Comparison of Various Classification Algorithms to Predict Hotel Booking Cancellations

K.Lakshmi Sairam [1], M.Nivas [2], K.U.S.Deveswar Reddy [3]
Vellore Institute of Technology, Chennai.

Email:
[1] lakshmisairam.kakarla2019@vitstudent.ac.in
[2]maddukuri.nivas2019@vitstudent.ac.in
[3]udaysurya.deveswar2019@vitstudent.ac.in

Abstract

Booking cancellations have a significant impact on the hotel industry. The percentage of cancellations is reaching almost 40%. If a person booked a room, and he/she canceled before the day of arrival it causes commercial loss to the management because to arrange the room for customers there will be some necessary work like cleaning that should be done. The main reasons for the unavailability of selected rooms,customer-centric reasons, parking spaces, etc. The objective of our project is to analyze and predict booking cancellations. The dataset which was used consists of the data of two hotels, by testing the attributes of the hotels we will find the various impacts that may be used to predict which bookings are likely to cancel. By using this data we will analyze and visualize the data which provides the infer about the performance of the two hotels by comparing their data. We have used Decision Tree, Random Forest, AdaBoost Classifier, and GradientBoost Classifier. From the F1 Score, we have found out that the Random Forest Model has outperformed all other algorithms. We have identified that lead_time,deposit_type, and ADR are the most important features that affect the booking cancellation.

Keywords: Decision Tree, AdaBoost Classifier, Random forest Classifier

I. Introduction

When lodgings attempt to protect themselves by utilizing administrations such as online booking websites, the burden at that point falls on OTAs. This benefit requires the OTA to pay for the reservation in case the booking is canceled and they cannot discover a modern visitor to the room. One thing is obvious, whether you're a lodger or an OTA, cancellations have a negative budgetary effect on your commerce. Last-minute cancellations are calculated and money related migraines for inns, but ones that have continuously been a necessary fiendish inside the industry. The point of the venture is to discover the reasons why they cancel the reservation and foresee whether a booking will be canceled. This forecast will be found by building the classification show.

The manager of the hotel confirmed that 250 out of 300 rooms have been reserved for a particular day. Various types of rooms have been reserved on that day. Even some people from other countries have been reserved about that. After two days, the manager observed that nearly 20 - 30% had been canceled. So, the manager wants to know which type of bookings are having a higher percentage of cancellations. And he wants to find a solution to avoid these problems. So, in this type of situation, our analysis and predictions will be useful to solve these constraints.

Our objective is to construct a model which was able to classify a booking as canceled or not. To do that we utilized information from the Kaggle. The information is initially from the article Inn Booking Request Datasets, composed by Nuno Antonio, Ana Almeida, and Luis Nunes for Information in Brief, Volume 22, February 2019. It comprises 119391 records and 36 traits. This information set will be utilized to make a client division examination to gain insights about the clients. We are going at that point to construct a classification demonstration to foresee whether or not a booking will be canceled with the most elevated exactness conceivable. This demonstrates a difference in inns to anticipate in case an unused booking will be canceled or not.

II. Overview

In this particular work, we started with exploring and visualization of data. By utilizing the information we analyze and visualize the information which gives the gathering almost the execution of the two hotels by comparing their information. The analytics which we performed will be used further for predictions using Decision Tree, Random Forest, AdaBoost Classifier, and GradientBoost Classifier. As our main motto is to predict the booking cancellation and the factors which lead to cancellations. Then by comparing the algorithms and calculating the F1 score, we will find the best performing algorithm and the factors that impact on booking cancellations.

III. Problem Statement

Cancellations are costly and damage the hotel's reputation. Our objective is to analyze the bookings of various types of rooms. Find the reasons for cancellations of reserved rooms. Predict the booking cancellations. we will find the various impacts that may be used to predict which bookings are likely to cancel. By using this data we will analyze and visualize the data which provides the infer about the performance of the two hotels by comparing their data.

IV. Literature Survey

The various research articles and publications we surveyed to understand the previous works done in the domain on booking cancellation predictions. In [1] Z.A.Andriawan et al have implemented and compared two random forest models in the CRISP-DM framework. Label encoding and one-hot encoding techniques were used and the generated random models were compared. Features are grouped together and the relative group importance on cancellation is explained. In [2] N.Antonio, L. Nunes, and A. de Almeida have implemented the Logistic regression Classification model. They have used optimization techniques such as AdaBoost to boost the gradient ascent algorithm. They have used L2 Normalization to prevent the overfitting of the model. Similar to the first article, they have also performed feature analysis.

Antonio, Nuno, and Ana[3] have proposed a deployment framework to integrate the machine learning and statistical model into the Central Reservation System used by hotels to manage and monitor their bookings. They have also explained how the data is collected, piped into the models and how models are selected, to produce the maximum accuracy. And doing this integration in real-time. In [4] Novakovic and Turina have implemented the KNN and Regression with AdaBoost Algorithm to predict hotel bookings. They also discussed the domain-specific drawbacks of these algorithms.

Falk, Martin; Vieru, Markku(2018)[5] have performed exploratory data analysis, produced various visualizations to understand the correlations between variables. Grouped variables and found relations between groups. Derived conclusions on the relative effects of the various variables on cancellation. Also provided a time-series based seasonality effect on the hotel booking cancellations.

V. Proposed Work

The Work consists of three phases. Data preparation, Analysis, and Model Building. In Data preparation we are going to deal with missing values and transformations by which we can achieve Accuracy, Consistency, and Uniformity. In the Analysis phase, analyze the bookings of various types of rooms. Find the reasons for cancellations of reserved rooms. And Visualize the data to get better insights. In the model Building Phase, Since the problem is under classification we chose classification algorithms like Logistic Regression, Random Forest Regression, and Support Vector Machine. After building the model, finding the accuracy is the foremost thing. So, we will choose the best fit model by comparing the accuracy.

