View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by International Hellenic University: IHU Open Access Repository

Anomaly Detection: Predicting hotel booking cancellations

Christos Timamopoulos

SID: 3308180021

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of
Master of Science (MSc) in Data Science

January 2020
THESSALONIKI – GREECE

Anomaly Detection: Predicting hotel booking cancellations

Christos Timamopoulos

SID: 3308180021

Supervisor: Prof. Apostolos Papadopoulos

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of
Master of Science (MSc) in Data Science

January 2020
THESSALONIKI – GREECE

Table of Contents

ABSTRACT

Booking cancellations have a momentous impact on the hospitality industry, as regards to demand management. In order to diminish the influence of cancellations, hotels apply severe cancellation policies and tactics, that may have negative results on the hotel's prestige and therefore its revenue. To minimize the impact of booking cancellations and improve the functionality of the hotel, a machine learning based model was developed. By using a dataset of a 4-stars hotel and approaching cancellation prediction as a supervised anomaly detection concept, it is exhibited that it is possible to develop a predicting machine learning model to forecast booking cancellations with overall accuracy 99%. The results of the research give the opportunity to the hotel manager to accurately predict demand through cancellations, produce improved forecasts and define better overbooking strategies.

Keywords: Booking cancellations, anomaly detection, literature review, hospitality industry, machine learning, rare class mining.

Chapter 1. Introduction

In the hospitality industry, the cancelation of bookings has a big influence on decisions regarding demand management. Booking cancelations have a direct effect on the output of accurate estimates, which is a crucial fact in terms of revenue management. In order to minimize the impact of cancelations, hotels apply strict cancelation policies and overbooking tactics that ultimately adversely affect the hotel's revenue and reputation. On the other hand, overbooking allows hotels to question the quality of service, which can lead to bad experience for the guest and have a negative impact on both the revenue and the credibility of the hotel.

In this study, we partnered with a 4-stars hotel resort to identify the characteristics of customers that are most probably to cancel their booking reservation. These kinds of concepts are usually resource-intensive and demand high-cost installation projects and such unforeseen cancellations hold a great risk to the partner company - hotel. It is important to emphasize that, to please any type of customer, hotels have a limited number of rooms and sell a perishable product. Customers who stay in suits could be considered as highly demanding, while customers who stay in double-rooms could be considered less demanding. By canceling a reservation, the hotel does not only lose a customer but also may have denied that specific room from another possible customer suited for that room. Bookings are indicative of a customer-hotel relationship (Talluri et al., 2004). By this reason, consumers have the right to use or cancel the service in the future before the service is provided. Even though previous bookings are considered the biggest predictor of the results forecast for a hotel (Smith et al., 2015), the option to cancel the reservation set at great risk the hotel itself. The hotel guarantees that rooms will be available to the customers, but simultaneously, it has to carry the cost of vacant capacity, in a scenario where, the customer cancels a reservation or doesn’t even show up (Talluri & Van Ryzin, 2004).

The development of a predictive booking cancelation model is therefore a high priority, with regard to what Chiang et al. (2007) also pointed out that revenue management could enterprise with mathematics and predictive models in order to make use of existing data and technology.

The aim of this study is, by the perspective of anomaly detection concept to build a machine learning model to identify the bookings that have high probability to be canceled, so our partner business can take measures to secure the bookings. Successful bookings are also identified as well as the cancellations, in order to take a step forward and also improve the sales pipeline. In addition to that, this paper covers an extensive literature work, regarding related work on booking cancellation and anomaly detection research. It also identifies the features of a hotel customer database that contribute to predicting booking cancellations. Finally, it demonstrates how data science can be applied to forecast cancelation of bookings within the context of hotel demand and revenue management.

Chapter 2. Literature Review

2.1 History of booking cancellation prediction

Mehrotra (2006) noted that precise demand forecasting is a key determinant of revenue management. Talluri (2004) have recognized the importance of revenue management forecasting by confirming that revenue management systems need a quantity forecast, and more precisely, " its performance depends critically on the quality of these forecasts". In addition to that, other authors like Morales & Wang (2010) and Ivanov & Zhechev (2012) acknowledged the crucial role of demand forecast where forecasting is crucial. Because of the need for predicted demand, the cancelation of reservations, as in the hospitality industry and other service industries that deal with advanced reservations, do not show the true demand for their services, as there are often a insignificant number of cancellations (Morales et al., 2010).

The cancelation of bookings is a well-known issue in the revenue management sector related to the service industries, and especially to the hospitality industry. With the growing effect of the internet on the way consumers search and purchase travel services in recent years (Noone & Lee, 2010), researches in this topic have been increased, and particularly on the subject of controls used to mitigate the effects of cancelations on revenue allocation, cancelation policies and overbooking (Ivanov, 2014; Talluri et al., 2004). It is important to mention that, in the hospitality industry, there is only a few literature on the booking cancelation forecast market. Among the related literature is the work of Huang et al. (2013), who in their case used data from restaurants. Another related work is Liu's one (2004), which used real data about hotels. Every other case uses Personal Name Record Data (PNR) which is a standard established by the International Civil Aviation Organization (2010). The use of PNR data is not an uncommon approach since work on cancelation forecasting is mostly available in Yoon et al., 2012; Lemke et al., 2009; Iliescu et al., 2008; Gorin et al., 2006)

The dominance of the airline industry in the booking cancellation forecast can be explained, as not only an extended operation of revenue management, but also has quite a high rate of cancellations on airline bookings, which indicates the 30 percent (Phillips, 2005) to 50 percent (Talluri et al., 2004) of all bookings. Although travel and hospitality are both service industries and can have many similarities, there are a few key points that distinguish them, which is the aspect that lures consumers to select their service providers. In hospitality industry key factors are the price, social reputation, quality of service, cleanliness, location, accessibility to transport hubs, while in airline industry the importance of the above factors changes and there may be others, like company's profile, safety reputation and loyalty programs (Chen et al., 2008; Park et al., 2006).

From the data science perspective, and especially in the machine learning field, supervised predictive modelling projects are usually divided into two categories (Hastie et al., 2001). The first category is called regression, and it simply defines the conditions in which quantitative outcomes are evaluated. For example, the prediction of reservations cancellation percentage of a company's total bookings. The second category called classification is considered when the outcome is a class or category. For example, the prediction of the possibility that a specific reservation "will be cancelled" or "will not be cancelled".

Even though, some of the published researches on prediction of booking cancellations consider it a classification question, most researchers approach it as a regression problem. Actually, Morales (2010) describes that "it is hard to imagine that one can predict whether a booking will be cancelled or not with high accuracy simply by looking at PNR information". Although, it is suggested in the next chapters that the classification of whether a room reservation will be cancelled is feasible. An additional reason to examine as a classification issue booking cancellation is that, from the class prediction results, it is possible to achieve quantitative results. For instance, the calculation of booking cancellation rate can be done by dividing the total number of bookings predicted as cancelled by the sum of bookings for that specific period of time.

According to Ivanov (2014), the registration of cancellations is an important factor for recognizing data trends and thus creating better forecasts, overbooking and cancellation policies. Talluri and Ryzin (2004) consider overbooking one of the most successful revenue management practices. Over the last years, some authors suggested rigid cancellations policies as effective tools toward cancellations, like financial penalties or payment in advance during the booking process (DeKay et al., 2004). At the present, these kinds of measures may have a negative effect on sales and revenue, as they are considered as sales inhibitor (Smith et al., 2015)

Sales forecasting is generally a complex process, as there are numerous phases and there are several participants at each phase. Buyers and sellers, for example, may not have the same goals and interests. Therefore, the sales forecast is a key factor in making managerial decisions as well, and inaccurate forecasts will result in great resource losses (Bohanec et al., 2017).

Customer cancellation is a classification concept in which machine learning techniques can be applied to enhance the accuracy of the predictions that a company can make of the concept if a customer cancels his or her reservation (Huang et al. 2013). Therefore, participants such as stakeholders and policy-makers are not only interested in the accuracy of the classification models, but need these studies as evidence to reinforce their opinions in decision-making situations. Thus, the interpretability of a prediction model is also a key factor, along with its accuracy (Bohanec et al., 2017). For this reason, while more sophisticated models, like SVMs and ensembled boosting methods, may indicate stronger predictive models, though they lack interpretability, such as of Logistic Regression, Nearest Neighbors models and Decision Trees (Caruana & Niculescu, 2006)

According to Kotsiantis (2007), it is important for a particular concept to be fully understand of the conditions under which a model can theoretically outperform the others. In customer cancellation, there is a rare limitation as the data is usually imbalanced, and for this reason, cases like these are approached as anomaly detection concepts. In general, an extremely low percentage of customers belong to that class and the minority class is usually the one that we are interested to predict (Zhao et al.,2005). Certain common examples alongside customer cancellation include fraud detection, intrusion detection and rare disease diagnosis (Chandola et al., 2007).

