CREATING AN INTEGRATED, ANALYTICAL SYSTEM FOR THE MODERN, DIGITAL BUSINESS

by

Colby Ford

A practicum submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Data Science and Business Analytics

Charlotte

2015

Copyright © 2015
Colby Ford
M.S. in Data Science and Business Analytics

ABSTRACT

"Big Data" is a current buzz word that many computer scientists, statisticians, and business professionals try and wrap their head around. Insight into a company's performance is crucial into understanding the next steps a company should take. However, this becomes increasingly difficult given the large amounts of data that a company may harbor with hopes that some of it will be useful.

I use a cloud-based solution to help automate the data mining process and expedite the path from analysis to discovery. By automating the process, future redundant analyses of new data takes far less time to perform and leaves analysts with the freedom for other data science endeavors. The key to this solution is that the components are integrated, meaning that human interaction should only happen at the beginning (to set up the desired analysis) and at the end (to interpret the result). Therefore, the previously individual links now form a cohesive chain that pipelines data into the processed state.

ACKNOWLEDGEMENT

It is with upmost gratitude that I acknowledge and show appreciation to Mariner, the company with which I studied to form the research around this document. Without the support of Phil Morris, Joe Guy, Peter Darragh, and the rest of the Mariner team, this project could not have been completed. I hope this study shines light on their motto (the dictum for this study), *Insight to Achieve*. Again, thank you for your support.

DEDICATION

I would like to dedicate this practicum and the research behind it to the 2014 National Blue Ribbon School, Caldwell Early College High School in Hudson, NC. Not only are they making a difference in the lives of students every day, they made a big difference in my life as well. As both as a student and an employee at Caldwell Early College High School, I have always received ongoing support for my education and personal achievement. I hope that my progress through higher education will be an inspiration for future students at Caldwell Early College High School to be Ready for College, Ready for Career, Ready for Life! Thank you, Caldwell Early College faculty and staff, for all that you do!

TABLE OF CONTENTS

INTRODUCTION

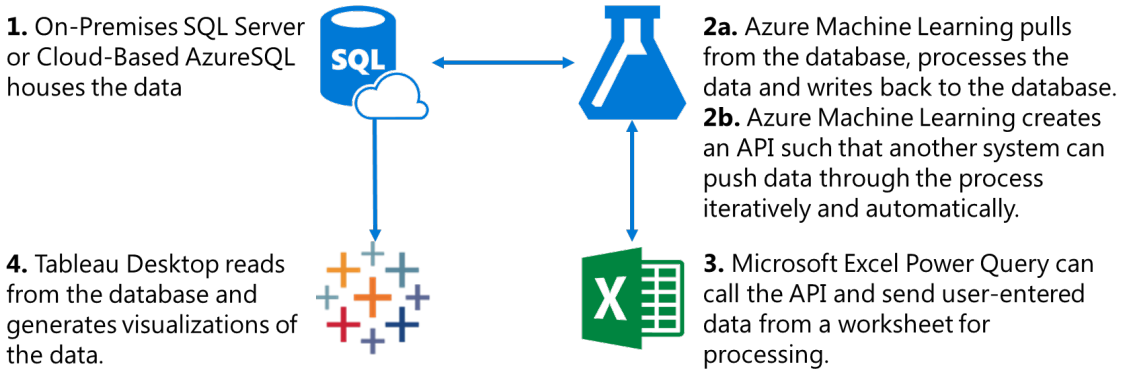
Thomas H. Davenport, an author and distinguished professor of Information Technology and Management at Babson College, describes analytics as having three distinct stages, namely Analytics 1.0, 2.0, and 3.0. Analytics 1.0, according to Davenport, is the "descriptive" stage where 2.0 and 3.0 are "predictive" and "prescriptive", respectively. He continues on by saying that companies usually fall within one of these three stages. That is, are they simply describing what happened, predicting what will happen, or prescribing actions to affect the future (Davenport, 2007)?

Many companies today are only beginning to get into that Analytics 2.0 realm and few are in the 3.0 stage. One foreseeable road block is the "silo" nature of business intelligence systems. A data warehouse, its connections and software, and people that manage it are completely separate from the business intelligence team that needs to pull data and maybe process the information through a statistical platform. Having the systems separated in distinct sections is detrimental to the flow of analytical processing.

