|  |
| --- |
| CI6227 dATA MINING GROUP PROJECT |
| The Forecasting of Airbnb’s Users’ Booking Destination |
| --based on multi-classification techniques of Voting Scheme and SVM |
|  |
| **Group Members** |
| CHENG SHANGYU - G1601216D |

Li PEI - G1602397A

Mulukutla MEENU - G1602411D

Rao RASHMITHA Ranganatha - G1602421B

|  |
| --- |
|  |

**Abstract**

This report is about investigation of the Airbnb dataset and development of model to predict users’ booking destination. The dataset originates from a progressing kaggle competition sponsored by Airbnb. We firstly did some exhaustive investigation on the dataset, investigated most components and gathered all elements we believed were valuable. During the process we portrayed and deciphered the hidden associations between attributes and target variable. Afterwards we developed a sensible model for this forecast assignment. To predict accurately as much as possible, we constructed a two-level classifiers model. The principal level is a paired classifier with Voting Mechanism joining linear, logistic and polynomial regressions together. The second level is a multiclass classifier which is the blend of SVM and multi-class one-against-rest logistic grouping. This procedure likewise included pattern recognition, feature transformation and extraction, model choice, thinking and depiction, and parameter tuning. Next we discuss the results and compare ours against other competing teams on the kaggle. At last, we suggest some possible improvements to make the prediction better in the future.

**Keywords**—Airbnb, Prediction, Kaggle, Two-level classification Model, Binary Classification, Multi-class Classification, Voting Mechanism

**Data Description**

In this challenge, we are given a list of users along with their demographics, web session records, and some summary statistics. Each participatory team was asked to predict which country a new user's first booking destination will be. There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. ‘Other’ can be any other country which is not included in the above list and NDF means the users cannot make up their minds yet. The training and test sets are split by dates. In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

There are six sets of data: train\_users (serving as a training set), test\_users (serving as a test set), sessions (web sessions log for users), countries (summary statistics of destination countries in the dataset and their locations), age\_gender\_bkts (summary statistics of users' age group, gender, country of destination) and sample\_submission (correct format for submitting your predictions). Of all, the first two datasets are the most important and their field descriptions are as follows:

|  |  |
| --- | --- |
| Field Name | Description |
| id | user id |
| date\_account\_created | the date of account creation |
| timestamp\_first\_active | timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up |
| date\_first\_booking | date of first booking |
| gender |  |
| age |  |
| signup\_method |  |
| signup\_flow | the page a user came to signup up from |
| language | international language preference |
| affiliate\_channel | what kind of paid marketing |
| affiliate\_provider | where the marketing is e.g. google, craigslist, other |
| first\_affiliate\_tracked | whats the first marketing the user interacted with before the signing up |
| signup\_app |  |
| first\_device\_type |  |
| first\_browser |  |
| country\_destination | this is the target variable you are to predict |

**Data Explorations & Preprocessing**

Based on our exploration of the dataset provided, a couple of interesting features have been found and summarized as follows:

1. Almost 90% of users fall into two categories: NDF, US

Figure One the Users Categories Distribution

1. User with non-date\_first\_booking is NDF, therefore no need to predict whether it is NDF.

Figure Two Column Graph of Booking Demographic Distribution Grouped by Date\_first\_booking

1. Because of a majority of US-user, it is better to tell US-user from others at first in order to avoid too generalization.

Figure Three Pie Chart of Users’ Nationality

1. The majority of the training data provided comes from the latest 2 years. In fact, if we limited the training data to accounts created from January 2013 onwards, we would still be including over 70% of all the data. This matters because, referring back to the notes provided by Airbnb, if we want to use the data in sessions.csv we would be limited to data from January 2014 onwards. Again looking at the numbers, this means that even though the sessions.csv data only covers 11% of the time period (6 out of 54 months), it still covers over 30% of the training data – or 76,466 users.

Figure Four the Number of Accounts Created by time

1. Booking month may be a useful feature (there may be some holiday factors which contribute to the cyclical pattern: two or three months trip planning ahead of major holidays such as Christmas or Thanksgiving may explain the surge in the booking volumes that occurs on a yearly basis)

Figure Five the Volume of Bookings in Each Month

Figure Six the Volume of Bookings in Each Week Day

1. If the difference between booking date and account creation date is less than 2, namely, 0 or 1,US-users tend to have larger ratio than Non-US-users. In this case, we came up with a useful feature that if difference between booking date and account creation date is less than 2,the feature vector is [1, 0],otherwise [0, 1]. The date difference < 11 count most of the date difference distribution, and it is the values of these date differences that have evident different properties among all the countries. So we created a 12-dimension vector, used the index as date difference itself if it was <= 10 or used index 11 otherwise. This feature representation gave us a good prediction result.

