Relevance of Analytics in the Context of Decision Making – Part 2

Prof. Ram Gopal

This is a paper that just got accepted in one of the top journals in the field called production and operations management and this is work with colleagues in the U.S. Tang Wong who's a doctoral student and David Bergman who is a professor. The title of this paper is Predictive and Prescriptive Analytics Location Selection of Add-on Retail Products. Again, the objective here as I mentioned in my primer is we are trying to merge the prescriptive and the predictive pieces together. So that's the key objective and it's a long title but the important phrase here is what is called add-on retail products. So let me start with what this add-on retail products are. There are several examples of add-on products. The basic idea is that an add-on product goes with what is called a base product. There's a base product and then there's an add-on product on top of that. So, when you buy the base product then you may buy add-on product. If you don't buy the base product then you do not buy the add-on product. Very classic example of that is if you go to the U.S. called gas stations in India petrol stations when you go to fill up patrol in the U.S. you called if you fill gas then many gas stations have options that you can add some additives to your fuel. The company that we work with is called Additec so they sell fuel additives. So, when you fill up gas in your car having filled up gas in the car then you have the option of buying additive product, fuel additive. What it does it makes your engine cleaner and it makes the engine more efficient. So, this additec product is an add-on product that you could buy after having bought a base product which is petrol or gas. See in the slide here is a number of other examples of add-on products.

So in our research some of the characteristics important are these locations is a brick-and-mortar setting that we have. So the different locations where people fill up the petrol or gas and different locations where this add-on product is sold. Now in our setting the other thing that is

important to know is that the seller of the base product is not the same as the seller of the add-on product. So the base product in this case petrol is sold by one organization, the add-on product , this Aditech is sold by another organization. So that's the basic setting and this industry partner that we worked with has over 1,200 locations all across the United States and for our research we looked, we chose two regions and I'll explain in a bit as to why we chose these two regions. Region one on the left hand side has 89 existing sites and region two on the right hand side has about 146 existing sites. And these two regions are about 600 miles thousand kilometers apart. Now the business objective for the industry partner was they were looking to expand and they wanted a location expansion strategy from us. So in the figure the ones in the green are the existing locations. These are existing gas stations where they sell Aditech products. The circles in red are other gas stations that are available and they do not sell the add-on product there and they're looking to figure out which of these candidate new locations that they should expand to. So that's the fundamental question. Now there's a lot research in location planning and what is different about our work is two things. One is we are focusing on add-on products. Much of the literature in the past has not done that. The second thing is the combining this predictive and optimizing framework. The prediction comes in terms of trying to predict sales in this new locations and the prescriptive part comes in terms of having predicted in all these locations which of those locations would be the best maximize my sales. That's the prescriptive piece.

Obviously if data is central to any analytics initiative. In our case, we have data from our industry partner. In the existing outlets these are the gas stations we are currently selling additive product. We have sales about the base product, the base product incorporates petrol, gasoline. So that's G and in fact we have access to each individual transaction that happens in that location. So what we did essentially was to aggregate that to the total sales of gasoline for one year and also the existing vacations we have sales for the add-on product which is Aditec.Now along with this we felt that it's important to

also look at the location demographics. It's not just how many, how much your sales are, the location also matters. This is an example of where you need external data. You will recall I mentioned that for many analytics initiatives having only internal data might not suffice. You need to add that, you need to add that external data to complement that. So in our case we have information about location; latitude longitude and also demographics, demographics for that location and in terms of demographics we had access to lots of data but we chose to just stick with two one is income in that location and population in that location.

So that's the data we have for existing locations. For the locations, the candidate locations these are other gas stations where I do not currently sell but I'm looking to expand to them possibly. For them you obviously have data on the location and demographics. For the base product sales. So here's another petrol station. I can actually go find out what the sales are but what we did was for our research purposes we did not go and ask them but rather we simulated. I will explain why we did that. And then in this new candidate locations my first objective would be if I would expand in that location I need to be able to predict my sales. That's the predict piece of it. And some summary statistics about these two regions. In region one the distance, average distance between locations was about 34 miles and for region two the distance was a little longer about 163 miles apart.