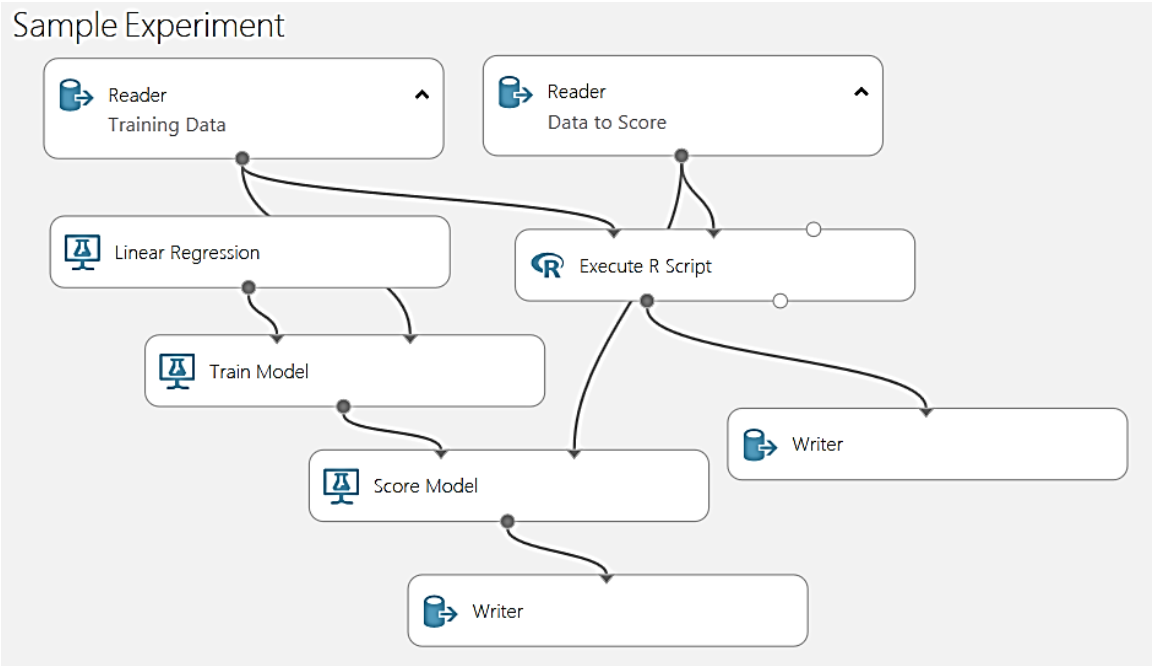
To combat this, I have used the Microsoft cloud solution, Azure, to do the heavy lifting for me. Azure includes many facets from virtual machines to storage to machine learning and more ("What is Microsoft Azure?", 2015). I use their machine learning system to process data in an integrated fashion to reduce time from data input to result output.

BACKGROUND INFORMATION AND PROCESSING DATA USING THE CLOUD

Data processing in the cloud using a solution like Microsoft Azure allows for a singular place for analysis to occur as well as results retrieval. The general process is as follows:



In the Azure Machine Learning environment, the analyst specifies the type of analysis to occur on the data. This ranges from prebuilt regression and classification nodes like linear regression and k-means clustering to custom nodes where custom R or Python scripts can be written.



In the above example, training data would be read from a database using the Reader node, then either the prebuilt Linear Regression node or the custom R script will run the regression and score the test data. Once the new data are score, they are then written back into the database with results using the Writer node. There is also the option to generate and API and attach Web service Inputs and Outputs instead of Reader and Writer. APIs can be called via external sources to run the machine learning experiment.

CASES

Case 1: Prediction and Classification of Utility Meter Performance

The largest energy provider in the South East has an issue with the performance of certain meters at customer locations. Most power meters perform as expected (reporting accurate usage), but there are some that report zero usage or have inflated or deflated reporting. This fault is serious on both ends of the spectrum. If a customer's meter is incorrectly reporting usage higher than what they actually used, it will take extra effort to issue a refund for the error. If the meter's report is incorrectly lower than it should be, the company is losing money at first, but the customer is still responsible for the charges once the error is caught. This leads to lower customer satisfaction. Moreover, incorrect reporting by faulty meters will throw off expected power consumption. In widespread, extreme cases, this can lead to a power shortage or power excess because of the miscalculation of projected power demand.

Currently, the regression and classification has been designed, coded, and implemented by another employee of Mariner, Wayne Snyder, a data analytics architect. He has built the system using the entire Microsoft SQL Server stack. SQL Server Analysis Services is the center for the calculations and then reports the classification for each meter back to the database. The issue discovered was that the analysis took a very long time to run. To help speed up this process, I have transformed and replicated the analysis process in the cloud with Azure.

