Prediction on Hotel Bookings Cancellation

Chetan Godase
Lakehead University
Department of Computer Science
Thunder Bay, Canada
cgodase@lakeheadu.ca

Anukaran Kathuria
Lakehead University
Department of Computer Science
Thunder Bay, Canada
akathuri@lakeheadu.ca

***Abstract*—Hotel bookings definitely vary based on the months and occasions, same with its cancellations, cancellations on a confirmed booking have a lot on impact on the time and money management. To overcome this problem, they tend to have cancellation policies with penalties, but they also overbook the rooms. But this might also create negative reviews if the bookings aren't cancelled, and multiple people show up for same room. Hence to avoid such problem we have developed machine learning model for predicting number of exact cancellations which are going to happen. We test out multiple ML models to find out which model suits best and gives almost perfect accuracy. For this ML model building we are using "Hotel Booking Demand" dataset which has 32 continuous and categorical features and has total of 119390 instances on work on. The fully trained model would be able to give predictions on whether booking is likely to be canceled or not and help the hotel to make more money by overbooking and to maintain its reputation too. Results shows a promising ML model which has a decent confusion matrix and accuracy and shows the likeliness of the booking to be canceled.**

Keywords—cancellation of bookings; classification models; machine learning; predictive modeling; prototyping; revenue management.

I. INTRODUCTION

In the hospitality business, cancellations on hotel bookings can have a drastic issue on its management decisions. To decrease or to avoid this difficulty which are occurred by booking cancellations, hotels tend to have cancellation policies with penalties, but they also overbook the rooms which is known as overbooking strategy. All the flights, trains bookings implement this strategy to decrease the risk of loss and to increase profit by adding more bookings. But this might also create negative reviews for the hotels if the bookings are not cancelled and multiple people show up for same room. This would create problems, usually if sometime as such happens hotels give out other room in other hotels for free of cost and with higher upgraded facilities, but this can be avoided too. Overbooking forces the hotel to deny service provision, which can be a terrible experience for the customer and have a negative effect on both the hotel's reputation and immediate revenue [4]. It can also mean future revenue loss from discontent customers who will not book again at the same hotel [1]. On the other hand, rigid cancellation policies, especially non-refundable policies, have the potential not only to reduce the number of bookings but also to diminish revenue due to the application of significant discounts on price [2].

In this paper, we are trying to use multiple models of classification technics to see which suits best for the dataset and give optimum results. In fact, Morales and Wang stated that "it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy" [6, p. 556]. But it can be possible to use ML, statistics, and visualization to get surprisingly good and accurate results.

By having a decent idea of come many bookings can be cancelled the management of the hotel can take accordingly act and take measures to mitigate potential losses and convert them to profits. This is possible because when bookings are cancelled the owner of the booking has to still provide some penalty charges for the cancellation, the severity can be dependent on how late the booking is getting cancelled. On the other hand, selling the same booking to someone else for the full price, (if peak time then may be higher). Also having an estimate on the booking, the management can arrange facilities as per the likings like room upgrades, breakfast, room services, and some other benefits. They can also manage the labor cost based on the estimation. This classification wont only helps with all the above estimations, but it can also help to develop new cancellation policies and booking prices as per the hotels profits.

To the extent of the authors knowledge, there are multiple documented prediction on hotel booking cancellations, but they all stick to only certain or maybe even type of classification technique which they find it convenient or utilitarian. But in this report, by developing a model with multiple famous and practical classification practices we desire to get very accurate results. This study proposes that using machine learning model on a big data set to develop a predictive model can be used as an excellent tool gimmick for hotel business owners to optimize costings, services, and revenue for the management.

II. BACKGROUND

Originally developed in 1966 by the aviation industry [7], revenue management was gradually introduced in other services industries, such as hotels, golf courses, restaurants, and casinos [7]. In the hospitality industry (rooms division), revenue management general definition was adapted to "making the right room available for the right guest and the right price at the right time via the right distribution channel" [1, p. 2].

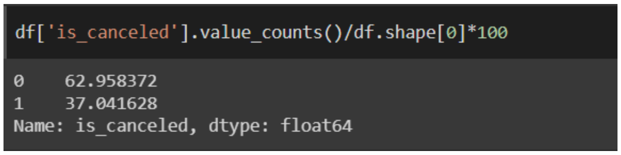
As in other service industries who have a fixed inventory and have a “perishable product”, the hospitality industry accepts bookings in advance [3]. The booking will act as a legit contract between the hotel and the owner of the booking who initiated the booking contract. This contract would come with some rights that the owner of the booking can use like the services from hotel on a fixed price on which booking was done, irrespective if the booking prices increase or decrease. But here is the catch that advance bookings are not certain and has a potential to be canceled. Here we consider both cancelled and no-show data as cancelled.

