How to deal with problematic customers?

***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.

TABLE I
STATE OF THE ART MODEL COMPARISON

Paper	Method used	Accuracy
Andriawan et al. [1]	RF	87.25%
Andriawan et al. [1]	LGBM	83.80%
Andriawan et al. [1]	Catboost	84.84%
Andriawan et al. [1]	XGB	82.27%
Chen et al. [3]	Catboost	100%
Chen et al. [3]	KNN	99.80%
Chen et al. [3]	Logistic regression	99.31%
Rakesh et al. [5]	DT	94.38%
Rakesh et al. [5]	KNN	94.64%
Satu, Md.S., Ahammed and Abedin [6]	GB	76.90%
Satu, Md.S., Ahammed and Abedin [6]	RF	75.61%
Satu, Md.S., Ahammed and Abedin [6]	XGB	79.15%
Satu, Md.S., Ahammed and Abedin [6]	DT	72.04%
Satu, Md.S., Ahammed and Abedin [6]	LR	76.91%
Satu, Md.S., Ahammed and Abedin [6]	KNN	75.90%
Satu, Md.S., Ahammed and Abedin [6]	GNB	74.47%

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 and a comparison can be found in Table I. 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. Moreover, 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 100% 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 was sourced from Kaggle and comes from 2017/2018 but mostly from 2018. None of the values were missing which meant there was no need for data cleaning. Firstly various graphs were used to visualise the data and to find a pattern, for example, there were two times more no cancellations than cancellations so RandomOverSampler was used for PCA to balance the data. Various methods were used for feature selection like a correlation matrix to find the most important features and create new ones. For each method, a baseline model was created which was used for benchmarking and to see how hyperparameters and feature selection affect the performance.

IV. RESULTS

Three machine learning techniques were applied in this study. K-Nearest Neighbors (KNN), Principal Component Analysis (PCA) with KNN and Random Forest (RF). The results are as follows in table II. The highest accuracy was achieved with RF with 91.49% accuracy. In the following subsections, each model will be described in more detail.

TABLE II
FINAL RESULTS

Model	Accuracy	rmse	F1 score
RF	91.49%	29.81%	91.48%
PCA	90.91%	30.15%	90.89%
KNN	88.54%	33.84%	88.34%

A. KNN

The first machine-learning technique used on the dataset was KNN. Firstly KNN was run without any hyperparameters to establish a baseline. Without doing any optimisation it achieved an accuracy of 80.49%. Afterwards, Standard Scaler was used and hyperparameters were chosen based on the GridSearchCV, it increased the accuracy of the model by 7.17% and the parameters used can be found in Table IV.

Afterwards, the bagging technique was used to improve the accuracy with max features 12 and n estimators 10. It resulted in a slight increase of 0.88% and Figure 1 shows how selecting max features affects the accuracy. The com-

TABLE III
KNN MODEL COMPARISON

Model	Accuracy	rmse	F1 score
KNN	80.49%	44.16%	80.11%
KNN with hyperparameters	87.66%	35.12%	87.50%
KNN with bagging	88.54%	33.84%	88.34%

TABLE IV
HYPERPARAMETERS USED FOR KNN

Parameter	Value
n neighbors	29
weights	distance
algorithm	kd tree
leaf size	4
p	1
metric	manhattan

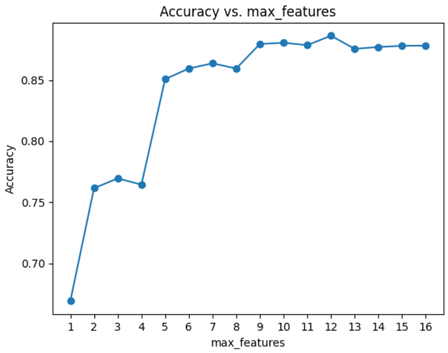

Fig. 1. max features affect on accuracy

parison of those three models can be found in Table III.

