ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse143952020.pdf
https://doi.org/10.30534/ijatcse/2020/143952020

The Prediction of Hotel Customer Loyalty using Machine Learning Technique

Youngkeun Choi [1], Jae Won Choi [2]
[1] Associate Professor, Division of Business Administration, College of Business, Sangmyung University Seoul, Korea, penking1@smu.ac.kr
[2] Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA, jxc190057@utdallas.edu

ABSTRACT

The purpose of this study is to find and analyze the prediction of hotel customer loyalty using machine learning technique so that hotel companies can use the study to formulate possible solutions for customer relationship management. The corresponding variable information is drawn from a third-party website, international challenge on the popular internet platform Kaggle (www.kaggle.com). And, this study uses decision tree which is a powerful and popular machine learning algorithm to this date for predicting and classifying big data. For this, this study essentially had two primary approaches. Firstly, this paper intends to understand the role of variables in hotel customer loyalty prediction modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. In these results, first, we can predict their loyalty with a lot of individual factors of hotel customers. Second, for the full model, the accuracy rate is 0.989, which implies that the error rate is 0.011. This study provides some originality and value. First, this study extends the existing literature by empirically examining the combined impact of the variables on hotel customer loyalty prediction modeling. S, this application helps hotel companies to manage the personal records of the customers and also this will make the decision faster if they have the report of the user already with them.

Key words : Tourism industry, Customer loyalty prediction, Customer relationship management, Machine learning, Decision tree, Artificial intelligence.

1. INTRODUCTION

Machine learning is a field of computer science in which patterns of artificial intelligence are identified and computational learning theories are learned [1]. Machine learning is typically a change in systems that perform tasks related to artificial intelligence (AI). These tasks include recognition, analysis, planning, robot control, and prediction. Explore the study and configuration of algorithms that can predict data. Machine learning is used to build programs with adjustment parameters to adapt to initial data and improve functionality. Machine learning is a technology that grows rapidly and works in the human mind. It represents a multistage record and can effectively address the selectivity dilemma [2].

Over the last decades, the fields of tourism, travel, hospitality and leisure have widely recognized the need for a customer-centric approach that primarily values tourists' needs, wants, preferences and requirements as major determinants in travel decisions in order to enhance both consumer satisfaction and the quality and memorability of the tourist experience [3]. Only very recently has an increasing amount of work related to the fields of data science grown to enrich these lines of research.

The increased profit from loyalty comes from reduced marketing costs, increased sales and reduced operational costs. Loyal customers are less likely to switch because of price and they make more purchases than similar non-loyal customers. Loyal customers will also help promote your hotel. They will provide strong word-of-mouth, create business referrals, provide references, and serve on advisory boards. Raman [4] states, loyal customers serve as a ``fantastic marketing force'' by providing recommendations and spreading positive word-of-mouth; those partnerships like activities are the best available advertising a company can get. Loyal customers increase sales by purchasing a wider variety of the hotel's products and by making more frequent purchases. Bowen and Shoemaker [5] found loyal hotel customers had higher food and beverage purchases than non-loyal customers. Finally, loyal customers cost less to serve, in part because they know the product and require less information

The purpose of this study is to find and analyze the prediction of hotel customer loyalty so that hotel companies can use the study to formulate possible solutions for customer relationship management. The methodology of the proactive approach and modeling techniques used in this paper can also be considered a roadmap for the reader to follow the steps

taken in this study and to apply a procedure to identify the causes of many other problems. The aim of this paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the hotel. The hotel customer loyalty prediction system can can automatically calculate the weight of each features taking part in decision making and on new test data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loyalty can be sanctioned or not. Hotel loyalty prediction system allows jumping to specific application so that it can be check on priority basis.

2. RELATED STUDY

2.1 Tourism industry and data science

Any tourism company (be it a hotel or an airline) needs to leverage its managerial and marketing strategies, tactics, and tools to achieve and maintain sustained competitive advantage. This is more critical in the current highly dynamic economic environment where competition is fierce and consumers are demanding and experienced. Increasingly, it is evident that it is extremely difficult, even for well-established companies, to cultivate and sustain a competitive advantage for a long period. We are going through an age of "temporary advantage" and "hyper competition" [6], where organizations need constant innovation to gain a temporary benefit and move ahead of the competition for a continued series of time periods [7].

