GA Data Science Final Project Ideas

Varun Dodla

Problem 1

Production Error to Team Assignment

During the release of our application to production, engineers are typically tasked with looking at the logs to diagnose errors and assign them to teams based on the error text (stacktrace). This could be automated using machine learning.

Problem 1
Data

We have historical data available in our database which shows errors tackled in previous releases over the last six months or more. Typical data format is as below:

Id	Source Class	Source Method	Stacktrace	TeamId

Most of the data is categorical and should be easy to translate to team but at the same time it is not a simple mapping problem. Also since data is text esp. the stacktrace column might need some preprocessing.

Problem 1
Hypothesis

While looking at stacktraces and predicting the team looks like a simple problem it is more involved as the stacktrace could be a combination of underlying errors and we have to look at all of them to make a meaningful prediction. That said some pre-processing should help us classify all the errors to an existing team. Also it might be tricky if a new error shows up for a team that historically never had errors assigned to it so we might have to have something like an Other/ Unknown category.

Problem 2
Predict Blood Donations

Using data from a blood donation drive vehicle which drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus.

(**Source**: https://www.drivendata.org/competitions/2/warm-up-predict-blood-donations/)

Problem 2
Data

The data for this problem is courtesy of Yeh, I-Cheng via the UCI Machine Learning repository. See format below:

Months Since Last Donation	Number of donations	Total Volume Donated (c.c.)	Months since First Donation	Made Donation in March 2007

All of the data is numerical and goal is to predict the last column as a probability.

Problem 2

Hypothesis

The factors given in the data may not be able to help us predict the chance for a donor blood donation next time around with 100% probability but should get us close. The assumptions being the donors are actually present when the vehicle drives by next time around.

Problem 3

Predict Occupancy Level of Belgian Trains

Trains can get really crowded sometimes, so wouldn't it be great to know in advance how busy your train will be, so you can take an earlier or later one?

(**Source**: https://inclass.kaggle.com/c/train-occupancy-prediction)

Problem 3
Data

This data is provided by irail. Data is provided as json text with separate training and test data.


```
{
 "querytype": "occupancy",
 "querytime": "2016-09-29T16:24:43+02:00",
 "post": {
        "connection": "http://irail.be/connections/008811601/20160929/S85666 ",
         "from": "http://irail.be/stations/NMBS/008811601 ",
         "to": "http://irail.be/stations/NMBS/008811676 ",
         "date": "20160929",
         "vehicle": "http://irail.be/vehicle/S85666 ",
         "occupancy": "http://api.irail.be/terms/medium"
 },
 "user_agent": "Railer/1610 CFNetwork/808.0.2 Darwin/16.0.0"
}
```


Data requires preprocessing to extract ids and such before we can start modelling on it.

Problem 3
Hypothesis

Train occupancy can depend on various factors that just historic data like the weather, holidays, etc. We should most likely augment the given data with data on public holidays and weather as an exploration. Assuming we have all that incorporated into our model we might be able to predict the occupancy levels with a good level of accuracy.