Machine Learning Engineer Nanodegree - Capstone Proposal

Bharat Krishna Raghavan

October 2018

Domain Background

The domain of the project is building a recommendation system for booking hotels via travel booking websites such as Expedia. Recommendation systems have been in use across various domains in e-commerce. They attained fame and popularity with the Netflix Prize which was a competition to build a movie recommendation system.

Two of the most popular techniques used in recommendation systems are collaborative filtering, which uses behavior from a large number of users and content-based filtering which uses a profile of user's set preferences to make recommendations[1]. Understanding of these techniques gives us a background on what a recommendation system comprises of.

The specific problem at hand in our project is to recommend hotels. This is a useful problem to solve since selecting a hotel can be a daunting task for a user since many hotels can be found in every destination, making it difficult to choose from. From the point of view of the booking website, recommending the right hotels would result in happier customers, which in turn results in more revenue.

Problem Statement

The problem is from the Kaggle competition "Expedia Hotel Recommendations"[2].

Expedia has in-house algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center, etc) are grouped together. These hotel clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data[3].

Our goal is to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups, based on their search and other attributes associated with that user event.

Datasets and Inputs

The dataset is obtained from Kaggle. The data is a random selection from Expedia. The training set contains 37,670,293 samples and 24 features. The test set contains 2,528,243 samples.

The train and test datasets are split based on time: training data from 2013 and 2014, while test data are from 2015. Training data includes all the users in the logs, including both click events and booking events. Test data only includes booking events[3].

Solution Statement

The solution to the problem is to come up with a good supervised learning model to predict the top five hotel clusters for a user.

Supervised learning algorithms such as Naive Bayes, Decision Trees, Ensemble Methods, K-Nearest Neighbors, Support Vector Machines and Logistic Regression would explored and their metric score would be calculated (see section "Evaluation Metrics" for more information on the metric to be measured).

Benchmark Model

A naïve predictor which would always return the top five most frequently occurring hotel clusters would be used as the Benchmark Model. Our model is expected to perform better than this naïve predictor.

Evaluation Metrics

The evaluation metrics defined in the Kaggle competition will be used[4].

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

where |U| is the number of user events, P(k) is the precision at cutoff k, n is the number of predicted hotel clusters.

Since we predict 5 hotel clusters for each user event, Mean Average Precision @ 5 would be a good evaluation metric.

Project Design

The project design would consist of the following steps:

1. **Install required software packages** such as Pandas, Numpy, Scikit-learn, etc.

2. **Exploratory Data Analysis and Visualization**
   The dataset would be studied by obtaining the descriptive statistics and visualized to create an intuition about the data. Plots such as the most frequently occurring hotel cluster, correlation between features.
3. **Data Cleaning and Preprocessing**
   Based on the analysis of the data, data cleaning would be performed to eliminate outliers and empty values. Preprocessing the data using techniques such as one-hot encoding would be performed if needed.
4. **Establishing a Training and Testing pipeline**
   - Fitting a Model
     Many supervised learning models would be fitted on the training data, following these steps:
     - Hyperparameter tuning to prevent overfitting and underfitting
     - Performing prediction on the test dataset
     - Selecting the best model
   - Based on the Evaluation Metrics, the best model will be selected.

References

[1] https://en.wikipedia.org/wiki/Recommender_system#Approaches

[2] https://www.kaggle.com/c/expedia-hotel-recommendations

[3] https://www.kaggle.com/c/expedia-hotel-recommendations/data

[4] https://www.kaggle.com/c/expedia-hotel-recommendations#evaluation