Analysing the Customer relationship management in Hotel chain

Analytical CRM

Anurag Abhay Singh
Msc. Data Analytics
x18104053
National College of Ireland

Abstract-**This paper proposes a valued business proposal by predicting the booking status from leads and sales records data. The proposed system has used statistical methods to understand relationship between booking demands and various factors to derive a relationship which will help hotel industry knows customer demands and predict booking state, which will automatically leads to efficient management of hotel services. For analysis Rapid miner, R, SPSS mathematical tools are used to calculate relative error rate, accuracy in predicting probability the customer booking status. Our paper is designed by mainly focusing hotel chain concern in handling unwanted cancelled of rooms by deriving certain output they can design their room management system**

Keywords—Rapid Miner, Logistic regression, Decision trees, data mining, machine learning.

I. Introduction

Customer relationship management (CRM) is tool to understand customer importance, concern by deriving relation in various terms. The CRM is used by company to understand to better communicate with customer to expand its business by holding on existing customer or increasing its customer base. The hotel industry is volatile industry because of uncertain behaviour of customer and fierce competition. The booking demands act as leads for hotel for arranging its service to match customer demands, but cancellation of booking after certain time or during arrival time leads to inefficient management plus void space of room can occur if there is no waiting list which affects hotels revenues. To avoid these pitfalls prediction is required by analysing the customer historical data by finding the factors which is creating such occurrences. The data mining will help in extracting features and showing probabilistic relation between features at certain events. Recent trend of online booking is latest to add misery to hotel, because of variety of option available on site on one click multiple booking and cancellation can happen at same time. Total cancellation worldwide in hotel industry is around 104 % [5].

II. Literature Review

Expert system is designed to recommend hotel services to guest by using fuzzy method by reviewing different online sites [1]. Existing booking system don’t recognize the needs of customer for services but the proposed system was designed considering question from Points of interest, activities and events on basis of which services can be categorized for visitors so that maximum booking can be achieved. The designed data warehouse gives the hotel management a visualised form dashboard content with historical data so that it can analysed it and serve better to customer.

Segmentation of customer can be done to decide optimal price for different customer on basis of prepayment done by customer [3]. The booking cost strategy is designed in such way that it is directly proportional to operational cost to avoid risk of booking cancellation, as one paid the higher cost while pre booking is done is less likely to cancel the booking and vice versa.

Machine learning is used to classify and identify the potential customer from pool of booking people to avoid last minute cancellation and overbooking which in turn helps in efficient management of booking system in [3]. The A/B testing mechanism was used to build predictive model for booking which might get cancel in future, and the model was evaluated in real time to suggest preventive measures to be taken when cancellation happens. The person contacted by hotel are less likely to cancel tickets than person who are not contacted at all.

The data warehouse is designed to generate business intelligence report in visualised form to understand relationship which is affecting the hotel booking system [4]. Reliable output achieved showing that using agent's option was best when it comes to booking.

The paper reviewed in this section clearly gives indicative direction about use of automation by researcher to make analysis process faster when we are dealing with virtual world, where there is lot of complexity in understanding the data and implementing in proper process to achieve predictive results.

III Dataset and hypothesis

The Dataset used for analysis is of hotel situated in Algarve region of Portugal contain information related to booking demand downloaded from science direct site and consist of 40,060 values and 7 selected features. The dataset booking demands and months are kept independent value and rest attributes are kept as dependent value to understand relationship. The hypothesis are developed as follows:

Ho- The Lead times, number of weekdays and weekend night is having significance on booking demand.

H1-The lead times, number of weekdays and weekend night don't have any significance on booking demand.

IV. Methodologies

The only way analysis is to understand the significance of one attributes on independent variable to which extent it has positive or negative correlation with one another. The transparent methodologies is used to show the results in visualised form including all attributes.

