IND5003 Group 16 Data Analytics Group Report

Group members:

Lin Xueshun, Chen Guo, Liao Tingbo, Zhang Hanxun, Wu Zongze

1. Introduction

- Brief description of the dataset
  The dataset would be analyzed in this project is obtained from Kaggle related to hotel bookings and encompasses various attributes that can offer insights into booking trends, customer preference, etc. This dataset offers a compelling exploration into the travel and hospitality industry, allowing us to understand customer behaviors that influence hotel booking and cancellations.

- Objectives of the analysis
  This project focus on conducting a comprehensive data analysis to support strategic decision-making, focusing on two main tasks: exploratory data analysis to understand the intricacies of our dataset and to identify key trends and patterns; building robust models to predict pricing dynamics and anticipate cancellation trends, offering the business-critical insights to optimize performance and customer satisfaction.

2. Exploratory Data Analysis (EDA)

- Overview of the dataset
  The dataset consists of 119,390 entries, each representing a hotel booking, and 32 columns representing various attributes such as hotel type, lead time, stay duration, the count of adults, children, or babies, and the booking status. The dataset encompasses both city and resort hotels, giving a broader understanding of trends and patterns across different hotel types.

- Distribution of key variables
  By simply looking, and checking the dataset's content using .head(), .info(), and .describe() function, we comes up few assumptions about the dataset:

  - For cancellation, we find that a significant portion of bookings were not canceled, suggesting a high follow-through rate from customers. Understanding the cancellation patterns can help in managing bookings and improving customer satisfaction.
  - For the lead time, we find most bookings are made relatively close to the arrival time. However, there is a considerable spread in the data, with some bookings made much earlier.
  - For the repeated guests, a majority of bookings are from new guests, but there is a segment of repeat guests, highlighting potential customer loyalty.
  - For the length of stay, both during the week and the weekend varies, with most guests opting for shorter stays. However, there are instances of longer stays, particularly on week nights.

  Therefore, we need to do the further data preprocessing and visualization to check more details of the dataset.

3. Data Preparation and Cleaning

- Handling missing values
  A comprehensive check for missing values was then conducted, leading to: dropping the 'company' column as 112593 of 119390 rows in total, nearly 95%, of the data is missing . Filling missing values in 'agent' with 0 and in 'country' with "Unknown". Removing rows with missing 'children' values. These actions ensured a cleaner dataset with no missing values.

- Cleaning duplicates and irrelevant data
  We also removed duplicate rows and addressed outliers and irrelevant data by:

  - Deleting bookings with all zero values in 'children', 'adults', and 'babies', since if the number of hotel guests in a reservation is zero,then the booking becomes meaningless.
  - Filtering out 'adr' values ≤ 0 or ≥ 400, as only 8 ‘adr’ of the elements are greater than 400. Therefore, we detect them as outliers. The ‘adr’ value lower than zero seems unreasonable, since adr represents the average daily rate. We remove those data as it’s meaningless.
  - Dropping the column 'reservation_status' as it provides the same information as 'is_canceled'.

- Data Type conversion
  We converted 'reservation_status_date' from string to datetime datatype and consolidated multiple arrival date columns into a single 'arrival_date' datetime column for an easier operation in the visualization stage later, and then dropped the original date columns.

4. Visual EDA

After cleaning and processing the data, we plotted a correlation heatmap to see how correlated the variables were. The details are as follows.

