2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) | 978-1-6654-0052-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIS53314.2022.9742843

Hotel Booking Cancelation Prediction using ML algorithms

M. Venkata Rakesh [1]
Student,
Department of CSE,
Sathyabama Institute of Science and Technology,
Chennai.
venkatarakesh4281@gmail.com

S. Prasanna Kumar [2]
Student,
Department of CSE,
Sathyabama Institute of Science and Technology,
Chennai
saikumar97460@gmail.com

Ms. Yogitha[3],
Assistant Professor,
Department of CSE,
Sathyabama Institute of Scinece and Technology,
Chennai
Yogihta.ravi1915@gmail.com

Ms. Aishwarya. R[4],
Assistant Professor,
Department of CSE,
Sathyabama Institute of Scinece and Technology,
Chennai
aishwaryarajkumar7@gmail.com

Abstract: Cancellations in bookings show a negative impact in the hospitality industry field while making management decisions. To prevent the negative impact of the cancellations, a lot of policies are implemented along with few overbooking techniques, this, in turn, can severely damage the income and reputation of that particular hotel. To prevent this situation, machine learning models have been developed. These models use previous data from the hotel and then it gets trained and predicts if the particular booking would get cancelled or not. Two hotels namely Resort hotel and City hotel have been considered, and then ML models are used to predict how particular actions taken by the hotel management shows impact on the hotel revenue and cancellations. This, in turn, makes management rethink about policies and their decisions. The Ml models will help management to predict the number of cancellations that may happen.

Keywords: Machine Learning, Artificial Intelligence, Power Bi, Linear Regression, Logistic Regression, Navies Bayes'.

I Introduction

The ultimate aim of this project is to predict hotel booking cancellations, and develop some Machine Learning models with some best accuracy. After Machine Learning models, use the same data to develop some interactive reports To ease user's task whenever they need any information they can get it from Power Bi reports and also they can get future predictions using Machine Learning models And to check the country-wise bookings and cancellations and also to view which month has more bookings . So, the proposed ML models will predict the future bookings based on this experience

To ease user's task whenever they need information they can get from the power bi report directly by just one click. And they can export the data in the form tables if needed. The hotel management has to build such a model that it should send all the details regarding a hotel to the right customer at right time depending upon his spending score [1]. Even though there is a bit of difference between cancellation and no-show, both are treated as cancellations for convenience. Cancellations mean the customer initially booked a room in a hotel but canceled again. No-show means the customer failed to check-in [2]. Some of the cancellations occur because of comprehensive reasons like changes in business vacations, change in timings, flight cancellations. But majority cancellations occur because of customers deal-seeking shows in figure1.

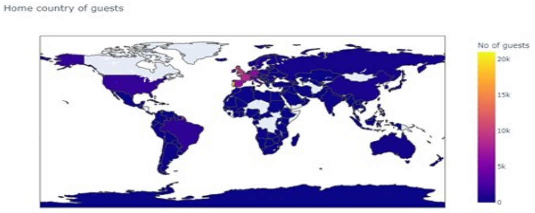


Fig.1 Customer Deal-seeking

Fig1 Shows that: By using plotly, a world map is created to know from which country more numbers of customers are visiting. People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe.

II Data and Methods

2.1 Linear Regression: Today in many fields Machine learning algorithms are widely used it solves complex tasks that cannot be solved by humans or by traditional machine approaches one of the simple and mostly used machine learning approach is linear regression. For exploratory data analysis Linear Regression is one of the algorithms that many people use. This is implemented in a place where there is a need to predict the output or forecast the information

By using one or more independent variables. This will find the relation between the input and independent variables with output or dependent variables by using a gradient descent algorithm. Regression is used in two cases, the first case is to make predictions based on the previous experience, and the second case is to find the relations between different variables. Experience means data in Machine learning.

2.2 K- Nearest Neighbor Algorithm: A super simple way of classifying data, it is also a supervised learning technique. And it is a parametric less Algorithm, which is without assumptions it takes up the data. Since it is not learning from data available for training this can also be called as lazy or idle learner but during classification the data will be stored and then it carryout some actions on the data. This algorithm will search for similarities at the time of classification.

