Capstone Project Jyothsna Krishnamurthy

Machine Learning Nanodegree 01-08-2017

Domain Background

This opportunity is taken from the Kaggle project

https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

The project is used to predict the destination(upto 5) a new user will make his next trip based on web session records provided.

Airbnb was founded in August 2008 and is based in San Francisco , California. The company's primary focus is to enable customers to book accommodations online or using mobile app from accommodations offered by other hosts. These accommodations can be someone's second home, extra living space etc. Airbnb offers unique accommodations that fits each customers price point. Airbnb has an international presence in about 65,000 cities and 191 countries.

A brief synopsis about the data science team in Airbnb:- Airbnb started a small data science team (with 1 person) in 2010. Initially the data science team was perceived as a team that mostly provided statistics like, How many listings we have in Paris? Or What are the top 10 destinations in Italy? While these numbers are important Airbnb recognized the importance of the sequence of events that leads up to getting those numbers. In their own words "A datum is a record of an action or event, which in most cases reflects a decision made by a person. If you can recreate the sequence of events leading up to that decision, you can learn from it; it's an indirect way of the person telling you what they like and don't like — this property is more attractive than that one, I find these features useful but those — not so much." Hence it was not just important to understand the numbers but the voices of the customers behind those numbers was a way to perceive the data.

Airbnb believes in a more proactive partnership between decision makes and the data science team. This enables a very closely knit cross collaboration between the data science team and the decision makers. Adding one more impressive line from their published article. Our distinction between good and great is impact — using insights to influence decisions and ensuring that the decisions had the intended effect. While this may seem obvious, it doesn't happen naturally — when data scientists are pressed for time, they have a tendency to toss the results of an analysis "over the wall" and then move on to the next problem.

From then on Airbnb has done a lot in the field of data science to enable them to be successful and focus more on data driven decision making. Below is a link to an article from where some of the quotes highlighted have been taken, it a great read!

https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/

The datasets used for this project are provided by Kaggle and Airbnb – link below

Link : https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

Below is a link to a white paper which uses machine learning implemented via random forest to solve a multi class – classification problem on classifying the type of disease

https://www.sciencedirect.com/science/article/pii/S2214317316300099

Problem Statement

This problem was picked from a Kaggle competition from the Kaggle website. The problem is to accurately predict where a new user will book their first booking. Airbnb has grown so much in the last few years because it has connected so many hosts to many customers looking for customized accommodations in specific prize ranges for their visit to a particular location whether the reason for travelling is vacation, business or something else. This unique experience has opened some new vistas in the travelling experience very different from the hotel options available in the past. It has also provided a competitive market and also allows hosts to make money from the facilities they have available.

My personal motivation for this problem is my passion for data , I love the fact that the company does not treat data as mere statistics or numbers. Data will tell you a story if you look into it carefully. From the article link in the domain background Airbnb has been true to this statement. Just like in everything they do, in this problem they have provided information about the web sessions of the users leading upto the decision point of choosing a particular destination. With the information provided in the sessions file I can train a model find patterns and use that trained model to accurately forecast where the customers destination will be. I look forward to using the data exploration skills and modeling skills acquired from my study in Udacity to see how I fair. I am a sql person and data analysis is not new but the interesting combination here is my newly acquired python skills along with the newly acquired modeling skills /methodology (data science modeling) to complete this project makes it challenging and interesting to me. It will amaze me at the end to see how I started out with just some data sets and tuning/predicting goals to be able to use some techniques to arrive at an optimal model.

The problem will be able to clearly make 5 predictions (1st one is the most probable one ordered by the remaining 4 less probable) of the location of the next destination of the user using what has been learned from the web sessions. From the problem statement the user can be predicted to go to one of the 12 destination countries 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found). The evaluation metric provided is NDCG (normalized discounted cumulative gain) @ k. Where k is the number of a prediction , 5 predictions are expected k takes a value from 1 thru 5. The country the person is predicted to choose first gets a NDCG = 1. The order of the remaining 4 countries and the NDCG value should be much less. The evaluation metric is discussed in detail in the evaluation section of this project proposal.

Datasets and Input

The following data sets are provided in Kaggle for this project

Link : https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

The data set consists of a list of users , their demographics and web session records, and some summary statistics. All the users from the dataset are from the USA.

The training and the test data sets are split by dates.

- train_users.csv - the training set of users
- test_users.csv - the test set of users
    - id: user id
    - date_account_created: the date of account creation
    - 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
    - date_first_booking: date of first booking
    - gender
    - age
    - signup_method
    - signup_flow: the page a user came to signup up from
    - language: international language preference
    - affiliate_channel: what kind of paid marketing
    - affiliate_provider: where the marketing is e.g. google, craigslist, other
    - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
    - signup_app
    - first_device_type
    - first_browser
    - country_destination: this is the target variable you are to predict
- sessions.csv - web sessions log for users
    - user_id: to be joined with the column 'id' in users table
    - action
    - action_type
    - action_detail
    - device_type
    - secs_elapsed
    -
- countries.csv - summary statistics of destination countries in this dataset and their locations
- age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination
- sample_submission.csv - correct format for submitting predictions.

Based on the initial data analysis the number of records in given for training are - 213451,
The classes distributed based on the target variable are as below

```
country_destination
```

```
NDF      124543
US        62376
other     10094
FR         5023
IT         2835
GB         2324
ES         2249
CA         1428
DE         1061
NL          762
AU          539
PT          217
```

The thought process is to split the data using Stratified Kfold to preserve the classes. Stratified Kfold splits the data into folds by preserving the percentage of classes in the fold, 1 fold is used as set each time.

