### 

**Capstone Project – Machine Learning**

**By Sumita Soundararajan**

***Foundations of Data Science Workshop by Springboard***

**Airbnb Recruiting - New User Bookings**

## ***Where will a new guest book their first travel experience?***



**Problem**

We are required to predict the new users’ first booking destination at the country level. There are 11 destinations to choose from, including Australia, Canada, Germany, Spain, France, Great Britain, Italy, Netherlands, Portugal, the U.S., and all the rest are labeled as” others”. Users who haven’t made a booking are categorized with the” NDF” label in the destination field.

**Supervised Classification problem**

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

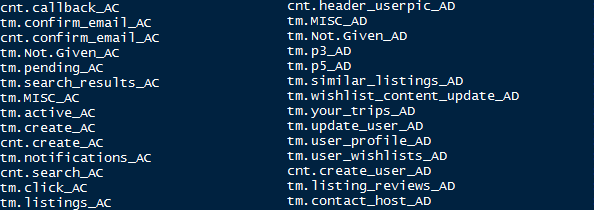
The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

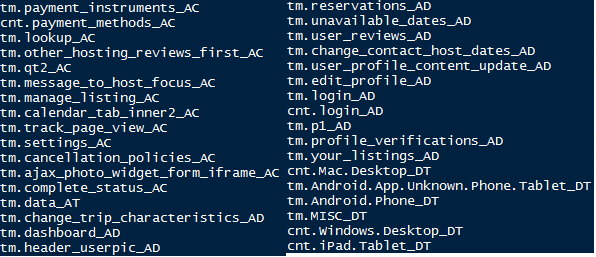
Supervised learning problems can be further grouped into regression and classification problems.

* Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Given that we have to predict the first booking destination (Y) based on many features available in the sessions, training and test user datasets (X), ours is actually a Supervised problem. Also, there are 12 country\_destination or classes to be predicted, which makes it a multi-class classification problem.

For our problem, the dependent or predicted variable (Y) would be country\_destination and the independent or predictor variables would be the 62 aggregated features from the Base Feature dataset (inner join on Sessions dataset with Training user dataset) shown in the screenshot below.





**Model Selection**

There is a total of 12 classes to be predicted. However, the data is not balanced because NDF and US counts for a large proportion of the data. The ratio of destination NDF to PT is 542 which shows that data is highly imbalanced between the classes.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PT** | **AU** | **DE** | **NL** | **CA** | **ES** | **GB** | **IT** | **FR** | **Other** | **US** | **NDF** |
| 83 | 152 | 250 | 247 | 440 | 707 | 731 | 979 | 1435 | 3655 | 20095 | 45041 |

A total of 66 level binary classifiers to separate the data based on destinations – NDF and US, NDF and FR, US and DE and so on would be built as Training set.

First the Logistic regression algorithm would be used to build a model since it works best for binary classification problems.

The second method will address this multi-class classification problem with supervised learning of [Decision trees](http://scikit-learn.org/stable/modules/tree.html#tree). Decisions trees predict a target variable by learning simple decision rules detected in feature data, and are appropriate here for some of the following reasons:

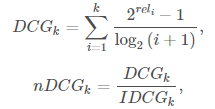
* can handle numerical and categorical data
* can handle multiple classes
* can handle class imbalance
* can handle missing data

The competition required a certain format on the submission where we had to list every user and the five most likely destinations which then after uploading to Kaggle website were evaluated using their NDCG system.

**Evaluation** **of Model**

To evaluate our solution, we will be using Kaggle’s submission system. The system gives us a score between 0 and 1 where 1 is the perfect solution. Kaggle uses NDCG (Normalized Discounted Cumulative Gain) which is a system for measuring the quality of rankings.

The NDCG calculation is shown as:



where is th is the relevance of the result at position

is the maxis the maximum possible (ideal) DCG for a given set of queries. *k* is the number of predictions and *rel* is either 0 or 1, its 1 for the correct prediction and 0 otherwise. All NDCG calculations are relative values on the interval 0.0 to 1.0.

Here, for every user’s prediction, we are listing at most 5 countries in order. The score of the predictions are calculated from the position of the correct prediction. Each user prediction gets a normalized score between 0 and 1 and the total score is then calculated from the average of all user prediction scores. The score is only calculated from the position of the correct prediction and any predictions with lower probability will not reduce the score of the user prediction.

For example, if the correct destination is predicted as the most probable destination then we would get a score of 1 and if the correct destination was not part of the predicted destinations we get a score of 0. If the correct destination was predicted as the 2nd most likely destination we would get a score of 0.632. [(21-1)/log2(2+1)] = 0.632.