The Linux Compete strategy: An analytics case study

Received (in revised form): 6th July, 2019

Marco Vriens
is Assistant Professor of Marketing at the University of Wisconsin — La Crosse. He has provided consultancy services to many Fortune 500 firms and has applied a range of analytical methodologies on a wide variety of business problems. He has a PhD in marketing and is the author of 'From Data to Decision: Handbook of the Modern Business Analysts' (Cognella Academic Publishing, 2018), and 'The Insights Advantage: Knowing How to Win' (2012), and editor of 'The Handbook of Marketing Research' (Sage, 2006).

University of Wisconsin — La Crosse, 1725 State St., La Crosse, WI 54601, USA
Tel: +1 608 518 8399; E-mail: mvriens@uwlax.edu

Chad Vidden
is an assistant professor at the University of Wisconsin — La Crosse, where he leads a data science and mathematical modelling research group that collaborates with local companies. He has a PhD in applied mathematics, with expertise in computational mathematics, data science and machine learning.

University of Wisconsin — La Crosse, 1009 Cowley Hall, La Crosse, WI 54601, USA
Tel: +1 608 785 5214; E-mail: cvidden@uwlax.edu

Abstract This paper examines how Microsoft dealt with the threat of an emerging competitor to its server operating system. The paper describes how a survey-based approach was used to identify the key drivers of customer brand preference and how Microsoft used multiple analytical approaches to quantify the potential threat. The context required both business understanding as well as forecasting and prediction accuracy. With these requirements in mind, the paper uses simulated data to demonstrate how, in this case, some approaches work better than others. The paper also briefly discusses the challenges encountered in getting management to act on the insights obtained through such analytical work.

KEYWORDS: advanced analytics, regression, decision trees, random forests, change management.

INTRODUCTION

In 2003, Microsoft had become a behemoth, marketing a broad range of products and services to consumers, IT professionals and software developers. For each audience, Microsoft had to understand different customer needs and compete with different business rivals. In addition, Microsoft was involved in several litigation battles, both with governments (the European anti-trust case) and other companies, concurrent with a wave of backlash from consumers who felt frustrated by the lack of alternatives on the market. At the same time, Microsoft's professional audiences were, to some extent, benefitting from Microsoft's dominance: many had been trained on Microsoft software and had Microsoft certifications that translated into economic (ie vocational) benefits.

IT professionals represented a key market for Microsoft as this audience was exposed to

a wide range of Microsoft products ranging from maintenance and troubleshooting software for desktop computers, to server and IT infrastructure products like Windows Server OS,[1,2] to the Office suite of products, software security products and, increasingly, business intelligence tools. Microsoft was facing competition not only from established companies like IBM and Oracle, but increasingly from many other directions, including Google (which was starting to challenge the Microsoft Office suite of products) and Linux — backed by companies like Novell and Red Hat that were challenging in the server operating system (OS) space, and VMware in the virtual server space.

Linux was a small open source server OS, originally derived from Unix. It was 'free' in more than one sense: anyone could download it without having to pay an initial purchase price or subscription fee, and it was free because IT professionals were able to make changes to it.[3] Although in 2003 its share of the US$6bn server OS market was dwarfed by that of Microsoft (Linux had under 3 per cent, while Microsoft had somewhere close to 75 per cent[1]), Microsoft recognised that Linux posed a potentially serious threat to its position.[4] Some even believed Linux could penetrate the desktop OS market, which would have had disastrous consequences for Microsoft.

After meeting with several enterprise customers in 2002, Jim Allchin, Microsoft's Group Vice President, made clear to his key executives that Linux had become a threat, stating quite unambiguously: 'We are not on a path to win against Linux'. [5] Microsoft's General Manager of Platform Strategy, Martin Taylor, recognised that even though Linux was very small, the company needed an in-depth analysis of the risk it posed in order to understand its severity and respond accordingly. To this end, he commissioned various research studies in order to develop an evidence-based campaign to contain the Linux threat.

