Proceedings of the International Conference on Industrial Engineering and Operations Management
Washington DC, USA, September 27-29, 2018

Using Analytics to Predict Project Management Success

Nichalin Summerfield, Juheng Zhang, Luvai Motiwalla, Tri Mai, and Kasey Mazza
Department of Operations and Information Systems
Manning School of Business
University of Massachusetts Lowell
Lowell, MA 01854, USA
Nichalin_Summerfield@uml.edu, Juheng_Zhang@uml.edu, Luvai_Motiwalla@uml.edu,
Tri_Mai@student.uml.edu, Kasey_Mazza@student.uml.edu

Abstract

IT projects normally face time and cost overruns challenges. Predicting which projects will not be completed by the expected end date (time overrun) and within the allotted number of hours (man-hours overrun) helps manage company's employee utilization and avoids overscheduling. This study applied a data analytics framework on IT project management dataset to predict the project completion time and overruns (in man-hours). We used linear regression and classification models to predict projects' performance and analyze the underlying factors causing project delay. Our predictive analysis used 131 variables which included 536 tasks, 138 resources, 69371 employee hours, 72 contractors that were assigned to 434 projects. We also calculated two new variables *closeness* and *betweenness* among project team members. Results showed that Decision Tree outperformed SVM, ANN, LDA, and logistic regression in predicting man-hours overrun. In addition, preliminary Social Network Analysis (SNA) indicates that *Avg-closeness* and *Avg-betweenness* did not improve prediction on the overall amount of time and man-hours overrun but improve the prediction on time overrun, but task, resource and contractor assignments variables were significant at p-value of .01. The models we used helps identify key predictors of project performance and provide insights into the company's resource management.

Keywords

Project Management, Data Analytics, Decision Tree, Logistic Regression, Social Network Analysis

Acknowledgements

We are grateful to a local company for providing their project management data for our analysis.

Biographies

Nichalin Summerfield is an Assistant Professor of Operations & Information Systems at UMass Lowell. She received her Ph.D. in Management with a concentration in Operations Management and a master's degree in Management Information Systems from the University of Arizona. Her research interests are in the areas of game theoretical analysis and business analytics in operations and supply chain management.

Juheng Zhang is an Associate Professor of Management Information Systems at UMass Lowell. She received her Ph.D. from University of Florida. Her research focuses on data analytics and examines information manipulation on decision makings. She has published in Information Systems Research, European Journal of Operational Research, Decision Support Systems, Information Systems Frontier, and other academic journals.

Luvai Motiwalla is Professor of Management Information Systems at UMass Lowell. He received his Ph.D. from the University of Arizona in 1989. He has two books, several journal articles in top MIS journals and recipient of four grants from NIH, NSF, DoE and Davis Foundation. He was the Chair of the OIS Dept. from 2008-11 and 2015-18.

Tri Mai is a graduate student in the MS in Business Analytics program at the UMass Lowell.

Kasey Mazza is a graduate student in the MS in Business Analytics program at the UMass Lowell.

Appendix *[optional]*

A.1 Introduction and Motivation

Many IT projects face time and cost overruns. Being able to accurately predict which projects will be completed by the expected end date and within the allotted number of hours will have long-term benefits for company in managing its resources and will result in better communication, less overscheduling, and manageable expectation.

We were given access to IT project management dataset of one of the well-known asset management companies in the United States. The company is constantly evolving and improving their technology to adapt quickly to the fast-paced financial markets. The company uses a proprietary software to track and manage their infrastructure and technology projects.

A.2 Methodology

After pre-processing the data to remove missing values, we applied a data analytics framework to find trends and patterns in the dataset and predict three project performance measures: 1) the overall amount of time, 2) whether a project would finish after the expected end date (time overrun), and 3) whether a project would exceed the allotted number of hours (man-hours overrun). Our analysis encompassed several variables including the count of tasks, allotted man-hours, contractors' headcounts, billing offices, and *closeness and betweenness* among project team members. We employed multivariate regression to predict the overall amount of time. Then, Decision Tree (DT), SVM, ANN, LDA, and logistic regression (LR) were applied to predict time and man-hours overruns.

