Journal of Data Science,17(3). P. 504 - 512,2019
DOI:10.6339/JDS.201907_17(3).0003

The Use of Predictive Modeling in the Evaluation of Technical Acquisition Performance Using Survival Analysis

David E. Booth[1], Ceyhun Ozgur[2]

Abstract

In the recent statistical literature, the difference between explanatory and predictive statistical models has been emphasized. One of the tenets of this dichotomy is that variable selection methods should be applied only to predictive models. In this paper, we compare the effectiveness of the acquisition strategies implemented by Google and Yahoo for the management of innovations. We argue that this is a predictive situation and thus apply lasso variable selection to a Cox regression model in order to compare the Google and Yahoo results. We show that the predictive approach yields different results than an explanatory approach and thus refutes the conventional wisdom that Google was always superior to Yahoo during the period under consideration.

Keywords: Innovation Performance, Acquisitions, Innovation Knowledge, Hazards Model, Cox Proportional Hazards, Survival Analysis, Google, Yahoo, predictive model, explanatory model.

1. Introduction

It is well known that management of innovation performance can be crucial to the success of a technology firm. Recently several authors have used survival analysis (also known as event history analysis) as a method to gauge the innovation performance of a firm. Most recently Datta and Roumani (Datta and Roumani , 2015) have attempted to do this by means of a proportional hazards Cox regression (Cox, 1972) model. We recognize that such a time-to-event study is quite a reasonable approach to take but unfortunately the authors were unaware of the statistical difference between predictive and explanatory models as described in the seminal work by Shmueli (Shmueli , 2010). This caused Datta and Roumani to confuse the techniques of explanatory modeling (which they used in their analysis) with the appropriate predictive modeling for the analysis of the proposed models. The present paper does the following:

1) Argues that this is indeed a predictive situation as described by Shmueli (Shmueli, 2010)
2) Treats the data analysis of the Cox Regression model, which Datta and Roumani treated by explanatory methods, by the appropriate predictive methodology (in this case adaptive lasso variable selection) and thus
3) Shows that the previous authors' conclusions were in error and
4) Draws the appropriate conclusions .

This paper thus provides a template for carrying out a statistical analysis of the performance effects of an innovation management program.

2. Model Development

It is well known that Google and Yahoo have used acquisitions as a major part of a strategy to manage innovation performance (Datta and Roumani, 2015). In particular, both have tried to increase their pace of innovation by means of acquisitions. It would be ideal if a way were available to measure the success of each company's program. Datta and Roumani (Datta and Roumani , 2015) have attempted to do this by means of a proportional hazards Cox regression (Cox 1972) model.

In order to do this, they posited a number of variables as being important to the company's success. They took the measure of innovation success to be the time to patent (TTP) and the time to launch (TTL) a product, where each was the first such event, that is the first such event of each type that happened after an acquisition measured in days, where a few observations were right censored. The data were found in publicly available data sources (Datta and Roumani, 2015). Because of the type of data a Cox regression seemed a reasonable way to start. They began by setting up a set of hypotheses to be tested using the parameters of the Cox model and proceeded to test the hypotheses (unfortunately confusing Cox regression estimation and ordinary least squares inferential methods). This is a typical (but somewhat confused) explanatory approach to modeling (Shmueli, 2010). However, for some time many statisticians have made the distinction between explanatory and predictive models. It is well known (Shmueli, 2010) that one type of model will not always replace the other. Further, the model building methods should often be different for the two types of models. We argue here that by the very method of a time-to-event study predictive modeling is the more appropriate approach. To see why this is so, let us consider the goal of the research. The main goal is to

determine which organization made better use of the acquisition process. With their explanatory model, Datta and Roumani (Datta and Roumani, 2015) conclude that Google outperforms Yahoo on both measures: TTP and TTL. Let us now consider if an explanatory model is the most appropriate.

