Project Milestone Report

Joe Marco

September 16, 2017

setwd("~/DataScienceFoundationsCourse/Capstone/Data Story")

### In this Project Milestone report I would like to hit on a few key points:

* An introduction to the problem (based on your earlier Capstone submissions).
* A deeper dive into the data set:
* What important fields and information does the data set have?
  + What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + What kind of cleaning and wrangling did you need to do?
  + Any preliminary exploration you’ve performed and your initial findings.
* How do I plan to move forward?
  + What approach am I going to take?
  + How has my approach changed since the beginning of this course?

### Let's revisit my dataset to provide some context

* Place yourself in a large coorporation's shoes for one second. It is critical that you must create, maintain, and deliver products (and or services) to your consumer to stay in business. My data is strictly focusing on "goods" or "products" and their associate key details:
  + Skus (unique product codes)
  + Product lead time (how long it takes from order to delivery of a certain product)
  + Product inventory (how much do i have in stock ready to ship to customers?)
  + Past sales and forecasted sales (so i can view and predict my revenue for my business)
  + etc.

### Being a business owner is no easy task, but product backorder issues directly contribute to lost customers

* Think about it, you find a product that you need or want, you go to order it and the business explains you will have to wait days/weeks/or months to get that product. You likely move on and find the same or similar product from another competition, or decide it is not really worth the purchase.
* The problem/opportunity we set out to overcome here is around the products within our business that have gone on backorder, and what might be causing this based off their specific characteristics
  + based off all of my product data can I predict if a product is going to go on backorder based off some similar associated entities between them?
* Value proposition behind this initiative
  + risk avoidance
  + brand loyalty
  + improved customer engagement/experience

### Now that we understand the context of what we are setting out to achieve, lets take a closer look at the data

* Recall we had all of the following characateristics for each product (sku) within our dataset, the items highlighted in red are the fields I have personally chosen to focus on while trying to achieve the above business proposition:
  + national\_inv - Current inventory level for the part
  + lead\_time - Transit time for product (if available)
  + in\_transit\_qty - Amount of product in transit from source
  + forecast\_3\_month - Forecast sales for the next 3 months
  + forecast\_6\_month - Forecast sales for the next 6 months
  + forecast\_9\_month - Forecast sales for the next 9 months
  + sales\_1\_month - Sales quantity for the prior 1 month time period
  + sales\_3\_month - Sales quantity for the prior 3 month time period
  + sales\_6\_month - Sales quantity for the prior 6 month time period
  + sales\_9\_month - Sales quantity for the prior 9 month time period
  + min\_bank - Minimum recommend amount to stock
  + potential\_issue - Source issue for part identified
  + pieces\_past\_due - Parts overdue from source
  + perf\_6\_month\_avg - Source performance for prior 6 month period
  + perf\_12\_month\_avg - Source performance for prior 12 month period
  + local\_bo\_qty - Amount of stock orders overdue
  + deck\_risk - Part risk flag
  + oe\_constraint - Part risk flag
  + ppap\_risk - Part risk flag
  + stop\_auto\_buy - Part risk flag
  + rev\_stop - Part risk flag
  + went\_on\_backorder - Product actually went on backorder. This is the target value.
* Although all the fields, in their own way have value to multiple business inquiries, the highleted red above are the fields I identified as most important to guage causation of backorders, and ultimate predictions based off inventory levels and forecasted figures.

### Limitations

* Now that we have an idea of the data attributes most important to the business case at hand, we can talk about some items that we CANNOT do with the data set given:
  + We do not have dollar sales figures, meaning no way to guage potential impact of revenue/profit based off backorder issues
  + We do not have "Source (supplier)" information to be able to determine which suppliers by name are contributing to backorder issues, although we could likely find some commonalities to assume which skus come from the same source

### Cleansing and Data Wrangling

* This dataset was provided by kaggle.com, by default the data itself was not as messy as we would see with real world data.
* I did some checking on unique skus and total counts of records to understand how much work needed to be done to the data based off trying to have a unique record for each row of the data
* I updated some field names for common terminology sake for future reference

### Exploration

* At the beginning of this Capstone, I thought I would be using all the data to narrow down the business insight I was trying to get to, however my recent explorations within the data have been narrowing to some of the following:
  + Comparing entities that could have some relations (sales, forecast, inventory, lead time, etc.)
  + Focusing in on the backordered list of products
  + Probability around backorders based on existing sku lists
  + Focusing specifically on inventory levels and lead time to see if these entities have major impact on how a product handles whethere or not it is at risk of backorder

### Moving Forward

* My planned approach moving forward is to really place my focus in the fields mentioned above in red and see if this is enough data and strong enough data to answer the business insight I am after (Can i figure out which products are at risk of backorder, and if so -- what can i do to prevent them from going on backorder)
* This plan has changed slightly since the beginning report as I originally planned to investigate all the skus and drive some insights, but my current plan is to really focus in on those products that went on backorder and try to drive out root cause of backorder and predictions if i have a new sku with some similar values. I also originally planned to try to drive out sources (suppliers) that were causing the issues with my sku backorder... While this is important, it is not nearly as important as figuring out what EXACTLY is causing my backorders (long lead time, too little inventory, etc.) as these are the problems I could have with any source, and a model that focuses on the larger problem would be more beneficial to the business group(s) outlined in the original proposal.