Managing Analytics Projects

Michael Greene[1], David Steier[2]
[1]Deloitte Consulting LLP, 200 Berkeley St, Suite 400, Boston, MA 02116
[2]Deloitte Consulting LLP, 225 West Santa Clara St., San Jose, CA 95113

Abstract
Analytics projects are often undertaken by statistical consultants in the business sector. Unfortunately many business analytics projects fail to achieve their goals. Often, we statisticians focus on aspects such as the lift curve or the model fit while ignoring critical aspects of managing the project, such as gathering requirements, engaging stakeholders, and choosing between infrastructure platforms. Statisticians could play many roles throughout the project life cycle and, we believe, improve analytic project chances for success. We illustrate these roles by describing business analytics projects, typical project activities and opportunities for statisticians. We will discuss positive (and negative) outcomes of these topics, illustrated through case studies and experience on analytics projects. The paper includes an extensive listing of red flags and warning signs and follow-up questions to assist the analytics practitioner or statistical consultant in navigating the dangers to successfully land the project.

Key Words: Statistical Consulting, Analytics Projects, Project Management

1. Introduction

In recent years, the business community has demonstrated unprecedented demand for use of analytics: the application of rigorous quantitative thinking to improve business decision-making and operations. Businesses now use predictive models for applications as diverse as targeting potential customers, underwriting financial risk, and even focusing retention efforts on employees who are the most likely to quit. This demand for analytics stems from the exponential increase in data collected, economic success of analytics-based business, the increased prevalence of basic statistical training in businesses, and the increasing accessibility of tools for data analysis and visualization.

Statisticians are naturally suited to play crucial roles throughout such analytics projects, from framing the problem to choice of model through the implementation of the analytics within the business. Unfortunately, when trained statisticians are involved at all, they often participate in only a small subset of these activities. And, perhaps not coincidentally, the majority of business analytics projects fail (Demirkan & Dal, 2014). Of course, failure can have many definitions. In this context, failure means not succeeding in meeting the business goals or objectives. In this paper, we describe the roles that statisticians can, and we believe should, play in managing analytics projects. This guide is not intended as a comprehensive discussion of analytics projects. Rather, we aim to help statisticians make analytics projects more successful with suggestions and recommendations on how to navigate the business environment.

The remainder of this paper consists of three main sections before concluding. First, we define business analytics projects and describe the importance of managing such projects. Second, we describe the stages of a typical business analytics projects and offer suggestions on the potential roles of statisticians in each of the phases of the project. Third, we provide recommendations for statisticians before, during and after the project including common warning signs and how to navigate potential issues.

2. Anatomy of the Business Analytics Project

Business analytics projects often differ from the projects a statistician undertakes in the academic or research worlds, having different goals, stakeholders, and risks that a statistician must understand to be successful. Because business analytics involves a much broader set of activities than testing a statistical hypothesis, we introduce definitions of "business decision-making" and "business analytics projects." We use these terms in discussing the typical phases of a business analytics project and the roles a statistician might play.

2.1 Key Definitions

Business decision-making chooses among potential actions to achieve business goals. Examples include: Which marketing plans will best help increase sales? At what price should a product or service be offered to maximize profits? Which people from a pool of potential candidates should be hired to achieve optimal performance in the organization? Which transactions are most likely to involve a heightened risk of fraud, waste, or abuse? In improving business decision-making, incomplete information is the norm, there may be more than one right answer, and it is not always obvious how quantitative methods should inform such decisions.

Statisticians, in contrast to business decision-makers, are generally more familiar with hypothesis testing, which uses statistical methods to accept or reject a quantitative hypothesis. A null hypothesis might be that an increased response to a direct marketing campaign over a control group could be due to chance. Examples from the business world include: does this marketing communication cause potential customers to respond at a higher rate than a control group? What is the confidence interval for the lift in sales due to the sales promotion? Did the introduction of a new benefit offering improve employee retention?