Following Shmueli (Shmueli , 2010), we wish to infer from the data which of Google or Yahoo has been most successful at using the adopted acquisition strategy as measured by TTL and TTP. In Section 1.5 of the work, Shmueli (Shmueli , 2010) says that "Laws connecting sets of variables allow inferences or predictions to be made from known values of some of the variables to unknown values of other variables." Thus here we argue that what we really want to do is to infer whether Google or Yahoo produced the best application of its strategy in the two cases TTP and TTL. We argue that knowing which of the firms is making the best use of the adopted strategy is a forward-looking concept. This is an important distinction because the two types of models in a particular instance are often not the same and the methods used to construct the model are often different. We conclude that a predictive model is most appropriate and thus proceed to use variable selection model building techniques. In particular, shrinkage methods of regression are suitable for predictive but not explanatory models Shmueli (Shmueli , 2010). We will show that the result of a predictive model analysis is not the same as an explanatory analysis and further we will show that the predictive model better answers the research question of which firm was more successful in applying each measure of a successful strategy. We further show that the distinction is important because the results of the two analyses are not the same. Further, perusal of Datta and Roumani (Datta and Roumani, 2015) shows that their model was built using explanatory modeling with Cox regression. The prediction equation was built with a Cox regression program. However, statistical inferences are dealt with as if the prediction equation was built with ordinary least squares regression. This is clearly incorrect and makes their conclusions invalid (Cox ,1972; Lee, 1992). Further, since the Datta-Roumani models are explanatory they are not necessarily the best predictors of TTP and TTL (Shmueli, 2010), while the Cox regression adaptive lasso predictors are. This is true because as Zhang and Lu (Zhang and Lu, 2007) showed they satisfy an oracle property. Thus the model that we propose here is superior to that proposed by Datta and Roumani (Datta and Roumani, 2015) because we use the methods of Zhang and Lu.

3. Data and Methods

In this section, we discuss the methods that were used in the predictive analysis. The data set was that used by Datta and Roumani (Datta and Roumani , 2015) and is described there. We then arranged the data into four subsets for analysis labeled as:

GTTP	YTTP
GTTL	YTTL

where the first letter represents the firm, Google or Yahoo, and the remainder of the name represents the dependent measure described. The size of these four sets is shown in Table 1.

Cox proportional hazards regression (Cox, 1972) was chosen as the base for building the predictive model. The reason was that we wished to know which firm was more efficient in terms of the goals of patenting and bringing products to market. Outliers were identified by

using the robust Cox regression estimator proposed by Faracomeni and Viviani (Faracomeni and Viviani , 2011) that is based on trimming. The 5% outliers were removed and set aside for further analysis. The reason for this is that we want to be sure that we have the best measure possible of the average performance of the two firms. The final predictive model was obtained by using the adaptive lasso procedure for Cox regression (Zhang and Lu, 2007) using a Bayesian Information Criterion (BIC). A comparison of BIC and AIC methods is given by Huang et al (2009). All final prediction equations were validated as described in Harrell (Harrell, 2001) using bootstrap cross validation with 150 bootstrap samples each, using Harrell's R packages. All computing was done in R. The Kaplan-Meier survival estimates were calculated with the final variable selected reduced Cox regression model as described in Lander (Lander, 2013). The predictors were those of Tables 2-5. The chosen predictors had at least one non-zero adaptive lasso coefficient. R programs and the data set are available from the authors.

4. Results

Outliers can cause serious errors in Cox regression (Faracomeni and Viviani, 2011). Because we are building a predictive model we want the final model to represent the majority of data points without the possible deficiencies introduced by outliers by using the Faracomeni and Viviani (Faracomeni and Viviani , 2011) procedure to ameliorate such outliers. We identified and removed the most serious outliers (5%). Because no other information was publically available about the observations removed nothing further was done with them. Table 1 shows the 5% outliers as defined in Faracomeni and Viviani's (Faracomeni and Viviani , 2011) algorithm.