Although most of the classification models, according to Chen et al. (2004), are meant to minimize the overall error and not focus on the minority class

Thus, two main approaches are used, in order to overcome the issue of imbalanced data, the resampling techniques and the cost-sensitive learning, which assigns high costs to misclassified instances.

Chawla et al. (2002) developed a well-known resampling technique named Synthetic Minority Over-sampling Technique (SMOTE). Commonly, in the oversampling technique, the minority class is over-sampled with replacement by random data points. However, in the SMOTE approach, the minority class is oversampled, based on its k-nearest neighbours, by creating new synthetic samples. In this way, the information is increased, along with the minority samples weight.

		Actual Class	
		Actual Positive	Actual Negative
Predicted Class	Classified Positive	TP	FP
	Classified Negative	FN	TN

Table 1: confusion matrix of a binary classification concept

Regarding data with class imbalance, Tang et.al (2009) discovered that overall accuracy is not the optimal model evaluation metric, as it is not capable to depict the misclassifications of rare positive samples, and also, grants the model a high total accuracy when all samples are predicted as negative. Thus, they introduced the use of Precision and Recall as the most appropriate ones.

With regard to class imbalance results, Tang et.al (2009) found that overall accuracy is not the optimum model evaluation metric, as it is not capable of depicting misclassifications of rare positive samples, and also provides the model with a high overall accuracy when all samples are expected to be negative. Thus, he introduced the use of Precision and Recall as the most appropriate ones. The table 1 depicts a

confusion matrix of a binary classification concept. True Positive (TP) and True Negative (TN) indicate that the actual class and predicted one are the same, while False Negative (FN) and False Positive (FP) indicates that negative and positive classes were misclassified. Accuracy, Precision and Recall (Larose, 2015) are described in detail in table 2.

Evaluation Metric	Formula	Description
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	Measures the proportion of true results among the total number of predictions
Precision	$\frac{(TP)}{(TP+FP)}$	Measures the proportion of True Positives against the sum of all positive predictions
Recall	$\frac{(TP)}{(TP+FN)}$	Measure of relevant predictions that are retrieved.

Table 2: Accuracy, Precision and Recall scores

2.2 Types of Anomalies

In the section, we look at some of the most basic and popular forms of Anomalies. Anomalies or unusual events can be categorized according to a number of parameters. Anomalies can be divided primarily into three groups according to Chandola et al. (2009), depending on the nature and viewpoint of anomalies.

2.2.1 Point anomalies

A point anomaly, also referred to as global anomaly, is observed when a data point shows a behavior which is different from that of the entire data set. Despite the fact that it is the easiest type of anomaly to be observed, the calculation that is chosen to deviate one point from the rest of the points is still a big problem. Hypothetically let's assume each node must have at least two "neighbors" nodes connected to it for a regular network. As illustrated in Figure 1, the nodes that compose the first group ‘V1’ are isolated points, while the second group ‘V2’ contains nodes that communicate with at least two neighbors. Thus, it can be assumed that group V2 represents a normal behavior, and on the contrary, group V1 represents an abnormal behavior.

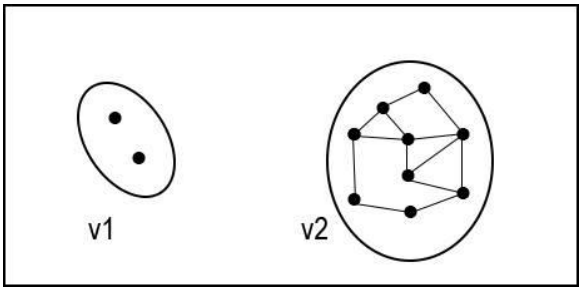

Figure 1. Point anomalies

2.2.2 Contextual anomalies

A contextual anomaly, also known as a conditional anomaly, occurs when a data instance diverges greatly from the rest of the data, with respect to a particular context. For example, if a temperature of 25 °C can be characterized as anomalous, it heavily depends on the time and location of the sampling. It can be considered as an anomaly if the sampling took place in winter in Greece. However, this temperature is completely normal in summers in Greece, so no phenomenon can be believed.

In the process of detecting contextual anomalies, there are two data instance attributes which define the entire data set:

- Contextual attributes: Those are the attributes which define the instance context. For example, the date and location in the above climatic example are the contextual attributes of the data instance.

- Behavior attributes: Generally, such characteristics describe an instance in such a way as to make it easier to identify its anomalous existence with respect to its meaning. In the above climatic example, attributes like degrees of temperature, wind, pressure and humidity could be characterized as behavior attributes

Based on Chandola et al. (2009), proximity based methods are usually used for contextual anomaly detection, as data instances may vary on their nature of whether they are anomalies or not, regarding a specific context each time.

2.2.3 Collective anomalies

Collective anomalies are detected when a collection of relevant data instances, within a data set, is anomalous with respect to the entire data set, but the values of the individual instances are not abnormal by themselves, in either a global or contextual perspective. A real-world scenario could be the cancellation of flights, as it may be considered as normal the cancellation of a flight in a time period of twelve hours, but if multiple flights start canceling one after the other, then as a complete group they are considered as outliers. Equivalently, in Figure 2 the group G of data instances denotes a collective anomaly regarding its density. The density of G group is unusually high with respect to the others, though each individual data instance that belongs to group G is not an outlier with respect to the other participants of the same group.

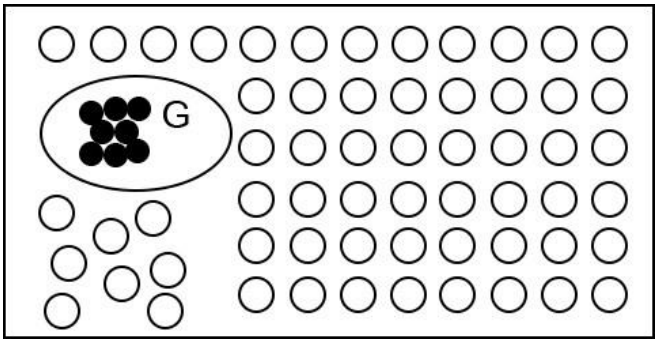

Figure 2. Collective anomalies.

2.3 Anomaly detection techniques

Anomaly detection is described as " an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." (Hawkins et al., 1980).

Chandola et al. (2009) indicate that, with regard to label availability, anomaly detection can be classified into three categories, as follows:

- Supervised methods
- Semi-supervised methods
- Unsupervised methods

The usage of the above methods depends highly on the availability of whether an expert labeled the data instances as normal and abnormal.

2.3.1 Supervised methods

Supervised methods approach anomaly detection as a classification problem with pre-labeled data, described as normal or abnormal. The main aim of these approaches is to allow the classifier to learn as efficiently as possible, and they can be set up in different ways. For example, it can be Support Vector Machine, also known as SVM, based (Ma et. al., 2003) (Ratsch et. al., 2002), Bayesian network based (Box et. al., 1968) (Abraham B et. al., 1979), neural network based (Brotherton T, 1998) (Augusteijn MF, 2002).

Dealing with supervised anomaly detection methods, one should keep in mind that imbalanced class problem arises, as abnormal data instances are quite more rare than normal data instances in a dataset. Specific techniques, such as oversampling, undersampling, or other artificial anomaly methods must therefore be applied (Lemaitre G, 2017). In addition to the above, significant focus should be put on recall metrics during the process of choosing a classification system to identify anomalies. By concentrating on memory, the goal becomes to identify as many anomalies as possible accurately as possible, rather than preventing false positives.

2.3.2 Unsupervised methods

Unsupervised methods are used when there are no pre-existing data labels and hidden patterns need to be found in data set. These kinds of approaches are usually studied as a clustering problem. Unsupervised approaches consider that normal instances form one or more clusters with specific attributes, thus normal instances are expected to follow a particular pattern. On the contrary, anomalies are expected to act in this way, as displayed in Figure 3.