This study and analysis show that using this dataset a model can be developed which can be used in real world predictions which can be used on empirical standpoint.

III. Methodology

The need to examine, evaluate and test the data in real world domain is very important and to provide and achieve such results we need to use skills such as machine learning, visualization, and data mining.

When analysis the data from the dataset, we noticed that there is great amount of cancellations. After analysis data for last 5 years, we found that almost 40% of the bookings are getting cancelled which is definitely not good for the hotel revenue and managements. The fig 1.1 shows the exact percentile of cancellations happening from the last 5 years.



```
df['is_canceled'].value_counts()/df.shape[0]*100

0    62.958372
1    37.041628
Name: is_canceled, dtype: float64
```

Fig 1.1 Cancellation percentile.

In fig 1.1 it tells the percentage of booking which were cancelled or not, ‘0’ indicates the percentage of bookings which were not cancelled and ‘1’ indicates the bookings which were cancelled. This analysis is for both type od hotels “resort hotels” and “city hotels”. After visualizaing the data defirently for both hotels we got more insight into it. The fig 1.2 shows cancellations and not cancellations based on types of hotels. It shows that the amount of cancellations on city_ hotels is really high compared to resort hotels, at the same time confirmed booking are also high in city hotels compared to that of resort hotels. This also concludes that city hotels are in more demand than compared to resort hotels.

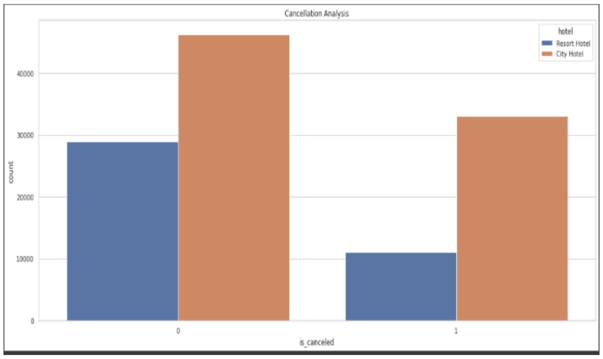


Fig 1.2 cancelled and not cancelled bookings.

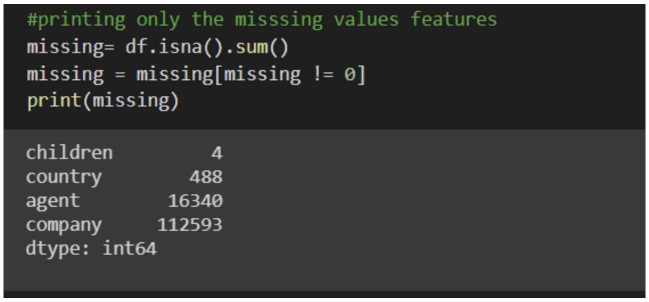
A. *Machine learning model*

1.Spliting the Data

The complete dataset must be divided into 2 parts training and testing. This is the most important step and must be done at the very beginning of your model building. As, if the splitting of data is done after all the preprocessing and model building then there might be information leak in the model which leads to misleading data and imperfect machine learning model. The splitting of data is done by the build-in function known as “train_test_split()” this function has multiple parameters, it also includes size into which the data must be split and its randomness of the selection of the instances. All the pre-processing, model building, and execution must be done on training dataset to avoid information breach in the ML model building. Test data set is kept untouched as these instances work as real-world problems on which we can check our model’s accuracy and usefulness.

2.Converting the categorical values

The data set contains lot of categorical features, which can’t be any use for the Model building. All the categorical features are converted to continuous features using the Label Encoder(). LabelEncoder() helps to assign a unique value for each unique items in the given feature and substitute the assigned values throughout the feature, that’s how all the features are passed one by one into a loop to get converted from categorical features into continuous features. This new data frame will be able to use all the information now for model building.



```
#printing only the misssing values features
missing= df.isna().sum()
missing = missing[missing != 0]
print(missing)

children        4
country       488
agent       16340
company    112593
dtype: int64
```

Fig1.3 Missing data in dataset.

