Statistical Learning-Classification

Project Title: Expedia Hotel Recommendation

Project Number:

Group Members:

Surname, First Name	Student ID	STAT 441	STAT 841	Your Dept. e.g. STAT, ECE, CS
		☑	☐	
		☑	☐	
		☑	☐	
Hanxin Zhang	20437677	☑	☐	CS & STAT

Your project falls into one of the following categories. Check the boxes which describe your project the best.

1. ☑ **Kaggle project.** Our project is a Kaggle competition.
   - This competition is active ☐ inactive ☑.
   - Our rank in the competition is *Expedia Hotel Recommendation*.
   - The best Kaggle score in this competition is ... ..., and our score is ....
2. ☐ **New algorithm.** We developed a new algorithm and demonstrated (theoretically and/or empirically) why our technique is better (or worse) than other algorithms.
3. ☐ **Application.** We applied known algorithm(s) to some domain.
   - ☐ We applied the algorithm(s) to our own research problem.
   - ☐ We tried to reproduce results of someone else's paper.
   - ☐ We used an existing implementation of the algorithm(s).
   - ☐ We implemented the algorithm(s) ourself.

Our most significant contributions are (List at most three):

(a) .

(b) .

(c) .

List the name of programming languages, tools, packages, and software that you have used in this project:
Python(numpy, sklearn, Rodeo), R(xgboost, caret), LaTeX, Bash

Contents

1 Background Introduction **2**

2 Methods **3**

2.1 Initial Attempts 3

2.2 New Approach for Data Engineering 3

3 Conclusion **4**

1

Background Introduction

1.1 Dataset

Expedia hotel recommendation is a Kaggle recommendation during the middle of 2016. In this competition, Expedia is looking to personalize hotel recommendations for each customers by predicting the right hotel group (out of a hundred) that a user is going to booked. The given dataset contains various factors like booking date and time, original and arrival destinations, and the submission is allowed up to five ranked hotel clusters to be evaluated. The training set contains 2013 and 2014 site traffics with 37 million events and 24 features, while the testing set contains 2015 traffics with 2.5 million events and 22 features.

1.2 Evalutation

In this Kaggle competition, the submissions will be scored on the Mean Average Precision at 5. The formula for MAP@5 is shown below.

$$MAP@5 = \frac{1}{|U|} \sum_{|U|}^{u=1} \sum_{min(5,n)}^{k=1} P(k)$$

This scoring method will weight the score based on the order of the submission. For example, if the correct hotel cluster is 3 and the submission is "3 10 52 98 75", then it would receive the highest score as the correct submission is ranked first. A MAP@5 score for pure random guesses is 0.022.

2

Methods

2.1 Initial Attempts

2.1.1 Data Preprocessing

The first attempts before trying out any methods is to filter out unnecessary data. Since the training set is very big (4 GB), it is necessary to filter out incomplete rows that contain missing values. Moreover, time stamp column is splitted into date and time columns so they can be comprehended by machine learning algorithms. Lastly, highly correlated variables were removed. For example, there were columns for continents and countries of a use. The continent column has been removed since it is obvoius that someone from France would be booking from Europe.

2.1.2 Algorithms

After reformatting the data, various machine learning algorithms have been implemented including Linear Disriminant, KNN, Decision Trees, Naive Bayes, XGBoost and AdaBoost. Unfortunately, the score only ranges as high as around 0.10.

2.2 New Approach for Data Engineering

3

Conclusion