However, the above hypothesis is not constantly true, as in some cases when dealing with collective anomalies, there are anomalous cases which form similarity clusters, as shown in Figure 2. So in this case, when the normal instances are scattered in contrast with anomalous instances, unsupervised methods tend to operate inefficiently, as they fall into the trap of false positives. In General, by dealing with unsupervised anomaly detection methods, two major challenges have arisen. First, an isolated instance of data can be regarded as anomalous, but this statement may typically be incorrect, as a data instance can be noise rather than an anomaly. Second, unsupervised approaches can sometimes be very time-consuming, as they are used to discover the clusters first, and then the anomalies. The key factor in that challenge is that usually the number of normal instances are far more than anomalies instances in a data set.

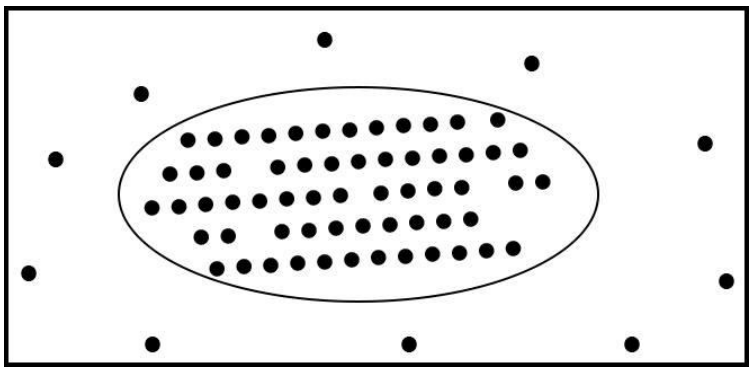

Figure 3: Unsupervised ‘clustering’ approach.

2.3.3 Semi-supervised methods

Semi-supervised methods use two sets of data, usually a small set of data, which are labeled as normal, and a large set of unlabeled data. By using the small set of labeled data, a classifier tries to recognize the unlabeled data, through the deductive derivation. A model is built for the normal data instances, in a way that the instances which don't fit the normal model are labeled as anomalies. This method is called self-training and is considered to be the easiest technique used in a semi-supervised approach. Another well-known approach called co-training and describes how two or more classifiers are deployed to train each other. In contrary to self-training, co-training is less sensitive to errors.

The challenge related to semi-supervised methods is that if the available small set of labeled data represent the anomalous instances, rather than the normal ones, the procedure of detecting every possible anomaly becomes extremely difficult for a model.

2.4 Proximity based anomaly detection

Proximity based anomaly detection techniques analyze and define every data instance as anomalous or normal, "with respect to its neighbors". Normal data instances are believed to have close proximity to their neighbors, because they adopt a trend of density where irregular instances are far away from their closest neighbours. Aggarwal (2013) suggests that proximity-based analytical techniques can primarily be divided into the following two categories: Distance-based techniques and Density techniques.

The following table 3, describes the general advantages and disadvantages of the proximity based anomaly detection methods:

Proximity based anomaly detection methods	
Advantages	*Disadvantages*
Simplest data mining approach.	It becomes difficult to handle and detect phenomena when we have many areas with very different densities.
Applicable to a number of domain.	The group of anomalies is difficult to detect, if they are present near each other.
An simple and straightforward solution is the identification of a distance or density metric, as the only major requirement for such methods.	Methods based on proximity are highly dependent on the proximity measures used for their efficient work which may not be accessible in certain circumstances..

Table 3. Advantages and disadvantages of proximity based anomaly detection methods

2.4.1 Distance based anomaly detection methods

Distance based anomaly detection methods determine anomaly score by using the distance between a data instance and its k neighbours. Anomalies based on distance are known as ' global anomalies. ' Mostly the distance from Mahalanobis, Manhattan or Euclidean is used as the metric distance. According to Aggarwal, while most of the distance based methods are constructed with the use of Euclidean distance, Mahalanobis distance is an excellent choice, “as it is all about the effective statistical normalization, based on the characteristics of a particular data locality” (Aggarwal, 2013). The concept of distance based anomaly detection methods does not consider any

underlying data distribution, as it also generalizes some of the concepts of distribution based methods.

Generally, distance based anomaly detection methods are amongst the most widely accepted and usually used methods in data mining and machine learning, as they totally depend upon the concept of local neighborhood (KNN) of the data points (Jin W, 2006). The above practice can also be described as Nearest Neighbor analysis, and it is applicable for either classification, clustering or most importantly anomaly detection.

A general approach for the method of detection of distance dependent anomalies is defined below. For each and every data instance the neighborhood of an instance is evaluated, calculated by the distance threshold. If an instance's neighborhood, o, loses out significantly on many instances from the whole data set, D, then the given neighborhood is considered an anomaly (Knorr, 2000).

The method quoted above uses two global parameters, d and β. Parameter d determines the maximum possible distance between instances that are part of an instance's neighborhood. To function as an anomalous node, the parameter β specifies the fraction threshold which defines the maximum number of instances that might belong in a neighborhood. As stated in Han J (2012), if d "with $d \geq 0$" is the distance threshold, β "with $0 < \beta \leq 1$" is the fraction threshold and dist (d, b) is the distance factor, then instance' o' is an anomaly if:

$$\frac{\|\{o' | \{\text{dist}(o, o') \leq d\}\|}{\|D\|} \leq \beta$$

Knorr (2000) suggests the nest loop method to be the simplest method for detection of anomalies regarding the distance. In this approach, an inner loop measures the β factor and determines if an instance is usual or anomalous, based on the amount of elements present in the instance's d-neighbourhood. Though this may be the easiest approach, it demands $O(n^2)$ time and it is supposed to be quite costly, especially when each instance is checked one by one, against the whole data set.

The Achilles heel of distance based approaches is that they fail to detect the local anomalies. In order to surpass this issue associated with distance based methods, density based methods are used. Anomaly detection methods based on density use more complex techniques to model data instances abnormality compared to distance-based methods. Such methods work by comparing the density of an instance to the density of its surrounding neighbours. Despite the fact that density based models may be evidence of stronger modeling toward anomaly detection, simultaneously they require quite expensive computations.

Hautamaki et al. (2004) suggested the Outlier Detection using In-degree Number (ODIN), probably the most straightforward density based method. However, Breunig et al. (2000) introduced the Local Outlier Factor (LOF), the most popular density based anomaly detection method thus far. The LOF method is an indirect way of detecting anomalies, and in fact its main idea is that the distance distribution between an instance and all the other instances will behave similarly with regard to the cumulative distance distribution for all the pair distances, if there are many other instances close to each other. The LOF score of an instance is defined as the ratio of the k-neighbors ' local reachability density of instance' o' to its own. The density of local accessibility used in LOF is a factor in the k-nearest neighborhood and the estimate of the distance of accessibility of instance' O.' As far as the anomalous instances are concerned, the LOF score is higher because the relative density of an anomalous node is greater than that of its neighbors, whereas the usual data instances are roughly the same.

The strong point of LOF over the simplest approach of ODIN proposed by Ramaswamy (2000), is that LOF score of a data instance indicates the density contrast between its density and those of its neighborhood (Breunig, 2000). Whether the density of an instance '*x*' is lower or the density of *x*'s neighbor is higher, the LOF score is larger which indicates that *o* has a greater degree of being an anomaly. LOF's weak point is that it fails to detect possible phenomena, the local density of which is very similar to that of its neighbours.

For that reason the Connectivity Based Outlier Factor (COF) method (Tang J, 2002), which improves the efficiency and effectiveness of the LOF approach, especially when the pattern itself has equal density of neighborhood as an anomaly (Zhang J, 2004).In a similar way, Influential Outlier (INFLO) method (Jin W, 2006) also focuses on the different variants of a neighborhood set. In more detail, INFLO employs the reverse k-nearest neighbors set (RkNN) to get all those points, which has instance '*o*' in its neighborhood set.

Long story short, the density based anomaly detection methods are computationally more complicated and therefore more costly than those based on distance. However, at the same time they are way better, as the density based methods analyze simultaneously the local density of the data instance being investigated and the local densities of its nearest neighbors.

