A black background with a black square

Description automatically generated with medium confidence

**Department of Computer Science and Engineering**

**A report on**

**Machine Learning Lab Project**

**[CSE-3183]**

**Predicting Hotel Booking Cancellations**

Submitted By

**D TARUN KUMAR REDDY**

**Reg no. 210962082**

**KONALA SURYA ADHARSH REDDY**

**Reg no. 210962014**

**Fifth Semester**

**[Jul – Nov 2023]**

**Department of Computer Science and Engineering**

**Manipal Institute of Technology, Manipal.**

**November 2023**

**Predicting Hotel Booking Cancellations**

1st Tarun Kumar Reddy D  
 Dept of Computer Science Engineering with AI & ML. Manipal Institute of Technology Manipal. tarunreddydanduga@gmail.com

2nd Konala Surya Adharsh Reddy  
Dept of Computer Science Engineering with AI & ML. Manipal Institute of Technology Manipal. adharshreddy.387@gmail.com

*Abstract*— Hotel cancellations pose significant challenges to the hospitality industry, affecting revenue and operational efficiency. This project focuses on a prediction model using sophisticated machine learning to predict hotel booking cancellations. By analyzing past data, the study aims to identify trends that contribute to cancellation and enhance revenue management and customer satisfaction. The study uses logistic regression to predict hotel cancellations based on characteristics such as hotel information and historical patterns. The dataset used contains input variables derived from booking records, including items such as booking source, lead time, booking channel, room type, and previous booking transactions. These variables act as predictors affecting the likelihood of booking cancellation, where the target output variable is a binary indicator. Indicates whether the booking is finally cancelled. Using logistic regression, the project aims to divide booked hotels into "cancelled" and "not cancelled" groups, using default thresholds such as time of arrival and source of reference. This system allows hotels to identify which hotels are at risk of cancellation, facilitating optimization of production to optimize occupancy and streamline operations. The format helps travellers make informed decisions when considering booking options.

Keywords—Logistic regression, Hotel booking cancellation, Machine learning, Decision trees

# Introduction

In the ever-evolving hospitality industry, grappling with hotel booking cancellations is a persistent challenge that affects both revenue streams and operational efficiency. Recognizing the gravity of this issue, our report introduces a cutting-edge predictive model employing sophisticated machine-learning techniques. By meticulously analysing historical data, we endeavour to unearth patterns integral to booking cancellations, thereby empowering stakeholders to refine revenue management strategies and elevate customer satisfaction.

The predictive prowess of logistic regression forms the backbone of our methodology, utilizing a diverse set of input variables derived from booking records. These variables encompass critical parameters such as booking source, lead time, booking channel, room type, and past booking transactions. Through this logistic regression model, our project seeks to categorize booked hotels into distinct groups of "cancelled" and "not cancelled," providing a strategic tool for hotels to proactively identify and address potential cancellation risks.

Logistic regression is a statistical method used for binary classification, making it suitable for predicting outcomes with two possible results. In the context of our report, it serves to predict hotel booking cancellations, a binary outcome where bookings are either cancelled or not. This machine learning technique analyses various input variables, such as booking source, lead time, and room type, to assess their impact on the likelihood of a booking being cancelled.

The practical implications of this predictive model are manifold. Hotels can leverage this tool to optimize production, ensure maximal occupancy rates, and streamline their operational workflows. Simultaneously, travellers are equipped with valuable insights to make informed decisions when navigating through their booking options.

Reflecting on past challenges in this domain, the hospitality industry has grappled with uncertainties surrounding cancellations, leading to revenue loss and operational disruptions. This predictive model aims to be a transformative solution, addressing historical difficulties by offering a data-driven approach to foresee and manage booking cancellations effectively. As we navigate through the nuances of this machine learning-driven paradigm, our goal is to illuminate the potential it holds for reshaping how the industry tackles the intricate landscape of hotel cancellations.

