4-30-2018

BUDGET PLANNING PROCESS & FORECASTING FOR IT ENTERPRISE

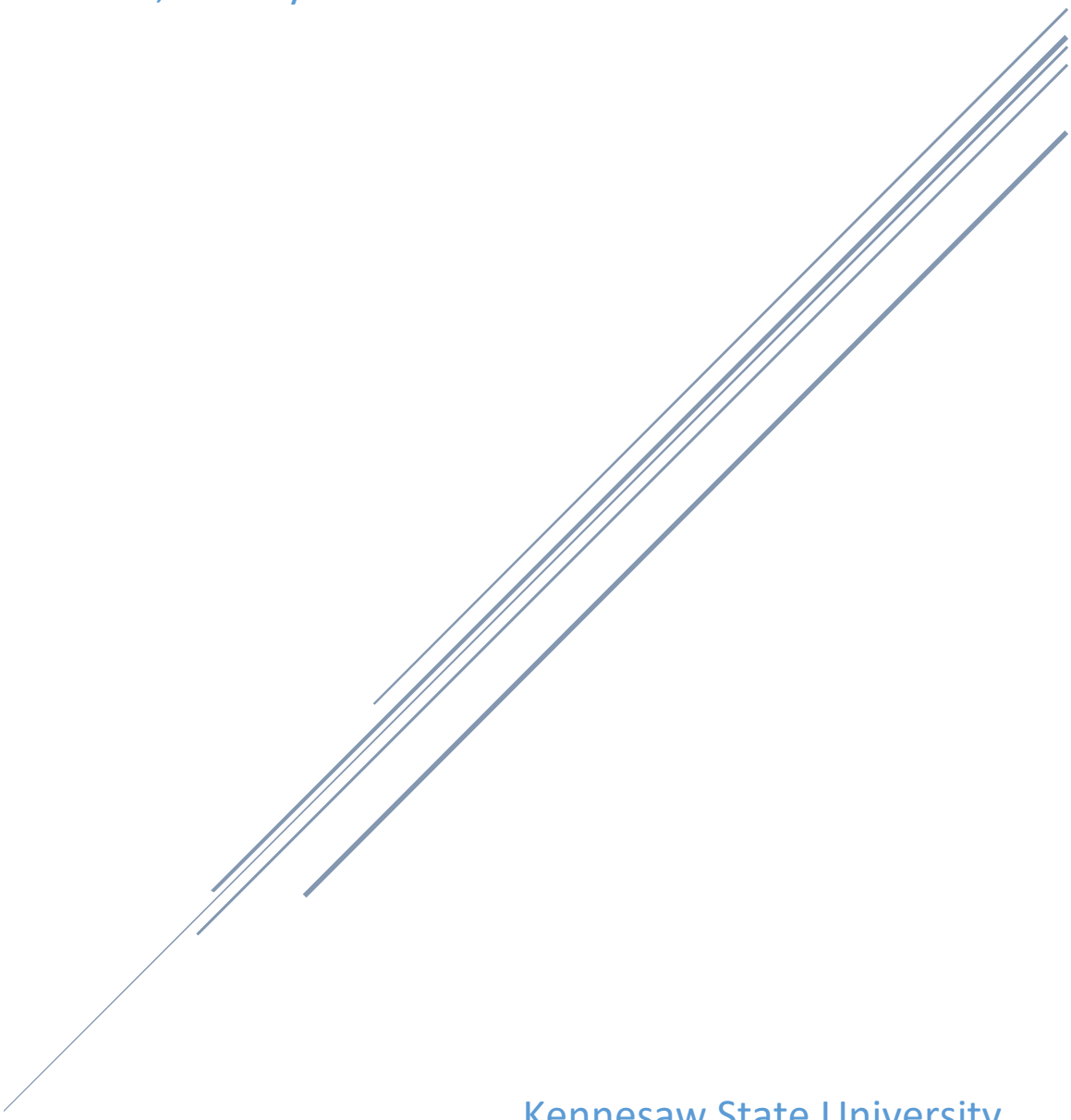
Joel Fillmon
Kennesaw State University

BUDGET PLANNING PROCESS & FORECASTING FOR IT ENTERPRISE

Joel Fillmon, Project Analyst

Rolf Erickson, Project Manager

Dr. Adeel Khalid, Faculty Advisor



Kennesaw State University

Senior Design Project

Blue text denotes additions for the IPR

Executive Summary

Executive Summary (Abstract, written LAST): project scope, achievements, etc. assuming that the project is completed successfully)

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

This project is, in short, a process improvement project to review the budgeting and budget forecasting efforts of the IT division of a large insurance company.