In this context, big data can make a difference for the data science of tourism companies, help them make better strategic and tactical decisions, and create value. This is the reason why big data is increasingly a crucial component of the wider data science umbrella. However, research on the role of big data for data science in the hospitality and tourism literature is still scant and highly fragmented. Single research activities often take place in a rather isolated manner and tackle a very specific aspect or research question without looking at the whole picture and embedding new work into the overall scholarly and practical context. Such research practices, however, are common during the emergence of new research areas or phenomena. Therefore, in the current phase of development of hospitality and tourism research leveraging big data and data science, it is important and even overdue, to provide a clear overview of the different facets and issues of the wide research domain of data science and identify, discuss and integrate existing research activities leveraging big data into the overall context of the focal research domain. This is important in particular for two reasons. First, to stimulate but also to systematize further research activities. Second, to provide informational bases and overview on current application areas and utilization potentials for companies and stakeholders in the tourism domain.

2.2 Hotel industry and customer royalty

Hotel industry is based on those businesses which lead to profitability only because of their customer service, customer loyalty and customer satisfaction. Like other industries, this industry has also deviated from its traditional way of doing business and is actually now becoming more customer-focused and aimed at developing a positive and satisfying relationship with their customers. There is huge interference of Information technology and it underpins the reason for efficiency, meeting customer expectation and accuracy in services. Information technology is the most significant factor that is helping management of hotel industry in lowering their cost, time accuracy and increasing the operational excellences

Customer loyalty is defined as commitment toward preferred products or services. Customer loyalty is established by supporting a particular organization, regularly patronizing a certain provider, and increasing the frequency of purchases [8]. Loyal customers impact the profitability and overall success of the organization in three distinctive ways: (a) repeat purchases of products or services generate income for the organization, (b) the cost of marketing, advertising and operations are reduced, and (c) the spreading of favorable news and recommendations of services to others. Based on the importance and benefits of customer loyalty, many service organizations, especially hotels, allocate substantial resources to measure and monitor customer satisfaction, corporate image, and service quality. In addition, customer loyalty also has been viewed as the main tool to retain existing customers; retaining customer loyalty costs substantially less to maintain compared to acquiring new customers. Hence, the recommended approach to measure customer loyalty in hotels is based on the context of repeat purchase intention, price sensitivity and recommendations to friends and relatives.

3. METHODOLOGY

3.1 Dataset

The corresponding variable information is drawn from a third-party website, international challenge on the popular internet platform Kaggle (www.kaggle.com), which provides data in the title of 'Hotel booking demand' that was uploaded by Jesse Mostipak. This data set contains a single file which compares various booking information between two hotels: a city hotel and a resort hotel. The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020. This data set contains the booking information of 119,386 customers for a city hotel and a resort hotel, and includes information such as when the booking was made, length of

stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data. This data set is ideal for anyone looking to practice their exploratory data analysis (EDA) or get started in building predictive models. The Kaggle asked participants to predict hotel booking demand. To help with algorithmic development, the organizers provided the types of a data stream for a large set of individual factors. These variables are listed and defined in Table 1.

Table 1: The variables in the dataset

Categories	Variables	Measurement
Individual factors	hotel	H1 = Resort Hotel or H2 = City Hotel
	is_canceled	Value indicating if the booking was canceled (1) or not (0)
	lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
	arrival_date_year	Year of arrival date
	arrival_date_month	Month of arrival date
	arrival_date_week_number	Week number of year for arrival date
	arrival_date_day_of_month	Day of arrival date
	stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
	stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
	adults	Number of adults
	children	Number of children
	babies	Number of babies
	meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
	country	Country of origin. Categories are represented in the ISO 3155–3:2013 format
	market_segment	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
	distribution_channel	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
	previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
	previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking
	reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons.
	assigned_room_type	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
	booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
	deposit_type	Indication on if the customer made a deposit

		to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
	days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
	customer_type	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
	adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
	required_car_parking_spaces	Number of car parking spaces required by the customer
	total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor)
Customer loyalty	is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0)

