IIT Bombay 2020 – 2021
Department : Mechanical Engineering
Course: ME 714 - Computer Int Mfg
Instructor: Prof. Soham Mujumdar

Supply Chain Management

19th April, 2021

Abstract

In this project, we aimed at solving 2 parts of a Supply Chain Management- the Economic and Delay Considerations. In the economic considerations part of the report, vaccine transport logistics were modelled using the Fixed Costs Capacitated Facility Location formulation. This was solved in Python with the help of the Pyomo Library and the GNU Linear Programming Kit. In the second part of this project, we have used several Machine Learning techniques in order to detect delayed deliveries of HIV medicines. This is a combined classification-regression problem. We have compared models such as Linear SVC, Random Forest,Multi-Layer Perceptron, Extra Trees ensemble using parameters such as RMSE, F1 score,R - Squared. Sklearn library in Python has been used for this purpose.

Authors:
Shardul Burde-18D100018
Vikrant Rangnekar-18D100023
Anshul Gupta-18D100005
Atharv Toraskar-180110089
Mitesh Pawar-18D100012

Contents

1 Introduction

Beginning from 1765 through the present day, we have seen a remarkable evolution. As we discovered different energy sources and, later, digital technologies, the entire landscape of the modern world has been transformed repeatedly. The step into production technology, which was completely different from the past, is called industrial revolution.

Industry 4.0 is the current trend of automation and data exchange in manufacturing technologies. Large-scale machine-to-machine communication (M2M) and the internet of things (IoT) are integrated for increased automation, improved communication and self-monitoring, and the production of intelligent machines that can analyze and diagnose issues without the need for human intervention. Industry 4.0 creates a “smart factory”.

1.1 Supply Chains and Product Consumption

Generally, the physical products we use in our personal and professional lives have a long journey. Before they reach their final destination, they are required to go through the supply chain process. The supply chain process is the flow of materials and information currency and service from the raw material suppliers through factories and warehouses to the end consumer. Supply Chain Management is the management of the flow of materials in the supply chain process. Supply chain management is very critical, especially in e-commerce, because when it is done well, the customer doesn’t notice that the supply chain process has occurred. They simply order the product online, and it arrives at their door in a short amount of time in perfect condition.

2 Economic Considerations of Supply Change Management

2.1 Introduction

One of the critical issues that borders on a vaccination drive in developing countries is centered on the efficient transport of vaccines from manufacturing plants. For an effective system, locating the collection points is very crucial. Governments and other stakeholders have made several investments but with limited results in combating the challenges associated with effective transportation. Thus, this review study focuses on the well-researched facility location problem (FLP).

FLP deals with the question of how to select from a given set of potential locations, a cost-effective subset of sites to place new facilities or retain existing ones. A facility could represent any service facility such as an electric power plant, hospital, food production plant, a warehouse (depot), petrol station, government office, etc. FLPs are an essential class of problems in logistic management. Facility location and the assignment of entities to such facilities usually determine the distribution pattern and the associated characteristics (e.g., time, cost, and efficiency) of the facility. The placement of one or more facilities and the assignment of customers in an optimal version not only improve the flow of materials and services offered by the facilities to customers but also utilize the facilities in an optimum manner, thus preventing the use of duplicated or redundant facilities . Technological advancement has made it somewhat easier to design cost-efficient facility locations for municipal solid waste collection especially for large areas.

The most common approach for formulating FLP models in the literature is through the use of integer programming (IP). IP is a form of linear programming in which some or all the decision variables are restricted to be non-negative integer values. When all the variables take on integer values, then it is called a pure integer programming problem. On the other hand, the mixed-integer programming (MIP) model is a variation where some variables are real, and some are integers, or at least one variable is an integer. It may have a binary variable, which can be used to identify if any entity is active or not by being assigned 1 or 0, respectively.

2.2 Fixed Costs Capacitated Facility Location Problem (FC-CFLP)

In an FC-CFLP, the objective is to minimize the fixed costs associated with the potential facilities. The FC-CFLP is a minisum problem because it seeks to minimize the sum of the cost of flow between facilities and customers. The fixed cost is a one-time expenditure that varies from location to location, which is expected to be recovered during the entire life of the facility. To formulate the problem as a mathematical model, the following sets are defined: [Ref 25]

C is the set of all n customers indexed with i;
F is the set of all candidate facilities indexed with j.
The fixed cost for opening facility j is c_j
The transportation cost from facility j to customer i is t_{ij}
a_j is the capacity of facility $j \in F$
β_i is the demand of customer $i \in C$
The decision variables are
$x_j = \{0, 1\} = \{facility\ close,\ open\}$
$y_j = \{0, 1\}\ = \{customer\ not\ assigned\ to\ facility,\ assigned]$

The mixed integer programming model of the problem is as follows :

$$min\left(\sum_{i\in C}\sum_{j\in F}(t_{ij}y_{ij} +\ c_j\, x_j)\right) \tag{1}$$

such that

$$\sum y_{ij} = 1,\ for\ all\ facility\ j \tag{2}$$
$$\sum \beta_i y_{ij} \leq a_j x_j,\ for\ all\ facility\ j \tag{3}$$
$$\sum y_{ij} \geq x_j,\ for\ all\ facility\ j \tag{4}$$
$$x_j, y_{ij} \in \{0, 1\},\ for\ all\ i, j \tag{5}$$

The objective function in Equation (1) minimizes the total cost (transportation and fixed cost) associated with an open facility.
The constraint in Equation (2) ensures that each customer is allocated to only one facility
Equation (3) defines capacity constraints, and it ensures that the total demand ofcustomer i assigned to facility j does not exceed the capacity of j
By constraints in Equation (4), anumber of open facilities must not exceed the total number of customers in the system.
Equation (5) is the integer constraint.