Figure Seven Distribution of Time Gap between Booking Date and Account Creation Date

1. Age could play a very important role in terms of predicting users’ behaviors hence a closer look must be taken in order to improve the performance. First of all, this field has a very high rate of NaN. Moreover, there are some outliers in the age set. It could be because the age input field was not sanitized or there was some mistakes handling the data. To handle this case, we preprocessed data to discard users with age over 100, or fewer than 5.

Figure Eight Age Distribution of Users

1. We group three parameters of first device type, first browser and signup-app together to reflect users' web browsing preferences since analyzing them separately won’t yield much significant results.

Figure Series Composition of first device type, first browser and signup-app among users

1. For the parameters of affiliate provider and first affiliate tracked, their separation gave us trivial results, so we combine them altogether as a 2- dimension vector, which could also reflect a user’s habit.

Figure Series Composition of affiliate provider and first affiliate tracked among users

1. We can develop a scoring method for evaluating the accuracy of our predictions. This validation method is that for every user’s prediction, we can list at most 5 countries in order. The higher the prediction is ranked on the list, the larger score we will assign to this prediction and vice versa.

**Model Evaluation**

1. The data is unbalanced because US counts for a large proportion of the data. If we used multiclass classification classifiers directly, the performance was a disaster. Therefore, we built a two-level classifier to separate the US and other countries first.

In the level one classification, the voting mechanism is employed to account for different configurations of scenarios. The implementation requires 2 out of 3 predictions to be yes in order to put the final prediction into the “yes” category and the same criterion holds true for the “no” category.

2. The first level is a binary classifier with Voting Mechanism combining linear, logistic and polynomial regression. The second level is a multiclass classifier which is the combination of SVM and multi-class one-against-rest logistic classification. Afterwards, we combine, sort and get top 5 unrepeated countries with highest probabilities. The results with the highest ranking will be compared first with the actual destinations and if correct, the score will be highest (1) and there is no need to consider the rest predictions. Otherwise, the score will be reduced stepwise for the following match.



Combine, sort, get top 5 unrepeated countries with highest probablities

Logistic: select top 5 countries with highest probabilities

3. To validate our model, we use 10-fold cross validation. In this way, each data can be used both to train and validate the model.

**Results Discussions**

The baseline for model performance must be established. Since the NDF and US count over 80 percent of the training set, the baseline only predicts these two countries. It predicts the NDF and US alternatively, like NDF, US, NDF, US, etc. The prediction score on the validation set was 0.78640. It means that 0.7864 is the minimum accuracy ratio we must achieve and any model adjustment must achieve good positive increase on the score.

We have tried the following combinations of different methodologies and the summary of the results is as follows:

|  |  |
| --- | --- |
| Model | Accuracy Ratio |
| Baseline (destination only choosing between U.S. and NDF) |  |
| ploynominal regression |  |
| logistic regression |  |
| linear regression |  |
| Voting Mechanism |  |
| SVM |  |
| multi-class logistic classification | |
| SVM + logistic classification |  |

**Conclusions:**

* Among three models, polynomial regression performs best. And as for linear and logistic regression, they perform a little worse. Nevertheless, when combining three models into the Voting Mechanism, it outperforms all three models.
* The ranking of our uploads stands as \* out of \*
* In VM, there are several different parameters for different member models; In linear regression, the parameter is the penalty for weight in regularization to avoid over-fitting and under-fitting problem. In polynomial regression, the parameter is the degree of polynomial features. In logistic regression, the parameter class weight is ’auto’, which would automatically adjust weights inversely proportional to class frequencies in the input data. As for SVM, the parameter sigma represents the penalty of the error term.
* According to above results table, VM outperforms other methods. And the reason for it is quite obvious that VM overcomes the flaws of the single model. For example, linear regression doesn’t perform well when the relation between features and result is not linear, but this defect doesn’t matter in logistic and polynomial model. As for logistic and polynomial regression, they may generate disappointing results when processing some extreme cases, which could be processed well on linear regression.
* For the further improvement, we may assign different weights to each model in VM in order to improve the accuracy. The road is still long ahead and we will keep honing our skills.