VJ Davey

CTBA Fall 2016

Lessons Learned Assignment

**Lectures**

* **Introduction to Business Analytics**
  + Work is never really completed. There is usually going to be a cyclical pattern as data must be cleaned and prepared, models need to be made that accommodate the data, and there needs to be evaluation and understanding of how the data can answer a business question.
  + Sometimes the question may not even be the right question to ask. Sometimes there may need to be a complete reevaluation of the problem at hand and the current model may be useless. Data may be useless if it is found to be lacking information that may be necessary when the problem or models must be changed as the process goes on.
  + Information can be drawn from just about everywhere and even things as simple as item placement on a shelf in a store can be examined and found to have drastic impacts on sales.
  + Good 80% solution in 20% of the time. Getting a sufficient answer is more important than getting the exactly right answer. Prototyping approach.
* **Pareto Analysis**
  + Large % of problem, difficulties or efforts caused by just a few of the possible root causes.
  + Sort items in order of greatest to least in terms of the metric of interest in order to do a Pareto Analysis. Calculate the percentage each item constitutes of the total, and compute the cumulative percentages following the established order. Plot on a column bar chart the data for each item, and mark the cumulative percentages with a line chart.
  + Excel tips:
    - Jump to end of page: End+ Down Arrow
    - Select all: SHIFT+CTRL+Down Arrow
    - Use Dollar Signs to freeze cell references
    - Double click lower right corner of a cell to copy a formula down to the cells beneath it
* **Data**
  + Ethics of web scraping. Automated scraping can crash or overload servers. Some websites will opt to create an API for people to use so that their servers aren’t overloaded by scrapers (e.g. Amazon)
  + Document Type Definition (DTD) is used to enforce standards with XML. It specifies required or optional elements and the allowable number of sub-elements.
  + Not all data is freely available. Some data is made in house, some will need to be purchased from other firms, some can be found on the internet. It is best to marry these sources to increase the value of analytics results.
* **Visualization**
  + Be careful of people who lie with graphs. The information presented can be accurate and true, but the manner of presentation may be manipulated to give readers an impression of a dataset that is exactly opposite of how they should be interpreting it. (e.g. truncated axes, flipped axes, things not drawn to scale, using the wrong kind of graph, mapping unrelated things on the same graphs to imply correlations where none exist, things which don’t total to 100%, etc).
  + Word clouds seem to be more style than substance
  + Matplotlib and Seaborn
    - Interesting packages. Will probably be useful through this program. Will probably want to get my hands dirty with and experiment with when I have more free time.
  + NASA – Listening to data
* **Metrics**
  + Has its importance with all facets of the business (HR, Marketing, Corporate, Quality Control, Supply Chain, Product/Service Development, etc.)
  + Many common metrics can be analyzed and evaluated over time and better decisions can be made or guided in an automated fashion through data analytics.
* **Confidentiality, Data Security, and Ethics**
  + Analytics is incredibly powerful. Data and technology can be used to find correlations and patterns which professionals wouldn’t even pick up on so easily.
    - Target Pregnancy
    - Diagnosing patients with diseases
    - Routing from origin to destination
  + Data can permit invasion of privacy
    - Problem with who owns data. Especially with cellphones.
    - Should companies have the right to sell data? What about protecting the data they have from hackers?
    - What about information a company learns about you from running analytics? Sometimes a company may actually be able to use patterns to discover things about you that you don’t know about yourself.
    - Uniquity
      * Only 4 data points are needed to locate a unique person in a supposedly anonymous dataset.
      * Hiding identifying information may not actually protect a person’s identity as that person most likely has a unique pattern in the dataset beyond their ID number. E.g. shopping or travel “fingerprints.” These can identify a person much like forensics can identify someone.
* **Internet of Things**
  + It has become less expensive to have things connected to the internet as time has gone by. Things including cars, phones, toys, and wristbands are all now connected to the web and feeding in data that can be analyzed.
    - Possible ethical concerns can arise. Some products like toys are meant for children, and as a result, lots of information from the unsuspecting can be collected with such products. Questions then can arise about what a company will do with that data, or who they may sell that data to.
    - Disney Magic Bands
  + It is cheaper than ever to build products that can sense changes in their environment and output information to the web, or be used to control things like a robotic arm. Raspberry Pi and related products.
  + Industrial Internet of Things (IIOT)
    - A company can feed information to all of its service machines and have a sort of group learning take place. When one machine fails, they can simultaneously have all the others know how it failed, so they know what to avoid or look out for with the other machines.
* **HR and People Analytics**
  + Use of analytics for probability predictions. HR and Customer retention.
    - Decision tree models. Can be easily implemented with a program and simulations can be ran to approximate probabilities of a person leaving.
    - Algorithms which group and cluster individuals may not be exact, but they can be very, very close.
  + Markov Models
    - Another way to represent the probability of something happening. A person or thing can be in one of several states. We try to determine the probability that a person switches from a given state to another. Memoryless. All probabilities for the switches together must total 1.
* **Verizon – Michael Antony Raj**
  + Done is better than perfect. Getting results is more important than extreme accuracy.
  + Focus on the right problem, be ready to question results and alter them afterwards. Focus on logic and identify the relevant data. Be clear, simple and precise in your communications.
* **AidData – Dan Runfola**
  + Using data analytics to determine where and why aid works or doesn’t work.
  + Computational Geography Methods.
    - Ideas and information might be there, but some of the data might be useless. Satellite camera data dates back to 1970s, but due to differences in image quality, not as much information can be gleaned from the older pictures if certain questions are asked (e.g. are there trees in area X?)
* **Capital One -Salil Singhal**
  + Location Analysis as an important tool
  + Can be used to allocate limited resources (e.g. dividing up multiple stores to several managers)
  + Identifying possible store locations.
    - Gravity Model
    - Piggybacking off of established shopping centers and setting up in areas which already attract the target client. Aim to make banking easy by setting up where the customer will be
    - Paying for store locations through rent instead of outright owning, to allow for an easy back out if something goes wrong.