The differences between the types of questions asked in business decision-making and in traditional statistics are addressed by John Tukey in the essay "We Need Both Exploratory and Confirmatory" (1980). Business questions that spawn analytics projects are questions that Tukey would describe as "not at all the kind of question that can have a statistically supported answer" (23). When a business person thinks "question," he or she envisions a broad, intuitive, question such as "why am I losing customers?" A statistician thinks "question" and envisions a null hypothesis. Tukey points out that part of the role of the statistician is translating the business question into one or more specific questions which might be answered through data analysis.

Business analytics is the use of data by businesses to improve business decision-making. (In a few cases, businesses may apply analytics for purposes other than decision-making, such as to advance the frontiers of knowledge in a research context, but this is a small

minority of cases and we focus on decision-making applications of analytics here.) The analysis of data in business may involve traditional statistical methods, but may also draw on other quantitative disciplines such as operations research or machine learning. Traditional measures such as goodness-of-fit may be used to assess results but may be complemented by speed, interpretability, resources required or other considerations.

A **business analytics project** is a time-bounded effort to coordinate statistical and other quantitative data analysis with non-statistical activities (e.g., technology implementation, process review, change management) to achieve business goals. A project structure is required because of the range of activities needed to ensure statistical hypothesis testing or other quantitative techniques lead to improved business decision-making. For example, the problem must be framed appropriately so the right choices are being considered in response to a particular business context. The right metrics must be chosen to measure business outcomes in response to the actions so relevant hypotheses can be created and tested. Data sources must be selected that contain the right input and output variables. The right modeling techniques must be chosen to analyze the data so decision makers can have confidence in the results. Results from statistical hypothesis testing may need to be integrated with other types of analytics. Often, the sponsor of the project is a business person with little statistical training. Analytical results must be communicated not only to those sponsors, but also to end-users: those who would use the analytics to improve decision-making. These activities are all non-trivial to execute, and may be as critical to project success as the analytics themselves.

Managing analytics projects systematically directs constrained scope and resources (i.e., people, time, money, and infrastructure) to achieve project goals. Efficiency of resource utilization is especially important when billing for analytics project consulting is based on hours worked. Relative to the academic environment, analytics projects in a business environment are also more likely to be constrained in scope, and the statistician often has to manage with imperfect data or less analysis than he or she would prefer.

Gartner analyses report that historically the majority of business analytics projects have failed (2013): that they exceed the budget of time or money allocated to them, or never achieve their business goals (Demirkan & Dal, 2014). A variety of reasons are given, ranging from data quality to lack of strategic vision or integration across the organization, insufficient training, or any number of operational reasons. If statisticians want to increase the likelihood that their analytics projects succeed, they have to be aware of these factors.

2.2 Typical Phases of an Analytics Project and Role of the Statistician

While each analytics project is unique, common types of tasks include data management and transformation, exploratory data analysis, data visualization, statistical modeling, reporting, and implementation in an operational system. These activities have been mapped by Davenport and Kim into three phases: Framing the Problem, Solving the Problem, and Communicating and Acting on Results (2013). Framing the Problem involves problem recognition and review of previous results. Solving the Problem involves variable selection, data collection, modeling and data analysis. Communicating and Acting on Results involves presenting the results (often with appropriate data visualizations) and taking action, perhaps in concert with other people and technology inside or outside the organization (customers, suppliers, partners, etc.). In this section, we describe each of these phases and how a statistician should be involved.

With one important exception, the stages from Davenport and Kim align with the categories of work that John Chambers describes in "Greater Statistics" (1993): preparing the data, analyzing the data, and presenting the data. The exception is in acting on results, which Chambers does not cover. In today's business environment, organizations realize value by implementing analytics. Although not required, many projects include changing a business or technology process by implementing an algorithm or analytical solution to change decision-making. Davenport and Kim capture these activities in the phase Communication and Acting on Results.