While a wide range of machine learning algorithms have been applied to project management [6, 7], social network analysis (SNA) has not yet been applied. A key factor to IT project success in close collaboration between all members of project team, therefore SNA is applied to derive the features that determine the level of connection between resources. The hypothesis is that there is a relationship between resources who have worked together on previous projects.

A.3 Exploratory Data Analysis Results

Exploratory data analysis was performed to summarize the data's main characteristics and to better understand the data and its distribution. The distribution of each numeric variable was calculated and can be seen in Figure 1.

	numTasks	numTasksAssigned	numHrsAssigned	numResourcesAssigned	numContractors	AverageCloseness	AverageBetweeness
count	434.000000	434.000000	434.000000	434.000000	434.000000	434.000000	434.000000
mean	18.569124	35.926267	2060.768001	19.322581	5.518433	0.567656	2219.046296
std	18.309655	48.701372	5583.299665	19.589131	8.837035	0.029707	1385.693432
min	1.000000	1.000000	3.000000	1.000000	0.000000	0.408742	65.707817
25%	7.250000	11.000000	159.875000	7.000000	0.000000	0.552479	1271.000112
50%	12.000000	20.000000	525.050000	13.000000	3.000000	0.570649	2042.250016
75%	23.750000	40.000000	1555.875000	25.000000	6.000000	0.585953	2887.109582
max	163.000000	536.000000	69370.990000	138.000000	72.000000	0.696821	13074.351950

Figure 1: Distributions of numeric attributes

Figure 2 shows the number of projects by actual finish date. It was observed that the number of projects finished tends to decrease in the summer and spike at the end of each year.

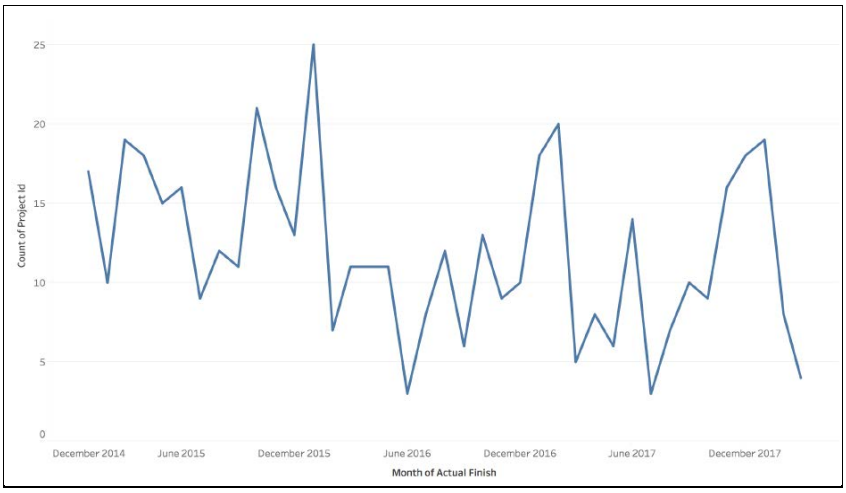

Figure 2: Trends in actual finish date over time

A.4 Predictive Analytics Results

A4.1 Predicting overall amount of time

The p-values from the linear regression is shown in Figure 3. "CreatedAndUpdatedSame" referred to whether a project was created and later updated into the project management software by the same employee, which was significant in our model. It was observed that the billing office fields as well as "averageCloseness" and "aveargeBetweenness" were the least significant variables in the model. This was an unexpected result. The features from social network analysis indicate how connected the resources are to each other. Researchers have long hypothesized that knowledge sharing and knowledge gaps play a significant role in a project's success [1,5]. However, this analysis rejects that hypothesis by showing that "averageCloseness" and "averageBetweenness" were not significant factors in predicting the actual number of hours that a project will take. A ridge model was created using the parameter of alpha equal to 0.1. This alpha was relatively low to avoid underfitting but not low enough to encounter overfitting. The ridge model was used because it regularizes the data. The data was split into training and testing sets, with the testing data representing 30% of the entire dataset. The training data was used to fit and train the model, while the testing data was used to validate the level of success of the newly created model. Overall the model resulted in a R^2 value of 0.741.

