How to deal with problem customers?

1st Michal Krezalek
Oxford Brookes University
Oxford, United Kingdom
19022526@brookes.ac.uk

***Abstract*—Abstract — This study investigates the factors impacting last-minute cancellations and no-shows in the hotel industry, a significant challenge causing potential revenue loss. Hotels can better prepare and mitigate the impact by identifying the most critical indicators of booking cancellations. The study builds on top of other research in the field and helps fill the gap. This study uses three different machine learning approaches and compares them to the current state of the art.**

***Index Terms*—Machine learning, hotel booking, predicting cancellation**

I. Introduction

The hotel industry plays a vital role in the global economy, yet it faces challenges like last-minute cancellations and no-shows, leading to potential revenue loss. There has been previous research exploring factors influencing cancellations which can be used to help with predicting which customers will cancel or not show up for booking [3]. The purpose of this study is to build on top of other research in the field and to identify other key factors contributing to booking cancellations [1][2][5][6]. This paper will explore a hotel reservation dataset found on Kaggle and use three different techniques, KNN, Random Forest and PCA and compare their accuracy with the current state of the art. By examining the dataset, this study seeks to provide insight for hotels to predict cancellations and no-shows better and implement effective strategies to minimise revenue loss. This paper will discuss the current state-of-the-art methodology and examine results from experiments.

II. LITERATURE REVIEW

Hotel booking cancellations can reach up to 20% of total bookings, resulting in lost potential revenue [1]. There can be many reasons for a person to cancel their booking ranging from unexpected occurrences to illness [5][6]. Various machine learning techniques were utilised to help hotels predict which customers are more likely to cancel their hotel booking. Andriawan et al. [1] used the Random Forest algorithm to predict the cancellation with an accuracy of 87.25%. The dataset (n=119,390) used in this paper came from two different hotels, which is insufficient to create an accurate model for every hotel. Furthermore, the research did not consider external factors affecting hotel booking demand, such as economic or political events. It did, however, consider one resort hotel and one city hotel with two different booking patterns. Satu, Ahammed and Abedin [6] used the same dataset as Andriawan et al. [1]; however, it used a different approach to the dataset, making it possible to compare the final results. Satu, Ahammed and Abedin [6], in their data preprocessing, detected outliers, missing values, and purified data, which led to a higher accuracy model as the final result; this is something that was missing in Andriawan et al. [1] research. Furthermore, Andriawan et al. [1] used three different feature selections with seven other machine-learning techniques making detailed results. Even though the preprocessing of Ahammed and Abedin (2020) was more comprehensive and thorough, it performed worse than the best method; XGB had an accuracy of 79.15%, while Andriawan et al. [1] had 87.25%. Machine learning can predict which customers are more likely to cancel their bookings and apply different cancellation strategies requiring more considerable cancellation fees [4]. This strategy is better than overbooking, which can lead to customer dissatisfaction and a bad reputation [5][6]. Rakesh et al. [5] achieved a very high accuracy compared to the other two papers, with 94.38% accuracy for DT and 94.64% for KNN. Rakesh et al. [5] and Satu, Ahammed and Abedin [6] used DT and KNN; however, their results differ vastly. It can potentially tell that the dataset is significant for training a model, and for some hotels, it will be easier to predict if a person will cancel their booking. Furthermore, Rakesh et al. [5] did not provide where the dataset came from, making it impossible to compare the sources. Andriawan et al. [1] and Chen et al. [3] found that the most significant factor in deciding if someone will cancel their booking or not show is the difference between bookings made and arrival time. Chen et al. [3] achieved an accuracy of 1 with the use of Catboost, which is attributed to the selection of key features, gradient boosting algorithm and quality and size of the dataset. To conclude, there is a need for more extensive datasets with data from various hotels rather than focusing on one or two hotels simultaneously. This will allow a more generic model to be applied to a broader range of hotels.

III. METHODOLOGY

The dataset came from Kaggle, and there were no missing data which meant there was no need for cleaning data. The dataset was created between 2017 and 2018 by the hotel. The dataset had some categorical columns which needed to be labelled. Firstly data was analysed, and machine learning techniques were applied without extra parameters. Afterwards, techniques were used to improve the algorithm and benchmarked with basic algorithms. Various feature selection methods were applied with minor improvements in accuracy.

TABLE I
STATE OF THE ART MODEL COMPARISON

Paper	Method used 2	Accuracy
Andriawan et al., 2020	RF	87.25%
Andriawan et al., 2020	LGBM	83.80%
Andriawan et al., 2020	Catboost	84.84%
Andriawan et al., 2020	XGB	82.27%
Ahammed and Abedin, 2020	GB	76.90%
Ahammed and Abedin, 2020	RF	75.61%
Ahammed and Abedin, 2020	XGB	79.15%
Ahammed and Abedin, 2020	DT	72.04%
Ahammed and Abedin, 2020	LR	76.91%
Ahammed and Abedin, 2020	KNN	75.90%
Ahammed and Abedin, 2020	GNB	74.47%
Rakesh et al., 2022	DT	94.38%
Rakesh et al., 2022	KNN	94.64%
Chen et al., 2022	Catboost	100%
Chen et al., 2022	KNN	99.80%
Chen et al., 2022	Logistic regression	99.31%

Each configuration of feature selection and hyperparameters was logged to track the progress of an algorithm.

IV. RESULTS

V. DISCUSSION

VI. CONCLUSION

VII. LEGAL, SOCIAL, ETHICAL, SECURITY AND PROFESSIONAL ISSUES

Fig. 1. Example of a figure caption.

REFERENCES

[1] Andriawan, Z.A., Purnama, S.R., Darmawan, A.S., Ricko, Wibowo, A., Sugiharto, A. and Wijayanto, F. (2020). Prediction of Hotel Booking Cancellation using CRISP-DM. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICICoS51170.2020.9299011.

[2] Chen, C.-C. and Xie, K. (Lijia) (2013). Differentiation of cancellation policies in the U.S. hotel industry. International Journal of Hospitality Management, 34, pp.66–72. doi:https://doi.org/10.1016/j.ijhm.2013.02.007.

[3] Chen, Y., Ding, C., Ye, H. and Zhou, Y. (2022). Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation. Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022). doi:https://doi.org/10.2991/aebmr.k.220307.225.

[4] He, Y.F., Wen, P.P., Lan, Y.Q. and Miao, Z.W. (2018). Hotel Cancellation Strategies Under Online Advanced Booking. [online] IEEE Xplore. doi:https://doi.org/10.1109/IEEM.2018.8607679.

[5] Rakesh, M.V., Kumar, S.P., Yogitha and Aishwarya., R. (2022). Hotel Booking Cancelation Prediction using ML algorithms. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). doi:https://doi.org/10.1109/icais53314.2022.9742843.

[6] [6] Satu, Md.S., Ahammed, K. and Abedin, M.Z. (2020). Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in the Hospitality Industry. 2020 23rd International Conference on Computer and Information Technology (ICCIT). doi:https://doi.org/10.1109/iccit51783.2020.9392648.