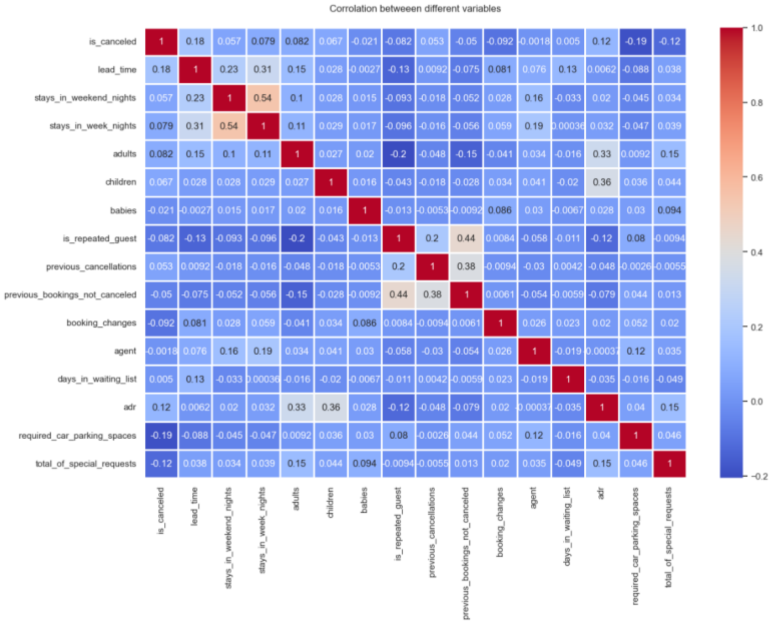

Figure 1 Correlation Heatmap of all variables

Then, we plotted a pie chart to see the proportion between Resort Hotel and City Hotel. Also, we created several bar charts to see the different canceled number, lead time distribution, stays in weekend nights and week nights between Resort Hotel and City Hotel. The details are as follows.

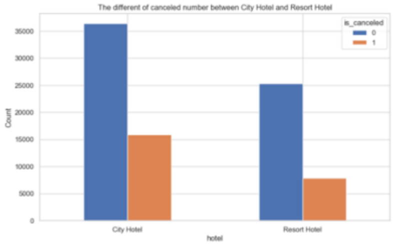

Figure 3 Cancelation of each hotel type

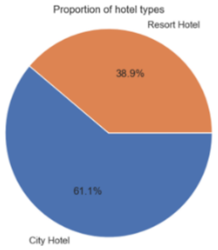

Figure 2 Proportion of hotel type

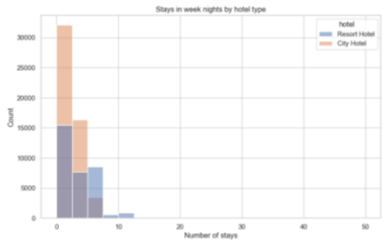

Figure 4 Stays in week nights

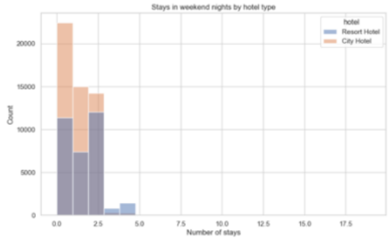

Figure 6 Stays in weekend nights

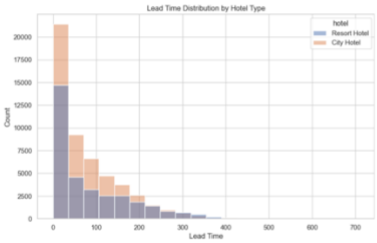

Figure 5 Lead time distribution

Finally, in order to explore the proportion of reserved rooms and assigned rooms are the same or not, we plotted two pie charts to see the difference between Resort Hotel and City Hotel. After that, we plotted visual map to see the distribution of guests from each county according to the data. Also, as to see which month is most expensive for reservation, we plotted a line plot to use average daily rate per person by month for Resort Hotel and City Hotel.

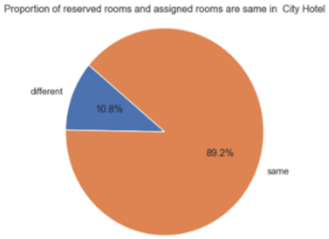

Figure 8 Proportion of reserved room type unchanged in city hotel

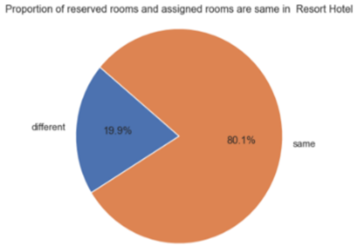

Figure 7 Proportion of reserved room type unchanged in resort hotel

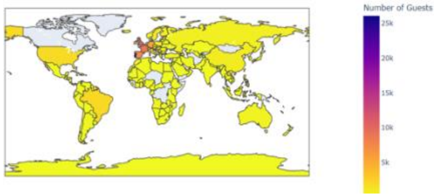

Figure 10 Number of guest on world map

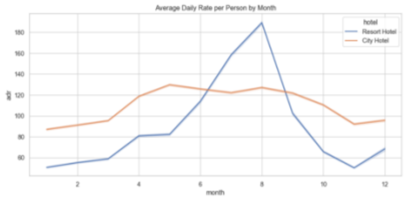

Figure 9 Average Daily Rate by month

5. Model Building and Evaluation

- Preparation for model building
  Before building the model, it’s necessary to encode categorical features in the dataset, so Label Encoding is used, transforming them into numerical format suitable for machine learning models. The affected columns include 'hotel', 'meal', 'country', etc. Additionally, we've extracted year, month, and day components back from the date fields 'reservation_status_date' and 'arrival_date', creating new columns for each component. The original date columns were subsequently dropped to streamline the dataset. Ensuring all features are in a compatible format for upcoming analyses and model training.