Table1: 5% Outliers Removed from the Data in Order for the Computational Cox Regressions to Represent Average Firm Performance

Data Set	N	Outliers
GTTP	63	10 13 36
GTTL	63	18 27 40
YTTL	55	13 23
YTTP	55	51 55

One of the major differences between explanatory and predictive modeling is the possible use of variable selection techniques to choose the independent predictor variables in the final selected predictive model. Because of its optimal predictive properties (Zhang and Lu, 2007; Lu, Zhang and Zeng, 2012; Lu, Goldberg and Fine, 2013) we chose adaptive lasso developed by Zhang and Lu (Zhang and Lu, 2007) to select the final predictive models for the data sets of Table 1 using a BIC criterion. All final prediction equations were selected by choosing the variables with non-zero adaptive lasso coefficients as shown in Tables 2-5. These prediction equations were successfully validated by bootstrap cross validation (Harrell, 2001) using 150 bootstrap samples each. Kaplan-Meier survival curves were calculated from the final prediction equations as described in Lander (Lander, 2013). In the case of the Kaplan-Meier curves, the selected prediction equations contained all variables for which at least one Table showed a non-zero adaptive lasso coefficient.

Table 2: Regression Coefficients – GTTP Data Dep. Var. – days to patent

X	Robust	Alasso
country	-0.63146	0
base	-0.79707	0
New_incremental	-0.20183	0
Related_or_not	0.00896	0
Product_process	-0.48540	-0.25342
(group) specialization	-0.19025	0

Table 3: Regression Coefficients – GTTL Data Dep. Var. – days to launch product

X	Robust	Alasso
country	-0.15148	0
base	0.10624	-.07999
New_incremental	-0.75734	3.28530
Related_or_not	-1.45981	0
Product_process	-1.63336	0
(group) specialization	-0.12635	0

Table 4: Regression Coefficients – YTTL Data Dep. Var. – days to launch product

X	Robust	Alasso
country	-0.28095	0
base	0.02071	0
New_incremental	-1.258578	-0.20215
Related_or_not	-2.458044	0
Product_process	-0.719901	0
(group) specialization	0.159702	0

Table 5: Regression Coefficients – YTTP Data Dep. Var. – days to patent

X	Robust	Alasso
country	-0.674438	-1.060182
base	-0.051255	0
New_incremental	0.786351	0
Related_or_not	0.858125	0
Product_process	-1.087786	-0.21000
(group) specialization	-0.166944	0

The analysis proceeds in the following manner. First, we know that outlying data points can cause errors in the conclusions drawn from a data analysis using Cox regression (Faracomeni and Viviani, 2011; Lander, 2013). In order to deal with this potential problem, we begin the analysis by looking for outlying observations in our data set using the robust Cox regression method of Faracomeni and Viviani (Faracomeni and Viviani, 2011) to identify the 5% outliers in the data. These results are indicated in Table 1. Again, because it is known that outliers are different from the majority of data observations, we removed the Table 1

outlying observations from the data sets and set them aside for separate analysis (Booth, 1984). Second, we then do variable selection on our four data sets after outlier removal in order to choose the optimal predictive model for each data set. Because of the many advantages of the adaptive lasso procedure (Zhang and Lu, 2007; Lu et al, 2012; Lu et al, 2013), we choose that method, using the R program of Zhang and Lu (Zhang and Lu, 2007), as reported by Boos (Boos , 2014) using a BIC. These results are given in Tables 2-5. The final reduced prediction selected variables were those without 0 adaptive lasso coefficients using a BIC. Equations were validated using bootstrap cross validation with 150 bootstrap samples as described in Harrell (Harrell, 2001) using Harrell's R packages. The Kaplan-Meier survival curves were calculated for each of the selected models (Aalen et al, 2008; Lander, 2013). The Google and Yahoo curves were then compared for the patent and product launch data sets to determine whether Google or Yahoo was more successful in those particular cases over the years studied. The lower of the two curves shows the firm that is the faster of the two in getting patents approved or launching products (Lee, 1992). All programs and data sets are available from the author (DEB).

