Kata Challenge

For: Ron McChesney, candidate for Data Science/Machine Learning position

From: Pillar Technology, Delivered at 8:04 AM on December 7, 2017

Posted on Github Acct: 8:00 AM on December 14th, 2017

Summary:

Client Problems/Challenges they would like Solved:

Mission is to build three models, to (1) identify how to minimize inventories stocked in warehouse(s) without running out of product for clients, (2) build an order form user friendly to wholesalers’ retail(stockist) clients with repeat orders and order management needs exceeding occasional or one time clients, and (3) relate the running operations of the business to recommendations to create a regional then global network of warehouses as the business expands revenue and its geographical footprint as defined by the location of its’ client base.

Process to Obtain Recommendation for Next Phase:

Think about the problems and possible solutions while evaluating the provided data set. Record three sets of thoughts and ideas as problem is better identified, data better understood, and new ideas emerge for possible solutions.

Initial Reaction (Thursday evening):

Model One:

The 1st part of the mission appears to be a standard inventory management model based on a constrained optimization model structured as a cost minimization equation. Given the small number of fields in the sample data set, supervised learning may be warranted, if we can identify a valid multivariate regression equation.

Model Two:

The 2nd part of the mission appears to be a front end tool created to maximize client satisfaction with the ordering process, along with provide information that will inform changes in inventory. This could include identifying to the client the number of items currently in stock, and provision to order what is on the shelf and place a backorder for the rest. In this sense, the values of each shelf item can be a negative number, with the <0 values indicating a backorder condition.

Both models 1 and 2 can be summarized for C-Suite executives as the firm using Microsoft PowerBI dashboards, that update as changes to inventory take place.

As there was one warehouse during 2011 (and we believe that to be the case as of late 2017), the existing business can be described using a single vector of items. For larger online ordering firms, such as Amazon, warehouse locations are in the 50-100 location range, therefore a matrix of locations to items might be something on the order of a mxn or 50x50,000 matrix.

Model Three:

The 3rd part of the mission appears to be a global scale location modeling exercise, where the best locations are selected for warehouse expansion based on a set of criteria. This can include transportation fees, time and related logistics, using straight line or actual route lengths. This can also include other push and pull factors such as taxes, regulation, infrastructure, business, education (labor) and security.

An approach would be to assume that location would be in one or more metropolitan areas that have a population of at least 500,000 persons, and in 2016 there were about 1,000 of these metros worldwide. Or, the model assumes a choice set of 1,000, which creates a 1,000 x 1,000 trading and distance matrix comprising of 1,000,000 cells (with the upper and lower regions of the matrix 499,500 cells each and a diagonal with 1,000 cells).

The model would be to use a variant of the Huff Model, to (1) identify low cost warehouse location rankings as a function of competitive trade with industry relevant retailers, and (2) weight this with a set of location attribute variables. Model output can be evaluated using the Nystuen-Dacey method, which will identify a global hierarchy of metros, from which strategic location decisions can be made (select a set of branches from a tree with higher nodal locations the initial set of warehouse sites).

Model output would be an estimate of a good choice set for warehouses, and an estimate of the market share that warehouse would have with other competitive warehouses selling similar goods in each metro markets. As competitors made moves to expand or recede their warehouse footprints, the impact on revenue for the firm can be simulated. The model also allows for estimates of total shipping or transportation costs that may vary with changes in cost inputs such as fuel costs, border restrictions, and other factors.

Integrated Model:

All three models can be related to each other, with an eventual goal of real time. Model 2 feeds into Model 1, and Model 1 output informs Model 3 scenarios. Therefore, as an example, given the firm and their set of clients, shifting geography and product mix over the course of say, one year in time, the location model will shift warehouse recommendations, preventing sub-optimal location decisions from being locked in, at a long term direct and opportunity cost measured in millions of dollars.

2nd Reactions (Saturday evening)

If we think of the client as having a set of about four thousand items, and each item is a shelf location in a warehouse, then each shelf can be thought of as a cell, and the catalog as a vector of cells. Each cell has a history, of items being moved into and out of the cell, as orders are placed and reorders are received. It might be that a blockchain model, that records this activity, such that each item has its own blockchain history, might result in some interesting results and possibilities. The chains can be then aggregated, to construct an invoice history. Or, the overall matrix can be deconstructed into a series of blockchains, and the series of blockchains can be reconstructed into an aggregate matrix.