- Consideration of PCA
  The Principal Component Analysis (PCA) was thought to conduct in this project at the beginning to reduce dimensionality of the dataset. A precursor to PCA is Bartlett's Sphericity Test, which assesses the presence of significant correlations between variables, a prerequisite for PCA. This test compares the covariance matrix of the data to an identity matrix to check if they differ notably. A p-value less than 0.05 typically indicates sufficient correlation for PCA. However, in this dataset, a p-value of 0.765, substantially above the threshold, suggests a lack of correlations, rendering PCA unsuitable. Consequently, PCA was not employed in our analysis.

5.1 Regression

Then, we focus on the problem of predicting the adr, which refers to the average daily rate, or simply, hotel price per night per individual. As the adr is continuous data, we would use regression models to handle this problem.

- Data preparation
  In the early phases of data preprocessing, scaling and normalization were not used because the main goal was to do general visualization. But it's important to recognize that these preprocessing steps are necessary when building regression models to handle features with varying scales, StandardScaler is used to do so. And the columns which are called ‘is_canceled’ are dropped because it is not related to average daily rate prediction. The column ‘adr’ is regarded as the target variable, and the rest of the data are features. Then the dataset was further divided into 70% training and 30% testing.

- Feature selection for regression model
  In hotel price prediction, some features might be correlated. Lasso can handle this well by zeroing out the less important features, helping in feature selection and improving model performance.We set the regularization parameter α equal to 0.1 and the threshold equal to 0.5. The result of feature selection is shown below.

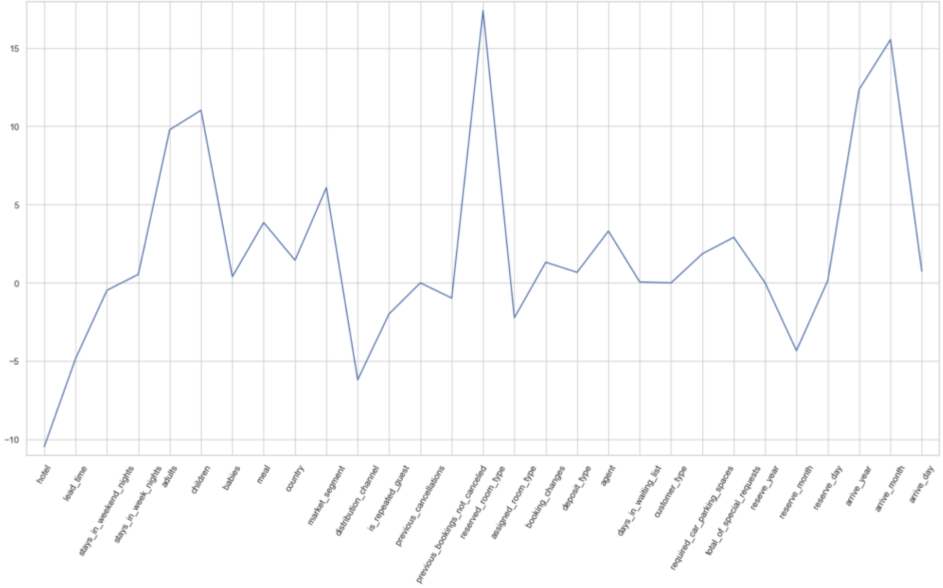

Figure 11 Lasso feature selection

- Training the models
  Linear regressor, ElasticNet regressor and RandomForest regressor are chosen to be the regression model. A Random Forest Regressor was chosen for this regression task due to its ability to handle large datasets and its robustness against overfitting. Linear regressor and ElasticNet regressor are the basic regression model, which are chosen to compare the result of RandomForest regression.

- Presentation of the result
  Root Mean Squared Error indicates the average magnitude of errors between predicted and observed values.The Coefficient of Determination is a commonly used regression analysis measure, often represented by R^2. A score close to 1 indicates a good fit. The table below show the value of RMSE and R^2:

	Root Mean Squared Error	Coefficient of Determination
Linear Regressor	38.29	0.414
ElasticNet Regressor	40.11	0.357
RandomForest Regressor	16.63	0.889

Table 1 Regression model result

5.2. Classification

We also tried to build our classification model to predict the likelihood whether the booking would be canceled by the residents.