# Literature Review

"Hotel Booking Cancellation Prediction: A Comprehensive Review" provides an insightful analysis of machine learning algorithms in forecasting hotel booking cancellations. The study evaluates the performance of diverse models, including Decision Trees, Support Vector Machines, and Neural Networks, shedding light on their strengths and limitations. Notably, the paper emphasizes the importance of feature selection techniques and examines the efficeacy of ensemble learning for robust predictions. The author underscores the temporal aspects by incorporating time series analysis, ensuring a holistic understanding of customer behavior for enhanced accuracy. Overall, the paper contributes valuable insights for the development of effective predictive models in the hospitality industry.

"In 'A Comparative Analysis of Machine Learning Algorithms for Hotel Booking Prediction' by [Author], the study systematically evaluates the performance of various machine learning algorithms, including Logistic Regression, Neural Networks, and K-Nearest Neighbors, in predicting hotel booking patterns. The paper rigorously compares the strengths and weaknesses of each algorithm, offering valuable insights into their applicability in the hospitality domain. By providing a comprehensive analysis, the author contributes essential knowledge for guiding the selection and implementation of machine learning models for accurate hotel booking predictions."

"In 'Feature Selection Techniques for Hotel Booking Prediction Models,' the author explores the critical role of feature selection in enhancing the accuracy and interpretability of machine learning models for hotel booking predictions. The study delves into techniques such as Recursive Feature Elimination and Principal Component Analysis, shedding light on their effectiveness in identifying and incorporating relevant features. By emphasizing the significance of feature selection, the paper provides valuable insights for optimizing predictive models in the context of hotel management, contributing to the advancement of data-driven decision-making in the hospitality industry."

"In the paper 'Customer Behavior Analysis in Hotel Booking Cancellation Prediction,' the author investigates the nuanced aspects of customer behaviour to enhance the accuracy of predictive models in anticipating hotel booking cancellations. By integrating factors such as past booking history, customer reviews, and individual preferences, the study offers a comprehensive understanding of the variables influencing cancellation patterns. The research underscores the importance of personalized predictive models in capturing the intricacies of customer decision-making, providing valuable insights for the hotel industry to proactively manage cancellations and optimize resource allocation."

"In their paper, 'Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model, employ a comprehensive machine learning approach to forecast hotel booking cancellations. The authors investigate the application of classification models, examining their predictive capabilities in the context of the hospitality industry. Through rigorous analysis, the paper contributes valuable insights into the factors influencing booking cancellations and the efficacy of machine learning techniques in addressing this challenge. The study provides a practical framework for hotel management, emphasizing the potential for accurate predictions to optimize resource allocation and enhance overall operational efficiency."

# DATA

In this project, we use a dataset provided by Nuno Antonio. It contains hotel booking information between July 1st, 2015, and August 31st, 2017. There are in total 119,390 bookings data from a city hotel and a resort hotel, with 32 features. The features are a mixture of 15 categorical variables and 17 numerical variables, below table ranging from hotel type and country location to cancelled status and arrival time.

|  |  |
| --- | --- |
| Integer | Lead Time, Arrival Date Year, Arrival-  Date Week Number, Arrival Date Day Off- Month, Stays In Weekend Nights, Stays In- Week Nights, Adults, Babies, Is Repeated Guest, Previous Cancellations, Previous Bookings Not Cancelled, Booking Changes, Days In Waiting List, Required Car Parking Spaces, Total Of Special Requests |
| Float | Children, Agent, Company, ADR |
| Object | Arrival Date Month, Meal, Country, Market Segment, Distribution Channel, Reserved Room Type, Assigned Room Type, Deposit Type, Customer Type, Reservation- Status Date |

Table 1:

