Hotel Booking Cancellation Prediction

Sneha Jayaraman
Computer Science
PES University
Bangalore, India
sneha.jayraman@gmail.com

Hemanth Alva
Computer Science
PES University
Bangalore, India
hemathalva2708@gmail.com

Thrupthi H M
Computer Science
PES University
Bangalore, India
thrupthijiyo22@gmail.com

***Abstract*— Hotel Booking Cancellation Analysis aims to predict whether a hotel booking will be cancelled. The dataset chosen is the Hotel Booking Demand dataset that includes booking transactions made by customers for two hotels - a City hotel and a Resort Hotel. Classifying a booking to likely to be cancelled or not is done using various models. Additionally, forecasting the booking cancellation behaviour by analysing the data as a time series is done using additional forecasting models. Analysis and results of the different classifiers and forecasting models are presented in the paper.**

Keywords—booking cancellation, classifiers, time series, binary classification

I. INTRODUCTION

Hotel booking cancellations affect the hospitality industry the most in terms of revenue and customer retention. Most hotels implement a rigid cancellation policy to do away with the negative impact on the top line figures. This strategy, however, sacrifices the flexibility available to the customers and results in customer dissatisfaction. Anytime free cancellation policy, on the other hand, will result in the hotel losing out on the profits of rooms that could have otherwise been sold if the cancellation was made well in advance. More recently, online travel agencies have become a common mode of booking that allows customers to book and cancel at any time, free of cost. This encourages overbooking. The hotels now face uncertainty and must devise strategies to profitably sell the rooms ensuring that they do not lose out on potential customers. The hotels are therefore tasked with creating a balance between profits generated and customer satisfaction.

One way to achieve this would be to devise revenue management strategies such as limiting the number of days for the cancellation to be a predefined fixed day from the arrival day i.e. for example 5 days before the arrival day. This, however, would not necessarily comply with the competitor strategy and the customers may prefer hotels that are more lenient in their policy. Demand in hotel rooms is dependent on various factors such as seasonality, holiday occurrences in addition to the facilities provided by the hotel itself. Predicting periods of high demand will help the organisation be better equipped. Further, implementation of dynamic pricing i.e. increasing the average daily rate on the high demand days would be possible only if the hotel could accurately forecast demand. The solution, therefore, would be to leverage the capabilities of data analytics and machine learning to identify customer behaviour and improvise demand-management decisions.

The bigger picture involves understanding customer groups and categorising customers that are likely to cancel their booking on a given day. Besides, identifying other features such as the distance of travel, booking mode, holidays that could affect the cancellation rate plays an important role in predicting demand. Identifying predictors that most impact the demand will help hotels and revenue managers make better-informed decisions. Providing perks or incentives to customers is a way of encouraging to book more. Using data science can therefore provide a huge competitive business advantage to the hospitality industry.

II. LITERATURE SURVEY

Predicting or forecasting booking cancellations is mostly considered a problem of regression since the parameter used to predict is the rate of cancellation rather than the individual booking cancellation probability themselves. [1] however considers this to be a case of classification and uses binary classification models such as boosted decision trees, decision forests, decision jungle, locally deep support vector machine and neural networks to forecast whether a given booking will be cancelled or not. A prediction model was devised for each hotel as the actions that cause a booking cancellation could be attributed to the services specific to a hotel in terms of its location or even the discounts that a hotel offers. The number of false positives (falsely equated to a cancellation) was chosen as an important parameter since the hotel would not like to lose potential customers not cancelling their bookings. The Decision Forest model performed the best overall with well over 90% accuracy. The models, however, predict which booking is likely to be cancelled on a daily basis. Moreover, the features that capture seasonality and trend were not used for modelling.

