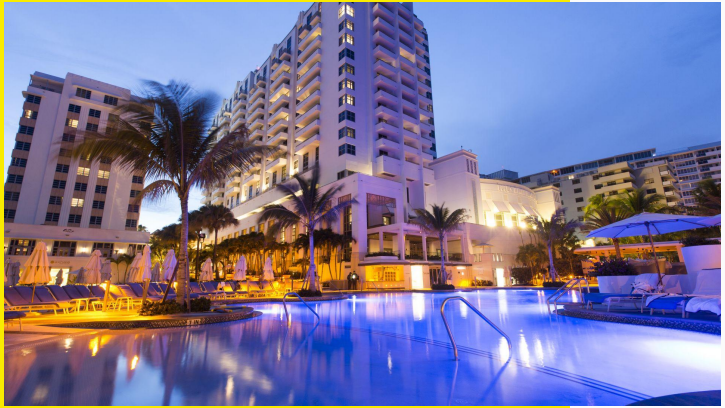
DataQuest 2023

Hunter Chen
Kiera Sobolewski

The Problem

Brescia Norton Hotel has lost $124 000 due to a growing number of room cancellations

An increase in room cancellations has also resulted in operational inefficiencies in the hotel



Our Goal is to build a model to predict cancellations at Brescia Norton Hotel

Approach

1. **Data Preprocessing**
   Cleaning data before building models
2. **Feature Engineering**
   Selecting and creating new features to include
3. **Model Selection**
   Selecting machine learning method to use
4. **Tuning of Hyperparameters**
   Using randomized and grid search to tune hyperparameters
5. **Applying the Model**

   Applying the model to the final data set

Categorical Variables

Meal Plan

What type of meal plan was chosen

Parking

Whether the customer requires a parking spot

Room type

You can describe the topic of the section here

Market Segment

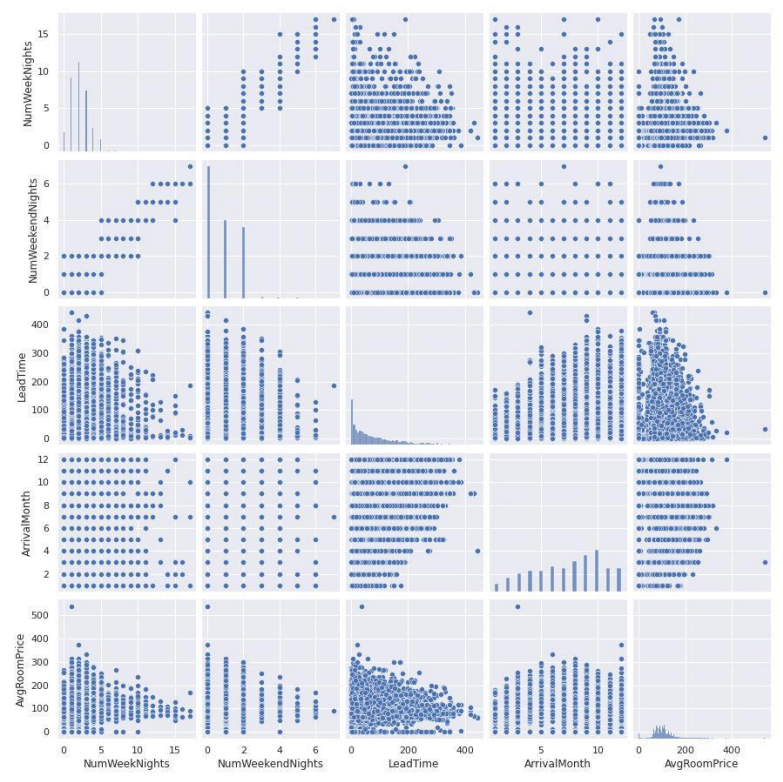
How was the booking made

Repeated Guest

Is the customer a repeated guest

Numerical Variables

- Number of Adults
- Number of Children
- Number of Weeknights Stayed or Booked
- Number of Weekend Nights Stayed or Booked
- Lead Time
- Arrival Date
- Number of Previous Cancellations
- Number of Previous non-cancellations
- Average Room Price
- Number of Special Requests Made



Data Cleaning

Removed non-relevant data

Removed Booking ID, Arrival Year and Parking

Fixed incorrect data

Found errors in meal plan type

Data Cleaning Cont’d

Converted Categorical Variables

- Using one-hot-encoding on meal plan and market segment
- Label encoding room types after binning

Feature Engineering

Created a new variables:

1. Total number of nights stayed
2. Total number of guests
3. Ratio of non-cancellation: cancellation
4. Month and day are concatenated into day of the year

Model Selection

We decided on using Random Forest for our model as it has a faster running time and produced better results.

Testing Accuracy

- Randomly split training data into two parts
- Trained on larger set and tested on smaller set

Random Forest

- Fits some number of decision tree classifiers on data sets, each tree with a random selection of features
- The number of trees is defined as a n_estimators

Tuning Hyperparameters

Used Randomized search on an increasingly tighter range of max depth and n_estimators then used grid search to finalize the max depth and n_estimators

range n_estimators: 50-1500
max depth: 15-45

→

n_estimators: 750-1250
max depth: 15-40

→

n_estimators: 900-1100 max depth: 20-35

Evaluated based on accuracy, precision and F1 score. In general the model performed better on higher n_estimators.

Final Model and Conclusions

- **Our final model had a max depth of 24 and n_estimators of 1000**
- **Our model had an accuracy of 89.5%, precision of 87.4% and F1 score 83.2%**
- **We suggest that the hotel uses this model to assess whether a customer will cancel their reservations and decide accordingly if they should allow the reservation.**