Data Science Approach for IT Project Management

Janis Grabis
Department of Management Information
Riga Technical University
Riga, Latvia
grabis@rtu.lv

Bohdan Haidabrus
Computer Science Department
Sumy State University
Sumy, Ukraine
haidabrus@gmail.com

Serhiy Protsenko
Department of Electronics, General and Applied Physics
Sumy State University
Sumy, Ukraine
serhiy.protsenko@gmail.com

Iryna Protsenko
Computer Science Department
Sumy State University
Sumy, Ukraine
protsenko3107@gmail.com

Anna Rovna
Computer Science Department
Sumy State University
Sumy, Ukraine
a.rovna@mk.sumdu.edu.ua

***Abstract*—Majority of the IT companies realized that ability to analyse and use data, could be one of the key factors for increasing of number of successful projects, portfolios, programs. Key performance indicators based on data analysis helps organizations be more prosperous in a long term perspective. Also, statistical data are very useful for monitoring and evaluation of project results which are very important for managers, delivery directors, CTO and others high level management of company. The Data Science methods could make more efficient project management in several of business problems.**

Analysis of historical data from the project life-cycle based on Data Science models could provide more efficient benefits for different stakeholders. Differential of the project data vector with target as an integral evaluation of the project success which allow for the complex correlations between separate features. Therefore, the influence of features importance and override creatures could be decreased on the target.

This study propose new approach based on Data Science providing more efficient and accurately project management, taking into account best practices and project performance data.

Keywords— Machine Learning, Data Analysis, Project Management, Business Processes.

I. INTRODUCTION

Nowadays, the ability to analyse and use data is one of the key factors affecting the organization’s ability of the IT companies to work effectively in the long perspective. The main factor affected the successful implementation of the projects could be described with so-cold project management triangle. The triangle consists of following components: time, budget and quality (or scope) [14, 19]. The are several ways to calculate indicators related to the deadlines:

- assessment of the task in hours of the developer who will be engaged in it (previously agreed with the developer himself);
- have the data on how many hours were actually spent on this task (this requires a time tracker);
- due date due to which the task should be ready.

From the other hand, it is possible to convert part of the time metrics to the project budget. The evaluation of the project time scale could be described as - *Start date*, *Due Date*, *Actual Date*. Specifically, the time metric can be measured as the number of deadlines per task, or the ratio of differences between *Start date* / *Due Date* and *Due Date / Actual Date*. The budget indicator is based on a preliminary assessment of the time and the actual time spent to the task [10].

Regarding the project budget, with respect of the price determination our research consists of three project options:

- *The fixed cost project* does not imply a deviation from the budget characteristics;
- *The time and Material project* means that the cost is linearly related to the hours spent, which means that deviations from the budget are identical with deviations from the deadlines. Thus, in these cases, there are virtually no budget metrics;
- *A Flexible fixed cost project*, allowing coordinate the budget changes, it is potentially not clear which part of the budget change should be attributed to the “merit” of a particular developer.

In general, it means that in the project management triangle for an individual developer, budget figures cannot be adequately representing, thus it is necessary to focus on the metrics of project quality and time.

Project success depends on a large number of factors. They for it is necessary to build and analyze the system of key performance indicators at all stages and phases of the project life-cycle. A project team, project

Print ISSN 1691-5402
Online ISSN 2256-070X
http://dx.doi.org/10.17770/etr2019vol2.4163

Published by Rezekne Academy of Technologies.

manager, delivery manager, company's CEO and CTO required to have information on the current status of the project. Important to provide all key information about project management and realization: project status, time delay, according to budget, risk monitoring and control, requirements status from the project owner, feedback from the stakeholders and the need to generalize all parameters of the specific project.

The aim of our research is to develop and implement an approach for predicting the evaluation of the project efficiency which based on an integrated estimation of all parameters during the project life-cycle. Analysis of historical data with Machine Learning (ML) models [7, 9] with aim to do initiation and planning phases and better coordination with the different stakeholders.

II. MATERIALS AND METHODS

Using Data Science Approach for IT Project Management can be divided into two parts:

- *The Best Practices Approach:* historical integrated analysis of the best practices and fails on the projects;
- *The Historical Data Approach*: analysis of historical data for project initiation and planning depend on the customer requirements and actual status of the project by using data.

According to *The Best Practices Approach* we have analysed *The Sales, Project, Production Methodology* of the IT Companies MindK (mindk.com) [22] and Already On (alreadyon.com) [23]. The documentation is developed as a support document for MindK and Already On to secure a common framework on how we indemnify an optimal project and development process.