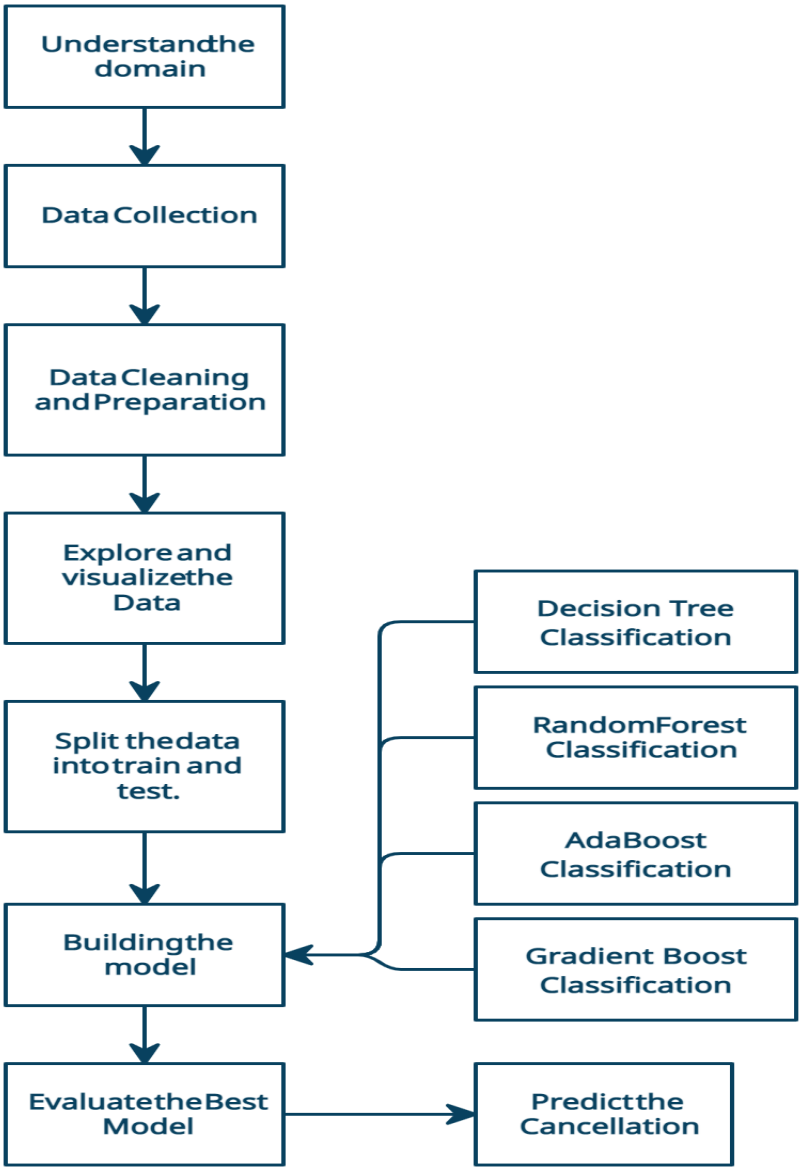

Fig-1. Proposed Workflow

VI. Feature Statistics

Feature statistics contain the relationship between various features and the target variable "is_canceled". It also contains the distribution and outliers plot.

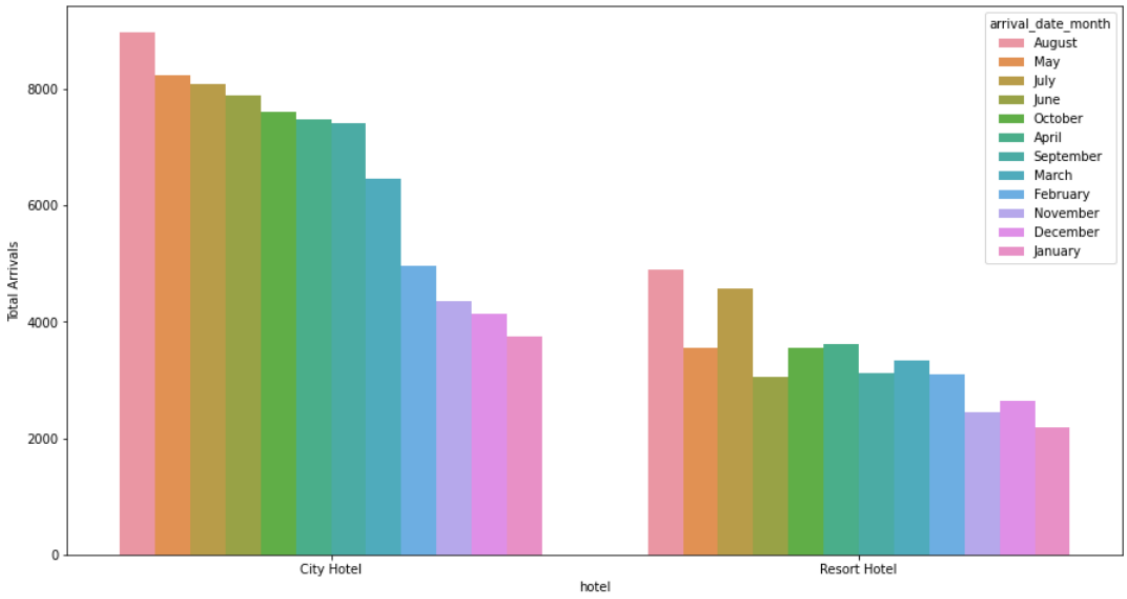

Fig-2: Month-wise bookings

The above bar graph is plotted to analyze month-wise arrivals for city and resort hotels. The plots vary with the number of bookings or arrivals at the city as well as resort hotels. City hotels get a greater number of arrivals than resort hotels every month. Both Resort hotel and City hotel had the highest number of arrivals in august. Both the hotels had the lowest number of arrivals in January.

