Predictive Analytics for Airline Customer Satisfaction

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

***Abstract*—Predictive analytics is now being applied across business domains to answer business questions and make data driven decisions especially with increase in data availability and computing resources. The airline industry can benefit highly from employing predictive analytics in handling customer satisfaction and understanding the impact of contributing factors as it is an important part of the business operation. This is the focus of this project. The design phase including the introduction of the data, ethical considerations, discussing the analytics strategy and preliminary visualisations are presented herein.**

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

I. Introduction

Owing to the rise in competition between businesses, most businesses have made customer satisfaction a priority as they are the focal point of any profitable venture. Businesses aim to deliver satisfactory experiences to their customers as studies have shown that customer dissatisfaction can have damaging impacts on the organization as information travels fast these days[1]. This is more the case in the airline industry as the competition is fierce and fight for customer loyalty is immense. In recent years, airlines have digitized most processes all in a bid to provide more efficient services and better product offerings to customers.
Customer satisfaction is very important as a satisfied customer is more likely to be a return customer and subsequently a loyal customer. With the advent of machine learning techniques and appropriate computing power, airlines can leverage predictive analytics to model and understand factors affecting customer satisfaction. This enables a customer-centric approach to operations. Analysis results can help deliver better customer service, optimize services and enable targeted product offerings all potentially leading to improved profits.

II. Aim and Objectives

In the airline industry, there are a number of factors that could affect customer satisfaction. This project aims to model these factors to predict customer satisfaction, understand and quantify their impact on potential customer satisfaction or dissatisfaction through machine learning techniques.

III. Background

This section outlines background information on predictive analytics and machine learning techniques powering it. Predictive analytics can be defined as the use of data and machine learning or statistical methods to predict future outcomes[1]. It encompasses an understanding of historical data and subsequent identification of the likelihood of future outcomes based on this knowledge. Machine learning techniques used to perform predictive analytics are broadly divided into supervised and unsupervised learning algorithms.

A. Supervised Machine Learning

For supervised machine learning techniques, the outcome variable for potential prediction is present in the dataset and can be predicted by other variables present in the dataset[2]. That, is, there is a clear target variable in the data for prediction. Supervised machine learning algorithms are further divided into classification and regression algorithms.
Regression algorithms are used for prediction when the target variable is numeric and continuous in nature[2]. Examples of prediction using regression algorithms in business domains include predicting house sale prices or sales forecasts as the predicted values are numeric and continuous in nature.
Classification algorithms are used for prediction when the target variable is categorical and discrete[2]. Some examples of classification algorithms include logistic regression (LR), decision trees, random forest (RF), naive bayes (NB) and support vector machine (SVM). Predicting customer satisfaction status is a classification problem as the predicted variable is categorical (satisfied or unsatisfied) and as such, classification algorithms are employed for the analysis.

B. Unsupervised Machine learning

Unsupervised machine learning algorithms are used when there is no clear target variable in the dataset for prediction[2]. They are mainly used for inference purposes to understand similarities or underlying patterns in the data. They are sometimes used as part of the data preparation process before use of supervised learning techniques. An example of application of unsupervised techniques for in business domains is customer segmentation for targeted marketing.

[1]https://www.sas.com/en$_ie/insights/analytics/predictive$ $analytics.html$

C. Evaluation Metrics

To evaluate the performance of machine learning algorithms for prediction or classification purposes, certain standard metrics are used. Regression problems have different metrics from classification problems. As this work focuses on a classification problem, only classification metrics are discussed.
Some of the classification evaluation metrics are derived from the confusion matrix. The confusion matrix is a diagnostic table for gaining insights into the errors being made by the model[2]. The evaluation metrics as discussed as follows;

TABLE I
CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	True Positive	False Negative
Actual(0)	False Positive	True Negative

- **Accuracy:** The accuracy of a model is the ratio of the number of correct predictions to the total number of predictions. When the data classes are highly imbalanced, this metric is not enough to judge model performance as a naive model will have a high accuracy without learning anything[2].

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

- **Kappa:** The kappa statistic adjusts the accuracy of a model by factoring in the possibility of making a correct prediction by chance[2]. For datasets with high class imbalance, the metric should be use to get a better sense of performance than accuracy. The higher the kappa value, the better the model.

