AN INTELLIGENT CLASSIFICATION OF HOTEL DATA USING MACHINE LEARNING ALGORITHMS

Fatima Ijaz*1
Department of computer science
Riphah International University
Lahore,Pakistan
fatimaijaz084@gmail.com

Misbah Iram
Department of computer science
GC University
Faisalabad, Pakistan
madah2912@gmail.com

Mushtaq Hussain
Department of computer science
GC University
Faisalabad , Pakistan
mushtaqabdi512@gmail.com

Tanweer Hussain
Department of computer science
GC University
Faisalabad, Pakistan
thussainbhakkar786@gmail.com

Abstract

Presently, there is an assortment of hotel booking systems. Conversely, the superiority of the systems and interconnection by means of an additional system is extremely unusual. Inside this article, the classification method and machine learning algorithms be put on hotel booking information intended for enhanced accomplishment. The projected method is base scheduled evaluate hotel services for distinctive kinds of hotel records attributes. K means Clustering and predication are apply on hotel booking data. In the evaluation find the nearest clusters and accuracy of machine learning algorithms. In predication find the every algorithm performance criterion factors which that accuracy, classification error, AUC, precision, recall, F measure, sensitivity, specificity and standard deviation of all these instances. In last construct the comparison of all these algorithms and find the best accurate algorithm.

Keywords: hotel booking, classification, navies bayes, decision tree, random forest, deep learning, fast large margin, generalized linear model.

1. INTRODUCTION

At this time the visiting the attractions are able-bodied consideration -out in the direction of be one of the for the most part significant cautious sectors in a lot of countries. This is for the reason that it brings incomes and creates innovative jobs. Online Booking Systems acquire a massive element appearing in this advertise as it is potential for tourists to naturally restraint features and availabilities of pour out rooms inwards diverse hotels approximately the sphere. This addict - gracious and well-organized explanation have determined nearly every one of the hotels in the direction of go away entirely computerized within stipulations of explore and booking rooms. The software developers have residential processor systems which are straightforward in the direction of exploit and deal with multifaceted applications approximating booking of rooms, preservation of room position, monetary coverage, economic investigation etc. These programs are effortless to manage and the users are provided by means of customer manuals to function. Individual of the large amount dangerous points for several hotels possibly will be its position and how distant it is from the magnetism tourist spaces. Classifiers are taught in the direction of decide the polarities of texts. Classifiers such as, random forest, Naïve Bayes (NB) classifier are the majority usually used models. Supplementary models such as generalized linear model, fast large margin container additionally exist worn with the machine learning approach for classification tribulations. Supervised learning method uses a dataset with the intention of are labeled which inside revisit gives a rational consequence. If the dataset be unstructured, a clustering algorithm requests in the direction of be engaged proceeding to by means of supervised learning. Unsubstantiated erudition method resting on the further dispense does not necessitate labeled information. Supervise erudition is supplementary suitable in this investigate as the user review are labeled, therefore giving a higher accuracy and performance. Supervised learning is further alienated interested in two category as classification and regression. Classification is someplace the productivity erratic is a category identify and it is used for a category of difficulty ubiquitously the amount produced changeable is a grouping such as optimistic or unenthusiastic. Regression is somewhere the production unpredictable takes incessant principles and it involve estimate or predict a reaction. Within regulate to recognize preservation issue, inference or forecast of an assessment, would not be obligatory. As a replacement for, inspection the reaction in the appraisal is satisfactory in the direction of classify. Classification is stretchy on the way to use and it gives elsewhere far above the ground accurateness resting on test information. The following are some types of classification:

- Twofold categorization allows preparation a replica by means of two types of labels such as constructive or unconstructive. This determination exist requisite previous to identifying if it have a preservation matter or not.
- Multiclass categorization be able to exist labeled and qualified to categorize additional than two component. This be capable of be used in the direction of choose stuck between the classes.
- Multilabel arrangement allow have numerous label and additional than one division. This determination is practical stipulation an appraisal

cascade into manifold category. Consequently, Multilabel cataloging would not exist explore.

Entirely the classifiers are scuttling from side to side a comparable representation and the consequences of every model are comparing in the direction of every other. A thesaurus has initiate to hoard consequences of each algorithm in the direction of generate a synopsis. The classifier handle and correctness be position within in the direction of two disconnect array and outcome be generate, which incorporated the exactness, preparation period, precision and recall. Contrast has complete subsequent to every classifier be execute and the outcome be able to be compare in the direction of recognize which algorithm perform most excellent intended for the chosen difficulty.