Taylor called an all-hands meeting with his leadership team — the director of advertising, director of public relations (PR), the Windows product director, and members from the licensing and pricing teams — to discuss what could be done to neutralise the Linux threat. Several opinions were offered: The discussions shown below are not the actual discussions that took place. The authors were not part of the actual discussions. However, based on our experience, the discussions used in the text are fairly common.

- The PR director recommended raising awareness about the reliability of Microsoft Server and setting aside budget for a 'Get the Facts' campaign. She pointed to initial case studies that had found Microsoft's product to be as reliable as its Linux equivalent but with total cost of ownership (TCO) that was as good or better. She believed that if IT professionals knew how well Microsoft Server performed against Linux, they would come out in favour of Microsoft.
- The licensing team strongly recommended better, less complicated licensing deals.
- The product director argued that IT professionals really care about security and reliability of operating systems and, in this regard, the Microsoft product fell short. She recommended focusing on the improvement of these aspects and investing in making the set of applications even stronger.

These recommendations made for an interesting starting point for a second meeting, at which the various assumptions and claims made would be explored in greater detail. Most importantly, however, Taylor needed to know what insights-based strategy would be best for neutralising the potential threat, and what was really driving IT professionals to prefer one product over the other.

ACTION: ANALYTICS

The marketing research team reanalysed the results from a recent usage and attitudes (U&A) study of approximately 300 IT

professionals in the USA. (It is worth noting that although the original data collection for this project was conducted nearly two decades ago, U&A studies still remain one of the most commonly used marketing research methods today.[6]) The study was also set up as a tracking study, with the survey executed over several time periods (eg twice a year, for a few years). For the purpose of the present paper, this study will be referred to as the 'Linux Compete A&U tracker survey'. Table 1 presents some of the key survey questions used in the analysis (simplified for educational purposes).

The present case study generated simulated data consistent with the results from the original dataset used by Microsoft in 2003. The paper uses a logistic regression model to represent the data-generating mechanism (DGM). In practice, the true consumer DGM is not known. For this reason, multiple analytical alternatives are usually conducted in order to identify which best represents the data at hand. For a description of the DGM see the appendix. Table 2 summarises the average (simulated) IT professional's perceptions of Windows and Linux on ten attributes.

Table 1: Outline of the Linux Compete U&A tracker survey

1: Were it totally up to you, what server OS brand would you prefer?	
(a)	Microsoft Windows Server OS
(b)	Linux
2: Please evaluate brand [insert first brand from q2, then second] on the following attributes:	
(a)	Reliability
(b)	Security
(c)	Total cost of ownership
(d)	Initial purchase price
(e)	Interoperability
(f)	Ease of use
(g)	Ease of licensing
(h)	Skills to support
(i)	Scalability
(j)	Applications

Table 2: The average values on the perceptions (ratings from 1–10) of Windows and Linux OS

Perception attribute	Microsoft	Linux
1: Reliability	7.0	5.9
2: Security	6.0	6.0
3: Total cost of ownership)	4.9	7.0
4: Initial purchase price	6.9	8.0
5: Interoperability	7.5	7.9
6: Ease of use	8.1	6.0
7: Ease of licensing	5.6	8.1
8: Skills to support	8.8	3.0
9: Scalability	7.9	4.1
10: Applications	8.1	4.0

The initial reporting of this survey was mainly based on descriptive statistics. This did not answer the question about what drives preference and what aspect of the value proposition has the greatest impact on market share. To get insight into this question and arrive at a compelling, actionable result requires a multivariate approach. In the present case, the approach consists of two components. First, a predictive model is developed. Predictive models can be used for either prediction/forecasting or understanding (interpretation) or both. If prediction is the sole purpose, one only needs to worry about how well the model predicts or forecasts, and not about interpretation. In the present case, however, both prediction and understanding are important. To be accepted by management, a model needs to predict with high accuracy and have an actionable interpretation. Secondly, to help with actionable interpretation a simulation of likely impact based on an on-par simulation is conducted. These two steps are discussed below.

Developing a predictive model

Given the binary dependent variable of the data (ie prefer Microsoft — yes or no),

one can consider a variety of analytical techniques (an overview of these techniques is available in many texts, eg[7]). This study compares four techniques: linear regression, logistic regression, decision trees (with Gini coefficient as splitting method), and random forest (with 500 trees each with a depth of 6). Both set size and depth (for random forests) were varied but did not yield better results.