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (2)$$

- **Precision:** Precision is the ratio of the truely positively labelled predictions to all the positively labelled predictions. In other words, It is the proportion of model predicted positives that truely positive [2].

$$Precision = \frac{True\ positives}{True\ positives + False\ Positives} \quad (3)$$

- **Specificity:** Specificity measures the proportion of the negative data points classified correctly [2]. It is the ratio of the true negatives to the sum of the true negatives and the false positives.

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ Positives} \quad (4)$$

- **Sensitivity:** Sensitivity shows the amount of correctly classified positive data points [2].

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} \quad (5)$$

- **F-measure:** The F-measure is calculated as the harmonic mean of the precision and recall.

$$F - measure = \frac{2 * precision * recall}{recall + precision} \quad (6)$$

IV. DATASET

The dataset for the project was obtained from kaggle submitted by [3]. The data was collated through a customer satisfaction survey. The customers rate various aspects of service on a scale of 0 through 5 depending on their level of satisfaction with the service. The variables are listed in the table below.

TABLE II
DATASET VARIABLES

Variable Name	Description	Data Type
satisfaction$_v$2	Overall satisfaction status	categorical
Gender	customer gender	categorical
Customer Type	Customer loyalty status	categorical
Age	age of the customer	continuous
Type of travel	type of journey embarked on	categorical
Class	class of cabin flown	categorical
Flight distance	Total distance if flight in km	numeric
Inflight wifi service	inflight wifi service ratings	ordinal
Departure convenience	customer ratings for departure	ordinal
Online booking ease	online booking ratings	ordinal
Gate location	gate location convenience	ordinal
Food and drink	food and drink ratings	ordinal
Online boarding	online boarding ratings	ordinal
seat comfort	seat comfort ratings	ordinal
inflight entertainment	entertainment ratings	ordinal
on board service	onboard service ratings	ordinal
leg room service	leg room service ratings	ordinal
Baggage handling	baggage handling ratings	ordinal
checkin service	checkin service ratings	ordinal
inflight service	inflight service ratings	ordinal
cleanliness	cleanlines ratings	ordinal
departure delay	departure delay time in mins	continuous
arrival delay	arrival delay time in mins	continuous

V. ETHICAL CONCERNS

As customer data is used to carry out predictive analytics, it is important to understand the ethical concerns surrounding the data use and validate the ethics of the project.
Customer data is collected and analyzed to offer better or customized services thereby improving customer experience and ensuring satisfaction. However, there are issues of privacy, data security and usage restrictions that need to be considered. The data used for the project is a combination of customer survey results rating different aspects of the airline service, customer specific information such as age and airline supplied information such as flight distance. As the surveys are assumed to be freely filled by the customers knowing the intended purpose of the survey is to improve customer service, using this data for analysis becomes both ethical and legal. However, there should be a limit to the application of results obtained from the analysis. Intrusive product offerings become unethical except the customer specifically signed up for them during the survey.

VI. Related Work

Researchers have looked into the use of machine learning models for predicting customer satisfaction across different business domains.

One of such was the work of Meinzer et. al. [1] using machine learning techniques to predict customer satisfaction in the automobile industry. They extracted and used 105 features from warranty, diagnostics, dealer system and general vehicle data for the analysis. They used four machine learning algorithms Adaboost, K-Nearest neighbours (KNN), support vector machine (SVM) and random forest (RF) for the prediction. The SVM radial basis kernel outperformed the other algorithms in terms of accuracy as it able to correctly classify 72% of the validation data. Though this work was insightful, it only used 'blackbox' models hence the relationship between the classification outcome and the predictors cannot be explained. To counter the problem of interpretability, Liu et. al.[4] proposed a local explanation method to interpret factor importance combined with a random forest classifier for the prediction of patient satisfaction from patient survey data. Unhappy patients were the focus of the explanations so as to find out the top reasons for dissatisfaction. The combination of a random forest classifier and local explanation resulted in an interpretable model that could be leveraged for quality healthcare. Customer reviews have also been used to measure satisfaction though they are largely qualitative. Lucini et. al. [5] employed a text mining approach to explore customer satisfaction in the airline industry from online customer reviews. They analysed over 55,000 reviews spanning across over 400 airline companies using latent Dirichlet allocation as a topic modelling technique to detect factors affecting customer satisfaction. They found out that type of cabin flown had huge impact on customer satisfaction while type of passenger impacted the least. Based on the analysis results, they were able to suggest methods of targeted improvement in customer service to achieve better customer satisfaction. This research was qualitative in nature as it only identified the factors affecting customer satisfaction by topic modelling of customer reviews but did not precisely quantify the impact of the factors on customer satisfaction. Also employing a text mining approach, Wang et. al.[6] analysed the impact of product attributes on customer satisfaction from online reviews for washing machines. They carried out sentiment analysis on the reviews to discern the satisfaction status of the customer. They then built a logistic regression model to understand the impact of some product properties on customer satisfaction. On further investigation of price as a factor affecting customer satisfaction, they found out that the relevance of design features on satisfaction was different between customers who purchased cheaper products than those who purchased more expensive products. The work was insightful and quantitative as it explained the impact of each factors on potential customer satisfaction and validated the use of machine learning models for this purpose. However, it did not take into account any of the customers personal features to understand how different features may appeal to different people. Using a naive bayes approach, Sanchez-Franco et. al. [7] were able to successfully classify customer satisfaction from online customer reviews. They analysed over 47,000 reviews in hospitality domain extracting common term unigrams in the reviews as features and the review star rating as the satisfaction level.