A) Linear Regression: SPSS tool is used to obtain linear relationship between one dependent and one or more independent variable. The output gained can be in several form like Durbin Watson factor, checking collinearity, error rate to check accuracy, co-efficient factor between variables, standard deviation from mean value, by keeping confidence interval at 95%. All these results will show the impact created on independent variable by dependent variable. The linear regression is used for prediction purpose by checking above all factors. The different analysis of output yields a better understanding and bringing transparency in process.

B) Logistic Regression: It's similar to linear regression but only difference is the relationship between one dependent variable and ratio level independent variable. Logistic regression are used for nonlinear variable to obtain predictive analysis. Rapid miner tool is used to obtain z-value, degree of co-efficient, error rate and p-value for analysis.

C) Naïve Bayes: It's a classification technique used by rapid miner tool to show probability of new customer falling in which category of booking status. Its shows majority of yes or no on basis of which customer booking status can be predicted. Naïve Bayes will mostly show the significance of independent variables on dependent variable and by showing its correlation in terms of probabilistic range.

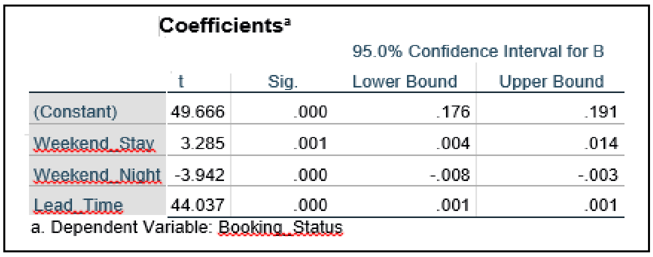
D) Decision trees: A decision trees has independent variable classified as multiple branches. Indirectly its function is to show division of variables in the range showing yes or no. If it's yes the value must be high and no then value must be less. Split is done in such way that includes maximum category of particular value to reduce informational entropy. It will classify as new booking will fall in which region of booking status on basis of range of independent variables. Rapid miner tool is used to import the observation and predict the report after analysis.

V. IMPLEMENTATION

This section mostly deals with statistical operation performed on datasets by various tools and results obtained which is used for business purposes and understand customer relationships. All steps and how model is evaluated with different parameters to have clear understanding of process and how to reuse same process.

Data Pre-processing- The selected dataset has 31 features which is reduced to 7 features like booking status, weekend stay, weekdays stay, Average daily rate (ADR), booking changes, lead time, arrival month and market segments. The booking status is encoded in SPSS software as ordinal dependent variable, market segments and arrival months are encoded as nominal variable. Rest all features are considered as independent variable as numeric value.

A)Linear Regression:-The data was loaded in SPSS and analysed directly by liner regression command showed output in various tables where dependent variable is booking status and Independent variables are weekends night booked, weekdays stay and lead time.

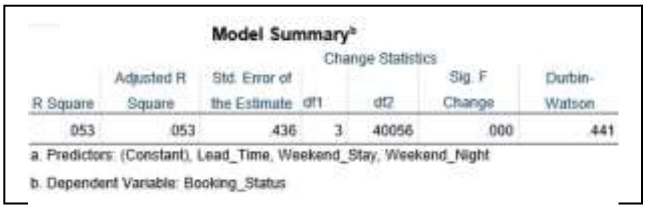


Coefficients[a]

	t	Sig.	95.0% Confidence Interval for B Lower Bound	Upper Bound
(Constant)	49.666	.000	.176	.191
Weekend_Stay	3.285	.001	.004	.014
Weekend_Night	-3.942	.000	-.008	-.003
Lead_Time	44.037	.000	.001	.001

a. Dependent Variable: Booking_Status

As the co-efficient output derived clearly show significance level of different selected variables on dependent variable. Only weekend stay is showing some value rest are 0. Only weekend stay is clearly impacting the booking status. Upper bound and lower bound of weekend stay is 0.004 to 0.014 respectively at 95% confidence interval that means 0.4% to 1.4% people in weekend stay are affecting the booking status.