3.Imputation

After converting all the data into continuous features now preprocessing can be done. In the given data set, there is a lot of missing values and it cannot be ignored as every data point is important. The fig1.3 shows the number of missing values in the dataset. If these missing values are removed, then it will not make a good Machine Learning model. Hence, we fill the values with artificial values from the data set. To impute the data this study uses KNNImputer(). This function helps to impute all the numerical features values. This function can only impute numerical features, hence conversion of categorical features to continued features is important before using KNNImputer(). This function will calculate the Euclidian distance from the data points and add artificial values with closest match found. This function has multiple parameters, one of which includes n_neighbors, by default its usually 5 but for convenience the study tunes the parameter to 2. And then use fit_tranform() on it to fill all the missing values with KNNimputer(). This deals with all the missing values in the dataset. Missing values can be checked by dataset.isna().sum() function.

3.Scaling

Feature scaling is one of the most important part of machine learning model and it is considered as a critical step in preprocessing of data before generating the machine learning model. For this study, MinMaxScaler() functions is used. This function will execute on each feature separately and set all the values in the range of 0-1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Fig1.4 Formula for Min max scaling

Scaling also reduces the problems of outliers in the data as it rebounds all the value between 0-1, the range can be changed based on the use on the user. This process of scaling the data by min max is called normalization of the data.

4.Oversmapling

For oversampling for this study the technique used is SMOTE (Synthetic Minority Over-Sampling TEchnique). It produces synthetic data and enters it into the dataset so that the imbalance class is now balanced. SMOTE() helps to complete this task. This will create class with less instances equal to majority class so that training can be done.

5.Correlation Matrix

The correlation matrix helps to check which features are important for our prediction class and which features are not, this is done based on output with a range of 0 to 1. In our study the feature “lead_time” has the highest correlation to the “is canceled” with correlation of 0.29.

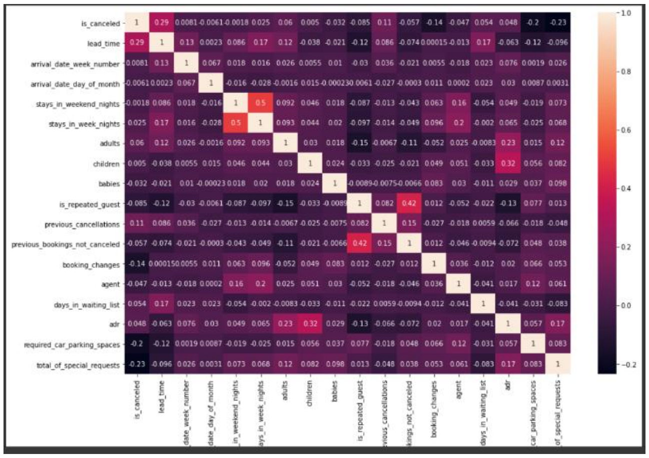
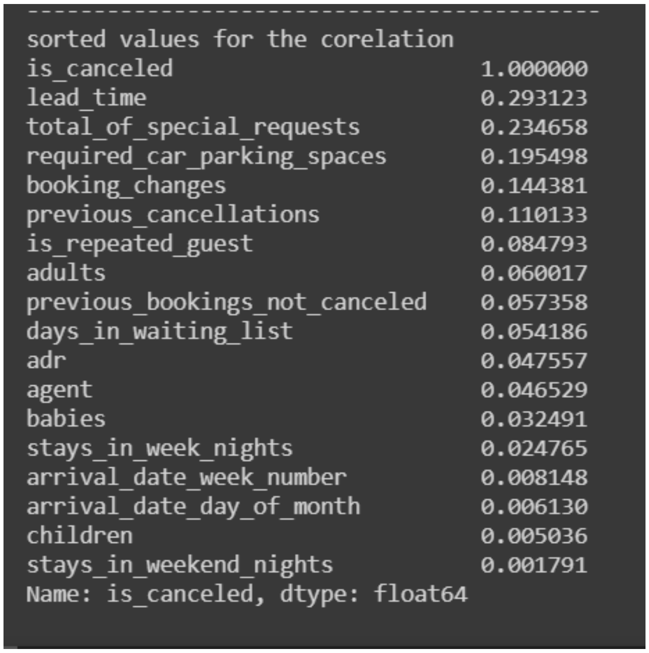


Fig 1.5 Heat map of correlation matrix

The fig1.5 shows complete heat map of the correlation for each features to other features. But for the study we consider only “is_cancelled” feature for prediction.



```
sorted values for the corelation
is_canceled                       1.000000
lead_time                         0.293123
total_of_special_requests         0.234658
required_car_parking_spaces       0.195498
booking_changes                   0.144381
previous_cancellations            0.110133
is_repeated_guest                 0.084793
adults                            0.060017
previous_bookings_not_canceled    0.057358
days_in_waiting_list              0.054186
adr                               0.047557
agent                             0.046529
babies                            0.032491
stays_in_week_nights              0.024765
arrival_date_week_number          0.008148
arrival_date_day_of_month         0.006130
children                          0.005036
stays_in_weekend_nights           0.001791
Name: is_canceled, dtype: float64
```

Fig 1.6 Correlation for “is_cancelled” features

The fig1.6 gives exact visualization of the importance of the features for the prediction and is sorted as per ascending order.

