Comparing Machine Learning Algorithms for hotel reservation status Analysis

Alwalid Abushanab
Computer Science
University Of Prince Edward Island
Machine Learning, Data Mining Assignment 2

March 2023

1 Abstract

Hotel reservation cancellation is a major source of revenue loss for hotels. However, machine learning algorithms can help reduce this loss by predicting whether a customer will cancel their reservation. This research paper focuses on developing and comparing different machine-learning models using a Hotel Reservations dataset. Specifically, K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, and Naive Bayes models were implemented, and their accuracy was compared to determine the best one. The results show that the Decision Tree model had the highest accuracy of 91.78%, indicating that machine learning algorithms can effectively predict customer cancellations. This, in turn, allows hotels to take appropriate measures to manage their bookings and resources.

2 Introduction

In this paper, we will try to build a model to predict whether a hotel customer will cancel their reservation. We are building the model to help hotels put a stop to or at least reduce the revenue loss caused by canceled reservations. In order to perform that, we will use four different machine learning algorithms, specifically K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, and Naive Bayes. we will evaluate the performance of each of these algorithms and then compare them against each other to find the best model for predicting if a reservation will be canceled or not. These models will be trained on a Hotel Reservation dataset containing information on reservations made by customers. The results of this research contain important findings for any hotel looking to implement classification models to improve their booking management processes and minimize their revenue loss.

3 Background

The dataset used in this paper is the ”Hotel Reservations Dataset” available on Kaggle [1]. This dataset contains 36275 observations, each with 18 features and a Booking_ID which serves as a unique identifier for each booking. (the following feature description is paraphrased from [1]) The features include no_of_adults and no_of_children, which indicate the number of adults and children included in the reservation. The features no_of_weekend_nights and no_of_week_nights indicate the number of weekend nights (Saturday or Sunday) and weekday nights (Monday to Friday) the guest booked or stayed at the hotel. The type_of_meal_plan feature indicates the meal plan chosen by the customer. The required_car_parking_space feature indicates whether the customer requires a parking space (0 - No, 1- Yes). The room_type_reserved feature indicates the type of room reserved by the customer, which is ciphered (encoded) by INN Hotels. The lead_time feature represents the number of days between the date of booking and the arrival date. The arrival_year, arrival_month, and arrival_date features represent the year, month, and date of arrival. The market_segment_type feature indicates the market segment designation. The repeated_guest feature indicates whether the customer is a repeated guest (0 - No, 1- Yes). The no_of_previous_cancellations and no_of_previous_bookings_not_canceled features indicate the number of previous bookings canceled and not canceled by the customer prior to the current booking. The avg_price_per_room feature represents the average price per day of the reservation in euros. Finally, the no_of_special_requests feature represents the total number of special requests made by the customer. The booking_status feature serves as a flag indicating if the booking was canceled or not, this is the target feature. The features type_of_meal_plan, room_type_reserved, market_segment_type, and booking_status are categorical, while all the rest of the features are numerical. The dataset did not contain any missing values. The observations in the data set are all from either 2017 or 2018.

Classification Algorithms used:

One of the machine learning algorithms used is K-Nearest Neighbour (KNN). KNN works by assigning a k value, which is an int representing how many neighboring points to look at. as the model is being trained, an observation will be assigned to a class by looking at its k neighbors. the most appearing class in these k neighbors will be assigned to the new observation.[2]

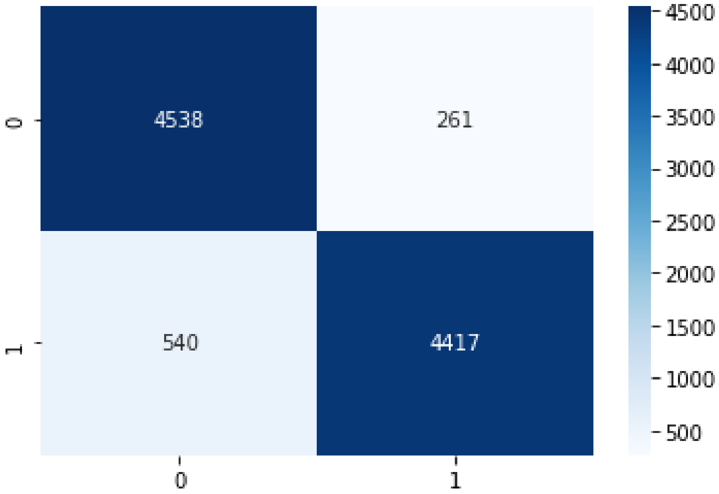
Another one of the machine learning algorithms used is logistic regression. Logistic regression is an algorithm that works by "modeling the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature."[3]. in the case of our dataset, logistic regression will model the probability of whether customers' reservations will be canceled or not which is a binary classification.