2.2.1 Why KNN?
For example, if there are two categories, A and B, if a new data arrives between A and B, using KNN it was easy to classify that new data belongs to category A or B.

2.2.2 Working:
Step1: Number of neighbors should be collected at first.
Step2: Euclidean distance will be calculated from every data point to N estimators.
Step3: According to the distance, the nearest neighbor to be taken is calculated.
Step4: Now allocate the data points to the group for which the number of neighbors is the highest or greater.
Step6: The model is ready.
Based on the EUCLIDEAN DISTANCE also it works.
EUCLIDEAN DISTANCE b/w points A and B given below (1)

$$d = \sqrt{}[(x2 - x1)2 + (y2 - y1)2 \qquad (1)$$

K- Value as a no particular way to determine, will try different ways and find the best output. Most preferred K=5.
If very less value of K is considered, it will lead to outliers in the model and also it gives noisy in output.
2.2.2.1 Advantages:
Implementation of KNN is quite simple
No discriminative function will be derived from data
If training data is more, the KNN will be more effective
2.2.2.2 Disadvantages:
The Standardization and normalization have to be completed prior to the implementation of KNN.
KNN algorithm is computationally expensive because it need to calculate distance for each data point of the dataset.

2.3 Decision Tree:

It comes under supervised learning, and it has also been used for Regression as well as classification problems. Mostly performable for classification.

- It maintains a tree structure.
- Internal nodes, data set, branches, decision rules are applicable.
- It takes the features of data sets in the form of nodes.
- Decision rules are taken as arms.
- One output will be given by every leaf node.
- In a decision tree there are two types of nodes: 1) Decision node. 2) Leaf node.
- Decision nodes are useful for taking up decisions, branches are there.
- Leaf node is an output for decisions made, no branches are there.
- In a decision tree, it tries to give a solution in all possible ways, and it's been represented in a graphical way, it is called a decision tree.
- A CART-algorithm is used to build a decision tree.
- CART is a classification, regression algorithm.
- Initially the decision tree starts with a DECISION NODE called as root node, then branches, if no branches it will be leaf node.

2.3.1 Why Decision Tree?

Reason 1: It thinks like a human and gives the decision.
Reason 2: Logic is easily understandable.

2.3.2 WORKING:
Step1: The starting of the tree should be a root which contains all the variables or features in the dataset.
Step2: To find the best attributes of the dataset, attribute selection or Principal component analysis (PCA) must be considered.
Step3: Later, the root to lot of sub-trees is divided using best attributes or features
Step4: Now, the decision tree which has the best feature is picked.
Step5: The process of making new decision trees are repeated using step3 until a stage where no more further trees can be build from the features in dataset is reached.

2.3.3 Rules for feature Selection:

1. After dividing the dataset based on attribute, the change in entropy is measured; the change is also called as information gain. Information gain calculates how much details will be provided by a feature in the dataset about a class. This information gain value is used to divide the node to build decision tree. Usually, the Decision tree algorithm tries to boost the IG or information gain value, and it splits the node which is having the highest Value of Information gain. Information gain can be calculate using below (4)

$$IG= E(S) - [(WA) * E(I)] \qquad (4)$$

Where,
E= entropy
WA= weighted average
I= each feature
IG= Information Gain
Entropy: The impurity present in a feature of the dataset is called entropy. It describes the randomness in data. Entropy can be calculated below (5)

$$s\_e - (16/40 * s1\_e + 24/40 * s2\_e) \qquad (5)$$

2. Gini Index:

This is the measurement of impurity or cleanness which is used in making decision tree by CART (Classification and regression Tree) algorithm.

The attribute having Low Gini Index should be given more preference when compared to High Gini Index.

The Gini Index will be used by CART algorithm to prepare binary splits

Gini Index can be calculated using the below formula (2) and (3):

$GI= 1- \sum jPj*j$ (2)

Gi=Gini Index (3)

2.3.3.1 Advantages of the decision Tree:

- For decision-related problems it is very useful.
- For a problem it helps to think about all possible ways.
- It won't require much data wrangling compared to other models.