2.3 Modelling Vaccine Logistics using FC-CFLP

2.3.1 Introduction

The daily number of new COVID-19 cases has crossed 200,000 as of today-19th April, 2021. We have various medically approved vaccines available for immediate vaccination, but vaccination centres keep running out of them. Keeping this in mind, we try to find an optimal solution of vaccine transportation and plant set-up locations, from an economic point of view.

In developing a model for the same, we make some assumptions:

1. There is only mode of transportation, and that is trucks.
2. The transportation cost per distance is the same for all routes.
3. Each location gets its vaccines from only 1 plant.
4. The fixed cost associated with opening a plant is the same for all.

Hence, we can use the FC-CFLP method to arrive at an optimal cost and network. Next we show the problem solving process behind arriving at the solution.

2.3.2 Step 1: Setting up the Data

ID	City	Population	Longitude(E)	Latitude(N)	x(km)=E*111	y(km)=N*111
MU	Mumbai	20,668,000	72.877426	19.07609	8089.394286	2117.44599
DE	Delhi	31,181,000	77.216721	28.6448	8571.056031	3179.5728
HY	Hyderabad	10,269,000	78.491684	17.38714	8712.576924	1929.97254
PU	Pune	6,808,000	73.856255	18.516726	8198.044305	2055.356586
CH	Chennai	11,235,018	80.237617	13.067439	8906.375487	1450.485729
GH	Ghaziabad	2866384	77.449791	28.667856	8596.926801	3182.132016
AH	Ahmedabad	8,253,226	72.585022	23.033863	8056.937442	2556.758793
NS	Nasik	2,123,000	73.789803	19.997454	8190.668133	2219.717394
BG	Bangalore	12,765,000	77.580643	12.972442	8611.451373	1439.941062
KL	Kolkata	14,974,000	88.363892	22.572645	9808.392012	2505.563595
ID	Indore	3,113,000	75.857727	22.719568	8420.207697	2521.872048
NG	Nagpur	2,940,000	79.08886	21.146633	8778.86346	2347.276263

We use 12 cities as the demand locations for vaccines.The population and coordinate data is given in the references 1-24

There are mainly 5 candidates for Vaccine Manufacturing Unit. Ref 25

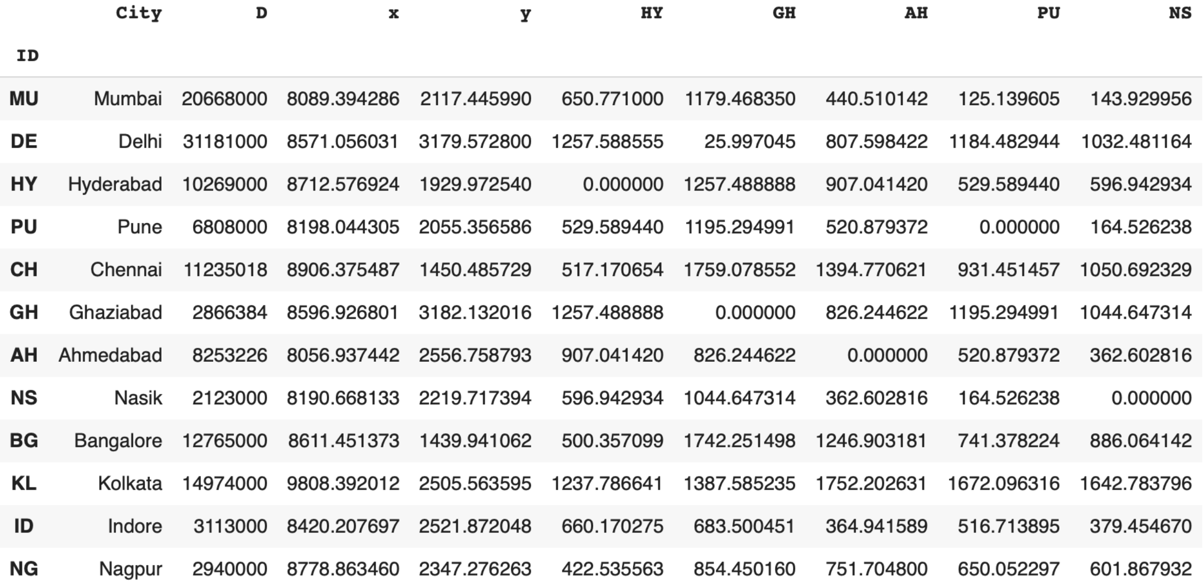
1. Bharat biotech International Ltd,Hyderabad (HY)
2. Biomed Pvt. Ltd, Ghaziabad (GH)
3. Cadila healthcare, Ahmedabad (AH)
4. Serum Institute of India, Pune (PU)
5. GSK Asia Pvt. Ltd.,, Nasik (NS)

Next, we approximate the 2 cost components.