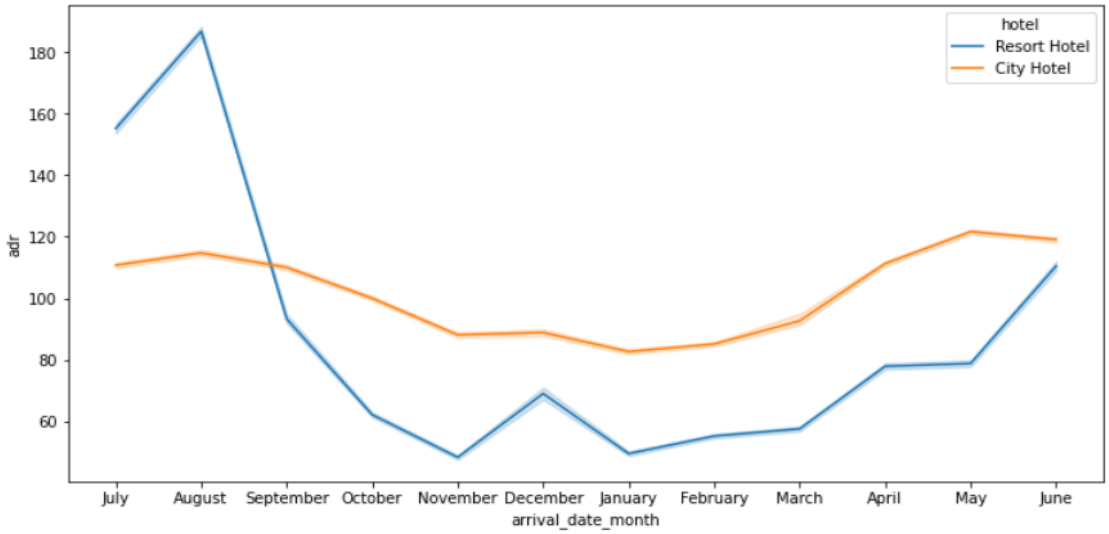

Fig-3: Month-wise average daily rate

The above graph is the month-wise average daily rate. It shows the comparison of people staying in different months. In-resort hotels the rate is increasing by every month till august. From August to December there is a wide decrease in the daily rate. In city hotels, the rate of staying is increasing but there is a wide range of increases for resort hotels. In the rainy season, there is a gradual decrease at city hotels but it was steep at resort hotels.

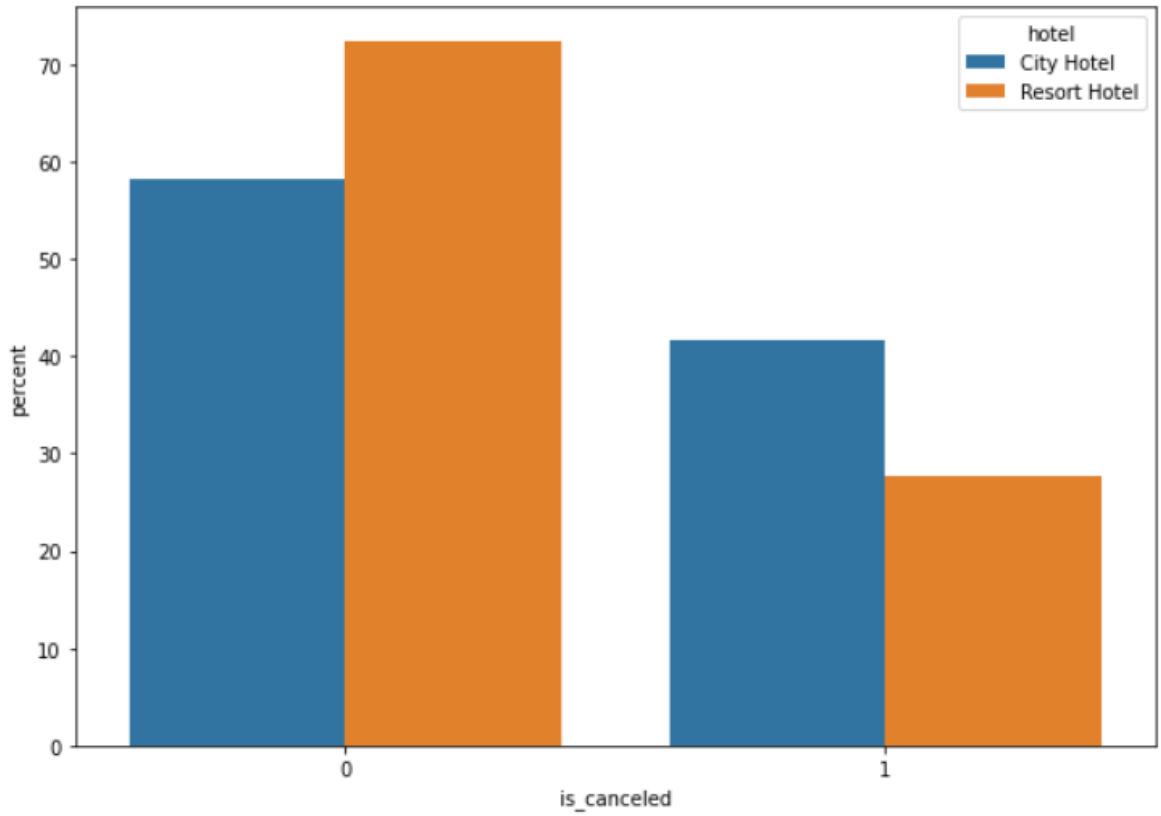

Fig-4: Rate of cancellations

This plot shows the rate of cancellations in hotels. So, the rate of cancellations has been greater in city hotels than resort hotels. Nearly 42% of city hotel bookings are canceled by the guests whereas resort hotels have only 27% of cancellations. Resort hotels had a greater number of bookings and less number of cancellations whereas city hotels had less number of bookings and a high number of cancellations.

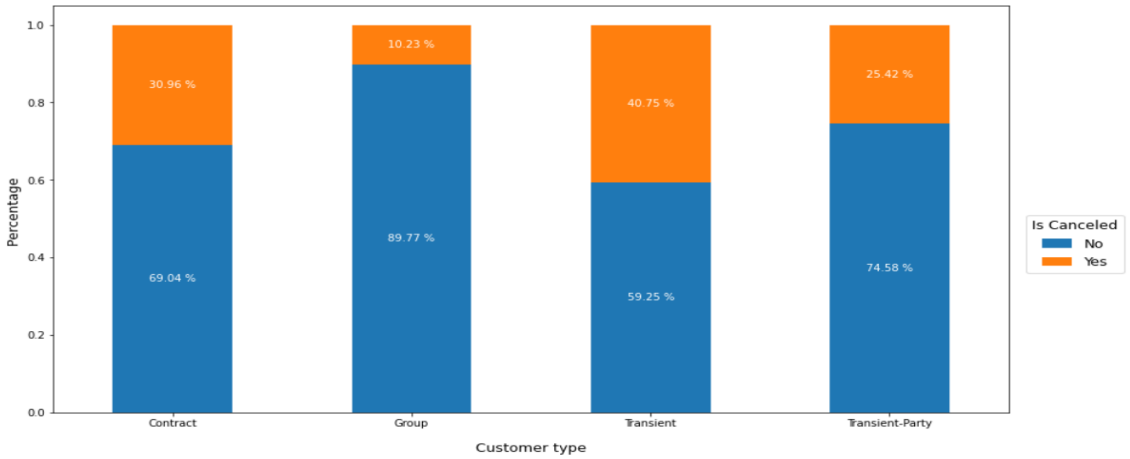

Fig-5: Cancellation rate by customer type

This graph shows the cancellation rate by customer type. There are four types of customers in bookings. Those are contract, group, transient, transient-party. The cancellation rate of group-type customers is 10% which is very less than other customer types. Contract base customers have a 31 % rate of cancellations, for transient and transient parties the rate of cancellations is 41% and 25% respectively. So, the hotels should maintain feedback from the transient party customers to find out the reasons for the cancellations.