Now this is a retail setting where there are these retail outlets in different locations. When you have a setting like that one of the first question that comes up is what's called spatial autocorrelation. Spatial autocorrelation basically talks about fact that the sales that I would have in my location not only depend upon the features of my location but also who my neighbors are. This becomes important in lots of retail settings. For example often times you may wonder in many cities around the world all the jewelers they tend to be close to each other in one district. It could be a jewelry district or diamond merchants would be in a one district called the diamond district. Same thing happens for textile and so on. Those

things happen because of autocorrelation. Sometimes there is a positive autocorrelation that is if I look at myself close to a competitor that actually helps my sales and sometimes the autocorrelation could be negative where if I'm too close to my competitor they might take away my sales. I would rather be further away. So one of the first questions you always ask is that is there spatial autocorrelation and how strong is that. That impacts both my prediction. It also impacts my prescription or my optimization. So there's a test called Moran's I test. It's if you are familiar with the concept of correlation an example is people often say that your income is correlated to the education levels, years of education you have. How do you test that correlation? Essentially you have two columns of data. One column of data on income. One column of data on years of education. Then you start to look to see if when your education levels are high your income levels would tend to be high or not. So based on that either they have a positive correlation or a negative correlation and something similar here. If on location to my neighbors also have if I have high sales but my neighbors also tend to have a higher sales. so that's the basic test. So we did Moran's I test and what we find is really interesting. So as I mentioned before there are two regions. In region one there was no spatial autocorrelation. In other words the sales in a particular location simply depend upon that location, that store's characteristics not the neighbors. Whereas in region two there is a pretty significant, pretty strong spatial autocorrelation. That becomes important. Now in terms of a predictive, prescriptive framework there's two steps. One is to build a predictive model then second is to select the right locations to expand to. Now if there is a spatial autocorrelation what that means is if I want predict sales of the particular location if there is spatial autocorrelation I also have to look at the neighbors. That is what this WG is, it's the weighted average of sales of neighbors. If there is no spatial autocorrelation then neighbors don't matter. It's just myself.

Prediction. 3 colleagues, we talked about number of different issues that one ought to be careful about in building predictive models. For us variable selection was

not a big issue because there are only a few variables. We simply had data about the base product sales and demographics essentially. So there's not too many variables. There's no need to further pare it down.

Now we had to make some interesting choices about model selection. To recall there are hundreds of different types of models for prediction but in our case because after we finish our prediction we need to build a model, a prediction model. I am sorry, a prescriptive model. Now to build that prescriptive model we need our prediction model to be close form like an equation. If it is not an equation then you cannot really optimize that. So we were forced to restrict ourselves to only those prediction models that are in closed form. That basically means we could only use essentially linear regression models and support vector regression and I mentioned what these two things are.

So we could not really use things like neural networks or decision trees or other models. There are two regions. There have been two predictive models. The question is which model do well. What we find is really interesting is that the model that does the best is what is called the radial kernel support vector regression. This has some non-linearities in it. So essentially what it tells us is that in both the regions there is non-linearity in it. In other words the relationship between the base product says that demographics and the neighbors on your predicted sales of add-on products tend to be intricate. It's a little complex. There's some non-linearity in hand in it and this predictive model captures that sense of non-linearity. We also ask the question okay we cannot use some other models because we cannot optimize but what if we use them? How well will they perform? And the answer here is that is for region two they will perform very poorly. Our model support vector regression is the best model. In region one again the same thing it's a little better here but still even if you use, if you compare with neural networks or with decision trees the model that we picked support vector regression with radial kernel and that was the best model. Now we have a predictive model. What do we do then? What's the next step? How

does that help me in terms of my key question which is which location should expand my business to?

