Accuracy of Predicting Hotel Type in Expedia: A Data Mining Approach Based on the Users' Attribute Information

Department of Computer Science, University of Victoria

V00	**Shirley Shu**
V00	**Ziyue Cheng**
V00	**Nannan Zhang**

Dec.4, 2019

Abstract

With the massive change in technology, amounts of people prefer to make their traveling plan through online tools. This project intends to analyze whether Expedia website and mobile app's hotel recommendation is efficiency, based on their search and other attributes associated with that users' demand. We argue a successful recommendation is strongly related to user's background, but not in a decisive way. Indeed, through a set of data mining techniques we present other variables like user location country and hotel country, that would help to identify key recommendations in the community. The dataset used to train and evaluate the results of the execution of the classifiers. Python and WEKA are the main tools used for data mining and data evaluation.

Table of Contents

1. Introduction

Currently, people always face plenty of different options when choosing products that they are interested in. In their daily operations, it is widespread to use search engines for their preferences. So, optimizing search engines is necessary. However, the abundance of information available make very difficult for users to make the best choice, and more importantly filtering those appropriate source. In the scope of the hospitality industry, Expedia [1] stands out as a well-known hotel ordering application with high activity. In fact, Expedia has established a mechanism of rank and introduce their hotel in a way that motivates and attracts customers to make a purchase. Therefore, how to choose the hotel ranking has become the most important, which is related to the users' experience when using the web page.

1.1 What is Kaggle

As an online community of data scientists and machine learners, Kaggle has a famous slogan, "your home for Data science," so thousands of companies would like to collaborate with Kaggle. They would publish their data sets when they want to build some new technology contributing to their product. Win prizes will be set to attract more data mining enthusiasts and practitioners. Such people join this competition not only for the bonus but also for improving their data mining skills through real-world problems. In addition, the data sets posted on the website can be considered close to real data, which makes our research conclusions of great value to real problems. In Kaggle, people not only can have various kinds of data set but also can discuss with people who have the same interests and goals with you.[2] Therefore, Kaggle is an excellent platform for data mining enthusiasts to demonstrate and improve their data science knowledge.

1.2 What is Expedia

Expedia is an online travel agency. The website and mobile app can be used to book airline tickets, hotel reservations, car rentals, cruise ships. It also provides vacation packages, including all services the customers needed during their vacation. Figure 1 shows the search screen on the Expedia homepage. Users can choose Destination, Check-in date, Check-out date, the numbers of adults and children, the numbers of rooms. With these conditions, Expedia can rank hotels and selects the best hotels to recommend to users.

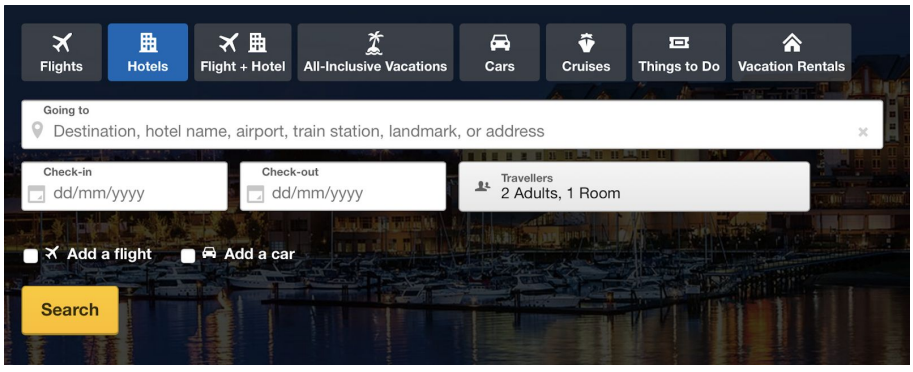

Figure 1: The homepage of Expedia.ca

2. Related Work

The study of search engines is a challenge around the hospitality industry's recommendation to their customers. However, we present this project as part of the Expedia Hotel Recommendations Competition on Kaggle that challenges researchers uncover interesting findings on Expedia Search Data. Therefore, we present an overview of different approaches towards this Question and Answer platform to highlight relevant aspects of our study.