1. Truck Freight Rate is Rs 35/km
2. Fixed Cost associated with setting up plant is 113 Million USD which is 8.4 Billion INR

2.3.3 Step 2: Creating Distance Matrix

We use the x and y coordinated to find distance of each city from each of the 5 candidates using euclidean norm.

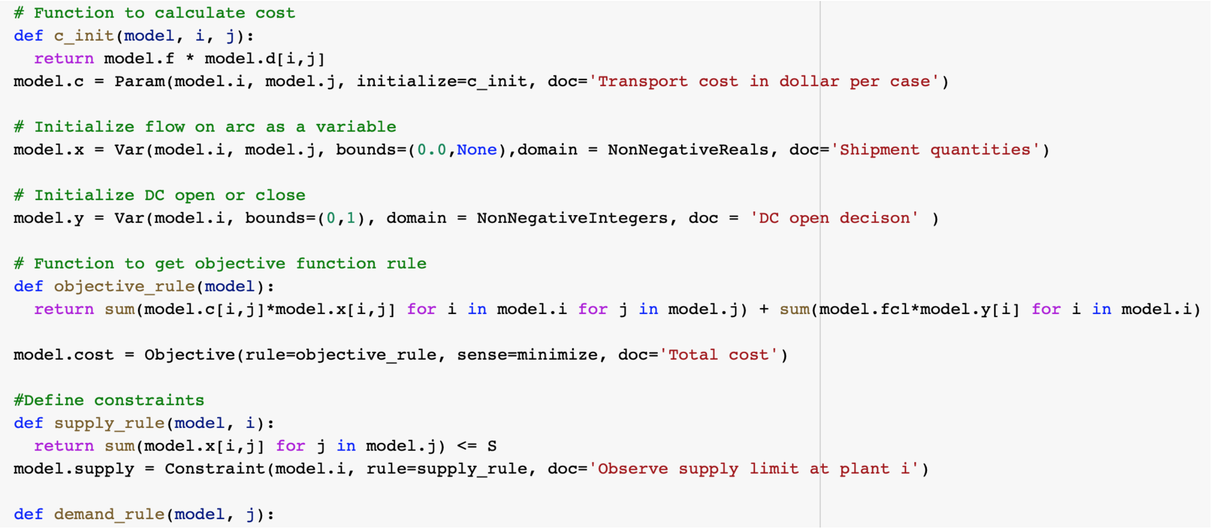


	City	D	x	y	HY	GH	AH	PU	NS
ID									
MU	Mumbai	20668000	8089.394286	2117.445990	650.771000	1179.468350	440.510142	125.139605	143.929956
DE	Delhi	31181000	8571.056031	3179.572800	1257.588555	25.997045	807.598422	1184.482944	1032.481164
HY	Hyderabad	10269000	8712.576924	1929.972540	0.000000	1257.488888	907.041420	529.589440	596.942934
PU	Pune	6808000	8198.044305	2055.356586	529.589440	1195.294991	520.879372	0.000000	164.526238
CH	Chennai	11235018	8906.375487	1450.485729	517.170654	1759.078552	1394.770621	931.451457	1050.692329
GH	Ghaziabad	2866384	8596.926801	3182.132016	1257.488888	0.000000	826.244622	1195.294991	1044.647314
AH	Ahmedabad	8253226	8056.937442	2556.758793	907.041420	826.244622	0.000000	520.879372	362.602816
NS	Nasik	2123000	8190.668133	2219.717394	596.942934	1044.647314	362.602816	164.526238	0.000000
BG	Bangalore	12765000	8611.451373	1439.941062	500.357099	1742.251498	1246.903181	741.378224	886.064142
KL	Kolkata	14974000	9808.392012	2505.563595	1237.786641	1387.585235	1752.202631	1672.096316	1642.783796
ID	Indore	3113000	8420.207697	2521.872048	660.170275	683.500451	364.941589	516.713895	379.454670
NG	Nagpur	2940000	8778.863460	2347.276263	422.535563	854.450160	751.704800	650.052297	601.867932

D is the population. Distance is given in km.

2.3.4 Step 3: Solving using Pyomo Library

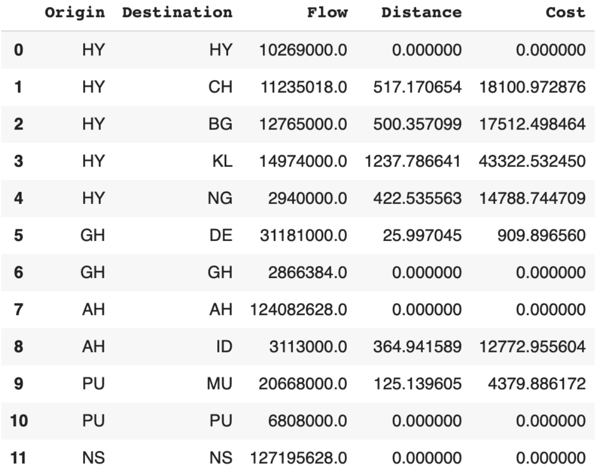
Pyomo is a Python-based open-source software package that supports a diverse set of optimization capabilities for formulating, solving, and analyzing optimization models. Using Pyomo, a user can describe an optimization model by specifying decision variables, constraints, and an optimization objective. Pyomo includes a rich set of features to enable modelling of complex systems, specifying a solver, and displaying the solution. Pyomo can be used to define general symbolic problems, create specific problem instances, and solve these instances using commercial and open-source solvers.