B. PCA

Afterwards, PCA was used and compared with KNN from the previous subsection with the use of the same hyperparameters and feature selection. PCA was accompanied by removing unnecessary features based on the correlation matrix that can be found in Figure 2. In total two new features were created and 13 were deleted. Number of components for PCA was calculated using cumulative variance and the result was to use seven out of seven features, Figure 3. Furthermore, RandomOverSampler was used to balance the dataset since there were two times more no cancellations than cancellations.

In the end, PCA optimisation improved model accuracy by 2.37% making it 90.91% and PCA results can be found in Table V.

C. Random Forest

RF is an ensemble machine learning method. For RF, various hyperparameters were tested with different results. Firstly RF was run without any hyperparameters or feature selection

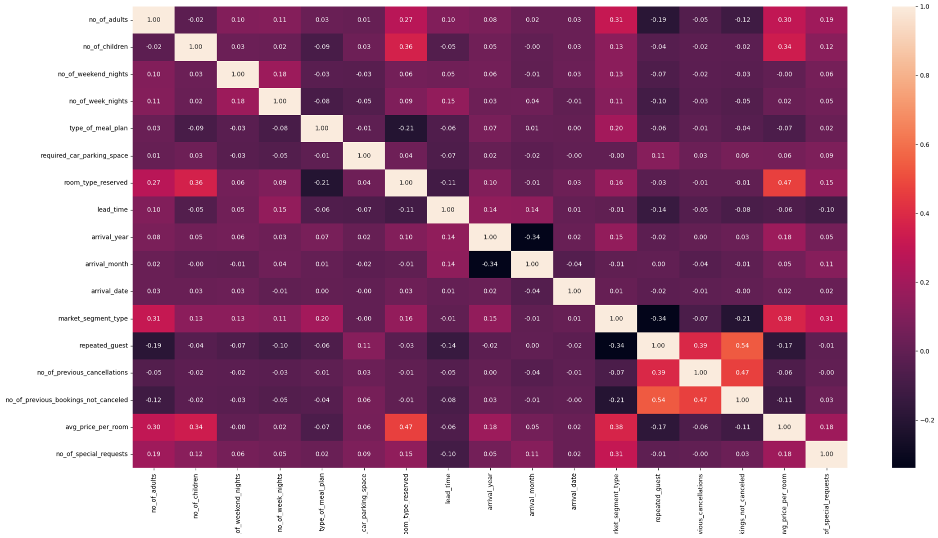

Fig. 2. correlation matrix for PCA analysis

TABLE V
PCA MODEL COMPARISON

Model	Accuracy	rmse	F1 score
PCA	90.66%	30.56%	90.64%
PCA with bagging	90.91%	30.15%	90.89%

TABLE VI
RANDOM FOREST MODEL COMPARISON

Model	Accuracy	rmse	F1 score
RF	91.28%	29.51%	91.21%
RF with hyperparameters	91.44%	29.25%	91.38%
RF with hyperparameters and feature selection	91.55%	29.06%	91.48%

and was used as a baseline for performance comparison. The comparison of three random forest models can be found in Table VI. Baseline results were above 90% which was making it challenging from the start to improve it. Afterwards, hyperparameters were selected based on the search using RandomizedSearchCV and GridSearchCV. Those two techniques only resulted in a 0.16% increase in accuracy. The parameters used can be found in Table VII. In the end with the help from PCA, the 15 most important features were

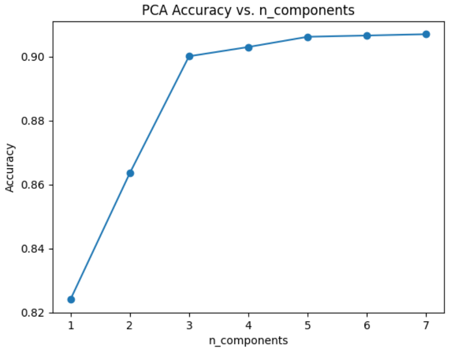

Fig. 3. PCA n components and accuracy

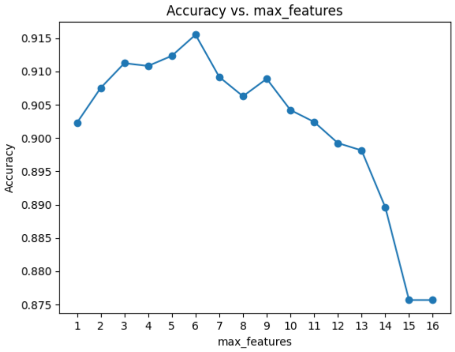

Fig. 4. max features affect on accuracy

selected which resulted in another slight improvement of 0.11%.

