Hotel Reservation Cancellation Prediction using Machine Learning

1. Introduction:

The advent of online hotel reservation channels has revolutionized the way customers book hotels, providing them with more flexibility and options to choose. However, it has also led to a significant increase in cancellations and no-shows, which can negatively affect a hotel's earnings and create operational overhead and challenges. This issue is typically addressed through strict cancellation policies, which can be inflexible and frustrating for customers. Moreover, hotel staff may be required to manually monitor the bookings and follow up with customers who have not shown up for their reservations, which can be a tedious, time-consuming, and costly affair.

To address these challenges, I have taken up this project to anticipate the likelihood of customer cancellations using machine learning techniques. The project uses an open-source dataset from Kaggle that includes various details related to customer behavior and booking history. The dataset will be preprocessed to ensure high-quality, and the machine learning models will be trained and tested to determine their accuracy in predicting cancellations. This approach offers a more flexible and effective solution for both hotels and customers, enabling hotels to better manage their bookings and potentially avoid losses due to no-shows.

The application of machine learning techniques to predict hotel reservation cancellations can be a valuable tool for the hospitality industry. It provides a way for hotels to balance the benefits of offering flexible reservation policies against potential losses incurred from cancellations and no-shows. This solution is particularly useful for hotels that offer flexible reservation policies as it enables them to optimize the revenue. The project's major contribution will be a predictive model that can be used by hotels to anticipate the likelihood of reservation cancellations and make informed decisions about managing their bookings. By keeping pace with changing customer needs and expectations, hotels can ensure their continued success in an increasingly competitive marketplace.

To conclude, the use of machine learning algorithms to predict hotel reservation cancellations offers a more effective and efficient solution for both hotels and customers. This approach enables hotels to manage their bookings more effectively, potentially reducing revenue losses due to no-shows, and provide better services to their customers. By leveraging the power of data and technology hotels can stay ahead of the competition and meet the evolving needs and expectations of their customers. With the completion of this project, hotels can look forward to leveraging the insights generated by the predictive model to optimize their operations and better serve their customers.

2. Literature Review

"Hotel reservation cancellations: analysis and prediction using machine learning algorithms" by Jasmina Novakovic and Snezana Turina (2021) presents a machine learning-based approach for predicting hotel cancellations using a dataset collected from a Portuguese market data. The authors used six classification algorithms and evaluated their performance using accuracy. Their results show that the Bagging classifier performed the best, achieving an accuracy of 90.37%.

"Predicting hotel booking cancellations to decrease uncertainty and increase revenue" by Nuno Antonio (2017) discusses the use of predictive analytics to forecast hotel booking cancellations. The study utilized a dataset of hotel reservation transactions from a Portuguese hotel and applied logistic regression to predict cancellations. The model was able to predict cancellations with an accuracy of 72.4%. The author suggests that the implementation of predictive analytics in the hotel industry can help reduce uncertainty and increase revenue by providing valuable insights for room availability and pricing decisions.

The research paper “Prediction of Hotel Booking Cancellation using CRISP-DM” by Andriawan, Z. A., et al. (2020) aims to predict hotel booking cancellations using machine learning techniques. The authors recommend using the CRISP-DM framework for predictive analysis and suggest that the Random Forest algorithm is the best machine learning model for predicting cancellations with an accuracy of 87.25%. The authors also found that the time difference between bookings made, and time of arrival was the most influential feature in predicting cancellation rates. Future work includes exploring different dataset preprocessing techniques, deployment strategies, and other modelling or hyperparameter tuning techniques to improve accuracy.