In this utility example, energy, data is collected about energy consumption and weather. Weather is measured in heating and cooling degree days, which is a measure of

how far above or below the temperature is from a normal comfort level ~65°F (Satel-Light, "Heating Degree Days").

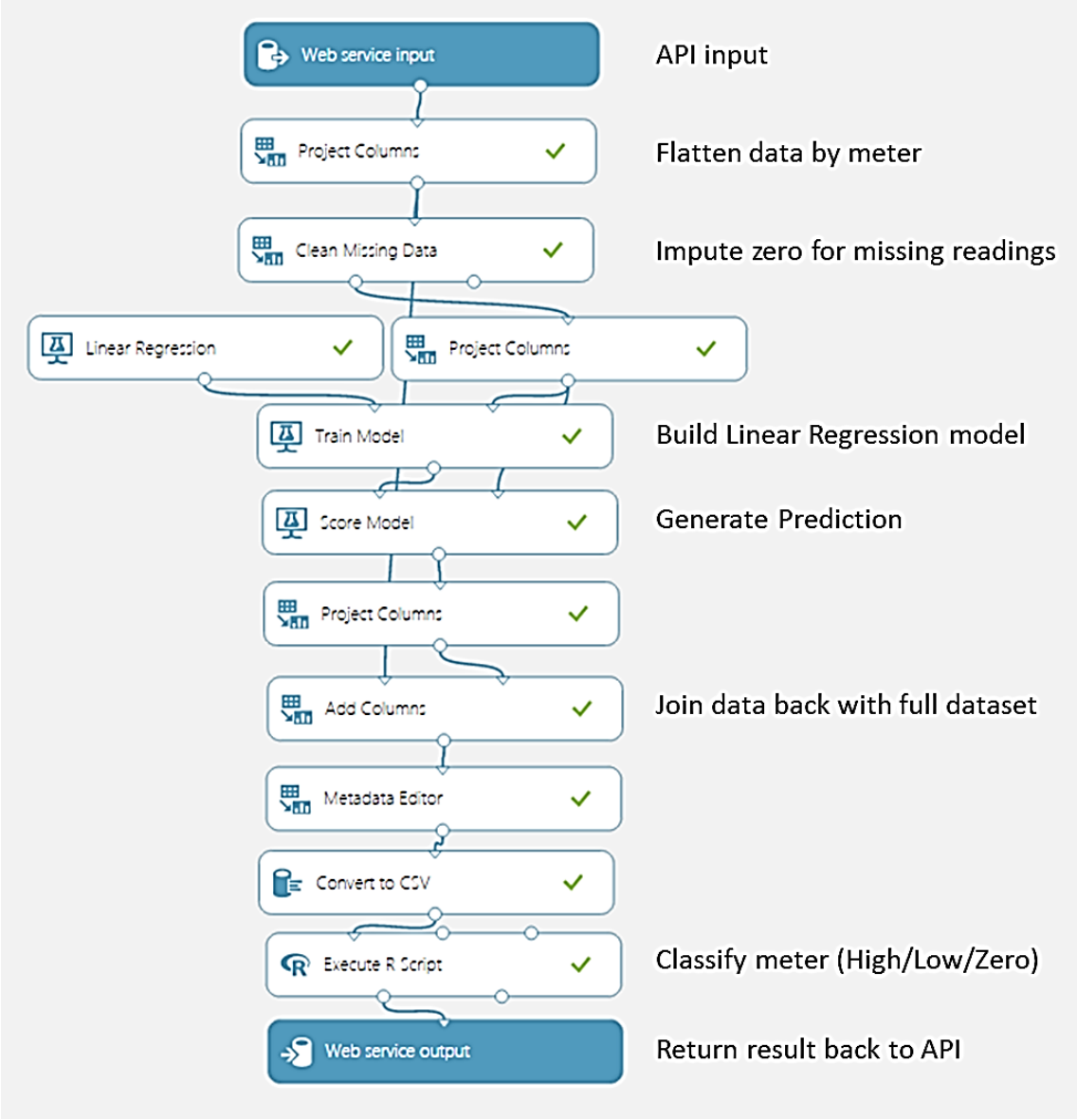
The data is sent to the AzureML system in the following format:

MeterID	Time	Usage	CDD	HDD
1	1	57	2	6
1	2	62	2	5
1	…	…	…	…
1	n	89	3	2

Then, a linear regression is built per meter ID to predict what the usage should be in current time period along with a confidence interval of an acceptable range. If the actual value is above or below the predicted range, the meter is classified as high or low, respectively. If the actual value is zero, the meter is classified as nonfunctioning.