A particularly interesting piece of related work is "The Surprising Psychology of Hotel Reservations," [3] a blog talks about how should the site presents pages so that people are more likely to book a hotel on the site. The most important thing is "Limit Your Options." Recent research suggests that consumers are six times more likely to buy something when faced with fewer options rather than an exhaustive list. In order to do this, Expedia.ca controls how the information is presented to the consumer, providing upgrade options as the guest progresses in the booking process. Therefore, in this report, we focus on which data is more important to users, especially when we consider how to select different hotels for different needs of the particular target population.

The other interesting point is called "Everyone else is doing it." The phenomenon of FOMO(Fear of Missing Out) is real and more rampant than ever. People all have FOMO. Expedia.ca shows how many people have already booked a stay within a recent span of time(10 people have booked a room within the last 12 hours) or by sharing how many of a specific rate or room type is still left in inventory(only 1 Queen Suites available for these dates). [4] The shows create a sense of nervousness, making users feel if they are not ordering now, then the room will be fully booked in the next second. Therefore, one of the things our team focuses on is the number of hotel bookings after searching.

3. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.[5] The data preprocessing is aiming at clean and data attributes separation. In order to do this, we have the following list to do:

1. **Raw data and structure analysis**: To identify the original data set and studies the elements according to the domain of the information
2. **Data transformation and reduction:** To transform original values into types in the domain of the study and identify the attributes which we need and delete useless information
3. **Data visualization and preparation:** To compute and calculate raw data to transform it into information, (i.e., normalization and aggregation)

4. **Training and Testing data set generation:** To generate the files required for the data mining process (i.e., ARFF files for the WEKA tool)

The details for each step are described in the next sections.