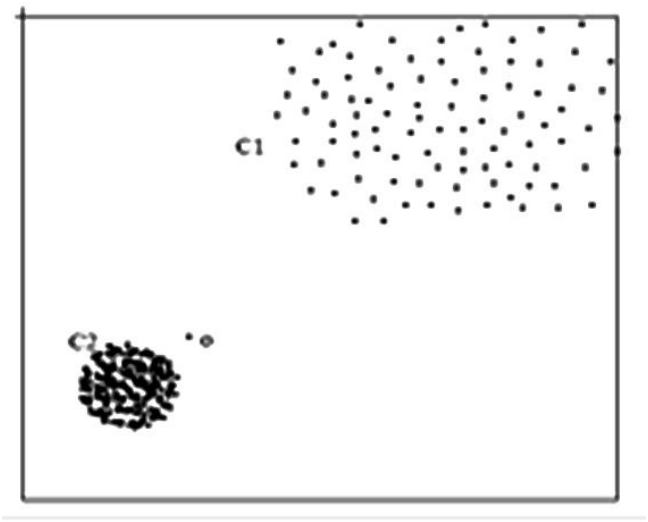

Figure 4.The advantage of LOF over Distance based Methods in anomaly detection (Challagalla et al, 2010)

2.5 Classification based methods

Classification based methods are defined by Han J et al. (2012) as supervised methods that are divided into two essential phases, the learning phase and the classification phase. In the learning phase, a trained data set of labeled instances is used to build up the classification model. This phase is also described as training stage. Subsequently, in the testing stage, the created classification model is used in the classification process to determine the class labels for the dataset. Regarding anomaly detection, the training data instances are classified as normal or anomalous, depending on their behaviour. Classical brute force methods may not be useful and effective in anomaly detection, since the number of anomalous data instances is much smaller than the number of regular data instances. However, specific classification-based methods can be applied either to one class (Moya, 1993) or to multiclass models. Some of the best suited for anomaly detection are discussed in the sections below.

2.5.1 Support Vector Machine

In Support Vector Machines (Cortes C, 1995), a hyperplane is used to distinguish the tuples of different classes from each other. The aim of SVM is to specify and select the greatest separating hyperplane among plenty of them. The Maximum Marginal Hyperplane approach (MMH) is considered one of the most accurate for classification.

Although SVM factions are a two-class model solution, it can also be viewed as a one-class approach by thinking that only a positive dataset is taken as a class, and the observed anomalies are treated as the other. Cortis and Vapnik (1995) used a vector machine model with one-class support to detect anomalous behaviours. Manevitz and Yousef (2002) have used a similar approach to classify various documents expressed in various formats. Another case of one class SVM is Ma and Perkins (2003), where time series novel data evaluated toward abnormalities, which were identified and ranked by providing a level of confidence to each anomaly. Piciarelli et al. (2008) used a one-class SVM clustering method to detect anomalous trajectories that were created by traffic monitoring and video monitoring.

2.5.2 Random Forest

Random Forest is a method of machine learning composed of bagging unpruned decision trees with a random selection of features at each split. At the beginning, Random Forest picks samples of n-tree bootstrap from the initial data, and then, to every bootstrap sample picked, the algorithm creates an unpruned classification tree. From that point on, the class that received most of the votes among all forest trees is used to identify the observation (Breiman, 2001)

2.5.3 Boosting methods

Schapire et al. (1998) proposed Boosting, and it can be described as an ensemble approach that combines multiple classificators. Boosting methods aim to improve performance of a set of weak classifiers into one strong classifier by providing sequential learning of the predictors. The first classifier learns in more detail from the entire dataset, the misclassified data instances are labelled and their weights are increased in order to have a higher probability of being in the posterior predictor training set. The posterior classifier therefore learns from training sets based on the performance of the preceding one. As a result of this approach, different classifiers have the possibility to be specialized in predicting different areas of the same data set (Graczyk et al, 2010).

Gradient Boosting is an ensemble approach that combines weak learners to build a single stronger learner, and typically a decision trees (Brownlee, 2016). The poor model makes a prediction at first, then some successive boosting phases forecast the residuals of error. Subsequently, by using the gradient descent approach, the error residuals are minimized. Specific hyperparameters for this algorithm can tune the individual decision trees or manage the boosting procedure based on the requirements, respectively (Jain, 2016). In addition, Gradient Boosting uses a weighted forecast description to provide a cumulative prediction (Gorman, 2017)

Extreme Gradient Boost, also known as XGBoost, is a decision tree ensemble technique, perceived as one of classification's most effective and efficient method (Chen and Guestrin, 2016). One of the most important strong points of XGBoost is that it handles overfitting, by using a set of parameters to make the model's formula fine tuned, along with making the training stage more resilient to noise.This feature makes XGBoost an ideal approach for anomaly detection cases. In addition to the above, Parameters include the subsample of instances to be used in each decision tree and the subsample of features to be used per decision tree. In the hospitality industry, Antonio, Almeira and Nunes (2017) used the XGBoost tree boosting machine learning model to build a classification model, powerful enough to daily predict cancelation likelihood in a fast pace, by using new data each day, along with past errors in predictions.

The following table 3, demonstrates the general advantages and disadvantages of the classification based anomaly detection methods:

Classification based anomaly detection methods	
Advantages	*Disadvantages*
Fast processing, especially in the testing phase, as a classification model has already been learnt which just needs to be analysed for testing process.	Heavy dependency and reliability on training dataset, which if not properly available may lead to the degradation of performance.
Difficulty in detecting group of anomalies as they occur close to each other.	Difficulty detecting group of anomalies as they occur close to each other.

Table 4. Advantages and disadvantages of classification based anomaly detection methods

Name	*Type*	*Description*
Num.	Numeric	ID of record
Customer	Categorical	Name of customer
ADR	Numeric	Average daily rate
Agent	Categorical	Brand of agent (if booked through an agent)
Reservation_Status	Categorical	Status of reservation (out: *the reservation is completed*, cancelled by guest, double booking, illness, no reason, no show: *guest did not show up*, wrong room type)
Room_Type	Categorical	Room type assigned to booking
Guests	Numerical	Number of guests
Total	Numerical	Income by booking, based on the type of room and number of nights the guests stayed at the hotel
Arrival_ month	Categorical	Month of arrival date
Nights	Numerical	Nights the guests had stayed at the hotel
Year	Numeric	Year of arrival
Arrival_day	Numeric	Day of month of arrival
Canceled	Categorical	Outcome variable: Binary value indicating if the reservation has been cancelled (0: no; 1: yes)
Arrive_weekend	Categorical	Binary value indicating that guest/guests came during the weekend (0: no; 1: yes)
Arrive_dayofweek_name	Categorical	Name of the arrival day (Monday through Sunday)
Departure_dayofweek_name	Categorical	Name of the departure day (Monday through Sunday)
Departure_weekend	Categorical	Binary value indicating if guest's departure was during the weekend (0: no; 1: yes)
Price	Numerical	Income by booking, based on ADR and number of nights the guest/guests were in the hotel

Table 5. Explanation of case study's attributes

The data used in this research collected from a 4-star hotel resort in Greece. The data set consists of 18 attributes and 4.301 instances for a period of time of 2 years that the hotel has been installed an information system to supervise and record its various operating processes. The features of the data set are discussed in table 5 above.

Domain awareness is essential, according to data science literature, in order to select the best attributes and escape any pit of a predictive model.

- The dimensionality curse: The dimensionality curse: high computational costs due to the relation between the amount of data and the high number of predictor variables
- Leakage: Based on the generated variables examining for possible future information leakage. For e.g., the "IsRepeatedGuest" variable acts as a binary predictor of the scenario: if a client has stayed in the hotel again before booking. In this case, the "IsRepeatedGuest" variable should have a value of 1 "yes". Otherwise, this vector will assume the value of 0 "no" in the first booking of that particular guest.
- Correlation: Guyon and Elisseeff (2003) describe this fact as "Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them". Thus, some variables were excluded from the learning phase, as they would be perfectly correlated. In our case, as they are illustrated in figure 6, the assignment of room numbers only take place during the arrival of the guest at the hotel. Therefore, all bookings which are not cancelled have assigned room number, while cancelled bookings do not have one.