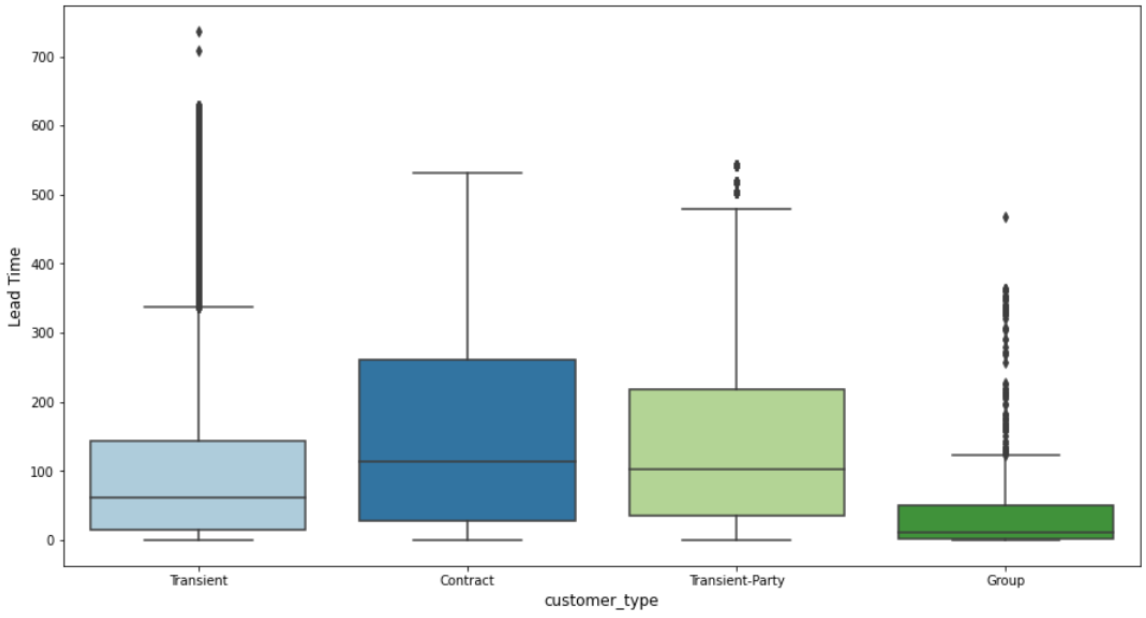

Fig-6: Distribution of lead time by customer type

This box plot shows the lead time by customer type. Lead time is the number of days that go by between the entering date of the hotel property and the arrival date. Transients and groups have a wide range of outliers. The customer type transient has the highest cancellation rate, reaching 41%. It can be concluded lead time is not having any impact on cancellations. Because contract customers and transient parties have a longer lead time than transient.

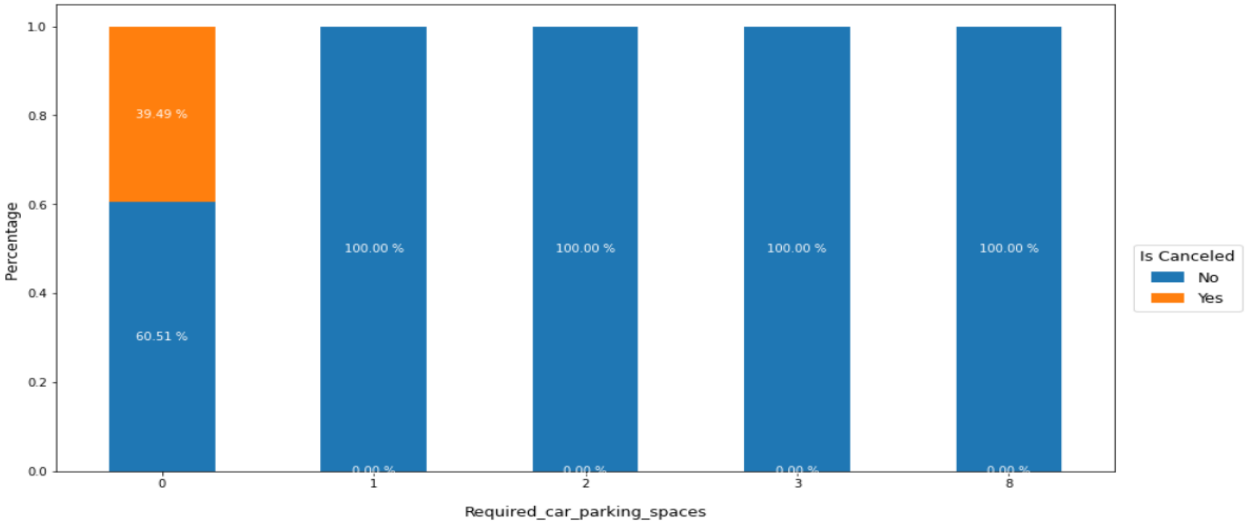

Fig-7: Cancellation rate by parking spaces

This plot shows the cancellation rate by required car parking spaces. The above graph shows that the required car parking spaces attribute can be used more predominantly for predicting whether the booking had the chance to cancel or not. The key from the above visualization can prove that Required car parking spaces will play a major role in cancellations. Nearly 40% of Customers who didn't request the parking spaces canceled their bookings. Other customers who requested parking spaces had no cancellations.