Even though the dependent variable is binary, the linear regression model is easy to understand and communicate. This is important when seeking buy-in from stakeholders. Logistic regression is specifically designed for a binary dependent variable[8] and comprises an underlying utility model and a choice model. The utility model is linear in its parameters but the conversion to binary choice is not. For both the linear and logistic regression, the DGM is very explicit. Decision trees[9,10] and random forests[11] are also suited for binary dependent variables. Decision trees work well when there are many variables and there is little or no theory to guide the variable selection[11] or when interaction effects are likely to help with the prediction. In the present instance, this is not fully the case: the number of variables is very modest and there is a decent notion of what variables to include (ie the variables in the survey). Although one might argue that interaction effects could be at play, the impact of interaction effects is typically not very large. It is worth noting that for both the decision tree and the random forest, the DGM is less straightforward. Sometimes researchers restate a decision tree as a linear model; this, however, is misleading as a given variable only accomplishes its effect in the branch in which it is appears; hence, it is conditional on other effects having kicked in already (unless it is the first variable that is split). In terms of prediction, random forests and decision trees are often (although not always) better than linear regression and logistic regression,[12–14] but this tends to come at the expense of being more difficult to interpret.

Table 3: Predictive (in-sample) accuracy results

Predictive model	In-sample predictive accuracy (%)
Linear regression	90
Logistic regression	91
Decision tree (Gini coefficient)	87
Random forest	99

The first aspect of the models to be evaluated here is (in-sample) predictive accuracy (see Table 3). The random forest performs best followed by the logistic regression model which fits very well. This is not surprising as it is consistent with the DGM that was used to generate the data.

Simulation of on-par scenarios

In Microsoft's case, the executives responding to the Linux threat needed more than just good predictions — they also needed to know what they could do to minimise the potential threat. The different methods shown here vary significantly in terms of interpretability. Interpreting the linear regression model is straightforward; the coefficients can be directly interpreted and any potential changes in perceptions (the independent variables) effects the dependent variable by multiplying the intended change with the perception coefficient. The interpretation of the logistic regression model is less straightforward because the impact of changes in an independent variable translate to a binary choice via a probability based output. To predict the impact of policy changes, one must specify the range over which improvements are desired. Typically, this can be calculated in Excel. Decision trees are even more challenging to interpret due to the lack of any coefficients at all and cannot not really be used for direct policy decisions. Figure 1 shows the decision tree derived for this dataset.