Solution Statement

Below is the outline of the solution :

To arrive at a good solution below are some of the steps I will implement

1. Explore the data to understand statistics
2. Perform some visualizations – numerical features may have some correlations
3. Observations on missing data
4. Separate the features and labels
5. Pre process the data
   a. Clean null values – impute using mean, median or mode (depends on data)
   b. Standardize domain values– example Gender column may have “Male”, “M” etc
   c. Outlier Detection
   d. Encoding categorical data - numerical encoding or one hot encoding
   e. Feature scaling
   f. Feature importance – I might try PCA on the data to see which feature offers the most variance in the data.

6. Then will use stratified fold to split the data into training and test to get the accuracy using one or more of the below models
   a. Multinomial logistic Regression
   b. KNN
   c. Will also try to use some ensemble models

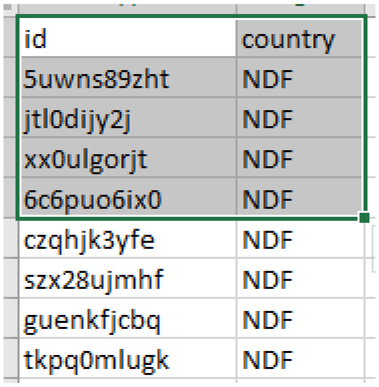
7. After comparing the results with the benchmark model will also run this on the test data set provided by Kaggle to see if they will score it.

The reason for trying to use one of the above classifiers is because given the data set we have to predict 5 places with the first being the most likely the user will choose to travel based on the characteristics of the web session data. Hence I think MLR , KNN can be used to predict this label.

Benchmark Model

To Benchmark the results , I plan to train a random forest model and get the accuracy results using that. Random forest does not require the feature normalization.

In Kaggle there is a submission page where even after the competition is complete if I submit my results in the submission format below , my submission will be scored.



id	country
5uwns89zht	NDF
jtl0dijy2j	NDF
xx0ulgorjt	NDF
6c6puo6ix0	NDF
czqhjk3yfe	NDF
szx28ujmhf	NDF
guenkfjcbq	NDF
tkpq0mlugk	NDF

One more thought is from discussions on Kaggle I have noticed that folks split the training data into training, validation and test. They use that test to benchmark. My only hesitation with this I am loosing the learning from the training data if I split it smaller.
This is my first attempt at Kaggle so I hope if I submit the submission in the format they provided my results will be scored.

Evaluation Metric

The airBnb has provided an evaluation metric for this project.

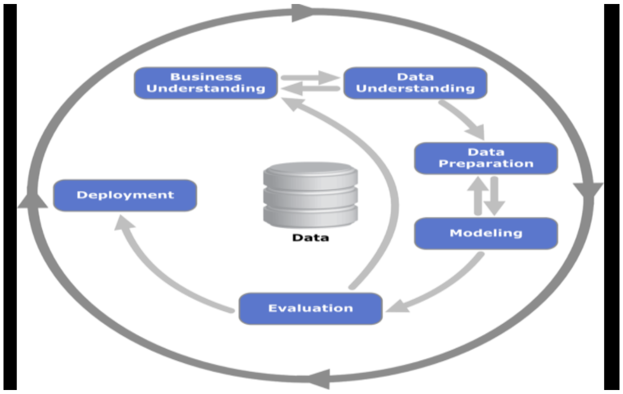
https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings#evaluation

AirBnb uses the NDCG (Normalized discounted cumulative gain) method to evaluate the predictions.

Project Design

I am going to try to follow the CRISP DM methodology for data mining. CRISP- DM stands for **C**ross **I**ndustry **S**tandard **P**rocess for Data Mining. The process has 6 steps

Picture taken from Wikipedia.org



Step 1 – Business Understanding

Work with business teams to understand the true nature of the problem. This step seems easy but comes with a lot of ambiguity and assumptions because the human factor is involved. You have to make every attempt to reach to the core business problem or metric that needs to optimized. Business users can sometimes be only explaining symptoms of a problem and not the whole problem. Some business users may only relay their perspective of the problem. It is the role of the data scientist tram to look past that and from conversations with several business users exactly identify or point out what needs to be solved. This may further lead or spawn into more projects trying to optimize several metrics. The art is to arrive at those individual metric focused goal and get it validated by business users.

Step 2 - Data Understanding

Perform data exploration to understand the attributes and data.

Step 3 -Data Preparation

This step can be generalized as Data Cleaning, Data Transformation and Data Reduction.

Data Cleaning

From the data exploration I may find data that have

<u>Missing Values</u>

These can be handled by a few ways

1. Ignore the tuple if the class label is missing
2. Use the attribute mean (or majority nominal value) to fill missing data
3. Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
4. Predict the missing value using a learning algorithm

<u>Identify outliers and smooth out noisy data</u>

Some strategies that can be employed are

1. Binning
2. Clustering
3. Regression

Correct Incorrect Data

Use domain knowledge where applicable documenting assumptions

Data Transformation

Possibly apply some of the below

1. Normalization
2. Aggregation
3. Generalization (less likely to use)
4. Attribute construction (less likely to use)

Data Reduction

Employ techniques like

1. Reducing the number of irrelevant attributes
2. Reducing attribute values by binning
3. Reducing number of row (less likely to use)

Step 4 – Modeling

Apply the chosen model technique or several models to the pre-processed data. The output of this will be the predictions of the countries (first 5 , 1 being the ground truth) the user is expected to choose as his destination for travel. Create the submission file as shown in the Benchmark Model section.

Step 5 – Evaluation

The file generated from the predictions will be uploaded to the leaderboard for evaluation. The process is iterative. If the results do not meet the expectations will have to go thru the Step 3 to Step 4 again applying other modeling techniques.