So the first step is you take all the candidate locations and you try to predict if I would expand to this possible location what would my sales be. Now to do that I need my base product sales in this candidate locations. As I mentioned before one simple option is to go to those candidate locations and get information about what the gasoline sales were, what the petrol sales were. Our industry partner did that but for the purpose of our research we wanted to simulate that and the reason to simulate that is you want to see under what conditions does the model does well, under what conditions does the model not do well. So we simulated that. So when we did the simulation we used a normal distribution to create, to simulate the sales in all of these locations. Now one thing we did was we tried different simulation settings and we varied the standard deviation. The variation of the distribution. Now if you're not mathematically inclined what that means is that when the standard deviation is large what that means is when you try to simulate sales in other locations what you essentially do is in some locations they'll be large sales. In some locations the sales will be very small. In other words if the standard deviation is high what you will find is that amongst the locations I'm looking at some will have large sales, some will have very small sales. and if the standard deviation is small the sales seem to be similar across all the location. Now when the standard deviation is large where you have some locations with large sales of the base product and some locations with the smaller sales problem becomes trivial right. Obviously you would go to locations there the gas sales are large. So when the standard deviation is large you expect the problem to be easy to solve. It's when the standard deviation is not so large the problem gets tricky. That's where you need to use your optimization. Again it's somewhat mathematical. I will not go too much into the details but the idea is that I have S locations where I'm currently selling the add-on product and I'm looking to expand to K locations and the objective is I want to pick those key locations to

maximize my F my F is my predicted sales overall. So that's the basic core objective.

Now depending upon what kind of a model that I use and depending upon whether there is spatial autocorrelation or not the model can be as simple to solve or complicated to solve. In the case that there are no spatial effects there's no spatial autocorrelation in many ways the problem becomes pretty simple. And this is a mathematical optimization model. Just to give you a sense of what this says my XI's are my decisions. Should I open in location I or not. The current location just set up as SI1 because I already have, I already sell in those locations. For sites Q these are the ones I'm looking at my candidate locations, my XI could be either 0 or 1. if it is 0 that means I'm not going to expand there. 1 means I'm going to expand there. There is constrain called summation XI is equal to K that means that I'm looking to expand to K locations and I want to find out what are the K best locations to expand. My objective is to maximize my sales which is LI times XI. When XI is 0 that means I don't use that location then there's no sales. when XI is 1 LI is the sales that I expect. Where does LI come from? LI comes from my prediction model. Simple for regression; more complicated for support vector regression.

Now if there's no special effects this is actually a pretty simple problem to solve from an optimization point of view. Essentially all you would do is for each of the locations that you are considering predict the sales. what would be my add-on sales in that location using our prediction model. Then once you have the predictive sales for all the locations you simply pick the K locations that have the highest sales. done. Those are the best locations to expand to. That's an easy problem.

When the special effects which happens a lot that's when it gets complicated. Why does it get complicated? It gets complicated because my prediction if I open in a particular location how much sales can I expect there depends upon whether I chose one of the neighbors or not. If I choose one of the neighbors sales will be different than if I didn't choose the

neighbors. So I can make these decisions independently or if I say it will be I have to do them jointly. And that's what creates all the complications. So here the sales depend not only on the location but also the neighbors where else am I going to expand to.

So this creates a ton of complications and this in fact is called cardinality-constrained optimization problem. It is pretty hard to solve. So when the problems become hard let's call NP-hard non-polynomial then typically come up with some kind of a heuristic approach. A solution that is not the best not optimal but based on the logic it should perform well and heuristic approach is pretty simple. Heuristic approach essentially was instead of looking at expanding to all K locations at the time just start with one. Pick one site and then pick the next site and then pick the next site. So that's how we do it and it is not optimal but in general heuristic approach even they're not the best did pretty well.