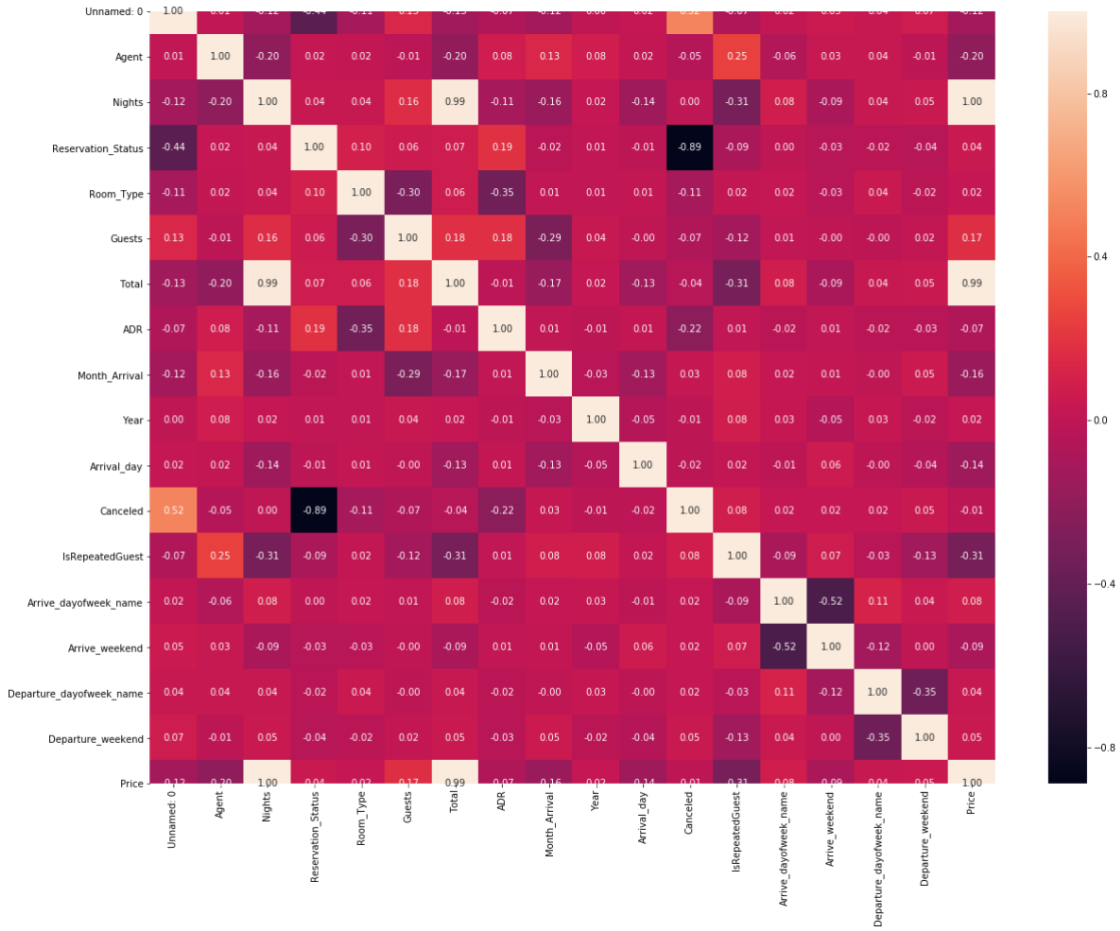

Figure 6. Correlation Matrix

The data sets were explored with the use of Python ‘3.7’ and Microsoft Excel. In the beginning, the first dataset (bookings) had 3796 observations, 18 columns and the second dataset (cancellations) 505 observations, 16 columns. The resulting pre-processed dataset had 4301 observations and 18 columns.

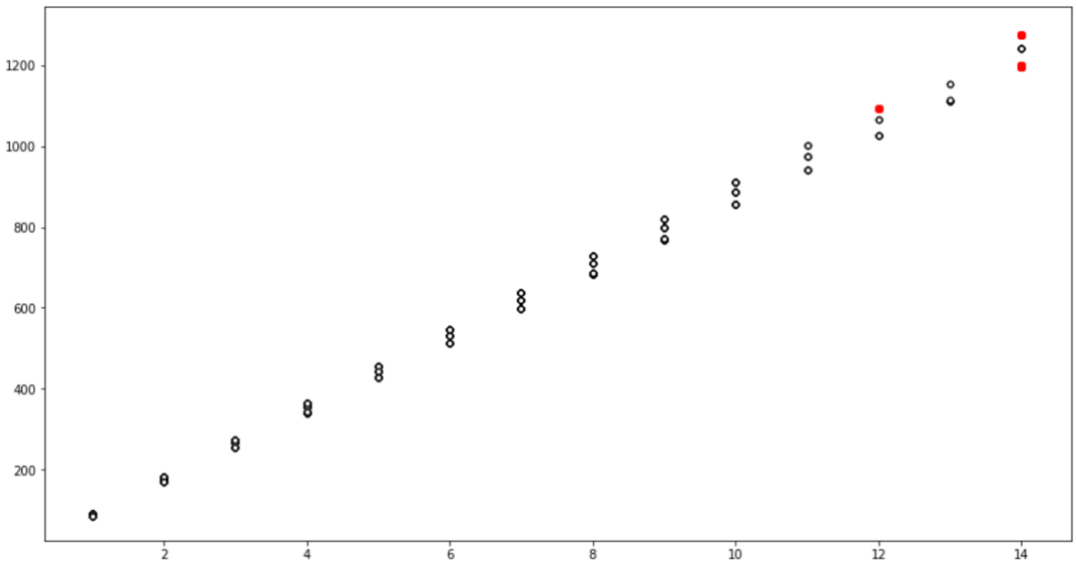

Figure 7. Elliptic Envelope between price income and nights stayed per reservation

During the preprocessing phase, unsupervised anomaly detection methods were performed to locate the possible outliers that could disorient our final model. As it can be observed in figure 7. Elliptic envelope detects outliers among guests who visited the hotel for 12 nights and 14 nights stayed. On the other hand, LOF (Figure 8) identify possible outliers at guests who stayed 9 nights and 11 nights.

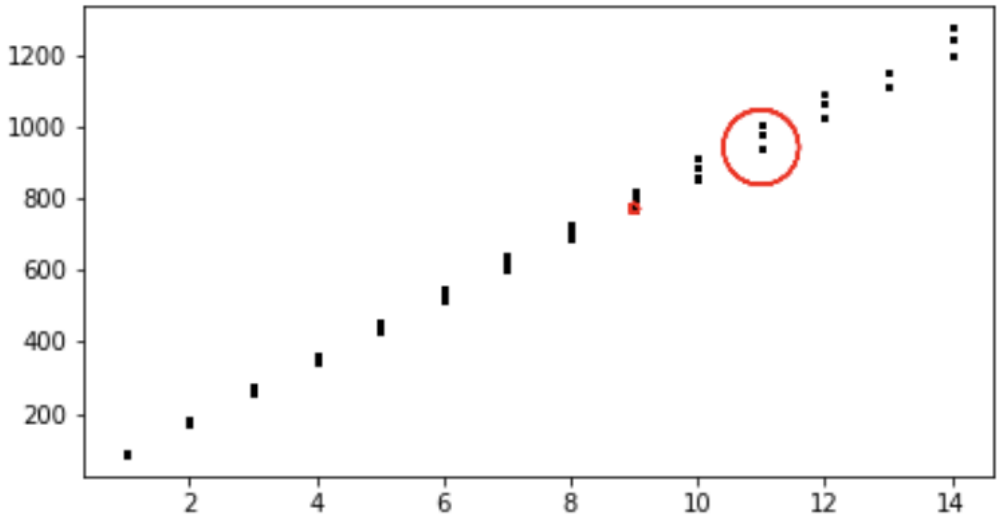

Figure 8. Local Outlier Factor between price income and nights stayed per reservation

According to our data, the lost income due to cancellations is estimated close to 169.097,80 euros in 2018 (268) and 151.411,45 euros (237) in 2019. We can observe a reduction of 11,56% on cancellations in 2019. In figure 9 are demonstrated the amounts of booking and cancellations during 2018 and 2019 season.