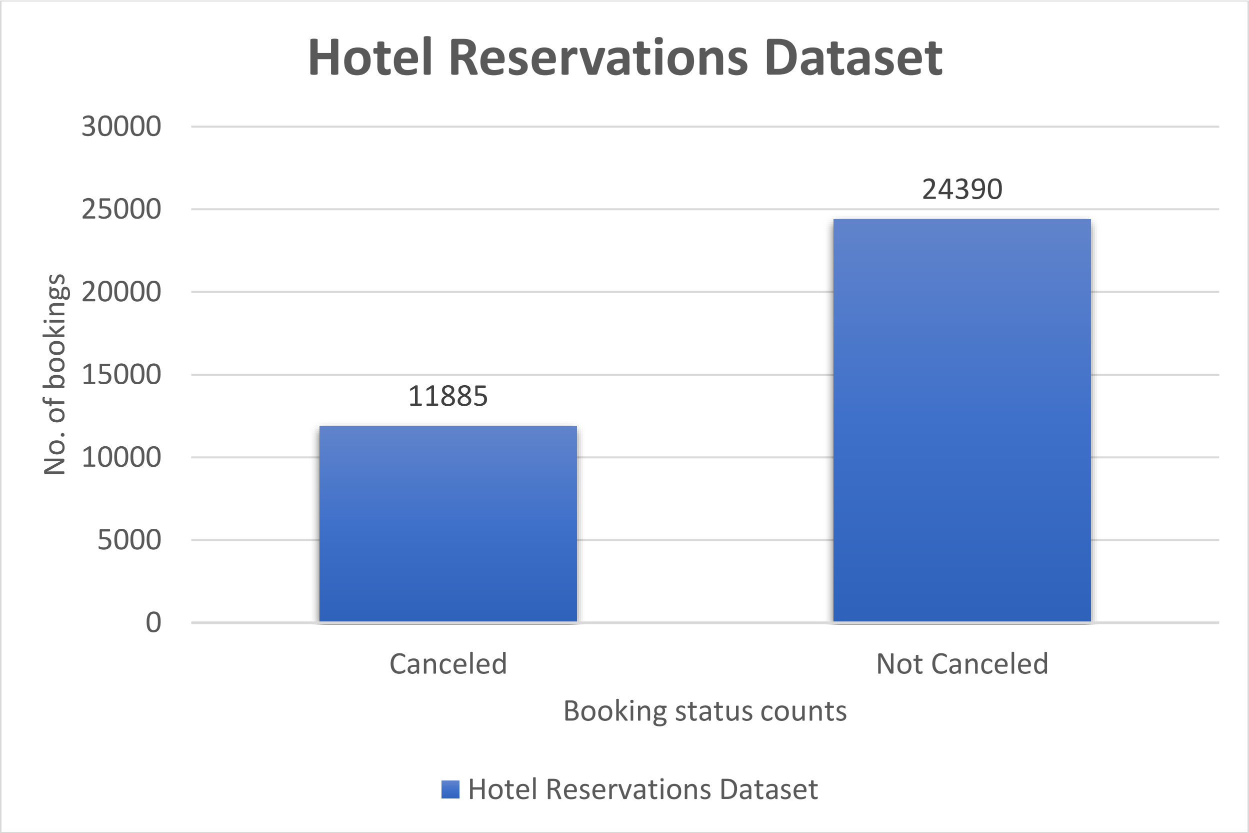
The paper “Solving the Problem of Class Imbalance in the Prediction of Hotel Cancelations: A Hybridized Machine Learning Approach”, Adil, M., Ansari, M. F., Alahmadi, A., Jei-Zheng Wu, & Chakrabortty, R. K. (2021) examines the problem of class imbalance in the prediction of hotel cancelations using a hybridized machine learning approach. The authors used a dataset from the Kaggle platform to build their model. The model was built using a combination of algorithms, namely Random Forest, Support Vector Machine and XGBoost. The results of the model showed that the hybridized machine learning approach achieved a higher accuracy rate than either of the individual algorithms. The model was able to achieve an accuracy rate of 90.6%, and the authors concluded that the hybridized approach was an effective solution to the problem of class imbalance in the prediction of hotel cancelations.

“Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation” by Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022), this paper discusses using hotel booking demand datasets to develop a model that can predict hotel booking cancellations. The authors analyze the dataset, prepare it for modeling, and train three different models to predict cancellations. They find that CatBoost is the most accurate model. The authors suggest that hotels should collect more information about their guests to increase the accuracy of the cancellation prediction model. Additionally, they recommend establishing cancellation policies and prioritizing reliable reservations. The authors note that adding more features about the hotels into the model can make the prediction more credible, and using more recent data in future work would be beneficial.