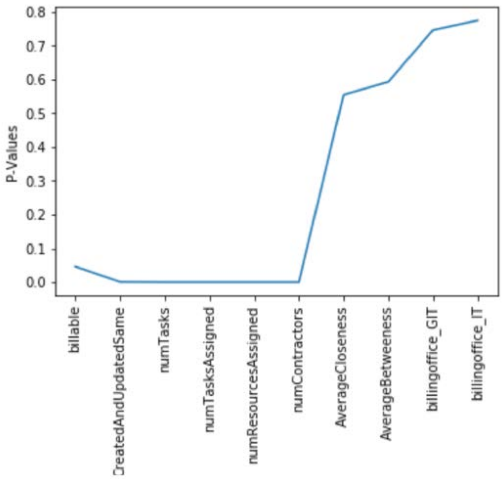

Figure 3: Significance of each attribute for regression

A4.2 Predicting whether a project will be late

All numerical variables were discretized by bucketing. New variables were created due to discretization. K-fold cross-validation was then used to avoid overfitting. Due to unbalance data (very high percentage of late projects) the synthetic minority over-sampling technique (SMOTE) were used [2, 8]. The accuracy, specificity, and sensitivity of the mode were 67.3%, 64.7%, and 69.9%, respectively. It should be noted that, contrary to the regression model, the social network analysis features were important in this model.