VII. Applicable Techniques

As can be seen from the works reviewed herein, classification algorithms such as Adaboost, KNN, SVM, random forests, logistic regression and naive bayes are all applicable predictive analytics involving prediction of customer satisfaction. Though some of these algorithms function differently, they are all able to perform classification to a high degree of accuracy. A major drawback is that most are 'blackbox' models that do not offer any interpretability. This problem is highlighted and solved in [4] using a local explanation method. However, the logistic regression algorithm offers full interpretability, the ability to quantify the impact of each of the contributing factors and also accurately classifies data to a large extent as evident in [6] hence would be a suitable model for the analysis.

A. Model evaluation

Noting the satisfied customers as positives and dissatisfied customers as negatives, false positives occur when the model predicts a dissatisfied customer as satisfied and false negatives occur when the model predicts a satisfied customer as dissatisfied. An airline would be most interested in dissatisfied customers to attempt to improve their services and keep their business, hence predicting dissatisfied customers as satisfied (false positives) are the more expensive errors that could lead to customer loss as palliative measures such as offers geared towards appeasing the customer might be forgone when it should not have been. This should be minimized as much as possible, hence, specificity becomes a very important evaluation metric and should be as good as possible.

VIII. Strategy

Some of the strategies employed by airlines to drive customer satisfaction can be grouped into pre-flight, inflight and post-flight services [2].

Pre-flight services include provision of seamless methods of booking flights, detail checking and destination exploration brochures. Charges are also made as open as possible. This gives the customer full knowledge of the trip as all the details are known to map out a suitable flight plan for whatever the reason of travel.

For inflight services, the strategy employed can be at a personal level as customers travelling for different reasons might prefer different aspects of the service. Comfortable seats and leg room are also important service aspects that are considered.

Post-flight services also play an important role in customer

[2]http://www.customerservice.ae/resources/featured/strategies-for-airlines-to-take-off-towards-customer-satisfaction/

satisfaction. Baggage handling, accommodation for cancelled flights, refunds, connecting flights are all made as seamless as possible for the customers.

Predictive analytics can be leveraged to improve these strategies and help airlines adopt a more customer-centric business approach[8]. It can play a vital role in enabling targeted offers, improving services, enhancing customer loyalty, acquiring new customers and increase competitive advantage. One of such customer-centric approaches is the analysis of customer satisfaction survey data as a strategy for improving services and subsequent customer retention. This can show areas of the operation that are most critical to customer satisfaction and need improvement. Other business relevant questions are also answered in this analysis. As mentioned previously, the perception of service between different type of travellers such as business, leisure, first time or group travellers can also be analysed as these groups of travellers might have different needs[8]. For example, the business traveller might be interested in connectivity speed while the leisure traveller might be more interested in inflight entertainment. Being able to factor in such granular details can be the difference between a satisfied customer and a dissatisfied customer. Predictive analytics facilitates the ability to identify opportunities to introduce services desired by customers and provide the right offers and messages to customers throughout the journey[8]. Other avenues of driving customer retention such as loyalty cards and discounts are sometimes not as cost effective. When these incentives are optimized with predictive analytics, they can show better results[8]. An example would be offering discounts to identified potentially dissatisfied customers as a palliative measure. This would reduce cost of such discount campaigns and potentially increase revenue as only a smaller amount of customers are discounted.

