Project Report

MSc in Data Analytics
Domain Application of Predictive Analytics

Ashish Rawat
X18185801

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

Predictive Analysis of Hotel Booking Cancellations

Ashish Rawat
Master’s in Data Analytics
X18185801

***Abstract* - In hospitality industry, revenue is negatively impacted by booking cancellations. Thus, hotels adopt measures like penalty on cancellation and overbooking which in some scenarios leads to unfavorable outcome. Acknowledging the need for a forecasting model which can accurately predict booking cancellation a logistic regression forecasting model is presented. The research utilizes the H2 dataset and evaluate the model on parameters like confusion matrix, accuracy, precision, recall and F1 score. The output of the classification models would guide hotel authorities in making better decision to increase revenue and improve their service.**

I. INTRODUCTION

The introduction of the booking system in the hospitality industry provided the customer with a facility to reserve hotel services for their use before the date of arrival [1]. As a result, it allowed hotel authorities to prepare their services to provision well in advance. Although it appeared mutually advantageous, it gave the customer the right to cancel the reservation before the arrival date. To cope up with this issue, hotels started the implementation of strict cancellation policies and overbooking methods [2]. But these approaches sometimes have adverse effects and results in destroying the reputation of the hotel and negatively influence future revenue.

At present, due to the availability of many booking channels people are subjected to multiple choices for a hotel of their preference. This and the low cancellation cost have encouraged customers to search for a better deal and in the process, they tend to book multiple hotels that are later canceled. Thus, resulting in an increased cancellation rate which fluctuates between 20 % to 60 % of the overall bookings [3].

Therefore, a need for a system which can effectively forecast future cancellations has risen. By understanding the factors affecting cancellation and getting better insights from customer booking patterns strategies can be developed that targets customers that are more likely to cancel. In this study I propose a logistic regression model to forecast cancellations and has utilized a real-world historic data of a Portuguese hotel.

II. LITERATURE REVIEW

Despite the relative importance of forecasting booking cancellations in a hospitality industry the issue has not been explored to that extent and only a few literatures on the topic are present. A brief review on previous works is given below:

[4] utilized the real historic data from four different hotels and adopted CRISP-DM, a frequently used methodology in the field of predictive analytics in their work. They focused on cancellations as the indicator and created a model which treated the issue as a classification problem. In the process, they critically explored their data and reviewed statistical concepts to avoid issues which are caused by curse of dimensionality, correlation and leakage of data. Insignificant attributes were discarded, and new relevant attributes were created by combining few attributes. For modeling the two-class classification problem following machine learning models were employed: Boosted Decision Tree (BDT), Decision Forest (DF), Decision Jungle (DJ), Deep Support Vector machine (DSVM) and Neural Network. They used K-fold cross validation for evaluating the model performance. The experimentation results indicated that in terms of accuracy, AUC and F1 score BDT and DF performed relatively better than the other models. But in terms of generating lowest number of false positive for “IsCancelled?”, DF clearly was the selected model.

[5] suggested the use of a complex machine learning algorithm XGBoost. And focused equally on the initial stage of the model development to avoid issues like data shift and data leakage. They employed historic data from 2 hotel and used convenience splitting method to shape train and test datasets. To achieve significant results features were carefully selected and reengineered. They evaluated the performance of the

model over traditional metric like AUC, accuracy, precision and F1 score) and introduced a new metric, Minimum Frequency (MF) to effectively evaluate the dynamic hotel booking data. Finally, the results highlighted the model's robustness towards noise (unknown data) and aversion from overfitting.

[6] illustrated dependence of customer attributes on booking horizon (time to service) to forecast cancellations at multiple stages of booking (seven-time stages). They suggested the use of multiple models insisted of at different stages of booking horizon. They utilized data of a major UK hotel consisting of more than 200,000 records. They used random forest assessment function to assess the importance of a variable at each time stages. Forecasting methods includes Average Cancellation rate (AVG), Seasonally Averaged Cancellation rate (SAVG), Logistic Regression (LR), Kernel Logistic Regression (KLR), C4.4 Probability Estimation Tree (C4.4), Minimum Squared Expected Error Tree (MSEE), Support Vector Machine (SVM) and Random Forest (RF). The evaluated the proposed framework by absolute error which they suggested a better metric than absolute difference and actual cancellation rate. Their results concluded high performance of multiple models-based framework against individual models which acknowledged the cancellation forecasting as a complex issue and dependent on time to service of booking.