Autoregressive Moving Average Model, EMD, EEMD, EEMD-ARIMA models were used to predict the hotel demand thereby taking into consideration the time series trend that influenced the bookings [2]. Time series models provide for a practical implementation taking into account the seasonal components. ARIMA model fitted well and produced a better forecasting accuracy than the EEMD-ARIMA for the study. A modified EEMD-ARIMA performed better with medium-term forecasting. For a hotel, the exact number of booking cancellations per day is of lesser interest. The degree of overall increase or decrease in cancellations weekly would be more meaningful from a business standpoint. Identifying the trend over a fixed period might help the hotels make better-informed decisions in their revenue strategy and cancellation policies [2].

[3] models the probability of cancellation as a probit model that takes into account several features that have an association with the target variable. This theoretical model shows the magnitude of the relationship between seasonal components and other external features on the booking cancellation. The empirical model can account for the large heterogeneity in the behaviour of customers and thereby determine the influencing attributes that could result in a booking cancellation that would otherwise be thought of as inevitable. A total of 6 hypotheses were formulated. These include possible reasoning to booking cancellations such as the impact of country of origin, mode of booking (online or onsite), the distance of travel, seasonality, size of booking, etc. Descriptive statistics is important in understanding the problem but only provides a theoretical framework on the most important exploratory variables.

III. PROBLEM STATEMENT

Cancellation prediction can be thought of as a probabilistic prediction problem wherein each booking can be classified as cancelled or not cancelled based on classification algorithms such as Logistic Regression, Decision Trees, Random Forest, KNN, SVM algorithm. Feature selection, dimensionality reduction techniques could be applied to identify the most important predictors affecting the cancellation probability. Formulating a theoretical model that identifies relationships between exploratory variables is important for revenue managers in terms of delivering on a profit margin. The rate of cancellation could be accurately forecasted using the supervised learning techniques above.

To account for the seasonality in the data and the impact of time in terms of seasons of fixed periodicity, an ARIMA model could be developed. Further, a hotel may only be interested in the approximate trend of cancellations for a particular period rather than the exact estimate of cancellation rate calculated for a day, let alone the classification of each booking in terms of the whether the booking is likely to be cancelled or not. Thus, a weekly approximate for the cancellation rate could be calculated that would be more beneficial for devising a revenue-management decision. Likewise, demand could be estimated on a weekly basis to check for when a hotel is probable to get a disproportionately large number of bookings.

IV. METHODOLOGY

A. Dataset

Hotel booking demand is a dataset that describes bookings made by customers. The dataset comprises of 2 hotels- a city hotel and a resort hotel along with 31 attributes that describe a booking and a target concept of booking cancellation. It has 12 categorical columns and 19 numerical columns. It includes features such as arrival date, lead time(number of days between booking and check-in), number of adults, children, babies, type of meal booked, country of origin, previous cancellations, previous bookings not cancelled by the user, type of room reserved, booking changes made after the booking, deposit type, travel agent that made the booking, ADR(Average Daily Rate), number of special requests, number of days the booking was in the waiting list, customer type and so on.

B. Data Pre-Processing

The dataset has missing values. The missing values are handled appropriately. The next step is to handle categorical variables. The categorical variables are One-hot encoded. The numeric columns are transformed using Standard Scaler so as to get a distribution with mean at 0 and variance at 1. This ensures an unbiased representation of all features. Features with higher values are, therefore, not considered more important than the other features in the model.

The column 'reservation_status' takes 2 unique values viz. 'Checkout' and 'cancelled'. Since this column is a potential leakage of the classification itself, it is appropriate to drop this column. In addition, the features 'reservation_status_date', 'assigned_room_type' and 'country' is also deemed not useful and is therefore, dropped.

The dataset is analyzed for the cancellation behavior. Since the hotels differ in their locations, cancellation behavior could be attributed to location-specific reasons. Hence, the dataset is split into two, each split representing either the city hotel or the resort hotel. Classifiers are built for the two hotels separately and evaluated for evaluation metrics.

C. Exploratory Data Analysis

Analysis of the dataset revealed the following.