6.Feature Selection and Engineering

Feature selections helps to determine which features can be useful for the prediction of the dataset and which aren't useful. For feature selection on the dataset we use the method selectfromModel(Lasso) for selecting all the features, we give alpha value to it, alpha values decides how many features can be selected for the training. The alpha value is usually very low, we have given it as 0.005, it ranges from 0-1. To find the best alpha value we can also find it by using cross fold validation. But for this test we found this alpha value serves better.

The output selects 7 features which it considers as important and then we shift the new selected features to new data set and train the model now. This is the end of all the preprocessing and cleaning of the data.

7. Training Models

For the study we consider multiple models for its accuracy and determine which model will be the bit fit for the prediction of the data set . This is a very complex dataset and has 10000+ instances to work on. The classification and prediction will surely be difficult and training the model and testing it for the accuracy which also be difficult . Here after a careful decision and best fit for the model, we have selected 5 types of machine learning models .

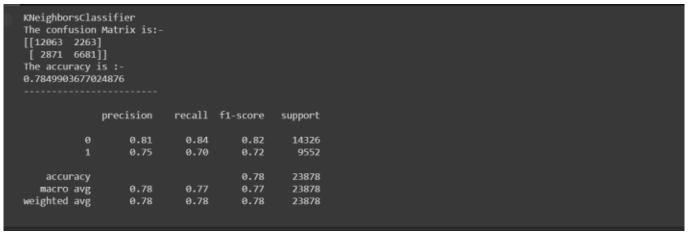
1.LogisticRegression()

2.DecisionTree()

3.GausianNB()

4.KNeighborsClassifier()

For the results we pass training dataset and train the model first and after training we pass the trained model to perform its prediction on the test dataset, And for the results we print the accuracy, f1 score, precision, recall and support. And to find all these scores we use confusion matrix to evaluate.



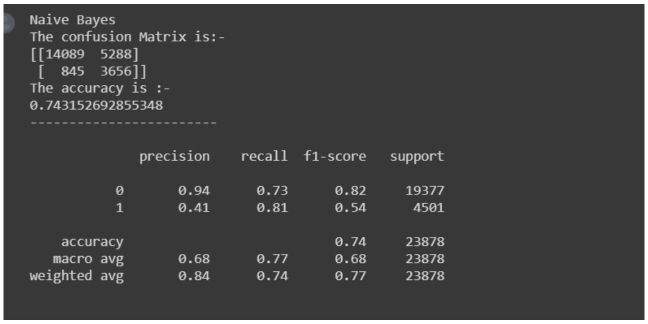
```
KNeighborsClassifier
The confusion Matrix is:-
[[12063  2263]
 [ 2871  6681]]
The accuracy is :-
0.7849903677024876
------------------------

              precision    recall  f1-score   support

           0       0.81      0.84      0.82     14326
           1       0.75      0.70      0.72      9552

    accuracy                           0.78     23878
   macro avg       0.78      0.77      0.77     23878
weighted avg       0.78      0.78      0.78     23878
```

Fig1.8 KNN model training and testing results



```
Naive Bayes
The confusion Matrix is:-
[[14089  5288]
 [  845  3656]]
The accuracy is :-
0.743152692855348
------------------------

              precision    recall  f1-score   support

           0       0.94      0.73      0.82     19377
           1       0.41      0.81      0.54      4501

    accuracy                           0.74     23878
   macro avg       0.68      0.77      0.68     23878
weighted avg       0.84      0.74      0.77     23878
```

Fig 1.9 Naive Bayes model training and testing results

Above fig 1.8 and 1.9 are some results of training and testing on the dataset. This output is shown after the training the model on its respective methods.

8.Results and Conclusion

After complete evaluation we noticed that the KNN and decision tree are good methods we achieved a accuracy of 78% and 77% respectively. All the above data is based on the dataset trained and tested which had feature selection and scaling and oversampling. We have also done oversampling and without feature selection the accuracy varies as the features not selected will definitely affects the training of the model and the results. We are submitting 2 types of the files where we do training with feature selection and oversampling and other is without feature selections and with oversampling. Below are some analysis we did but didn't find important to add into the project.