3.1 Raw data and structure analysis

The dataset collected to train is the *train.csv* file provided by Kaggle: Predict hotel type in Expedia.[5] Because of the huge dataset, we select 10000 entries. The attributes information is shown in Figure 2 below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 24 columns):
date_time                    10000 non-null object
site_name                    10000 non-null int64
posa_continent               10000 non-null int64
user_location_country        10000 non-null int64
user_location_region         10000 non-null int64
user_location_city           10000 non-null int64
orig_destination_distance    6271 non-null float64
user_id                      10000 non-null int64
is_mobile                    10000 non-null int64
is_package                   10000 non-null int64
channel                      10000 non-null int64
srch_ci                      9993 non-null object
srch_co                      9993 non-null object
srch_adults_cnt              10000 non-null int64
srch_children_cnt            10000 non-null int64
srch_rm_cnt                  10000 non-null int64
srch_destination_id          10000 non-null int64
srch_destination_type_id     10000 non-null int64
is_booking                   10000 non-null int64
cnt                          10000 non-null int64
hotel_continent              10000 non-null int64
hotel_country                10000 non-null int64
hotel_market                 10000 non-null int64
hotel_cluster                10000 non-null int64
dtypes: float64(1), int64(20), object(3)
memory usage: 1.8+ MB
```

Figure 2: Attributes information

This file contains 24 attributes. These attributes shows the users status, including where are the users from and where are they going, how many people and children are involved during this trip. Also, these attributes present the hotel status, the hotel country, destination, is booking or not, the criteria for the users decide whether or not to book the particular hotel.

3.2 Data transformation and reduction

We mainly focus on whether the attributes that appear on the Expedia homepage are important to travelers when they book a hotel. Therefore, we remove some unnecessary attributes. For example, users ID, which we have little to do with the hotel recommendation. Another example, is_mobile, it means travelers use mobile or computer to search, which is also unnecessary

attribute for our research. We only keep the relevant information to evaluate our prediction and Table 1 shows the relevant attributes that we selected and their description.

Table 1: The list of attributes and their description

Attribute	**Description**	**Typing to measure**
User_location_country	ID of customer's country (from 1 to 100)	Experience
Is_package	1 if booking/click was part of package, 0 otherwise	Experience
Hotel_country	The country of hotel	Experience
Srch_children_cnt	Number of children	Experience, Information from expedia homepage
Srch_adults_cnt	Number of adults	Experience, Information from expedia homepage
Srch_rm_cnt	Number of rooms users intend to book	Experience, Information from expedia homepage
Hotel_cluster	ID of hotel cluster	Experience
Is_booking	1 if a booking after searching, 0 if not booking after searching	Experience

3.3 Data visualization and preparation

