfied customer as dissatisfied. As the airline strategy is mostly focused on identifying dissatisfied customers, false positive errors become very costly and need to be limited as much as possible as they lead to misidentification of dissatisfied customers as satisfied. Owing to this, the specificity score becomes a very essential metric to evaluate the model as it indicates the level of false positive errors being made. The higher the specificity score, the better the model performance. The random forest model also performed well in this regard as it had a specificity score of 0.94. The sensitivity score indicates the level of false negative errors made. In this analysis, they are the less expensive errors as an error classifying a satisfied customer as dissatisfied will not lead to any loss of customers. The model still performed well with a sensitivity score of 0.98.

TABLE II
Random Forest Confusion Matrix

	Predicted(satisfied)	Predicted(dissatisfied)
Actual(satisfied)	21436	467
Actual(dissatisfied)	1012	15932

TABLE III
Model Result

Model	**Accuracy%**	**Precison**	**specificity**	**sensitivity**
Random forest	96.19	0.96	0.94	0.98

B. Feature Importance

The importance of each of the features on the outcome was obtained from the model results. The results indicate the features influencing customer satisfaction in descending order of importance. From Fig. 2, it is seen that online boarding, inflight wifi service, type of travel, travelling business class and inflight entertainment are the top 5 features impacting customer satisfaction. This means that improvement in these aspects of service will likely see an immediate increase in customer satisfaction levels. Travelling in the Eco plus class, gender, departure delay and food and drink were the bottom factors affecting customer satisfaction.

From the data exploration phase, it was seen that only customers in the business class were more satisfied with services than neutral or dissatisfied. This was explored further to ascertain the drivers of customer satisfaction for each of the flight classes as this might be the reason business class flyers were more satisfied than the other classes of passengers. From, Fig. 3, it is seen that in the eco class, inflight wifi service, online boarding, ease of online booking, flight distance and age were the top 5 factors affecting the satisfaction status of the customer while gender, gate location, cleanliness, departure/arrival time convenience and food and drink were the factors least affecting customer satisfaction.

On the other hand, from Fig. 4, it is seen that business class flyers satisfaction were most impacted by online boarding, inflight wifi service, seat comfort, inflight entertainment and legroom service while they were least impacted by gender, departure or arrival delay and food and drink.

For customers flying eco plus, the most important factors impacting satisfaction were inflight wifi service, online boarding, type of travel, and ease of online booking while the least important factors were gender, gate location and departure delay.

From the feature importance results for each of the classes, it can be seen that different factors drive satisfaction for different classes of travellers and this could be an explanation for why only those in business class were more satisfied than dissatisfied.

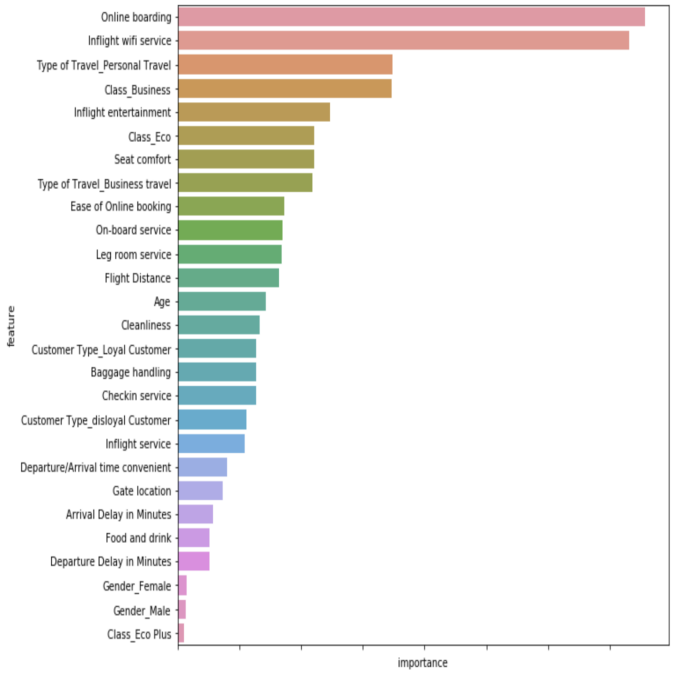

Fig. 2. Overall Feature Importance

V. Conclusion

Customer satisfaction is known to be a very important concept for businesses as they aim to acquire and retain customers to maximise profits. This work leverages predictive analytics to predict customer satisfaction by modelling affecting factors for an airline. Explaining the impact of each of the factors on the outcome is another important part of this work as it helps the business know where to focus efforts to see immediate improvement gains. The random forest algorithm was employed for the analysis and it performed satisfactorily with an accuracy score of. 96.19% and precision and recall of 0.96 and 0.98 respectively. In the analysis, the satisfied customers are denoted as positives and dissatisfied customers as negatives. As airlines are more interested in dissatisfied customers, classifying dissatisfied customers as satisfied are