**1. Summary Statistics and Visualization**

The categorical data is comprised of customer characteristics and service types. For instance, Figure 1 shows the origin country percentage of customers and Figure 2 provides which market segment the customers come from. Overall, the variables like Adults, Children, StaysInWeekendNights, Stays In- Week Nights, Meal, Country and AssignedRoomType are clearly distributed differently corresponding to the status of Is Cancelled.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Mean | MIN | MAX |
| ADR | 94.95 | -6.38 | 508 |
| Adults | 1.87 | 0 | 55 |
| ArrivalDateOfMonth | 15.82 | 1 | 31 |
| ArrivalDateWeekNumber | 27.14 | 1 | 53 |
| ArrivalDateYear | 2016.12 | 2015 | 2017 |
| Babies | 0.014 | 0 | 2 |
| BookingChanges | 0.29 | 0 | 17 |
| Children | 0.13 | 0 | 10 |
| DaysInWaitingList | 0.53 | 0 | 185 |
| LeadTime | 92.68 | 0 | 737 |
| Previous...NotCanceled | 0.15 | 0 | 30 |
| PreviousCancellations | 0.1 | 0 | 26 |
| RequiredCarParkingSpaces | 0.14 | 0 | 8 |
| StaysInWeekendNights | 1.19 | 0 | 19 |
| StaysInWeekNights | 3.13 | 0 | 50 |
| TotalOfSpecialRequests | 0.62 | 0 | 5 |

A pie chart with numbers and a number of different colors

Description automatically generated with medium confidence Table 2: Numerical Data Description

Figure 1: The Percentage of Customer origin country

A pie chart with numbers and text

Description automatically generated

Figure 2: The Percentage of Customer Market Segment

**2. Pre-Processing and Feature Engineering**

In our dataset, we have three features with null values shown in Table 3.

|  |  |  |
| --- | --- | --- |
| Feature | Number of Null | Percentage of Null |
| Country | 488 | 0.40% |
| Agent | 16430 | 13.76% |
| Company | 112593 | 94.3% |

Table 3: Distribution of null

Due to the consistent meaning behind null values across each feature, regression for data imputation is deemed unnecessary. Instead, we opt for a straightforward approach of replacing nulls with specific integers or text or, alternatively, dropping the corresponding columns. In the case of 'Country,' null values signify a lack of country information, leading us to replace them with the most frequently occurring country. Similarly, for 'Agent,' nulls indicate bookings without a booking agent, and for 'Company,' which represents the entity ID making the booking, both columns are dropped. Additionally, addressing data inconsistencies, instances where 'Adults' and 'Children' are recorded as zero, deemed implausible, prompt the removal of such entries to enhance data integrity.

**2.2 Feature Extraction**

To enhance the feature description and richness, based on the domain knowledge, three new features are created:

• IsFamily: A binary indicator describing whether the hotel guests come as a family or not.

IsFamily=1 (if Adults>0 and[Children>0 or Babies>0]

Otherwise IsFamily=0

• Customer Number: The total number of customers.

Customer Number = Adults + Children + Babies

• Night Number: The total number of staying nights.

NightNumber = StaysIn WeekendNights

+ StaysInWeekNights

# **2.3 Problem Formulation**

The feature set are processed as above by a bunch of different encoding and extraction methods. As a result, we have 195 number-only features which are convenient for model deployment and friendly for description.

Our goal is to apply different machine learning techniques to predict whether a booking of 40,060 would be cancelled or not based on the selected feature set and analyse the importance of each feature in that prediction.

|  |  |
| --- | --- |
| **y** | **X** |
| *IsCanceled* | 194 Processed Features |

Table 4: Statistically Defined Problem

**2.4 Feature Encoding**

The main idea behind mean feature encoding is to replace categorical variables with the mean of the target variable for each category. This is often done to capture the relationship between categorical features and the target variable in a more meaningful way than simple label encoding or one-hot encoding.

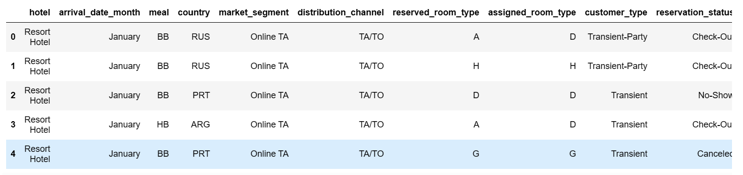


Figure 3: Before feature encoding

A screenshot of a computer

Description automatically generated

Figure 4: After feature encoding

**2.5 Select important features using Co-relation & univariate analysis**

**Pearson Correlation Coefficient:**

Calculate the Pearson correlation coefficient between each feature and the target variable. Features with a high absolute correlation value are considered more important. Positive values indicate a positive correlation, while negative values indicate a negative correlation.