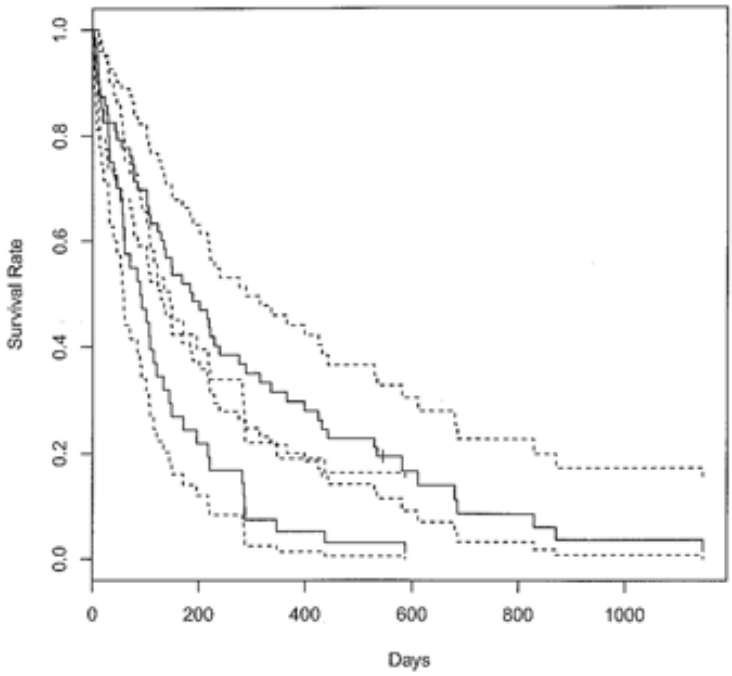

Figure 1: Kaplan-Meier Curve for Time to Patent The lower curve is Yahoo

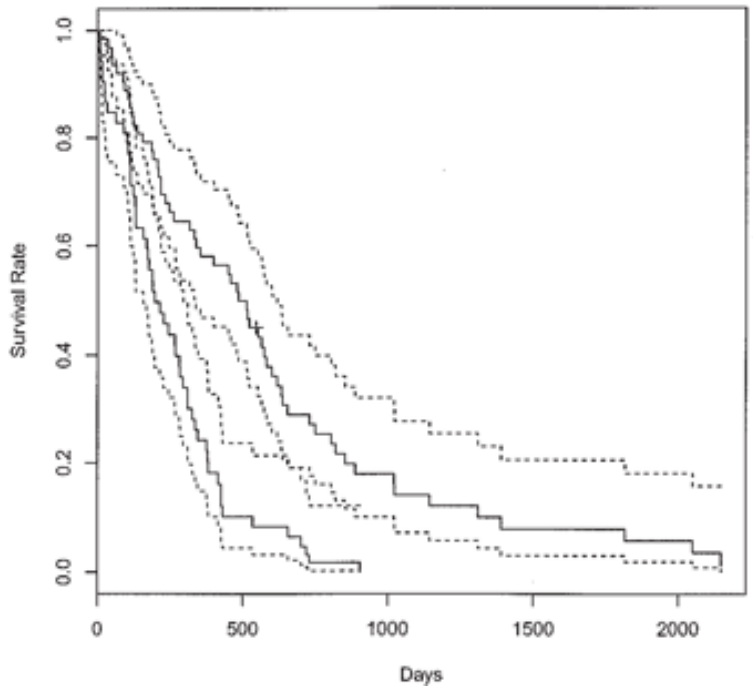

Figure 2: Kaplan-Meier Curve for time to launch product Google is the lower curve

5. Discussion

The results of the analysis are shown in Tables 1-5 and Figures 1 and 2. It is interesting to note the similarities of the Tables. In the patent case (Tables 2 and 5) the major predictor selected was Product_Process for both Yahoo and Google. While country was selected for only Yahoo both variables were included in the rest of the analysis. In the product launch case (Tables 3 and 4) new_incremental was selected for both but size of knowledge base was also included in the further analysis even though that selection was specific to Google. It is reassuring to note that the major predictor variables selected in the patent and launch cases were essentially the same for both firms indicating the major drivers were the same in both cases.