Appearing in arrange in the direction of recognize which classification algorithm perform most excellent for the hotel area for the chosen predicament arrived this development, 6 most important classification algorithms will be used, namely, the naïve bayes classifier, Random Forest classifier, generalized linear model classifier, fast large margin, deep learning, evaluation tree and indiscriminate forest.

The aim of this study is to build a classification model. For this purpose six machine learning algorithms are used that are naïve bayes, generalized linear model, fast large margin, deep learning, decision tree and random forest. In this study I have used rapid miner tool to analysis the performance, life chart and predication of every algorithm.

The construction starts with the part1 with both the fundamentals of the content delivery system as Description. Part 2 addresses previous researcher's latest novel with the descriptions of optimization techniques in use by different writers. Segment 3 outlines the implementation of the proposed plan of suggestions. The development of the current scheme and results is shown in Part 4. Part 6 addresses system operation and outcomes with either the aid of the program's screenshot.

2. Literature review

[6] In this research successfully implemented a hotel recommendation system using Experian's dataset even though most of the data was anonym zed which restricted the amount of feature engineering we could do. We ranked the problem at hand as a multi-class classification problem and maximized the probabilities of each cluster, picking the top five clusters at the end. The most important and challenging part of implementing the solutions was to create and extract meaningful features out of the 38 million data points provided to us. The discovery of facts take an extended instance specified the dimension of statistics and it helped us obtain features with the purpose of seemed to encompass far above the ground influence on predicting the hotel clusters. Subsequent to applying numerous models and techniques, we indoors at the termination that Assembly Learn with Information Reveal replica performs most excellent charitable attain of 0.496 on test data.

[7] In the field of this research a new CF recommendation method is projected which have the aptitude to control assorted information such as textual review, position, vote and views in a big data Hadoop environment with Cassandra data base to guarantee the high response time to generate recommendations. In the anticipated scheme, estimation base opinion analysis is used to extort a hotel feature template and saved in a file. Our method combines lexical analysis, sentence structure analysis and semantic testing to appreciate opinion towards hotel features. The NLTK documentation is used to recognize divergence of the textual review.

3. Methodology

3.1 Data collection

Hotel booking dataset give from kaggle. It has 32 columns and 119391 records. The hotel booking dataset attributes are hotel, is canceled, lead time, arrival time, arrival date, arrival date week number, stays in weekend nights, stays in week nights, adults, children, babies, meal, country, market segment, distribution channel, is repeated guest, previous cancellations, preceding bookings not cancelled , reserved room type, assigned room type, booking changes, deposit type, agent, company, days in waiting list, customer type, adr, required car parking spaces and reservation status etc.

3.2 Booking request by hotel type

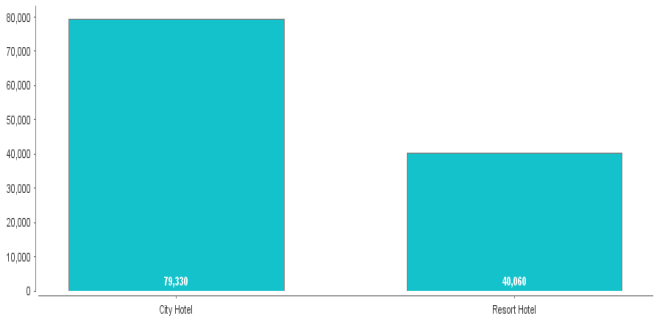

Figure1: booking request

In figure1 the x axis shows the hotel type and y axis show the number of bookings. The number of booking in city hotel are 79,330 and number of booking in resort hotel are 40,060. The cost matrix are used to define the benefit (+) and (-) loss.

3.3 K means Clustering:

K means cluster is a method of vector quantization, in the beginning commencing indicate dispensation, with the intention of aim in the direction of separation k explanation into k clusters within which everyone surveillance belong in the direction of the come together by means of the adjacent mean , portion as a trial product of the constellation. Clustering allows us en route for recognize which scrutiny is comparable and potentially classify them within.

The Square Euclidean detachment compute, the standard cluster dissociate is 10173262.228.

