NYC check-ins Rich Dataset

- BY AASHISH DUBEY

8 JUNE 2020

Introduction

This dataset contains check-ins in NYC collected for about 10 month (from 12 April 2012 to 16 February 2013). It contains 227,428 check-ins in New York city. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories). This dataset is originally used for studying the spatial-temporal regularity of user activity in LBSNs. This data is issued by Foursquare, you can find it here.

Business Problem

Most of the marking and advertising companies are looking for a mass audience which fits according to their product.

For example, if a company is advertising about a trip, it would be better to target an audience who are looking for such kind of thing. Advertising for such kind of things at beaches or places where people come to relax has a great chance of getting a conversion, and boosting the sales. But very few of them are practicing an approach like this, which makes me to drive a solution.

Business Solution

We can use this data and analyse it to see the insights that on what day or at what time, majority of people like to visit and this could be beneficial for the people who are running advert campaigns. The insights from this data will help them to target a mass audience, as per the requirement of their campaign.

For example, they can decide a launch plan of some new appetizer in a restaurant where people visit more frequently, or a brand can reach to a potential business partner as per their requirement.

Data

This dataset includes long-term (about 10 months) check-in data in New York city and Tokyo collected from Foursquare from 12 April 2012 to 16 February 2013.
It contains two files in csv format. Each file contains 8 columns, which are:

1. User ID (anonymized)
2. Venue ID (Foursquare)
3. Venue category ID (Foursquare)
4. Venue category name (Fousquare)
5. Latitude

6. Longitude
7. Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
8. UTC time

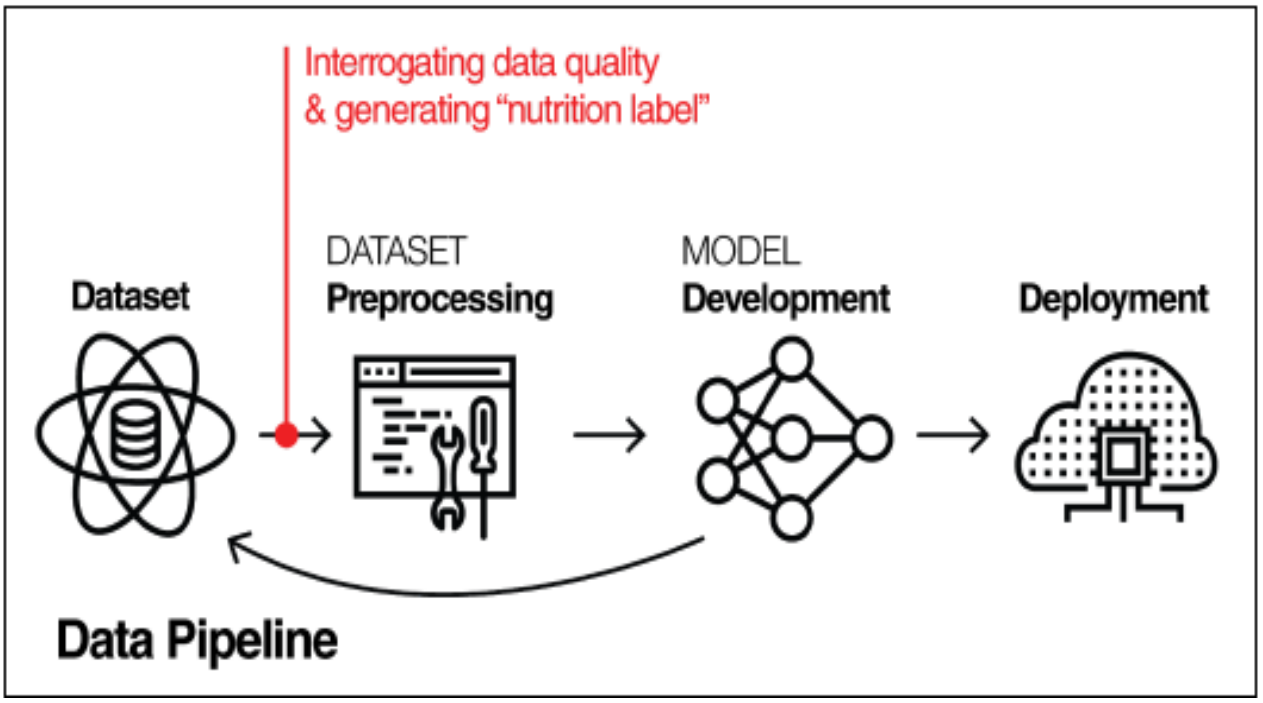
The file dataset*TSMC2014*NYC.txt contains 227428 check-ins in New York city