The Strategy & Planning group receives budget forecasts from several hundred projects that involve contribution from IT resources. These are submitted on a rolling basis. From this information, the overall IT budget and expenditure forecasts for given periods of time (e.g. - fiscal year, calendar year, quarter, month, etc…) are composed. From these forecasts, the budgetary committees of the company formulate the authorized budget expenses for fiscal and calendar years.

The Strategy & Planning group requested a review of this budget intake process with a goal to improve its accuracy. The method of analysis is open-ended, but the objective is to better understand how to accurately estimate future expenditure.

This project is important because the outcomes of past efforts to estimate IT expenditure have been mixed. For example, the 2017 annual IT budget forecast was ~**10% less than actual expenditure** at year end.

One consequence of this error is that millions of dollars, (actual amount redacted) which had been assigned to particular projects, went unused. These unused funds could have been allocated to other projects which, due to presumed resource constraints, were rejected, delayed, or reduced in scope and effectiveness.

This is an incredible opportunity cost in progress for the entire organization. Each project that is not allowed to go forward is a delay in achieving the outcome of that project. The domino effect of this is hardly quantifiable without intimate knowledge of all of the projects rejected (a list of which I have yet to locate, if it exists), but one can imagine that the amounts in dollars alone could be staggering.

The formal objectives of this project are listed below. The first list are the original objectives. The second list are the updated objectives which were reimagined as project execution progressed.

Original Objectives:

a. Build forecast accuracy score for submitted project budgets
b. Develop forecast accuracy model for incoming initiative budget data
c. Construct revised 2018 IT budget forecast
d. Build automated process to repeat analysis for future use
e. (Stretch Goal) Replicate process for similar dataset
f. Lay groundwork and provide supporting data for process improvement project related to improving division-wide budget forecasting
g. (Possible Outcome) Find areas of the organization or methods of practice where consistent inaccuracies occur
    i. (Stretch Goal #2 (likely out of domain)) Recommend methods to address these areas of concern

Updated Objectives

a. Generate S-Curves (cumulative sum of project expenditure over time) for use in goodness-of-fit tests.
b. Define a method for determining whether a submitted budget forecast follows a reasonable (typical) S-Curve.
c. Recommend future steps for continuing this analysis

The process observation was conducted in the form of extracting data from SQL Server as an MS Excel file output. The analysis was performed using the R programming language inside of the RStudio IDE (integrated development environment). Some visualizations were constructed using Tableau for use in Design Review presentations.

Chapter 2 – Literature Review

The literature review portion of this progress was an ongoing effort over the entire course of the project. The method of discovering appropriate literature was entirely based on web searches for key words such as “project forecasting,” “budget forecasting in R,” or “goodness of fit tests in R.” This turned out to be a reasonably effective method and provided a wide range of publication formats for review. Quora.com and stackoverflow.com provided great technical opinions from active research communities for choosing analysis methods and how to implement analysis with R.

Blogs were also a great source of information for project management context, and academic journals helped to provide explicit formulation of statistical analyses.

It is truly amazing how many examples, tutorials (especially for technical, computer-based analysis) and instructive content there is, and how active research communities are on the internet. Below is a representative list of resources used, but it is by no means comprehensive. Through the course of this project at least 50, possibly as many as 100 different resources were explored for context and instruction in relevant areas. Some of these resources will also be repeated in the references section of this paper, but many contributed creating combined perspectives for approaching the challenges at hand, though on their own may not have been comprehensively helpful.

- Murmis, G. M. (1997). "S" curves for monitoring project progress. *Project Management Journal, 28*(3), 29–35.
- Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". *Journal of the American Statistical Association*. American Statistical Association. 69 (347): 730–737
- http://www.maxwideman.com/guests/s-curve/intro.htm
- Introduction to Multivariate Regression Analysis
- E C Alexopoulos Hippokratia. 2010 Dec; 14(Suppl 1): 23–28.
- https://earlygrowthfinancialservices.com/bottom-up-vs-top-down-forecasting-realistic-financial-planning/
- http://seaopenresearch.eu/Journals/articles/SPAS_4_56.pdf
- https://www.sciencedirect.com/science/article/pii/0169207095006478
- http://www.academia.edu/6679293/THE_ACCURACY_OF_THE_BUDGET_FORECASTING_IN_LOCAL_GOVERNMENTS_IN_POLAND
- https://repositorio.ucp.pt/bitstream/10400.14/18798/1/Master_Thesis_Marcus_Wienhold_final.pdf
- https://brage.bibsys.no/xmlui/bitstream/handle/11250/169653/valuckas%202012.PDF?sequence=1
- https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials

- https://www.sas.com/en_us/insights/analytics/machine-learning.html

Chapter 3: Approach and Organization

Problem Solving Approach

One of the key challenges is the large number of projects being evaluated (~950, give or take, depending on if any have been added or removed month-to-month), and with that, the large quantities of erroneous or "bad" data. Outliers exist, but sometimes they are due to "bad" data, and sometimes they are due just to being typical outliers. This makes it difficult to report summary statistics and generalized findings.

So, for the above reason, doing a bottom up approach is much more labor intensive than originally anticipated. A bottom-up forecasting approach requires analyzing each individual project, then aggregating the outputs to form the portfolio forecast. A top-down approach, on the other hand, entails analyzing at the entire portfolio as a whole without looking at individual projects (Investopedia).

Bottom-up analysis is still the ultimate goal of this endeavor, but for the purpose of this project is infeasible given the time constraints. It is now understood that it will take longer than the time available for this portion of the project. Instead, the project has become a basis for formulation of recommendations for ongoing analysis which will include both bottom-up and top down efforts.

In this spirit, much of the labor for this project (all of it except for report writing and presentation preparation) has been observing the process, gathering observational data (in Excel extract format), and manipulating it to construct a data-flow that can be easily used to enable future regression analysis via a machine learning.

Requirements

- Generate S-Curves (cumulative sum of project expenditure over time) for use in goodness-of-fit tests.
- Define a method for determining whether a submitted budget forecast follows a reasonable (typical) S-Curve.
- Recommend future steps for continuing this analysis

System Overview

The processes being analyzed are twofold.

1) The creation of budget forecasts for individual IT projects

2) The creation of the overall IT portfolio budget forecast based on the aggregation of said individual project estimates

Both of these processes have inherent variability. Individual project budgets are created by hundreds of different project owners with varying degrees of experience, aptitude and methods for generating these budgets.

Forecasts then inevitably have varying degrees of accuracy relative to actual expenditure.

The variability in the accuracy of individual projects generates inaccuracy in the aggregate forecast.