Furthermore, a decision tree model was also used. Decision trees are used to classify observations based on the response to a set of questions.[4] The questions are based on the features, they would be ordered based on a mathematical equation. The mathematical equation determines how well the feature separates the observations. after selecting a feature, the algorithm will look into the next feature that separates the observations even further based on the response of the first feature.[5] This will be done until all features that will but most if not all observations in a class.

Last but not least, Naïve Bayes was used. The following definition is from ChatGBT: Naive Bayes is based on Bayes' theorem, which calculates the probability of a certain event occurring based on prior knowledge of related conditions. The "naive" assumption made in this algorithm is that all input features are independent of each other, which simplifies the calculations. To classify a new data point, Naive Bayes calculates the probability of it belonging to each class, based on the frequencies of the features in the training data for each class. The algorithm then assigns the new data point to the class with the highest probability.

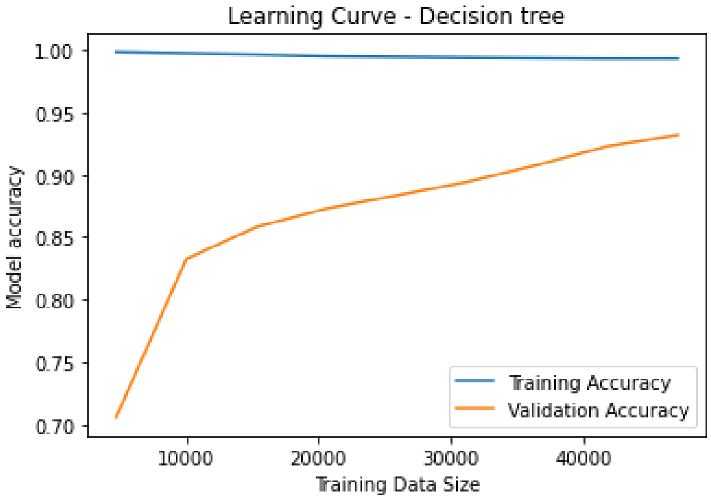
4 Analysis results

The model with the highest reported accuracy is the Decision tree model. It has an accuracy of 91.78%. To produce these results, we tested the model accuracy on 80% of the raw data, the accuracy reported after testing the model with the remaining 20% of the data was 86.24%. This accuracy is considered to be good, but further data processing was done to try and increase it even more. In order for us to increase it, we started by scaling the training data. Scaling is the process of transforming the different features into a similar scale.[6] However, the scaled data reported a lower accuracy with the same data split. This decrease in accuracy is most likely to be due to the noise or variant being amplified after scaling the data. Next, we performed feature selection and PCA separately on the raw training data. In feature selection, the data features were reduced from 17 to 9, almost half. And using PCA, the data was reduced to 2 dimensions. However, just like scaling, both feature selection and PCA reported a lower accuracy than the raw data. We then tried one last technique, which was Over-Sampling. Over-sampling takes the small groups in the data, it was the "Not canceled" group in our case, and creates more samples by duplicating existing ones to balance the dataset.[7] Over-sampling the data produces a higher accuracy than the raw data, it produced an accuracy of 91.78%. Then, we produce the confusion matrix and learning curve to analyze the results further.



