Hotel Booking demand and cancellation with Predictive Analysis

Santosh Omprakash

DSC 630 – Predictive Analytics

Bellevue University

Executive Summary

Hotel booking cancellation is a major revenue issue for hotels. Booking cancellations not only has impact on hotel revenue and affect pricing but can also impact overbooking situations and could affect hotel's online social reputation. The objective is to build model that can classify booking as cancelled or not using Machine learning. Machine Learning is a branch of Artificial Intelligence (AI) where applications are developed to learn from data to improve accuracy without any programming done to do so.

The data is obtained from Hotel Booking Demand Datasets for the analysis. The dataset is from real bookings in European market. The data obtained in raw form is cleaned and formatted. Visualization tools are used to create images that provides important insights about the dataset. For instance, cancellation before certain number of days will be looked as it is important to so that hotels can arrangements well ahead of time. With the help of machine learning algorithms, different models are trained and performance score of each model is determined. The model that has the best score is selected for fine tuning.

From all the models tested, it is discovered that the best performing model is Random Forest. The tuned model has around 87% accuracy in predicting hotel cancellation.

By using the prediction, hotels can act on high cancellation probability and associated revenue losses. Hotels can manage their business accordingly. Better net demand forecasts can be made, overbooking, cancellation policies can be improved and the revenue can be increased. There are not many models available to predict hotel cancellations. Hence, it is recommended for business to use this model and grow their business.

Introduction

The possibility of booking hotel is based on various factors. For hotels, booking cancellations are major problem. The data available in European market shows that 50% of hotel bookings made on OTA booking.com in 2018 were cancelled. Some hotel booking websites provides option to customers to book now and cancel late without any cancellation fees. This could result in customer booking more than one hotel and later decide which to choose. Due to this, there is loss in company revenue as the rooms booked won't be available to other until it is cancelled. This poses some question like what hotels can do to reduce uncertainty and maximize the revenue. Some certain aspects can be done with revenue management techniques. For example, increasing number of days until when the customer can cancel the booking before the arrival data without any charges which would give then more time to resell the rooms. The steps taken is not straightforward as other hotel may not have restrictions, so it is important to analyze this doesn't affect in any negative way.

To solve this problem, data science and machine learning algorithms may be use predict if any specific reservation would be cancelled. In this project some key question will be analyzed like percentages or booking per year, bookings cancelled per year, booking ratio between hotel and city hotel, length of stay in hotel, busiest month for hotels, most booked based on accommodation type.

Problem Statement

In hotel industry, majority of the reservations are made through Online Travel Agencies (OTA) like booking.com. Data from European market shows that around 50% of bookings made on

booking.com were cancelled. This is due to options available in OTA's where customers are encouraged to book now and cancel later free of charge when needed. This would result in customers booking more than one hotel needed.

Data science and machine learning techniques are used to accurately predict the hotel reservations that will be cancelled. Dataset of hotel bookings will be chosen and the data was cleaned to make sure there are no null or missing values. Some the date columns are formatted to have them in consistent format. Data visualization tools Exploratory Data Analysis (EDA) is used to understand and get better idea of the behavior and get insights from the data. Five machine learning models are used for analysis – Baseline, Logistic Regression, K Nearest Neighbors, Support Vector Machine (SVM) and Random Forest. The model with highest accuracy is selected for predictive analysis and evaluated.

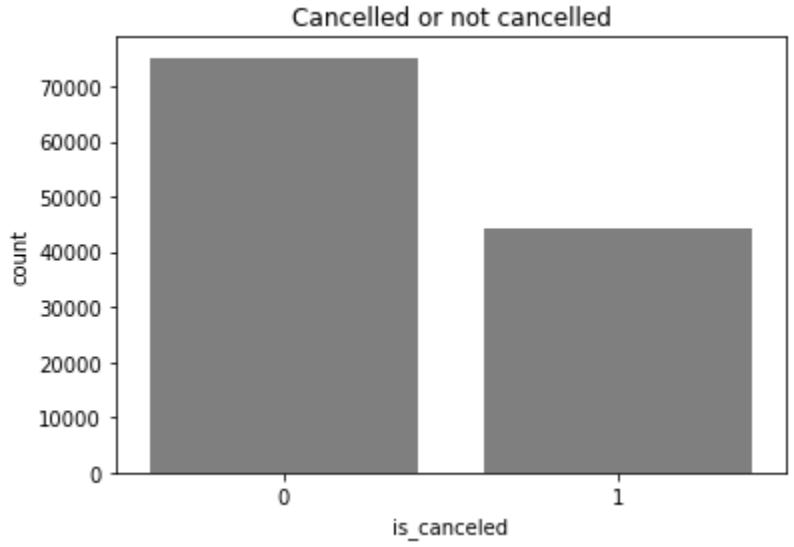
It can then be determined on how machine learning can be used to select best model for the data. The selected model will be tested and tuned. Parameters are adjusted for Random Forest model to improve optimize accuracy, improve operation for better prediction. Based on the results, the model can be ready for deployment. This can be used to improve hotel's revenue.