This client might be best off staying in a niche, and hence either an expansion with smaller size warehouses in several other locations in Britain, or an expansion to cover Europe (which does have ten times the population of England), might be recommended, instead of a global scale expansion model.

More specifically, if the firm is a 15 million dollar a year revenue company, with most of its client trade occurring in the United Kingdom, and the objective is to expand using small warehouses that carry 20 percent of catalog items that represent 80 percent of sales, then each new smaller warehouse might be in the order of a 3 million dollar a year business expansion. Five of these warehouses could be located in Britain, assuming the competitive market would allow for this expansion. Given increased demand for same day delivery, the main location would remain in London, but the five smaller warehouses could be located in markets with at least 3 million residents. An example of this might be small warehouses in Bristol, Birmingham, Manchester, Newcastle Upon Tyne, and Glasgow. Two hour delivery windows might cover the majority of the British population and locations of the network of retailers (stockists).

3rd Reactions (Monday evening)

An inventory model that combines with a transportation model for receipts from vendors and delivery paths to clients, and a history of a shelf item using blockchain model might better inform not only Model One, but also Model Two and Model Three.

The data set is probably best separated into a few different matrices. The first is a matrix of invoice numbers and stock numbers. The second is a matrix of invoice numbers and customers. The third is a matrix of stock numbers and customers. These can be created from the existing data set, and while they will create sparse data sets (on purpose), the evaluation of these may suggest some ideas for how to create an order prediction model.

Also, for the prediction model, after clients are clustered into infrequent invoices, moderate invoices, and frequent invoices, we probably are best creating a separate model for each of the moderate or medium activity clients, rather than trying to run a model off aggregate data set. Easier to input run and evaluate. Once we have what we think is a good model, we can then try running it as a batch process, with output being range of predictions or a distribution of predictions for the study set of perhaps 2,000 clients. Then we can try different forms of the model, to get a boost in performance.

Need to study the Minkowski and Hamming distance measures, which might be helpful for one or all of these models. Know Euclidean, Great-circle and Manhattan distances, but these two are new to me.

**Recommended Solutions**

Model One for Inventory Management:

**MIN-Plus with Blockchain History Model Training**

Model Two for Client Side Management:

**Cluster and Custom Neural Networks with Markov Chain Transition Assessment**

Model Three for Business Development Management:

**Huff-Global Competitive Warehouse Model with Nystuen-Dacey Hierarchical Assessment**

Client Benefits:

Increased information about the major elements of the business.

Increased worker productivity and reduce operating expenses.

Reduced Inventory Stocks and increased cash flow.

Increased Revenue.

Improved probability to achieving pathways toward significant business expansion.

Therefore, cash flow savings resultant from an inventory management minimization strategy may be only a relatively small element of the overall value to the client from implementing all three models.

Factors to consider:

Inventory management is actually a chained set of inventories, each with a management strategy to minimize stocks while also minimizing shortages. This includes manufacturers who wish to fulfill orders from wholesalers without leading to production line stockpiles. Warehouses want to receive items, and then repackage and send to their retail clients with minimal inventory levels, to control cash flow and to minimize losses due to other factors such as spoilage, damage, shifts in market demand, and pilferage. Retailers wish to stock items in a way that customers prefer (big pyramid of oranges in a produce section to select from, etc.) while minimizing inventory and risks of loss. Customers may have their own inventory at home (always want to have a dozen eggs in the fridge, etc.) The wholesaler inventory minimization problem is one element of a larger value change optimization problem.

The actual company related to this data set was identified using the description product codes. Since the data may have been intended to be anonymous, we will not reveal them here. However, a review of their West London location and their online web site suggests a few important factors: (1) they are not in an optimal location unless the majority of their clients are located in the London area; Leicester in the midlands has the best access to the United Kingdom market, (2) The term ‘Stockist’ refers to retailers, (3) The web site may not be state of the art but it is not bad, (4) company does not appear to be positioned for large scale rapid growth, unless there is a change in ownership or major shift in business development objectives, and (5) This is a low margin business selling basic craft store/dollar store types of goods.

Pillar Technology Consulting Work and possible Product Development

Steps

Agile Sprint Week Zero. Fee: $0

Agile Client Discovery Project Consisting of One Month and Four One-Week Sprints. Fee: $36,000

Agile Client Engagement/Implementation Project Consisting of Four months and Twenty- One Week Sprints. Fee: $288,000

The Discovery project might consist of a team of three persons, assigned 20 hours a week each, for four weeks. The team might be a front-end designer to create a inventory dashboard, a customized order screen that is of different designs depending on what cluster group a client is in, when they are logging in, and a warehouse scenario dashboard, a coder to process databases and calculate outputs, and a data scientist to identify the best models and related algorithms. Objective is to fully model out the three models so all the moving parts can be viewed and reviewed.