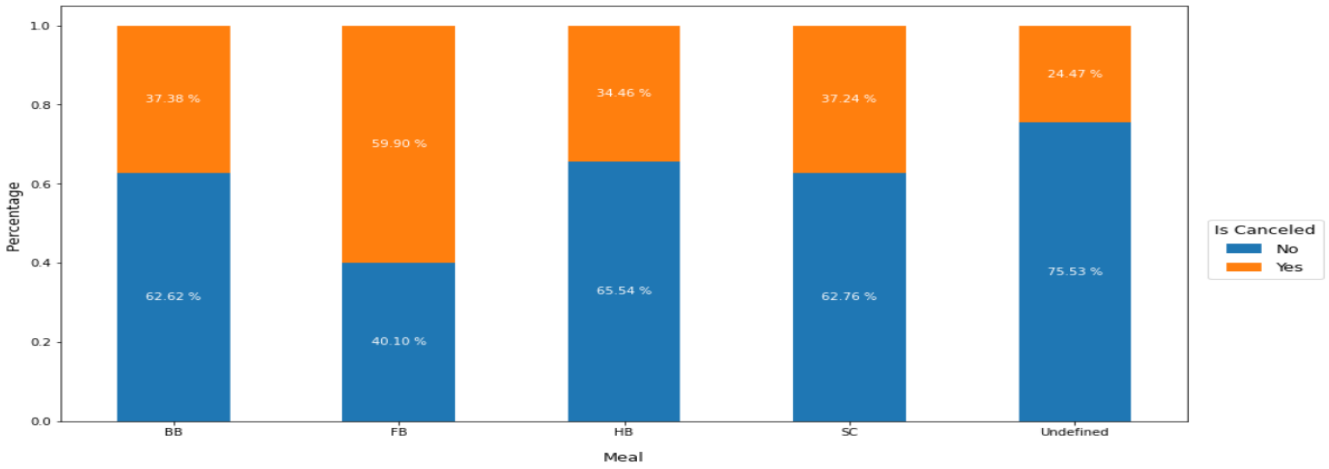

Fig-8: Cancellation rate by meal type

The plot shows the cancellation rate by meal type. Customers who selected full board meals have canceled most of their reservations. Customers who didn't mention any of the types of cancellation rates are not much greater.

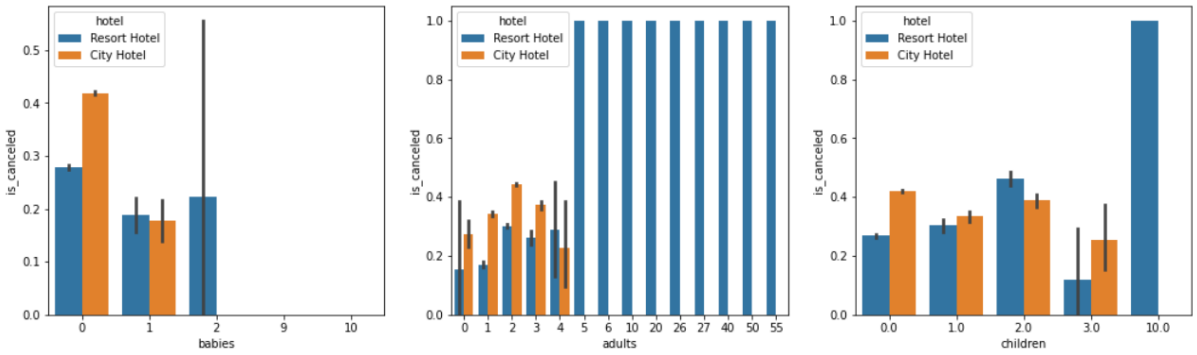

Fig-9: Comparison of cancellations by age groups

These subplots are the comparison of the cancellations by the number of babies, children, and adults in the hotels. In-resort hotels, most of the cancellations were made by adults whereas in city hotels the rate of cancellation is almost equal for different groups of ages.