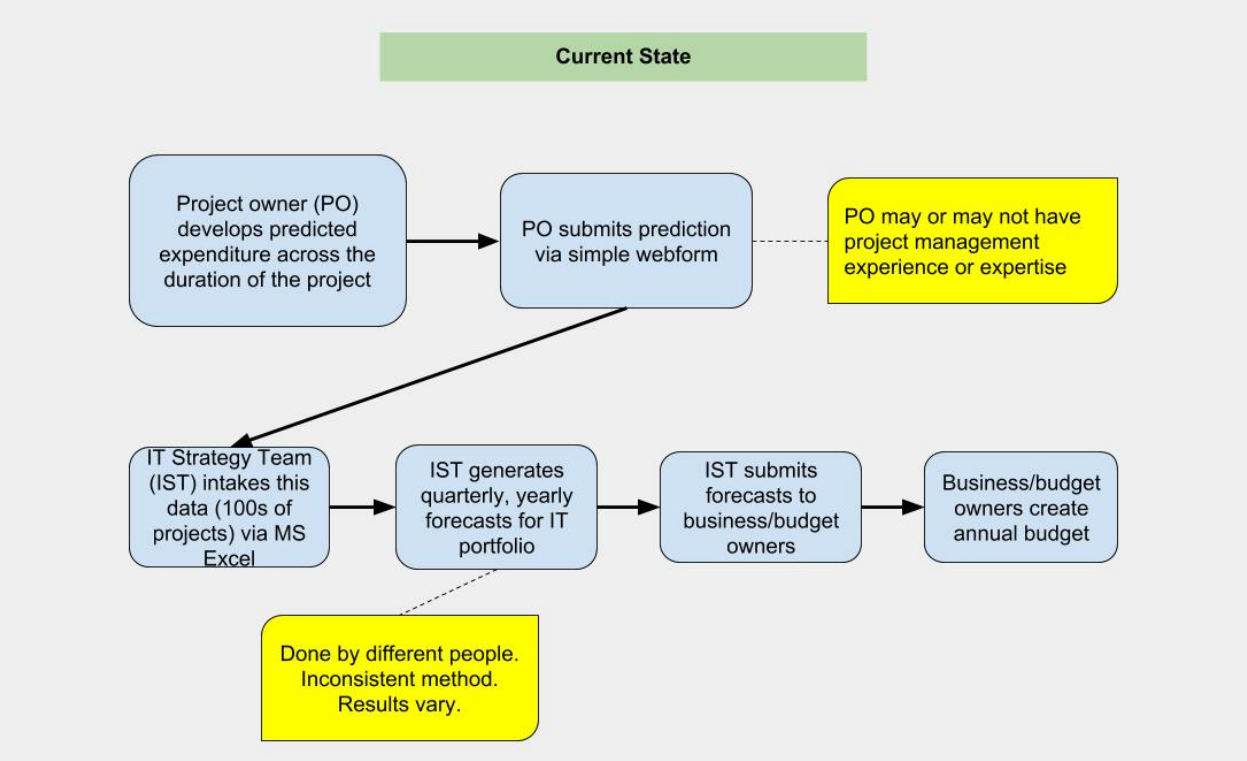
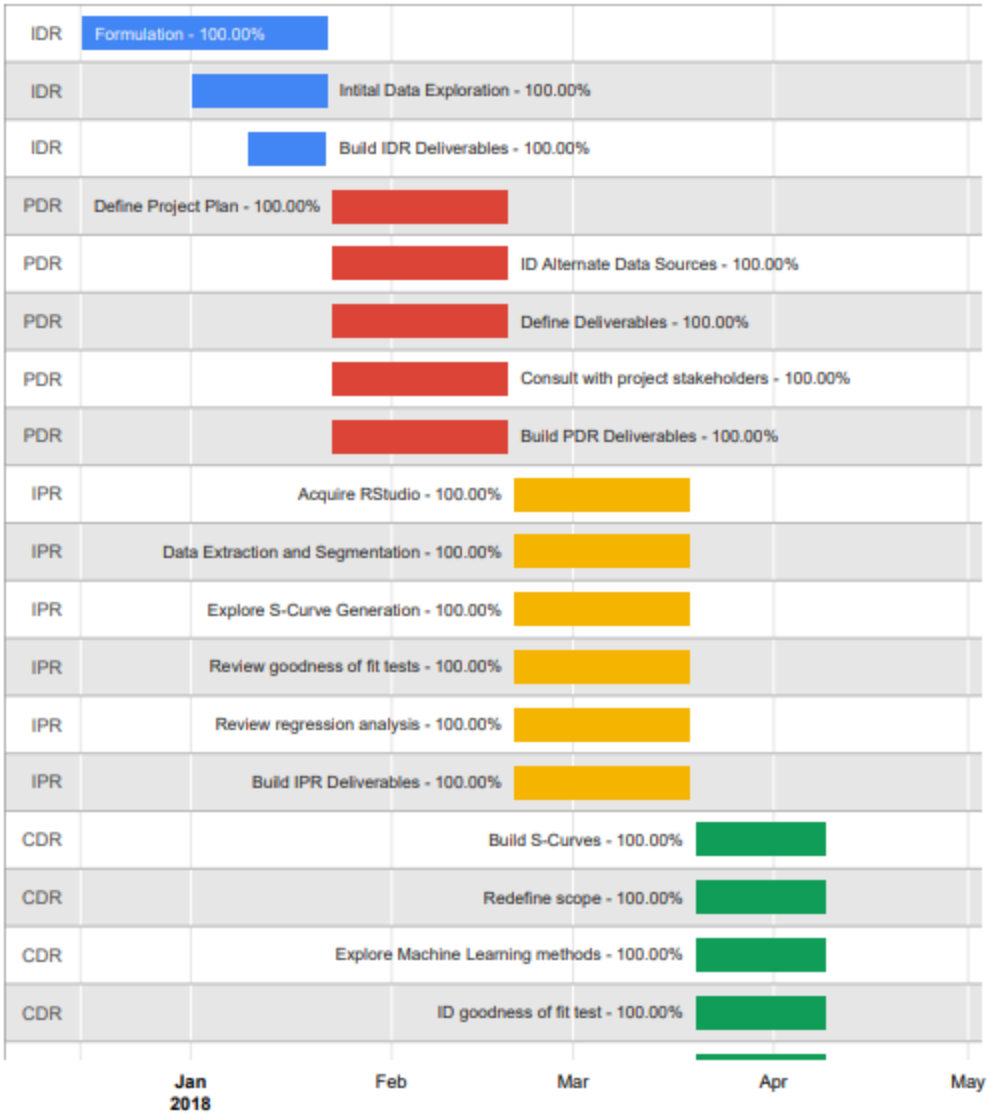