The study presented in [7] contributes towards gaining a deeper insight in understanding the effects customer and their booking behaviors have on cancellation. To determine cancellation probability, they implemented a probit model which performs regression between binary values and used data from 9 hotels in Finnish Lapland. The empirical results concluded that a large gap between booking to arrival date exhibits a higher risk of cancellation and features like season, travel group and country of residence has also portrayed a significant relation with cancellation. Importance of channel attribute has also been highlighted and depending on the selected mode of channel cancellation probability has proven to vary significantly.

III.. METHODOLOGY

In predictive analytics it is important to adopt an appropriate approach to build a reliable forecasting model that can offer valuable feedback and help address business problems. Therefore, in this research CRISP -DM (cross industry standard process for data mining) based methodology has been incorporated with reference to [1] and [8].

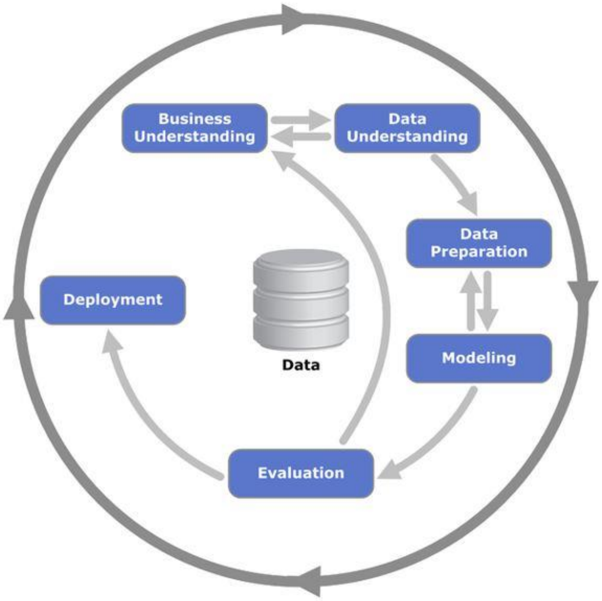


Figure 1: CRSIP-DM Process Flow

A. Data Description

The hotel booking system data of a hotel in Portugal has been utilized for this research (H2 dataset) [9]. The dataset contains records for a period of 2 years starting from July 2015 till the end of August 2017. It consists of total 31 variables (categorical and numerical) including the binary output variable; booking confirmed or cancelled. Figure 2 representing the cancellation ratio in the H2 dataset indicates a high cancellation rate of around 42% with respect to the total bookings made.

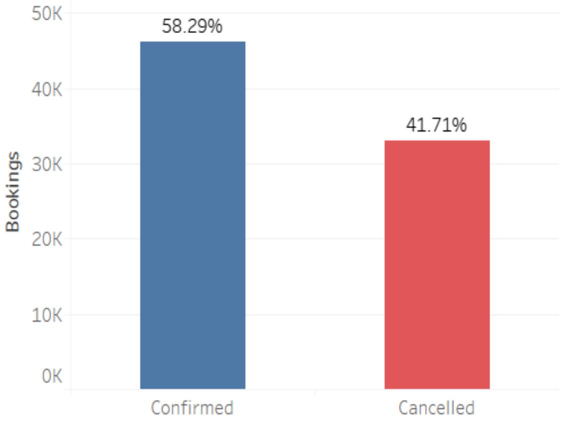


Figure 2: Cancellation Rate of H2 dataset.