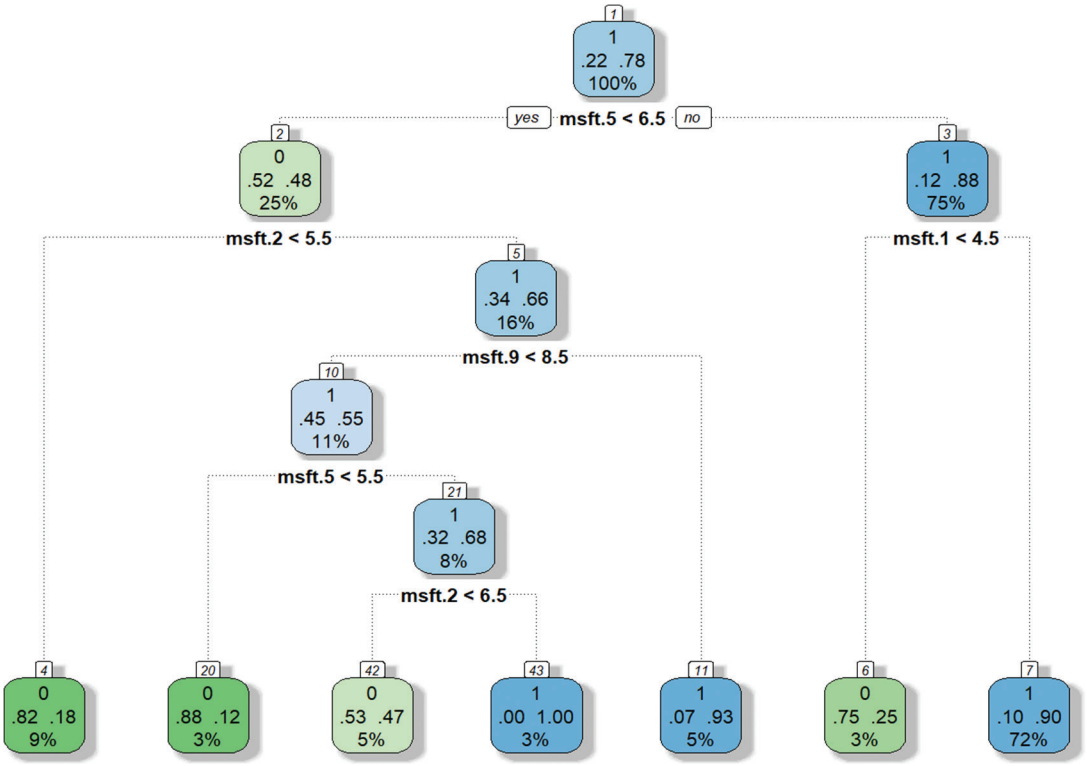

Figure 1: Linux vs Microsoft decision tree

As one can see from Figure 1, this tree is still relatively easy to interpret. As decision trees get deeper, interpretation becomes more difficult, and one typically see effects that cannot be intuitively explained.

One way to circumvent the coefficients or effects directly is to run simulations. In the Linux case, it was decided to do an on-par simulation. For those attributes on which Microsoft had an advantage, the study simulated what would happen had Linux improved its performance to match Microsoft and vice versa (ie for those attributes in which Linux had an advantage, a simulation was run to model what would happen were Microsoft to catch up).

This simulation approach works well for linear regression and logistic regression but does not work well with decision trees and random forests. A tree splits a variable in two at each branch. This creates limitations from a simulation point of view. It means that some continuous improvements in an independent variable will show no effect on the preference share. Impact can only be identified when simulations correspond with how the tree splits the variable. For example, if the decision tree splits the skills to support variable between a score less than 5 versus a score higher than 5, then only simulations that make the jump from less to more than 5 will show an effect. In contrast, with linear and logistic regression, an effect will always be observed. Furthermore, quantifying the importance of the independent variables, especially in the presence of interaction effects, is not straightforward with decision trees. The random forest model is even harder to explain and understand as it would require the inspection of multiple trees.

Table 2 shows that Linux has an advantage on TCO (7 vs 4.9), so in this case, one would simulate what would happen if Microsoft's TCO perception

could be improved to 7 as well. A similar analysis is conducted for all significant variables in the various approaches. The results of these simulations are shown in Table 4.

Table 4 provides several interesting observations. First, although the various approaches agree largely in terms of what variables are significantly predicting preference share, they vary substantially in terms of the impact they predict if a brand improves on a certain variable. One can see from Table 4 that the Linux skills to support perceptions variable appear to have a big impact under the linear and logistic regression model, but no impact in the decision tree or random forests models; as stated previously, only if the simulation aligns with the splits is it possible to observe any effect.

Table 4: Simulation results based on the various models

On-par simulation	Change in share gain
Linear regression	
Microsoft reliability	7%
Microsoft interoperability	–2%
Microsoft scalability	29%
Linux skills-to-support	23%
Logistic regression	
Microsoft reliability	19%
Microsoft interoperability	–6%
Microsoft scalability	36%
Linux skills-to-support	33%
Decision trees	
Microsoft reliability	6%
Microsoft interoperability	No gain
Microsoft scalability	3%
Linux skills-to-support	Variable not statistically significant
Random forest	
Microsoft reliability	No gain
Microsoft interoperability	No gain
Microsoft scalability	No gain
Linux skills-to-support	No gain

Validation

Several analytical approaches were applied to validate findings. The results cannot be considered consistent; in other words, the policy implications would probably be different. So, how to choose? Practically speaking, there are three ways to improve the likelihood that any action based on results will work as intended: (1) avoid over-fitting, (2) apply triangulation and (3) use multiple sets of data.[15,16]

