Machine Learning Engineer Nanodegree

Capstone Proposal

Neethu Wilson
October 13th, 2018

Proposal

Domain Background

Airbnb is an online home rental platform based in San Francisco that lets people list, find, and rent short-term lodging. In recent years, Airbnb has evolved its peer-to-peer model to give hosts the technology tools they need to run a seamless, sophisticated operation.

Using Machine Learning on the available data, can be used to generate new insights to improve customer satisfaction, adding new customers etc. The project that I will be working is posted by Airbnb in Kaggle to predict New User Bookings.

Problem Statement

New users on Airbnb can book a place to stay in 34,000+ cities, across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking and better forecast demand.

In this project I will be applying Machine Learning classification problems to predict the new user booking using the data provided here https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings.

Datasets and Inputs

Kaggle has provided below set of files as part of the competition.

1. **train_users.csv** - the training set of users
2. **test_users.csv** - the test set of users. Both the files have the below fields
    1. id: user id
    2. date_account_created: the date of account creation
    3. 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
    4. date_first_booking: date of first booking
    5. gender
    6. age
    7. signup_method

8. signup_flow: the page a user came to signup up from
9. language: international language preference
10. affiliate_channel: what kind of paid marketing
11. affiliate_provider: where the marketing is e.g. google, craigslist, other
12. first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
13. signup_app
14. first_device_type
15. first_browser
16. country_destination: this is the **target variable** you are to predict

3. **sessions.csv** - web sessions log for users, which contains the below fields
    1. user_id: to be joined with the column 'id' in users table
    2. action
    3. action_type
    4. action_detail
    5. device_type
    6. secs_elapsed
4. **countries.csv** - summary statistics of destination countries in this dataset and their locations
5. **age_gender_bkts.csv** - summary statistics of users' age group, gender, country of destination
6. **sample_submission.csv** - correct format for submitting your predictions

Solution Statement

As per the given problem statement, the model needs to predict the first location where a new user of Airbnb will book. The location will be one among the 12 countries provided. So this is a multi-class classification problem when we need to classify the first booking to one of the 12 locations.

Rather than going for a simple decision tree, I have decided to go with XGBoost. This uses boosting Ensemble Methods. In this ensemble method, a number of weak trees are generated in order to generate a strong tree.

Benchmark Model

This problem already exists as a Kaggle Competition. It is therefore possible to benchmark this submission against the already submitted entries. As of writing this proposal, there are 1462 submissions against the Kaggle Competition. I therefore propose to benchmark my model against the entry which currently stands at 700th position in the private leader board.

The scores are reproduced here

Private Leader Board

700	**Kuber@IITB**	0.86986

Public Leader Board

619	**Kuber@IITB**	0.86583

Evaluation Metrics

As the problem is taken from Kaggle, evaluation metric is already known. The evaluation metric for this competition is NDCG (Normalized discounted cumulative gain) @k where k=5.

NDCG is calculated as:

$$DCG(k) = \sum_{i=1}^{k} \frac{2^{rel(i)} - 1}{\log 2(i+1)}$$

$$\text{nDCGk} = \frac{DCGk}{DCGk}$$

where reli is the relevance of the result at position i.

IDCGk is the maximum possible (ideal) DCG for a given set of queries.

All NDCG calculations are relative values on the interval 0.0 to 1.0.

For each new user, you are to make a maximum of 5 predictions on the country of the first booking. The ground truth country is marked with relevance = 1, while the rest have relevance = 0.

For example, if for a particular user the destination is FR, then the predictions become:

- [FR] gives a NDCG=21−1log2(1+1)=1.0
- [US, FR] gives a DCG=20−1log2(1+1)+21−1log2(2+1)=11.58496=0.6309

Project Design

One-Hot Encoding and Label Encoder

From a high level structure of the provided files, it is very clear that there are many categorical columns like gender, language, affiliate_channel etc. As decision tree can not handle string categorical values, we can use One-Hot Encoding or Label Encoder to convert string input to numerical values.

cross_validation

For handling under or over fitting, Its important to do cross validation. Here I am planing to use K-fold validation technique. In cross validation, a set of data from the training data is kept aside for validating the model. The main drawback of this is that, a set data that we could have used for training is lost, due to cross validation step. To avoid these, we use K-fold validation technique were model trains K-times on the different combination train and validation dataset.

Xgboost (eXtreme Gradient Boosting)

XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. We could also do feature selection in Xgboost.