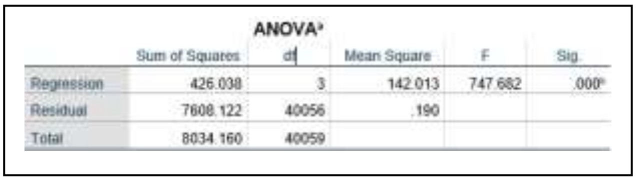
Model Summary[b]

R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics df1	df2	Sig. F Change	Durbin-Watson
.053	.053	.436	3	40056	.000	.441

a. Predictors: (Constant), Lead_Time, Weekend_Stay, Weekend_Night

b. Dependent Variable: Booking_Status

The second obtained calculation shows that 3 degree of freedom for dependent variable. R square and adjusted square R rate is 0.053 this means variables completely fits in model and is not over fitted. Only 5.3% case can be explained. However such low results indicate that model is less affective. Durbin Watson output is 0.441 which indicates there is positive autocorrelation between dependent and independent variables.

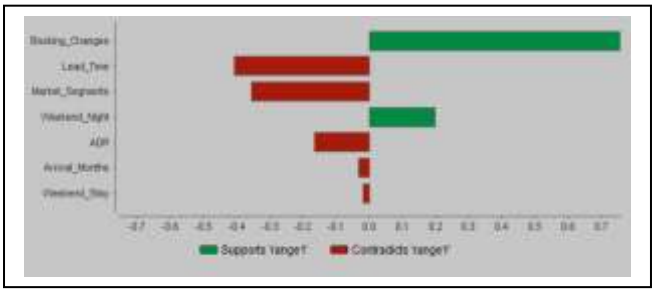


ANOVA[a]

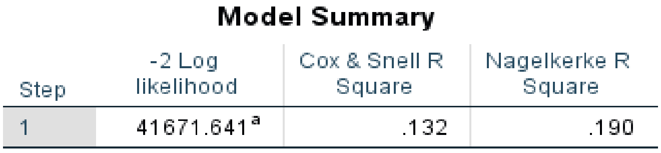
	Sum of Squares	df	Mean Square	F	Sig.
Regression	426.038	3	142.013	747.682	.000[b]
Residual	7608.122	40056	.190		
Total	8034.160	40059			

The result shows p-value 0.00 which is less than 0.05 value therefore null hypothesis which state there is significance relation between weekend nights, weekday stays and dependent variable booking status is rejected or it can be other way stated by previous one that significance level might be so small to be considered.

B) Logistic Regression:-The dataset converted by SPSS is directly loaded into Rapid miner tool. And doing step by step operation keeping booking status as dependent variable operation is set into auto model performance.



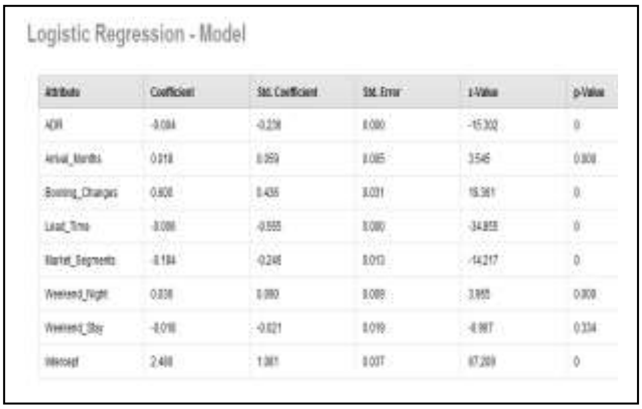
After SPSS strategy was changed to include more variables as independent variables by assigning booking status as range 1 the result is shown in form of graph above. The green bar indicates the variable support to extent to dependent variable and the red bar contradicts the support. From pattern it's clearly seen as booking changes and weekend night has huge significance on booking status. Lead time what was thought to affect booking status has the highest contradiction. The results up to 70% booking status is influence the booking changes. Major strategy can be built around it how many times person changes booking decisions. Therefore lead time can be removed from consideration. Keeping same consideration for evaluation SPSS operation was performed. Cox & Snell R square is 0.132 which states that its perfectly fit model. Obtained Nagelkerke R square also points in same direction.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	41671.641[a]	.132	.190