We simply graded some attributes so that it can be better represented and visualized in the following data analysis, especially using Decision Tree. We classify the attributes of hotel_cluster, Srch_children_cnt, Srch_adults_cnt, Srch_rm_cnt. Figure 3 shows the correlation of the attributes. Red area means two attributes are closely related and blue area means there is little correlation between these two attributes. Through that we can prove these attributes are independent. Figure 4 shows the classification of the attributes and Table 2 introduce the detail range of the classification.

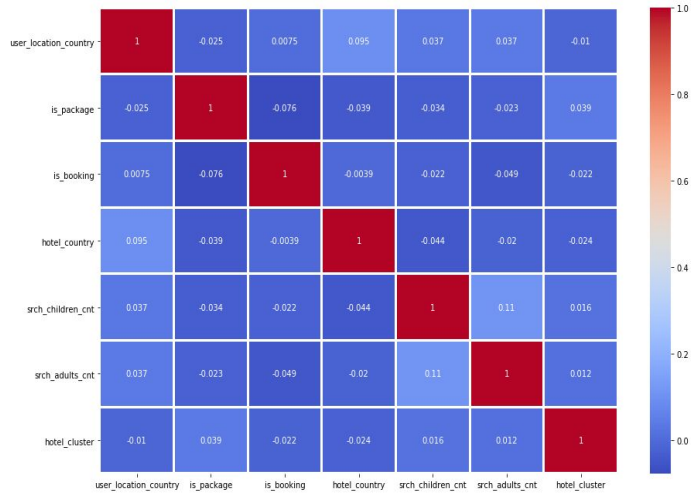

Figure 3: correlation of attributes

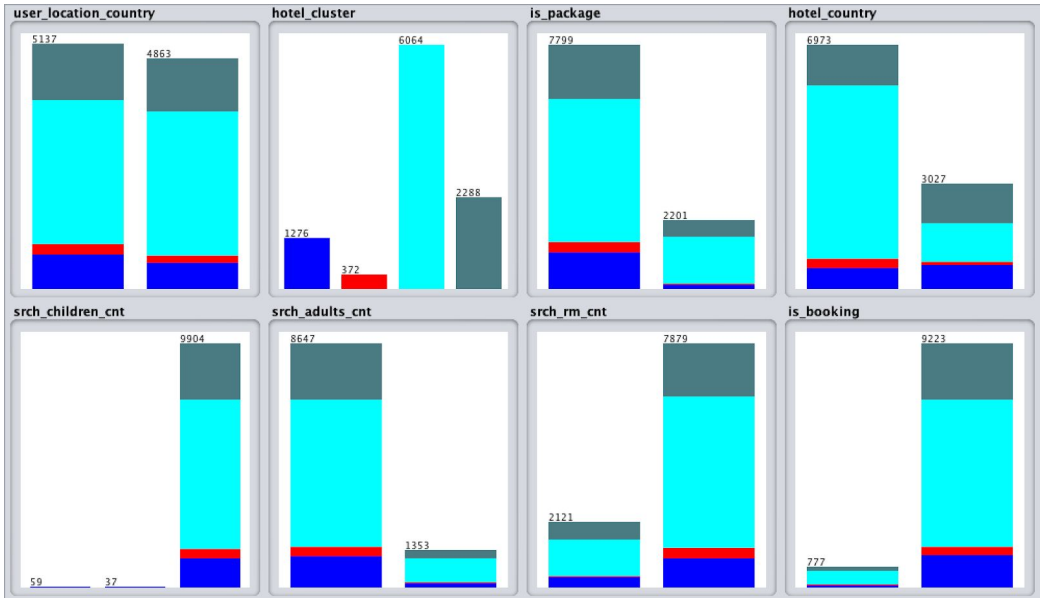

Figure 4: Data Visualization from WIKA

Table 2: Data Classification Group

user_location_country	Group	Is_package	Group
[0,50)	1	0	1
[50,100]	2	1	2
Hotel_country	**Group**	**Srch_children_cnt**	**Group**
[0,50)	1	0	1
[50,100]	2	(0,2]	2
		(2,9]	3
Srch_adults_cnt	**Group**	**Srch_rm_cnt**	**Group**
0	1	0	1
(0,2]	2	(0,1]	2
(2,9]	3	(2,9]	3
Hotel_cluster	**Group**	**Is_booking**	**Group**
[0,25)	1	0	1
[25,50)	2	1	2
[50.75)	3		
[75,100)	4		

3.4 Training and Testing data set generation

Finally, we transform the information from the CSV file to ARFF file ready to be imported in WEKA.

4. Data Mining

The data mining process aim to find hidden information from raw data by implementing algorithms and filters to get better results. For this, after preprocessing the data, our team make advantage of WEKA to classify the attributes in order to identify those that are significant in the relationship of hotel cluster that users will book or not.

The steps are as follows:

1. **Arff Loader:** This element loads the information into WEKA to be processed. In our case from a CSV file into an ARFF file, that is compatible with WEKA.

2. **Class Assigner:** This step selects the Class we want to predict. In this case, we test the is_booking and hotel_cluster separately.

3. **Cross Validation Fold Maker:** This step is in charge of delivering the training set and the test set to each classifier. In this case we used 7 classifiers. (i.e., Naive Bayes, Logistic Regression, J48, ID3, Classification Via Regression, JRip, and IBk)

4. **Classifier performance evaluator:** This step evaluates the results of the classifier and creates an output.

5. **Views:** The results of the classifier are presented in the information viewer prior analysis.

4.1 Test for *is_booking* classifier

Our purpose is to know whether the information users given in the Expedia homepage is helpful or not for Expedia to recommend a satisfying hotel for customers. *is_booking* is chosen to test because customers do book the hotel when *is_booking* equal to 1, which means expedia's recommendation hotel satisfies users.

After transform the CSV file to ARFF file, we executed 7 classifiers that are suitable for our data set: Naive Bayes, Logistic Regression, J48, ID3, Classification via Regression, JRip and IBk. The results of the executions are described in Table 3 showing that ID3 stands out over the others with 92.2647% accuracy.

Classifiers	**Accuracy %**
Naive Bayes	89.9706 %
Logistic	92.1176 %
J48	92.2647 %
ID3	92.17%
Classification Via Regression	92.1765 %
JRip	92.2647 %
IBk	92.2059 %

Table 3: Results of the classifiers available in WEKA

4.1.1 Result of *is_booking* classifier

Given that ID3 is a tree classifier it seems more comfortable to interpret while analyzing the data. Figures 5 and 6 show the complete output of the classifier in WEKA. According to the WEKA output, we manually created the decision tree resulting of the classifier shown in Figure 7.