2.3.3.2 Decision Tree Disadvantages:

- Since its having many number of layers it will become complex
- Using Random Forest algorithm over fitting issue can be resolved.
- It requires more time for model training or it can be said as computational complexity

2.4 Naive Bayes Classifier: This belongs to the probabilistic classifier family and it is based on Bayes' Theorem. This is also based on kernel density estimation along with simple

Bayesian network. Naive Bayes is a type of classifier that is highly scalable and it also requires a lot of variables that are linear in the problem. For the construction of classifiers, a simple technique is used.

Like Naive Bayes. There is no other particular algorithm for training some classifiers, but here a particular feature is assumed as independent over any other feature.

For example, if a fruit like mango which is yellow, round, and has a diameter of 14cm. Now the naive Bayes' algorithm consider that all these features should contribute equally or independently for finding whether the fruit is mango or not.

Naive Bayes' classifier is used very efficiently for some supervised learning probability models. Naive Bayes uses conditional probability. $k=(c*i) / l$

Where,

Posterior=K, Prior=c, Likelihood=I, Evidence=l

Interestingly the model or algorithm accuracy is not correlated with the number of independencies

III Literature survey

A) Predicting cancellations in hotel bookings using machine learning algorithms is simple and common thing in current scenario. These cancellations might bring lot of economic losses to management,

Hence it is important to predict this cancellations. This base paper is helpful in explaining how machine learning is applied in this situation to identify which booking is going to get cancelled

And this in term prevent losses. For accuracy it should be evaluated in real environment.

B) Hotel industry is also get affected by the IT industry growth. But the change and impact is quite slow. Many people are doing research in testing and implementing new AI technology along with learning environment in hotels.

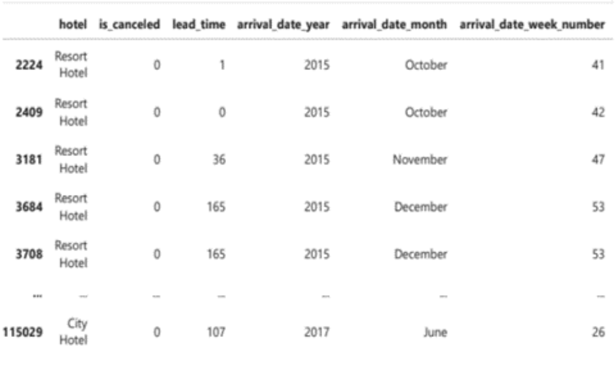
Machine learning is quite trending these days.

C) The reviews given by customers or Visitors play a major role in reputation as well as revenue system on the hotel.

But most of the people won't read all the comments or reviews given by previous customers of the hotel. These overlooked comments will be grouped and taken necessary steps and emotional analysis is done. Then improvement in the service is done.

IV Proposed work

The main aim of the hospitality industry is to provide the quality services to its customers which in turn leave a positive impact in their minds. This will create some positive reputation and reviews. Customer satisfaction is the main aim of this hospitality industry.



	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
2224	Resort Hotel	0	1	2015	October	41
2409	Resort Hotel	0	0	2015	October	42
3181	Resort Hotel	0	36	2015	November	47
3684	Resort Hotel	0	165	2015	December	53
3708	Resort Hotel	0	165	2015	December	53
...	...	...	...	...	...	...
115029	City Hotel	0	107	2017	June	26

Fig 2: Monitored data

4.1 Power bi: Is a Microsoft data analytics tool. This will help any user to prepare interactive reports using power view. It also allows for data cleaning using power query. Using power bi service, the dashboard can be shared to the team or with client, it can periodically load the newly added data and also it will send report to the end user at frequent intervals of time

4.2 Components in Power BI

- Power Bi has big collection of applications which can be used in mobile devices or on Power Bi desktop application as SaaS product.
- Power BI Desktop is a software that can be used for free and on-premises.
- Power Query will be used for data cleaning.
- Power Pivot is for data modeling.
- Power View is for developing interactive charts.

The different components of Power BI are meant to let users create and share business insights in a way that fits with their role shows in figure 4.

4.3 System architecture and methodology

The below is the system Architecture of the project shows in figure 3.