**Books and Articles**

* Automate This!
  + Algorithms are everywhere.
    - When used properly, their use with machines or software can be equivalent to incredibly cheap labor from a worker that is always giving maximum effort and has no need to sleep or take breaks.
    - Peterffy and Wall Street in 1987
    - Can extend to pretty much every part of a person’s life.
      * Medical Diagnosis
      * Sports
      * Geographical networking
* 7-11 Japan Case
  + Different kinds of analyses and sources of data a company can use to their advantage: point of Sale data, trends on item sales, scrap trend analysis, stock-out ranking, weather analysis, shipping analysis, location-trade area analysis.
  + Benefits from heavy analysis of company operations included:
    - Higher inventory turns
    - Decreased shipping costs
    - Maintained sales and market share in times of economic decline
    - Decision making benefits (knowing when to discontinue selling a product, knowing when to create a new original product, knowing when to put certain items out for sale, how to arrange items in the store, knowing when and where to open new stores, etc.)
* Optimization Case
  + Companies may adhere to business practices that aren’t exactly the most beneficial for them in the long run. Different sectors of a company may focus on their own work and their place in the system without thinking too much about how they effect the overall production rates for a company.
    - Minimizing cost at every sector or stage of a products development may not be the optimal solution for some situations.
    - Hierarchical decision making processes have an important role in separating decisions made regarding capacity and inventory.
  + Tradeoff between capacity and inventory. Having a fixed capacity can hurt in the long run. Problems could include suboptimal ROA and high inventory hold costs.
  + Return on Operating Assets (ROOA) and its use as a metric of how well a company is performing with a certain level of capacity.

**My Work**

* xmlAssignment.py
  + Pretty intuitive and easy to grasp.
  + Introduced to lxml and BeautifulSoup parsers
* htmlAssignment.py
  + Messier than the xml and json assignments. HTML data can be a pain to navigate and collect. Need to carefully examine HTML structure more so when scraping than the others.
* jsonAssignment.py
  + Intuitive and easy to grasp much like XML.