This methodology consists of the three general part: Sales/Initiation Process, Project Process and Production Process. The Sales process define the following steps: Identification, Quality, Offering, Negotiation, Pre Project, Signing, Specification phase. The Sales process always ends with a signed specification and acceptance criteria. The data which we can use to analyze from the Sales/Initiation Process are at the RASCI-roles matrix (Table 1) (variation of RACI):

- R - Responsible – Persons involved to achieve a task.
- A - Accountable – Persons ultimately accountable for the correct and thorough completion of the deliverable or task, and the one to whom Responsible is accountable. If there is no Responsible on the project, Accountable does the work.
- S - Support – Resources dedicated to Responsible. Unlike Consulted, who may provide input to the task, Support will assist in completing the task.
- C - Consulted – Persons who are not directly involved in a process but provide inputs and whose opinions are sought.
- I - Informed – Persons receive outputs from a process or are kept up-to-date on progress, often only on completion of the task or deliverable.

Roles in the Sales/Initiation Process are:

- Product Owner (PO)
- Sales person
- Business Analyst (BA)
- Project Manager (PM)
- UX specialist (Designer)
- Solution Architect
- QA specialist
- Development Team (Tech Lead, Developer, HTML-Developer, QA, System Administrator)
- HR manager.

Documents of Sales/Initiation Process related to: Contract, Specification, Acceptance criteria, Project Plan.

Table 1 RASCI matrix

Stage	Planning phase		
	To do	RASCI roles	Artefact/ Outputs
Resource planning	Resource availability analysis (human resources, software, etc.)	PM) – A HR manager – S	Team members list
	Plan acquisition of additional resources (if required)	PM – A HR manager – S	
	Plan education of new human resources (if required)	PM – A	
	Schedule resource involvement	PM – A	
Product backlog planning	Prepare a detailed work breakdown structure	PM – A Team – R	WBS, Gantt Chart
	Prepare Gantt Chart (if required)	PM – A Team – R	
	Assign tasks to team members (if required)	PM – A Team – R	
	Prepare a detailed work breakdown structure	PM – A Team – R	
Budget planning	Prepare a detailed hours break down based on the WBS and the final estimate	PM – A Team – R	Hours break-down structure
Risk management planning	Prepare a risk management plan (mitigation measures, responsible people, deadlines, etc) based on the risk management report	PM – A Team – C (PO - C	Risk management plan Risk status tracking document
	Prepare a risk status tracking document, which will be updated during the project period	PM – A	
	Prepare a risk management plan (mitigation measures, responsible people, deadlines, etc) based on the risk management report	PM – A Team – C PO - C	
Project planning	Prepare a detailed project plan based on the WBS, the resource schedule, the risk management plan	PM - A	Approved project plan
	Present to the client and approve	PM - A	
Final approve	Check that everything is ready to start development and receive final approval from the client	PM – A	Approve from the client

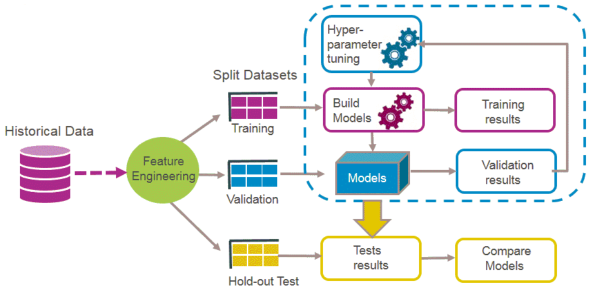
The Project process includes the following activities: Mandate / Scope, Risks, Milestones, Specification, Economy, Change Orders, Test and QA process and period. The documents of this Process are very useful for data analysis: Weekly reports, Change requests (added to specification), Go Live plan and Signed acceptance criteria [20].

And the last, but not list the Production Process is defined by the following activities:

- Acceptance from Customer,
- Invoicing,
- Support,
- Change Orders.

PM's daily duties include many different types of Business Processes and could be very regulated and clear described with using templates and tools according to the IT company standard. In our research we focused on the most important documents from the project life-cycle, which was used as an input data for analysis, in particular:

- *Project Charter*: the document created by PM and PO and consist of: General Information (Project name, Project manager, Data), Project Overview, Project Objectives, Project Type, Resource Costs and Estimates, Main stakeholders, Attachments.
- *Communication and Escalation Plan***:** how communication will be managed during Project Lifecycle, customer contacts, corporate and supplier contacts, etc.
- *Risk Management: Risk Matrix*, *Risk Register*. Including a list of risks which described by several criteria's: ID, Risk Description, Impact, Risk Value, Mitigation, Trigger, Owner, Deadline.
- *Stakeholder Sheet:* Formalize the expectation and standards of internal and external customers, suppliers, employees, sponsors, beneficiaries, etc.