3. Data Exploration

The source of this dataset is from Kaggle.com where it was published by Ahsan Raza, a Dataset Expert on Kaggle and a co-owner of a service-based company named AWJ International LLC, under Attribution 4.0 International (CC BY 4.0) License. From the arrival_year in the dataset, we can figure-out that the data might be collected between the years of 2017 and 2018. There are a total of 36,275 rows in the dataset.

Also, the distribution of the target i.e., booking_status was as below with 11,885 bookings as canceled and 24,390 bookings not canceled,



Below are the list of features and their respective descriptions from the Hotel Reservations Dataset Kaggle,

1. **Booking_ID**: unique identifier of each booking
2. **no_of_adults**: Number of adults
3. **no_of_children**: Number of Children
4. **no_of_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
5. **no_of_week_nights**: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
6. **type_of_meal_plan**: Type of meal plan booked by the customer:
7. **required_car_parking_space**: Does the customer require a car parking space? (0 - No, 1 - Yes)
8. **room_type_reserved**: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
9. **lead_time**: Number of days between the date of booking and the arrival date
10. **arrival_year**: Year of arrival date
11. **arrival_month**: Month of arrival date
12. **arrival_date**: Date of the month
13. **market_segment_type**: Market segment designation.
14. **repeated_guest**: Is the customer a repeated guest? (0 - No, 1 - Yes)
15. **no_of_previous_cancellations**: Number of previous bookings that were canceled by the customer prior to the current booking

16. **no_of_previous_bookings_not_canceled**: Number of previous bookings not canceled by the customer prior to the current booking
17. **avg_price_per_room**: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
18. **no_of_special_requests**: Total number of special requests made by the customer (e.g., high floor, view from the room, etc.)
19. **booking_status**: Flag indicating if the booking was canceled or not.

4. Evaluation Metrics

Various metrics, such as Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrix, can be used to evaluate the performance of a binary classification model like the hotel reservation cancellation dataset. However, it is crucial to consider the class distribution of the dataset when evaluating models on imbalanced datasets. For instance, the hotel reservation cancellation dataset has an imbalanced distribution of classes, with 67.24% being 'Not_canceled' and 32.76% being 'Canceled'. Using Accuracy as the sole performance metric in such cases can be misleading, as it may result in high accuracy due to the model predicting the majority class for most instances. Thus, F1-score is recommended as a more reliable performance metric that considers both precision and recall, making it more informative for evaluating model performance on imbalanced datasets.

In conclusion, to comprehensively evaluate the performance of the binary classification model on the hotel reservation cancellation dataset, it is recommended to report both F1-score and accuracy of the model. Although some research papers may have used accuracy as the sole performance metric, it is crucial to consider the class distribution of the dataset when evaluating model performance on imbalanced datasets. Therefore, reporting both accuracy and F1-score will provide a more comprehensive evaluation of the model's performance.

5. Baseline Methodology

For a Baseline model, I have used two different baseline models to compare the performance of my final model. Firstly, I had implemented a simple straightforward positive class baseline where all the predicted test sample data were classified as ‘Canceled’ (1) after performing a train test split using the *train_test_split()* method with the *test_size* parameter as 0.2 i.e., 80% for training and 20% for testing. Finally, for calculating the results, I have used the *accuracy_score()* and *f1_score()* methods in the scikit-learn and the results from them were 33.3% and 0.5.

Secondly, using a simple Logistic Regression model with a train-to-test split ratio of 80:20 would provide us with a better baseline for classification in this problem. To implement this baseline, I excluded certain categorical features such as 'room_type_reserved', 'type_of_meal_plan', and 'market_segment_type'. Additionally, I used integer encoding for the target variable 'booking_status'. Finally, I obtained the results using the *accuracy_score()* and *f1_score()* methods from the sklearn library of scikit-learn package. And the results of accuracy and F1 score from this baseline were 78.14% and 0.62. As discussed above in the Evaluation metrics section 4, we would be reporting both the accuracy and F1 scores for this problem.