Other factors outside the ones considered in this project can affect customer satisfaction. Such factors include customer specific or behavioural information such as current mood can affect potential satisfaction but those are hard to control or model. The analysis process is illustrated in Fig. 1. The

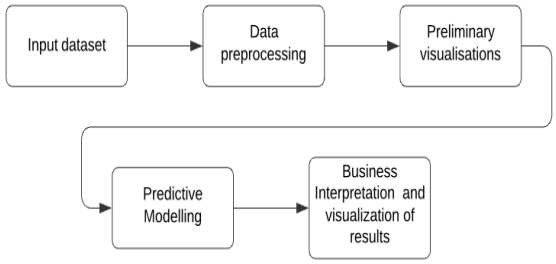

Fig. 1. Analysis Strategy

input dataset is preprocessed and cleaned. Then preliminary visualizations are done to get prior insights into the data before finally predictive modelling and then interpretation of results.

IX. Visualizations

Exploratory data analyis (EDA) is an important part of the predictive analytics process. The data is represented in plots and tables to expose relationships between variables, gain preliminary insights to the data and uncover interesting patterns. The plots are done using python 3.7.

Fig. 2 shows the summary of the customer ratings for variables inflight wifi service, depature/arrival time convenience, ease of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on board service, baggage handling, checkin service and inflight service on a scale of 0 to 5. It can be seen that the ratings for each of the variables were fairly positive as the plots were tending towards the right. This shows that the customers gave approximately satisfactory ratings though there is still a lot of room for improvement in these services. From the dataset, it can also be seen that there are 3 classes of flights, eco, eco plus and business class and most of the people in the survey flew business class while eco plus was the least as shown in Fig 3.

Moving on to more granular EDA, Fig. 4 shows the number of satisfied and neutral or unsatisfied customers. The neutral or dissatisfied customers were more than the satisfied customers hence there is need for major improvement to change these numbers. In terms of use for predictive analytics, the classes are slightly imbalanced and this is noted and can be corrected through undersampling or oversampling so as to produce a truly representative model without bias to any class. As shown in Fig. 5, only in the business class did the number of customers satisfied outweigh the dissatisfied customers with 69.4% to 30.6%. In the eco and eco plus classes, customers were generally less satisfied as only 18.8% and 24.6% were satisfied in both classes respectively. This raises the question, why are people in business class more satisfied than those in other classes?. This can be answered by the fact that airlines generally offer better services to business class customers but this will need further investigations during analysis. However, it is evident that service improvements are needed for other other classes of customers. Investigating the relationship between variables, the flight distance may be related to the ratings for on board and inflight services. This relationship is shown in the box plots in Figs. 6 and 7. It is seen that long distance travellers rated those services the least scores. This suggests that current level of service may be sufficient for short distance flights but need to be improved for longer flights. For customer specific information, the distribution of the customer ages in the survey is shown to be fairly normally distributed. On, further investigation of the impact of customer age on satisfaction status, it is seen that older customers are more likely to be satisfied than younger customers. This relationship will be further quantified during the next phase of analysis.