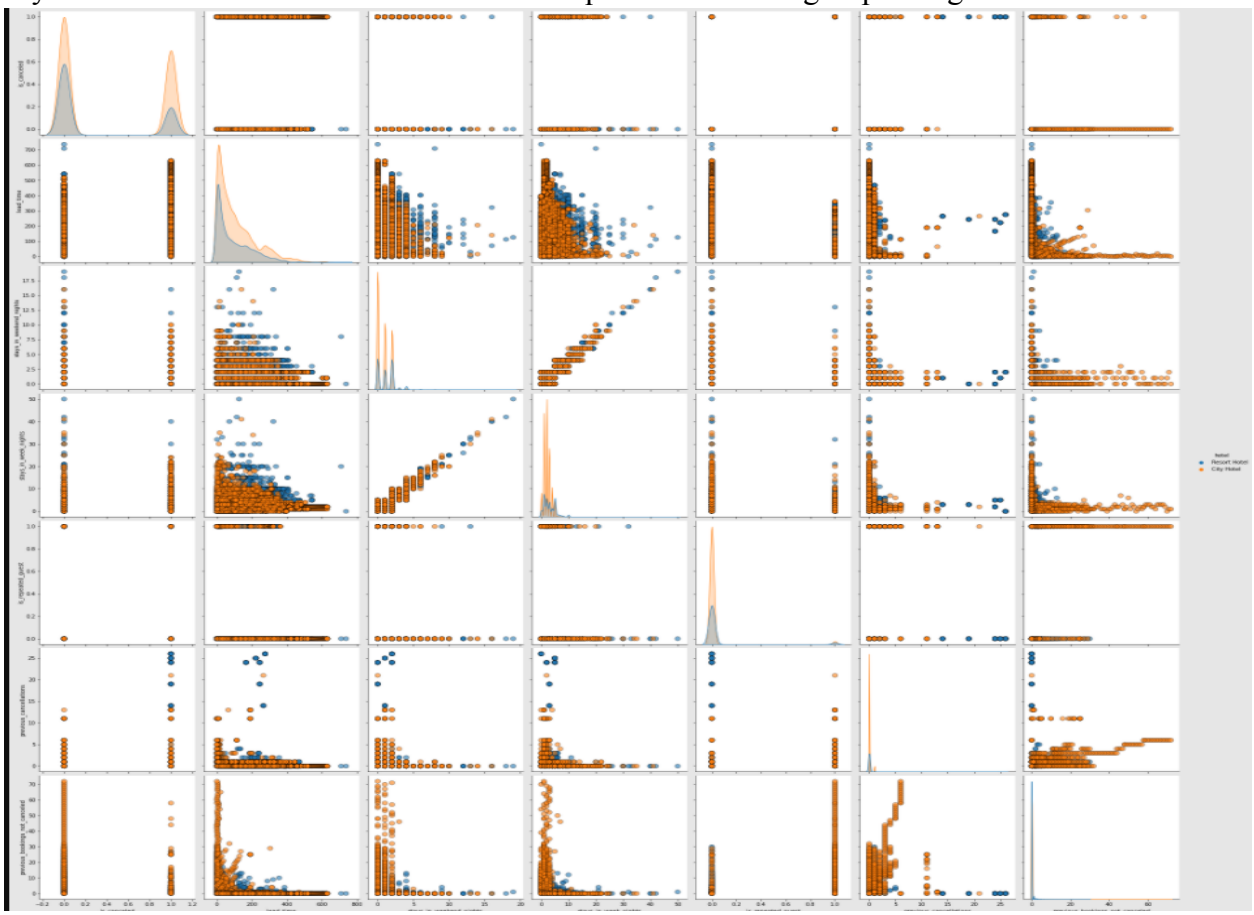

Fig-10: Pairwise Distributions

This above chart shows the pairwise bivariate distributions among the various variables in the dataset. It shows the various relationships between the attributes. It is used to understand the best fit of features like people staying on weeknights and weekend nights. The rate of cancellations after booking changes, etc.

Below are the boxplot of each column and their respective frequency distribution. Using boxplot outliers can be removed. If there are any outliers then the accuracy of the model will decrease. So, the outliers were removed using log transformations. In some of the columns, if the percentage of records containing outliers is less than 1% they were removed to increase the accuracy of the model. And in some columns like Adults, Children, Booking changes, Required car parking access, etc. There was no interquartile range this is because of the more uniqueness in the data. This is the reason to plot the frequency distribution graph. Where in the adult's column most of them are between 0 to 5. In the children column, most of them are 0 and so on.
Fig- 27,30 and 33 show the outliers in lead time, stays on weekend nights, and the babies column. The outliers in lead time were removed by using log transformation. Whereas, stays on weekend nights and the babies column outliers are less than 1%. Hence, the records containing outliers were eliminated.