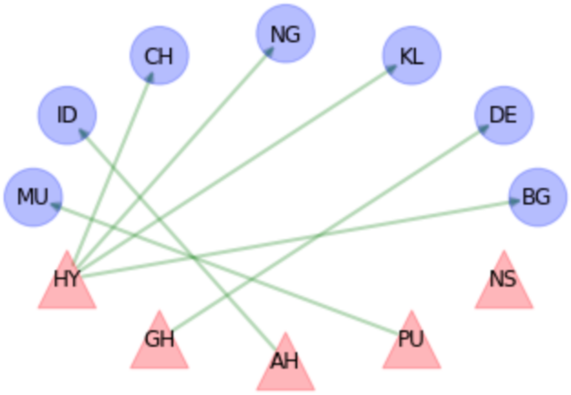
```
# Function to calculate cost
def c_init(model, i, j):
  return model.f * model.d[i,j]
model.c = Param(model.i, model.j, initialize=c_init, doc='Transport cost in dollar per case')

# Initialize flow on arc as a variable
model.x = Var(model.i, model.j, bounds=(0.0,None),domain = NonNegativeReals, doc='Shipment quantities')

# Initialize DC open or close
model.y = Var(model.i, bounds=(0,1), domain = NonNegativeIntegers, doc = 'DC open decison' )

# Function to get objective function rule
def objective_rule(model):
  return sum(model.c[i,j]*model.x[i,j] for i in model.i for j in model.j) + sum(model.fcl*model.y[i] for i in model.i)

model.cost = Objective(rule=objective_rule, sense=minimize, doc='Total cost')

#Define constraints
def supply_rule(model, i):
  return sum(model.x[i,j] for j in model.j) <= S
model.supply = Constraint(model.i, rule=supply_rule, doc='Observe supply limit at plant i')

def demand_rule(model, j):
```

Pyomo supports a wide range of problem types, including: Linear programming,Quadratic programming, Nonlinear programming and Mathematical programming with equilibrium constraints. We use the glpk solver function defined in the Pyomo library to solve the mathematical programming problem defined. The GLPK (GNU Linear Programming Kit) package is intended for solving large-scale linear programming (LP), mixed integer programming (MIP). The glpk solver helps us to get the solution of the mathematical optimization problem which is defined.

2.4 Results



	Origin	Destination	Flow	Distance	Cost
0	HY	HY	10269000.0	0.000000	0.000000
1	HY	CH	11235018.0	517.170654	18100.972876
2	HY	BG	12765000.0	500.357099	17512.498464
3	HY	KL	14974000.0	1237.786641	43322.532450
4	HY	NG	2940000.0	422.535563	14788.744709
5	GH	DE	31181000.0	25.997045	909.896560
6	GH	GH	2866384.0	0.000000	0.000000
7	AH	AH	124082628.0	0.000000	0.000000
8	AH	ID	3113000.0	364.941589	12772.955604
9	PU	MU	20668000.0	125.139605	4379.886172
10	PU	PU	6808000.0	0.000000	0.000000
11	NS	NS	127195628.0	0.000000	0.000000

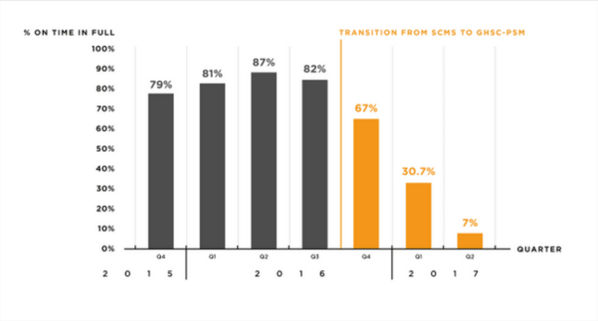


As can be seen from the optimal network flow chart, for the given parameters, all 5 candidates are active. But Nashik supplies to itself only. Rest of the candidates supply to themselves and others. The total cost sum that is minimised comes out to be Rs. 131975,94,92,163 or 131975 crores .Here, arrow length is not scale able to distance between two cities. We will now discuss delay considerations in a supply chain using Machine Learning Concepts.

3 Supply Chain Analysis Using Machine Learning

3.1 Case Study Overview

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 prolong life and 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 2020, only 19.5 million received this life-saving treatment. Timeliness of HIV medicines procurement is critical to the program's efficiency and impact in saving lives, controlling, and eventually eliminating HIV. Delays in the 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 critical factors in predicting whether HIV drugs are delivered on time or not. It will then use these factors to predict how long delays are likely to be, thus allowing HIV/Supply Chain program managers to know when and which products are likely to be delayed, as well as the extent of the delay so that they can take mitigating action to save lives and avoid additional supply chain costs.