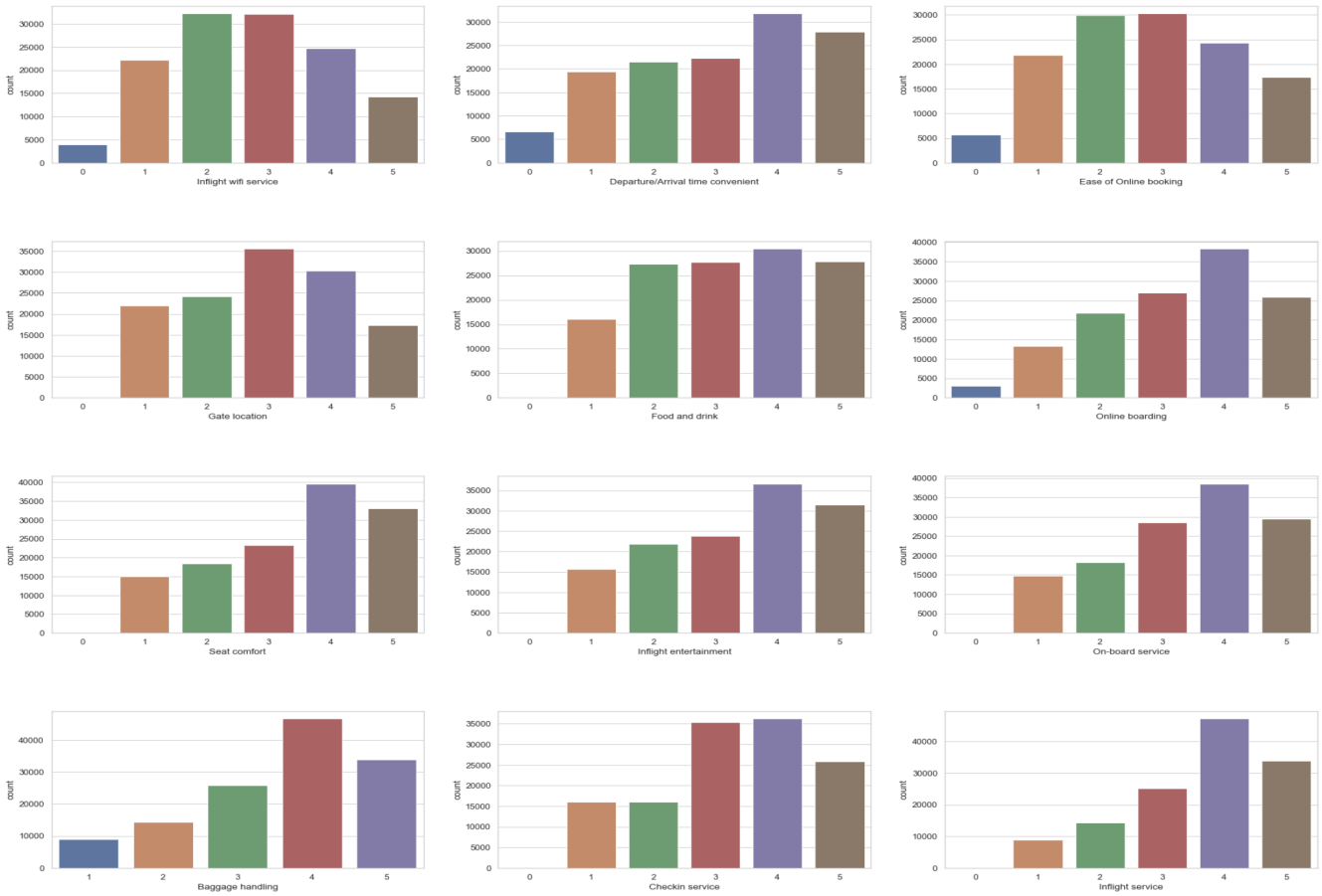

Fig. 2. Overall ratings

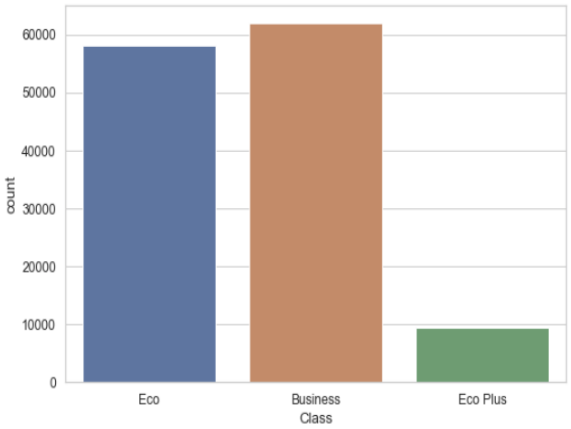

Fig. 3. Flight classes

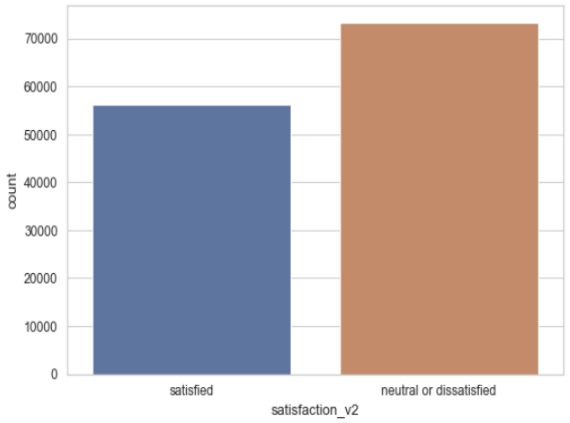

Fig. 4. Satisfaction status

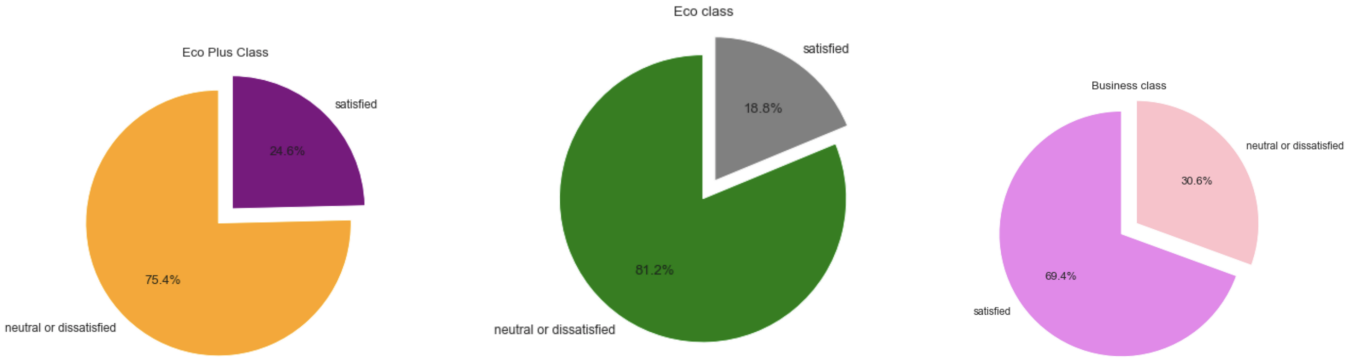

Fig. 5. Satisfaction status across different flight classes

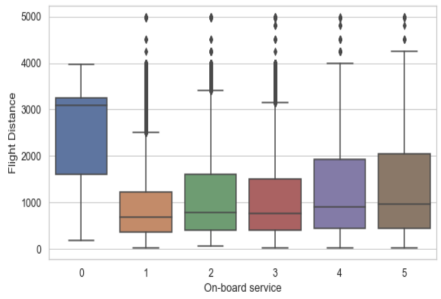

Fig. 6. Flight distance vs Onboard services rating

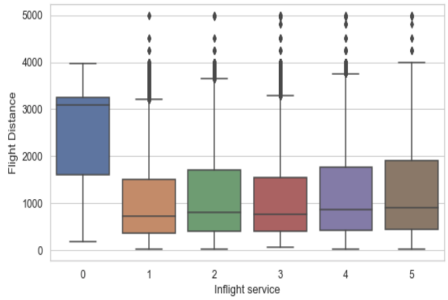

Fig. 7. Flight distance vs inflight services rating

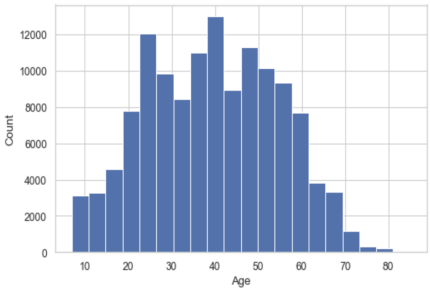

Fig. 8. Customer Age Distribution

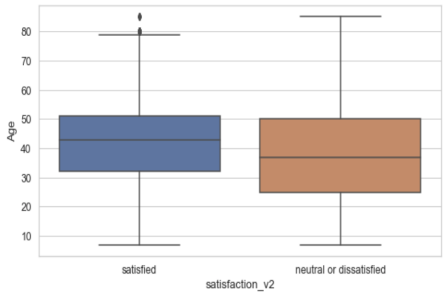

Fig. 9. Customer Age vs satisfaction status

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

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

[4] 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]

[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, pp. 1-11, Mar. 2018. doi:10.1016/j.elerap.2018.03.003

[7] M. J. Sanchez-Franco, A. Navarro-Garcia, and F. J. Rondan-Cataluna, "A naive bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services", *Journal of Business Research*, vol. 101, pp. 499-506, Dec. 2018. doi:10.1016/j.jbusres.2018.12.051

[8] S. Warner, and S. Sen, ”How predictive analytics elevate airlines’ customer centricity and competitive advantage”, 2011. [Online] Available:https://www.cognizant.com/InsightsWhitepapers/How-Predictive-Analytics-Elevate-Airlines-Customer-Centricity-and-Competitive-Advantage.pdf [Accessed on: Mar. 18, 2020]