To compare my model’s performance with that of the previous studies conducted on a similar dataset by Nuno Antonio et al. [2] had achieved an accuracy of 72.4% by using Logistic Regression, and Jasmina, N et al. [1] had achieved an accuracy of 90.37% by using Bagging Classifier.

6. Proposed Methodology

My proposed methodology would begin with data preprocessing, which involves handling missing values, removing irrelevant features, and encoding categorical features using integer or one-hot encoding as required. Missing values can be imputed using different techniques, including mean imputation or machine learning models such as KNN imputation. As the dataset consists of categorical features such as 'room_type_reserved', 'type_of_meal_plan', and 'market_segment_type', it is not appropriate for training any Machine learning model directly. Thus, I plan to perform feature engineering using integer or one-hot encoding to transform the data into a suitable format for training the model.

Following the preprocessing stage, I plan to split the dataset into 70:30 ratio for the training and testing sets. Since there is limited information available online about the dataset, I am assuming that it is an Independent and Identically Distributed (IID) dataset and performing exploratory data analysis could help in gaining more insights. Based on my initial analysis, I do not think that there are any features with significantly large scales in the dataset. Therefore, I do not plan to perform feature scaling. However, I will still evaluate if feature scaling using normalization on the 'arrival_year' column would have any impact on the models.

Finally, I plan to evaluate various machine learning algorithms, including but not limited to Logistic Regression, SVM, KNN, Random Forests, and XGBoost, to measure their accuracy and F1 score metrics and compare their respective results. In addition, I will utilize stratified cross-fold validation and GridSearchCV techniques to select hyperparameters and optimize the model performance.

7. Methodology

In my study, I have split my data into training and test sets using an 70:30 ratio with stratified turned on. To ensure that each fold had roughly the same number of classes, I have used 5-fold stratified cross validation technique to evaluate my models. I have tried different machine learning algorithms, including Logistic Regression, KNN, SVM, Random Forests, and Extreme Gradient Boosting (XGBoost).

To tune the hyper parameters of my models, I have used grid search with F1-score for scoring and specified a wide range of values for each hyperparameter. Then, I tried evaluating some of the possible combinations of hyperparameters to find the optimal set. The hyperparameters that got me the best results were newton-cg solver with L2 penalty and C = 1000 for Logistic Regression, K = 1 for K – Nearest Neighbors, no of estimators = 500 with entropy algorithm for Random Forest and learning rate of 0.095 with max depth of 12 and no. of estimators = 200 for Extreme Gradient Boosting.

8. Results

In this project, I have trained several machine learning models on a binary classification task and evaluated their performance using F1-score and accuracy metrics. I started by setting a simple positive class baseline that randomly assigned labels to samples with an 80:20 train-test split, which gave an F1-score of 0.5 and an accuracy of 33.3%. I then trained a logistic regression baseline with default hyperparameters, which gave an F1-score of 0.62 and an accuracy of 78.14%.

I then experimented with various machine learning algorithms, including Logistic Regression, Support Vector Machines, KNN, Random Forest, and XGBoost. For each model, I tried different hyperparameters to find the best combination that maximized the F1-score on the validation set. I also evaluated each model on the test set to get an estimate of its generalization performance.

The best performing model on the test set was the Random Forest model with 500 estimators and entropy criterion, which achieved an F1-score of 0.8417 and an accuracy of 89.99%. The XGBoost model with learning rate of 0.095, max depth of 12, and 200 estimators came in second with an F1-score of 0.8414 and an accuracy of 89.91%. I also experimented with a Random Forest model with L2 normalization, which gave an F1-score of 0.8432 and an accuracy of 90.12% on the test set, making it my final model. Overall, my Final model Random Forest with L2 normalization outperformed the simple positive class baseline and the logistic regression baseline by 68.64% and 36% respectively w.r.t the F1-score and were competitive with other models reported in the literature on this task. As mentioned earlier, the bagging classifier by Jasmina, N et al. [1] had achieved an accuracy of 90.37% and my best performing Random Forest model achieved an 90.12% accuracy which was relatively close.