2.2.1 Framing the Problem

Framing the problem means understanding the business context motivating an analytics project in order to address it by analytic methods. When referring to "business context," we mean the specifics of the problem being faced by the person(s) making the business decision: the choices available among which a decision must be made, who is making the decision, and the value of an improved decision. Statisticians must absorb all this context, performing research into what types of analytics have been used in similar problems, and working with stakeholders to refine the problem into one or more exploratory analyses or statistical hypotheses to be tested, where outcomes are measurable and data are available.

A business person identifying a problem where analytics might help does not usually start with a null hypothesis. Rather, the business person identifies an issue, what Tukey calls "an idea of a question" (23). These issues are informed by deep knowledge of the business yet are less specific than a statistician needs to properly frame a problem. Examples are: sales are sluggish, employee attrition is up, or our marketing isn't working anymore. The business owner may not understand the root cause or even how analytics might help. The statistician has an important role in converting these issues into questions which might be answered through data analysis. What are the precise questions the business person is trying to answer? What form would an answer to a question take, and what hypotheses do those potential answers translate into? What data would we need to test those hypotheses? How would we change the business process based on the analysis? Additionally, the statistician may have to persuade skeptics that the proposed analysis will answer the business questions.

Beyond ensuring the problem is framed precisely, statisticians should leverage their experience to point out potential ethical dilemmas in the application of the solution. For example, it is highly likely that models to predict when a mortgage-holder might default will show that home location is a significant predictor of default risk. Yet denying a mortgage application solely on the basis of location (also known as redlining) is illegal, because of the possibility of racial discrimination (Federal Reserve, 2014, 1). The time to flag such potential issues is before analysis starts.

For a statistician, playing such a variety of non-technical roles might seem uncomfortable or outside his or her comfort zone. As an example, in a response to Breiman's "Two Cultures" essay (Breiman, 2001), Brad Efron captures this feeling by saying "sample sizes have swollen alarmingly while goals grow less distinct" (218). Yet imprecision and ambiguity is more the rule than the exception in a business environment, and the analytics process must take into account such imprecision, or the project is at risk from the earliest days. It is worth remembering Tukey's advice that "finding the question is often more important than finding the answer," (23) and Chambers advice "Better to have an approximate answer to the right question than a precise answer to the wrong question" (1).

In one project, we were tasked with building a model that would predict which customers of a services business would be most likely not renew their service contracts. We had assumed that the results of the model would be used to target marketing efforts at customers at high risk of leaving and apply measures designed to retain them such as discounted pricing or improved contract terms. It was only after the project was well underway that we discovered the business did not have any data on the effectiveness of potential retention marketing efforts for different types of customers. Worse yet, the whole process of contract renewals had been outsourced to a third party call center under terms that would not allow for customized offers. Thus, we learned that what we had originally envisioned as a fairly straightforward churn prediction problem needed to be reframed as a challenge to re-think the entire contract renewal process.

2.2.2 Solving the Problem

In the context of analytics project, solving the problem means selecting data sources, gathering data, performing exploratory data analysis and visualization iteratively with statistical modeling. This process can involve multiple rounds and iterations are usually needed before a satisfactory result is obtained. A statistician is usually more comfortable during this phase but it is important to keep in mind the business context that constrains the analysis, such as the data sources to be used and the modeling techniques applied. Example constraints include availability of data sources, how the model could be used in production, or acceptance by the business.

As more data sources have become available in recent years, considerations of privacy have risen in importance. Data privacy concerns are particularly germane when dealing with sensitive types of personal data such as health or financial data. But browsing, shopping, location data are all potent sources for inferring behavior and should also be handled with care. The statistician may be in a position to determine whether subjects have granted consent for data about them to be used in the analysis, and issue warnings when privacy or ethical issues may arise.