B. Data Processing

The raw H2 data consists of irregularities which required pre-processing and variable transformation before using it in a predictive model. Pre-processing steps involved removal of missing values, omitting irrelevant levels (unknown) of a variable and fusing minority levels to form a comparatively relevant level in the dataset. It was followed with conversion of data types of variable, since data when imported consisted of variables of incorrect data types. After the primary processing steps, new columns ('ArrivalDate') were formed by combining columns like 'ArrivalDateMonth', 'ArrivalDateYear' and 'ArrivalDateDay', to eliminate multicollinearity and accordingly specific features were selected using Filter selection methods to reduce the curse of dimensionality present due to high number of variables in the data. Since the data contains categorical variable, dummy variables were generated to represent these variables using Pandas "get_dummies" function. And finally, for model training and evaluation the data is divided using random sampling in the ratio of 70:30, where 70% of the data is used to train the model and the rest 30% is utilized to test the performance of the model.

C. Machine Learning Model

The output variable in the H2 dataset is a binary variable which indicates whether a booking is cancelled or not. Thus, treating the problem as binary classification problem I've implemented a model using Logistic Regression algorithm.

Logistic regression is a supervised classification algorithm which is used when the output variable contains binary value (Yes/No or 0/1). It is represented with sigmoid function as shown in figure 3 and expressed through equation 1 [10].

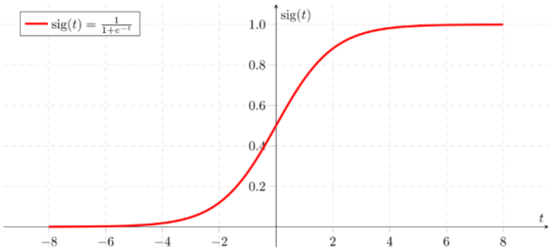


Figure 3: Logistic regression (Sigmoid Function)

$$\boldsymbol{P = \frac{e^{(\beta_0+\beta_1 x)}}{1+ e^{(\beta_0+\beta_1 x)}}} \qquad (1)$$

Where P is the dependent output which indicates the estimated probability of an event (0 or 1) to occur and x is independent variable. β_0 is the constant and β_1 represents the slope.

IV. EVALUATION

To determine the performance of the logistic regression model following evaluation metrics were used [11]:

1. Confusion Matrix

It is a table which summarizes the actual and predicted observation with respect to the output. It is utilized to determine the error the model is making and specifies the type of error. A confusion matrix is represented as shown in figure 4.

	Actual Value	
Predicted Value	TP	FP
	FN	TN

Figure 4: Confusion Matrix

Where,

True Positive (TP): Actual and predicted values are true.

False Positive (FP): Actual value is true while predicted value is false.

False Negative (FN): Actual value is false while predicted value is true.

True Negative (TN): Actual and predicted values are false.

2. Precision

Precision is calculated through equation 2 and indicates the proximity of positive predicted value with positive actual value.

$$\boldsymbol{P = \frac{TP}{TP+FP}} \qquad (2)$$

3. Recall

Recall is expressed through equation 3 and represents the ration of correctly determined actual positives.

$$\boldsymbol{R} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+NP}} \qquad (3)$$

4. F1 score

F1 score determines the balance between precision and recall required. It is a function of both the metrics and expressed through equation 4.

$$\boldsymbol{F1\ Score} = \boldsymbol{2}\frac{\boldsymbol{P*R}}{\boldsymbol{P+R}} \qquad (4)$$

5. Accuracy

It determines the overall performance of the model that is the number of correctly classified instances. It is represented through equation 5.

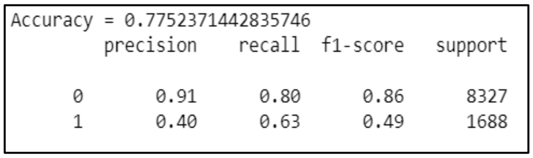
$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \qquad (5)$$

V. RESULTS

The results represented in Table 1 and Figure 5 were obtained after running the logistic regression model.