3.2 Decision tree

Among various analysis techniques, decision tree (DT) is a powerful and popular machine learning algorithm to this date for predicting and classifying big data [9]. They are used for both classification and regression problems. Now a question might arise why we are willing to use DT classifier over other classifiers. To answer that question, we can bring about two reasons. One being, decision trees often try to mimic the same way the human brain thinks, so it is quite simple to understand the data and come to some good conclusions or interpretations. The second reason can be, Decision trees allow us to see the logic for the data to interpret rather than being a black box algorithm like SVM, NN, and others. It has the specialty of being simple and clear, easily becoming one of the favorites among programmers of this generation. Now that we have discussed why the decision tree is good to let us look further into what actually is the decision tree classifier. To start a decision tree is a tree where there are a bunch of nodes, and each node represents a feature (attribute), each link (branch) represents a decision otherwise known as a rule and each leaf of the tree represent an outcome otherwise known as categorical or continues value. The idea is to create a tree for the entire data and get an outcome at every leaf. Now we are a bit more familiar with what is a decision tree. Let us go ahead and discuss how we can build a decision tree classifier. The decision tree can be made based on two different algorithms. One being the CART (classification and Regression Trees) and the other being ID3 (iterative Dichotomiser 3).

For ID3 first, we take the x value in the column and a y value, which stays at the last position of the column and only has “YES” or “NO” value. For the chart above, we have (outlook, temp, humidity, windy) as our x values and play, which Only has two options either ‘YES’ or “NO” is at the last position of the column or is our y value. Now we need to do the mapping of x and y. As we can see, it is a binary classification problem, so let us build the tree using the ID3 algorithm. Now to create a tree, we need a root node at first, and we need to pick one first to be the root node. A general rule of thumb is to choose the feature which has the most influence on the value y first as the root node. Then we move on and choose the next most influential feature to be the next node. Here we are going to use the concept of entropy, which is the measure of the amount of uncertainty in the data set. We need to calculate the entropy for all categorical values for the binary classification problem. So to sum it all up, we can say that we need to compute the entropy for the data set first. Then for every attribute/feature, we need first of all to calculate entropy for all the categorical values, then take the average value information entropy for the current attribute and finally calculate how much we have gained for the current attribute. After that, we need to pick the highest gain attribute and repeat until we get our desired tree. Now that is the process of ID3.

As we have discussed above decision tree classifier has been made on another algorithm know as CART short for classification and regression trees. In this algorithm we use Gini index as our cost function used to evaluate splits in the dataset. Here our target variable is indeed a binary variable so it will take two values (yes and no). And as we all know there can be 4 combinations. Now we need to figure out the Gini score which will give us a good idea of how we can split the data. If we can get the Gini score of 0 we can consider it to be a perfect separation whereas worst case scenario would a split of 50/50. Now the question arises how we can calculate Gini index value.

Now, if the target variable is a categorical variable with multiple levels, the Gini index will still be similar. So the steps for this method are the first compute of the Gini index for data-set. Then for every feature, we need to calculate the Gini index for all categorical values and take average information entropy for the current attribute and, in the end, calculate the Gini gain. After we are done with that, we can pick the best Gini gain attribute, and we need to repeat until we get our desired tree. And that is how the decision tree algorithm works.

DT classification methods involve building tree models that consist of a series of predictors. Each of these predictors (attributes) within a training set is split repetitively until pure subsets are obtained. This process of repetitive splitting is influenced by a particular entity's (i.e., customer) characteristics. The basic anatomy of a DT comprises of both a leaf node and a decision node. The leaf node represents a predictor variable and signifies the point where binary splits transpire. Leaf nodes are also known as internal nodes. The decision node, also known as the terminal node, represents the output variable (binary outcome variable) and graphically is depicted as the end of the branch. It is the terminal node that serves as the basis for churn prediction for it reports the category with the majority of cases. Extant literature has revealed that four major DT machine learning algorithms are commonly utilized: 1) Classification and Regression Trees (CART) 2) C4.5 3) chi-squared automatic interaction detection (CHAID) and 4) C5.0. DTs serve as the foundation of other tree methods like random forests and ensemble forests, which essentially involve aggregating multiple decision trees.