Once appropriate data sources have been selected, the data must be explored and prepared for analysis. This stage is critical to perform before starting modeling as much of the value from business analytics projects come from insights gained through exploratory analysis of data. Imperfect data appear in almost all analytics projects. The statistician must specify data transformations, handle imputation of missing values, or deal with outliers and other quirks of the data. For example, if data on all customers are not available, the statistician should understand if a bias exists before creating models and generalizing results. If key pieces of data are not available or less populated as expected, the statistician should carefully evaluate and implement data imputation as needed. On projects where data from multiple sources are brought together, the statistician should indicate what to do when data for an observation is available in only one dataset. Should the record be excluded? Will a bias be introduced? How the statistician manages the issue during the project has a disproportionate effect on project success.

Sometimes findings from initial data analysis lead to reframing the problem. During a recent project one of the authors worked on an assessment of the impact of the introduction of 2D barcodes on vaccine labels on the quality of data regarding vaccination administrations in electronic health record (EHR) systems. The business question was whether adding 2D barcodes on vaccination products is effective in improving data quality. The statistical questions were harder to state precisely. Multiple data elements are recorded

in the EHR system: manufacturing lot number, expiration date, and product code. How should we define quality: accuracy of information or completeness? Once these questions were answered we set forth an analysis plan, defining how we planned to analyze and what statistical tests we would perform. During exploratory analysis of the data, we saw some promising signs: data quality appeared higher when 2D barcodes and scanners were in use. Unfortunately, once we executed the analysis plan, we found no improvement in quality.

Further exploration uncovered the cause: a relationship between how we defined data quality and how we defined the unit of observation in the analysis. Initially, we aggregated observations by vaccine (e.g., flu vaccine Brand A) and data quality for all observations for that vaccine was compared before and after scanners were available to record information from the barcode. The data was tainted, however, because the data quality on the name of the vaccine was correlated with data quality on the key elements being analyzed. In other words, if the person recording the vaccine entered poor quality data, that person was likely to introduce error into multiple fields at once in the EHR system. Conversely, people recording high quality data on the vaccine name were likely to record high quality data in the key data elements. Thus aggregation of observations by vaccine would not have uncovered the effect on data quality we were looking for. After exploratory analysis, we had to go back to re-aggregate the data by a different dimension: the providers administering the vaccines, which made the effect visible.

Once data sources are selected and the data prepped, the final step is performing the analysis. The analysis may draw on a variety of disciplines, some of which are in the realm of traditional statistics while others are not. The analysis may include a statistical model, predictive model, machine learning technique, or other quantitative model. Examples include generative models, regression analysis, statistical tests (e.g., t-test, Chi-Sq test), decision trees (e.g., CART, CHAID, and Random Forest), support vector machines, and artificial neural networks. The statistician is often the expert on the project team choosing the model, fitting the model to data, and analyzing the data.

The choice of analytical technique is a critical part of the statistician's role on the project. The statistician must look ahead to how the results are intended to be used and who will be making these decisions. Intuitively, the technique and model should align with the data, provide valuable results and insights, and help solve the problem. The statistician must consider both predictive power and usability when choosing the algorithm. If clarity and explainability of the model is needed for users who may not have statistical training, it may be preferable to use simple linear models, even if a non-linear model is slightly more accurate. Depending on the business context, it may be important for the client to be able to re-calibrate the model at a later date, necessitating a simpler model to minimize future costs.

We have worked with insurance clients where the most predictive models have been non-linear models, for example neural networks. Analysis on hold-out sets suggested these were able to predict insurance losses better than generalized linear models. To our surprise, clients rarely adopted these models. As clients see it, the five to ten percentage point increase in predictive power is rarely worth the increased complexity of understanding and implementation.

A few years ago we met statisticians from a large document storage and services provider who had hired a consultant to build a model for client retention. Unfortunately, none of them understood the stochastic gradient descent algorithm the consultant built for them.