So in terms of performance. How well did our methods perform? The first thing it I should do is to explain the metric. When you say better by what measure, by what metric. So the metric we use is the following. Let me explain that with the simple example. Let's say the current sales in the existing site is 80. That's my current sales and suppose I'm looking to expand at 10 locations. Their baseline strategy if they did not have this prediction they did not have this optimization model they're just going to go with simple common sense what would they do. The baseline approach which is a naive approach which doesn't use any analytics is simply to say that look I have this 50 possible locations. Let me go to the look into expand it 10 sites or the 50 if I need to pick 10 let me pick 10 with the highest gasoline sales. higher gasoline sales should mean higher add-on sales. so I am just going to pick the top 10 in terms of their base product sales. so that's the baseline approach. Now suppose you go with that. Then let's say you would go then from ED that's current sales to 100. That gain of 20 is because you're expanding. Now suppose you say that instead of using this baseline naive approach suppose let's do prediction let's do optimization and if you do all of that and let's say the sales now become 110. So you're going from 80

which is no expansion to 100 which is expansion but with no analytics but with analytics you go from instead of 100 you go to 110. So the gap from 100 to 110 is the additional value that you created because of analytics.

So our metric is that expansion, simple expansion gives you 20, 80 to 100. Analytic is taking you from 100 to 110 that's 10. So the value of analytics is 10 over 20 which is 50%. So that's kind of the metric. So and the results you see in the table here. What you see here is therefore both region one and region two. In all cases the results are positive. Again what I should emphasize here is that these are results purely from analytics. There's nothing else you did. What you did was to predict, predict well and you optimize, you optimize well. By doing that you realized significant gains and in fact a couple of other points I should make here is that for region two, region two is where there is spatial autocorrelation if your expansion strategy is to create lots of new locations that's when the gains become larger and larger and larger. And as I mentioned before as the uncertainty goes up like standard deviation was from 2 to 4 to 6 then performance gains drop a little. Again the reason for that is the uncertainty is high then you end up with having some locations in the large sales some location small sales. there the problem is a bit more, more trivial. This visually you can see this as well. So on the left hand side the blue dots are the ones that you would choose. If you're just going with baseline strategy and no analytics. On the right hand side is if you use analytics right and if you expand these are the locations you would go to. As you can see there are significant differences. In fact there's a location here which I'm highlighting here that you would use just because that's a location with large gasoline sales that you would not use if you use analytics. So analytics gives you more intelligence and so that it gains a larger. Then visually same thing happens with the region two as well. There are locations that you pick here that are large gasoline sales that analytics tells you do not pick that and that happens because of both demographics and also because of spatial autocorrelation which it accounts for.

We did some robustness in all cases. The models performed well. It's the last slide. Closure, essentially what it says is that this whole approach of predictive, prescriptive framework to solve your business problems can yield significant value especially on this spatial autocorrelation if you have a large-scale expansion strategy. Anyway that's where we are going to end here and at this point in time I going to stop sharing my screen and see if there any questions.

It looks like there's another program question. I'm just going to read the question out. It says what is the average GMAT score required to apply for the program? Again that's a I do not personally know the answer to that but it's a question that can be answered by the program itself. A few questions. One is please explain the baseline versus the predictive model equation in the context of my research? Well the baseline, so in our case the baseline is simply there's no prediction with the baseline. So if the decision to make is I want to expand to 10 new gas stations which station should I pick. The baseline approach would simply say that you look at all available gas stations and look at the sales pick the 10 largest, pick the location to the ten largest gasoline sales. that's a baseline model. The prediction model comes from the support vector regression which was the best model. So that tells you what are the sales in each of the location then you do the optimization. It tells you what are the 10 best locations to buy and those 10 best locations might not be the same 10 best locations that come from the baseline model. In fact they are different.

Okay. The next question is a more technical question. How do you ensure spatial autocorrelation is not biased with non-normal data? How do you sanitize the data? I'm not exactly sure fully understand the question. To compute spatial autocorrelation you do not need the assumption of normality for the data and in terms of sanitizing the data essentially we had sales in each of the locations in the neighbors and we have a distance metric which is basically gradient distance between locations and that should enable us to compute the spatial autocorrelation. Again I'm not exactly sure what the question is. Okay. The next question is what methods were used to simulate sales? Please come again on that

