Capstone Project Proposal

Domain Background

Using data to aid decision making has become an integral part of many companies. Internet companies have huge amount of records of users' behavior on their website or mobile app. Series records of actions represent the decisions of customers and could tell a lot about how they like the product. Leveraging this sort of data could give the company a competitive advantage. Information extracted from data could be used in decision making. This project will be a great practice to get familiar with real problem a Data Scientist deal with in a business setting.

Problem Statement

The Capstone Project I propose to work on is the Airbnb New User Bookings Kaggle competition (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings). The problem is to predict the destination country of a new Airbnb user's first booking. Available data are demographics of existing users, records of users' web sessions as well as some basic summary statics.

This problem is a multi-class classification problem. In the end, the result from my analysis and modeling will an ordered list of 5 predictions for each user in the testing data set. Evaluation of the result is going to be the Normalized discounted cumulative gain @ k = 5 (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings#evaluation) or the public leaderboard score on Kaggle website. This problem is highly reproducible since Airbnb provides booking services to existing and new users as long as their business exists and they could always run models developed in this project on their growing database.

Datasets and Inputs

Datasets to this project is provided on Kaggle website (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data). The file descriptions provide detailed information of the dataset. Here I will briefly discuss the characteristics of the datasets and how they are going to be used in the project.

Training users' dataset contains 213,451 rows and 16 columns of data. Each row contains information of a customer: date when account was created, time stamp when the customer is first active (which could be earlier than date_account_created), date of first booking if any, gender, age, sign-up method, sign-up flow (the page a user came to sign-up from), sign-up app, language preference, affiliate channel (what kind of paid marketing), affiliate provider (where the marketing is), first affiliate tracked, first device type and first web browser. Target variable is the country of first booking destination has 12 possible outcomes: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL', 'DE', 'AU', 'NDF' (no destination found) and 'other'. Difference between 'NDF' and 'other' is that 'NDF' means no booking, while 'other' means there was a booking but the country is not included in the list.

Sessions dataset contains 10,567,737 rows and 6 columns of data. Each row contains the action, action type, action detail, device type as well as the seconds elapsed since the previous action for each user's action. The sessions dataset contains data only as early as 1/1/2014, while users' dataset dates back to 2010.

Countries dataset containing the latitude and longitude of the 10 destination countries, as well as their distance from the US and the area of that country. Age and gender dataset contains the distribution of age and gender grouped by each destination country.

Training users' dataset, sessions dataset, countries dataset as well as the age and gender distribution dataset will be used to learn classification model to predict the target variable.

Solution Statement

I am going to apply various classification algorithms to this problem. According to the discussions I read on the Kaggle website, effective models for this problem are: logistic regression, tree-based algorithms (gradient boosting trees), as well as deep learning models (Multilayer Perceptron).

Benchmark model

A benchmark model for this project is one script posted by Sandro Vega Pons who won the third place in the competition (https://www.kaggle.com/svpons/script-0-8655) where he applied simple feature engineering and XGBoost model.

For benchmarking with ensemble models stacking various single model, another script from him (https://www.kaggle.com/svpons/three-level-classification-architecture) could also be used as benchmark.

Evaluation Metrics

The metric used for evaluation is NDCG (Normalized discounted cumulative gain) @ k = 5. Mathematical formula is detailed on the Kaggle website (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings#evaluation).

Project Design

In this project, I'll first explore the data sets and use visualization to build intuition of the datasets. For example, I'll compare the distributions of the categorical features between training and testing data.

I'll also add derived features based on the existing ones. For example, I will add the time elapsed between the time stamp of first active of a user to the date when account is created. I believe this kind of new features provide interpretation of raw data and help models to make better predictions.