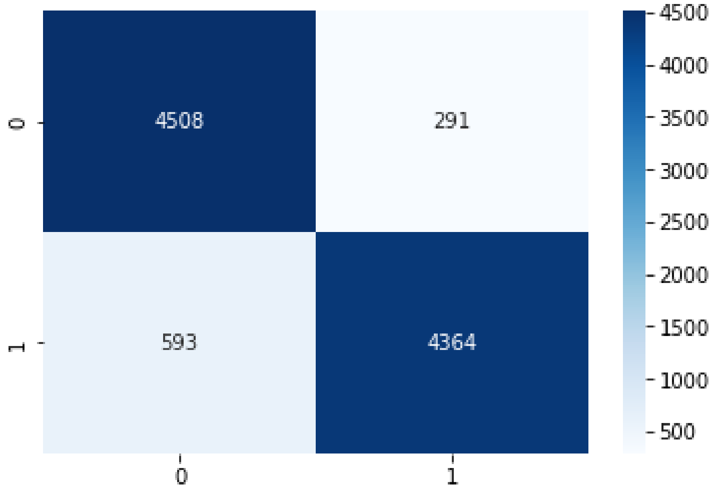
Confusion Matrix: 0 represents Canceled Reservations, and 1 non-canceled

The above Confusion Matrix shows that the model predicted a total of 8,955 observations correctly (predicted 4538 to be Canceled, and they were. and 4417 to be not canceled, and they were not). However, it mistakenly predicted 801 observations. the majority of the mistakenly predicted observations were false negatives (almost twice as much as false positives) false negative is when the model predicts non-canceled reservations as canceled.



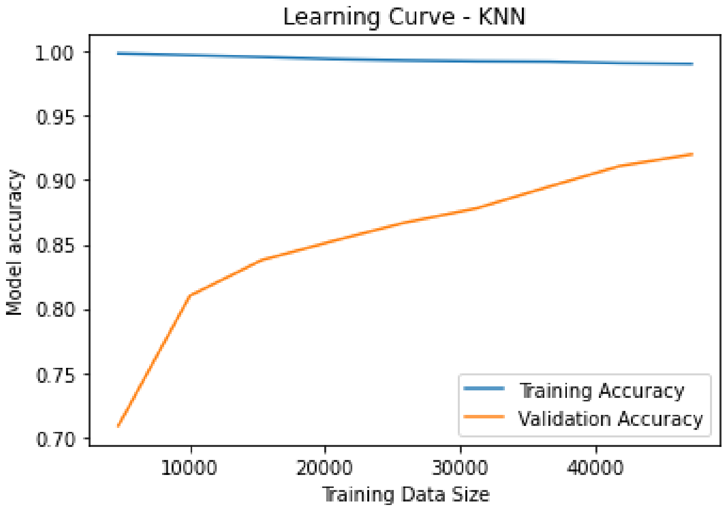
The learning curve indicates that the model is performing extremely well (with almost 100% accuracy) on the training data. Which might be an indication of overfitting (the model is learning the noise). However, based on the confusion matrix result and since the validation accuracy is also somewhat high (91%), the model is most likely to not be overfitting.

KNN model reported a 90.93% accuracy. We followed the same steps taken for the decision tree model. The best accuracy was obtained from the scaled and then Over-sampled data at k = 1. while Both feature selection and PCA produced a lower accuracy than the scaled data (without over-sampling). The following is the confusion matrix for the Knn model.



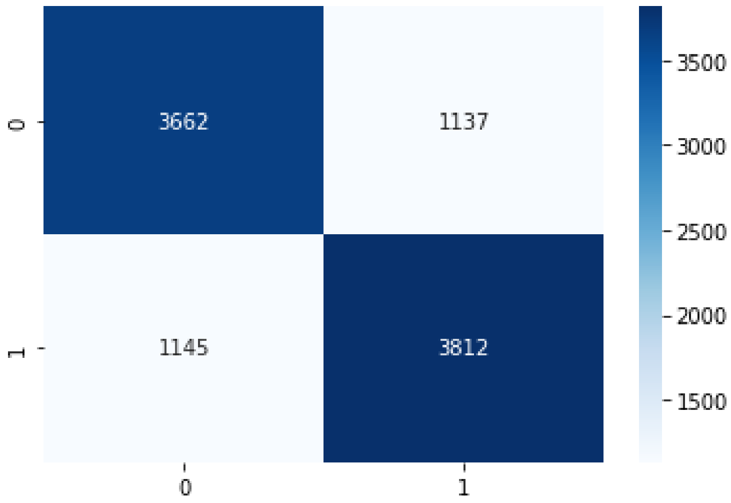
Confusion Matrix: 0 represents Canceled Reservations, and 1 non-canceled