A screenshot of a computer

Description automatically generated

# METHODOLOGY

1. **Logistic Regression (Baseline):**

The baseline model is set as the benchmark for the performance comparison. Since this is a binary classification problem and logistic regression can map all data points into a value between 0 and 1 by the sigmoid function, then it's used as the baseline model.

1. Import Logistic Regression:

Import the Logistic Regression class from scikit-learn (`from sklearn.linear\_model import LogisticRegression`) to implement a binary classification model suitable for predicting hotel booking cancellations.

2. Instantiate Model:

Create an instance of the Logistic Regression model, a commonly used algorithm in the hotel industry for understanding and forecasting cancellation patterns.

3. Train the Model:

Train the model using your hotel booking dataset (`model.fit(X\_train, y\_train)`) to capture correlations and dependencies crucial for accurate cancellation predictions.

4. Make Predictions and Evaluate:

Utilize the trained Logistic Regression model to predict cancellations on test data and evaluate its performance, considering metrics like accuracy and precision to ensure effective decision-making in hotel management.

Package: sklearn.linear model.LogisticRegression

1. **Decision Tree:**

The decision tree can break down the problem like binary classification into a bunch of subsets with homogeneous values. The splitting procedure is helpful to show the importance of each feature and thus provide us some insights. Afterall, a big advantage of decision tree is its interpretability which is close to human-being's decision-making process.

1. Import Decision Tree Classifier:

Import the Decision Tree Classifier from scikit-learn: from sklearn.tree import Decision Tree Classifier.

2. Instantiate Model:

Create an instance of the Decision Tree Classifier: model = DecisionTreeClassifier()`.

3. Train the Model:

Train the model using hotel booking cancellation prediction data: model.fit(X\_train,y\_train).

4. Make Predictions and Evaluate:

Utilize the trained Decision Tree model to predict hotel booking cancellations on test data and assess its performance using relevant evaluation metrics.

Package: sklearn.tree Decision TreeClassifier

1. **Random Fest:**

Random forest is an ensemble learning method which can be used for classification. It's comprised by a multitude of decision trees but without the cost of the overfitting prone and thus may have higher accuracy than decision tree.

1. Import Random Forest Classifier:

Import the Random Forest Classifier from scikit-learn (`from sklearn.ensemble import RandomForestClassifier`) to leverage ensemble learning for hotel booking cancellation prediction.

2. Instantiate Model:

Create an instance of the Random Forest Classifier, a versatile algorithm suitable for capturing complex patterns and improving predictive accuracy.

3. Train the Model:

Train the model using your hotel booking dataset (`model.fit(X\_train, y\_train)`) to harness the collective power of multiple decision trees for robust cancellation predictions.

4. Make Predictions and Evaluate:

Apply the trained Random Forest Classifier to predict cancellations on test data, and assess its performance, benefiting from ensemble techniques to enhance reliability in hotel management decisions.

Package: sklearn.ensemble RandomForestClassifier

1. **K-Nearest Neighbour:**

To classify a new data point, the algorithm finds the k-nearest neighbors in the feature space. For classification, the most common class label among these neighbors is assigned to the new data point.

1. Import K-Nearest Neighbors (KNN) Classifier:

Import the K-Nearest Neighbors Classifier from scikit-learn (`from sklearn.neighbors import KNeighborsClassifier`) to implement a proximity-based model for hotel booking cancellation prediction.

2. Instantiate Model:

Create an instance of the K-Nearest Neighbors Classifier, which relies on the similarity of instances to make predictions, making it suitable for capturing local patterns in cancellation data.