Sample R Code for Classification:

```
#Calculate Standard Deviation by oeter (one meter per row)
data$UsageSD <- apply(usage,1,sd)
#Calculate 95% Confidence Interval
data$HighBand <- (data$Usage_Per_Day_Hat + (2 * data$UsageSD))
data$LowBand <- (data$Usage_Per_Day_Hat - (2 * data$UsageSD))
#Classify each meter (Binary Result)
data$IsActHigh <- data$Usage_Current > data$HighBand
data$IsActLow <- data$Usage_Current < data$LowBand
data$IsActZero <- data$Usage_Current == "0"
```



This prediction and classification model generated an API that can then be called from a C#, R, or Python script or from within Microsoft Power Query. By doing this, the classification process is dynamically integrated into the database system, which decreases the time it takes to get from data collection to an analytical result.

Case 2: Forecasting of Inventory Demand

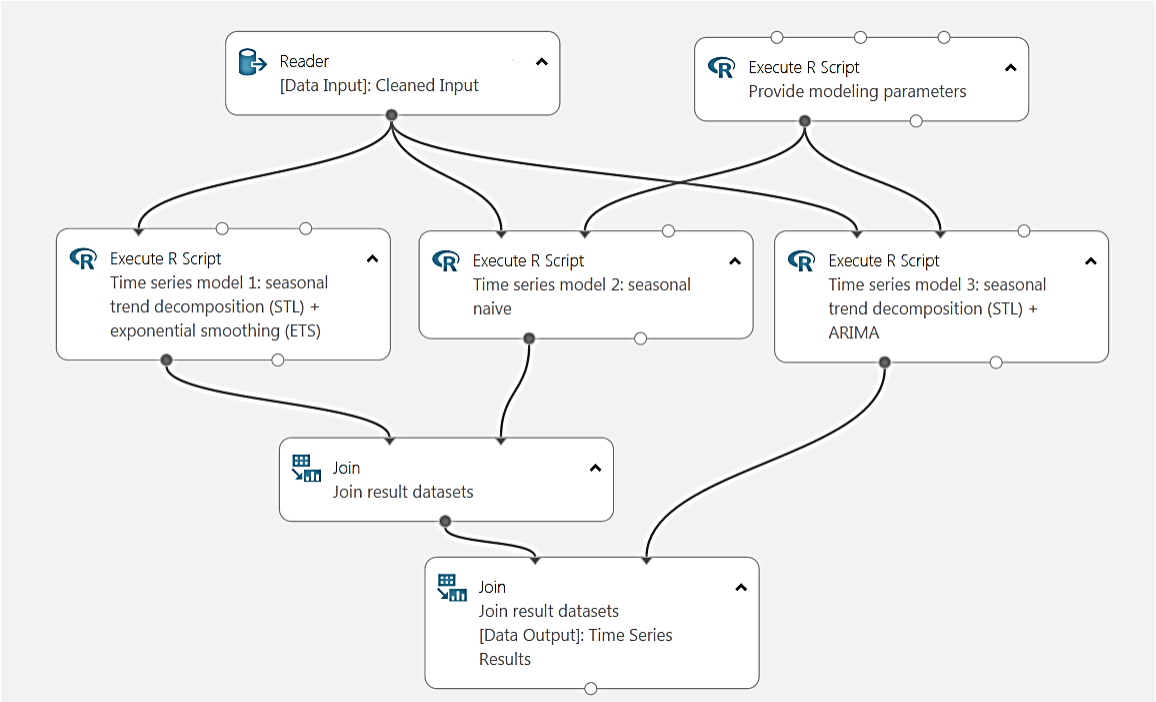
Note: This case study was funded by the Microsoft Internet of Things project.

A household hardware design company purchases its inventory based on sales reported from big box retailers. Currently, the company has information about how many of each of their items was sold from each of the retailers. Using this, the company has a static forecasting method to help predict how much stock they need to order from their manufacturer in the future.