- Feature selection for classification models

Before fitting the model, in order to simplify the model and possibly enhance its predictive capabilities, we tried to do the feature selection for classification models as well. SelectKBest is chosen to be used as it focuses on selecting features based on their statistical relevance to the target variable, ensuring that the most pertinent features are used for prediction, which is crucial in a classification task like cancellation prediction. The top 25 features were chosen for additional modeling procedures based on the scores, the result plotting is shown below:

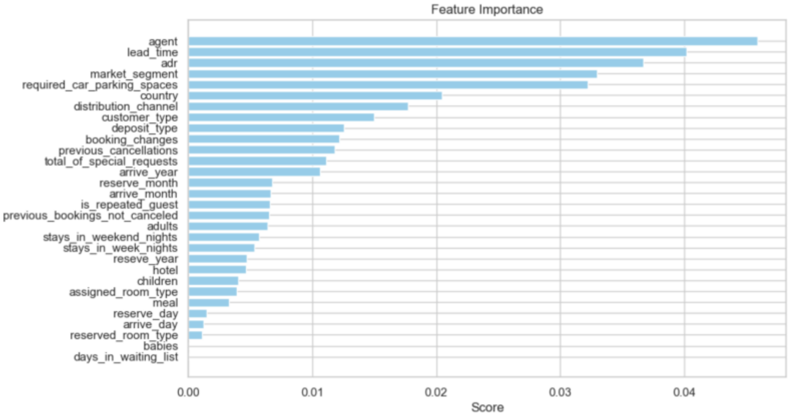

Figure 12 SelectKBest feature selection

- Data preparation
  This time, we use the column ‘is_canceled’ as the target variable, and the rest of the data as features. The dataset is divided into 70% for training and 30% for testing as well. Standardization is also necessary to apply as many classification models are also sensitive to varied scaleds.

- Training the models
  Three classification models, including KNN, logistic regression, and random forest, are introduced to handle this cancellation problem. Each of the model’s hyperparameters are well-tuned using the method of GridSearchCV or RandomizedSearchCV.

- Presentation of the result
  The result is shown in table as follow: Where precision focuses on the accuracy of positive predictions, recall highlights the model’s ability to find all relevant cases, and F1-score balances precision and recall, providing a single metric for model evaluation.

	Precision	Recall	F1-Score
KNN	0.84	0.84	0.83
Logistic Regression	0.93	0.92	0.91
Random Forest	0.91	0.91	0.91

Table 2 Classification model result

6. Results and Discussion

- Findings from the EDA
  In the visual exploratory data analysis, we observed low correlation coefficients among most variables, with the exception of a moderate positive correlation (0.54) between stays on weeknights and weekend nights. The data shows a higher prevalence of City Hotels at 61.1% compared to Resort Hotels at 38.9%. City hotels experience a greater rate of cancellations, while Resort hotel bookings are characterized by longer lead times and extended stays throughout the week and over weekends. The geographical analysis reveals a concentration of guests from Western Europe. Pie chart visualizations indicate that guests are more likely to be assigned the room they reserved, especially in City Hotels, where the discrepancy is less compared to Resort Hotels. When examining the average daily rate per person by month, Resort Hotels generally charge more than City Hotels, with May being the peak for City Hotels and August for Resort Hotels.

- Interpreting the results of the Regression models.
  According to the result of three models performance, the Random Forest Regressor performed best among three regressors in predicting the ‘adr’ for hotel bookings. With an R^2 score of 0.889, the model captures a significant portion of the variance in the data. The RMSE of 16.63, while not negligible, is acceptable given the range of `adr` values. Future work could focus on hyperparameter tuning or exploring other regression models to further improve performance.

- Interpreting the results of the classification models
  The classification models show robust performance in predicting hotel cancellations. K-Nearest Neighbors (KNN) achieves consistent results with scores around 0.83-0.84, indicating good accuracy. Logistic Regression, however, excels with a precision of 0.93 and an F1-score of 0.91, suggesting it is the most accurate model in predicting true cancellations while minimizing false positives. Random Forest matches Logistic Regression in overall accuracy (F1-score of 0.91) but is slightly less precise. These high metric scores across the models reflect their effective hyperparameter tuning and their strong predictive capabilities, with Logistic Regression being the standout for its precision in this scenario.

7. Conclusion

In conclusion, our project explored hotel booking patterns, pinpointing such as cancelation disparity favoring City Hotels and more extended lead times and stays at Resort Hotels. Our predictive modeling not only accurately forecasted cancellation likelihoods, with Logistic Regression leading the pack, but also adeptly anticipated pricing trends via the adept Random Forest Regressor. These actionable insights offer a strategic compass for the hospitality industry to refine pricing, improve booking systems, and tailor customer engagement, paving the way for future enhancements through richer data integration and analytical sophistication.