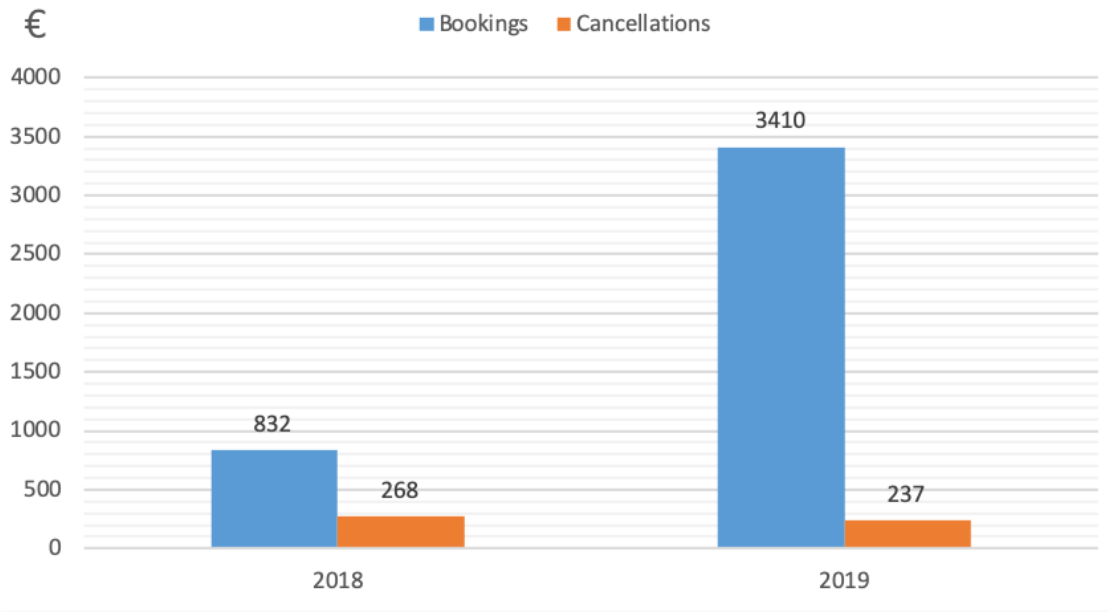

Figure 9: Cancellations & bookings per year

Chapter 4. Methodology

4.1 Data characterization and used methods

As previously mentioned, this paper uses real booking data from a hotel located in the Chalkidiki, one of Greece's most famous resort areas. Data extends from 2018 to 2019. The hotel needed anonymity as planned, so the two data sets, bookings, and cancellations were properly updated and redesigned. Some information on facilities and services are given, in order to better understand the demand of the hotel. Summer months are considered high season, from July through September. During the low season the hotel closes temporarily, from early October to late April.

The study is broken down into 4 specific stages:

- Data exploration
- Pre-processing and data cleaning
- Model building
- Model evaluation and comparison

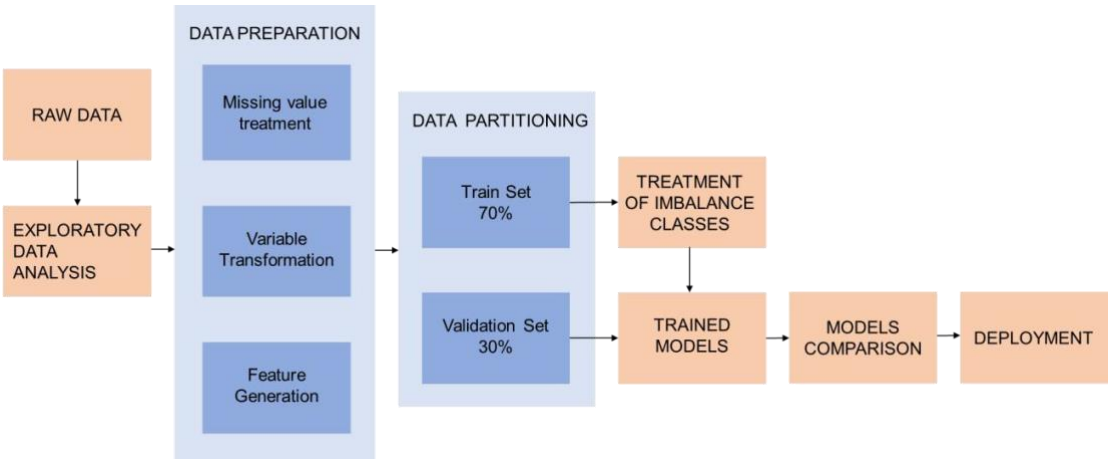

Figure 10: Methodology process flow

4.2 Data understanding

During the first stage of data exploration, we investigated the two data sets toward to some specific patterns:

- The interrelationship between Bookings and Cancellations data sets
- The interrelationship between the features of both data sets
- Distribution and fill-rate of their attributes
- Associations between the predictor ‘cancelled’ and the response variables

4.3 Data preparation

After understanding the data, we performed the following tasks:

- Analysis of the missing values
- Elimination of high correlating features or almost zero variance
- Data validation, in terms of verifying the correctness of data and deal with anomalies
- Feature transformations, including specific encoding and normalization of features data

Feature Engineering: During this phase, additional features were created through the existing ones. The feature generated was inputted directly into the model and evaluated for its importance in the analysis.

Some of the created features:

- Average Daily Rate: ADR metric indicates how much revenue is made per room. ADR is widely used in the hospitality industry and it is perceived as one of the key performance indicators (KPI) in the industry (American Hotel & Lodging Association, 2014). ADR is calculated by taking the average revenue earned from all rooms and by dividing it with the number of rooms occupied in a specific period of time.

  Formula:

$$\text{Average Daily Rate} = \frac{\text{Rooms Revenue Earned}}{\text{Number of Rooms Sold}}$$

- Reservation Status: The reason for canceling a booking reservation. The bookings that weren't canceled hold the value "out"
- Arrival/departure weekend: Indication of whether a guest arrived of left the hotel during weekend.
- Room Type (in cancellations dataset): The attribute Room type was absent in cancellations, as rooms are assigned when the guest arrives in the hotel. Thus, in order to discover the possible room type of a cancelled reservation; total cost of a canceled reservation was divided by the nights and through the combination of that price value, the average price per room type and an IF statement, possible room types were obtained to cancelled reservations.

4.4 Modeling and Evaluation

4.4.1 Treatment of Class Imbalance

The imbalance between the classes that are Cancelled (1) and not Cancelled (0) is approached by using the two following techniques:

a. Over-sampling the minority class
b. Resampling using Synthetic Minority Over-Sampling Technique (SMOTE)

The SMOTE resampling technique is used for oversampling the minority classes in a data set. Particularly, it defines the k-nearest neighbors for the minority class and also generates synthetic minority class instances. In our case, SMOTE resampling approach is applied only on the training set. On the contrary, if SMOTE technique was performed on the entire data, training and testing sets, the data in the train set would include quite a few data from the validation set, thereby, this would definitely inflate the precision and recall of the model.

4.4.2 Models

The study requires to classify the bookings of a hotel as canceled and not canceled, which projects the customer's attitude toward a room's reservation. Jupiter Notebook was the tool used to build the models, which run on Python 3.7.

Different algorithms showed different results, so new models were created using a number of different algorithms, and then we selected the best ones. Considering that the label "canceled" is a binary attribute, the following two-class classification models were selected:

- Logistic Regression
- Naïve Bayes
- K-Nearest Neighbors
- Support Vector Machines
- Decision Tree
- Random Forests
- Gradient Boosting Machines
- Extreme Gradient Boost

In this work, cross-validation was used, in particular k-fold cross-validation, a well-known technique for model evaluation (Hastie et al, 2001). The main objective of k-fold cross-validation is to randomly partition the data of the given sample into subsamples of k size. Although this model evaluation technique may be computationally inefficient, it encourages the structure of models that are not overfitted and can be applied to independent data at the same time (Smola and Vishwanathan, 2010). In our case, we selected 10 folds to divide the data set, a very common number of folds in the applied machine learning field (Smola and Vishwanathan, 2010). For each of the 10 folds, performance measurements of the selected models are calculated, and then the mean score is calculated to determine the global output of each algorithm. Table 6 also contains the mean score by each of the algorithms.

4.4.3 Model Evaluation

As this is a binary classification issue with a class imbalance, Precision and Recall metrics are used as this study's performance metrics. According to Larose et al. (2015) Precision and Recall are defined as follows:

- ***Precision***: Indicates when a model expects an instance to belong in a class, and how often it actually falls within that particular class. For example, 90 percent precision means the used model ranked 90 times correctly out of 100 times that an instance belongs to a certain class.

  Formula: $$\frac{(True\ Positive)}{(True\ Positive + False\ Positive)}$$

- ***Recall***: Indicates the times when an instance is classified actually into a class, and how many times that it is correctly classified into that class. For example, 90 percent recall means that out of 100 instances that are classified into a given project, the model correctly classified the 90 of them into that project.