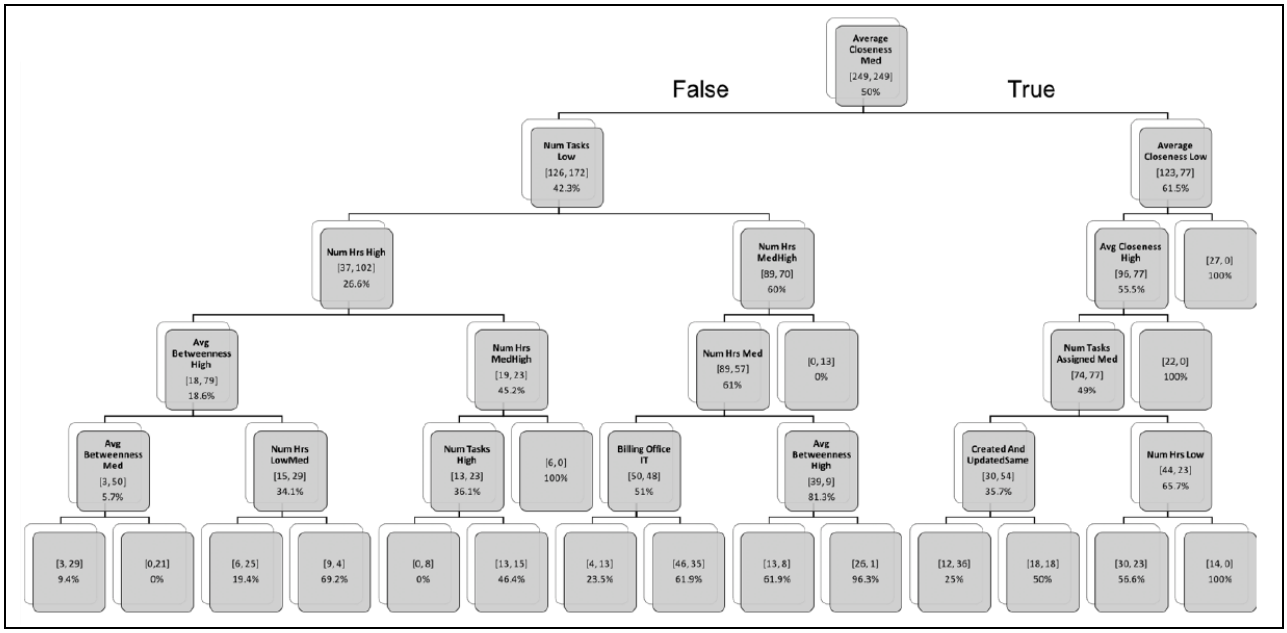

Figure 4: Rules created by decision tree classifier

A4.3 Predicting whether a project will exceed the allotted number of hours

Since the target variable has imbalanced population (139 projects over the baseline and 295 projects is not over the baseline), random over sampling (ROS) was applied to rebalance the minority class so that the machine learning algorithms can provide more accurate result [4, 6]. 139 projects with minority classes were randomly replicated so that it would balance to the other class. After balancing process, the dataset contained 590 projects: 295 projects over the baseline and 295 projects not over the baseline. A variety of machine learning algorithms were used to predict the target variable to find the best performing model for this dataset. The results are shown in Table 1.

	Accuracy	Sensitivity	Specificity
Decision tree	0.683	0.705	0.661
LDA	0.634	0.661	0.607
Logistic	0.612	0.607	0.617
SVM	0.509	0.62	0.475
ANN	0.500	0.298	0.702

Table 1: Results of various classification models for predicting project hours

The decision tree classifier performs the best in terms of accuracy and sensitivity and it scores second-best for specificity. The decision tree classifier generates a set of rules that can be easily understood. The rules are shown in Figure 11. For example, the top end node has 41 on-time projects and 6 late projects, which is very good in comparison to current reality. This node includes projects that have less than or equal to 198 baseline hours, less than 18 or equal to tasks assigned, belongs to the Information Technology billing office, and has less than or equal to 8 contractors assigned.