```
=== Run information ===

Scheme:       weka.classifiers.trees.Id3
Relation:     classified-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    10000
Attributes:   8
              user_location_country
              hotel_cluster
              is_package
              hotel_country
              srch_children_cnt
              srch_adults_cnt
              srch_rm_cnt
              is_booking
Test mode:    10-fold cross-validation
```

Figure 5: ID3 Classifier output for hotel_cluster (part 1)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9217               92.17   %
Incorrectly Classified Instances       783                7.83   %
Kappa statistic                         -0.0012
Mean absolute error                      0.14
Root mean squared error                  0.2663
Relative absolute error                 97.6256 %
Root relative squared error             99.4818 %
Total Number of Instances           10000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.626     0.120     0
                 0.999    1.000    0.922      0.999    0.959      -0.007   0.626     0.946     1
Weighted Avg.    0.922    0.922    0.851      0.922    0.885      -0.007   0.626     0.882

=== Confusion Matrix ===

    a    b   <-- classified as
    0  777 |    a = 0
    6 9217 |    b = 1
```

Figure 6: ID3 Classifier output for hotel_cluster (part 2)

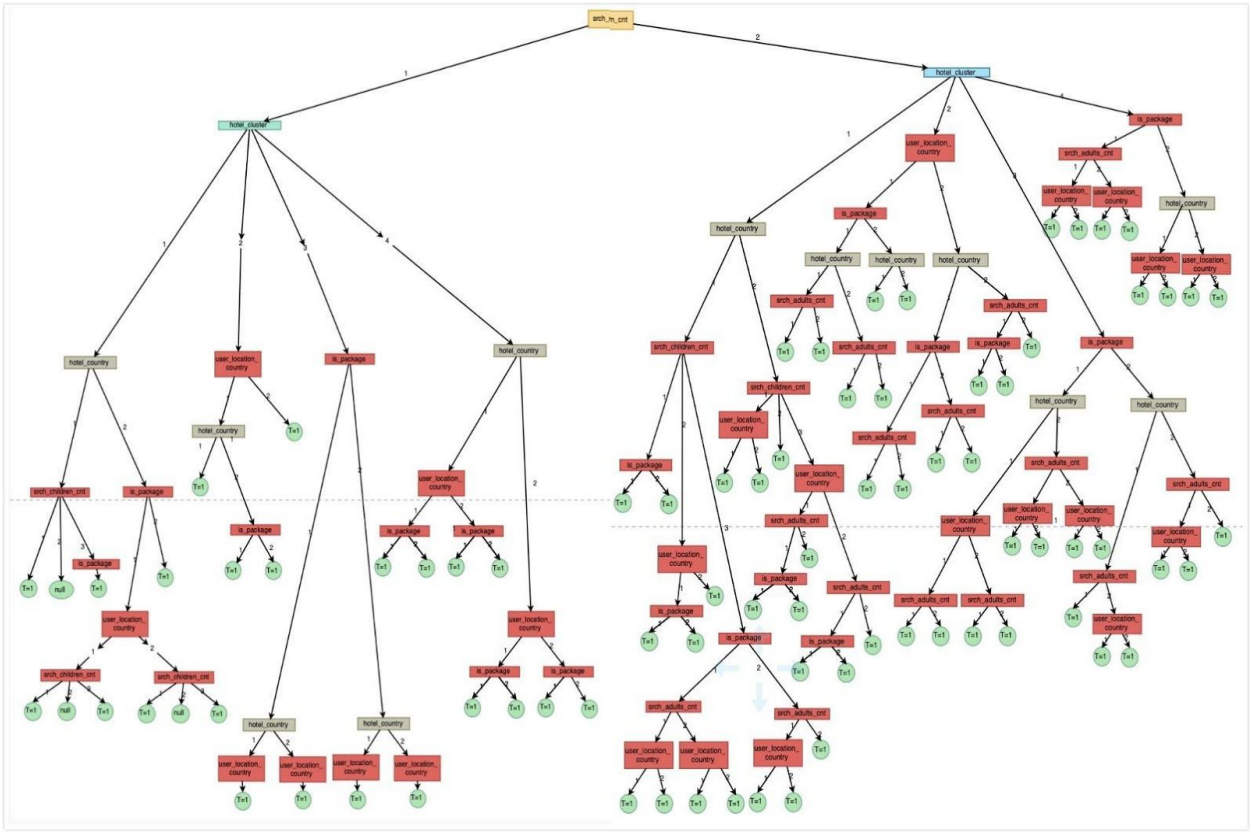

Figure 7: ID3 Decision Tree of is_booking classifier

4.1.2 Observation of is_booking classifier

Figure 6 exhibits the values distribution according every attribute. As noted in Figure 7, the first level in ID3 tree is divided by *srch_rm_cnt* and the second level is divided by *hotel_cluster*. We can see that the right part has the heavier than the left part, so the room count is one of the most important attributes that related to the recommendation. However, it is a big and mass decision tree, and each variable almost shows same times in the tree, so we cannot clearly find which variable have higher rank to this model. Therefore, we decided to make one more test to analyse the hotel_cluster classifier.

4.2 Test for hotel_cluster classifier

We do more analysis, testing for *hotel_cluster* as it is the second level in our ID3 decision tree which has been shown in section 4.1. We want to know which variable in the Expedia homepage play an important role to help Expedia make a closed prediction.

After transform the CSV file to ARFF file, we executed 7 classifiers that are suitable for our data set: Naive Bayes, Logistic Regression, J48, ID3, Classification via Regression, JRip and IBk. The results of the executions are described in Table 2 showing that J48 stands out over the others with 63.86% accuracy.

