Siddharth Kushwaha

[sid234k@gmail.com](mailto:sid234k@gmail.com)

+91 9945301305

**ONLINE HOTEL BOOKING ANALYSIS**

* 1. INTRODUCTION



An online hotel booking agency dataset where users visit the page, search their requirements then they are redirected to a page with several hotels satisfying their request with the total price below with their competitor total prices. Now user decides whether to click on the book link. And then one can decide whether to proceed and book it or not.

1.2 DATA CLEANING

There are total 2380057 rows and 54 features in which there are 1 datetime, 30 categorical variables and 23 continuous variables. user\_hist\_stars, user\_hist\_paid, log\_click\_proportion, booking\_value and mostly competitor data have null values more than 50% of data which can be dropped during model building time but can be useful during EDA. Price\_usd and distance\_to\_dest have majorly right skewed data which needs to be treated before modelling.

Upon checking correlation coefficient of all features, clicked and booked were found highly collinear i.e. if a person is clicking only then he can book the hotel. Hence because of this dependency clicked and booked have 0.78 correlation coefficient. Collinearity affects the estimated coefficients and p-values not the predicted values. The coefficients swing wildly based on which other independent variables are in the model.

Booked and clicked are possible few options to predict through the model. But due to some reasons it will lead to false accuracy. False reduction in accuracy means it has a high biasing error. It suggests that the model is only learning one classification i.e. it is more concerned towards one label only suggesting huge imbalance in dataset. Upon finding the possible target variables it was highly imbalanced. Clicked and booked over 90% unbalance. So now the model mostly will learn about 0 having a very high False Positive Value reducing precision. Hence, reducing overall F1 score.

There were some features which can be combined together to reduce total number of dimensions for better results for EDA and modelling. Created a column for total travellers by adding total number of adults and kids and a column is\_converted where if clicked and booked then he is converted otherwise not, if not clicked then value is null.

A feature named price\_usd can be price per night or price for the whole stay. Only confirmed value is booking\_value which is only available when a user has booked the hotel. So to convert this price\_usd to the mean price a new column price\_usd\_corr was created suggesting average mean prices of the hotel per room per night. There were several steps followed to create price\_usd\_corr i.e.

* *Find maximum permissible value of each room for length\_of\_stay 1 for each category of star rating by segregating and that maximum permissible is 0.99 quantile value.*
* *Now for each star rating hotel where (room or length\_of\_stay) > 1 if the price\_usd greater than the maximum permissible value then to average out the mean price use formula (price\_usd)/(num\_rooms\*length\_of\_stay).*
* *Created total\_price equals to the total price for the trip for the user and check with booking\_value if the RMSE is low or not, was found to be less than 1 suggesting closeness to the data.*

**site\_id, user\_country\_id, user\_hist\_stars** and **user\_hist\_paid** consists of all the user data information and **listing\_country\_id , listing\_id, listing\_review\_score, listing\_stars** and **listing\_position** consists of all the hotel information and some hotel related data with competitor data was secured in a way that the information loss was minimum. Hence, there were no data privacy or security issues in the dataset.

1.3 EDA (Exploratory Data Analysis)

In this dataset each row does not represent each unique user\_id. Each search\_id’s whole response is extracted and got the timestamp of when it was extracted or when the activity was over. So grouping them together to find the conversion rate and click-through rate.

Click through rate can be calculated as ratio of number of unique users visited by total number of impressions or visits on the site/webpage.

Conversion rate can be calculated as ratio of number of unique users who were converted by total number of impressions or visits on the site/webpage.

**Click-through rate = 4.03%**

**Conversion rate = 2.77%**

Created two important features from timestamp which could be useful like weekday and month which can be easily extracted from timestamp.

The observed frequencies for listing\_review\_score variable match the expected frequencies for the listing\_star proved by chi-square test between them. Hence, only high star rating around 3-5 with review score between 3.5-4.5 generates maximum conversions and click responses.

Features which have more than 50% null values in them were dropped. Now features with more than 5% but less than 50% were predited through Random Forest Regressor to fill null values in them, they are **distance\_to\_dest** and **location\_score2.** These null values are needed to be treated to fit in the model.

Upon searching for outliers, **distance\_to\_dest** and **total\_price** were having highly right skewed data. To treat outliers power transformation was used with box-cox transformations. If not treated then it will lead to a bias error or increase in FP (False Positives) and FN (False Negatives).

A screenshot of a cell phone

Description automatically generated

![A screenshot of a cell phone

Description automatically generated]()

Some important findings in Tableau-

* Users are spending more money in major chain hotels.
* Users doesn’t bother much for promotion on hotels except for the competitor1 hotel rates.
* Lowest sales on new year 2013, probably because mostly people encourage home meetings and every Saturday the sale is observed to be lowest.

1.4 MODEL SELECTION

We can predict either clicked or booked from the dataset. CTR (Click-Through Rate) is 4.03% which is more than the average CTR of all webpages/site all over world. Comparatively, CVR (Conversion rate) of dataset is much less than CTR and is around average CVR which we can work upon to improve. Now we are predicting booked, whether the user is going to convert by booking or not.

Booked has very high imbalance so the prediction is going to be false accurate i.e. False Positives (FP) while predicting booked will be high. Our main objective is to increase the f1-score as much as possible as our evaluation metric as we have to reduce the FP and FN here as having both of them will reduce the performance of model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Regression  (without SMOTE)  (cutoff = 0.03) | Logistic Regression  (with SMOTE)  (cut-off = 0.66) | Logistic Regression  (with balanced class weight)  (cut-off = 0.67) |
| F1 Score | 0.09 | 0.18 | 0.18 |
| Accuracy | 0.61 | 0.84 | 0.84 |
| Precision | 0.05 | 0.10 | 0.10 |

Clearly without using any imbalance technique the FP were so high that it was dragging the precision score down. After applying SMOTE precision score increased slightly. Best performance for Logistic comes under balanced class weight with an **auc\_roc score** of 81.91. But still the false positives were way too high. Hence going for Random Forest which will use Bagging and create several trees to check whether the data point predicts booked or not increasing it by oob\_score which is equivalent to three cross fold validation.

|  |  |
| --- | --- |
|  | Random Forest Classifier  (with SMOTE) [cut-off = 0.44] |
| F1 Score | 0.10 |
| Accuracy | 0.69 |
| Precision Score | 0.05 |

Even with Random Forest, SMOTE and undersampling f1-score is not even increasing above 20%. Hence, reducing the training data now.

1.5 CONCLUSION

To predict booked, we can modify the data by only selecting data from a condition where clicked is 1 because if a person clicked on the site only then can he book the ticket. Hence, separating clicked=1 data and using it as a training data to predict on remaining dataset for booked.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Regression  (without SMOTE)  (cutoff = 0.65) | Logistic Regression  (with balanced class weight)  (cut-off = 0.22) | Random Forest Classifier  (with max\_depth=8)  (cut-off = 0.6) |
| F1 Score | 0.67 | 0.90 | 0.89 |
| Accuracy | 0.6 | 0.86 | 0.85 |
| Precision | 0.68 | 0.85 | 0.85 |
| Auc\_Roc | 0.6 | 0.87 | 0.87 |

Clearly from above we can either go for Logistic Regression with balanced class weightage or Random Forest Classifier for predicting booked.

Top 10 important features **random sort, total\_price, booking\_window, length\_of\_stay, listing\_position, price\_usd\_corr, log\_historical\_price, num\_total, location\_score2 and site\_id** which a firm should focus on to convert more users.