ML life cycle model for historical data analysis.

Based on *The Sales, Project, Production Methodology* the massive of data with target as an integral evaluation of the project success which allow for the complex correlations between separate features. This approach, which is presented in "Fig. 1", allows to describe all complex correlations between different project features, to consider the features importance and to zero the importance of those features which have a weak effect on the target. It means, that the PM should consider such features as a low-priority [7].

The disadvantage of this approach may be the complexity of presenting different projects in one vector space. Because there are no identical projects and there is an another problem which associated with small number of observations in multidimensional space. Considering these remarks, using of classical ML models such as linear models, random forest or boosting will not always give good results but certainly more adequate than using deep learning based on small datasets.

III. RESULTS AND DISCUSSIONS

In our research, it is very important step of multidimensional project vectors is the initial preprocessing and features selections for this stage, we have used a statistical approach based on Bayesian statistics Automatic Relevance Determination (ARD) [11, 18]. As a result, we received smaller dimensional vectors with features that were of greater statistical weight and in this case the use of classical ML models gave a more adequate result.

Example 1. We have the original data - only 30 points (data on 30 projects). And each project has 30 signs. And the task was to create a regression model (for example, to predict project sales volume according to location, type, sales area, configuration, number of features and other project parameters).

Building ordinary linear regression under such conditions will be pure insanity. Let us further exacerbate the problem by the fact that only 5 signs really matter, and the rest are completely irrelevant data [21].

Thus, let the real dependence be represented by the equation:

$$Y = W \cdot X + e \qquad (1)$$

where, e - is a random normal error and the coefficients W are equal [1, 2, 3, 4, 5, 0, 0, ..., 0], that is, only the first five coefficients are nonzero, and the signs from the 6th to the 30th generally do not affect the real value of Y. We only have data - X and Y - and we need to calculate the coefficients W:

```
# prepare the data, separate target
and form Xtrain та Ytrain
Xtrain=data.iloc[:,1:-1]
Ytrain=data.iloc[:,-1]
#Check the size of our massive
Xtrain.shape, Ytrain.shape
```

Using the *ARDRegression()* and will see, which features are important and influence to the target:

```
ard = ARDRegression()
ard.fit(Xtrain, ytrain)
# The higher of coefficient value, so
it is more important for the target
print (ard.coef_)
```

And the results are:

```
 [ 1.92557895e-03  -9.51988416e-04
-2.13868725e-04  -7.21210062e-04
 1.51514818e-02  4.13840050e-01
5.14818026e-01  8.82658008e-04
  9.09728493e-04]
```

Sort by decrease the value of coefficients, for analyzing the feature importance:

```
ard_df=pd.
DataFrame(columns=[‘Features_name’,
‘best_ard.coef_’ ])
ard_df[‘Features_name’]=Xtrain.columns
ard_df[‘best_ard.coef_’]=ard.coef_
ard_df=ard_df.sort_values(by=’best_
ard.coef’, ascending=False)
ard_df
ard_df.to_csv(‘features importance_
ard.csv’)
ard_df
```

Thus, having only 30 points in a 30-dimensional space, it is possible to build a model that identical the real dependence.

For comparison, the coefficients calculated using ARD regression:

```
array([ 1.92557895e-03, -9.51988416e-
04,
-2.13868725e-04, -7.21210062e-04,
1.51514818e-02,  4.13840050e-01,
5.14818026e-01,  8.82658008e-04,
9.09728493e-04])
```

The ARD are very popular for using in the different kernel-methods [8], for example Relevance Vector Machine (RVM) – this is Support Vector Machine (SVM) with ARD. It is also convenient in classifiers, when it is necessary to evaluate the significance of the available features from the projects.

According to *The Historical Data Approach* it is necessary to analyse the current historical project data from the company’s analytical systems: an issue tracking systems, a team collaboration software, internal Customer Relationship Management (CRM) system, proprietary Human Recourses Management (HRM) system (for tracking HRM processes), code repository (GitHub, GitLab, etc.), the internal financial management system and others. The project processes proposed as acyclic graph-theoretical model, specifically Probabilistic Graphical Models (PGM). Such network may represent probabilistic relationships between estimates and inputs, it can be used to calculate the probabilities of evaluation of successful project realization with the different inputs that will be taken into account with their weights.