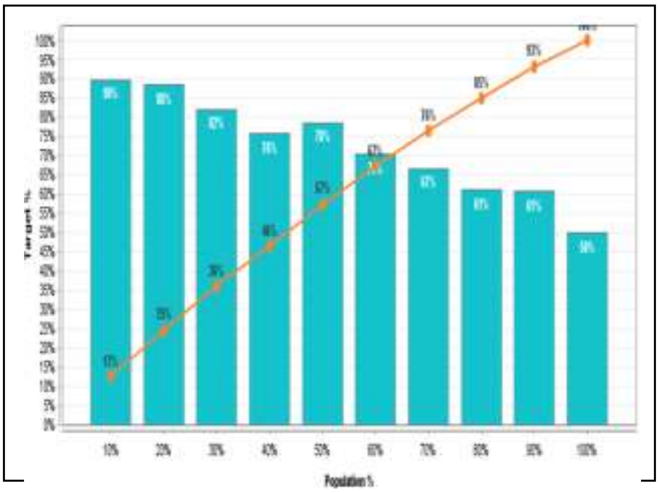
The rapid miner has also given the standard error rate and calculation used to generate graph



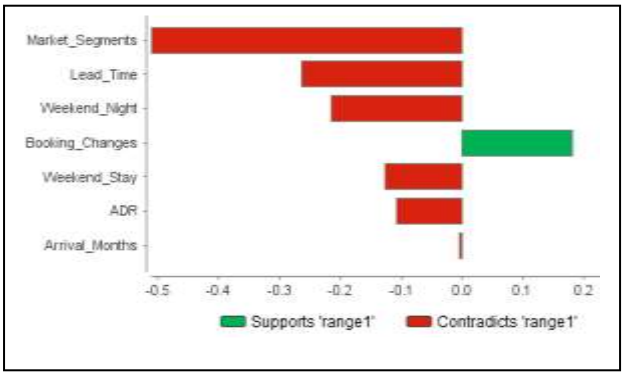
Logistic Regression - Model

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
ADR	-0.004	-0.238	0.000	-15.302	0
Arrival_Months	0.018	0.059	0.005	3.546	0.000
Booking_Changes	0.608	0.436	0.031	19.361	0
Lead_Time	-0.008	-0.555	0.000	-34.858	0
Market_Segments	-0.184	-0.248	0.013	-14.217	0
Weekend_Night	0.036	0.080	0.009	3.953	0.000
Weekend_Stay	-0.018	-0.021	0.019	-0.987	0.334
Intercept	2.488	1.381	0.037	67.209	0

C) Naïve Bayes:- The rapid miner tool were set with same parameter of defining booking status as dependent variable as previous one. The probability of person booking status will be yes or no is given by chart population vs target, to classify in which category booking will fall.

