Machine Learning for Hotel Booking Cancellations Prediction

Vicky Kuo
York University

Keywords: forecasting, prediction, machine learning, explainable AI, bayesian optimization

1. Introduction

Predicting hotel booking cancellations is a significant concern for hotel managers as it has a direct impact on their revenue and occupancy rate. In this study, we aim to develop a predictive model that not only accurately predicts hotel booking cancellations but also provides explainability to the stakeholders.

To fulfill this objective, we use a combination of the XGBoost algorithm and Bayesian Optimization (BO) technique. XGBoost is a highly effective tree-based model that has the capability to handle large datasets and produce accurate results [1]. The XGBoost algorithm has been proved to be the best model for reservation cancellations [2]. However, the performance of XGBoost depends heavily on the selection of hyperparameters, which can be challenging to optimize. This is where the BO algorithm [3] comes in, allowing us to iteratively update the probabilistic surrogate function and identify the optimal set of hyperparameters for XGBoost.

In addition to accuracy, explainability is also a crucial factor in this study, and we apply the explainable AI (XAI) approach with the SHAP framework [4] to analyze feature importance and understand the relationships between features and the target variable.

2. Methods

The methods employed in this study encompass several important steps. First, exploratory data analysis is critical in understanding the data and preparing it for modeling. This involves cleaning the data, removing missing values, correcting inaccuracies, and ensuring that the data is in the correct format. Next, descriptive statistics and visualizations are generated to understand the high-level structure of the data. These statistics and visualizations are used to identify relationships and patterns in the data, understand the distribution of each feature, and impute any outliers.

Data engineering is the next step in exploratory data analysis, which involves normalizing the data to ensure that the features have the same scale. Additionally, the correlation between features is analyzed to identify and remove highly correlated features, which reduces the risk of overfitting and improves the performance of the model. The data is then split into training and testing sets, which are used to train and evaluate the machine learning model.

The XGBoost model's hyperparameters are optimized using BO. We define a list of XGBoost hyperparameters and their range of values, which serve as the search space for the Bayesian optimization algorithm. Then, we prepare an objective function for the performance evaluation and run the Bayesian optimization algorithm using the defined search space, objective function, and XGBoost model. The Bayesian optimization algorithm iteratively adjusts the hyperparameters and evaluates the performance of the XGBoost model until the optimal hyperparameters are found.

*vickytc@yorku.ca

Bayesian Optimization Algorithm

Assuming goal is to maximize unknown function f(x) on data D:

for n loops do
-> select new x_{n+1} by optimization of α which is an acquisition function $x_{n+1} = \max \alpha(x; D_n)$
-> get new observation y_{n+1} from objective function
-> augment data $D_{n+1} = \{D_n, (x_{n+1}, y_{n+1})\}$
-> update model
end for

Figure 1. Bayesian Optimization Algorithm [3]

Finally, we apply the XAI-based SHAP technique to analyze the feature importance and understand how these features affect the target variable. The feature importance of the XGBoost model will be analyzed to determine which features have the most impact on the predictions. This information will be used to gain insight into the relationships between the features and the target variable, which can help in understanding why certain predictions are made by machine learning models.

The SHAP technique provides a unified measure of feature importance that considers the impact of each feature on the target variable while accounting for the effects of other features. It measures the importance of each feature by computing the average contribution of the feature value to the difference between the actual prediction and the baseline prediction. This allows us to identify the most influential features in the model and analyze their impact on the model's performance.

In addition to analyzing the feature importance, visualizations will also be used to understand the relationships between the features and the target variable. These can help improve the interpretability and transparency of the model by showing how changes in feature values can impact the predictions. Visualizations can also be used to identify and explain complex patterns in the data that are difficult to understand through feature importance analysis alone.

References

[1] *XGBoost*. Wikipedia, 2022, September 25. URL: https://en.wikipedia.org/w/index.php?title=XGBoost&oldid=1112145594.

[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] *Hyperparameter optimization*. Wikipedia, 2023, January 29. URL: https://en.wikipedia.org/w/index.php?title=Hyperparameter_optimization&oldid=1136348545.

[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.