The Implementation project might consist of a team of six persons, assigned 20 hours a week each, for sixteen weeks. Each sprint would be designed to produce a unit of workable code, with the final three models and integration of the three models to work as one overall business management product, created at the end of the backlog completion and burn down of the project.

Data Set

The data set is a listing of order or sales transactions. It forms a 541,910 row by 8 column matrix.

The data tells the following story: A medium size firm with about 10 million pounds (about 15 million dollars, and in general I will refer to generalized comments relative to dollars instead of euros or pounds, although the dataframe matrix is expressed in terms of 2011 pounds) of revenue, an average price per unit sold of about $2.70, and about five million units shipped during 2011. Thousands of clients, thousands of items, with most items being prices for less than ten dollars per unit. Prices for the same unit vary, perhaps due to different buying programs. Most clients are located in the United Kingdom (91.5%), with some business in France and Germany and other parts of Europe, and trivial amounts in the rest of the world. Over 99% of sales are in Europe. Not a global company, rather a Western European company. Most clients are repeat buyers and tend to buy a dozen or s items with every invoice. No correlation between unit price and quantity sold; overall a relatively homogenous distribution of clients, prices, units shipped and revenue.

The data does not reveal inventory stocks, requiring inference to guess at what the inventory stock might be, from which we can assess to what extent modeling inventory management with minimize stocks while minimizing probability of shortages due to unanticipated large orders, can result in a lower average inventory that can free up some level of cash flow resources. Assuming a 50 day aggregate inventory stock, an inventory estimate can be made, which would provide some idea of the impact of a modest reduction in inventory levels resulting from a modeled inventory management system that would recommend a stock level for each item in the warehouse.

The data is in relatively good shape, with quirks in each of the eight variables or fields. There are negative values which may indicate returned items, accounting items that do not belong in the records, including bad debt provisions, some missing values that could be interpreted in several ways. Overall, some data cleaning results in 529,064 out of 541,909 usable records, or 97.63% records retained.

The first row are field descriptions, telling us that this data consists of: (1) an invoice number, (2) a stockcode number, (3) a text description of product items, (4) quantity of units ordered, (5) data of invoice expressed in day and minute, (6) UnitPrice, (7) a Customer number, and (8) the Country from which the product is being ordered.

