Machine Learning Engineer Nanodegree

Capstone Proposal

Tichakunda Mangono
September 4th , 2017

Proposal

Domain Background

More than **36.7 million**[1] people in the world were living with HIV in 2016 and every year, about **1 million** people worldwide die from AIDS-related causes. While this death rate has decreased significantly (by 38%) since 2001 and continues to decline, about **1.8 million** people became newly infected in 2016 alone. The epidemic disproportionately affects low income countries in Eastern and Southern Africa and women, adolescents and key populations like female sex workers and LGBTQ individuals are the most affected groups. There is currently no cure or vaccine for HIV and while several prevention methods exist, their efficacy is reduced by several factors, including economic and psycho-social factors. Fortunately, it has been shown that treatment can not only prolong life but also prevent the spread of HIV as it lowers the viral load of people living with HIV to a non-infectious level. However, of the **36.7 million** people living with HIV in 2016, only **19.5 million** were receiving this life-saving treatment.

The largest buyer of drugs and testing and laboratory kits for HIV treatment is The President's Emergency Plan for AIDS Relief (PEPFAR), a US government program which spends about **$9.5 Billion** per year on procurement of essential medicines to fight HIV/AIDS around the world. It is very important that these procurements arrive on time and in full to meet the needs of People Living with HIV (PLHIV) around the world. Thus, it is important to know whether or not HIV drugs are delivered on time and if not, what is the extent of the delay. This study will use publicly available data from PEPFAR over the years 2006-2015 to determine the factors influencing timeliness of pharmaceutical deliveries as well as use these factors to develop a model that can predict if and by how long a particular HIV commodity will be delayed in delivery. While more and more supply chain analysis has begun to incorporate machine learning it is especially aimed at demand forecasting as opposed to predicting the lead-time directly. However, the approaches taken in some academic studies[2] e.g. SVMs and RNNs can be

[1] http://www.unaids.org/en/resources/fact-sheet
[2] https://www.researchgate.net/publication/222928270_Application_of_machine_learning_techniques_for_supply_chain_demand_forecasting

adopted for this study as well. Similar problems like predicting flight delays[3] and improving flight efficiency have also been solved using machine-learning.

I am personally invested in this work as because I come from Zimbabwe, a country that has a 15% HIV prevalence rate and where tens of thousands of people die of AIDS-related illnesses every year. Timeliness of HIV procurement is critical to the efficiency and impact of the program in saving lives, controlling and eventually eliminating HIV.

Problem Statement

Delays in supply of commodities result in extra costs in terms of storage, coordination and most importantly lost lives in the case of HIV medicines. This study will use publicly available supply chain data to determine the most important factors in predicting whether HIV drugs are delivered on time or not. It will then use these factors to predict how long these delays will be, thus allowing HIV program managers to know **when and which products are likely to be delayed**, as well as **the extent of the delay** so they can take mitigating action to save lives and avoid additional supply chain costs.

Datasets and Inputs

This study will use data from The President's Emergency Plan for AIDS Relief (PEPFAR) program's Supply Chain Management System (SCMS) data made publicly available online through the website: https://data.pepfar.net/additionalData . This is a very detailed dataset with over 10,000 observations of unique HIV medicines/products with **33 feature columns** of product details, country, manufacturer and shipment details including - order, purchase and delivery dates. Two additional columns for **the target variables** ("on-time" a binary viable and "delay", a continuous variables will be derived from the existing date-time columns). While the data has some limitations where products are sometimes consolidated into large shipments to save on costs, the availability of anticipated delivery and actual delivery dates makes this appropriate for this study. See appendix for list of dataset features/inputs.

Solution Statement

As a proposed solution, this study will first explore **classification machine learning algorithms** to determine whether a particular product was delayed or not. It will then use **regression analysis** to predict the length of the delay using the subset of the data which the classification predicts will be delayed. This will maximize the utility of the complete model since it follows the natural decision-making process – a program manager would normally care about the products that will be delivered late and within those, focus on the ones that will likely have the longest

[3] https://www.kaggle.com/c/flight

delays first, thus allowing them to prioritize supply chain/logistics management and solve the biggest problems first.

To select best model, both the classification and regression versions of these models will be explored evaluated against the benchmarks (see **"Benchmarks"** section below): i) Random-Forests ii) XGBoost iii) Support-Vector Machines (SVM) and iv) Recurrent Neural Networks (RNNs). Random-Forests and XGBoost are proven **high-performing ensemble** algorithms which can do **automatic feature extraction** while SVMs perform very well with **high-dimensional data** and can **detect non-linear relationships** if the right kernel is used. Finally, RNNs are useful for high-dimensional time-series data. Please see **"Project Design"** section for workflow and overview of algorithms. The above advantages of these algorithms are well-suited to the selected dataset which has several categorical columns which will increase dimensionality and potentially be non-linearly related to the target variable after data transformation. Finally, the data is well-suited for this overall approach since our target variables is well-defined on the data i.e. it can be determined by data on scheduled versus actual delivery dates, thus allowing us to quantify and measure the problem and solution. The study results will be applicable to future instances of supply chain orders, and thus it is applicable to future occurrences of similar supply chain data observations.

Benchmark Model

Since the proposed solution model is combination of two algorithms working together sequentially, the benchmark model will also require a two-part benchmark. In order to make clear objective comparisons, the same model, **Random Forest will be used as the benchmark for both classification and regression**. The study will use the default versions of the Scikit-Learn implementation of these models.

