A survey of 3 Capstone Projects

A BRIEF OVERVIEW OF PROJECTS

K. Rajesh Jagannath |Springboard Career Track Data Science| 11/9/2017

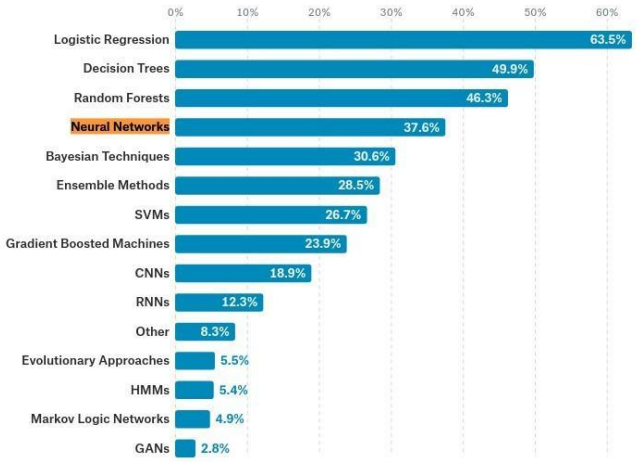
Introduction

This document is for Spring Data Science Career track Capstone 1 project. It outlines 3 possible projects of which one will be chosen.

CRITERIA FOR SELECTION

To begin with, I had a few criteria to select my capstone project.

- Strike a balance between simplicity (implementation) and complexity (advanced ML methods) enough to show-case in my portfolio of projects.
- It needed to be a Kaggle dataset – I do not have one to show-case. Data Validation, Cleaning and Preparation takes time and a lot of work with data set providers which are often deemed proprietary. So, a data set from a data aggregator would be preferable.
- Business understanding takes time and a lot of work. I want to deepen my understanding of one of the top ML techniques used by data scientists as per the Kaggle Survey “Kaggle's "The State of Data Science and Machine Learning" 2017



PROJECT 1: INSTACART MARKET BASKET ANALYSIS

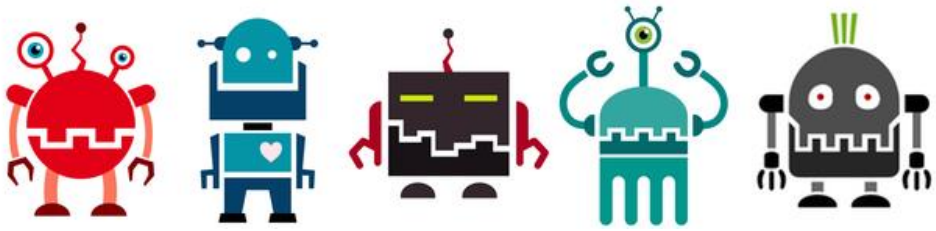
Website : https://www.kaggle.com/c/instacart-market-basket-analysis

Instacart is one of the top online grocery delivery services. They have open sourced 3 million orders data-set that has been anonymized. The goal of this project would be to use data on customer orders over time to predict which previously purchased products will be in a user's next order.

Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

PROJECT 2: FACEBOOK BOT OR HUMAN CHALLENGE

This project is particularly appealing as it a anomaly detection classification project that as a variety of applications in different fields and not specific to a certain industry such as Retail. Human bidders on the site are becoming increasingly frustrated with their inability to win auctions vs. their software-controlled counterparts. As a result, usage from the site's core customer base is plummeting.



In order to rebuild customer happiness, the site owners need to eliminate computer generated bidding from their auctions. Their attempt at building a model to identify these bids using behavioral data, including bid frequency over short periods of time, has proven insufficient.

The goal of this competition is to identify online auction bids that are placed by "robots", helping the site owners easily flag these users for removal from their site to prevent unfair auction activity.

The data in this competition comes from an online platform, not from Facebook

PROJECT3: FACEBOOK CHECKINS PREDICTION

Website : https://www.kaggle.com/c/facebook-v-predicting-check-ins

.

The goal is to predict which place a person would like to check in to. For the purposes of this project an artificial world was created consisting of more than 100,000 places located in a 10 km by 10 km square. For a given set of coordinates, the task is to return a ranked list of the most likely places. Data was fabricated to resemble location signals coming from mobile devices, giving one a flavor of what it takes to work with real data complicated by inaccurate and noisy values. Inconsistent and erroneous location data can disrupt experience for services like Facebook Check In.

Data Wrangling : High effort