GreenBook Methodology

The Misrepresentation of Statistical Accuracy – Unethical, but Fraud?

At what point does misrepresentation of statistical outcomes journey into fraud? This blog is going to address a reoccurring and vexing issue - the misreporting of a segmentation.

By:Michael Lieberman
Founder & President / Multivariate Solutions

- Posted: July 4, 2022 07:07 am

Ethics in marketing research is extremely pertinent due to:

- The excess of contact marketers have with the public
- The reliance on the research product in decision-making

Among all aspects of business, marketing is closest to the public view, societal analysis, and scrutiny. Consequently, it has created an impression that marketing is an area prone to unethical practices.

This is particularly true in the statistical, predictive analytics, and data science areas which is a reason for concern. Often clients are not able to reproduce and validate results. They have unrealistic expectations of statistical outcomes. Disappointment with actual results is a common hazard in the statistical consulting business. Juking the stats when clients are not pleased to receive unexpected results or are quick to set the statistical bar too high, has become an ethical issue.

This blog is going to address a reoccurring and vexing issue, i.e. the misreporting of a segmentation Typing Tool Reclassification Rate (TT Rates). It is an issue where holding high ethic standards and reporting the true TT Rate comes with risks. For example, being dismissed from a project because previous data analysts had no qualms about reporting outcomes that are not mathematically possible.

Overview of segmentation and TT Rate

Suppose a company conducts a segmentation based on, say, 125 ratings of various aspects of a product, lifestyle, and preferences. The client decides to go with six segments as their preferred solution.

The company then requests a mathematical model to reclassify future respondents – that is, a screener of, say, 12-15 of the most predictive questions to put new respondents or a large database into one of the six baskets.

The original segmentation was based on a sample of 1000 respondents, all of whom have original classification. The model then applies the reclassification model to the *same* 1000 respondents. The percentage of respondents whose new classification *matches* the original classification is the TT Rate. For example, if 60% of the new classification matches the original segmentation the TT Rate is 60%. This is used as a measure of validity for the model It is often reported to the end client as a measure of the strength of the Typing Tool.

Segmentation in market research

The goal for Mangiare International – a fictive international convenience store megacorp – was to position itself as an upscale, quick-meal alternative to its competitors. In other words, it wanted consumers all over Europe to take the extra five minutes to drive to a Mangiare for a quick hot meal rather than pull over to the nearest convenience store they spotted on the road. To accomplish this, it hired Aniva Research, a mock marketing research company to perform a market segmentation. Aniva Research intended to find the highest spending Mangiare customers and craft a media strategy to attract these valuable consumers.

Aniva Research created a survey with 125 variables that would be used to segment Mangiare customers. Ultimately Aniva Research settled on six segments. A reclassification tool was a required deliverable for the project. I was replacing the outgoing Director of Marketing Sciences at Aniva Research. The project directors were adamant that I perform the segmentation *exactly* the same way as the departing employee – including the reclassification typing tool.

Aniva Research requested to create two equations. One is based on 25 statements other on 14 statements.

TT Rate methodology

The most effective method for reconstituting a structure is to use exactly the same materials used to build the original. If one thinks of segmentation as a structure, those materials are the data used to calculate the final version of the segmentation. Simply put, the 125 statements.

Whenever I embark on a reclassification scheme, I move the segment solution and the structural variables into a separate file, open that file in R stat, and run a simplifying model training and tuning across a wide variety of modeling techniques. Among them, are discriminant analysis, nested regression analysis, and R Statistical Caret prediction procedures.

To create a baseline, I used all 125 variables in a model to re-create the six-segment model. This gives me a ceiling, the *maximum possible* TT Rate. In the case of Mangiare, we reached an extremely high 85% TT Rate. Of course, Mangiare cannot deploy a 125-questionnaire screener for future studies. Also, they cannot apply such a large reclassification screener to the millions of data records in the Mangiare database.

The solution, then, is to find the most predictive subsets of variables for the maximum effect.

Using the same model training and modeling techniques, I found the maximum reclassification with the most effective 25 variables is *70.1%*. The maximum reclassification for 14 variables is *53.2%*. To put that in perspective, if randomly re-assigned respondents into six segments, they would most likely get 17% correct. Using our reclassification procedure, we improved the accuracy of the model for 25 variables by a lift of (70.1/17) by 410%. Using 14 variables, we improved the lift of the model by (53.2/17) or 315%.

The pitfalls of arbitrary expectations

Aniva Research was not happy with my reclassification model. They claimed that their departing statistician regularly reported TT Rates of 80% using 25 variables and 70% using 14 variables. I was dismissed from the Mangiare project. Aniva Research claimed to have found another statistician that was able to meet its expectations.

Here is the rub. Given the parameters of the segmentation, the typing tool classification rates that the other consultant reported are *mathematically impossible*. What Aniva Research wanted, what it had become used to, was that its data scientist was able to remove 100 statements, or 80% of the variables, with absolutely no loss of predictive accuracy. In other words, it builds just as solid a structure with 80% less material.

The calculated probability of that being true is 0.0000000086%. Not zero, but very close.

It is actually quite easy to juke the TT Rate:

- Simply report the requested TT Rate
- Remove half the data to jiggle the TT Rate
- Use 25 variables to re-create the segmentation and then report the TT Rate using the maximum (25) variables to attain the TT Rate requested. This is rare, but I have seen it.

Unless the replacement consultant *changed the parameters*. In other words, moved the goal posts. This is akin to using different materials to *manufacture* the same structure. Which is *unethical* and *fraudulent.*

Unethical…but does it constitute fraud?

It is unethical for data scientists to misreport any finding. It is also commonly given that, often, results cannot be independently verified. However, the question of fraud is not as clear.

In the case of Mangiare, the Aniva Research segmentation output is an expected, perhaps an outstanding deliverable. It does meet Mangiare's requirements and reveals a valuable segmentation that is shown to have a higher value for the company. The reclassification typing tool is likely the most efficient. It is just not as accurate as reported to Aniva Research by their data scientist.

The journey from unethical into fraud could well depend on how the typing tool is reported. If Aniva Research reports the TT Rate to the management of Mangiare, and the company then uses it financially, the line has been crossed into fraud. The Aniva Research executives, if caught, could reasonably claim ignorance.

Takeaways

This situation has arisen for me on a few occasions in my career. On those occasions, the reactions from clients like Aniva Research have been, not unexpectedly, silent. It is not in the interest of the research firm to admit its unreasonable expectations or that a third-party consultant has come upon unethical practices within its organization.

This is, unfortunately for the third-party statistician, a losing situation. Any relationship with a client like Aniva Research is lost. The statistician has nothing to gain – and perhaps quite a bit to lose – to contact an end client like Mangiare directly.

Tips for damage control: Keep a paper trail of emails. Pen detailed explanations to the management of firms like Aniva Research on exactly what transpired. I report the highest possible TT Rate given the project and the likelihood of their expected rate, shown above as near zero. Be sure to make the memo on the record. Other than that, there is little to be done.

Conclusion

Things sometimes go wrong. Clients will refuse to listen. Project goals will change. Go with the flow. Events like this study are out of your hands, particularly when the research marketing teams are unwilling or unable to take action.

Move on to the next project.