Figure 1- Current Budget Creation Process

Gantt Chart



Roles & Responsibilities

Analyst – Joel Fillmon

- Research problem objective and methods of meeting this objective
- Perform analaysis
- Construct report deliverables

Project Manager – Rolf Erickson

- Help guide the searching and provisioning of access to data within company data storage
- Act as an advisor on managing scope of the project
- Provide feedback on what the relevant working groups in the organization could be able to do, and would be interesting in doing, in the future with findings and labor from this project

Budget & Materials

The resources available and required for this project are listed here. They consistent entirely of computer software, with the sole exception of the company laptop which enables access to sensitive information. The cost of these items (many of which are free) is essential negligible as SQL Server, Microsoft OneNote, and the laptop are provided as a part of the overall IT environment, and not specific or exclusive to this project.

- SQL Server
- Tableau
- R
- Microsoft OneNote
- Company provided laptop w/ VPN access to financial information

Chapter 4 – Data Management and Segmentation

Structural Refinement of Dataset

The data provisioned for this project exists in an MS Excel file which consists of 22228 rows and 88 columns. This number of rows changes each month as projects are added or removed. Therefore, the data segmentation must by capable of dynamically adapting to the changing file.

Of these rows and columns, as of 4/30/18, only 3231 rows are of interest, and 64 columns.

By reducing the data to contain only qualitative data and a single column for a matching key (project ID), the information can be dynamically updated, and in the future, programmatically attached to any number of related categorical variables.

Additionally, quantitative calculations are made much simpler and easier to follow in the code when the management of categorical variables is kept to a minimum. So the process shown below reduces the data to the matching key, and monthly expenditure (and estimated expenditure) of projects. By retaining the key, previously removed, or soon-to-be-generated categorical variables can be reassigned (or assigned) to the appropriate project IDs.

Below is a sample of code used, written in R.

This example demonstrates the following:

- an import of previously extracted aggregate data,
- the removal of 20 categorical variables,
- the subset creation of the quantitative entries into three categories entitled Actuals, Target, and Revised.
- Finally, project duration and total cost is calculated by row calculation.

```
#import file
test_set <-
read.csv("C:\\Users\\AF37263\\Documents\\@ForecastProject\\hde_mva_wave2_feb\\hde_m
va_wave2_impact.csv")

#Remove categorical variables and build new table
Essential_Data <- test_set[c(6, 13, 21:81)]

#select only Data where "Metric" = One-time IT Labor
ED1 <- subset(Essential_Data, Metric == 'One-Time Implementation Cost - IT Labor (IT)')

#replace NA with "0" (zero) (If desired)
ED1[is.na(ED1)] <- 0



#Sort by "purpose" (target, actuals, revised)
ED1_Actuals <- subset(ED1, purpose == 'Actuals')
ED1_Target  <- subset(ED1, purpose == 'Target')
ED1_Revised <- subset(ED1, purpose == 'Revised')

#Just Initiative & Expenditure (remove purpose and metric columns)
ED1_Actuals <- ED1_Actuals[c(1, 4:63)]
ED1_Target  <- ED1_Target[c(1, 4:63)]
ED1_Revised <- ED1_Revised[c(1, 4:63)]

#new column that sums total expenditure across rows
ED1$Total_Cost <- rowSums(ED1[,4:63])

#new column that displays count of columns with values per row ((in months))
ED1$Project_Duration <- rowSums(ED1 != 0)
```

Chapter 5 – Modeling & Analysis

S-Curves

An S-curve is defined as:

"A display of cumulative costs, labor hours or other quantities plotted against time. The name derives from the S-like shape of the curve, flatter at the beginning and end and steeper in the middle, which is typical of most projects. The beginning represents a slow, deliberate but accelerating start, while the end represents a deceleration as the work runs out."(Garland)

- S-Curves allow the progress of a project to be tracked visually over time, and form a historical record of what has happened to date.
- Analyses of Percentage S-curves allow project managers to quickly identify project percentage **growth**, and percentage **slippage (time delay)**
- Comparison of the Target Revised, and Actuals S-curves quickly reveals if the project has grown or contracted
- **Growth** (Target or Actuals S-curve finishes above Baseline S-curve) or **Contraction** (Target or Actuals S-curve finishes below Baseline S-curve) in scope
- Additionally, whether duration of the project will **increase** (finish later) or **decrease** (finish earlier) can be identified.