We now consider the Kaplan-Meier (Lachin 2011; Lee, 1992) curves for the above models. For the Kaplan-Meier curves, in the Time to Patent case (Figure 1) Yahoo is the lower curve while in the Time to Launch case (Figure 2) Google is the lower curve. In both cases there is very little overlap between the 95% confidence bands. In the Kaplan-Meier curves the lower the curve the faster the event happens (Lee, 1992). Hence Yahoo was faster patenting but Google was faster launching products. Because Google was undoubtedly the market leader during this period (Datta and Roumani, 2015) this would indicate that the speed of bringing products to market was a key factor in the competitive leadership at this point in time given that both firms had similar strategies subject to the limitations of the analysis to be described in a moment. This results in variance with the Datta and Roumani (Datta and Roumani, 2015) result based on a somewhat confused explanatory model analysis. We suggest based on the previous discussion that the predictive model is to be preferred and hence that the result reported here is to be preferred.

There are, of course, limitations to this analysis. These are mostly caused because some information is proprietary. The biggest two such pieces of information that are missing are the amount of money spent on research and development for each of these products and the amounts spent for marketing the products. Either or both of these sources of funds could have shortened the time for patenting and launching the products. Thus varying amounts spent could have an effect that should have been included in the Kaplan-Meier curves. However, the amounts spent are known only to the two firms.

6. Conclusion

Based on our analysis we conclude that Google was more effective at bringing products to market and that Yahoo was more effective at getting patents during the periods studied and that predictive models are most appropriate for this application.

7. Acknowledgements

We thank Sara Viviani and Wen-bin Lu for the use of their R programs and Yaman Roumani and Pratim Datta for the use of their data and the idea of using a survival based analysis. We thank the editors and reviewers for their helpful comments.

References

[1] Aalen, O., Borgan, O., and Gjessing, H. K. (2008), Survival and Event History Analysis. New York, NY: Springer Science, Business Media, LLC.

[2] Boos, D. (2014), An Adaptive lasso in R. 2/9/2014 http://www.Stat.NCSU.edu/~boos.Var.Select/lasso.adaptive. html

[3] Booth D. (1984), Some Applications of Robust Statistical Methods to Analytical Chemistry, Ph.D. Dissertation, University of North Carolina at Chapel Hill.

[4] Cox, David R. (1972), "Regression Models and Life Tables." Journal of the Royal Statistical Society Series B, 34(2), 187–220.

[5] Datta, P., and Roumani, Y. (2015), "Knowledge Acquisition and Post-Acquisition Innovative Performance: A Comparative-Hazards Model." European Journal of Information Systems 24, 202-226.

[6] Faracomeni, P. and Viviani, S. (2011), "Robust Estimation for the Cox Regression Model based on Trimming." Biometrical Journal 53, 956-973.

[7] Marvin, G. (2017), Report: Google earns 78% of $36.7B US search ad revenues, soon to be 80%, https://searchengineland.com/google-search-ad-revenues-271188

[8] Harrell, F. (2001), Regression Modeling Strategies. New York, NY: Springer Science.

[9] Huang, J. S. Ma, H. Xie and Zhang, C. H. (2009), "A group bridge approach for variable selection." Biometrika 96(2), 339-355.

[10] Lachin, J. M. (2011), Biostatistical methods: the assessment of relative risk, 2nd ed., New York, NY: John Wiley and Sons.

[11] Lander, J. (2013), R for Everyone. New York, NY: Addison Wesley.

[12] Lee, E. T. (1992) Statistical Methods for Survival Data Analysis, 2nd Edition. New York, NY: John Wiley and Sons.

[13] Lu, W., Zhang, H. H., and Zeng, D. (2013), "Variable selection for Optimal Treatment." Decision Stat. Meth. Med. Research 22, 493-506.

[14] Lu, W., Goldberg, Y., and Fine, J. P. (2012), "On the Robustness of the Adaptive Lasso to Model Misspecification." Biometrika 99, 717-731.

[15] Shmueli, G. (2010), "To Explain or To Predict?" Statistical Science 25(3), 289-310.

[16] Zhang, H. H. and Lu. W. (2007), "Adaptive Lasso for Cox's Proportional Hazards Model." Biometrika 94, 1-13.