TABLE VII
RANDOM FOREST PARAMETERS

Parameter	Value
n_estimators	31
max_depth	33
max_features	6
min_sample_split	5
min_sample_leaf	2
bootstrap	False
criterion	entropy
class_weight	None

One of the experiments that were conducted was to find the most optimal number of max features. The results can be found in the Figure 4. Using six features gives the highest accuracy.

In all of the three methods used bagging improved the accuracy slightly.

V. DISCUSSION

The result of this study clearly shows that it is possible to predict who is more likely to cancel their booking. Also this paper also agrees with other papers that the feature "lead time" is the most significant factor when it comes to predicting booking status.

RF in this research achieved the highest accuracy and, compared to the current state-of-the-art, ranks the highest [1] [6]. Despite not considering feature selection fully explored, the RF classifier delivered exceptional performance and hyperparameter tuning only marginally improved the performance (Table VI). RF without any hyperparameters had an accuracy of 91.28%, and the extensive search for hyperparameters and feature selection only increased by 0.27%.

In contrast, the KNN classifier without PCA performed poorly compared to the current state of the art. Feature selection was not explored with KNN so that it can be compared with PCA and see an increase in performance. Chen et al. [3] achieved 99.80% while Rekesh et al. [5] achieved 94.64% accuracy using KNN. This is most likely due to the differences in the dataset and features that made it easier to make predictions, like country of origin, which was not available in the dataset used in this research. Nevertheless, it scores significantly better than Satu, Md.S., Ahammed and Abedin [6] with a 12.64% increase.

The PCA-based approach was better than the non-PCA KNN because it created two new features from existing ones and deleted the ones that were not important. There were 17 features at the beginning, and that number was cut down to only seven features showing that very few are needed to make a prediction. For example, the feature "year" should not be considered since it will make the model less accurate in the future.

By creating a baseline it was easy to compare how hyperparameters and feature selection are improving the model performance. Also, it showed in a tangible way how each change was making the performance increase or decrease. In the end, the highest increase in performance because of tunning was with KNN and the smallest improvement was with RF.

This research has explored three different machine-learning techniques. However, there are plenty of things that can be done to improve the performance of the models. For example, feature selection was not fully explored, or data was not fully balanced with RF or KNN, which might improve the model's performance.

Overall this study confirms that with the suitable dataset and the right models, it is possible to predict who is more likely to cancel their booking. Hotels can use those methods with their booking dataset and train the models specifically for their dataset since all hotels are different. To create a generic model for predicting hotel cancellation, every hotel would need to collect the same kind of data and a lot of external factors like economical or political would need to be taken into account. For now, it is not possible to create such a generic model.

VI. Conclusion

To conclude, it is possible to predict who is more likely to cancel their booking with a high probability. Hotels can use this knowledge to prepare better and have a flexible cancellation policy that charges more for people more likely to cancel. This research explored three different approaches to that problem and, in all three, achieved high accuracy. The highest was a Random Forest, with an accuracy of over 90%. Furthermore, this paper also aligns with previous research that the most significant factor in predicting booking cancellation is the lead time to arrival day.

VII. Legal, Social, Ethical, Security and Professional Issues

A potential legal issue with machine learning is data privacy. When applying machine learning, data should be fully anonymised. It should not be possible to link the real identity to the dataset. Depending on the data used for machine training, some groups of people might be mistreated, for example, by age or race. The algorithm should be transparent in that aspect, and it should be possible to explain its logic and be free from bias. Professionally, hotels must balance optimising revenue using those algorithms and client satisfaction. Machine learning algorithms should also be monitored and evaluated to ensure proper working.

VIII. Appendix A: GitHub link to source code

https://github.com/MigthyMike/COMP7032-Data-Science-and-Machine-Learning

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] 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.