Data Exploration

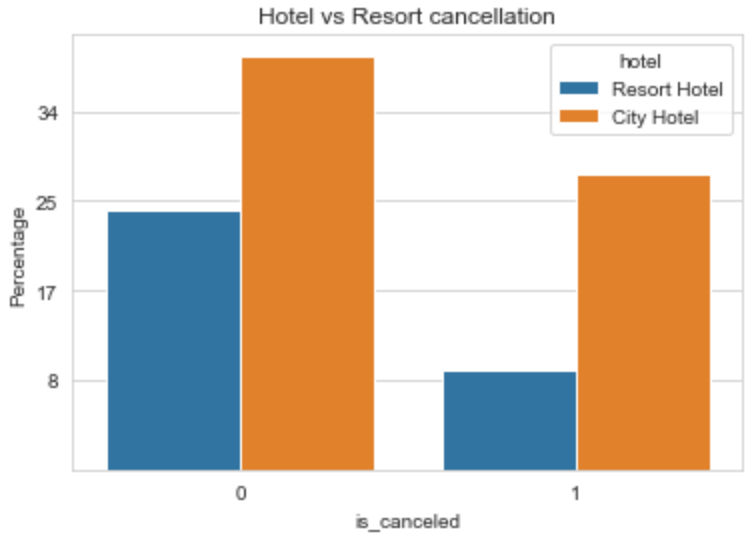
The hotel booking dataset is selected from Kaggle. The first approach is to review the data and perform Exploratory Data Analysis to analyze data and use some techniques to understand the behavior. Initial research shows that there are some missing values in some variables. They were identified and missing values were replaced.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 5 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   is_canceled  119390 non-null  int64
 1   children     119386 non-null  float64
 2   country      118902 non-null  object
 3   agent        103050 non-null  float64
 4   company      6797 non-null    float64
dtypes: float64(3), int64(1), object(1)
memory usage: 4.6+ MB
```

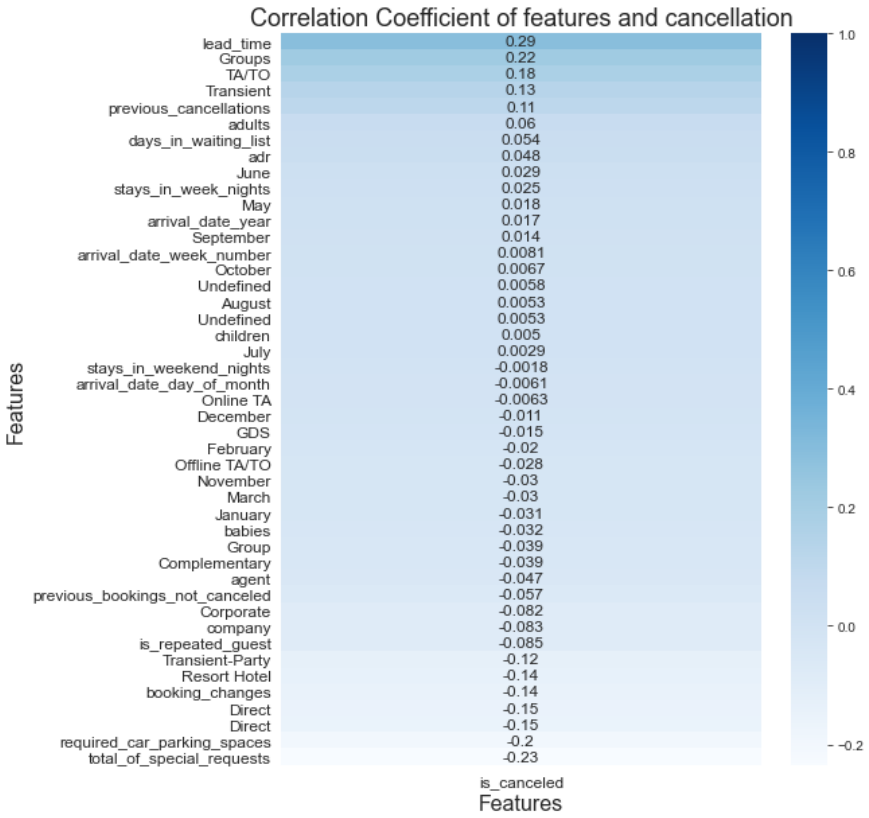
The dataset has data from city hotel and resort. Some important variables in dataset are – Hotel – Indicates if its Resort or City hotel, is_cancelled – has values 0 and 1. 0 is not cancelled and 1 is cancelled, lead_time, arrival_date_year, arrival_data_month, adults, children, babies, deposit_type, market_segment.



There are no personally identifying data in the dataset. Most of the bookings is expected to by online. Other market segment will be analyzed to see if they have any impact on cancellations. Lead_time is the important variable to research. Hence, visualization will be key on this to understand the data trend.



Exploratory data analysis of lead_time is performed and is shown below. Feature scaling is performed using StandardScaler.



Correlation coefficient of features and cancellation shows that Lead time is highly correlated features on whether booking will be canceled or not. This is valid because, as the number of days between booking made and the arrival date increases, customers have more time to cancel the reservation and the time for an unforeseen circumstance is more that could affect travel plans.

It is interesting to note the number of special request feature. When the number of special requests is more, the likelihood of booking to be canceled decreases. This indicates that having engagement with hotel before arrival and making the customer needs heard could make the customer less likely to cancel the booking.

Modeling

Libraries or packages that is needed for the machine learning model is installed. The dataset is be first cleaned to remove any missing values or replace any missing values. Using EDA, correlation of features and cancellation is understood.

The overall process is as follows –

The dataset is split into training and test sets and feature scaling is applied. Baseline model is created. Training and prediction of multiple models are done and is compared against baseline model to choose the best model using accuracy. Hyperparameters tuning is done using GridSearchCV. Model is retrained using hyperparameters and prediction is made.