3.2 Proposed Solution

This study used a combined model which uses classification machine learning algorithms to predict whether a particular product is delayed or not and then use regression analysis to indicate the length of the delay using the subset of the data which the classification predicted will be delayed. For selecting the best model, both the classification and regression versions of the following models will be explored evaluated against predetermined benchmarks of the Random Forest model with default parameters in SciKit-Learn: i) ExtraTrees ii) XGBoost iii) SupportVector Machines (SVM), and iv) Multi-Layer Perceptron (MLP). Random-Forests, ExtraTrees, and XGBoost are proven high-performing ensemble algorithms which can do automatic feature extraction. Simultaneously, SVMs perform very well with high-dimensional data and can detect non-linear relationships if the correct kernel is used. Finally, MLPs are useful for high-dimensional time-series data. These algorithms' above advantages are well-suited to the selected dataset, which has several categorical columns that 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 are well-defined on the data. For example, data on scheduled versus actual delivery dates can be determined by delay occurrences and duration, allowing precise quantification and measurement of the problem and solution. This study's results will apply to future instances of supply chain orders. Thus it applies to future occurrences of similar supply chain data observations and valuable for planning purposes.

3.3 Evaluation Metrics

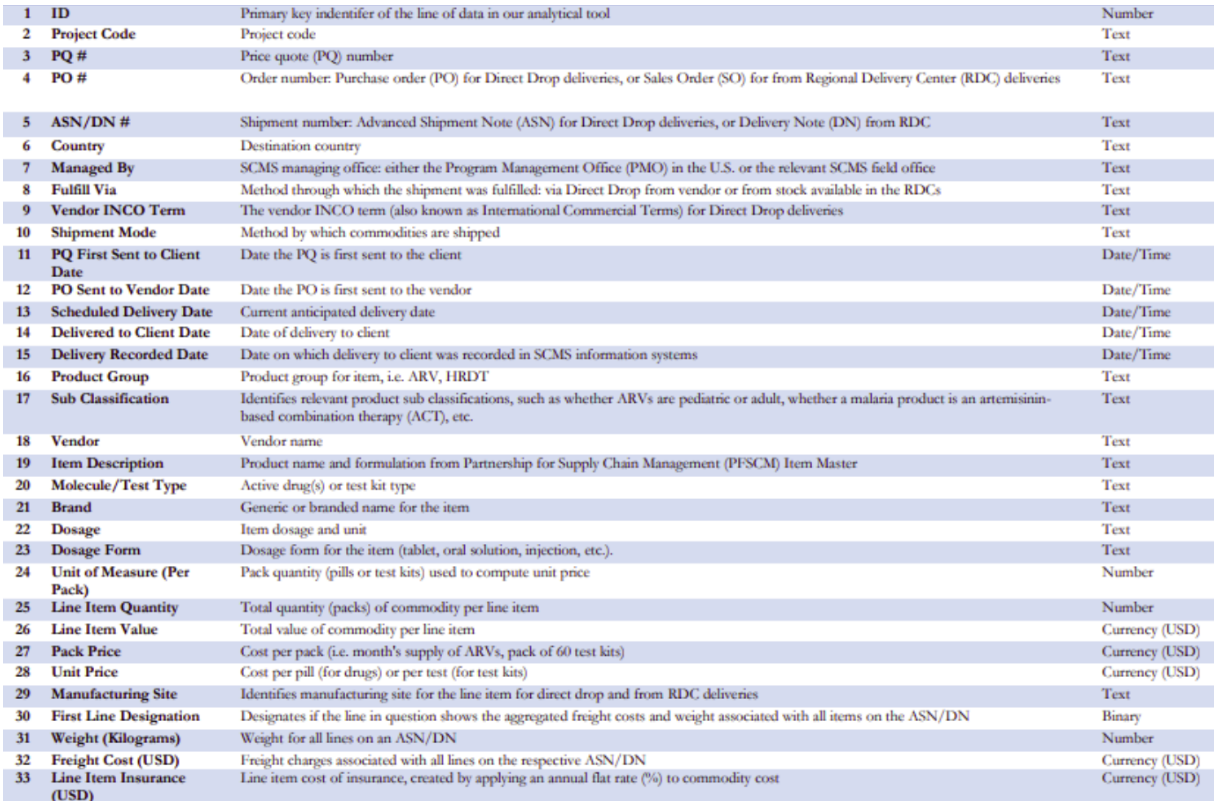
The resulting combined models will be evaluated based on 4 metrics: Recall and F1-Score for classification, to balance the recall/precision trade-off, mainly because the dataset is unbalanced with a ratio of 1:9 between the positive and negative class respectively. For the regression part of the model, 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.

1. Recall: measures the success rate of correctly labeling the positive items i.e. what proportion of the positive labels did we successfully identify? This is important because it tells us what proportion of delays we can actually predict. Recall = True Positives / (True Positives + False Negatives)
2. F1-Score is an average (harmonic mean) 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)

3. R-squared is the “coefficient of determination” that measures the amount of variation in the data explained by the model, again as a percentage/fraction of total variation.
4. 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. Here, “y-hat” are the predicted values and “yi” are true values of the target variable. “n” is the number of observations in the dataset.