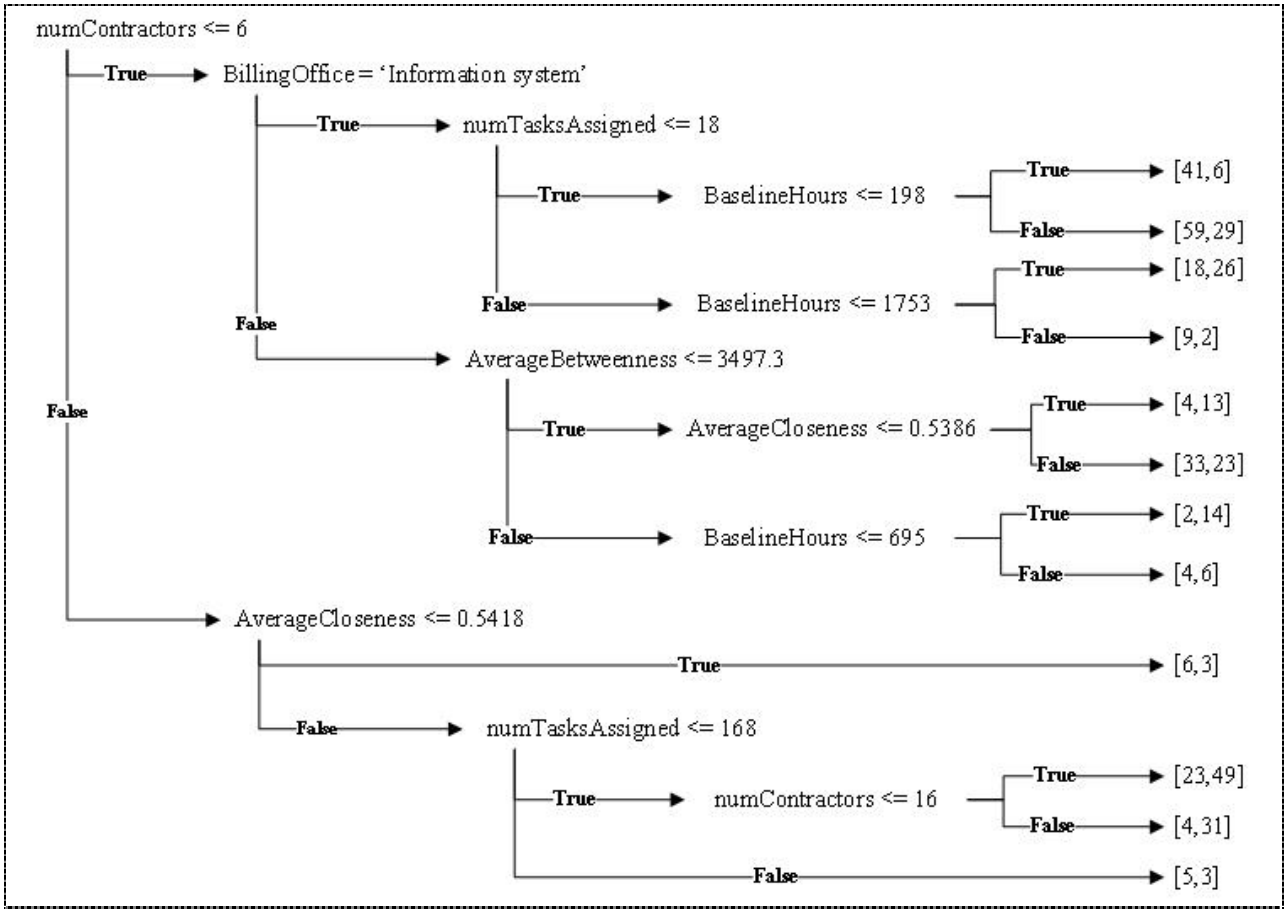

Figure 5: Rules generated by decision tree classifier

Predicting overall amount of time	Predicting whether a project will be late	Predicting whether a project will exceed the allotted number of hours
numContractors	numHrsAssigned	numContractors
numTasks	numTasks	numTasksAssigned
numTasksAssigned	AverageCloseness	numTasks
numResourcesAssigned	numTasksAssigned	BaselineHours
CreatedAndUpdatedSame	AverageBetweeness	CreatedAndUpdatedSame

Table 2: Top 5 variables for each model

References

[1] Atkinson, Roger. (1999). Project management: cost, time, and quality, two best guesses and aphenomenon, its tie to accept other success criteria. International Journal of Project Management. Vol. 17, No. 6, pp. 337-342.
[2] Boldi, Paolo & Vigna, Sebastiano. (2013). Axioms for Centrality. Internet Mathematics. 10.10.1080/15427951.2013.865686.
[3] Brocke, H., Uebernickel, F., & Brenner,W. (2009). Success factors in it-projects to providecustomer value propositions. In 20th Australasian Conference on Information Systems, Australia.
[4] Costantino, F., Di Gravio, G., Nonino, F. (2015). Project selection in project portfolio management: An artificial neural network model based on critical success factors. International Journal of Project Management. Vol. 33, pp. 1744-1754.
[5] Cooke-Davies, Terry (2002). The 'real' success factors on projects. International Journal of Project Management, 20(3), 185–190.
[6] Dvir, D., Ben-David, A., Sadeh, A., and Shenhar, A. J. (2006). Critical managerial factors affecting defense projects success: A comparison between neural network and regression analysis" Eng. Appl. Artif. Intell., vol. 19, no. 5, pp. 535–543.
[7] Gingnell, L., Franke, U., Lagerström, R., Ericsson, E., and Lilliesköld, J. (2014). Quantifying Success Factors for IT Projects—An Expert-Based Bayesian Model," Inf. Syst. Manag., vol. 31, no. 1, pp. 21–36.
[8] Marchiori, Massimo and Latora, Vito. (2000). Harmony in the small-world. Physica A: Statistical Mechanics and its Applications, 285(3-4):539 – 546.
[9] MFS. https://www.mfs.com/en-us/what-we-do/corporate-fact-sheet.html.
[10] Reel, J. S. (1999). Critical success factors in software projects. IEEE Software, 16(May/June),18–23.
[11] Xia, B. W., & Lee, G. (2004). Grasping the complexity of is development projects. Communications of the ACM, 47(5), 68–74.
[12] Yeo, K. (2003). Critical failure factors in information system projects. International Journal of Project Management, 20(3), 241–246.