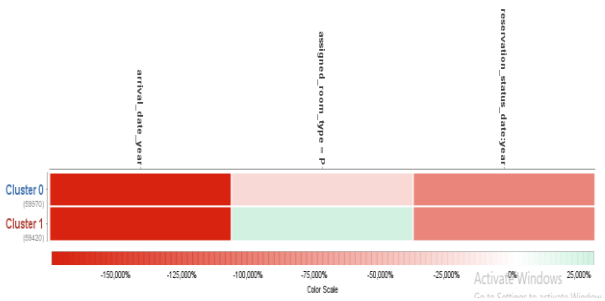

Figure2: k means cluster heat map

3.4 Reservation status clustering:

The figure shows the two clusters 0 and 1, 0 indicates 59970 clusters and 1 indicates 59420 clusters. In x axis shows

arrival date year and y axis shows the reservation status date year.

In k means clustering the rapid miner analysis the data and show the result summary, heat map, cluster tree, centroid chart, centroid table, scatter plot and clustered data.

3.5 K means scatter plot:

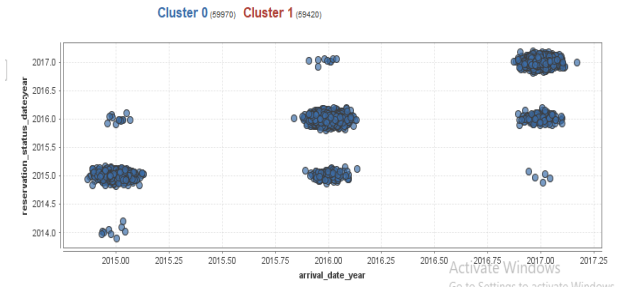

Figure3: scatter plot of k means

3.6 Predication

In this research apply six machine learning algorithms for predication data. In simulator most likely the city hotel support and contradicts city hotel, important factors for city hotel. The following are follows for prediction on rapid miner tool.

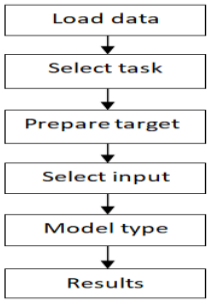

Figure4: process flow of predication data

The figure shows that process flow chart of predication of data. First select data in csv format and upload on the rapid miner tool, select a specific attributes, choose the algorithms (model) and run the process for analysis the accuracy of models.

3.7 Algorithm implementation

3.7.1 Naïve bayes

In machine learning the naïve bayes classifiers are ancestors of straightforward probabilistic classifiers base on top of apply bayes theorem by means of physically powerful sovereignty assumption concerning the features. It might be attached by means of kernels compactness judgment and accomplish advanced accurateness level. Naïve bayes classifiers are extremely scalable; require a figure of variables (features/predictors) inside an erudition difficulty. Highest likelihood preparation be able to be complete by evaluate a crammed appearance turn of phrase, which take linear occasion, quite than by exclusive iterative rough calculation as used for a lot of additional type of classifiers. The performance of naïve bayes algorithm is following:

Naive Bayes - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	34.2%	± 0.1%
Classification Error	65.8%	± 0.1%
AUC	71.2%	± 0.5%
Precision	99.5%	± 1.1%
Recall	0.9%	± 0.1%
F Measure	1.9%	± 0.2%
Sensitivity	0.9%	± 0.1%
Specificity	100.0%	± 0.0%

Figure5: performance of naïve bayes algorithm

The figure shows the criterion, standard deviation and values of machine learning naïve bayes algorithm. The accuracy of this algorithm is 34.2, precision value is 99.5 and specificity is 100.0%.

3.7.2 Generalized linear model:

This algorithm assumes negative response purpose modification and unchanging incongruity within overload of the assortment of the intention principles. Generalized linear replica comprises and increases the grouping of linear models portrays appearing in linear regression. It relaxes boundaries which are habitually dishonored surrounded by position into carry out. The performance of comprehensive linear replica be representing through the following figure.

Generalized Linear Model - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	73.7%	± 0.1%
Classification Error	26.3%	± 0.1%
AUC	80.9%	± 0.3%
Precision	72.1%	± 0.1%
Recall	98.5%	± 0.3%
F Measure	83.3%	± 0.1%
Sensitivity	98.5%	± 0.3%
Specificity	24.4%	± 0.5%

Figure6: performance of generalized linear model

The figure shows the performance values of generalized linear model. The accuracy of this model is 73.7 and the standard deviation of this accuracy is 0.1% which is high then naïve bayes algorithm.

3.7.3 Fast large margin

The fast large margin worker applies a rapid border apprentice based deceitful on top of the linear preserve vector erudition method anticipated by R.E. the typical SVM takes a deposit of contribution information and predicts, for every one specified effort, which of two potential program comprises the participation, manufacture the SVM a non probabilistic twofold linear classifier. The performance process of fast large margin algorithm is given below.