Ironically, the statisticians ended up building a simpler regression model to understand the output of the more complicated model!

Leo Breiman writes about these considerations when he discussed "Occam's Dilemma" and the "Rashomon Multiplicity" as part of choosing algorithms (Two Cultures, 206-208). Occam's Dilemma states that we should choose as simple a model as possible, but no simpler. Breiman uses what he calls the Rashomon Multiplicity to illustrate his advice on model interpretation. In the movie *Rashomon*, the same events are shown through the perspective of multiple characters. Each interpretation appears different, yet each applied to the same underlying event. In analytics, the Rashomon Multiplicity occurs when many potential models with very different variables are equally "good" but one particular model may be chosen due to a quirk of the data. Due to the importance of insights and interpretation, the model selection is critical to gaining business acceptance. Unfortunately, the business owner may not understand that multiple models are nearly equivalent but tell different stories. The statistician must choose analytical techniques that provide stable insights.

We frequently encounter the Rashomon Multiplicity on projects. For example, on a recent project with a national discount retailer we built a model to explain store performance based on population demographics of the area around each store. The business believed that it was necessary to analyze hundreds of demographic attributes. As one might expect, these data had high levels of multicollinearity. We found that if the data were resampled, different predictors would appear more important. However, each set of variables told a very different story about the predictors of store performance.

To overcome this issue, we employed the robust variable selection techniques Random Forest (Breiman, 2001), Elastic Net LASSO (Zhou & Hastie 2005), and Principal Components Analysis (Jolliffe, 2002). These methods assist with variable selection and dimension reduction but might violate Occam's Dilemma as discussed above. Ultimately, we leveraged these techniques to explore the data and winnow down the list of potential predictors. The final model for the client was a simpler linear regression model based on a subset of the predictors identified by these exploratory techniques. The result was a simple and understandable model.

2.2.3 Communicating and Acting On Results

Communicating and acting on results involves a range of activities such as from documentation of the models, data visualizations, presenting the findings to various stakeholders, training end-users, and working with technical staff to feed analytical results into other systems (for example CRM and marketing automation systems). Here the statistician must take care to keep in mind the audience and stakeholders. User studies, such as those performed by ethnographers trained in anthropological field research methods, may be required. Additionally, to ensure end-user adoption training, change management, and extensive communication may be required. Without these investments, adoption will likely suffer.

As statisticians, our interest often focuses on Solving the Problem, but business leaders often consider Communication and Acting On Results the more important phase. In fact, we have never seen a project have too much communication. When we communicate more frequently, produce interim or preliminary results, and openly discuss early findings, we have more successful projects. When we remain silent and produce results only at the end, the business person may not understand the subtle analysis we perform, and projects lose

business urgency and even fail. With frequent communication and a preview, the client can see our progress and remain excited about potential findings. The statistician plays an important role in translating the analytical results so they can be communicated to business stakeholders, for example in in a slide presentation. While p-values and test statistics may be valuable to a statistician, we have to find more intuitive ways to communicate with business users. Without a clear understanding of assumptions and caveats, the business owner might extrapolate the findings or generalize inappropriately.

Documentation is a key part of this phase. Because of the role the statistician has played in each project phase, he or she often has a unique depth of understanding of the analytics, assumptions, and calculations necessary for implementing the analytics or acting on the results. Implementing the analytics in a business environment also may require additional documentation and specifications on data handling such as how to impute missing values, or how to handle extreme values. Additionally, the statistician might need to document pseudo-code, or the logic required to transform the raw data into the independent variables used in the model. Without detailed pseudo-code, IT professionals can misinterpret the algorithm while coding the models, reducing or reversing the benefit.

Acting on Results is the final part of this phase. Even if analysis, documentation, and communication are perfect, if the model does not impact decisions, the analytics work has little value. For example, marketing campaigns might be focused on those with the highest likelihood to purchase products, underwriting efforts and safe driver training could be focused on the drivers with the highest chance of an accident, or bonuses could be given to valuable employees who are likely to quit. In each case, the model a statistician built needs to be closely integrated into a technology system to act on results.