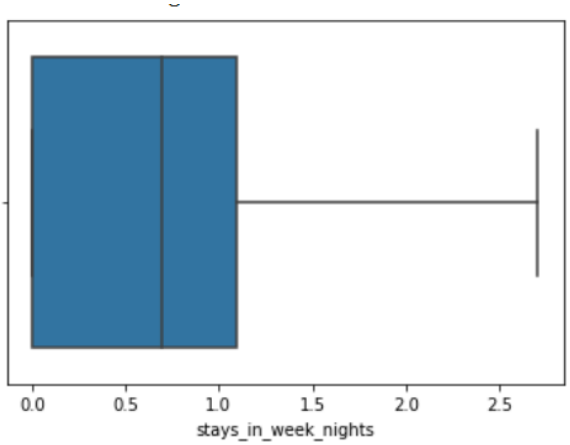

Fig-11: Stays in weeknights Box Plot

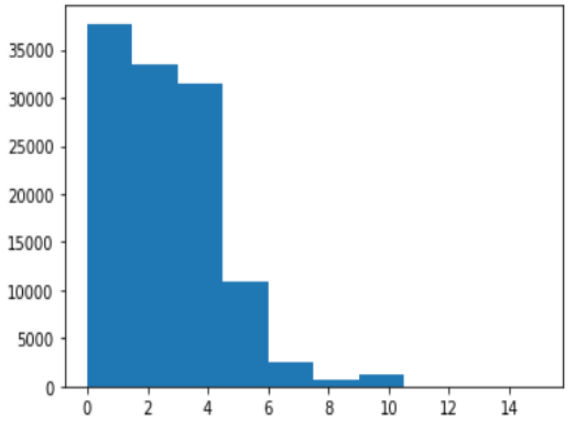

Fig-12: Stays in weeknights Distribution

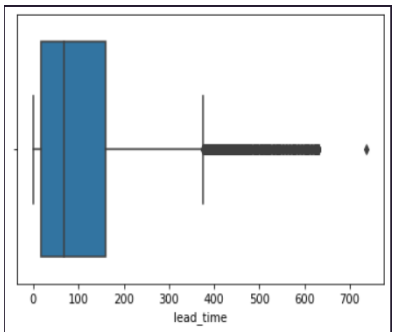

Fig-13: Lead Time Box Plot

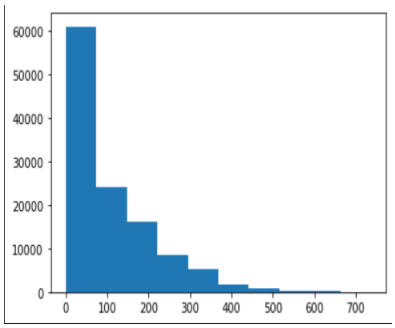

Fig-14: Lead Time Distribution

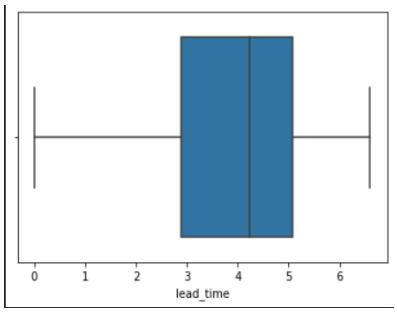

Fig-15: Lead Time Logarithmic Transformation

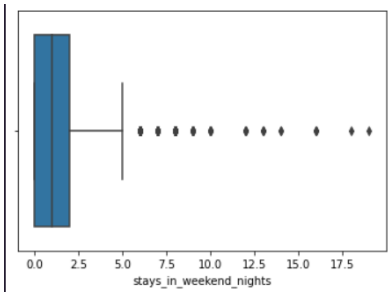

Fig-16: stays in weekend nights Box Plot

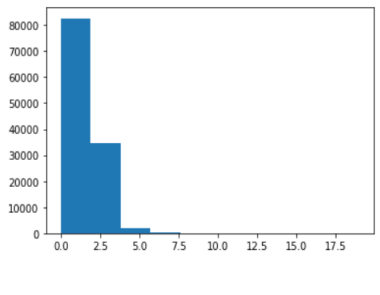

Fig-17: Stays in weekend Nights Distribution

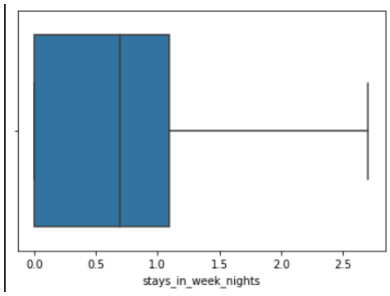

Fig-18: Stays in weekend Nights Outliers Elimination

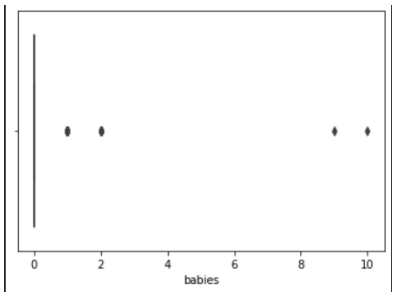

Fig-19: Babies Box Plot

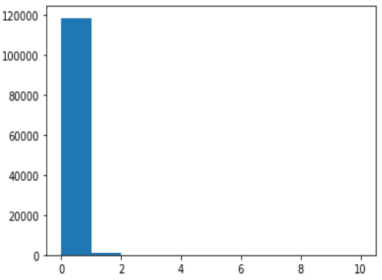

Fig-20: Babies Distribution

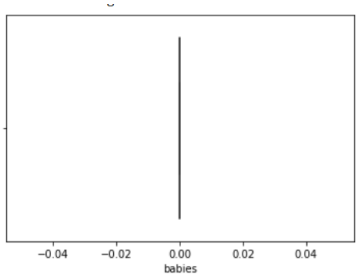

Fig-21: Babies Outliers Elimination.

VII. Comparison of Various Classification Algorithms

For predicting the values and accuracy the techniques used are decision tree classifier, random forest, Adaboost classifier, and gradient boost classifier. A Decision Tree is a simple representation of classifying illustrations. It may be Administered Machine Learning where the information is persistently part agreeing to a certain parameter. In this strategy, a set of preparing cases is broken down into smaller and smaller subsets whereas at the same time a related decision tree gets incrementally created. At the conclusion of the learning handle, a choice tree covering the preparing set is returned. The key thought is to utilize a choice tree to parcel the information space into cluster (or thick) districts and empty (or inadequate) regions.

Random forest classifier makes a set of choice trees from the subjectively chosen subset of the planning set. It at that point aggregates the votes from differing choice trees to select the extreme lesson of the test challenge. As the title recommends, Arbitrary Woodland may well be a classifier that contains a number of choice trees on distinctive subsets of the given dataset and takes the ordinary to form strides the prescient accuracy of that dataset.

After a long time, boosting calculations picked up enormous ubiquity in data science or machine learning competitions. Ada-boost or Flexible Boosting is an equip boosting classifier. It combines distinctive classifiers to expand the precision of classifiers. AdaBoost is an iterative gathering technique. AdaBoost classifier builds a strong classifier by combining various

ineffectively performing classifiers so simply basically will get a tall accuracy strong classifier. The basic concept behind Ada boost is to set the weights of classifiers and prepare the data test in each accentuation such that it ensures the precise figures of unordinary perception.

In Gradient Boosting, each marker tries to advance on its forerunner by reducing the botches. But the captivating thought behind Angle Boosting is that instead of fitting a pointer on the data at each cycle, it truly fits an unused marker to the remaining botches made by the past predictor. The learning rate might be a hyperparameter that's utilized to scale each tree's commitment, giving up the inclination for prevalent change. In other words, we copy this number by the expected regard so that we do not overfit the data.

From the Statistical Analysis of the features, we have observed that there are a lot of outliers in the features. Most of them are due to the IQR being 0. So clearly we can expect that linear algorithms which are vulnerable to outliers are not going to perform as expected. We have compared Decision Tree, Random Forest, Adaboost Classifier, and GradientBoost Classifier.

Random Forest is a combination of randomly generated decision trees whose results are combined to generate a final probability. Adaboost and GradientBoost are specific types of Linear Regression Classifiers with either weight normalization or Gradient Normalization to prevent overfitting of the model and generally increase the accuracy of the model.

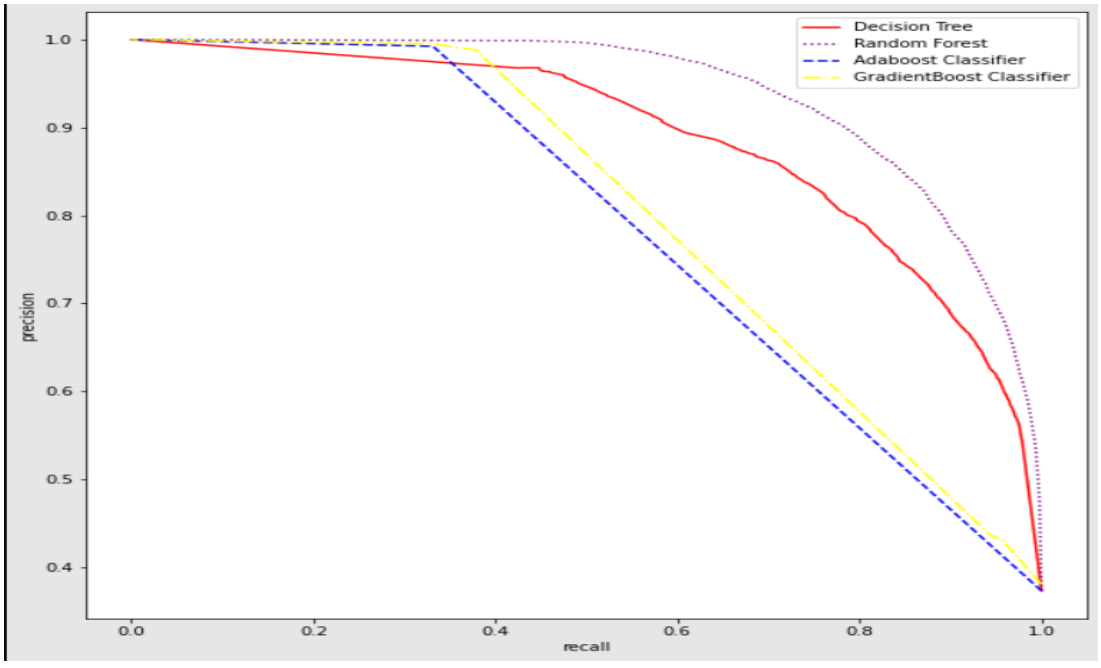

Fig-22: Precision-Recall Curve for chosen algorithms

Model	Accuracy	AUC	F1-Score

Decision Tree	85.8	0.8875	0.791
Random Forest	90.1	0.9390	0.843
AdaBoost Classifier	75.9	0.7865	0.4968
Gradient Boost Classifier	63.6	0.803	0.4423

Table-1: Comparison of Chosen Classifier Algorithms.