Below are my results charted in a table for each of the Machine Learning models that I had used in this study with their respective performance metrices like F1-score and Accuracy,

Type of ML Algorithm	**Hyper parameters**	**Validation F1-score**	**F1-score**	**Test set Accuracy**
Simple positive class baseline (with 80:20 train-test splits)	-	-	0.5	33.3%
Logistic Regression baseline (with 80:20 train-test splits)	Defaults	-	0.62	78.14%
Logistic Regression	Defaults	-	0.6727	80.36 %

Logistic Regression	C = 1000
	solver = newton-cg
	l2 penalty
		0.6732	0.6871	80.77%

Support Vector Machines	Defaults	-	0.5371	76.38%
KNN	Defaults	-	0.6830	80.80%
KNN	K = 1	0.6980	0.7032	80.36%
Random Forest	Defaults	-	0.8391	89.86%
Random Forest	n_estimators = 500 entropy criterion	0.8370	0.8417	89.99%
XG Boost	Defaults	-	0.8328	89.42%

XG Boost	learning rate = 0.095
	max_depth = 12
	n_estimators = 200
		0.8361	0.8414	89.91%

Random Forest with L2 Normalization (Final Model)	n_estimators = 500 entropy criterion	-	0.8432	90.12%

9. Discussion

The one thing that helped me a lot was picking the right machine learning model and here in this scenario, the ensemble techniques like the bagging classifier Random Forest and the boosting classifier Extreme Gradient Boosting did better job on generalizing the dataset properly. Also, I have observed that with the Hyper parameter tuning, I was able to improve the performance of the models but not by a huge margin. Another thing that helped me was doing the specific feature enggaveg like performing integer encoding for room_type_reserved column and one-hot encoding for market_segment_type and type_of_meal_plan columns gave me better results than simply integer encoding all the categorical columns.

Interestingly, I have tried few feature scaling techniques like StandardScaler, MinMaxScaler, etc., and Normalization techniques like L1, L2 and max and out of all of them the default L2 Euclidian normalizer helped me consistently achieve the 90% accuracy score and 0.84 F1-score but percentage change in scores without the normalization were relatively less i.e., 0.14% and 0.18% respectively w.r.t to accuracy and F1-score. Another interesting thing that I have noticed was that SVMs take a lot of time to train compared to other algorithms with relatively large datasets with increasing amounts of features.

10. Conclusion

Through this project, I have learned on how to approach a machine learning project from performing the data analysis to identify skewness of output classes, feature engineering and to finally select different models and identify the appropriate ones and perform fine tuning on them to get the best possible results. Also, I have learnt how to implement other machine learning models like Random Forests and XGBoost with grid search to get the best possible results and there is a lot more to learn in-regards to the continuously evolving machine learning techniques. In future, I would like to tryout on other models like Extra Trees Classifier, Neural Networks and ensembling multiple models to further improve the performance. Overall, I would like to conclude that ensemble learning techniques like Random Forests and Extreme Gradient Boosting were able to obtain the best possible results on this hotel reservation cancellation dataset properly.

References:

1. Jasmina, N., & Snezana, T. (2021). Hotel reservation cancellations: analysis and prediction using machine learning algorithms. International Academic Journal, 2(1), 4-13.
2. Nuno, A., Ana, A., & Luis, N. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. Tourism & Management Studies, 13(2), 25-39.
3. Andriawan, Z. A., et al. (2020). Prediction of Hotel Booking Cancellation using CRISP-DM. In 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, (pp. 1-6). doi: 10.1109/ICICoS51170.2020.9299011.
4. Adil, M., Ansari, M. F., Alahmadi, A., Jei-Zheng Wu, & Chakrabortty, R. K. (2021). Solving the problem of class imbalance in the prediction of hotel cancelations: A hybridized machine learning approach. Processes, 9(10), 1713. doi:10.3390/pr9101713.
5. Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022). Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation. In Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022) (pp. 1363–1370). doi:10.2991/aebmr.k.220307.225.