One of us worked with a commercial lines insurer to construct and implement predictive models for identifying risky insurance policies. The system would provide recommendations to underwriters based on the analytics. After the system was implemented, actuaries at the insurer discovered that underwriters were not following the recommendations. The actuaries identified the one hundred riskiest policies where the underwriters had ignored the recommendations of the models, and calculated these policies had cost the company over ten million dollars in the first year. The problem was misaligned incentives: the underwriters were paid by the volume of policies, not the quality. Once the results were presented to the company executives, underwriter incentives were changed and adoption of the analytics improved accordingly.

This example demonstrates the importance of measuring and monitoring the impact after the initial analytics have been completed. As statistical consultants, we often must leave a project before we see results. However, we can help our clients by providing recommendations and advice for post-project activities that ensure project goals are achieved.

2.2.4 Putting It All Together

Every analytics project involves project-enabling activities that cross the three stages; statistical consultants can assist. These activities are planning and budgeting, assembling a project team from internal and external resources and designing and implementing a technology infrastructure. The latter may be critical if the application involves “big data,” where the size, variety, or speed at which data arrive threaten to overwhelm traditional database systems. In such cases, the architecture of the appropriate platform to store and analyze the data may require very specialized talent to design. We worked with one

regional cable provider to analyze customers and viewership patterns. With the transactional database reaching into the hundreds of billions of records, we could not easily leverage traditional regression models to analyze patterns. The database, available software, and volume of data dictated some of the analytics we were able to perform. To address the challenges, we collaborated with the client to develop Bayesian models that could be implemented in the available systems and meet the goals of the project.

All of the project phase identified by Davenport and Kim are highly iterative. Exploratory data analysis is crucial for success in analytics projects as business leaders often do not have access to data and the associated insights. In modeling, sometimes results do not live up to expectations. Models may not predict as well as hoped and a randomized control trial may run short of expectations. As the statistician looks at the data, he or she might learn that the original framing of the problem cannot work or a different set of data sources or analytics techniques may be required. The statistician should communicate clearly and often with the results and preliminary indications.

3. Analytic Project Red Flags and Questions

In our experiences as statistical consultants working on analytics projects, we have encountered a variety of frequently occurring issues that undermine the chances for project success. In this section, we provide some examples of warning signs or red flags to watch for and questions to ask before, during, and after analytics projects.

3.1 Red Flags Before the Start of A Project

Table 1 below detail some of the red flags that we look for when starting projects, and potential root causes which may underpin the observable symptoms.

Table 1: Red Flags to watch for in starting an analytics project

Red Flag	Potential Issue
Requirements are very high level and do not specify completion criteria.	The business owner(s) may not fully understand the goals or have competing goals. If these priorities are not clarified, it may be difficult to reach consensus on when the requirements are met and the project is done.
Unrealistically high expectations of project benefits.	The recent popularity of analytics has led some people who do not understand the underlying techniques to make unwarranted claims. Data quality, sample bias, and variance are often ignored. If the client asks for economic predictions multiple years out, be wary.
Multiple project sponsors (with different agendas).	When sponsors from different departments are involved the project might have multiple sets of goals and requirements, potentially in conflict.

Red Flag	Potential Issue
Nobody proposing the project has actually seen the data.	In businesses, decision-makers might be divorced from the actual datasets necessary for performing analysis. Assumptions about data distributions and quality often prove invalid, and analytics may not be able to deliver a sponsor's expectations.
A key group or person already tasked with solving the problem is not involved.	Businesses sometimes hire consultants to validate results. When internal resources producing the results are not involved, you may be entering a hostile environment.
Key project members have left or are leaving the client or project.	The departure of key members can signal changes in sponsorship or that support for the project is not as deep as might have been assumed.
No representation of end user(s) on project team.	When end users are not represented, there is significant risk that the analytics will not be used due to lack of acceptance or buy-in.
No consideration of privacy when dealing with sensitive data sources or results.	Data breaches or misalignment with consumer expectations are more likely to happen when sensitive data are carelessly handled and distributed.
Lack of legal review of project risks.	Analytics projects are relatively new to the business community, where risk and legal precautions may not yet be well understood.