This dataset is acquired from here

Following is the citation of the dataset author's paper:

Dingqi Yang, Daqing Zhang, Vincent W. Zheng, Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Trans. on Systems, Man, and Cybernetics: Systems, (TSMC), 45(1), 129-142, 2015

Methodology



Data driven decision making systems play an increasingly important and impactful role in our lives. These frameworks are built on increasingly sophisticated artificial intelligence (AI) systems and are tuned by a growing population of data specialists3 to infer a vast diversity of outcomes: the song that plays next on your playlist, the type of advertisement you are most likely to see, or whether you qualify for a mortgage and at what rate. These systems deliver untold societal and economic benefits, but they can also pose harm. Researchers continue to uncover troubling consequences of these systems. Data is a fundamental ingredient in AI, and the quality of a dataset used to build a model will directly influence the outcomes it produces.

Like the fruit of a poisoned tree, an AI model trained on problematic or missing data will likely produce problematic outcomes. Examples of these problems include gender bias in language translations surfaced through natural language processing, and skin shade bias in facial recognition systems due to non-representative data. Typically the model development pipeline begins with a question or goal. Within the realm of supervised learning, for example, a data specialist will curate a labeled dataset of previous answers in response to the guiding question. Such data is then used to train a model to respond in a way that accurately correlates with past occurrences. In this way, past answers are used to forecast the future. This is particularly problematic when outcomes of past events are contaminated with (often unintentional) bias.

Result

Location based social networks have attracted millions of users and massively contains their digital footprints. We have crawled a part of these digital footprints from Foursquare in order to study the problems of personalized location recommendation and search. This dataset includes check-in, tip and tag data of restaurant venues in NYC collected from Foursquare from 24 October 2011 to 20 February 2012. It contains 3112 users and 3298 venues with 27149 check-ins and 10377 tips (written in English).
The proposed datasheet includes dataset provenance, key characteristics, relevant regulations and test results, but also significant yet more subjective information such as potential bias, strengths and weaknesses of the dataset, API, or model, and suggested uses. As domain experts, dataset, API, and model creators would be responsible for creating the datasheets, not end users or other parties. We are also aware of a forthcoming benefits that cannot be ignored by a business firms and advertising industries, as this data plays a very important role in deciding our customer base.

Conclusion

Large Scale Enterprises are rapidly adopting machine learning for driving their business in several ways. Automation of several tasks is one of the key **future** goals of the industries. As a result, they are able to prevent losses from taking place. With the rise of artificial intelligence (AI) and machine learning (ML), organizations are demanding faster insights to remain competitive. Remarkably, the same technology **advancements** that drive this urgency are also the key to unlocking better efficiency in **data science** work. Most technology research firms are tracking the self-serve data analytics trends. These trends are making it possible for the average enterprise and the average business user to leverage sophisticated analytics, algorithms and techniques without the skills of a data scientist. The solutions are easy to use and allow the average user to build on their core business skills and see data and make decisions in a meaningful way. Think of it as data democratization. These new solutions for advanced analytics and augmented analytics, self-serve data preparation, smart visualization and assisted predictive modeling will allow data scientists to focus on strategic initiatives and business users to leverage sophisticated tools without a lot of training so the enterprise will get rapid ROI and low TCO. Does Data Democratization Result in Data Anarchy and Bad Business Decisions? This article will helps us to understand how data literacy can help the enterprise and the business users and how these advancements in data analytics will bring data democracy to the organization.