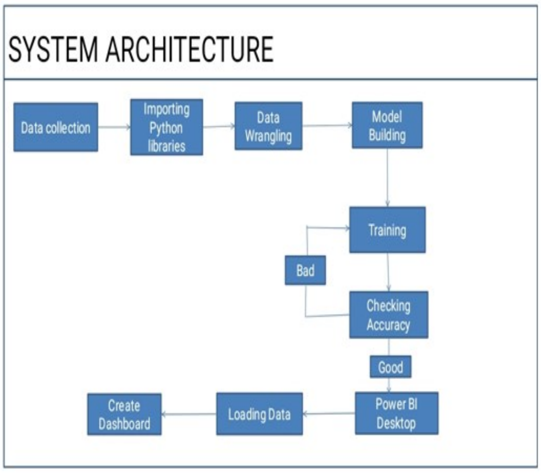


Fig 3: System Architecture

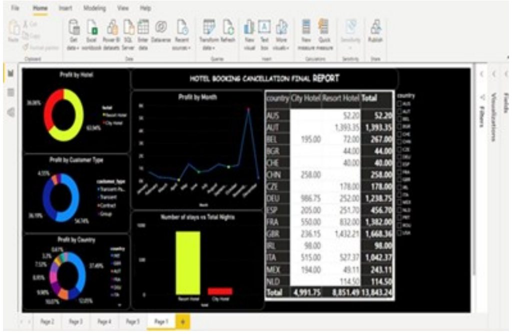
1) First, the data is assembled and then stored in the working directory.

2) By using python libraries, data wrangling and data analysis are performed.

3) Then, modeling the data depending on the requirement is done.

4) After all these steps, the machine learning models are built.

5) Then, the model is made to interact with the real world and accuracy is checked. If the accuracy is good, the power bi dashboards are built. If not, the training of the model will be done again and then the modeling of data will be done depending on the requirement.

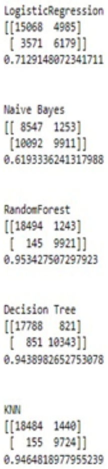


country	City Hotel	Resort Hotel	Total
AUS		52.20	52.20
AUT		1,393.35	1,393.35
BEL	195.00	72.00	267.00
BGR		44.00	44.00
CHE		40.00	40.00
CHN	258.00		258.00
CZE		178.00	178.00
DEU	986.75	252.00	1,238.75
ESP	205.00	251.70	456.70
FRA	550.00	832.00	1,382.00
GBR	236.15	1,432.21	1,668.36
IRL	98.00		98.00
ITA	515.00	527.37	1,042.37
MEX	194.00	49.11	243.11
NLD		114.50	114.50
Total	4,991.75	8,851.49	13,843.24

Fig 4: Components of Power BI(Business Insight)

V Result

At this stage it allows us to know various predictions for the data shows in figure 5.



```
LogisticRegression
[[15068  4985]
 [ 3571  6179]]
0.7129148072341711


Naive Bayes
[[ 8547  1253]
 [10092  9911]]
0.6193336241317988


RandomForest
[[18494  1243]
 [  145  9921]]
0.953427507297923


Decision Tree
[[17788   821]
 [  851 10343]]
0.9438982652753078


KNN
[[18484  1440]
 [  155  9724]]
0.9464818977955239
```

Fig 5: Predicted Data for hotel cancelling

Data sets : The data are collected from the kaggle .The dataset contains features like 'hotel', 'arrival_date_month', 'meal', 'market_segment', 'distribution_channel', 'reserved_room_type', 'deposit_type', 'customer_type', 'year', 'month', 'day', 'cancellation'. The machine learning algorithms are built and then an interactive dashboard is created using POWERBI.

VI Conclusion

The machine learning model is built and is trained using the past data of the hotel. In this way it can understand the patterns of cancellations. So, the proposed ML model can now predict which booking is going to get cancelled and prevent the loss to hotel management field. And interactive dashboards are built using powerbi. So the end client from hotel side can now be able to see and filter the data and can make on time decisions on their own without human interference.

References

1. Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and

increase revenue. *Tourism & Management Studies*, *13*(2), 25-39.

2. Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, *89*, 102546.
3. Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020, December). Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.
4. Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*.
5. Adil, M., Ansari, M. F., Alahmadi, A., Wu, J. Z., & Chakrabortty, R. K. (2021). Solving the problem of class imbalance in the prediction of hotel cancelations: A hybridized machine learning approach. *Processes*, *9*(10), 1713.
6. Antonio, N., de Almeida, A., & Nunes, L. (2017, December). Predicting hotel bookings cancellation with a machine learning classification model. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1049-1054).
7. Saputro, P. H., & Nanang, H. (2021). Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets. *Journal of Applied Data Sciences*, *2*(1), 40-56.
8. Lee, H., Yang, S. B., & Chung, N. (2021). Out of sight, out of cancellation: The impact of psychological distance on the cancellation behavior of tourists. *Journal of Air Transport Management*, *90*, 101942.
9. Novakovic, J., & Turina, S. Hotel reservation cancellations: analysis and prediction using machine learning algorithms. *ACADEMIC JOURNAL*, 4.
10. Alotaibi, E. (2020). Application of Machine Learning in the Hotel Industry: A Critical Review. *Journal of Association of Arab Universities for Tourism and Hospitality*, *18*(3), 78-96.
11. Ansari, A., Shaikh, A., Mapkar, S., & Khan, M. (2019, April). Cancellation Prediction for Flight Data Using Machine Learning. In *2nd International Conference on Advances in Science & Technology (ICAST)*.
12. Antonio, N., de Almeida, A., & Nunes, L. (2019). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, *60*(4), 298-319.
13. Pereira, L. N., & Cerqueira, V. (2021). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, 1-18.
14. Putro, N. A., Septian, R., Widiastuti, W., Maulidah, M., & Pardede, H. F. (2021). PREDICTION OF HOTEL BOOKING CANCELLATION USING DEEP NEURAL NETWORK AND LOGISTIC REGRESSION ALGORITHM. *Jurnal Techno Nusa Mandiri*, *18*(1), 1-8.
15. Song, K. S. (2021). Simultaneous statistical modelling of excess zeros, over/underdispersion, and multimodality with applications in hotel industry. *Journal of Applied Statistics*, *48*(9), 1603-1627.
16. Putra, M. S. T., & Azhar, Y. (2021). Perbandingan Model Logistic Regression dan Artificial Neural Network pada Prediksi Pembatalan Hotel. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(1), 29-37.
17. Pinheiro, A. B., Pinto, A. S., Abreu, A., Costa, E., & Borges, I. (2020, October). The Impact of Artificial Intelligence on the Tourism Industry: A Systematic Review. In *International Conference on Tourism, Technology and Systems* (pp. 458-469). Springer, Singapore.
18. Abreu, A., Costa, E., & Borges, I. (2020). The Impact of Artificial Intelligence on the Tourism Industry: A Systematic Review. *Advances in Tourism, Technology and Systems: Selected Papers from ICOTTS20, Volume 1*, *208*, 458.
19. Petricek, M., Chalupa, S., & Melas, D. (2021). Model of Price Optimization as a

Part of Hotel Revenue Management—Stochastic Approach. *Mathematics*, *9*(13), 1552.

20. Antonio, N., de Almeida, A., & Nunes, L. (2019). An automated machine learning based decision support system to predict hotel booking cancellations. *An automated machine learning based decision support system to predict hotel booking cancellations*, (1), 1-20.
21. Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2015). Business intelligence for cross-process knowledge extraction at tourism destinations. *Information Technology & Tourism*, *15*(2), 101-130.
22. Azhar, Y., Mahesa, G. A., & Mustaqim, M. C. (2021). Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritme Random Forest. *Jurnal Teknologi dan Sistem Komputer*, *9*(1), 15-21.
23. Putranto, Y., Sartono, B., & Djuraidah, A. (2021). Topic modelling and hotel rating prediction based on customer review in Indonesia. *International Journal of Management and Decision Making*, *20*(3), 282-307.
24. Benítez-Aurioles, B. (2018). Why are flexible booking policies priced negatively?. *Tourism Management*, *67*, 312-325.
25. Schwartz, Z., Uysal, M., Webb, T., & Altin, M. (2016). Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. *International Journal of Contemporary Hospitality Management*.