- 62.9% of the total bookings were cancelled.
- Months of July and August receive the highest number of bookings. January has the least number of bookings.
- Months of July and August have the highest number of booking cancellations. Thus, the number of booking cancellations follows the volume of bookings for the months.
- On categorizing the customers based on their country of origin, we see that Portugal makes up for 22.2 per cent of bookings that are not cancelled (highest for a region/country). The next in the list is Great Britain and then France. On looking at the bookings that were cancelled, Portugal makes up for 59.1% of the total booking cancellations. Next, comes Great Britain and Spain. Therefore, Portugal and Great Britain have the highest percentages of cancellation and non-cancellation.
- On comparing the deposit types for cancellations, we see that the customers that have not made a

deposit are more likely to cancel a booking than deposits made with no refund or refundable policies.

- Comparing the customer types for cancellations, we see that transient customers have a higher probability of cancelling a booking when compared to bookings on a contract or group.
- Most of the transient customers make bookings with no deposit policy. Group customers are a minority.
- An average Daily Rate is a statistical unit that is often used in the lodging industry. The number represents the average rental income per paid occupied room in a given time period. City hotel has a higher average ADR than the resort hotel.
- Average ADR is lowest in the month of January. Hence in January when the hotel receives the least number of bookings, the price is kept to the minimum. This month would ideally be more profitable for the customer.
- 14.4% of the repeated guests have cancelled their bookings. 37.7% of the first-time guests have cancelled their bookings. This shows that repeated guests have a lower probability of cancelling bookings.
- Couples (no children) form the majority of the customers. As expected in accordance with the popularity of the hotels among people types, couples show a higher probability of cancellation.
- Longer the Lead Time, the higher is the probability of cancellation

Correlations between the target variable and the numerical features showed the presence of only a moderate correlation. Lead time (number of days between the booking and check-in) is seen to be correlated with the target variable the most when compared to other numeric features. The spearman correlation between the target variable and the deposit type is 0.48. This is in accordance with the behavior observed as stated above.

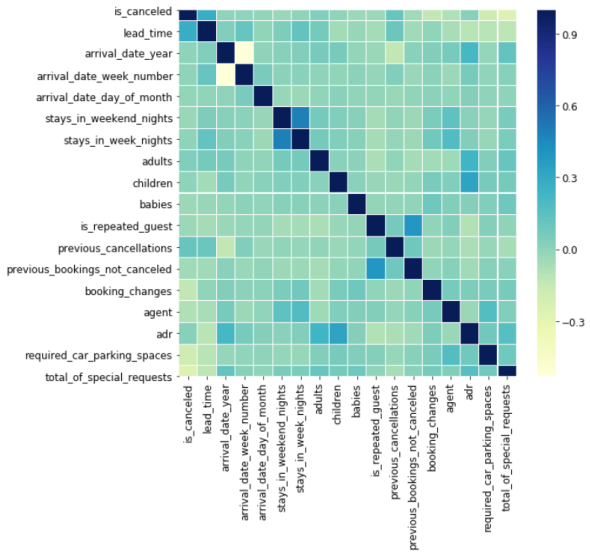

Fig. 1. Visualisation of the Correlation Matrix

D. Building Classifiers

The data is split into training and testing data in order to be able to build and test the models. The split is in the 80-20 ratio.

The classifiers implemented are Logistic Regression, Decision Trees and Random Forest Ensemble method. Grid search hyperparameter tuning was done using sklearn to exhaustively look for parameter combinations and select the best parameter values for the Random Forest Classification model.

The feature importance visualization shows that the agent used for booking, number of days of stay, and number of previous cancellations are the three most important features that contributed to model building.