To help increase the accuracy of the prediction, I used the Microsoft Research team's multi-algorithm approach. Editing the code provided by them, I fit the script to the company's data structure. The multi-algorithm approach generates an exponential smoothing model, a seasonal ARIMA model, and a naïve Bayes model, compares the accuracy of each of the models' predictions, and selects the best model for that data. In the future, this case will be expanded to automatically compare the results and see which models prove to be more accurate over time. Then, the stability of this solution will ensue. Also, this case will be implemented to write back to their data warehouse such that the purchasing team can pull the results and make immediate ordering decisions based on them.

The desire for a cloud-based solution increases as the complexity of the predictions can overwhelm human workload with hands-on calculation. For example, not every item will have the same ordering period. That is, one item may require a 3-month lead time while another set of items require longer. This is complicated for a standalone forecasting system to handle since these variables are static and are required before

analysis can take place. By using the R scripts in AzureML, the lead times are included in the data as a variable and can be passed through as the calculations will be done per item and not as one lump of sales. This way, all the predictions can be rolled up together to predict sales demand by category or for the entire company.



This approach is dynamic in that the machine learning system may choose a different model today from the model it will choose next quarter. A better prediction will lead to a more accurate order from the manufacturer, minimizing excess stock (that would otherwise have to be sold at a drastic discount to get rid of) while still having enough stock to meet consumer demand and therefore improving company profit.

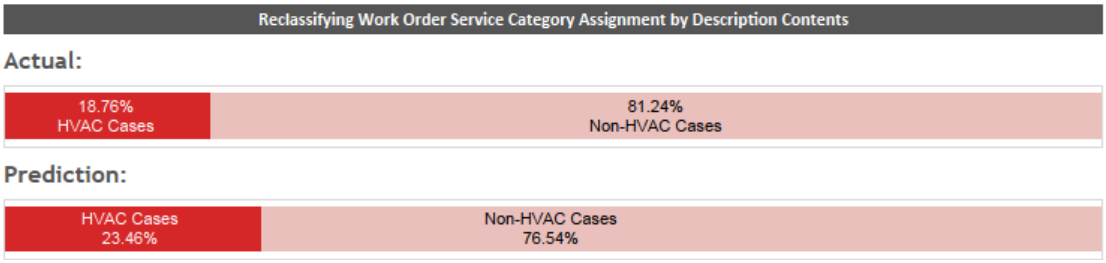
Case 3: Textual Analysis and Reclassification of Work Orders

A large, global real estate management company receives thousands of work orders every day from the inhabitants of its buildings. Each of this company's client's buildings use separate systems for each of preventative and reactive work orders. Note that preventative work orders are those that are scheduled in the system to maintain the operability of the building's assets. Reactive work orders are those that an individual may submit into the system manually when a problem arises. Both types of work orders could have to do with HVAC, locks and keys, security, plumbing, etc.

Since reactive work orders are those that are entered manually by one of a multitude of users, some errors or inconsistencies are expected. However, this company has experienced a grave misclassification of work orders in both location and maintenance category. That is, a user may put their work order in as {location: general, category: other}, but type in the {description: "It is very cold in my office on level 5."}. This should have been entered as {location: level 5, category: HVAC} by the enterer.

Using the AzureML system, I pulled data from the SQL Server and ran it through the tm package in an R node. The tm package is for text mining documents in R. In this case, the "documents" were the individual work order descriptions. This generated a term-document matrix, which then reports terms to be sorted and grouped by category. For correctly classified work orders, I used the terms found by the text mining to reclassify the work orders that were not classified.