	Actual Value	
Predicted Value	37617	9232
	4309	28144

Table 1: Confusion Matrix obtained.



```
Accuracy = 0.7752371442835746
          precision    recall  f1-score   support

     0         0.91      0.80      0.86      8327
     1         0.40      0.63      0.49      1688
```

Figure 5: Evaluation metrics obtained.

A. Quantitative Results

The logistic regression model implemented in this research achieved an accuracy of 77.5 % which is relatively good considering the dynamic nature of the hotel booking dataset. Confirmed Bookings represented with ‘0’ in figure 5 achieved high values for precision, recall and F1 score which indicates high performance of model in classifying confirmed cases. On the contrary, Cancelled Bookings represented with ‘1’ obtained low values which indicates the model inefficiency in accurately determining cancellation records.

B. Qualitative Results

Assessing the observation of confusion matrix, it can be said that the model developed is quite efficient in classifying confirmed booking records. Since, the research focusses on forecasting possible cancellations, relatively low value of FP and FN as obtained was expected. It signifies the error, model made in forecasting cancellation records. From business perspective, strategies revolving around these errors should be developed to increase profits and provide efficient service to customers.

VI. CONCLUSION

This research proposed the implementation of logistic regression model to forecast hotel booking cancellations. By critically exploring the hotel data significant features that affect the cancellations were identified. Insights on the effectiveness of the service currently provided by the hotel authorities were also assessed and implied adoption of trends in demand. And finally, the model achieved fairly good results in determining the cancellation records with low error rates.

VII. REFERENCES

[1] K. Talluri and G. Ryzin, The theory and practice of revenue management. Boston, MA: Springer-Verlag US, 2005, pp. 1-10. [Accessed on: April 1, 2020].

[2] K. Talluri, I. Karaesmen, G. van Ryzin and G. Vulcano, "Revenue management: Models and methods", Proceedings of the 2009 Winter Simulation Conference (WSC), 2009. Available: 10.1109/wsc.2009.5429322 [Accessed on: April 1, 2020].

[3] D. Romero Morales and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining", European Journal of Operational Research, vol. 202, no. 2, pp. 554-562, 2010. Available: 10.1016/j.ejor.2009.06.006 [Accessed on : April 1, 2020].

[4] N. Antonio, A. Almeida and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and

increase revenue", Tourism & Management Studies, vol. 13, no. 2, pp. 25-39, 2017. Available: 10.18089/tms.2017.13203 [Accessed on: April 4, 2020].

[5] N. Antonio, A. de Almeida and L. Nunes, "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations", Data Science Journal, 18(1), pp.32, 2019. Available: http://doi.org/10.5334/dsj-2019-032 [Accessed on: April 4, 2020]

[6] D. Romero Morales and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining", European Journal of Operational Research, vol. 202, no. 2, pp. 554-562, 2010. Available: 10.1016/j.ejor.2009.06.006 [Accessed on: April 5, 2020].

[7] M. Falk and M. Vieru, "Modelling the cancellation behavior of hotel guests", International Journal of Contemporary Hospitality Management, vol. 30, no. 10, pp. 3100-3116, 2018. Available: 10.1108/ijchm-08-2017-0509 [Accessed on: April 6, 2020].

[8] A. Shi-Nash and D. Hardoon, "Data Analytics and Predictive Analytics in the era of Big Data", Internet of Things and Data Analytics Handbook, pp. 329-345, 2016. Available: 10.1002/9781119173601.ch19 [Accessed on: April 6, 2020].

[9] N. Antonio, A. de Almeida and L. Nunes, "Hotel booking demand datasets", Data in Brief, vol. 22, pp. 41-49, 2019. Available: 10.1016/j.dib.2018.11.126 [Accessed 24 February 2020].

[10] S. Swaminathan, "Logistic Regression — Detailed Overview", Medium, 2018. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. [Accessed on: April 7, 2020].

[11] I. Lindgren, "Evaluation Metrics for Classification Problems in Machine Learning", Medium, 2020. Available: https://towardsdatascience.com/evaluation-metrics-for-classification-problems-in-machine-learning-d9f9c7313190 [Accessed on: April 7, 2020].