1. INVOICENO or invoice number is a six digit code, starting at 536365 and ending with 581587. While this might suggest 45,222 invoices generated, not all numbers in the sequence are used, possibly due to discarded order or some other factor. The invoice numbers indicate that orders consists of one or more items, generally in the range of ten to twenty items in a typical order. While the data is missing some of the numbers in the sequence, it does not contain any blank or null values.
2. STOCKCODE or the item number is either a five digit code, or a six digit code with five numbers and one letter. There are some quirks in the code as it relates to text string descriptions, and some items are virtually identical to each other except for minor differences, but it would appear that there are about 3,866 different stock codes representing 3,866 distinguishable items.
3. DESCRIPTION or the description field is text typically about 5-7 words, for a variety of inexpensive craft and decoration related goods. Many of the items are similar, with variations in design and color. There does not appear to be a pattern in the Stockcodes and Descriptions (which are two perfectly correlated fields), in the sense one might find in a retail store that has goods stocked in aisles according to some categorical typology.
4. QUANTITY refers to the number of items of the same stock code ordered on the same invoice. The number is an integer that ranges from 1 to 80,995, with a median value of 6. The data does not reveal items for which zero orders might be placed during a year. Presumably these items would be discontinued, but they do represent an unknown portion of the total inventory stock. Looking at the distribution with quantity, while records where the quantity ordered is less than 10 items represents 75% or most of the orders, they only represent 21.7% of the total units sold during the year. Orders of 11-99 units represent 45.6% of total units sold, and orders equal or greater than 100 units represent 32.7% of units sold.
5. INVOICEDATE or the Date Stamp is the date in which an item has an invoice number. The data includes day, month, year, and time. We do not know if this refers to the date in which an order is made, the date the order if fulfilled and shipped, or the date in which an invoice is generated. These could be three separate dates or roughly the same if orders are generally placed and shipped and billed on the same day. Also, it might be the invoice is generated when an online order is placed, and this might occur before the goods are boxed and shipped, as opposed to after. The field, does allow one to separate the data into units and revenue per month. This shows that there is a clear seasonality where orders increase in the months of October and November (as if gearing up for the holiday season). Also the data starts in 2010 and goes partway into December, 2011. For a 12th month evaluation, we must use the time period December 1, 2010, to November 30th, 2011, which is not an optimal time series. Also, seasonality must be adjusted if forecast or projects into the 2nd year using this one year of data, is to have any veracity.
6. UNITPRICE is a currency number, that we assume but do not know for sure, is represented in British pounds. The range is .04 for a Pencil to 649.50 for a 60 Piece Wicker Picnic Basket (seems a bit high unless it is 60 baskets). Most items are in the 1 to 10 pound range, with an average of 2.94 pounds for the 397,574 items whose invoice record includes a customer number. The firm makes money selling millions of low priced units to thousands of clients. We assume the price is the price sold to their clients, as opposed to the cost of goods sold price paid to suppliers before markup, but we do not know that for sure.
7. CUSTOMER is a five digit number. About thirty percent of the records do not have a customer number. While it might be tempting to simply delete these records, the rest of the fields look like normal invoices, items selection and related, and the missing fields could be missing for a number of reasons; hence for now they are left in. Of the records that have client numbers, there are 4,335 clients, and it appears that the average client purchases about 100 items over the course of a year, in invoice batches of five to twenty items per order.
8. COUNTRY or the shipping destination lists the name of the country the client is located in, although we do not know for certain if this represents billing or shipping addresses. There are a total of 36 countries listed. The United Kingdom represents 91.5% of the cell entries, while European countries represent 99.4% of the cell entries. There are some global entries for the USA, Brazil and Singapore, but they are trivial. In fact, the real company states on their web site that they do not currently sell to the United States.

Normal distribution or Skewed?

If sales are not normally distributed, that the use of the mean for modeling will create error, and median or mode might be considered as alternative measures of central tendency.

A rank size distribution and 80/20 Rule:

If 80% of revenue results from 20 percent of items, then while the importance of variety is not challenged (for example, and ice cream store with 30 flavors, most popular vanilla, etc, … will have a least popular flavor that is maintained or rotated with other least popular flavors to preserve choice to their clients), one might obtain insight studying the 20 percent and modeling the patterns it presents, and then viewing the remaining 80 percent of items or 20 percent of revenue.

In the case of this data set, 820 of the items in the catalog represent 80% of sales during that year. Therefore, if this number is consistent over time, then smaller warehouses can be added to the company, with 20% of the stocked items, or can be one-fifth the size of the main warehouse, and will cover 80% of sales revenue, with the remaining 20% made up by separate shipments from the main warehouse.

Most Active Client 17841

Of the set of clients in the data set, client 17841 was the most active. Over twelve months, they ordered $38,416 of goods, with 120 invoices, 21,492 units, with 168 units per invoice. They show a clear seasonality with increased orders in October and November, 2011. One idea with modeling data to predict order behavior, is to sequence the invoices into a chain, and record how many items are ordered repetitively versus new. While 17841 does appear to order similar items in general, the choice set is so large that most of the movement from one invoice to the next is different items. The challenge then, is to look at a chain of invoices, in order to arrive at a sequence, where some items are ordered every invoice, some every month, some every quarter and so forth.

It might be the case that the most active clients order too often to have a good predictive model for their t+1, t+2…t+n orders. Medium size clients that perhaps order one time per month, and fewer items in total, might be more repetitive in their behavior. One time order clients are impossible, the only thing we can model is the probability they return at a future time.

Five Questions:

1. Given what you have available, what algorithms, techniques, or models did you use? What are the trade-offs you saw in making your decision? If you had more time, what avenue do you believe offers the most promise?

The data set is promising in a number of respects. First, it is a good representation of transactional data in general. Transactional data can include a range of activities, from a unit of work done, to a bank record, to an order form. This data should exist in large volume for most companies, and should be relatively accessible going back a number of years.

For the questions being asked, the data can be used for part of the inventory question, part of the customer order experience question, and part of the warehouse expansion question. Maximum unit orders over the course of a year can be used to create a first run at an inventory level. That is simply, for each item in the catalog, the stock level should be the number that is the maximum order placed during the course of the prior year. The sum of these levels will provide a total inventory that can be measured in number of units, cost of sales dollar volume, and sales value.