3.4 Data Preparation

We have used the PEPFAR program’s Supply Chain Management System data which is publicly available. It has 10,000+ data items over 33 features such as product details, the country, the manufacturer and the shipment details. The two target columns are “on-time”, a binary variable for classification and “delay” , a continuous variable for regression.



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)

1. **Handling Missing Data:** 3 columns - Dosage,Line Item Insurance and Shipping Mode had missing values (1736, 287 and 360 respectively) which have been imputed using mean or mode methods. For the missing dates, estimates are made using other data columns. For example using expected delivery dates, purchase order dates are extrapolated, which are further used to estimate purchase quote dates. Assuming the weight of standard items should be constant, the missing weight values are imputed using item quantity * item weight. Freight cost is proportional to the weight and hence is computed depending on whether the items were single or bundled.
2. **Feature Engineering:** Time data (like year, month, week day)was captured using pandas library date-time functionality. Numeric data like counts, proportions sums were calculated at country-year, factory-year, vendor-year, brand-year levels which were then merged with the item-level data. In shipment configuration, using string functions and regular expressions, the shipments were divided into single or bundled. The target value for regression; “delayed” was computed as difference of scheduled delivery date and date of delivery to the client. External features on logistics and country fragility(which also required some cleaning) were also obtained.

3.5 Exploratory Data Analysis and Feature Selection:

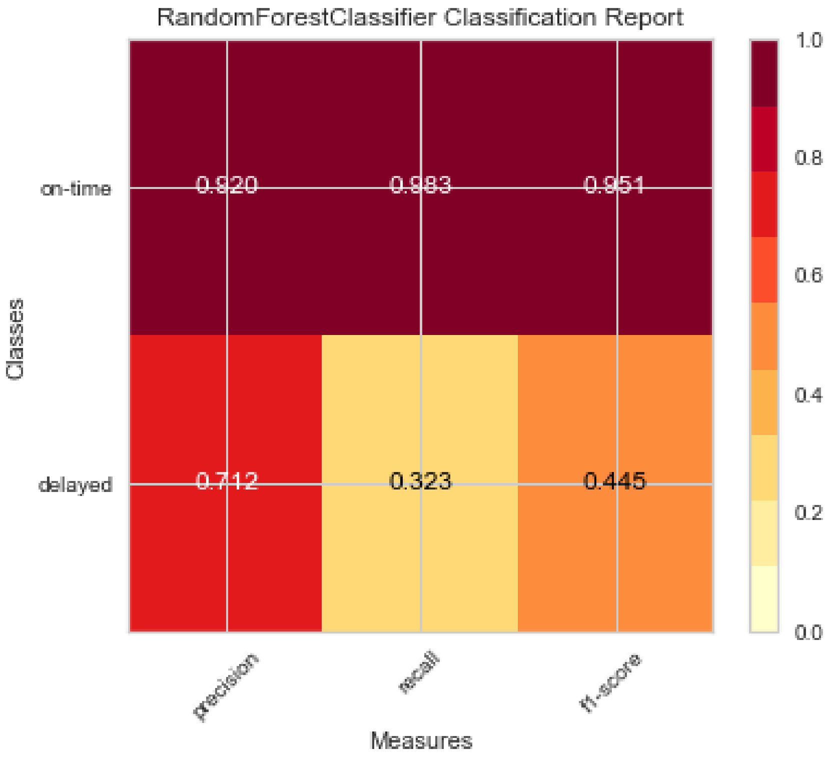
Although there are a lot of useful features, some of them are highly correlated, using all of which would prove challenging and wasteful for linear regression. However, Random Forest Classifiers don't face any trouble dealing with such features.

Findings of the EDA.

1. The years 2010-2014 had above average delays with January having month wise higher delay rate, and weekends having higher delay rate.
2. Price, Weight, Freight Cost are some of the features which have high correlation.
3. The value and quantities of products have moderate correlation with the delays, with vendor metrics dominating in correlation.
4. The deliveries from fragile to fragile countries have longer delays.
5. Better logistics of country of origin lead to lesser delays.

3.6 Model Benchmark

The solution model is a combination of two algorithms working together sequentially; thus, 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. Showing Results



Classification benchmark results for SciKit-Learn Random Forest algorithm

1. **Classification results** [Recall: 0.33 F1-score: 0.445 Total: 134 instances of delayed delivery correctly identified]
2. **R-squared: 0.85** [RMSE: 13 days]

3.7 Model Selection

After preprocessing the data through a pipeline for logarithm, standard scaling, one-hot and label encoding as well as oversampling to balance the classes, the following models were compared:

1. **Classification**:LinearSVC, SVC, KNeighborsClassifier, LogisticRegressionCV, LogisticRegression, SGDClassifier, BaggingClassifier, ExtraTreesClassifier, RandomForestClassifier, MLPClassifier ,GaussianNB, LinearDiscriminantAnalysis

2. **Regression**: LinearSVR, SVR, KNeighborsRegressor, LinearRegression , SGDRegressor, BaggingRegressor, ExtraTreesRegressor, RandomForestRegressor, MLPRegressor Final Models selected: Extra Trees Classifier for classification and Extra-Trees Regressor for regression

LinearSVC: The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

KNeighborsClassifier: KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). Random Forest uses a subset of data using bootstrap aggregation. Multiple decision trees are grown and a weighted average of the outputs is given as final output.