The KNN model's Confusion Matrix shows that the model predicted a total of 8,872 observations correctly (predicted 4508 to be Canceled, and they were. and 4364 to be not canceled, and they were not). However, it mistakenly predicted 593 observations. the majority of the mistakenly predicted observations were false negatives (almost twice as much as false positives) false negative is when the model predicts non-canceled reservations as canceled.



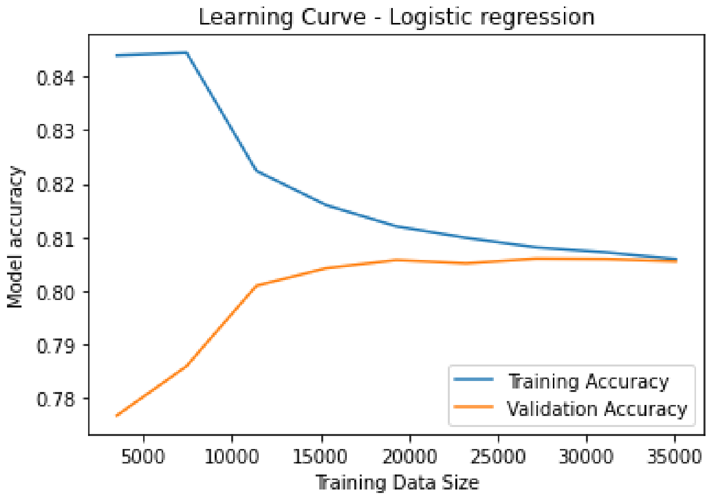
The learning curve indicates that the model is performing extremely well (with almost 100% accuracy) on the training data. Which might be an indication of overfitting (the model is learning the noise). However, based on the confusion matrix result and since the validation accuracy is also somewhat high (90%), the model is most likely to not be overfitting.

Logistic Regression model reported a 79.88% accuracy. We followed the same steps taken for the decision tree model. The best accuracy was obtained from the scaled data. Over-sampling, feature selection, and PCA when performed on the scaled data all produce lower accuracies than the scaled data.



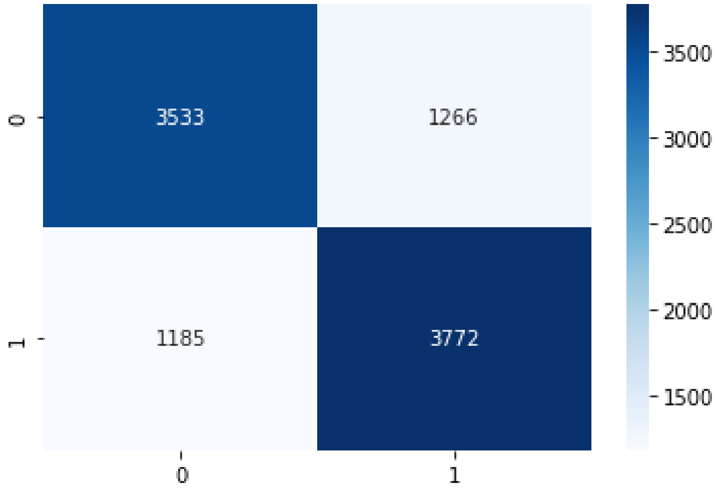
Confusion Matrix: 0 represents Canceled Reservations, and 1 non-canceled

The above Confusion Matrix shows that the model predicted a total of 7,474 observations correctly (predicted 3662 to be Canceled, and they were. and 3812 to be not canceled, and they were not). However, it mistakenly predicted 2,282 observations. The false positive and the false Negative are extremely close to each other, so the model appears to be equally classifying observations from either class wrongly.