If the model had been run in 2011, given perfect knowledge of what occurred with the half million item orders, then the error rate in which a client would place an order that would be out of stock and require a backlog order delay would be zero. Error rates for 2012 and beyond would be a function of how much variability in the data exists beyond the one year example. We would expect some, given the company may be expanding, and consumer preferences for the items may shift.

Given more, time, I would look at the whole value added chain, and the logistics related to length of time to reorder items, whether minimum orders need to be placed, can some orders from clients be relayed to manufacturers and have direct shipment, and a host of business logistics and environment factors that can speed up or impede the flow of goods from manufacturers through the wholesaler to the set of retailer clients. Also, if we know what the wholesaler’s tolerance for out of stock events is, we could lower the stock values, knowing in most months an out of stock event would not occur. If lost orders exceeds the savings in lower inventory, then the strategy would not be recommended.

For a better order experience, that is a function of being able to predict what will be ordered in the future based on past pattern. In order to explore this the data was sorted by client, and we took the largest client and checked their pattern of ordering. It turns out that the prior invoice of items is not a good predictor of the next invoice. This might be due to this client ordering every few days, and having a large catalog to choose from. If we aggregate the invoices by month, then the month to month prediction is better than the invoice to invoice prediction, but again due to the large choice set the catalog provides the accuracy rate is low.

I think that a neural net model that looks at a chain of orders, ie, not just the preceding n-1 but the n-2, n-3, n-4 etc, that predicts the current invoice items based on the pattern of prior orders for a number of preceding invoices. That is the model I would look at. For one time order clients, the model would not be relevant, and it might be the case that the largest clients are very hard to predict, due to the wide variety of items they order. It may be the case that the most predictable clients are those which are perhaps order one time per month, and tend to order the same small basket of items. Neither the smallest or largest clients, medium size clients might be most easy to predict, and safest to introduce suggestions to encourage them to slowly expand their order baskets.

1. Which features are important in the data that is provided? How did you select them? What justification do you have for their usefulness?

The invoice numbers, stock item numbers and client numbers are very important in order to follow the process of sales going on at the warehouse. The date is important to segment that data into weeks, months and quarters to look for seasonality in the order pattern. Unit price is needed to calculate total revenue by invoice, client, and the company as a whole. Quantity of units is needed to have some idea of fluctuations in inventory stock items. Country of order shipments is needed to determine whether firm’s market is best characterized as local, regional or global. In this case, all of the fields provide important information.

Which features are unimportant? Why?

Descriptions are text strings that correlate perfectly with stock numbers, and

in that sense are redundant. However, the text is needed to have some idea of

what the warehouse is stocking; otherwise it could be parts to a nuclear

reactor. Unit price also tells us information about the type of business this is;

inexpensive consumer items.

1. How did you assess the accuracy of your implementation?

At this stage I have not identified the best model structure for clustering clients into categories, from which one would employ a second model to predict order behavior. One could segment the data set into training and test data, and there are enough records (over half a million) to do that.

1. Do you have sufficient data and information to accomplish what the mock client is requesting?

The data set in isolation of company processes, provides some idea of what is being sold, that there are thousands of clients, thousands of items in the catalog of warehoused items, and thousands of invoices, with frequent activity throughout the year. It does not include enough information about where suppliers are located, and more precise information about where clients are located, to sufficiently model a warehouse expansion scenario. The absence of actually inventory stocks, means while inventory levels can be estimated, they can’t be confirmed to see if real cash flow savings might be realized. An A/B test would be needed to create a static order form, and a smart order form, and then observe order behavior by clients to see if speed of order is decreased and sales volume of orders are increased.

If not, what other data or information would you want to ask for, and why would you want it? In other words, what would having that data or information allow you to do that you otherwise would not be able to do?