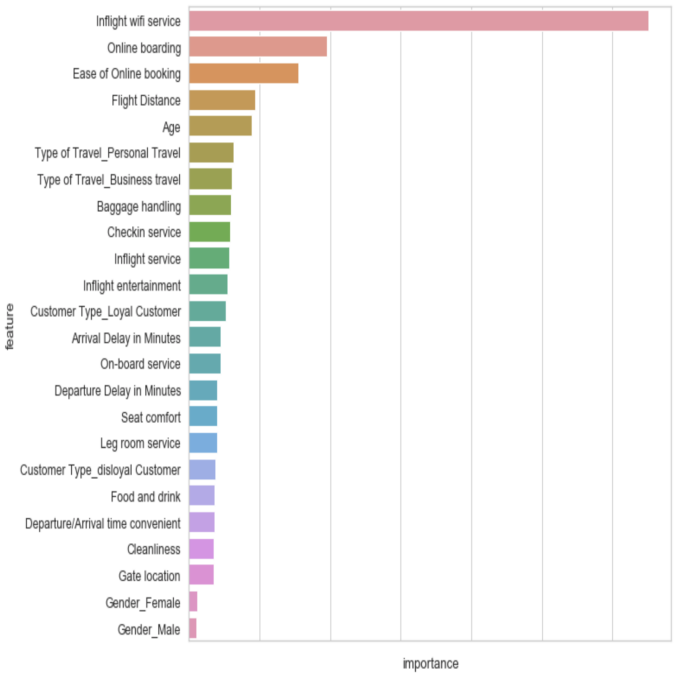

Fig. 3. Feature Importance for the Eco class

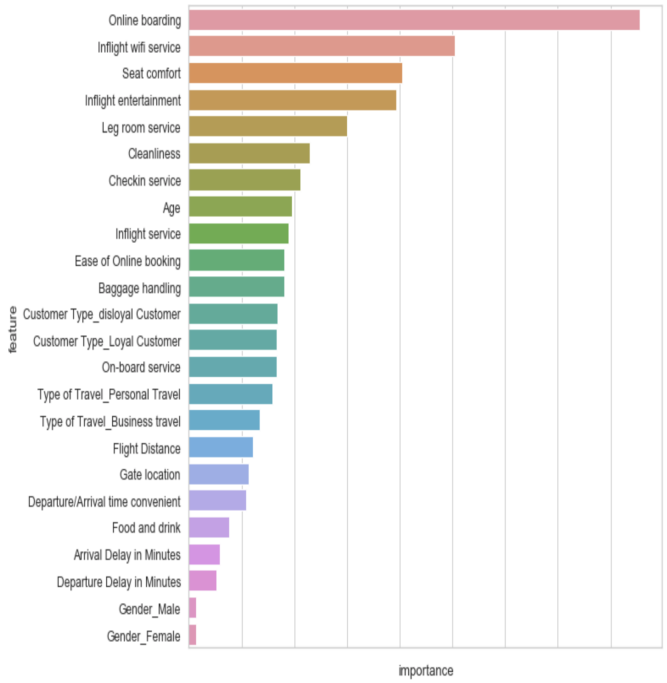

Fig. 4. Feature Importance for the business class

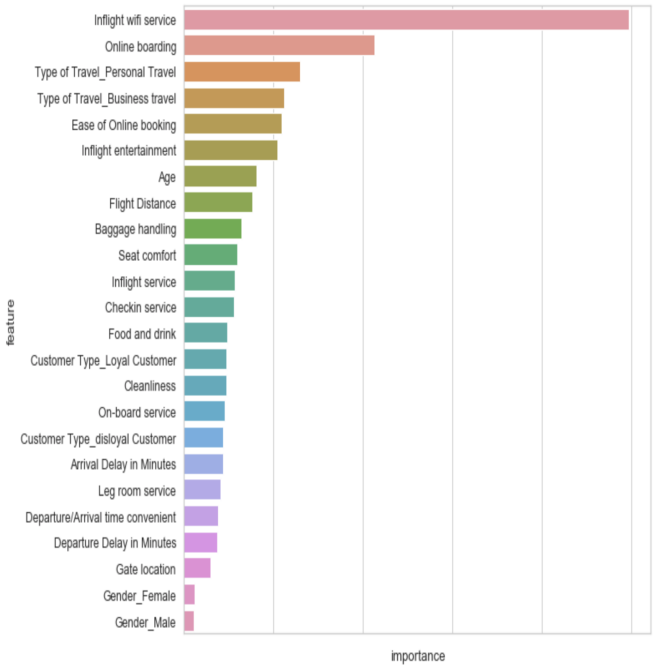

Fig. 5. Feature Importance for the Eco Plus class

false positive errors and can lead to loss of the customers business thereby defeating the main purpose of the analysis. Due to this, specificity is an important metric to evaluate the model as it indicates the level of false positive errors made. The model had a specificity score of 0.94 which was quite satisfactory. The analysis was further broken down for 3 different flight classes, eco, eco plus and business class to identify the factors driving customer satisfaction in each of those flight classes. Inflight wifi service and online boarding were common amongst the top factors influencing satisfaction in the 3 flight classes. Ease of online booking and was a common factor affecting passenger satisfaction amongst eco and eco plus passengers while business class passengers were mostly impacted by seat comfort and leg room service. This variation in tastes for different flight class customers could be an explanation for why business class passengers were mostly satisfied than dissatisfied. This work can be leveraged as a decision making tool for airlines to optimize services and increase passenger satisfaction leading to increased revenue. Further analysis can be done to identify the features affecting customer satisfaction for different type of travellers as a business traveller might be more interested in inflight wifi services to keep up with latest business information while a leisure traveller might be more interested in inflight entertainment.

For future work, other machine learning approaches could be applied to further eliminate errors and improve performance.

REFERENCES

