A survey of 3 Capstone Projects

A BRIEF OVERVIEW OF PROJECTS

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

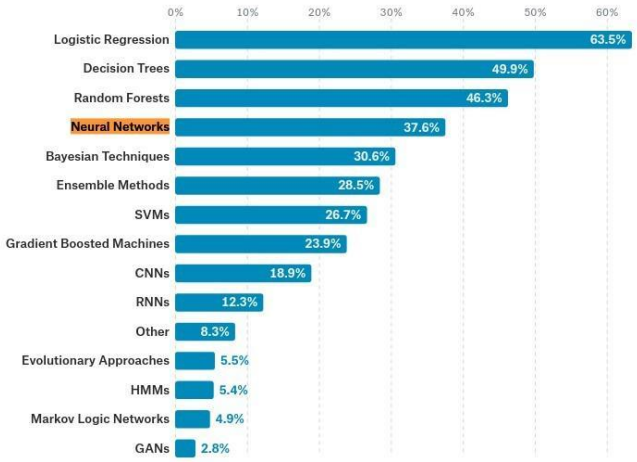
Introduction

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.

1. What is the problem you want to solve?
2. Predict based on use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order
3. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?
    a. Client would be Instacart. The client will implement a recommendation system to display in their website as the orders are filled
4. What data are you going to use for this? How will you acquire this data?
    a. The data is available on Kaggle

The dataset for this competition is a relational set of files describing customers' orders over time. The goal of the competition is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders. For more information, see the blog post accompanying its public release.

File descriptions

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Most of the files and variable names should be self-explanatory.

aisles.csv

```
aisle_id,aisle
1,prepared soups salads
2,specialty cheeses
3,energy granola bars
...
```

departments.csv

```
department_id,department
1,frozen
2,other
3,bakery
...
```

order_products__*.csv

These files specify which products were purchased in each order. order_products__prior.csv contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product. Note that some orders will have no reordered items. You may predict an explicit 'None' value for orders with no reordered items. See the evaluation page for full details.

```
order_id,product_id,add_to_cart_order,reordered
1,49302,1,1
```

```
1,11109,2,1
1,10246,3,0
...
```

orders.csv

This file tells to which set (prior, train, test) an order belongs. You are predicting reordered items only for the test set orders. 'order_dow' is the day of week.

```
order_id,user_id,eval_set,order_number,order_dow,order_hour_of_day,days
_since_prior_order
2539329,1,prior,1,2,08,
2398795,1,prior,2,3,07,15.0
473747,1,prior,3,3,12,21.0
...
```

products.csv

```
product_id,product_name,aisle_id,department_id
1,Chocolate Sandwich Cookies,61,19
2,All-Seasons Salt,104,13
3,Robust Golden Unsweetened Oolong Tea,94,7
...
```

sample_submission.csv

```
order_id,products
17,39276
34,39276
137,39276
...
```

5. In brief, outline your approach to solving this problem (knowing that this might change later).
    a. I will follow the approach outlined in http://blog.kaggle.com/2017/09/21/instacart-market-basket-analysis-winners-interview-2nd-place-kazuki-onodera/
    b. Predict Re-orders : which previously purchased products will be in the next order? Depends on both the user and the product

c. Predict None – will the user's next order contain previously purchased product ? This depends on the user alone
d. Find probabilities from the above 2-steps.
e. Use this probability to predict if User A will re-purchase in the next prder

6. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.
    a. Code in python
    b. A slide deck and
    c. A paper

PROJECT 2 : MACHINE LEARNING REPOSITORY : DIABETES 130-US HOSPITALS FOR YEARS 1999-2008

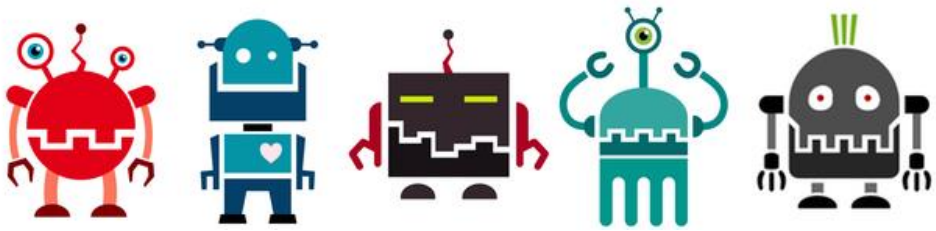
Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. The Project will explore the factors that will improve the outcome of in-patient care.

Website : https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

1. What is the problem you want to solve?
    a. Explore the link between diabetes and hospital re-admission and mortality
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?
    a. The client would be any Health insurance or Hospital that would be interested in increasing the outcome of in-patient care (prevent re-admission, mortality) and decrease costs (for the hospital).
3. What data are you going to use for this? How will you acquire this data?
    a. The data is available in UCI Machine Learning Repository
    b. Data Format is described in https://www.hindawi.com/journals/bmri/2014/781670/tab1/
4. In brief, outline your approach to solving this problem (knowing that this might change later).
    a. Multivariable logistic regression will be used to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates such as demographics, severity and type of the disease, and type of admission.
    b. Reference : https://www.hindawi.com/journals/bmri/2014/781670/
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.
    a. Code in python
    b. A slide deck and
    c. A paper

PROJECT 3: 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