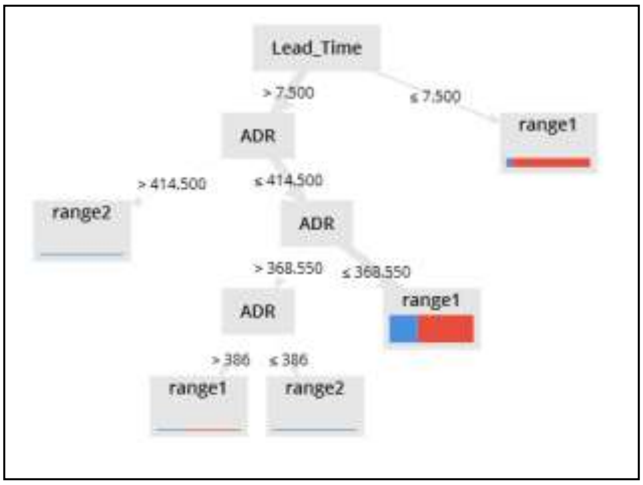


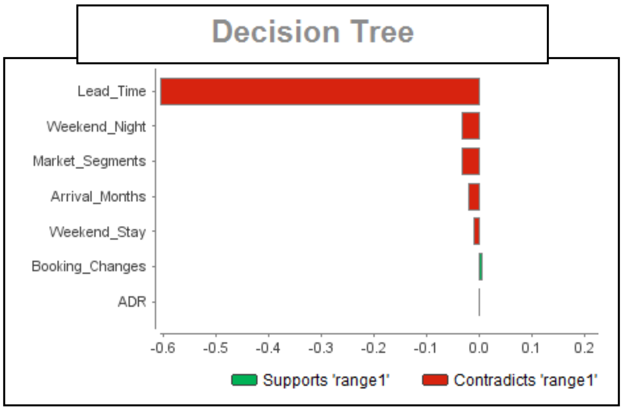
The lift chart has more area under the curve i.e. around 67% which shows its good fit for model. Around 67% of total population can be targeted and understand in model about booking status in future. It can be confirmation for booking or decline. The results is interpreted till the line touches the bar in graph and ending point is last prediction status.



Naïve buyer shows booking changes variable significantly affects the booking status of person. The amount of time booking changes happen influences decision of bookings. Close to 20 % booking changes have significance on 67% of booking status. Arrival Months have also somewhat influence on booking status as it is close to zero shown in graph. We need to consider it for future use to make our model more helpful for analysis purpose by seeing how many time booking changes are with respect to arrival month.

D):- Decision Trees:- The classification technique used to classify how booking decision impact hotel revenue and how you can classify the booking will be yes or no. Lead time is classified as variable by Decision Trees which affects the dependent variable which states that average 7 and ½ day can be used to analyse the booking status of person which is classified yes or no. The probable is yes that same person is going to spend more than 414.5 dollars weekly defined ADR. That ADR is again classified into range which shows the person maximum people out of ADR 414.50 on the average are going to spend 368.550. So by sure if booking happens within 7.5 days the average earning per week for hotel will be 368.550. This how model is classified by maximum no of trees having probability of yes lead time affects booking status therefore lead time is tree point and ADR is shown as branches of that tree. As seen lead time is having huge contribution in affecting decision of booking time there it's classified as main variable which is narrow done to ADR which shows signs of support in influencing booking status. This how classification trees works presenting analysis on classifying dependent variables on basis 2 variable in 2 dimension.





Conclusion

The study was conducted to analyse the hotel industry relation with customer by using different statistical tools. Aim was to check booking status pattern with different variables to have clear understanding and improve the booking status to full occupancy by improving it. The results obtained clearly shows the lead time, weekday stay, weekend night and booking changes by customer have impact on booking status. Only feature which was assumed but has no impact is type of customer from different business segments even after being calculated by different tools. The results show booking changes of 20% feature control 67% of booking status population. And out of these population mostly preferences is weekend_night time. The lead time is also important parameter to understand the booking scenario as one with lead time less than 7.5 days have higher chances of booking room and spending around 368.50$ weekly. This analysis can help business organization to reconsider and design some policy to suit parameters as stated above which will indirectly increase efficiency of hotels in terms of revenue and occupancy.

References

[1] B. Walek, O. Hosek, and R. Farana, "Proposal of expert system for hotel booking system," *Proc. 2016 17th Int. Carpathian Control Conf. ICCC 2016*, pp. 804–807, 2016.

[2] Z. W. Miao, T. Wei, and Y. Q. Lan, "Hotel's online booking segementation for heterogenous customers," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2016–December, pp. 1846–1850, 2016.

[3] N. Antonio, A. De Almeida, and L. Nunes, "Predicting hotel bookings cancellation with a machine learning classification model," *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, vol. 2018–January, pp. 1049–1054, 2018.

[4] A. S. Girsang *et al.*, "Decision support system using data warehouse for hotel reservation system," *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018–January, pp. 369–373, 2018.

[5] Delgado, P.(2019). Cancellations on Booking.com: 104% more than on the hotel website. Expedia, 31% more.