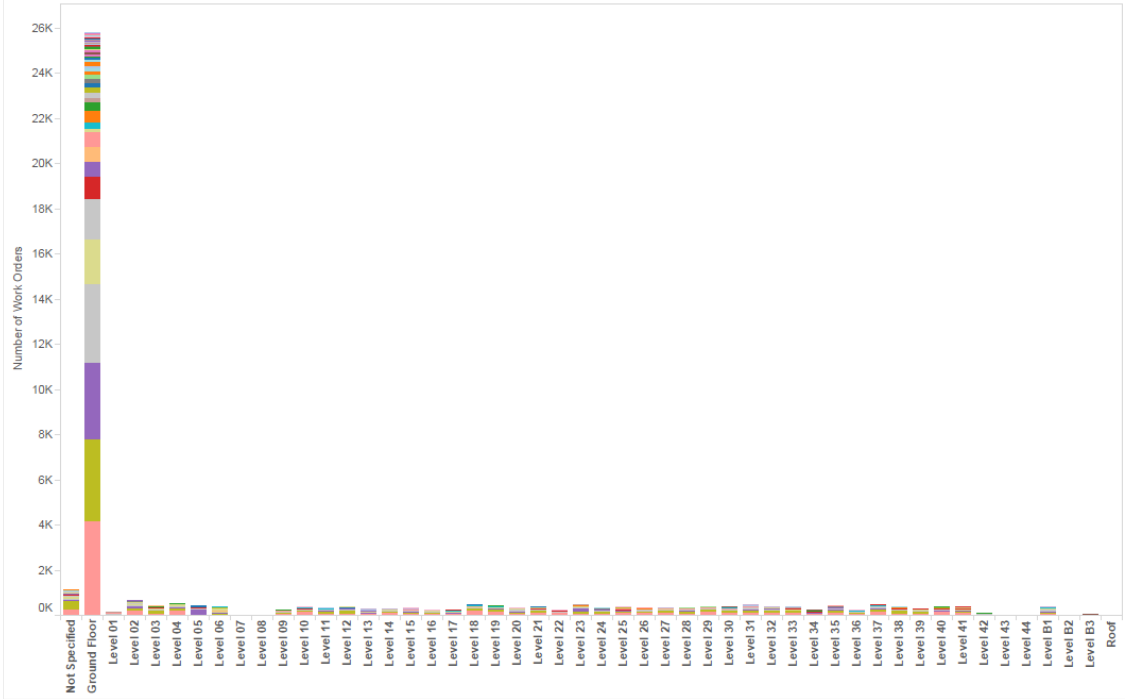
Sample results from reclassifying using HVAC terms:



The results were interesting in that by just using a few terms that the text mining picked out, the machine learning system reclassified ~5% of the work orders, which is over 1,000. This result may be even more notable with more training data or by including more terms.

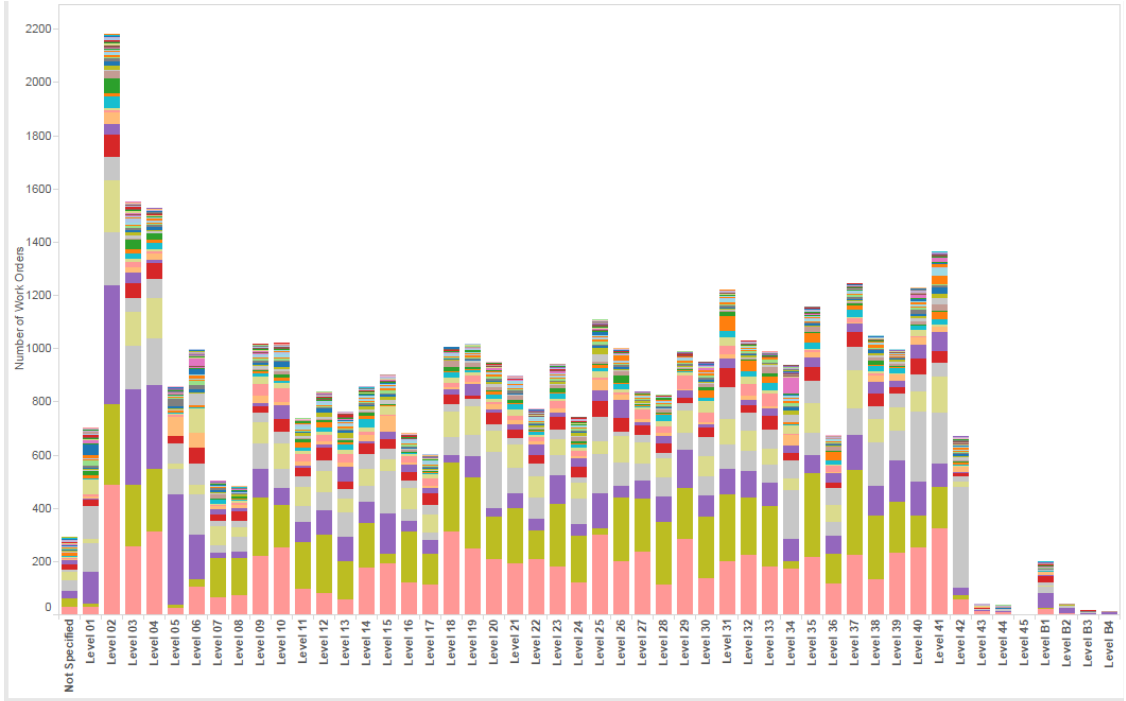
For level information, I used a substring script to pull out the word "level" and then three characters after it. (For example, running the script on "It is too hot in the hall on level 14." would return "level 14".) Note that all descriptions were tokenized and the letter cases were all changed to lower case. This way, "Level 3!" returns a similar result as "level 3".