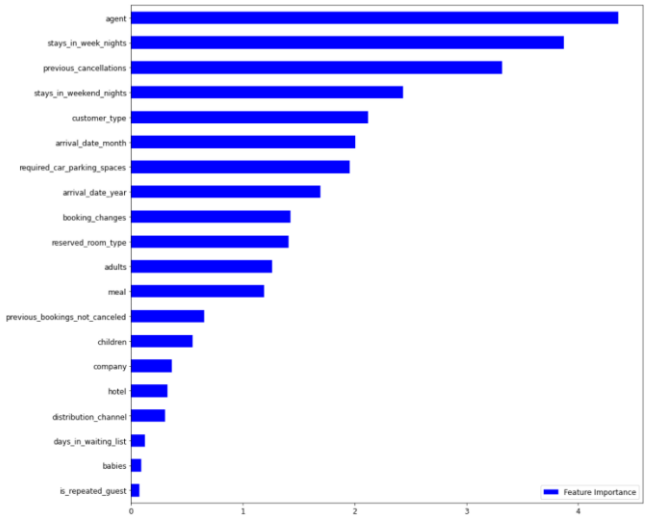

Fig. 2. Visualisation of Feature Importances

Estimator performance was validated using the k-Fold cross-validation technique using 4 splits to ensure that the performance in the testing set was not due to splitting issues. The accuracy and F1-score for the models after k-Fold cross validation is as follows.

MODEL	AVERAGE F1-SCORE	ACCURACY
LOGISTIC REGRESSION	0.481	0.625
DECISION TREES	0.835	0.836
RANDOM FOREST	0.857	0.860

Fig. 3. Accuracy Table

The ROC Curve for the three models is as seen below. The curve revalidates the Random Forest as the best performing model on the data.

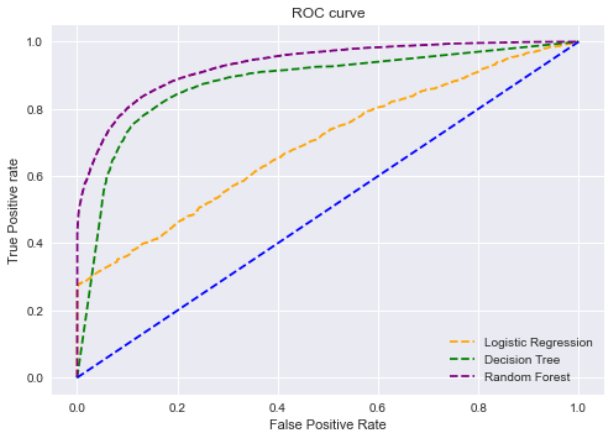

Fig. 4. Receiver Operating Characteristics Curve

The data consists of records for two hotels – a Resort Hotel and a City Hotel. Since the causes of a booking cancellation can be specific to a hotel in terms of its location or the facilities provided the hotel, it is apt to model a hotel individually.

The dataset is now divided on the basis of the hotel into two. Training and testing of the model are done individually for each hotel. The train-test split is in the 80-20 ratio.

The classifiers modeled are that of Logistic Regression, KNN, Decision trees and SVM. The accuracy and F1-score obtained for the models for the Resort Hotel is as follows.

MODEL	AVERAGE F1-SCORE	ACCURACY
LOGISTIC REGRESSION	0.7680	0.7873
KNN	0.8074	0.8168
DECISION TREES	0.8260	0.8240
SVM	0.8032	0.8112

Fig. 5. Accuracy Table for Resort Hotel

The accuracy and F1-score for the models on the City Hotel is as follows.

MODEL	AVERAGE F1-SCORE	ACCURACY
LOGISTIC REGRESSION	0.7886	0.7956
KNN	0.8262	0.8278
DECISION TREES	0.8283	0.8293
SVM	0.8028	0.8063

Fig. 6. Accuracy Table for City Hotel

E. Time Series Analysis

To take into account the features that capture trend and seasonality, we have done time series analysis for both daily data and data aggregated into weeks.