Additional Data Needed For Best Outcome

1. Inventory number on company balance sheet. Prefer monthly general ledger values for at least five to ten years, but annual would be helpful.
2. Client addresses to create a dot map of client locations. This would be for the ship to locations of orders. No need to reveal client names, and the longitude and latitude locations can be made fuzzy to prevent easy identification. An example of fuzzy would be to take a set of locations, and randomly adjust latitudes and longitude values such that a shift would take place in any of a randomly selected 360 degrees in direction for about several kilometers.
3. Since most of the clients as of 2011 were in the United Kingdom, a warehouse expansion strategy might look at multiple warehouse locations in England and Scotland, Wales and Northern Ireland if there are sufficient orders occurring in those parts of the Isles.
4. The location of manufacturers that supply and/or are contracted to supply the thousands of items in the catalog. This is needed to understand the time needed to restock based on shipping distance and logistical routing factors.
5. Examples of what order form data look like when received.
6. Time stamps of orders received, orders fulfilled and shipped, and orders delivered to clients.
7. For items, cost of goods sold and sales prices (which appear to vary by client and volume ordered).
8. Do the stock code numbers have a categorical logic to them, either assigned by the firm, or from major manufacturer/suppliers, or are they simply sequential and added as the company evolved and slowly expanded their catalog?
9. Any company wisdom regarding the many factors and steps involved in the operation of the warehouse, fulfillment of orders, and factors leading to order cancellations, returns, and other environmental impedances.
10. Business planning: what are the growth objectives for the next five years?
11. Skype Interviews: Interview C-suite separately to obtain perspectives from production, accounting and sales, interview three to seven warehouse employees, and three to seven clients.
12. Does the client view the impacts of Brexit to negatively affect sales in Europe? Is there a plan to shift focus of geographic market expansion?

What if you cannot get that data? (The client does not have it or is unwilling to supply it.) What can you do with the data that you have, and what are the limitations on what you can do with it?

The existing data does allow one to model the relationship between invoices and items, items with price and quantities, and customers, invoices, and item selection. The data does reveal a lot about the nature and operations of Interesting Goods, and provides ideas about what types of modeling might reveal interesting structural aspects of the business operations. It is not sufficient to create an inventory management tool, order entry tool, or spatial expansion scenario plan.

1. Which of your generated artifacts would you use when collaborating with the development team to implement your proposed solutions? How would you use them, and what value do they bring in using them in those ways?

I would identify three specific equations that can be used to solve the clients three objectives. Next, I would create three flow charts of steps involved in data inputs, model calculations, and estimated outputs. These flow charts would then be converted into three working algorithms expressed as Python code lines, from which a conversion to Scala or whatever program(s) would be used to create workable code. Whatever the best practices are to take a conceptual idea, and translate that to workable code.

Inventory Management and the Warehouse Industry

Academic literature on optimizing inventory flows is extensive and goes back many decades. This literature has been produced from a number of different departments and schools of thought. Some represents application of promising theory, and some has been discredited as examples of poorly thought out central planning distribution systems. Good ideas can be found in Just In Time literature, and there is a large volume of research in academia, government agencies and business industry associations that can be mined for ideas.

Amazon as an example.

Amazon has rapidly grown from 74.452 billion in 2013 to 135.987 billion in 2016. Their cost of sales (cost to obtain the goods that they then resell to their online customers) decreased from 73 percent to 65 percent of revenue. Their inventory increased from 7.411 billion in 2013 to 11.461 billion in 2016. The turnover rate can be estimated as cost of sales divided by inventory. This rate actually increased from 7.3 in 2013 to 7.7 in 2016, meaning the average item sat in a warehouse shelf for 50 days in 2013, decreasing to 47 days in 2016.

This suggests that Amazon freed 630 million dollars from inventory stocks to cash from 2013 to 2016. In their case, as one of the world’s largest wholesalers/retailers, whatever set of inventory management techniques and changes that were applied resulted in a significant amount of cash that can be used for other elements of the business operations.

It should be noted that inventory is a stock value that is reported on a balance sheet, not an expense that is reported on an income statement. As such, decreases in inventory represent a one-time gain in cash, not a decrease in operational expenses.

Noting that a more elaborate review of wholesale inventory metrics over time is warranted, simply applying Amazon numbers to the company Interesting Goods indicates: A 15 million dollar (10 million pound) company, with 65% costs of sales, would have a roughly 9.75 million dollar/7.7 = $1,266,233 inventory during 2011. An improvement from a 7.3 turnover rate would result in a one-time cash savings of $69,383.

Assuming a client willingness to share 12.5% of business gain with their consultant advisor, the maximum fee that this company would likely accept for a custom inventory management strategy based on data science and machine learning application would be about $8,673. At a consulting billing rate of $150 per hour, this project would need to be completed in 60 hours. Therefore, they might not be the best candidate for applying a research based methodology to inventory reduction. An existing, quick to implement solution might be a better recommendation.