Evaluation Metrics

This model will be evaluated based on 3 metrics: **F1-Score** to balance our recall and precision, especially because the dataset is unbalanced with a ratio of **1:8** between the positive and negative class respectively. For the regression part of the model, the **R-squared** and **root mean-squared deviation (RMSD)** will be used to evaluate how well the regression model can predict the direction and length of delays in HIV medicine deliveries.

i) **F1-Score** is an average of the recall and precision scores.

F1-Score = 2*(Recall * Precision) / (Recall + Precision), *where* ***Precision = True Positives/ (True Positives + False Positive)*** *and* ***Recall = True Positives / (True Positives + False Negatives)***

ii) **R-squared** is the "coefficient of determination" which measures the amount of variation in the data that is explained by the model, again as a percentage/fraction of total variation.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Here, r represents R-squared, n is the number of observations and x and y are the feature and target variables respectively.

iii) **RMSD** measures the average size (absolute value) of the error that the model makes when predicting continuous target variables e.g. days late/delay in this case.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}.$$

Here, "y-hat" are the predicted values and "y*i*" are true values of the target variable. "n" is the number of observations in the dataset.

Project Design

The project workflow will consist of 6 distinct stages described below:

1. **Load Data**
   - Identify and use the relevant libraries, software
   - Export load and transform the data into usable format.
2. **Data Cleaning/Transformation**
   - Understand the data ; deal with the date columns and conversions
   - Account for legacy issues and changing contexts within the data
   - Deal with missing values and dates not captured
3. **Feature-Engineering**
   - Dates: engineer the target date variables and features (year, month, day)
   - Manufacturer/factory address, country and continent – use googlemaps API
   - Distance between destinations: can be a good predictive feature
   - Freight-cost and weight: untangle complex, consolidated shipment data
4. **Exploratory Data Analysis**
   - Figure out any pairwise associations or variable distributions
   - Charts and visualizations, correlation factors and factor plots
5. **Develop the Predictive ML Models**
   - Model selection using Train-Test Split and/or cross-validation evaluation methods
     - *Classification candidate models* – Logistic Regression , Random-Forest classification, Support-Vector Machine (SVM) classification, and Recurrent Neural Network (RNN)
     - *Regression candidate model*s – Linear Regression, Random-Forest Regression, SVM, and RNN
   - Model fine-tuning/Improvement of the best model GridSearch/Cross-Validation and tuning Hyper-parameters
6. Document **results** formulate **conclusions** and outline the **implications**

Table 1: Available Input Features in the PEPFAR Supply Chain Data Set

#	FieldName	FieldDescription	DataType
1	ID	Primary key indentifer of the line of data in our analytical tool	Number
2	Project Code	Project code	Text
3	PQ #	Price quote (PQ) number	Text
4	PO #	Order number: Purchase order (PO) for Direct Drop deliveries, or Sales Order (SO) for from Regional Delivery Center (RDC) deliveries	Text
5	ASN/DN #	Shipment number: Advanced Shipment Note (ASN) for Direct Drop deliveries, or Delivery Note (DN) from RDC	Text
6	Country	Destination country	Text
7	Managed By	SCMS managing office: either the Program Management Office (PMO) in the U.S. or the relevant SCMS field office	Text
8	Fulfill Via	Method through which the shipment was fulfilled: via Direct Drop from vendor or from stock available in the RDCs	Text
9	Vendor INCO Term	The vendor INCO term (also known as International Commercial Terms) for Direct Drop deliveries	Text
10	Shipment Mode	Method by which commodities are shipped	Text
11	PQ First Sent to Client Date	Date the PQ is first sent to the client	Date/Time
12	PO Sent to Vendor Date	Date the PO is first sent to the vendor	Date/Time
13	Scheduled Delivery Date	Current anticipated delivery date	Date/Time
14	Delivered to Client Date	Date of delivery to client	Date/Time
15	Delivery Recorded Date	Date on which delivery to client was recorded in SCMS information systems	Date/Time
16	Product Group	Product group for item, i.e. ARV, HRDT	Text
17	Sub Classification	Identifies relevant product sub classifications, such as whether ARVs are pediatric or adult, whether a malaria product is an artemisinin-based combination therapy (ACT), etc.	Text
18	Vendor	Vendor name	Text
19	Item Description	Product name and formulation from Partnership for Supply Chain Management (PFSCM) Item Master	Text
20	Molecule/Test Type	Active drug(s) or test kit type	Text
21	Brand	Generic or branded name for the item	Text
22	Dosage	Item dosage and unit	Text
23	Dosage Form	Dosage form for the item (tablet, oral solution, injection, etc.).	Text
24	Unit of Measure (Per Pack)	Pack quantity (pills or test kits) used to compute unit price	Number
25	Line Item Quantity	Total quantity (packs) of commodity per line item	Number
26	Line Item Value	Total value of commodity per line item	Currency (USD)
27	Pack Price	Cost per pack (i.e. month's supply of ARVs, pack of 60 test kits)	Currency (USD)
28	Unit Price	Cost per pill (for drugs) or per test (for test kits)	Currency (USD)
29	Manufacturing Site	Identifies manufacturing site for the line item for direct drop and from RDC deliveries	Text
30	First Line Designation	Designates if the line in question shows the aggregated freight costs and weight associated with all items on the ASN/DN	Binary
31	Weight (Kilograms)	Weight for all lines on an ASN/DN	Number
32	Freight Cost (USD)	Freight charges associated with all lines on the respective ASN/DN	Currency (USD)
33	Line Item Insurance (USD)	Line item cost of insurance, created by applying an annual flat rate (%) to commodity cost	Currency (USD)