The first step towards model selection is the selection of performance metric. Since the dataset is not highly imbalanced, I have used Accuracy as a metric.

I have used Baseline model accuracy as base and compared it with different models – Logistic Regression, KNN, SVM and Random Forest. Below are the accuracy scores.

	Accuracy
Baseline	62.54
LogisticRegression	78.80
KNearestNeighbors	81.62
SVM	82.01
RandomForest	86.22
OptimizedRandomForest	87.26

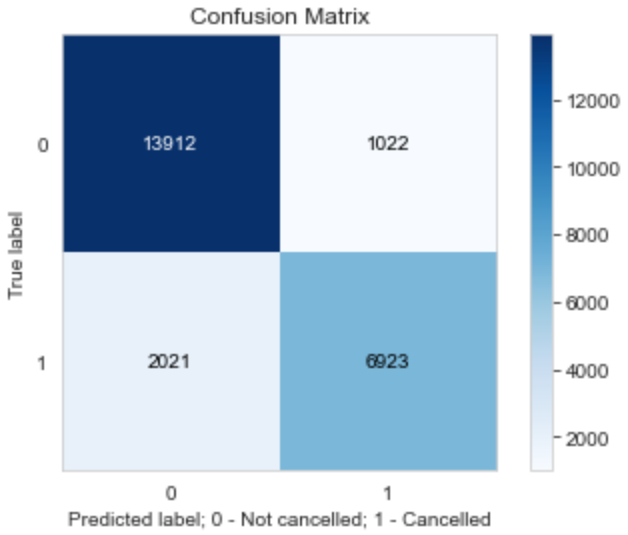
The data will be split into train and test. 80% of the data is train data and 20% is test data. In order to determine the best model, train dataset will be used. The test data will then be used for model evaluation. This is supervised classification problem as the goal is to predict if reservation will be cancelled or not.

The Random Forest model is tuned with hyper parameters using Grid Search and retrained to see if improves the accuracy score. The Grid SearchCV on random forest classifier has improved the accuracy score.

Results

Based on the accuracy scores, Random Forest model is selected and the parameters are tuned which makes it more efficient. A model is selected so that it behaves well with current data and predict the hotel booking demand and cancellation for any future data.

Below is the confusion matrix. Top left is the true negative, top right is the false positive, bottom left is the false negative and bottom right is the true positive.



Overall, 87% of the booking were classified correctly. Specifically looking at each category, 77% of canceled booking and 93% of not canceled booking are correctly classified. The booking predicated to be canceled that are actually cancelled is 87% and the booking predicted as not canceled that are actually not canceled is 87%. Confusion matrix shows that 1022 bookings that the model predicted to be canceled were not cancelled.

Looking at the confusion matrix, we see that there are 1022 bookings that our model predicted to be canceled that were not actually canceled. It means that in 4.3% of cases, a guest could arrive but the hotel may not be ready for them or there could be risk in overbooking if they were looking for replacement guest. There are 2021 bookings that model predicted to be not canceled that were actually canceled. It means that in 8.5% of scenarios, the hotel could be allocating rooms to wrong reservations.

References:

https://www.kaggle.com/jessemostipak/hotel-booking-demand

https://www.igi-global.com/chapter/using-data-science-to-predict-hotel-booking-cancellations/170956

https://www.linkedin.com/pulse/u-hotel-booking-cancellations-using-machine-learning-manuel-banza