Booking cancellation is dependent on the time and the season of the year. The data shows a time-varying trend. To capture this seasonality in the data, forecasting models such as AR, MA, ARIMA models were built. Another notable aspect of cancellation prediction is that a hotelier would want to observe the trends in the cancellation behavior rather than the exact number of booking cancellations made per day. Hence, weekly predictions which show the overall trend of customer behavior was formulated, in addition to predictions on a daily scale.

a. Weekly Time Series:

Data is aggregated into weeks and the total number of booking cancellations for each week is tallied. An ARIMA model was initially built. The integration component (d) of the ARIMA model helps in converting the non-stationary time series to a stationary time series. Difference stationarity is used as the time series shows seasonality. An order of differencing of 1 was sufficient to convert it to a stationary series with a confidence of 99% as indicated by the dickey-fuller test. The PACF and the ACF plots were plotted so as to get the AR lags p and the MA lags q respectively.

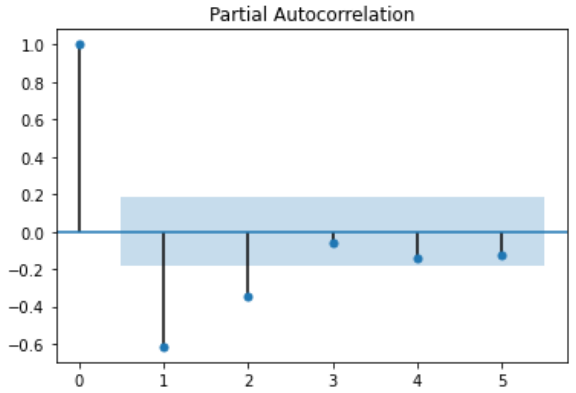

Fig. 7. PACF Plot

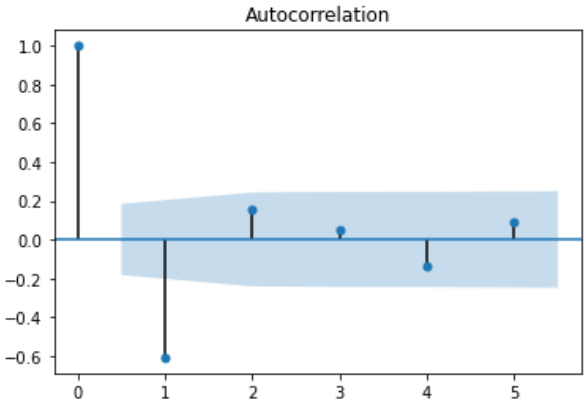

Fig. 8. ACF Plot

ARIMA(2,1,1) models were used as estimators for the two hotels separately. The predicted series is compared with the original series. RMSE for the resort hotel and the city hotel is 150.62 and 252.16 respectively.

b. Daily Time Series:

The number of booking cancellations is now modelled on a daily basis using an extension of the ARIMA model called SARIMA. This model takes into account the seasonal terms. The evaluation metric RMSE on the test data came up to be 10.32.

F. Analysing Results

Out of the models trained on the entire dataset, the Random Forest Model performed the best. The logistic regression model performed poorly and one of the reasons to this behavior is that it is capable of drawing only linear decision boundaries. Clearly the target concept does not vary linearly with the independent feature variables. The decision tree performed considerably well on the data with an accuracy of 83.6% and F1-score of 83.5%. The F1-score combines recall and precision. Recall is an important metric in predicting booking cancellations. The hotel managers want to make sure that a potential booking cancellation is predicted accurately so as to take necessary steps ensuring no loss in revenue. Additionally, the accuracy of the positives should be as high as possible. An F1-score that equals the accuracy score shows that the model is able to predict a non-booking cancellation just as accurately, given that the input data is slightly unbalanced. The Random forest classifier gives slightly better results than the decision tree model. Though a random forest model takes the average of the predictions from multiple decision trees, in this case it outperforms by only a negligible amount.

Another observation is that the evaluation metrics performance for the two hotels is consistent. This could be attributed to the consistent size of the two datasets, in addition to accurate definitions of the features.