The process of binary splitting an attribute relies on selecting the right attributes to split. Correct attribute selection is dependent on calculating either entropy measures (C4.5) or choosing the Gini criterion (CART) based on the type of DT algorithm. DT analysis is quite popular due to simplicity, graphical layout, and ease of interpretation. DTs provide an appropriate schematic to model both quantitative and qualitative decision making questions without needing to create dummy variables or transformations. Moreover, DTs are also able to monitor non-linearities and are easy to compute. However, DTs also have their disadvantages. DT results may not always be as predictively accurate as other methods. Furthermore, minor changes in the dataset can result in non-robust predictions. However, this classification technique has been used frequently to model churn.

3.3 Data mining models

To survive in an increasingly competitive marketplace, many companies are turning to data mining techniques for decision prediction analysis. To manage customers effectively, it is important to build a more effective and accurate decision prediction model. Statistical and data mining techniques have been utilized to construct decision prediction models. The data mining techniques can be used to discover interesting patterns or relationships in the data, and predict or classify the behavior by fitting a model based on available data. In the case where the learning dataset and the test dataset are separated for machine learning, the test dataset must satisfy the following requirements. First, the training dataset and the test dataset must be created in the same format. Second, the test dataset should not be included in the training dataset. Third, the training dataset and the test dataset must be consistent in data. However, it is very difficult to create a test data set that meets these requirements. In data mining, various verification frameworks using one dataset have been developed to solve this problem. This study uses the Split Validation operator provided by RapidMiner to support this. The operator splits the input dataset into a training dataset and a test dataset to support performance evaluation. This study selects relative segmentation among the segmentation method parameters of this operator and uses 70% of input data as learning data.

3.4 Performance evaluation

Performance assessment uses training data to determine how well the generated model works. Performance measures can be divided into technical performance measures and heuristic measures. The technical performance measures to be used in this study show performance results by generating models from training data, processing test data into models, and comparing the class labels of original verification cases with predicted class labels. Measuring technical performance can be divided into supervised and unsupervised learning. The supervised learning used in this study is classified and regressed. The data used for this learning and test all have original class values. The performance is obtained by comparing and analyzing the original class values with the prediction results.

The classification problem is the most common data analysis problem. Various metrics have been developed to measure the performance of classification models. For classification problems of category type, accuracy, precision, recall, f-measure are used a lot. RapidMiner includes Performance (Classification), which measures performance indicators for common classification problems, and Performance (Binominal Classification), which provides performance indicators specific to binomial classification problems. The table 2 shows how these indicators are calculated.

Table 2: Key performance indicators of binomial classification

		Actual class (as determined by Gold Standard)	
		True	False
Predicted class	Positive	True Positive	False Positive(Type □ error)
	Negative	False Negative(Type □ error)	True Negative

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

4. RESULTS

4.1 Decision tree

Figure 1 shows the classification tree for the full model after pruning the tree using cross-validation to avoid overfitting. The key variables in the full model analysis consist of 28 ones, as shown below, based on the criterion established with each of these variables. Hotel, is_canceled, lead_time, arrival_date_year, arrival_date_week_number, arrival_date_day_month, stays_in_weekend_nights, stays_in_week_nights, adults, previous_cancellation, previous_bookings_not_canceled, deposit_type, days_in_waiting_list, and adr influence is_repeated guest.
Tables 3 illustrate each of the confusion matrix measures. For the full model, the accuracy rate is 0.989, which implies that the error rate is 0.011 Among the customers who are predicted not to be repeated guest, the accuracy that would not be repeated guest was 99.43%, and the accuracy that would be repeated guest was 81.79% among the patients who are predicted to be repeated guest.