  Formula: $$\frac{(True\ Positive)}{(True\ Positive + False\ Negative)}$$

Chapter 5. Results

The models were evaluated based on their Accuracy Mean, F1 and scores. Table 6 below displays the models and their accuracy, precision, recall, F1, AUC scores for the data collection of validation:

Algorithm	***Imbalance Technique***	***Mean Score***	***Accuracy***	***Precision***	***Recall***	***F1***	***AUC***
Logistic Regression	*Oversampling*	0.999734	0.990427	0.99	0.99	0.99	1.00
	SMOTE	0.999337	1.00	1.00	1.00	1.00	1.00
Naïve Bayes	*Oversampling*	0.916189	0.806571	0.81	0.81	0.81	0.83
	SMOTE	0.871726	0.892358	0.73	0.80	0.76	0.89
KNN	*Oversampling*	0.998270	0.979513	0.98	0.98	0.98	1.00
	SMOTE	0.997337	0.994684	0.98	0.99	0.99	1.00
SVM	*Oversampling*	0.983769	0.928428	0.94	0.93	0.93	-
	SMOTE	0.991364	0.967441	0.88	0.98	0.93	-
Decision Tree	*Oversampling*	1.00	0.999068	1.00	1.00	1.00	1.00
	SMOTE	1.00	0.999335	1.00	1.00	1.00	1.00
Random Forest	*Oversampling*	1.00	0.950246	0.95	0.95	0.95	1.00
	SMOTE	0.996020	0.994684	0.99	0.98	0.99	1.00
Gradient Boosting	*Oversampling*	1.00	0.999068	1.00	1.00	1.00	1.00
	SMOTE	1.00	0.999335	1.00	1.00	1.00	1.00
XGBoost	*Oversampling*	0.999867	0.998935	1.00	1.00	1.00	1.00
	SMOTE	1.00	0.999335	1.00	1.00	1.00	1.00

Table 6. Results of the different algorithms

Cross-validation scores were auspicious for almost every algorithm. The lowest accuracy mean result was 87.17%, while most algorithms are similar to 99 percent of the mean accuracy score. When AUC is considered as the metric of evaluation, the results are even better, except Naïve Bayes, all the other algorithms reported 100 percent scores, which are regarded excellent values based on Zhu et al. (2010). F1 scores also shown great result, where Decision Tree, Gradient boost and XGBoost achieved the perfect 100% score, while the worst results were denoted by Naïve Bayes and Random Forest.

In terms of accuracy, Decision Tree and Gradient Boost algorithms were the best. In terms of precision, Decision Tree, Gradient Boost, XGBoost were the best algorithms. It is important to highlight that Logistic Regression almost reach the results of others way more sophisticated algorithm, an evidence that Logistic Regression could be called the most efficient algorithm in this case, in terms of energy efficiency and performance. The final models were built with Logistic Regression, Decision Tree, Gradient Boost and XGBoost for a final assessment.

This is done routinely in the creation of predictive models for machine learning, the data set has been divided into two class-conscious subsets. For training, a subset with 70 percent of the total data was used, and another with the remaining 30 percent used for testing the built model. In addition, Tuning was applied for each algorithm during the training process to examine a number of variations of the parameters of each algorithm and analyze the best parameters to be used in each situation. The results of the three algorithms for the test sets are described in Table 7.

Algorithm	***Imbalance Technique***	***Accuracy***	***Execution Time***	***Class***	***Precision***	***Recall***	***F1***
Decision Tree	*SMOTE*	0.999335	*1.4 sec*	*Not Canceled*	1.00	1.00	1.00
				Canceled	1.00	0.99	1.00
Gradient Boost	*SMOTE*	0.999335	*7.1 sec*	*Not Canceled*	1.00	1.00	1.00
				Canceled	1.00	0.99	1.00
XGBoost	*SMOTE*	0.999335	*30.8 sec*	*Not Canceled*	1.00	1.00	1.00
				Canceled	1.00	0.99	1.00

Table 7

Table 7 above provides a description of the results of each selected model:

- All chosen classifiers reached absolute accuracy in the “canceled” and “not canceled” classes, meaning that if a booking is classified as “canceled”, it is extremely likely to be cancelled.
- In same manners, all algorithms recorded high values of Recall and specifically 99%. This fact implies that the 99% of the cancels were predicted out of all the canceled bookings
- According to results, Decision Tree classifier outperformed all the other classifiers, as it achieved the same excellent Accuracy of 99.93%, but also it was almost 5 times faster than the second-best Gradient Boost algorithm

Chapter 6. Discussion - Conclusion

Throughout this study, we created a machine learning model to predict whether a reservation will be completed successfully, or whether the customer will renege on the reservation and cancel the reservation. Through data analytics techniques and the application of unsupervised anomaly detection techniques, it is possible to identify outliers that could reduce the quality of results. In addition, we are capable to understand the features’ predictive relevance, by using data visualizations, along with the implementation of the mutual information filter.

Several classification models are applied for this binary classification concept. All models built reached overall accuracy over 87.1%. This shows that in our situation, the decision tree algorithm, using data sets with the precisely defined attributes, is a great technique for creating predictive models for booking cancellations. These findings also confirm Chiang (2007) statement “as new business models keep on emerging, the old forecasting methods that worked well before may not work well in the future. Facing these challenges, researchers need to continue to develop new and better forecasting methods”.

The imbalance in target class ‘Canceled’ is handled by using SMOTE. We find that, in our case study, simple minority class oversampling, and the more sophisticated SMOTE technique have almost equal results in combination with any of the applied classification model.

The models are evaluated by comparing the overall accuracy, execution time, precision and recall scores of predictions. By deploying the greatest performing model, Decision Tree classifier with SMOTE approach, the hotel can predict the booking cancellations by 99.93% and save 160.254,625 euros annually.

Such models motivate hotel management to take more action on reservation classified as “canceled”. In the same time, these models encourage the development of a more detailed approach to demand and revenue forecasting.

Future work can be applied on the largest hotel chains in the world, like Marriot International, which use big data platforms to collect data from a variety of their operations. One of the main objectives of these platforms can be dynamic pricing automation that contributes in optimizing room prices, which take advantage of global economic factors.

Chapter 7. Limitations of the study

Further research can be done by using data from additional hotels. A greater variety of hotels could lead to a better understanding of the prediction of booking cancelations, and furthermore examine how well do the used classifiers perform in a generalized concept of booking cancellation prediction in hotel industry.

Additionally, research may also use features from external data sources, such as currency exchange rates, competitive intelligence (prices and social reputation) and weather information, to boost model efficiency and measure the impact of those external features on booking cancellations.

Finally, a great limitation could be the security concern, which has intensified by GDPR. Hotels that record personal customers' data have higher responsibilities of safeguarding them. This fact effects the credibility of these kinds of research.

REFERENCES

1. Abraham B, Box GEP. (1979) Bayesian analysis of some outlier problems in time series. Biometrika;66(2):229–36.
2. Aggarwal CC, Yu PS. (2001) Outlier detection for high dimensional data. ACM Sigmod Rec.
3. Ajay Challagalla, S.S. Shivaji, Dhiraj, DVLN Somayajulu, Toms Shaji Mathew, Saurav Tiwari, Syed Sharique A. (2010). "Privacy preserving outlier detection using hierarchical clustering Methods, 34th Annual IEEE Computer Software and Application Conference Workshops.
4. American Hotel & Lodging Association, (2014). Uniform System Of Accounts For The Lodging Industry With Answer Sheet (ahlei) (11th Edition) (ahlei - Hospitality Accounting / Financial Management). Educational Institute.
5. Anderson, C. K. (2012). The impact of social media on lodging performance. Cornell Hospitality Report, 12(15), 4–11.
6. Angiulli F, Pizzuti C. (2002) Fast outlier detection in high dimensional spaces. In: Principles of data mining and knowledge discovery. Springer. p. 15–27.
7. Antonio, N., de Almeida, A. and Nunes, L. (2017). Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).

8. Augusteijn MF, Folkert BA. (2002). Neural network classification and novelty detection. Int J Remote Sens.
9. Barnett V, Lewis T. (1994) Outliers in statistical data, vol. 3. New York:Wiley.
10. Bohanec, M., Robnik-Šikonja, M., Borštnar, M. K. (2017). Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. *Organizacija 50(3)*.
11. Box GEP, Tiao GC. (1968) A Bayesian approach to some outlier problems. Biometrika; 55(1):119–29.
12. Breiman, L. (2001). Random Forests. Machine Learning, 5-32.
13. Breunig MM, Kriegel H-P, Ng RT, Sander J. (2000) LOF: identifying density-based local outliers. ACM Sigmod Rec. p. 93–104.
14. Breuning, M., H-P. Kriegel, R. Ng, and J. Sander. (2000) LOF: Identifying Density-Based Local Outliers. In Proc. of 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, Texas, 93-104.
15. Brotherton T, Johnson T, Chadderdon G. (1998). Classification and novelty detection using linear models and a class depend-entelliptical basis function neural network. In: Proc. of IEEE world congress on computational intelligence on neural networks, vol.2.
16. Brownlee, J. (2016). A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. Retrieved from Machinelearningmastery.com: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
17. Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. 23rd International Conference on Machine Learning. Pittsburg, PA.
18. Chandola V, Banerjee A, Kumar V. (2009) Anomaly detection: a survey. ACM Comput. Surv. 2009.
19. Chandola V, Banerjee A, Kumar V. (2007). Outlier Detection-A Survey, Technical Report, TR 07-017, Department of Computer Science and Engineering, University of Minnesota.
20. Charu C. Aggarwal, Phillip S. Y. (2005) An effective and efficient algorithm for higher dimensional outlier detection

21. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 321-357.
22. Chen, A. H., Peng, N., & Hackley, C. (2008). Evaluating service marketing in airline industry and Its Influence on student passengers' purchasing behavior using Taipei–London route as an example. Journal of Travel & Tourism Marketing, 25(2), 149–160.
23. Chen, T and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 785–794. Available at http://dl.acm.org/citation.cfm?id=2939785 [Last accessed 22 December 2019]. DOI: https:// doi.org/10.1145/2939672.2939785
24. Chiang, W.-C., Chen, J. C., & Xu, X. (2007). An overview of research on revenue management: current issues and future research. International Journal of Revenue Management, 1(1), 97–128.
25. Cortes C, Vapnik V. (1995) Support-vector networks. Mach Learn;20(3):273–97.
26. David MJ. Tax. (2001). One-class classification: concept-learning in the absence of counter-examples. ASCI dissertation series, vol. 65.
27. DeKay, F., Yates, B., & Toh, R. S. (2004). Non-performance penalties in the hotel industry. International Journal of Hospitality Management, 23(3), 273–286.
28. Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics, pp. 1189–1232.
29. Gorin, T., Brunger, W. G., & White, M. M. (2006). No-show forecasting: A blended cost-based, PNR-adjusted approach. Journal of Revenue and Pricing Management, 5(3), 188–206.
30. Gorman, B. (2019). A Kaggle Master Explains Gradient Boosting. Retrieved from Blog.Kaggle.com: http://blog.kaggle.com/a-kaggle-master-explains-gradient-boosting/
31. Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K. (2010). Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS, vol. 5991, pp. 340–350. Springer, Heidelberg.

32. Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics 1969.
33. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157–1182.
34. Han J, Kamber M, Pei J. (2012) Data mining concepts and techniques. 3rd ed. Elsevier.
35. Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics Springer, Berlin.
36. Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics Springer, Berlin. Retrieved from http://statweb.stanford.edu/~tibs/book/preface.ps
37. Hautamaki ki V, Karkkainen I, Franti P. (2004) Outlier detection using k-nearest neighbour graph. ICPR. p. 430–433.
38. Hawkins DM. (1980) Identification of outliers, vol. 11. Springer.
39. Hayes, D. K., & Miller, A. A. (2011). Revenue management for the hospitality industry. Hoboken, NJ, USA: John Wiley & Sons, Inc.
40. Huang, H.-C., Chang, A. Y., Ho, C.-C., & others. (2013). Using artificial neural networks to establish a customer-cancellation prediction model. Przeglad Elektrotechniczny, 89(1b), 178–180.
41. Iliescu, D. C., Garrow, L. A., & Parker, R. A. (2008). A hazard model of US airline passengers' refund and exchange behavior. Transportation Research Part B: Methodological, 42(3), 229–242.
42. International Civil Aviation Organization. (2010). Guidelines on Passenger Name Record (PNR) data. Retrieved December 7, 2019, from:https://www.iata.org/iata/passenger-datatoolkit/assets/doc_library/04-pnr/New%20Doc%209944%201st%20Edition%20PNR.pdf
43. Ivanov, S. (2014). Hotel revenue management: From theory to practice. Varna, Bulgary: Zangador.
44. Ivanov, S., & Zhechev, V. (2012). Hotel revenue management–A critical literature review. Turizam: Znanstveno-Strucnicasopis, 60(2), 175–197.
45. Jain, A. (2016). Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python. Retrieved from Analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

46. Jin W, Tung AKH, Han J, Wang W. (2006). Ranking outliers using symmetric neighborhood relationship. Adv Knowl Discov Data Min., 577–93.
47. Jin, W., A. K. H. Tung, J. Han and W. Wang (2006). Ranking Outliers Using Symmetric Neighborhood Relationship. PAKDD'06, 577-593.
48. John GH. (1995). Robust decision trees: removing outliers from databases. In: Proc of KDD, 174–9.
49. Knorr EM, Ng RT, Tucakov V. (2000). Distance-based outliers: algorithms and applications, 237–53.
50. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, 249 - 268.
51. Larose, D. T., & Larose, C. D. (2015). Data Mining and Predictive Analytics. John Wiley & Sons, Inc.
52. Lemaitre G, Nogueira F, Aridis C. (2017) Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 1-5
53. Lemke, C., Riedel, S., & Gabrys, B. (2009). Dynamic combination of forecasts generated by diversification procedures applied to forecasting of airline cancellations. In IEEE Symposium on Computational Intelligence for Financial Engineering, 2009. CIFEr '09 (pp. 85–91).
54. Li K-L, Huang H-K, Tian S-F, Xu W. (2003). Improving one-class SVM for anomaly detection. Int Conf Mach Learn Cybernetics,5–81.
55. Liu, P. H. (2004). Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In I.Yeoman & U. McMahon-Beattie (Eds.), Revenue management and pricing: Case studies and applications, 91–101
56. Lumini A, Nanni L. (2009). Ensemble of on-line signature matchers based on OverComplete feature generation. Expert Syst Appl, 36(3):5291–6.
57. Ma J, Perkins S. (2003) Time-series novelty detection using one-class support vector machines. In: Proc of the international joint conference on neural network, vol. 3.
58. Manevitz LM, Yousef M. (2002) One-class SVMs for document classification. J Mach Learn Res, 139–54.
59. Mehrotra, R., & Ruttley, J. (2006). Revenue management (seconded.). Washington, DC, USA: American Hotel & Lodging Association (AHLA).

60. Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. European Journal of Operational Research, 202(2), 554–562.
61. Moya MM, Koch MW, Hostetler LD. (1993) One-class classifier networks for target recognition applications. NASA STI/Recon Tech. Rep. N, vol. 93.
62. Noone, B. M., & Lee, C. H. (2010). Hotel overbooking: The effect of overcompensation on customers' reactions to denied service. Journal of Hospitality & Tourism Research, 35(3), 334–357.
63. Park, J.-W., Robertson, R., & Wu, C.-L. (2006). Modelling the Impact of airline service quality and marketing variables on passengers' future behavioural intentions. Transportation Planning and Technology, 29(5), 359–381.
64. Phillips, R. L. (2005). Pricing and revenue optimization. Stanford, CA, USA: Stanford University Press.
65. Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data.ACM Press, 427.
66. Ratsch G, Mika S, Scholkopf B, Muller K. (2002). Constructing boosting algorithms from SVMs: an application to one-class classification. Pattern Anal Mach Intell IEEE Trans, 24(9):1184–99.
67. Savage D, Zhang X, Yu X, Chou P, Wang Q. (2014). Anomaly detection in online social networks. Soc Networks, 39:62–70.
68. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5), 1651–1686.
69. Smith, S. J., Parsa, H. G., Bujisic, M., van der Rest, J.-P. (2015). Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. Journal of Travel & Tourism Marketing, 32(7), 886–906.
70. Talluri, K. T., & Van Ryzin, G. (2004). The theory and practice of revenue management. Boston, MA, USA: Kluwer Academic Publishers.
71. Talluri, K. T., & Van Ryzin, G. (2004). The theory and practice of revenue management. Boston, MA, USA: Kluwer Academic Publishers.
72. Tan, P-N, Steinbach M, Kumar V. (2006) Introduction to Data Mining. Pearson Education, Inc..

73. Tang J, Chen Z, Fu AW-C, Cheung DW. (2002). Enhancing effectiveness of outlier detections for low density patterns. Adv Knowl Discov Data Min, 535–48.
74. Yoon, M. G., Lee, H. Y., & Song, Y. S. (2012). Linear approximation approach for a stochastic seat allocation problem with cancellation & refund policy in airlines. Journal of Air Transport Management, 23, 41–46.
75. Zhang, J., M. Lou, T. W. Ling and H.Wang. (2004). HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data. VLDB Conference.
76. Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. Lecture Notes in Computer Science, pp. 300-306.
77. Zhu, W., Zeng, N., Wang, N., & others. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland, 1–9.
78. Dallemand, J. (2020). Why is Marriott the Big Data analytics leader in hospitality? Blog.datumize.com. Available at: https://blog.datumize.com/big-data-analytics-in-hospitality-marriott-international-case-study#smooth-scroll-top [Accessed 27 Jan. 2020]