Over-fitting can be dealt with, at least partially, by using an out-of-sample fit measure. This can be done in a variety of ways. One can 'hold out', say 10 per cent of the data, then estimate the model based on 90 per cent of the data, and evaluate how well the model fits the 10 per cent that was held out, ie not used for estimating the model. A generalised version of this approach is known as k-fold validation. Using this approach will typically provide a more modest fit value. For example, Mullainathan and Spiess developed models to predict the value of houses.[17] Using the random forest model yielded an 85 per cent in-sample fit, but only a 45 per cent out-of-sample fit. This method is used to screen out models that are over-fitted. The same authors also executed a k-fold validation, demonstrating that a significant variable used in one partition may not show up in another partition. The more consistently different folds identify a certain variable as significant, the more confident one can be about the importance of that variable.

Triangulation is another approach sometimes recommended[16] which consists of using multiple sets of data to confirm findings. This process increases the generalisability of the results. In the Linux case, for example, the quantitative analysis was followed by focus groups that discussed a number of the variables identified in the

predictive models. Through these focus groups it was identified that 'skill factor' was an important impediment to the wider adoption of Linux.

ACTIVATION AND ACTION

When the results were first presented, Microsoft's chief Linux strategist did not believe the predictions and insights. To confirm some of the conclusions, he requested follow-up focus groups and other ways to validate the insights. For example, the analysis was repeated using data from other waves of the survey (ie other time periods in which the survey had been executed). Microsoft also executed a further survey to measure market share. This market share tracker was executed in some of the same countries as the Linux Compete U&A tracker. Modelling from the Linux Compete U&A tracker was used to predict Linux market share in countries that both studies had in common, and a high degree of accuracy was observed between our predictions and the results from the other study.

It took 4–6 months to get the Linux Compete team to accept the insights and get them to act on the results. Actions that were considered or taken included: (1) publicise the benefits of Microsoft skills, (2) further improve Microsoft training and certification, and (3) offer academic institutions significant discounts against Microsoft products, possibly including free software for students. In the latter case, the cost of providing gratis software for students could have been measured and weighed against the loss of losing market share (whether or not this was done, however, is beyond the present authors' knowledge).

CONCLUSION

The business case was informed by the strong assumptions made by Microsoft's executives. Staff were assigned to find out the facts about Linux vis-à-vis Microsoft. With in-person conversations and focus groups, IT professionals were found to be complaining regularly about the price of Microsoft's products, how confusing the licensing agreements were, and that the software was not as reliable or secure as they wanted. In these areas, Linux was said to perform better. Secondly, Linux was popular among people who passionately believed in a future away from corporate greed with software for the people by the people. IT professionals also liked that they could make changes to Linux to customise it to their environment. (Interestingly, however, the feature 'customisation options' was not included as a potential decision variable in the study.)

Microsoft, however, believed that some of these sentiments were not based on facts. As such, it commissioned a number of studies to measure how Microsoft really stacked up against Linux and to show that Linux was not 'really free'. The results of these case studies informed the subsequent 'Get the Facts' campaign, which Microsoft explicitly intended not to be a 'traditional marketing' campaign. Although this campaign was successful, it left out a key factor in the competitive battle. Contrary to the assumptions of the PR director and the licensing team, it was not total cost of ownership, or the licensing agreement, or the reliability of the OS that constituted the major threat to Microsoft — it was the skills factor. Without, multivariate predictive analytics, this insight would have been very hard to uncover. It became more urgent as Microsoft realised that 'prospective IT professionals — those still in college', were being trained on Linux, which colleges were using because it was free. In other words, within three to five years, employers would be recruiting IT graduates with a natural preference for Linux.

From a modelling perspective, this study has demonstrated that it pays to evaluate multiple analytical alternatives, to increase the likelihood of selecting the model most consistent with the DGM. The study has

also shown that one cannot fully rely on prediction. Popular machine-learning algorithms such as random forests and decision trees may do well prediction-wise, but can be hard to translate into credible and actionable recommendations.