Disadvantage of the Bayesian network graph is than it does not consider into account the historical changes in the project. It may possibly correct by using the Long Short-Term Memory (LSTM) neural networks, where inputs will be historical data from the Bayesian network [6]. This will allow time to consider changing the status of implementation of our project and make more accurate predictions of its performance, it is this approach we believe can give better results.

IV. CONCLUSIONS

Our results show that proposed method could be the complexity of representing as the one of vector data model and lack of on a many-dimensional space. There for convention Machine Learning models such as: linear models, random forest, boosting are not always give reliable results but definitely more effective than using Deep Learning on the small data-sets. The initial processing and features selection are the important step in the case of many-dimensional vectors of the projects, using statistical approach based on the Bayesian probability Automatic Relevance Determination. Preliminary test shows the perspective of applying many tools and templates for project management, namely: Priority Matrix, Work Breakdown Structure, Gantt Chart, Project Budget, Risk Matrix, etc.

The implementation of our research allows more precise do project planning according the best practices, historical data and data on completed previous projects. In the next step, our research will allow to the project status threats and risks in the real time. The company’s high level management would have an access information about the economic analysis, from the project initiation phase to feedback from all internal and external stakeholders.

References

[1] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion“. IEEE Trans. Pattern Anal. Mach. Intell., 12:629– 639, 1990.

[2] C. L. Philip Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” Inf. Sci. (Ny)., vol. 275, pp. 314–347, 2014.

[3] J. Alex Stark, “Adaptive Image Contrast Enhancement Using Generalizations of Histogram Equalization“ IEEE Transactions on image processing, 2000.

[4] N. Bhargava, A. Kumawat, R. Bhargava, “Threshold and binarization for document image analysis using otsu’s Algorithm “,International Journal of Computer Trends and Technology (IJCTT) – volume 17 Number 5 Nov 2014.

[5] D. F. Rogers and J. A. Adams, “Matematical elememnts for computer graphics”, 2001.

[6] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., and Zhifeng Chen, e. a. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[7] Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instancebased learning algorithms. Machine Learning, 6(1):37– 66.

[8] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46:175–185.

[9] Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). SpringerVerlag New York, Inc., Secaucus, NJ, USA.

[10] Boehm,B.W.,Defense,T.R.W.,Group,S.,Boehm,H.W., Defense, T. R. W., and Group, S. (1987). A Spiral Model of Software Development and Enhancement. Computer (Long. Beach. Calif)., 21(May):61–72.

[11] Burch, C. (2010). Django, a web framework using python: Tutorial presentation. J. Comput. Sci. Coll., 25(5):154– 155.

[12] Freedman, D. (2005). Statistical models: theory and practice.

Inza, I., Larranaga, P., and Sierra, B. (2002). Feature Weighting for Nearest Neighbor by Estimation of Distribution Algorithms, pages 295–311. Springer US, Boston, MA.

[13] McConnell, S. (1996). Rapid Development: Taming Wild Software Schedules. Microsoft Press, Redmond, WA, USA, 1st edition.

[14] Project Management Institute (2004). A Guide To The Project Management Body Of Knowledge (PMBOK Guides). Project Management Institute.Ruder, S. (2016). An overview of gradient descent optimization algorithms. Web Page, pages 1–12.

[15] Tahir, M. A., Bouridane, A., and Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using hybrid tabu search / k-nearest neighbor classifier. Pattern Recognition Letters, 28(4):438 – 446.

[16] The Bull Survey (1998). The bull survey. London: Spikes Cavell Research Company.

[17] Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints, abs/1605.02688.

[18] Wilson, J. M. (2003). Gantt charts: A centenary appreciation. European Journal of Operational Research, 149(2):430 – 437. Sequencing and Scheduling.

[19] Chenarani, A., Druzhinin, E.A., Kritskiy, D.N.: Simulating the impact of activity uncertainties and risk combinations in R & D projects. Journal of Engineering Science and Technology Review 10(4), 1-9 (2017).

[20] Zou, K. H., Tuncali, K., and Silverman, S. G. (2003). Correlation and simple linear regression. Radiology, 227(3):617–628.

[21] MindK IT Company, March 2019. [Online]. Available: https://www.mindk.com/ [Accessed: March. 01, 2019].

[22] AlreadyOn IT company, March 2019. [Online]. Available: https://www.alreadyon.com/ [Accessed: March. 01, 2019].