Classifiers trained on the split dataset did not perform better than the ones trained on the whole dataset. One reason could be the dataset size itself. Splitting the dataset reduced the size by approximately half for each of the hotels. Nevertheless, the accuracies obtained were considerably good. Different hotels could receive different booking cancellation rates influenced by hotel specific factors such as cancellation policies, services and average customer satisfaction. Hence, the models trained on the individual hotels are a true representation of the cancellation behavior.

Forecasting models implemented on the data for the weekly scale were the AR, MA, and ARIMA models. The models were implemented separately for the two hotels. The dickey-fuller test showed that the series was stationary with a confidence of 95%. Log transformation was even then applied that made the data stationary with a confidence of 99%.

On plotting the PACF and the ACF plots, the lags p and q was found to be 2 and 1 respectively. The AR and the MA model showed higher RSS than a combined ARIMA model. Thus, an ARIMA model was the best choice.

Forecasting on daily data was done using the SARIMA model. The model was tested on 200 records. It performed reasonably well with RMSE of 10.32. One reason is that the SARIMA model considers seasonality. From this, we can conclude that seasonality plays an important role in the booking cancellation behavior.

V. RESULTS

The Logistic Regression model has the least accuracy of 78% and 79% for the two hotels respectively. This is because it can only construct linear boundaries. Additionally, it is a probabilistic model that highly depends on classification cutoff. In this case, multiple bookings lied around the classification cutoff, that made the model error-prone.

The K-Nearest neighbours model was implemented with the k-value as 9. The accuracy vs the number of neighbours plot showed the best accuracy for 9 neighbours which was 0.81 and 0.82 for the two hotels respectively. The KNN model has a relatively lower accuracy than the Decision Trees classifier. This could be due to the large number of input features that the KNN had to train on. Moreover, KNN is based on a majority vote system, where it consults its nearest neighbourhood for classification. Thus, it would perform poorly when the distances of the nearest neighbours don't comply with that of the test data.

With the SVM model, we got an accuracy of 81% and 80% respectively. The accuracy here maybe a bit off the accuracy values for the other models due to incorrect or faulty data recorded that would have reduced the maximum distance from the decision boundary.

The result analysis for time series is done in the previous section of the paper.

VI. CONCLUSIONS

Out of all the different classifiers implemented, the maximum accuracy of 86% was achieved. This proves that booking cancellation prediction indeed gives near accurate results when robust ML models are used. Fine-tuning the hyperparameters by automated techniques such as grid search tuning helps in increasing the overall accuracy of the models.

The forecasting models built took advantage of the seasonality inherent in the data. Weekly predictions help the hotel managers be better prepared for the upcoming week in

terms of the average upward or downward trend in the cancellation behaviour from the previous week. An improvement in the prediction could be done by using robust models such as the EEMD-ARIMA model.

Contributions

Name	Contribution
Sneha Jayaraman	• Data Pre-processing • Exploratory Data Analysis • Binary Classification Models- Logistic regression, Decision Trees, Random Forest (For the entire data as a whole) • Time Series Analysis – Weekly forecast, daily forecast • Project Report
Hemanth	• Data Pre-processing • Exploratory Data Analysis • Classification Models- Logistic regression, KNN, Decision Trees, SVM. (For the two hotels - individually) • Project Report
Thrupthi	• Data Pre-processing • Exploratory Data Analysis • Time Series Analysis – Weekly forecast, daily forecast. • Decision Trees • Project Report

REFERENCES

[1] Antonio, Nuno & De Almeida, Ana & Nunes, Luís. (2017). Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue. Tourism and Management Studies. 13. 25-39. 10.18089/tms.2017.13203.

[2] Muzi Zhang Junyi Li Bing Pan and Gaojun Zhang. (2018). Weekly Hotel Occupancy Forecasting of a Tourism Destination. MDPI

[3] Falk, Martin & Vieru, Markku. (2018). Modelling the cancellation behaviour of hotel guests. International Journal of Contemporary Hospitality Management. 10.1108/IJCHM-08-2017-0509.