APPENDIX: THE DATA-GENERATING PROCESS

- Manually choose 20 variable (questions 1–10 for both Microsoft and Linux) means and variances and generate 300 random respondents on these 20 variables.
- Manually choose significant variables with desired coefficients.
- Variable Microsoft Reliability with coefficient 1.8
- Variable Microsoft Security with coefficient 2.1
- Variable Microsoft Interoperability with coefficient 2.3
- Variable Microsoft Scalability with coefficient 0.9
- Variable Linux Skills to Support with coefficient −0.5 (note negative coefficient implying inverse relation with respondent Microsoft preference)
- Generate dependent variable (binary choice indicated Microsoft preferred) from simulated dataset by only considering significant variables with coefficients.
- Probability = σ(C^T X) where σ(t) = 1/(1 + e^t) is the sigmoid function.
- Dependent variable is generated by comparing probability to a binomial distribution.
- Append dependent variable to full dataset (20 independent variables, 300 respondents).
- Build logistic regression model on full dataset.

References

1. Wikipedia (n.d.) 'Windows Server 2003', available at: https://en.wikipedia.org/wiki/Windows_Server_2003 (accessed 11th July, 2019).
2. Microsoft Corporation (2005) 'Results of operations for fiscal years 2003, 2004, and 2005', available at: https://www.microsoft.com/investor/reports/ar05/staticversion/10k_fr_dis.html (accessed 11th July, 2019).
3. Stallman, R. (2009) 'Viewpoint: why "open source" misses the point of free software', *Communications of the ACM*, Vol. 52, No. 6, pp. 31–33.
4. Kolakowski, N. (2004) 'IDC sees double digit growth continuing for Linux', 8th December, available at: http://www.eweek.com/article2/0,1759,1737068,00.asp (accessed 11th July, 2019).
5. Schestowitz, R., (2009) 'Microsoft's Jim Allchin: "I am Scared [of GNU/Linux]" (Analysts Cartel Part II)', available at: http://techrights.org/2009/01/18/allchin-scared-of-gnu-linux/ (accessed 11th July, 2019).
6. Vriens, M., Brokaw, S., Rademaker, D. and Verhulst, R. (2019) 'The marketing research curriculum: closing the practitioner-academic gap', *International Journal of Market Research*, available at: https://journals.sagepub.com/doi/full/10.1177/1470785319843775 (accessed 11th July, 2019).
7. Vriens, M., Chen, S. and Vidden (2018) 'From Data to Decision: Handbook for the Modern Business Analyst', Cognella Academic Publishing, San Diego, CA.
8. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) 'Applied Logistic Regression', Wiley & Sons, Hoboken, NJ.
9. Kass, G. (1980), 'An exploratory technique for investigating large quantities of categorical Data', *Applied Statistics*, Vol. 29, No. 2, pp. 119–127.
10. Magidson, J. (1994), 'The CHAID Approach to Segmentation Modeling: Chi-Squared Automatic Interaction Detection', in: R.P. Bagozzi (ed.), 'Advanced Methods of Marketing Research', Basil Blackwell, Cambridge, MA, pp. 118–159.
11. Breiman, L. (2001), 'Random forests', *Machine Learning Journal*, Vol. 45, No. 1, pp. 5–32.
12. McCarty, J.A. and Hastak, M. (2007) 'Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression', *Journal of Business Research*, Vol. 60, No. 6, pp. 656–662.
13. Larasati, A., DeYong, C. and Slevitch, L. (2015) 'Comparing neural nets and ordinal logistic regression to analyse attitude responses. *Service Science*, Vol. 3, No. 4, pp. 304–312.
14. Lee, K., Park, J. Kim, I., Choi, Y. (2018) 'Predicting movie success with machine learning techniques: ways to improve accuracy', *Information Systems Frontiers*, Vol. 20, No. 3, pp. 577–588.
15. Ehrenberg, A.S.C. (1990) 'A hope for the future of statistics: MSOD', *American Statistician*, Vol. 44, No. 3, pp. 195–196.
16. Saffo, P. (2007) 'Six rules for effective forecasting', *Harvard Business Review*, Vol. 85, No. 7/8, pp. 122–131.
17. Mullainathan, S. and Spiess, J. (2017) 'Machine learning: an applied econometrics approach', *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 87–106.