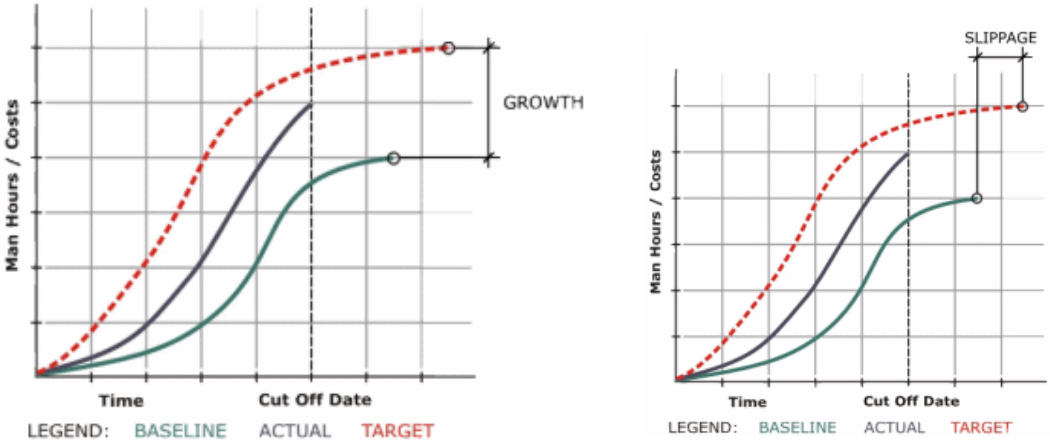


Figure 2- Example S-Curves

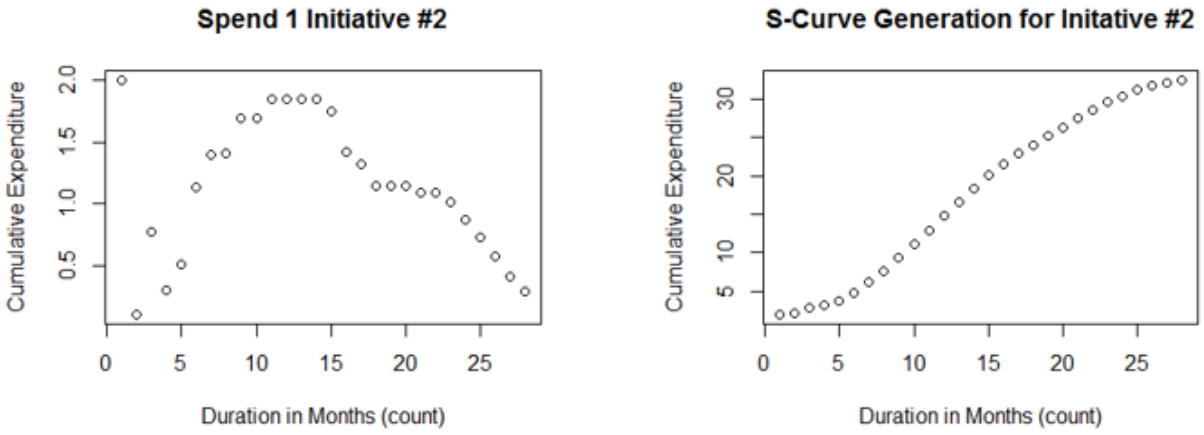


Figure 3- R-Generated Curve (Initiative #2)

As shown above in Figure 3, Initiative #2 has a relatively typical S-Curve when compared to the examples in Figure 2. Below, Figure 4 (using Tableau) overlays Initiative #4 (Blue) with Initiative #37 (Green)