Insert Figure 1 here

Table 3: Performance evaluation

	True 0	True 1	Class precision
Pred. 0	34443	198	99.43%
Pred. 1	214	961	81.79%
Class recall	99.38%	82.92%	

5. CONCLUSION

The main purpose of this paper is to test the accuracy of models and develop a new model to predict the hotel customer loyalty**.** To recap, this study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in hotel customer loyalty prediction modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. Based on the findings reported above, a series of implications are drawn.
Concerning the first goal, the findings of the study suggest that assessing the role of variables is complex and that their influences vary according to the classification methods employed. The decision tree methods highlight the explanatory power as most important to the analysis. Therefore, collectively no unanimous conclusions can be drawn about which explanatory variables are most critical to hotel customer loyalty prediction for all the methods employed in totality. Yet, the findings of this study do shed some additional light on the customer's profile. The hotel companies should be seeking to predict hotel customer loyalty on the classification methods employed.
This study provides some research contributions and practical contributions. First, this study extends the existing literature by empirically examining the combined impact of the variables on hotel customer loyalty prediction modeling. A lot of studies have been reported about hotel customer loyalty prediction, but no one can say that researchers can create a universal human tool to predict hotel customer loyalty. Hotel customer loyalty prediction is so complex and connected to so many elements that researchers tend to use fewer elements and ignore the effects of other factors. Customer demographics are often changed and constantly monitored, which can cause problems with hotel companies and compromise personal information. Some studies looked at age, gender, and geographic location. But researchers are still unable to express cultural and behavioral factors that can affect hotel customer loyalty prediction. This study contributes to the literature regarding hotel customer loyalty prediction by providing a global model summarizing the hotel customer loyalty prediction determinants of customers' individual factors. Second, the methodology used in this paper can be viewed as a roadmap for the reader to follow the steps taken in this case study and to apply the one-day procedure to identify the causes of many other problems. This paper attempts to come up with the best-performing model for predicting hotel customer loyalty based on a limited set of features, including customers' individual factors. Machine learning technique such as decision tree, along with feature importance analyses, is employed to achieve the best results in terms of accuracy. With this methodology, this study identified a pattern of hotel customer loyalty prediction.
Practically, this application helps hotel companies to manage the personal records of the customers and also this will make the decision faster if they have the report of the user already

with them. Basically, a prototype of the model is described in the paper which can be used by the organizations for making the correct or right decision to recognize the loyalty of the hotel customers. Futhermore, this paper provides practical implications to the managing authority of hotel company. Because whole process of prediction is done privately, no stakeholders would be able to alter the processing. Result against particular customer can be send to various departments of hotel so that they can take appropriate action on application. This helps all others department to carried out other formalities.

In the proposed system, I have a database to store the records of the customers, and when the count of the customers increases means, more data will be generated, and the storage will become a problem. Therefore, in a future release, there will be a cloud facility to store all the records in the cloud. Therefore, the data protected safely and can be retrieved from anywhere if we have the right to access the data. The smart device will be synced with our application in the future release. Therefore, the customer's real-time transaction record will be monitored, and in the case of booking needs, the hotel companies will get alerted.

In the future, the machine learning model will make use of a larger training dataset, possibly more than a million different data points maintained in an electronic transaction record system. Although it would be a huge leap in terms of computational power and software sophistication, a system that will work on artificial intelligence might allow the financial practitioner to decide the best-suited decisiont for the concerned customers as soon as possible. A software API can be developed to enable health websites and apps to provide access to the customer free of cost. The probability prediction would be performed with zero or virtually no delay in processing.