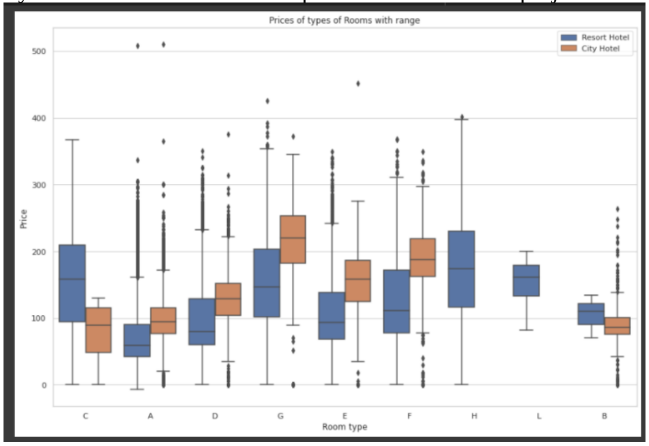


Fig1.10 Price differences for each hotel rooms

REFERENCES

[1] R. Mehrotra and J. Ruttley, *Revenue management (second ed.)*. Washington, DC, USA: American Hotel & Lodging Association (AHLA), 2006.

[2] S. J. Smith, H. G. Parsa, M. Bujisic, and J.-P. van der Rest, "Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry," *J. Travel Tour. Mark.*, vol. 32, no. 7, pp. 886–906, Oct. 2015.

[3] K. T. Talluri and G. Van Ryzin, *The theory and practice of revenue management*. New York, NY: Springer, 2005.

[4] B. M. Noone and C. H. Lee, "Hotel overbooking: The effect of overcompensation on customers' reactions to denied service," *J. Hosp. Tour. Res.*, vol. 35, no. 3, pp. 334–357, Nov. 2010.

[5] N. Antonio, A. Almeida, and L. Nunes, "Predicting hotel booking cancellation to decrease uncertainty and increase revenue," *Tour. Manag. Stud.*, vol. 13, no. 2, pp. 25–39, 2017.

[6] N. Antonio, A. de Almeida, and L. Nunes, "Using data science to predict hotel booking cancellations," in *Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry*, P. Vasant and K. M, Eds. Hershey, PA, USA: Business Science Reference, 2016, pp. 141–167.

[7] W.-C. Chiang, J. C. Chen, and X. Xu, "An overview of research on revenue management: current issues and future research," *Int. J. Revenue Manag.*, vol. 1, no. 1, pp. 97–128, 2007.

[8] L. Garrow and M. Ferguson, "Revenue management and the analytics explosion: Perspectives from industry experts," *J. Revenue Pricing Manag.*, vol. 7, no. 2, pp. 219–229, Jun. 2008.

[9] C. Hueglin and F. Vannotti, "Data mining techniques to improve forecast accuracy in airline business," in *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, 2001, pp. 438–442.

[10] B. Freisleben and G. Gleichmann, "Controlling airline seat allocations with neural networks," in *Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, 1993*, 1993, vol. iv, pp. 635–642 vol.4.

[11] C. Lemke, "Combinations of time series forecasts: When and hhy are they beneficial?," Bournemouth University, 2010.

[12] J. Subramanian, S. Stidham Jr, and C. J. Lautenbacher, "Airline yield management with overbooking, cancellations, and no-shows," *Transp. Sci.*, vol. 33, no. 2, pp. 147–167, 1999.

[13] M. G. Yoon, H. Y. Lee, and Y. S. Song, "Linear approximation approach for a stochastic seat allocation problem with cancellation & refund policy in airlines," *J. Air Transp. Manag.*, vol. 23, pp. 41–46, Aug. 2012.

[14] Zvi Schwartz, Muzaffer Uysal, Timothy Webb, and Mehmet Altin, "Hotel daily occupancy forecasting with competitive sets: a recursive algorithm," *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 2, pp. 267–285, Jan. 2016.

[15] L. N. Pereira, "An introduction to helpful forecasting methods for hotel revenue management," *Int. J. Hosp. Manag.*, vol. 58, pp. 13–23, Sep. 2016.

[16] W. Caicedo-Torres and F. Payares, "A machine learning model for occupancy rates and demand forecasting in the hospitality industry," in *Advances in Artificial Intelligence - IBERAMIA 2016*, 2016, pp. 201–211.