[1] S. Meinzer, U. Jensen, A. Thamm, J. Hornegger, and B. M. Eskofier, "Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry," *Artificial Intelligence Research*, vol. 6, no. 1, pp. 80-90, Dec. 2016. doi:10.5430/air.v6n1p80

[2] N. Liu, S. Kumara, and E. Reich, "Explainable data-driven modelling of patient satisfaction survey data," *in 2017 International Conference on Big Data (Big Data),Boston, MA, USA, December 11-14, 2017*, pp. 3869-3876. [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/document/8258391[Accessed on: Feb. 25, 2020]

[3] A.K. Singh, M. Yoo, and R.J. Dalpatadu, "Determinants of Customer Satisfaction at the San Francisco International Airport," *Journal of Tourism Hospitality*, vol.8, no. 398, pp. 2167-0269, Dec, 2019. doi:10.35248/2167-0269.19.8.398

[4] H.B. Sankaranarayanan, B.V. Vishwanath, and V. Rathod, "An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees," *In 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, September 23-25, 2016*, pp. 285-290. [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/document/7813672 [Accessed on: Feb. 25, 2020]

[5] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto, and M. J. Anzanello, "Text mining approach to explore dimensions of airline customer using online customer reviews," *Journal of Air Transport Management*, vol. 83, pp. 101760-101712, Mar. 2020. doi:10.1016/j.jairtraman.2019.101760

[6] Y. Wang, X. Lu, and Y. Tan, "Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines," *Electronic Commerce Research and Applications*, vol. 29, no. 5, pp. 1-11, Mar. 2018. doi:10.1016/j.elerap.2018.03.003

[7] E. Sezgen, K.J. Mason, and R. Mayer, "Voice of airline passenger: A text mining approach to understand customer satisfaction", *Journal of Air Transport Management*, vol. 77, no. 34, pp.65-74. June 2019. doi:10.1016/j.jairtraman.2019.04.001

[8] D. John, "Passenger satisfaction," 2018. [Online]. Available: https://www.kaggle.com/johnddddd/customer-satisfaction [Accessed on: Feb. 20, 2020]

[9] B. Lantz, *Machine learning with R*, 3rd ed. Packtz, Birmingham UK, 2019.