Modeling Options:

Review Machine Learning methods (source: <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/> )

1. Linear Regression (supervised learning): Aggregate data set too large for optimization model, would need a heuristic approach and gradient descent. Data set with 8 fields do not provide a dependent/independent variable structure for regression testing. Inventory level = function (starting inventory, predicted orders, distribution of quantities ordered, replacement or restocking time, other…)

2. Logistic Regression (supervised learning): Rather than a positive sigmoid function, a negative logistics curve can be used to model outcomes that fall between 0 (shortage) and 1 (excessive inventory stock). Data would need to be normalized, and risk metrics specified.

3. Decision Tree: Clients seem to know what they want when they place order; not sure how valuable this modeling approach might be.

4. SVM: Support Vector Machine; might be find as a method to identify three to five client type clusters based on the items they order or the mix of items they order, frequency of ordering, and related.

5. Naïve Bayes:We can try this with chaining together a sequence of invoices, but with t-1, t-2,…t-n levels evaluated instead of just prior event.

6. KNN: K-Nearest neighbors; typically I think of this in terms of actual geography, for example, human settlement patterns are found to cluster, sprawl and fragment, and major clusters can be identified (New York City, etc.) For non geographic data, might be some sort of networked distance or groups with relational links. Not sure in the case of a series of invoices; although I do think the clients will be clustered into five major groups: (1) one-time, (2) infrequent and low units, (3) infrequent and high units, (4) frequent and low units, and (5) frequent and high units.

7. K-Means: Probably what I would try first to cluster the data into major client types.

8. Random Forest: Interesting for classifying networks, not sure it would find use here, but that is only my first impression.

9. Dimensionality Reduction Algorithms: If we can obtain client specific data (warehouse client would need to provide their list of clients per terms on an ND), then the information we might mine about them might reveal insight to their buying behavior and increase the predictability of an ordering model.

10. Gradient Boosting Algorithms: first get a model working, then look at procedures to improve model performance.

A. GBM

B. XGBoost

C. LightGBM

D. CatBoost

Additional Modeling Ideas (Ron):

11. Blockchain as a method to aggregate and decompose matrix data sets.

12. Neural Network with Back Propagation, evaluated with Markov Chains to predict probability of clients changing categories.

13. Huff Competition Model Wholesaler Version with Nystuen-Dacey Hierarchical Estimation.

Review Agile Scrum processes:

Pillar tends to favor week long sprints. Team sizes vary, and people may be on several teams at the same time. Flexible and fast moving.

Sprint Planning: assume these occur Monday for half a day during start of project, and then for one hour following Monday morning retrospective.

Creation of Backlog; Product owner after consulting with Stakeholders

Daily Scrum: 15 minute meetings for Agile Team.

Sprint Review: assume these occur generally on Friday afternoons.

Sprint Retrospective: assume these occur Monday morning following Sprint Review

Review Pillar Three Objectives:

1. Create: Create a solution that anticipates where the frontier of inventory management will be in next ten years. Warehouse location strategy designed to outflank anticipated moves by competition, and create flow of goods to create good relations with vendors and suppliers and excellent relations with clients.

2. Transform: Take all the things that our client is doing right, creating value for their clients, and elevating that value by implementing processes to help the warehouse clients, the retailers, create a better experience for their clients the consumers of craft and decoration products.

3. Innovate: Develop a process that allows the warehouse to network with its’ stakeholders, increasing revenue and lowering expenses as a result of greater transparency and interaction. From factory to table.

Workflow Evolution and Migration Path

Ron the researcher will need to develop his skill set and align it a bit away from a specialization in spatial modeling, visualization and interactive mapping, toward parallel skill level with the general Data Science and Machine Learning community, and toward the Agile/TDD processes that Pillar Technology employs. This will allow me to increase productivity as I work effectively with the teams formed to solve various consulting service and product development missions. While developing and then excelling with the set of parallel processes, I will also go orthogonal and perpendicular to the general processes and pathways, which will allow me to identify new opportunities, new processes and new products.

Ron in 2017:

Theory: Academic Journal Articles, inventing new equations, Net Logo

Visualization: ArcGIS Maps, Patterns, Excel 10% View, square cells, and Visual

basic Macros.