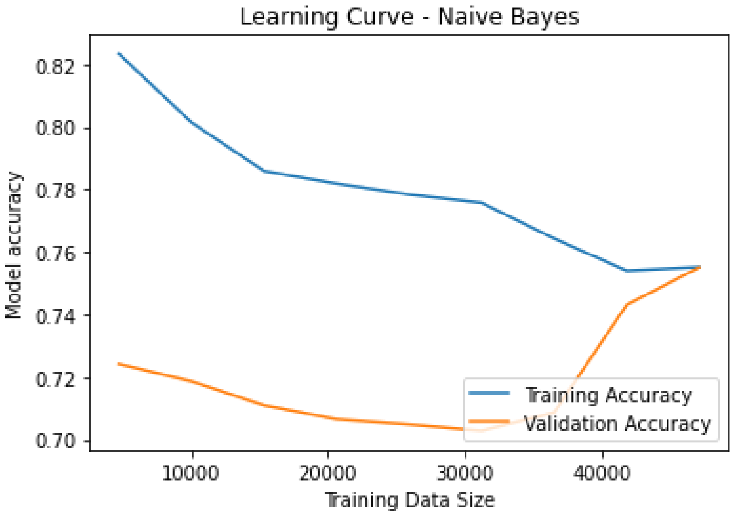


The learning curve indicates that the model is performing poorly (with almost 81% accuracy) on the training data. Which might be an indication of underfitting (the model is unable to capture the differences between the 2 classes). Based on the confusion matrix result and since the validation accuracy is also low (around 81%), the model is most likely to be underfitting.

The naive Bayes model produces results similar to the Logistic regression model, but with lower accuracy.



Confusion Matrix: 0 represents Canceled Reservations, and 1 non-canceled



5 Conclusions

In conclusion, hotel reservation cancellation is a significant source of revenue loss for hotels. However, machine learning algorithms can be used to predict whether a customer will cancel their reservation, which can help hotels reduce this loss. In this research paper, four different machine learning models, K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, and Naive Bayes, were implemented and evaluated using a Hotel Reservations dataset. The results showed that the Decision Tree model had the highest accuracy of 91.78%, indicating that machine learning algorithms can effectively predict customer cancellations. These findings are significant for any hotel looking to implement classification models to improve their booking management processes and minimize their revenue loss.

6 Resources

1 Kaggle link

2 UPEI CS4120 intro lecture

3 https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/

4 https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/#::̃text=A%20decision%20tree%20is%20a,that%20contains%20the%20desired%20categorization.

5 UPEI CS4120 decision tree lecture and slides

6 UPEI CS4120 scaling

7 https://www.pewresearch.org/fact-tank/2016/10/25/oversampling-is-used-to-study-small-groups-not-bias-poll-results/#::̃text=Oversampling%20is%20the%20practice%20of,too%20small%20to%20report%20on.

7 ChatGBT prompts

1- i have done some analysis on a hotel dataset to try and classify whether a costumer will cancel their reservation or not, i made a knn, logistic regres-

sion, decition tree, and Naive Bayes models.
give me an outline for the abstract

2- revise the following: The dataset we use for the paper is called "Hotel Reservations Dataset" which can be found on Kaggle [1]. This dataset contains 36275 observations. Each observation 17 features, a label and the target value. The target value is the status of the reservation (canceled or not). And the 17 features are as follows: number of adults, number of children, number of weekend nights, number of week nights, type of meal plan, required car parking space, room type reserved, lead time, arrival year, arrival month, arrival day, market segment type, repeated guest, number of previous cancelation, number previous bookings not canceled, average price per room, and number of requests. 3- include the following describtion of the features: Booking_ID: unique identifier of each booking no_of_adults: Number of adults no_of_children: Number of Children no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel type_of_meal_plan: Type of meal plan booked by the customer: required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes) room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels. lead_time: Number of days between the date of booking and the arrival date arrival_year: Year of arrival date arrival_month: Month of arrival date arrival_date: Date of the month market_segment_type: Market segment designation. repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes) no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros) no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc) booking_status: Flag indicating if the booking was canceled or not.

3-given the following paper, produce a conclusion paragraph: (gave it the entire paper starting from the abstract and ending at the end of the results and analysis section)