REFERENCES

1. J. Simon. **Artificial intelligence: scope, players, markets and geography**. *Digital Policy, Regulation and Governance*, Vol. 21, No. 3, pp. 208-237, 2019.
2. Y. Le Cun, Y. Bengio, and G. E. Hinton. **Deep learning**. *Nature*, Vol. 521, pp. 436-444, 2015.
3. A. Correia, M. Kozak, and J. Ferradeira. **From tourist motivations to tourist satisfaction.** *International. Journal of Culture, Tourism and Hospitality Research*, Vol. 7, No. 4, pp. 411-424, 2013.
4. P. Raman. **Way to create loyalty.** New Straits Times, Kuala Lumpur, 17 August,1999
5. J. T. Bowen and S. Shoemaker. **Loyalty: a strategic commitment?**. *Cornell Hotel and Restaurant Administration Quarterly*, February, pp. 12-25, 1998.
6. R. A. D'Aveni, G. B. Dagnino, and K. G. Smith. **The Age of Temporary Advantage**. *Strategic Management Journal*, Vol. 31, No. 13, pp. 1371-1385, 2010.
7. M. M. Mariani, M. Di Felice, and M. Mura. **Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations**. *Tourism Management*, Vol. 54, pp. 321-343, 2016.
8. H. Wilkins, B. Merrilees, and C. Herington. **The determinants of loyalty in hotels**. *Journal of Hospitality Marketing & Management*, Vol. 19, pp. 1–21, 2009.
9. J. M. Gonzalez-Cava, J. A. Reboso, J. L. Casteleiro-Roca, J. L. Calvo-Rolle, and J. A. M. Pérez. **A Novel Fuzzy Algorithm to Introduce New Variables in the Drug Supply Decision-Making Process in Medicine**. *Complexity*, https://doi.org/10.1155/2018/9012720, 2018.

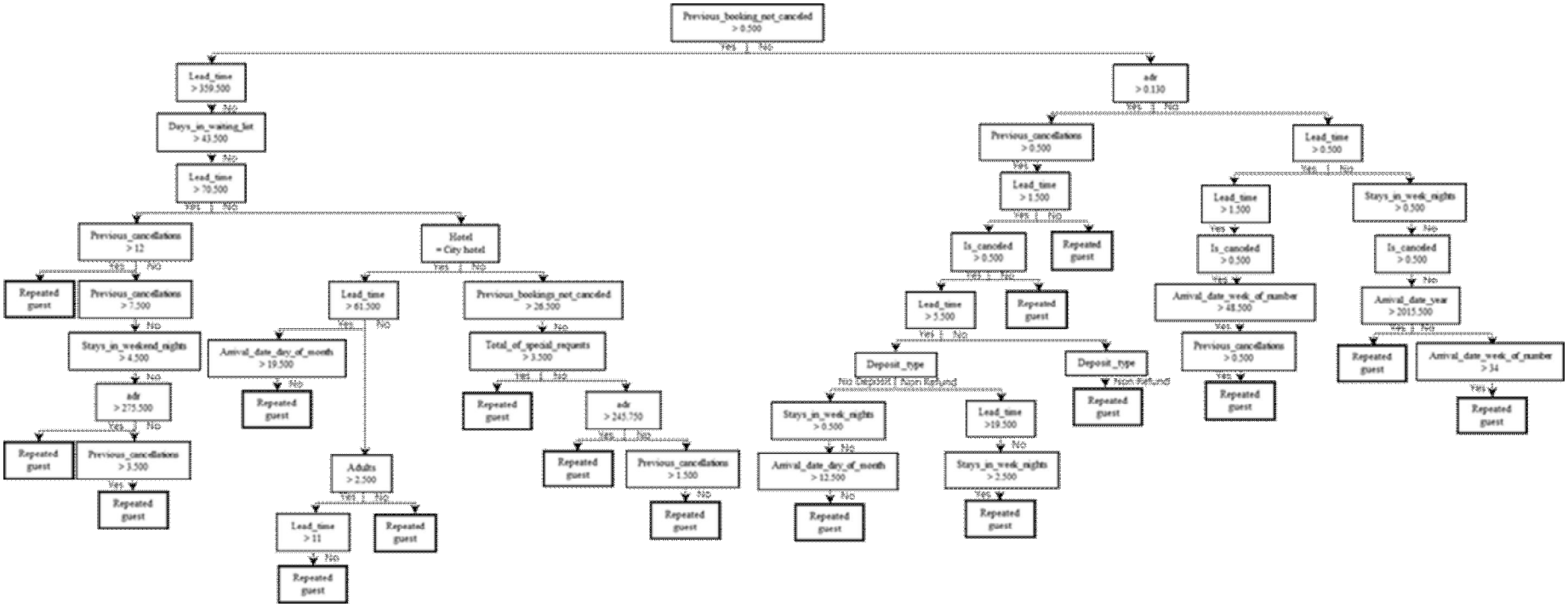

Figure 1: Classification Tree for the Full Model