Statistics: Stata (<https://www.stata.com/new-in-stata/> )

Coding: Leaflet for interactive mapping, Python only for data scraping

**Ron in 2018:**

**Theory: Google patent search, inventing new equations, Azure Machine**

**Learning Studio**

**Visualization: Microsoft PowerBI**

**Statistics/Coding: Python with add-ons and translations into Scala**

Model References

**MIN-Plus with Blockchain History:**

[**http://niaohe.ise.illinois.edu/IE598/pdf/IE598-lecture13-case%20study%20on%20logistic%20regression.pdf**](http://niaohe.ise.illinois.edu/IE598/pdf/IE598-lecture13-case%20study%20on%20logistic%20regression.pdf)

[**https://bitcoin.org/bitcoin.pdf**](https://bitcoin.org/bitcoin.pdf)

**Cluster and Custom Neural Networks with Markov Chains**

**Neural Networks for predicting consumer behavior:**

[**http://www.sciencedirect.com/science/article/pii/S2212567114004924**](http://www.sciencedirect.com/science/article/pii/S2212567114004924)

**Markov Chains**

<https://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter11.pdf>

**Huff-Global Competitive Warehouse Model with Nystuen-Dacey Hierarchical Assessment (note: in this case consumers are retailers who sell a category of product in a geographic market, which can range from local to global, and stores are warehouses that obtain a competitive market share of revenue as a function of a number of factors including size of warehouse, prices and other factors).**

<https://www.esri.com/news/arcuser/1003/files/huff.pdf>

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.828.6384&rep=rep1&type=pdf>

**Nystuen-Dacey:**

[**https://deepblue.lib.umich.edu/bitstream/handle/2027.42/45977/10110\_2005\_Article\_BF01969070.pdf?sequence=1**](https://deepblue.lib.umich.edu/bitstream/handle/2027.42/45977/10110_2005_Article_BF01969070.pdf?sequence=1)

Other References

Amazon Inventory in 2016: <http://www.nasdaq.com/symbol/amzn/financials?query=balance-sheet>

Bertazzi, Luca, and Speranza, Grazia. Models and Algorithms for the Minimization of Inventory and Transportation Costs: A Survey. New Trends in Distribution Logistics, Springer-Verlag, 1999.

Goddard, Phil, Mniszewski, Susan, Neukart, Florian, Pakin, Scott and Reinhardt, Steve. How Will Early Quantum Computing Benefit Computational Methods? SIAM News, December, 2017, pages 3, 5.

Kaggle Dataset work: Carrie: <https://www.kaggle.com/carrie1/customer-insights>

Kaggle Data Science Survey: <https://www.kaggle.com/surveys/2017>

Kaggle Dataset work: F.Daniel: <https://www.kaggle.com/fabiendaniel/customer-segmentation>

Kaggle Discussion regarding machine learning using Python: <http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/>

Kolda, Tamara G. Sandia National Laboratories. Tensor Decomposition: A Mathematical Tool for Data Analysis. SIAM News, December, 2017, page 8.

Leicester, England best place to site a warehouse: <http://www.indiebeautydelivers.com/guide-to-finding-the-best-location-in-the-uk-for-your-warehouse/>

Nash, Gerald. Let’s Build the Tiniest Blockchain (In Less than 50 Lines of Python): <https://medium.com/crypto-currently/lets-build-the-tiniest-blockchain-e70965a248b>

Github: <https://gist.github.com/aunyks/8f2c2fd51cc17f342737917e1c2582e2>

Petersen, Kaare Brandt and Pedersen, Michael Syskind. The Matrix Cookbook, November 15, 2012: <http://www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf>

Rezaei, Hamid Reza. Designing an Intelligent Agent Using Machine Learning for Inventory Monitoring. 2012 3rd International Conference on Information Security and Artificial Intelligence. IPCSIT volume 56, 2012.

Udoh, Samuel S., Obot, Okure U., George, Uduak D. and Islong, Etebong B. Modified s- S Inventory Model using Artificial Neural Network. International Journal of Scientific and Engineering Research, volume 3, Issue 5, May, 2012.

United States Department of Labor, Bureau of Labor Statistics, Industries at a Glance, Warehousing and Storage: NAICS 493. <https://www.bls.gov/iag/tgs/iag493.htm> .

Wikipedia; Agent-based model: <https://en.wikipedia.org/wiki/Agent-based_model>

Wikipedia; Hamming distance: <https://en.wikipedia.org/wiki/Hamming_distance>

Wikipedia, Minkowsi distance: <https://en.wikipedia.org/wiki/Minkowski_distance>

Wikipedia; Supply chain optimization: <https://en.wikipedia.org/wiki/Supply_chain_optimization>