Extra Trees is like a Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not best splits. So in summary, ExtraTrees:

— builds multiple trees with bootstrap = False by default, which means it samples without replacement

— nodes are split based on random splits among a random subset of the features selected at every node

In Extra Trees, randomness doesn't come from bootstrapping the data, but rather comes from the random splits of all observations. ExtraTrees is named for (Extremely Randomized Trees). Final Models selected: Extra Trees Classifier for classification and Extra-Trees Regressor for regression.

3.8 Model Comparison

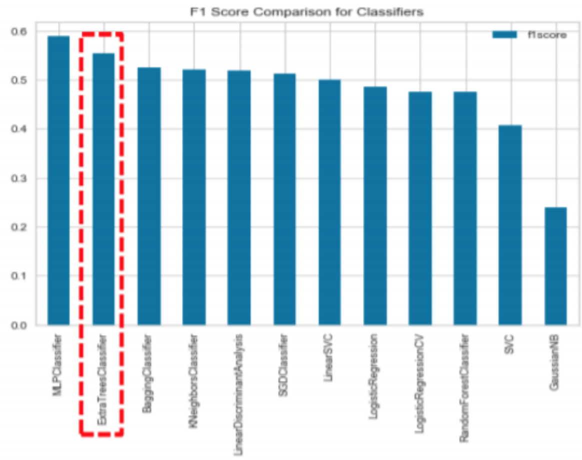
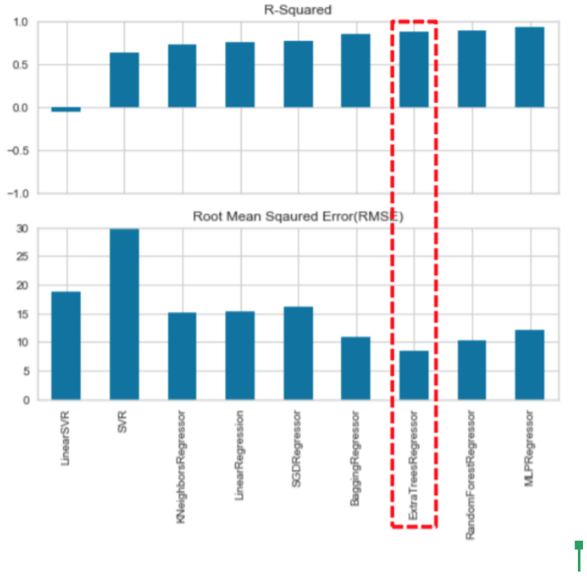


Figure 1 – F1 Score comparisons for classification model selection

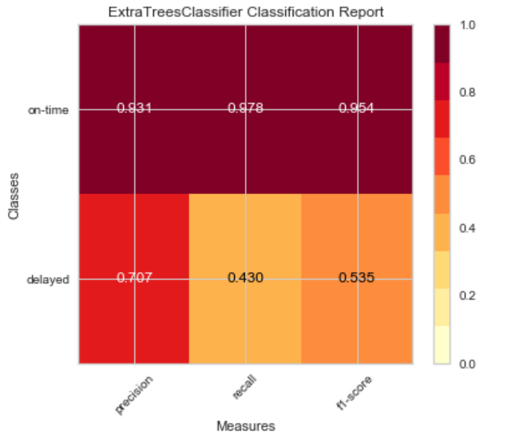


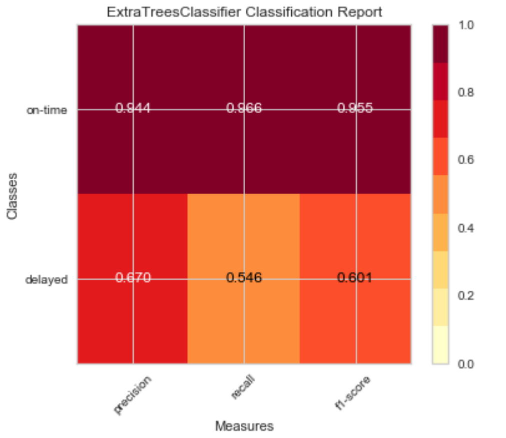
R-squared and RMSE comparison for model selection

Figure 2 – R-squared and RMSE comparison for model selection

3.9 Model Improvement and Fine-tuning

In general, the higher the n-estimators, max-features and max of depth, the more able the Extra Trees model is to predict accurately. Since it works with randomly-chosen cut points, it requires more trees to reduce the variance, a significant (but not too large) number of features on which to split (so as to capture most of the signal available in the data), as well as a good number for maximum depth to allow the model to capture all the nuances in a particular feature (thus increasing bias). All these features increase the accuracy of the model.