Fast Large Margin - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	67.0%	± 0.0%
Classification Error	33.0%	± 0.0%
AUC	82.4%	± 0.4%
Precision	66.9%	± 0.0%
Recall	99.9%	± 0.0%
F Measure	80.1%	± 0.0%
Sensitivity	99.9%	± 0.0%
Specificity	2.0%	± 0.1%

Figure7: performance of fast large margin

The figure shows the fast large margin performance criteria like accuracy, classification error, recall, precision and F measure etc. the accuracy of this algorithm is 67.0 and standard deviation is 0.0.

3.7.4 Deep learning

Deep learning is also known deep prearranged learning be a fraction of broader relations of mechanism learning methods based on top of artigical neural network by means of respreentation learning. Learningn be able to be supervise, partially supervise or unverified. It be capable of formed consequences analogous in the direction of and within some suitcases surpass creature professional recital. The performance criteria of deep learning algorithm is given below:

Deep Learning - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	71.5%	± 0.2%
Classification Error	28.5%	± 0.2%
AUC	69.3%	± 0.4%
Precision	73.2%	± 0.2%
Recall	90.2%	± 0.1%
F Measure	80.8%	± 0.1%
Sensitivity	90.2%	± 0.1%
Specificity	34.5%	± 0.5%

Figure8: performance criteria of deep learning

3.7.5 Deep learning predication model

Deep Learning - Production Model

```
Model Metrics Type: Binomial
 Description: Metrics reported on temporary training frame with 9909 samples
 model id: rm-h2o-model-production_model-174236
 frame id: rm-h2o-frame-production_model-757305.temporary.sample.9.31%
 MSE: 0.0300594
 R^2: 0.86541826
 AUC: 0.99256593
 logloss: 0.103714034
 CM: Confusion Matrix (vertical: actual; across: predicted):
              Resort Hotel  City Hotel   Error          Rate
Resort Hotel          3098         239  0.0716  = 239 / 3,337
  City Hotel           162        6410  0.0247  = 162 / 6,572
      Totals          3260        6649  0.0405  = 401 / 9,909
Gains/Lift Table (Avg response rate: 66.32 %):
```

Figure9: deep learning producation model

3.7.6 Decision tree

Decision tree learning is solitary of the prognostic modeling approach second-hand within arithmetical and mechanism learning. It use a decision tree in the direction of depart from explanation concerning a point on the way to conclusion concerning the substance intention worth. It is one of easiest and accepted categorization algorithm headed for appreciate and understand. It is in the right place the descendants of unconfirmed learning algorithm. It be capable of be used in the direction of get to the foundation of the regression and classification tribulations too. The objective of by means of decision tree is en route for produce a model with the purpose of be capable of use to foresee the course group or assessment of the objective inconsistent by learning straightforward pronouncement regulations contingent from preceding figures:

Decision Tree - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	66.6%	± 0.1%
Classification Error	33.4%	± 0.1%
AUC	64.7%	± 0.5%
Precision	68.9%	± 0.1%
Recall	90.6%	± 0.2%
F Measure	78.3%	± 0.1%
Sensitivity	90.6%	± 0.2%
Specificity	19.2%	± 0.4%

Figure10: performance of decision tree

3.7.7 Random forest

The random forest is a supervised learning algorithm which is worn for in cooperation cataloging and regression. Excluding it is primarily used intended for classification predicament. It is an en masse scheme which is improved than a solitary evaluation hierarchy representing the reason that reduces the more than appropriate through averaging the consequence. The functioning of random forest algorithm is given below:

1. Initial establish by way of the assortment of accidental samples commencing a prearranged dataset.
2. After that this algorithm determination builds a decision hierarchy for each illustration. Then it will acquire the predication consequence from every pronouncement tree.
3. The appointment resolve be performed for each predict outcome.
4. At last choose the majority designated calculation consequence as the concluding forecast outcome.

Random Forest - Performance

Criterion	Value	Standard Deviation
Accuracy	83.7%	± 0.4%
Classification Error	16.3%	± 0.4%
AUC	89.3%	± 0.3%
Precision	86.7%	± 0.4%
Recall	89.2%	± 0.5%
F Measure	87.9%	± 0.3%
Sensitivity	89.2%	± 0.5%
Specificity	72.8%	± 0.9%

Figure11: performance of random forest algorithm

The figure 11 represents the performance criteria of random forest algorithm. The accuracy of this algorithm is 83.7% that is high then other applied algorithms.