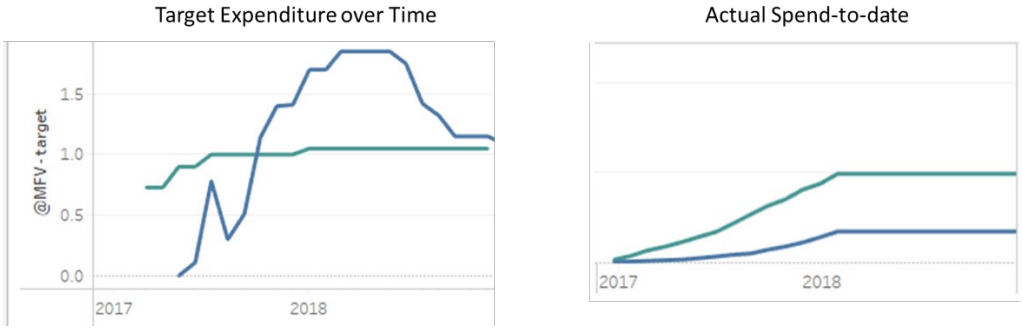


Figure 4- Comparison of Health vs. Unhealthy Projects

The Blue Target Expenditure and Actual Spend curve reflect the appropriate behavior of Figure 3.

- Actual spend to-date plateaus in accordance with target reducing over time.

The Green Actual Spend Curve is displaying appropriate behavior as well, but as you can see, the Target Expenditure plots vary greatly between Green and Blue.

This is an example of a poor Target Expenditure Estimate. If the estimate were accurate, the Green Actual Spend curve would form roughly straight line from bottom left, to top right of the display.

S-Curve Generation with 'R'

```
# remove columns 2 & 3 (totals)
Sigmoid_Actuals <- subset(ED1_Target, select = c(-2,-3))

# Transpose Table from Row to Column Vectors
Trans_Sigmoid_Actuals <- t(Sigmoid_Actuals)

# Isolate Iniative #2 (in column 1)
Reduction <- subset(Trans_Sigmoid_Actuals, select = (c(1)))

#Test_Index <- which(Reduction != 0)
Test_Plot_Data <- subset(Reduction, select = (r(Test_Index)))

# This one extracts the values are not zero
Test2_Index <- Reduction[Reduction > 0, ]

# plots each value  ## plot(x, y, )
plot(Test2_Index, , main = "Spend 1 Initiative #2"
                        , xlab = "Duration in Months (count)"
                        , ylab = "Cumulative Expenditure")

# plots cumulative sum
plot(cumsum(Test2_Index), , main = "S-Curve Generation for Initative #2"
                                    , xlab = "Duration in Months (count)"
                                    , ylab = "Cumulative Expenditure")
```

Comparing S-Curves

To meet goal number two, a method to compare these generated curves programmatically is required. It is easy enough to look at graphs such as Figure 4, but to do this comparison for 900+ projects is far too labor intensive for frequent or on-demand reporting.

By using a Goodness-of-Fit test, a score can be applied to how well these curves match a baseline curve. Since the x-values, or durations, of the project curves vary, a test which can accommodate this variation is required. Literature review uncovered the use of the Kologorv-Smirnov (KS) Test for this application.

The KS test is a nonparametric test calculates the distance between the empirical distribution function (ECDF) of the sample, and the cumulative distribution function of the baseline. This distance score respresents how closely the two match (Stephens). See figure 5 for a visualization.

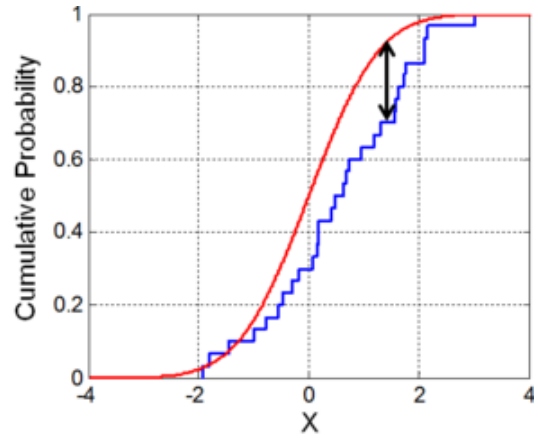


Illustration of the Kolmogorov–Smirnov statistic. Red line is CDF, blue line is an ECDF, and the black arrow is the K–S statistic.

Figure 5 – Kolorogov-Smirnov Test (Smirnov Test, Wikipedia)

The drawback to the KS test is that it is very sensitive. Slight variations can give results outside of typical goodness-of-fit test responses, and it is easy to find Type I errors (False positive of the null hypothesis) in hypothesis testing. However, with more repetitions of this test to well-fitting curves, it will be possible to better understand how to interpret these results in the context of this analysis.

