To Run Code: Run Jupyter notebook, and open TripAdvisor.ipynb file.

It consists of code, documentation, and explanation.

Requirements*: Python 2.7 and above, Scikit-Learn, Ipython, XGBoost, Numpy, csv, Pandas .*

*Results is in **output.csv**, consisting of user's next prediction, and in **output2.csv** is user's next up to 5 predictions.*

ImportantPoints:Modelfinalized:XGBoostClassifier

Experimented with Random Forest, KNN, and Rule Based Approach. Experimented with hybrid method which will return the hotel present in both most frequent + Machine Learning output. Didn't give good results.

1. *First Step is to read data and see what kind of values are present in data.*
2. *Apply Methods: After Observing Data we have applied two methods: Machine Learning Algorithms and Rule Based Approach:*

   Machine Learning Algorithms: As Data is low in parameters, plus the data entries is also not enough, its better to go for Non-Parametric Equation as parametric one might overfit. So, We can think of KNN, Decision Trees(Random Forest, XGBOOST). With KNN, it was working good, when there was no constraint on repeating the user's history, but again it was giving good accuracy at K=1, but very unstable. Which is why, the idea of Ensembles look best.

   Rule Based Approach: As, we could see, gender didn't play any role in determining the famous hotels. The only other parameter that left is continent of user, and if that even doesn't play any role, it will mean its totally random selection. So, One experiment to test a frequency per continent results has also been done.

3. *After Experimenting with both approaches, it has been realized that rule based method isn't performing better than Machine Learning Algorithm. Also, Both Random Forest, and XGBoost are giving almost same results.*
   WHY XGBOOST?
   I chose XGBoost as it was giving better results in some points, plus, Boosting is based on weak learners (high bias, low variance). Boosting reduces error mainly by reducing bias (and to some extent variance, by aggregating the output from many models). On the other hand, Random Forest uses fully grown decision trees (low bias, high variance). It tackles the error reduction task in the opposite way: by reducing variance, and for given dataset we need to reduce bias.

4. *Also, I have used 2 methods to return the hotel which is not present in user's history. First, which only consider Machine Learning Result. Second Hybrid Method which consider Both Frequency of Hotel as per continent, and Machine Learning Result. Upon Experimenting, I couldn't get good results with hybrid method, and decided to go with normal Machine Learning Approach. Maybe we need to determine proper weightage of what needs to given to frequency and Machine Learning Result.*

5. *The Optimal Parameters have been obtained, Via Grid Search CV and predictions based on optimal parameters have been saved, in csv files.*