Do not underestimate how far people's assumptions can be from reality around data and statistics before starting an analytics project. One of us has encountered a controller who swore there could never be entries above a certain threshold in his general ledger. Yet we found them. Automatically generated entries calculated on the basis of erroneously entered data were incompletely unwound when the original data errors were corrected. Misunderstandings can also be quite fundamental to managing expectations around analytics: one person asked us to make predictions – but couldn't understand why we needed historical data to do so.

3.2 Red Flags During the Project

Table 2 describes some red flags we watch for during an analytics projects.

Table 2: Red Flags to watch for while managing an analytics project

Red Flag	Potential Issue
Not enough time in project schedule for testing, iteration, and documentation.	Analytics projects where technology resources are not involved often skimp on the time necessary to implement a robust application. This shortfall can lead to cost and time overruns later.

Red Flag	Potential Issue
Access to the data is constantly delayed.	Lack of access can be a sign of low-priority for the project set by client managers, and the data may never become available.
Key elements of the data are missing or otherwise much worse quality than expected.	Additional data collection or cleansing time may be necessary that puts the project timeline at risk.
Single source for key project elements (especially data, modeling skills).	The time of key client personnel is critical to project success. A single resource who is the expert on data, modeling skills etc. is likely to have many competing priorities.
Major unanticipated shifts in scope.	A sign of deeper problems, major changes in scope might be due to competing priorities, change in sponsorship or other issues. Major scope changes can derail an analytics project success quickly.
Lack of process or a schedule for status reporting and follow-up.	When a client is not able to dedicate time to learn about the status of a project, it may be a sign that the project is low priority.
Analytics runs take much longer than expected.	If the computing resources, software, or code are not sufficient for the task at hand, or an inefficient algorithm is used, result quality may be lower because fewer analytic approaches can be tried.
Reliance on vendor claims with no independent reference-able experience.	In projects where multiple vendors may be competing to serve a client, it is not uncommon for vendors to inflate their experience. If a vendor has a responsibility that cannot be fulfilled, the project may be at risk.
No case studies of where the technique has worked before.	Sometimes key stakeholders, often those without actual analytics experience, believe that a certain approach or analytics technique will work when it has not been tried.
Key stakeholders are unavailable or do not answer calls.	When key stakeholders are not available, the project is likely to fail.
All data used for training – no hold-out test set.	In analytics projects, employing a hold out set is a key to defending against overfitting. Without enough data for a hold out set, results in production may not match expectations.

Beware the situation when the "data guy" goes missing. Most businesses have knowledgeable experts who know which databases, tables, and fields contain useful data – and which do not. On one project, we ran into significant issues by not engaging the steward of a critical data source early in the project. The individual went on vacation while we performed some analysis. When he returned, the individual reported to the project sponsor that our analysis was invalid because we did not properly use the data he had provided.

Although not the only difficulty during the project, this issue set back our timeline as we had to revisit how we leveraged the data in the analysis.

3.3 Questions to Ask After the Project

Post-project activities are important because analytics projects almost never end with the actions taken on the basis of the first model produced. Even when projects are successful, there will be feedback on how to improve the model or present the results to the end user. Future opportunities might arise to recalibrate the model or expand the scope of the analytics to other parts of the organization, or even outside the organization. The statistician who has performed the initial analysis is often in a good position to inform decisions, incorporate the feedback, recalibrate the model, and expand the scope.

