Canadian AI 2023
5—9 June 2023, Montréal

Machine Learning for Hotel Booking Cancellation Prediction

Vicky Kuo
York University, Toronto, Ontario, Canada

Keywords: prediction, machine learning, principal component, explainable AI, bayesian optimization

1. Introduction

Accurate demand forecasting is essential for hotels to optimize their operations, reduce costs, and improve customer satisfaction [1]. The high rate of booking cancellations in the hospitality industry poses a challenge to demand forecasting, as inaccurate predictions can result in overbooking or underbooking, leading to financial losses and unsatisfied customers. Machine learning has emerged as a powerful tool for predicting hotel booking cancellations, with the XGBoost classifier being a commonly used model for generating accurate results in hospitality datasets [2].

This study proposes an approach that combines oversampling, principal component analysis (PCA), hyperparameter tuning with Bayesian Optimization (BO) [3], and explainable AI (XAI) [4]. The proposed research aims to develop a machine learning model that not only accurately predicts hotel booking cancellations but also provides stakeholders with an explanation of how the prediction was made.

2. Literature Review

The hospitality industry is a significant global industry that faces several challenges, including the high rate of booking cancellations. Accurate prediction of booking cancellations can enable hotels to optimize their operations, minimize costs, and enhance customer satisfaction. Therefore, machine learning has emerged as a powerful tool for predicting hotel booking cancellations.

The XGBoost classifier is a popular machine learning algorithm for solving classification problems [5]. It has been widely used in various fields, including natural language processing, computer vision, and finance. XGBoost is a gradient boosting framework that can handle large datasets and provides high accuracy in classification tasks. Several studies have found XGBoost to be the most effective method for analyzing hotel booking cancellation datasets.

According to the finds of Antonio et al. [1], hoteliers can improve estimating and forecasting in revenue management by using machine learning, which offers higher accuracy in a more timely manner and, most importantly, in a more practical manner that is less reliant on subjective judgements or speculative thinking.

In a study conducted by Satu et al. [6], the GB and XGB classifiers were more frequently used to produce the most effective prediction of hotel booking cancellations.

Khair et al. [7] also highlights the effectiveness of machine learning techniques in predicting hotel booking cancellations with XGBoost emerging as a top contender.

To the best of our knowledge, no study has attempted to enhance hotel booking cancellation forecasting by reducing noise with PCA and employing the BO on XGBoost.

*vickytc@yorku.ca

3. Methods

The aim of this study is to develop a machine learning model that accurately predicts hotel booking cancellations and fills the gap for the still-missing case studies of the use of XAI. To achieve this goal, the study proposes an approach that combines data preprocessing, oversampling, PCA, hyperparameter tuning with BO, and XAI.

3.1. Data Preprocessing

The dataset used in this study was obtained from Kaggle, an internet website. It contained 36,275 rows and 19 features, including booking details, booking source, and booking status. The data contains both numerical and categorical variables. Since the dataset does not have duplicate records and missing values, the dataset will be only preprocessed by converting categorical values to numerical ones and normalizing the data with log transformation considering the data followed a power law distribution. To train and evaluate our models, we will split 70% of the data for training, and use the other 30% for testing.

3.2. Oversampling

To address the issue of imbalanced classes in the training data, where the minority class represents cancellations (32.8%), the SMOTE (Synthetic Minority Over-sampling) technique will be used. SMOTE is an oversampling technique that generates new synthetic samples in the minority class by interpolating between minority class instances and their nearest neighbors in the feature space. The purpose of generating synthetic samples is to increase the number of minority class instances in the training data, which can help to balance the class distribution and improve the performance of the model on the minority class.

3.3. Principal Component Analysis (PCA)

PCA is a popular technique used to identify patterns in high-dimensional data by reducing the number of variables while preserving most of the variance in the original data. To remove noise and reduce the dimensionality of dataset, PCA will be used in this study to reduce the dimensionality of the data into the lower dimensional data with non-correlated attributes.

3.4. XGBoost with Bayesian Optimization (BO)

To optimize the hyperparameters of the XGBoost classifier, we will employ the BO to minimize the loss function while searching for the optimal hyperparameters. We will then train the XGBoost model on the training data using the optimal hyperparameters.

3.5. Explainable AI (XAI)

One of the downsides of ML models is a lack of result interpretation. [8]. To address this issue, we will use XAI techniques such as SHAP (SHapley Additive ExPlanations) values and visualization analysis to provide stakeholders with a clear understanding of the model's decision-making process. The SHAP values method helps to explain the contribution of each feature to the final prediction. The use of XAI techniques will enhance the trustworthiness of the model and provides stakeholders with valuable insights into the factors that drive hotel booking cancellations.

References

[1] d. A. A. N. L. Antonio N. "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior". In: (2019). DOI: 10.1177/1938965519851466.

[2] M. S. Satu, K. Ahammed, and M. Z. Abedin. "Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry". In: (2021). DOI: 10.1109/ICCIT51783.2020.9392648.

[3] J. Snoek, H. Larochelle, and R. P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Ed. by P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. 2012, pp. 2960–2968. URL: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

[4] I. Ahmed, G. Jeon, and F. Piccialli. "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where". In: *IEEE Transactions on Industrial Informatics* 18.8 (2022), pp. 5031–5042. DOI: 10.1109/TII.2022.3146552.

[5] T. Chen and C. Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

[6] M. S. Satu, K. Ahammed, and M. Z. Abedin. "Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry". In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. 2020, pp. 1–6. DOI: 10.1109/ICCIT51783.2020.9392648.

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

[8] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens. "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach". In: *European Journal of Operational Research* 218.1 (2012), pp. 211–229. URL: https://EconPapers.repec.org/RePEc:eee:ejores:v:218:y:2012:i:1:p:211-229.