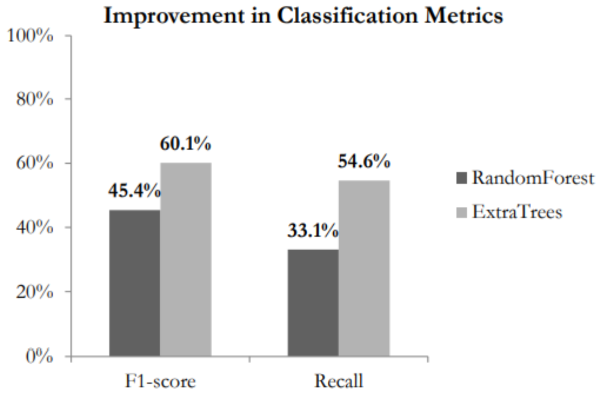
Final hyper parameters chosen:
Classification : ExtraTreesClassifier (n-estimators=900, max-features= 50 , criterion= 'entropy',max-depth= 50, random-state=121)
Regression : ExtraTreesRegressor(n-estimators=900 ,max-features= 50,max-depth= 50, random-state=121)

3.10 Final Model Improvements

Supplier-side factors such as origin country stability, vendor and brand volumes as well origin country logistics environment explain significant variation in the data. Together with customer/receiver-side factors, they explain about a third of the variation in the data. The rest of the variation is due to product level characteristics like volumes, value/price, and weight as well as how they evolve over time (time-series). Of note is the significant auto-correlation where vendors who have delayed items in the past as well as more recently are more likely to delay deliveries again in the future. These insights were used for feature selection for the final classification and regression models.

The Extra Trees Classifier and Extra Trees Regressor were selected as the best algorithms for the classification and regression tasks respectively. Both algorithms outperformed the benchmark Random Forest and several other algorithms. Here are the observed improvements.

Metric	RandomForest	ExtraTrees	Improvement
F1-score	45.4%	60.2%	33%
Recall	33.3%	50.6%	52%
Precision	71.1%	74.3%	5%
R-Squared	80.7%	86.3%	7%
RMSE (days)	15.91	11.97	25%



4 Annexure

1. Economic Considerations of Supply Chain Management Code: https://colab.research.google.com/drive/1c1VQT0IXqajJ0ZzwHqN339hN0OtWbbvn?usp=sharing
2. Economic Considerations Data
   https://drive.google.com/file/d/1u3PMFfaeQheQeZeJ0rIwdGiDx2iYhFib/view?usp=sharing
3. Supply Chain Analysis Using Machine Learning code: https://drive.google.com/file/d/17kiq986Mh_H3irxMgUu0KBAOSM3oIelW/view?usp=sharing

5 References

1. https://www.macrotrends.net/cities/21206/mumbai/population
2. https://www.macrotrends.net/cities/21228/delhi/population
3. https://www.macrotrends.net/cities/21275/hyderabad/population
4. https://www.macrotrends.net/cities/21371/poona/population
5. http://www.populationu.com/cities/chennai-population
6. https://indiapopulation2020.in/population-of-ghaziabad-2020.html
7. https://worldpopulationreview.com/world-cities/ahmedabad-population
8. https://www.macrotrends.net/cities/21350/nashik/population
9. https://www.macrotrends.net/cities/21176/bangalore/population
10. https://www.macrotrends.net/cities/21211/calcutta/population
11. https://www.macrotrends.net/cities/21278/indore/population
12. https://www.macrotrends.net/cities/21347/nagpur/population
13. https://www.latlong.net/place/mumbai-maharashtra-india-27236.html
14. https://www.latlong.net/place/new-delhi-delhi-india-2441.html
15. https://www.latlong.net/place/hyderabad-telangana-india-1136.html
16. https://www.latlong.net/place/pune-maharashtra-india-563.html
17. https://www.latlong.net/place/chennai-tamil-nadu-india-2284.html
18. https://www.latlong.net/place/ghaziabad-uttar-pradesh-india-13888.html
19. https://www.latlong.net/place/ahmedabad-gujarat-india-1187.html
20. https://www.latlong.net/place/nashik-maharashtra-india-2367.html
21. https://www.latlong.net/place/bangalore-karnataka-india-499.html
22. https://www.latlong.net/place/kolkata-west-bengal-india-257.html
23. https://www.latlong.net/place/indore-madhya-pradesh-india-2528.html
24. https://www.latlong.net/place/nagpur-india-622.html
25. Facility Location Problems: Models, Techniques, and Applications in Waste Management-Olawale J. Adeleke 1 and David O. Olukanni 2,*
26. https://www.onupkeep.com/answers/maintenance-history/four-industrial-revolutions
27. https://www.coursera.org/learn/optimize-supply-chains-analysis-google-sheets/
28. https://www.coursera.org/learn/optimize-supply-chains-analysis-google-sheets/
29. https://tallyfy.com/root-cause-analysis-rca/#:~:text=A%20single%20root%20cause%20can%20use%20can%20have%20multiple%20effects.&text=The%20root%20cause%20was%20the,%E2%80%9D%20you’re%20looking%20for
30. https://marutitech.com/machine-learning-in-supply-chain/
31. https://cdsco.gov.in/opencms/export/sites/CDSCO_WEB/Pdf-documents/biologicals/facilitiesLIST.pdf