whether standard deviation being higher is good or bad? What methods were used to simulate sales? Again we try to simulate sales on candidate location. The candidate locations what are the baseline gasoline sales. as I said we could ask them but we didn't. So we estimated it. So the way we simulated is to use basic Monte Carlo simulation which is a it's a normal distribution. So we take all the existing sites. So we have the data on those sites and then we compute the mean and we take that mean and we go to new locations we assume that on average the sales are the mean. And then the standard deviation. When the standard deviation is high that means when you randomly pick values from a from distribution with the large standard deviation the numbers that you would get tend to be either large or small. So that's essentially what happens. So when the standard deviation is large when you randomly pick the sales of these locations some numbers will be large some numbers will be small. What does that mean that means that in some of the locations of gasoline sales will be large. In some locations they'll be small. So the problem becomes more easy because when you have some locations with large sales you naturally tend to pick those not the ones with small whereas if the standard deviation is smaller then most of them will have very similar sales. then you have to study auto correlation, demographics a lot more closely that's where it gets more complicated.

You next question is can you please advise what are the few best industry standard methods to deploy predictive models into production? I'm not exactly sure what are the few best industry standard methods. I'm not exactly sure I don't know if you're talking about what type of predictive models. So typically in most cases regression based models do really well and to extend that there are no constraints to using black box kind of approaches like neural nets. They are used as well. In terms of methods to deploy predictive models into production. Not exactly sure what the question is but I can take that question offline or you can email me.

Next question is this model reusable in the context of other add-on products for instance can this model be

used to predict which service provider a new customer would sign up for on the basis of TV sales? That's an excellent question. I think the approach is obviously reusable. You would essentially through the same steps in terms of another context like you're talking about which customers will sign up based on TV sales. Yes you can use the same approach but the data would be different obviously. You may have different types of X variables. You obviously will not have location and most likely and you may not have demographics. You may have other kinds of data. Also in terms of what prediction models work the best it may not be support vector regression. It could be something else so but the overall framework is certainly reusable.

Okay. I will be here for another few minutes to answer any further questions that may come up. Again thank you very much. I think lots of great questions. I certainly enjoyed the session. I hope you did as well and hope you got some value out of it. Again I mention one more time that seriously consider looking at the DM program I think it's an outstanding program and I think you'll benefit a lot from it. A few other questions. Okay. One question here was which test is better between Moran's I and Mantel's to check for a spatial autocorrelation? Great question. I do not know the answer to that. I have to look it up. I have a vague recollection of what the Mantel's approach is but I have to look it up. I am not exactly sure.

Great case discussed. Thank you. Another question any pricing analytics cases you came across that you can share the problem statement for? I have to look at it. I don't remember offhand but obviously pricing analytics is huge in marketing analytics. There are lots of interesting analytics applications there. Can you please throw some light on various data cleaning techniques available? Yeah I think that's as I mentioned before data cleaning is actually a very critical part of your whole predictive modeling project and in fact what people claim is that data cleaning takes up typically anywhere from 70% to 80% of the time in terms of the whole analytics. So it's an important step. Please throw some light on some various data cleaning techniques that are available. There lots of them. I think data cleaning there

are a number of issues that you see a lot frequently. One is with missing data that becomes an important issue. How do you deal with missing data. If you ignore it that may create bias. If you don't ignore it what do you then do. The issues also with errors in data that becomes a huge challenge because oftentimes you are not exactly sure where the errors come from and how to clean them. That creates a lot of challenges. It also challenges in terms of data types. Sometimes data that you want should be numeric like age but the data you have may not come in that format. So you have to deal with that. Data reduction is a huge issue as well because sometimes you have too much data. So you'll have to lose some significant [and pick a selection and for that we use techniques like principle components analysis and so on. Now if you go to any of the standard analytics tools like SAS or our particular or Python the lots of libraries available that will go through on the data cleaning stuff. It's certainly worth looking into that.

Okay. I have taken up almost hour and a half of your time. At this point in time I'm going to end the session. Again thank you. Thank you very much for attending the session and thank you for the great questions and all the best. Take care. Bye. Bye.