Before reclassification:



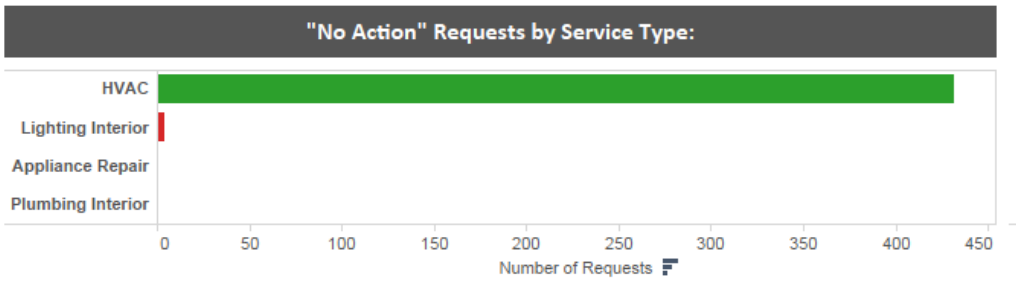
(Note that most all work orders are under “Ground Floor”, which is the default selection.)

After reclassification:

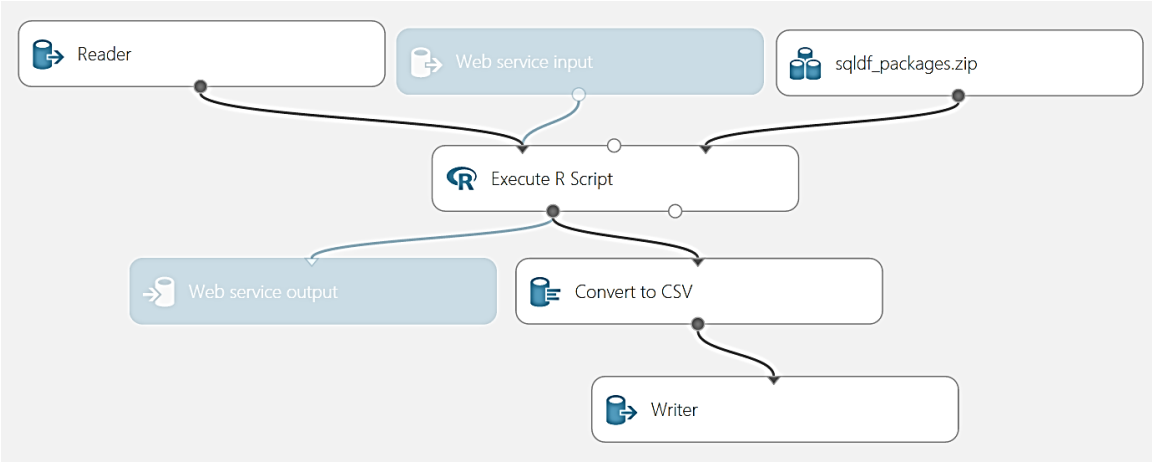


(Note how much the distribution evened out.)

Finally, after reclassification, I noticed that there were many comments left by the maintenance person that were eluding to “no action necessary”. After reading through some of the maintenance comments and noticing that there were quite a few where nothing was done once the maintenance person got to the service location, I decided to run the substringer on those common phrases to see if there is a trend between requestor (the person who put in the work order) or level or service category (HVAC, plumbing, etc.) and the number of “no action needed” comments by maintenance personnel.