From table 1 the algorithms can be compared based on the F1 Score and AUC(Area Under Curve). F1 Score is the harmonic mean of Precision and Recall. We have observed that on a data split of 80%, the random forest has the highest testing accuracy of 90%. Based on Both AUC and F1 Scores, the random forest model has outperformed the other models. The data for Random Forest and Decision Trees are transformed accordingly. The categorical data is label encoded. Similarly, for Linear Regression-Based Algorithms we have eliminated some outliers and also the categorical data is encoded using one-hot encoding.

Based on the feature importance from the Random Forest Model, we can identify the most important features and we can also compare and improve those features.

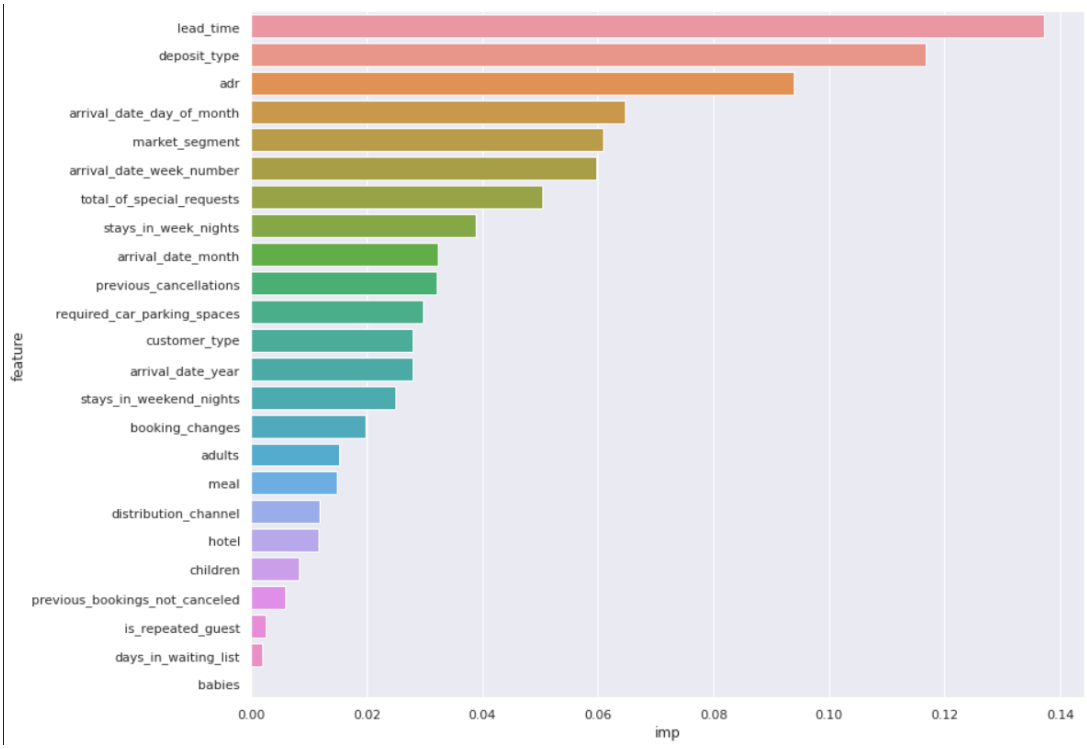

Fig-23: Feature Importance for Random Forest Model

From figure 37 it can be found that lead_time is the most important factor in predicting booking cancellation. So having a low lead_time can decrease the cancellation rating. The deposit_type is the second most important factor. The three types of deposit types accepted by the hotels are No Deposit, Refundable, Non-Refund. From above it can be observed that the cancellation rate is

low for the Refundable deposit type. The ADR is the third most important feature. ADR is the average daily rate. It is the ratio of the sum of lodging transactions by the total number of staying nights. In general, having low staying nights and high lodging transactions increases the ADR.

VIII. Conclusion

Booking cancellations is economically and reputation wise damaging to the hotels, but cancellations are inevitable, but if we can predict to an extent, what would be the likelihood of booking being canceled and integrate this function into the booking system, it would be able to schedule the booking smartly. In the data set used in this project, customers' previous visits and previous cancellations are also included, which gave the algorithm the power to be contextfull. In this paper we have compared four algorithms and Random forest was chosen as the best fitted model based on its AUC for precision recall curve, F score and testing accuracy. Based on this mode, we have identified that lead_time and deposite_type are the most influencing parameters of the booking cancellations.ADR(Average Deposit Rate) must be balanced for efficient booking scheduling.

IX. References

[1] Z. A. Andriawan et al., "Prediction of Hotel Booking Cancellation using CRISP-DM," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020, pp. 1-6, DOI: 10.1109/ICICoS51170.2020.9299011.

[2] N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 1049-1054, DOI: 10.1109/ICMLA.2017.00-11.

[3] Antonio, Nuno & De Almeida, Ana & Nunes, Luís. (2017). Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue. Tourism and Management Studies. 13. 25-39. 10.18089/tms.2017.13203.

[4]Hotel reservation cancellations: analysis and prediction using machine learning algorithms Jasmina Novakovic1, Snezana Turina2.International Academic Journal, Vol. 2, Issue 1. 2021

[5] Modeling the cancellation behavior of hotel guests Falk, Martin; Vieru, Markku Published in INTERNATIONAL JOURNAL OF CONTEMPORARY HOSPITALITY MANAGEMENT DOI: 10.1108/IJCHM-08-2017-0509 Published: 08.10.2018