Chapter 5

Results and Discussions

The construction of these curves is complete for all projects, but tuning the KS Test is still underway. Once this is complete, the results can be used to generate a baseline curve that fits typical project behavior. It is likely that there will be groupings of project behavior, and that there will not be just one definition for "typical." In this case, a regression analysis can be applied to understand if certain categorical variables are related to these groupings of behavior.

The regression will be multi-variate. In the multiple linear regression model, Y has normal distribution with mean:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta\rho X\rho + \sigma(Y), \quad sd(Y) = \sigma \text{ (independent of X's)}$$

The model parameters $\beta_0 + \beta_1 + + \beta_\rho$ and σ must be estimated from data.

β_0 = intercept, β_1 β_ρ = regression coefficients, $\sigma = \sigma_{res}$ = residual standard deviation (Hippokratia).

Potential categorical variables for regression analysis include;

- By cost type (labor, hardware, etc…)
- By duration/magnitude of project
- By internal organizational structure (two types of classification)
    - "Control Tower"
    - Initiative

From this analysis, causal, or predictive factors may be identified.

Chapter 6: Conclusions

<u>This project started with three aims;</u>

1) Identify the points in the budget process at which generate variation between the estimated and actual project expenditure.
2) Identify and quantify variation in project budget forecasts.
3) Develop a method or process that can correct for the variation and predictively model future budget expenditure

In addressing these aims, 1) and 2) have been sufficiently accomplished. It is understood that project variation derives from inconsistencies in budget estimation by project owners. The general, and percent differences between these projects estimates and their actuals is also recorded within generated data using R.

However, it has become more and more apparent that aim 3) is a more pronounced challenge than originally anticipated.

Therefore there are two final thoughts and recommendations.

1) Data manipulation and cleaning is incredibly time consuming (and sometimes difficult). If the data can be collected and recorded in a specific format beforehand, it would make analysis much more efficient.
2) Contemplating 900+ projects piecemeal is untenable for rapid analysis. Therefore, a programmatic and automated process should be developed to this end. The data this process works from should be built from a machine learning application that learns how to compare S-Curves.

The estimation for budgeting this recommendation is as follows:

2-3 employees ($35-50/hr)

20-25 hours/week for 6-8 weeks.

10 hours/week for continuing work for ~16 weeks

= $11,200 - $15,000

References

(n.d.). Retrieved from https://www.investopedia.com/exam-guide/cfp/theory-portfolio/cfp4.asp

Garland. (n.d.). Retrieved from http://www.maxwideman.com/guests/s-curve/what.htm

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". *Journal of the American Statistical Association*. American Statistical Association. 69 (347): 730–737

Introduction to Multivariate Regression Analysis
E C Alexopoulos Hippokratia. 2010 Dec; 14(Suppl 1): 23–28.

Appendix A: Acknowledgements

Rolf Erickson, for keeping the scope of the project under control.

Dr. Adeel Khalid, for keeping the schedule of the project on track.

Appendix B: Contact Information (Student and Advisor Contacts)

Student

- Joel Fillmon
    - E-mail: joel.fillmon@gmail.com

Manager

- Rolf Erickson
    - Email: rolf.erickson@anthem.com

Faculty Advisor

- Dr. Adeel Khalid
    - Email: akhalid2@kennesaw.edu

Appendix C: Reflections

- The Educational Experience
    - Incredibly robust. First opportunity to deal with data extraction at this level of complexity. Was able to leverage previous coursework in R, but had to do a lot of self-teaching, attending study lab in Data Analysis department, and general trial & error.
    - Also working with scoping a project. The original ambitions were possibly doable in the timeframe, but I think only if I was focused on this full-time without other coursework and job responsibilities.
- Challenges Faced
    - Getting access to the data. Spent a couple of weeks early on trying to gain access to databases that I do not have authority to access. Changed tactics to

used data extracts in an Excel file, but this had its challenges because the format was not the easiest to use. Instead of being able to write custom SQL queries from the database which arranged the data in easy to use formats, I spent tens of hours doing data manipulation with R and am still facing challenges with this format. There have been a couple of breakthroughs recently on this front though. As I gain more experience with R and with manipulating the data I am better understanding what sequences of action to perform for efficient data segmentation.

Appendix D: Contributions

- Joel Fillmon contributed entirely to each section of this and all deliverables, as well as all technical contributions