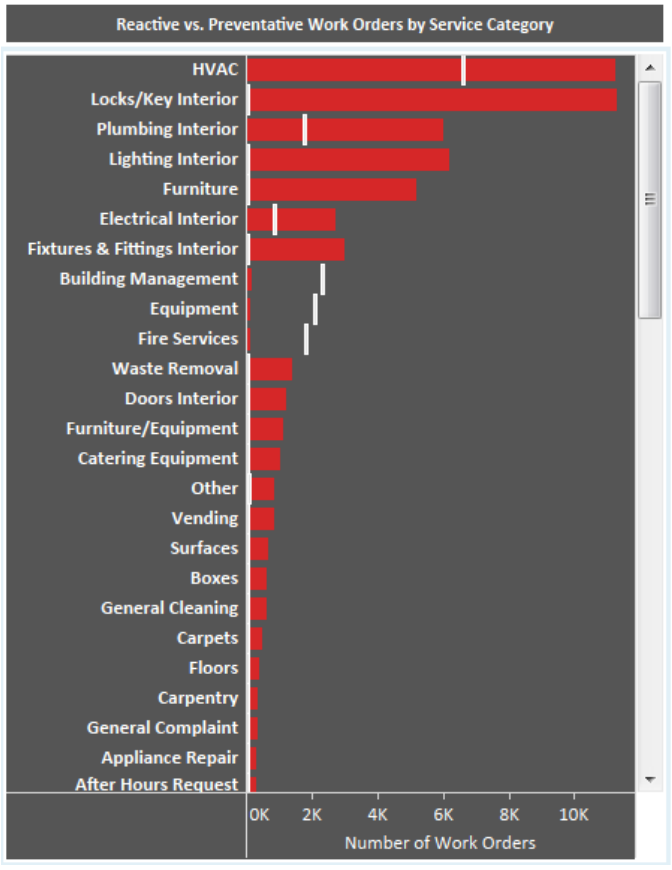


This uncovered that there is a huge issue with HVAC work orders being unnecessarily opened and where money is spent on sending maintenance person to unnecessarily investigate.



Here, I use a dual approach for classification. Notice from above that the Reader node is used to pull current work orders from the SQL Server as well as the Web service Input is used so that the work order system can call the API for new work orders.

Visualizing the relationship between reactive and preventative work orders was also important to this company. Preventative maintenance is scheduled such that, in theory, this will reduce the spontaneous issues and therefore reduce reactive work orders that are requested by the individuals working in the building. From the analysis of preventative versus reactive work orders, it was obvious that there are areas of improvement between efforts spent on certain areas in regards to preventative maintenance.



The white line represents the number of preventative maintenance work orders and the red bar represents the number of reactive work orders. Note that there are some areas

where reactive greatly exceeds the preventative (HVAC and Plumbing Interior, for example) and there are some where it would seem that there is too much preventative maintenance (Building management, Fire services, and Equipment). It would seem that money for preventative maintenance should be better appropriated to fit the needs based off of the number of reactive work orders.

By bringing attention to and correcting such issues, the company will spend less time, money, and resources figuring out where work orders are being requested, what for, and what is being done once the maintenance person gets there; thereby ensuring they are analytically equipped to staying analytically informed about their company.

Note about the Software Used in Cases

Mariner (mentioned in the acknowledgement) is a Microsoft Gold Partner in Data Analytics and a Silver Partner in Intelligent Systems and Data Platform. They are also a partner with Tableau. The software solutions used in the cases come from either Microsoft or from Tableau. There is no personal endorsement specifically for these products unless otherwise noted.

CONCLUSION

Business is no longer solely about marketing, products, services, and customer relations. Being informed has become increasingly more crucial. To do this, companies are rapidly hoarding data about every customer, transaction, and contact in hopes to have the data necessary to possibly tell us something. However, many companies do very little with their data.

This is a sad thought since, according to a Harvard Business Review, companies that make data-driven decisions outperform competitors by being 5% more productive and are 6% more profitable (McAfee, "Big Data: The Management Revolution"). It seems that this is an easy win to simply be more data-driven.

The problem comes into play when a company fails to know how to use the data or at least use it in a timely manner. If the marketing department needs to know about customer segments, but they needed to know 3 months ago when the design team was creating the advertisements, slow data analysis is useless. This is why creating an integrated, analytical system for the modern, digital business is so important.

The cases aforementioned illustrate the immense potential of using a cloud-based solution for rapid processing. Reducing the time between data retrieval to analytical processing to results can be a great game-changer in the business world. From a prediction and classification to quick visualization, using tools that allow for insight into your data can open your eyes to an entirely new frame of mind about decision making in your company; and it is as simple as connecting the flow of data from input to results.

REFERENCES

Davenport, T., & Harris, J. (2007). Competing on analytics: The new science of winning. Boston, Massachusetts: Harvard Business School Press.

Heating Degree Days. (n.d.). Retrieved April 22, 2015, from http://www.satel-light.com/runlib/soda/hddhelp.htm

McAfee, A. (2012, October 1). Big Data: The Management Revolution. Retrieved April 22, 2015.

What is Microsoft Azure? (2015). Retrieved April 18, 2015.