Table 4: *Results of the classifiers available in WEKA*

Classifiers	**Accuracy %**
Naive Bayes	63.97 %
Logistic	63.67 %
J48	63.86 %
ID3	63.89 %
Classification Via Regression	62.79 %
JRip	62.81 %
IBk	63.87 %

4.2.1 Result of hotel_cluster classifier

Given that J48 is a tree classifier it seems more comfortable to interpret while analyzing the data. Figures 5 and 6 show the complete output of the classifier in WEKA. Finally, Figure 7 shows the tree resulting of the classifier.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     classified-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    10000
Attributes:   8
              user_location_country
              hotel_cluster
              is_package
              hotel_country
              srch_children_cnt
              srch_adults_cnt
              srch_rm_cnt
              is_booking
Test mode:    10-fold cross-validation
```

Figure 8: J48 Classifier output for hotel_cluster (part 1)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6386               63.86   %
Incorrectly Classified Instances      3614               36.14   %
Kappa statistic                          0.2376
Mean absolute error                      0.2521
Root mean squared error                  0.3554
Relative absolute error                 89.6632 %
Root relative squared error             94.8002 %
Total Number of Instances           10000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.075    0.000    1.000      0.075    0.140      0.257    0.670     0.305     1
                 0.000    0.000    0.000      0.000    0.000      -0.003   0.539     0.043     2
                 0.902    0.617    0.692      0.902    0.783      0.341    0.668     0.704     3
                 0.359    0.153    0.410      0.359    0.383      0.216    0.630     0.340     4
Weighted Avg.    0.639    0.409    0.641      0.639    0.580      0.289    0.655     0.545

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   96    0  659  521 |    a = 1
    0    0  306   66 |    b = 2
    0    0 5468  596 |    c = 3
    0    2 1464  822 |    d = 4
```

Figure 9: J48 Classifier output for hotel_cluster (part 2)