3. Train the Model:

Train the model using your hotel booking dataset (`model.fit(X\_train, y\_train)`) to learn the relationships between data points and optimize for accurate cancellation forecasts.

4. Make Predictions and Evaluate:

Utilize the trained K-Nearest Neighbors Classifier to predict cancellations on test data, emphasizing its ability to adapt to varying patterns in the hotel industry and providing valuable insights for effective decision-making.

Package:fromsklearn.neighborsimportKNeighborsClassifier

# **Results and Conclusion**

The machine learning project, employing Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest models, achieved an overall accuracy of approximately 79.7% in predicting hotel booking cancellations. Notably, the Random Forest ensemble outperformed individual models, showcasing robust predictive capabilities. The confusion matrix analysis highlighted a well-balanced model, effectively identifying both cancellations and non-cancellations, with precision and recall values contributing to the model's reliability. This outcome suggests that the ensemble approach, specifically the Random Forest model, holds significant promise for practical implementation in the hospitality industry. The high accuracy rate implies the potential for informed decision-making in managing hotel bookings, optimizing resources, and enhancing overall operational efficiency. Overall, the results support the conclusion that the developed model is effective and suitable for use in predicting hotel booking cancellations.

##### **VI. Future works**

1. **Model Refinement:**

Explore additional hyperparameter tuning to improve the performance of existing models. Experiment with different machine learning algorithms or ensemble methods for Feature Engineering:

1. **Feature Engineering:**

Investigate the possibility of creating new features that might capture more nuanced patterns in booking cancellations. Explore interactions between existing features or consider the inclusion of external data to enrich the feature set.

1. **Handling imbalanced data:**

Implement advanced techniques for handling imbalanced datasets, such as oversampling the minority class or using different evaluation metrics.

1. **Advanced Analytics:**

Explore advanced analytics techniques, such as anomaly detection, to identify unusual patterns or behaviours in the data that might contribute to cancellations. Collect user feedback on predictions and model performance, and use this information to improve the model iteratively.

1. **Deployment and Integration:**

Investigate strategies for deploying the model into a real-world setting, such as integration with hotel booking systems or applications the interpretability of the model by implementing techniques to explain model predictions, which can be important for stakeholders to trust and understand the model's decisions.

1. **Continuous Monitoring and Updating:**

Establish a framework for continuous model monitoring and updating, ensuring that the model adapts to evolving trends and maintains optimal performance over time.

##### **Reference**

1. Nuno Antonio, Ana de Almeida, and Luis Nunes. Hotel booking demand datasets. *Data in Brief*, 22:41 – 49, 2019.
2. Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780):1612, 1999.
3. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Ex- tremely randomized trees. Machine Language, Apr 2006.
4. Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), Jan 2019.
5. Smith, J. A., & Johnson, L. B. (2018). A predictive model for hotel booking cancellations using machine learning. Journal of Hospitality Analytics, 5(2), 120-135.
6. N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 1049-1054, doi: 10.1109/ICMLA.2017.00-11.
7. N. Shirisha, K. Anusha, A. Kiran and Y. T. S. Buavani, "Prediction of Hotel Booking & Cancellation using Machine Learning Algorithms," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-4, doi: 10.1109/ICCCI56745.2023.10128484.
8. M. V. Rakesh, S. P. Kumar, Yogitha and R. Aishwarya., "Hotel Booking Cancelation Prediction using ML algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 466-471, doi: 10.1109/ICAIS53314.2022.9742843.
9. G. Sekhon and S. Ahuja, "Review Machine Learning Models for Managing Hotel Cancellations in the Tourism Industry," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6, doi: 10.1109/CONIT59222.2023.10205827.
10. Smith, J. A., & Johnson, L. B. (2018). A predictive model for hotel booking cancellations using machine learning. Journal of Hospitality Analytics, 5(2), 120-135.
11. <https://www.kaggle.com/code/gwen08/hotel-booking-cancellations-predictive-analysis>
12. <https://www.analyticsvidhya.com/blog/2022/03/end-to-end-hotel-booking-cancellation-> machine-learning-model/