When a project concludes, the statistician may also help develop an independent objective evaluation of the impact of an analytics project, to learn lessons from the experience that may be useful in the next project. We recommend revisiting the original decision to be improved that was the target of the analytics effort and asking some hard questions. Table 3 below includes some of these questions and considerations.

Table 3: Questions to ask after the project is complete

Question	Considerations
Are the results being used to improve the decision?	When decisions are not being improved, little value accrue to the client from the analytics.
Are the end users applying the analytics as intended?	End users not applying the analytics might be indicative of misaligned incentives or less buy-in than expected from key stakeholders.
Have the decision improvements met expectations (e.g., accuracy, cost effectiveness, stakeholder feedback)?	Expectations play a key part in defining success. Analytics which do not meet expectations may constitute a failed project.
Have there been any unintended consequences?	Unintended consequences such as privacy issues, behavioral changes, or increased workload can result and have negative impact on other aspects of the business.
Are there follow – on opportunities?	If the project was successful (or even if it was not), there may be additional follow on opportunities with the client.

As with any project associated with significant investment, there will be commensurate political pressure to evaluate the project as successful, or if declaring success is a lost cause, to sweep failures under the rug and move on. How to respond to those pressures is a personal choice, but those people who have asked the hard questions to assess the "ground truth" will be able to make an informed decision on how to proceed.

4. Conclusions

The root cause of analytics project failure depends as much on decisions made regarding the management of the project as decisions made during the execution of the analytics. Due to depth of knowledge, statisticians should play a bigger role in managing analytics projects. Many statisticians shy away from key management decision to the detriment of their projects. We firmly believe that if statisticians were more involved in all aspects of delivering analytics projects that more of these projects would succeed. Indeed, we have seen a correlation between successful analytics projects and the involvement of skilled statisticians in all phases of managing analytics projects – Framing the Problem, Solving the Problem, and Communicating and Acting on the Results.

We strongly encourage would-be statistical consultants to familiarize themselves with project management, client management, and technology implementation. Statisticians who invest time in these areas will find their investments pay off with more successful projects and hopefully additional statistical consulting projects in the future.

Acknowledgements

We would like to thank Jim Guszcza for his advice and suggestions for research of Tukey, Chambers, and Breiman. Also, we would like to thank MaryJo Smith, Isabella Ghement, and Ralph Turner for their words of wisdom and editorial eyes during the drafting of the paper.

References

Breiman, L. (2001). Random forests. Machine learning 45, no. 1. 5-32.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16, no. 3. 199-231.

Chambers, J. M. (1993) Greater or lesser statistics: a choice for future research. Statistics and Computing 3, no. 4. 182-184.

Chambers, J. John Tukey and 'Software'. Memories of John Tukey. Retrieved on July 29, 2015: http://ect.bell-labs.com/sl/tukey/tributes.html#chambers1

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review 69.1. 21-26.

Davenport, T., & Jinho K. (2013) Keeping up with the quants. Harvard Business School Press: Boston.

Demirkan, H. & Dal, B. (2014). Why do so many analytics project fail? Key considerations for deep analytics on big data, learning and insights. Analytics Magazine July/August 2014. Retrieved July 29, 2015. On the World Wide Web: http://www.analytics-magazine.org/july-august-2014/1074-the-data-economy-why-do-so-many-analytics-projects-fail

Federal Reserve (2014). Consumer Compliance Handbook. Retrieved July 29, 2015: http://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_fhact.pdf

Gartner, Inc. (2013). Gartner Predicts Business Intelligence and Analytics Will Remain Top Focus for CIOs Through 2017. Retrieved Dec. 16, 2013. http://www.gartner.com/newsroom/id/2637615

Jolliffe, I. (2002). Principal component analysis. Springer-Verlag: New York.

Tukey, J. W. (1980). We need both exploratory and confirmatory. The American Statistician 34, no. 1. 23-25.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 2. 301-320.