October 26, 2019

To: Danielle Sherman, CTO

From: Kevin Burr, Analyst

Subject: Product Profit and Customer Preference Predictive Analytics Lessons Learned

Danielle,

Here is a summary of lessons learned from our recent tasks using R for predictive analytics.

**Lesson: Getting Started with R**

This activity created a foundation for core skills in R to run a simple predictive analytic pipeline. This lesson was particularly useful because we ended up with a baseline pipeline from which we can work every analytical inquiry going forward. While simple in its functionality, the pipeline lays out the standard flow for best practice approach to data science problems. This can, and was, embellished with more complexity as we worked through more complicated examples.

Three lessons learned from this activity:

* Google has the answers – when there are errors that are non-obvious, use Google. Search for the exact error message. Most likely someone else has encountered this issue and posted a solution.
* Take your time – many mistakes I make or overlooked in this example were a product of rushing. If you take your time and proceed carefully, following all the best practice steps and really reading to understand the issues, R is not really that complicated.
* Spend time getting to know the language – R is really powerful and flexible. There is a wealth of information on the internet. Take the time to learn the packages you apply before and as you apply them.

**Lesson: Predicting Customer Brand Preference**

This activity created a solid understanding of the methodology for predicting customer preferences. Using the C5.0 and Random Forest models to predict how customers would have responded to the customer survey was a great way to learn how to run models in R and tune them. This process is also useful in predicting other attributes in other use cases where the output is a factor, or binomial. This continued to build upon the standard pipeline, and added some complexities to the outline that can definitely be leveraged in future projects.

Three lessons learned from this activity:

* The outputs from the models in R can be quite verbose, but if you take the time to really understand it, there is some really insightful information to help drive to conclusions. I found the Variable Importance function particularly useful. I thought of it as a much easier why to get to an understanding of what attributes are driving outcomes, something that took a lot more effort in creating and assessing decision trees.
* Resamples is a great function – before I even got to the point of executing this function I was wondering in my head how I would collate all the model output from each model tested to easily compare and contrast model performance. This function is fantastic since it outputs comparative analytics side by side for each model. This is a great snapshot for performance evaluation.
* Don’t make assumptions and be sure to run the data out of the box before deciding on a model. It’s possible out of the box is accurate enough. Out of the box is also a great baseline to determine whether or not any data cleansing or engineering is making the potential outcome of the model better or worse.

**Lesson: Multiple Regression in R to Predict New Product Volume**

The most useful aspect of this activity was running multiple algorithms. This activity reinforced the process, best practice, of the standard pipeline. Using multiple algorithms to predict product volume for new product introduction is also practical.

Three lessons learned from this activity:

* The pipeline remains the same – no matter which algorithm we needed to run, nor the general problem or use case we were trying to solve, we still followed the same pipeline. I started to play with the order of things and format and other potential changes, and really came to conclude I was only making it more complicated.
* Run multiple algorithms inline – we don’t need a different script altogether to run different algorithms. In fact, it’s better to run them all in the same script. This was something I played with for a little bit to see if it was easier to manage or keep track of where I was in each process. Since all the data wrangling/cleansing/scaling has to be the same for each model run, there really is no reason to start breaking things apart into different files.
* Spend some time researching the models you’re using – there are a lot of parameters for some of these models to tune. Really spend some time getting to know how these work. Run dummy tests to “play” with different values for the parameters and understand the model impact. Once you have a good understanding, then start to combine them one at a time. If you start tuning to many things all at once, it gets really complicated quickly and difficulty to figure out what change had what affect.

**Lesson: Basket Analysis with Apriori Model**

This lesson was useful because it is so practical. Half way through this lesson I was already considering other product data sets and applications for this lesson that would yield interesting insight into market dynamics. This lesson also stretched me a bit in terms of reading and researching how to use the plots and image functions. The R plotting capability is fairly full featured, but there were some instances where the plot output was unreadable. I had to spend quite a bit of time researching how to make better use of these functions to produce useful visualizations. That was a good exercise because in the end I learned more from that extra research than simply following the steps in the lesson.

Three lessons learned from this activity:

* Visualizations in R can be difficult to use – but they appear typically sufficient. I do now see why other products in the market are used for visualization and R is used for the data engineering, cleansing, wrangling and model building/execution. This lesson helped me understand a little better how different products in the market can be used together to solve and present solutions for different use cases.
* There is a certain amount of art as there is science to this work – the more experience we get with the standard process and then the tuning of models, the better I am getting at understanding how to start out and tune the algorithms with which we are working. Sometimes I need to be patient and keep trying a variety of inputs, charting the outcomes and comparing until I find a best possible result for the data set with which I am working. There is a lot to read online regarding the process steps. For me, I need to spend a little more time reading the resources online before diving in to a problem, to ensure I have a better understanding of the models, or functions, or tools I’m planning on using.
* One of the reasons I’ve chosen to pursue analytics skills development is to determine whether or not I would find this enjoyable as a career path – or something to infuse into my daily work. I am learning that I really enjoy this work. I want to work on more complicated problems. I learned there are a lot of complex workflows, cleansing routines, data wrangling processes I can build in R. That is really exciting to me. S, my lesson learned here is that because this activity is practical, and directly applicable to what I do for work, I can see the value and enjoy imagining the possibilities of applying this new skill set in the future to different use cases. Seeing that we can do some really interesting and insightful things in R with very little practical experience makes me wonder what we can do the more we learn how to apply this tool to real world examples.

This completes my retrospective of these activities we recently executed. It’s been very insightful and productive.

Kevin