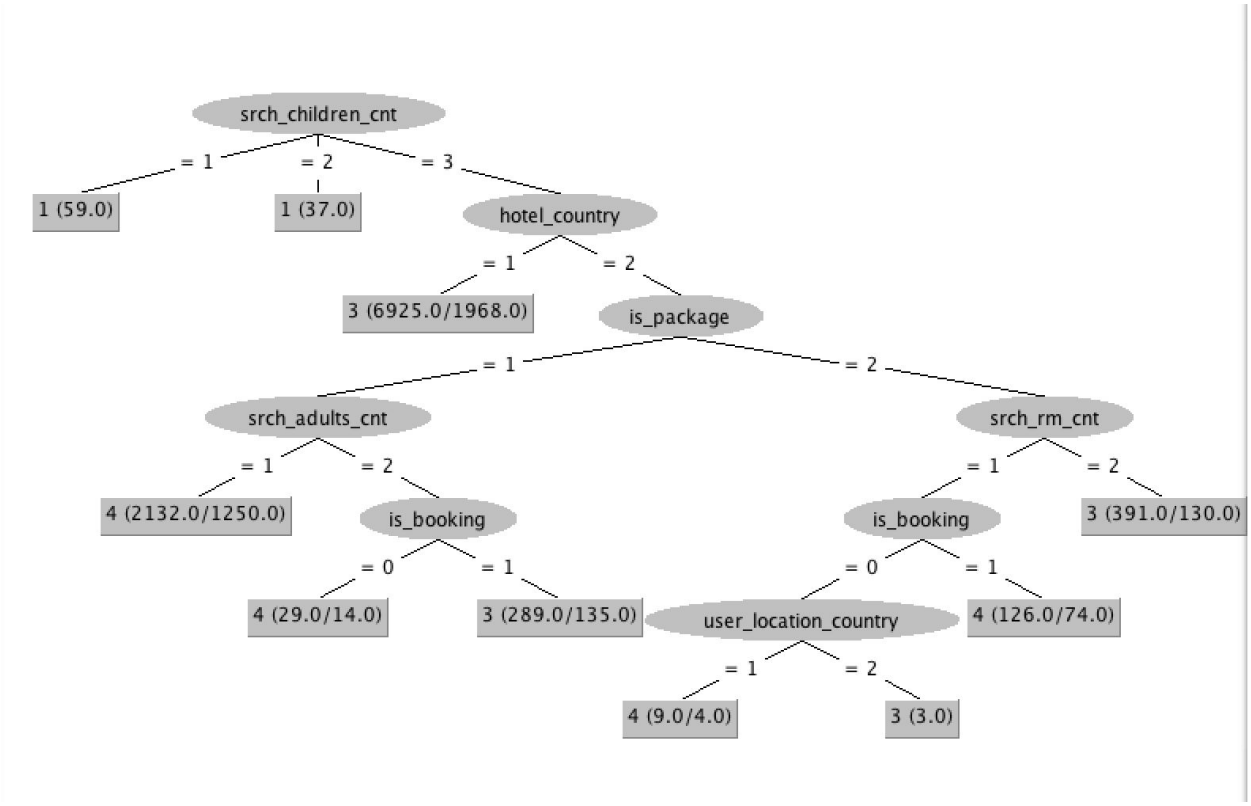

Figure 10: J48 Tree

4.2.2 Observation of hotel_cluster classifier

Figure 7 exhibits the values distribution according every attribute. As noted in Figure 8, *srch_children_cnt* is the most important attribute to classify the information due to the fact that the density of the hotel cluster is in the first part of the plot. This means that *srch_children_cnt* in group 1 and 2 tend to have hotel cluster in group 1 compared with other hotel cluster. *hotel_country* is another important attribute. Users prefer to have hotel cluster in group 3 when *hotel_country* is in group 1. However, *is_package*, *srch_adults_cnt* and *user_location_country* are not such important than *srch_children_cnt* and *hotel_country* as they are close to the leaves in decision tree. When users not searched with a package (*is_package* = 1) and adults count less than 2 (*srch_adults_cnt* = 1), users tend to choose hotel cluster in group 4. When users searched with a package (*is_package* = 2) and room count more than 2 (*srch_room_cnt* = 2), users tend to choose hotel cluster in group 3.

5. Conclusion

As mentioned in Section 1, for the hospitality industry is very important to have efficient and attractive recommendation to their customers. Search engines present an essential role as they directly connected to users. However, due to the massive search section, it can be very noisy to find the most appropriate attributes.

As shown on the homepage, expedia do have some attributes considered as significant search section. Our research based on them and argued that the hotel cluster that users choose are related to these attributes. As a consequence, we propose to mine the available information to find those additional attributes that predict whether users booking the hotel and which hotel they would choose.

We mined the data from Expedia data set, available on kaggle website. Using python packages like pandas, numpy, matplotlib, and seaborn we modified the raw data and performed some cleaning and transformation procedures to get the suitable data for our approach. Later, we query the data to create the proper CSV files to imported in the data mining tool WEKA.

Furthermore, we implemented different classifiers over the data and finally we chose ID3 given the acceptable accuracy over 90% for *is_booking* classifier and J48 for *hotel_cluster* classifier. As a result the ID3 tree exposed that all the attributes shown on the expedia homepage related to the booking percentage. In order to get a more specific answer, we test the *hotel_cluster* classifier. As a result the J48 tree exposed interesting findings such as the attributes that influence the most in the hotel recommendation, such as *srch_children_cnt* and *hotel_country.*

Finally, we concluded that hotel clusters that customers would choose based on different travel purpose. For expedia, the number of children and hotel country are significantly related to customers' choice.

Acknowledgments

This report is a project of 2019 Data Mining course SENG474 and it was presented as part of the final project of the course. We thank Dr. Alex Thomo for his guidance and helping us solve our problems during the course.

References

[1] Expedia Homepage, [Online]. Available: https://www.expedia.ca

[2] “your home for Data science”, [Online]. Available: https://en.wikipedia.org/wiki/Kaggle

[3] R. Lund, “The Surprising Psychology of Hotel Reservations”, , [Online]. Available: https://www.huffpost.com/entry/the-surprising-psychology_b_9663506

[4] “The Surprising Psychology Behind Successful Hotel Websites” , [Online]. Available: https://www.hospitalitynet.org/news/4071180.html

[5] “Data Preprocessing on Expedia Hotel Dataset”, [Online]. Available: https://www.kaggle.com/ajay1216/practical-guide-on-data-preprocessing-in-python