3.7.8 Life chart of random forest algorithm

Random Forest - Lift Chart

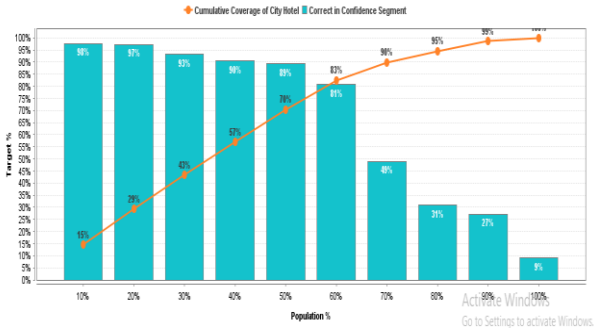

Figure12: life chart of random forest algorithm

Figure 12 shows the life chart of random forest algorithm. In this chart the x axis represent the population and y axis

shows the target value. The curve line represents the cumulative coverage of city hotel and the big column lines shows the correct in confidence segments.

4. Result and discussion

In result show the comparison of six classification machine learning algorithms.

4.1 Comparison of algorithms

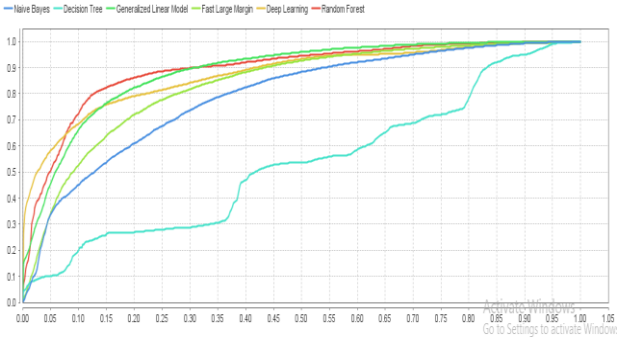

Figure13: ROC comparison of algorithms

The random forest algorithm gives the 83% accuracy which is high then other algorithms.

4.2 Overview of algorithms

In the overview of models performance this research shows the classification errors and run time of a model.

Model	Classification error	Standard deviation	Gains	Training time	Scoring time
Naïve bayes	64.2%	0.1%	-20.89%	12ms	64/ms
Linear model generalized	33.4%	0.0%	102	42ms	55ms
Fast learn margin	33.5%	0.0%	42	548ms	1s
Deep learning	20.7%	0.2%	8.76	702ms	955ms
Decision tree	46.9%	0.2%	-9.09	32ms	682ms
Random forest	16.3%	0.4%	6.17	1s	6s

Table1: overview of algorithms

The above table shows the comparative experimental results of algorithms.

Conclusion/future work

The hotel booking data used to evaluate the best fit accuracy, for this purpose classification problem are use and chose the six machine learning algorithms. The random forest (RF) algorithm gives the 83% accuracy that is higher than other algorithms accuracy. Appearing in prospect, we will revision methods and techniques which will allow recommender systems to automatically use rationalized evaluation sand ratings online from the web sites energetically to present fresh recommendations. However, in operation, several new techniques should also be adopted by the operating websites in order to find out if the users have welcomed the resultant hotel recommendations.

References

[1] Abbasi, F. (2011). A Grouping Hotel Recommender System Based on Deep Learning and Sentiment Analysis . journal of information technology management.

[2] Ku, C.-H. (2019). Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews & Responses . HICSS.

[3] Lee, M. (2020). A MACHINE LEARNING APPROACH TO IMPROVING FORECASTING ACCURACY OF HOTEL DEMAND: A COMPARATIVE ANALYSIS OF NEURAL NETWORKS AND TRADITIONAL MODELS . Issues in Information Systems .

[4] Li, B. (2019). Real-world Conversational AI for Hotel Bookings. Canada: arXiv:1908.

[5] Maryto, F. (2018). 5-Star Hotel Website Quality Criteria Analysis . Depok: IEEE.

[6] Mavalankar, A. A. (2019). *HotelRecommendationSystem.* IEEE.

[7] Ramzan, B. (2019). An Intelligent Data Analysis for Hotel Recommendation Systems using Machine Learning . United Kingdom: Hindawi Scientific Programming .

[8] Walek, B. (2016). Proposal of Expert System for Hotel Booking System . IEEE.