# Data Verification of Telecommunication Projects for Risk Assessment Models

Ayşe Buharali Olcaysoy
Yildiz Technical University & Turkcell Technology
Istanbul, Turkey
ayse.buharali@turkcell.com.tr

Oya Kalipsiz
Yildiz Technical University
Istanbul, Turkey
kalipsiz@yildiz.edu.tr

***Abstract*—Every project is important and has some risks. However, the definition of risk needs to be clarified, since risk is usually mixed with other project concepts such as constraint and problem. Risk represents future uncertain events with a probability of occurrence and a potential for loss. It is possible to reduce or even eliminate this negative impact with risk management methods. In risk management, risk models are created to discover possible project risks in early stages of the project. Risk models can also be used to predict the success of a project in the very beginning. In order to create a predictive risk model, a large dataset is required. In this work, we created a risk dataset containing 357 projects of a telecommunication company using their records spanning two years between 2010 and 2012.**

***Keywords-Software Project Management; Software Risk Management.***

## I. INTRODUCTION

Each year Standish Group publishes CHAOS manifesto, which shows percentage of successfully completed software projects in global companies. According to CHAOS manifesto, only 39% of the projects were completed successfully (delivered on time, on budget, with required features and functions) in 2012. This ratio is still very low, even though it is increasing compared to previous years [1]. Risk management plays an important role in raising this ratio. Increasing the number of academic studies on different aspects of risk management also shows the awareness of the importance of risk management. Verna et al. [2] conducted a survey to investigate risk and risk mitigation strategies in global software development in 2013. In this survey, authors investigated 37 papers reporting 24 unique global software development projects. They also reported that the number of studies on risk management on global software development is increasing every year.

Risk assessment is an important part of risk management. Risk assessment enables a project manager to evaluate the possible risks in the early phases of project life cycle. In risk assessment, a model is constructed to discover possible risks or to evaluate the effects of risks on the progress of the project [3]. Our aim is to create a predictive risk model to discover and analyze risks of a software project in the early phases. In order to create such, a model a large dataset is required. However, to our best knowledge, there is no such dataset.

In this study, we collected a risk dataset related to internal projects of a company, which is operating in the Turkish telecommunication market. In order to create a clean dataset, several preprocessing and feature selection methods were performed. Consistency of the created risk dataset is verified through clustering and statistical distribution.

The remainder of this paper is organized as follows; Section II summarizes the existing work on risk analysis and assessment. In Sections III and IV, we give information on our proposed risk assessment model and detailed information about data preprocessing, feature selection and verification steps. We conclude this paper with Section V.

## II. RELATED WORK

Many risk management studies refer to the studies of Pretty and Briand [4]. They developed a tool, namely METRIX, for software risk analysis and management. The tool employs a modeling technique that is based on the Optimal Set Reduction algorithm [4].

Foo and Muruganantham have developed SRAM (Software Risk Assessment Model). SRAM is determined based on the results of the survey on the outcomes of past projects. The quality of the project in SRAM, time and cost of the criteria identified nine critical elements of risk relationship [5]. This value is determined only according to the risks associated with the internal dynamics.

In 2006, Jiamthubthugsin and Sutivong proposed a risk assessment model [6], which is based on an assumption that evolutionary cycles can be modeled by Weibull's family distribution. The factors used in the model are requirement volatility, staff productivity, software complexity and development time.

The study published in 2008 by Gupta and Sadik, provided a software risk assessment and forecasting model, SRAEM (Software Risk Assessment and Estimation Model) [7]. Using this model, a near-success of software project with

ISBN: 978-1-61208-449-7

accuracy can be estimated. This model not only performs a risk assessment, but also estimates the risks of software project.

Risk management also plays an important role in software architecture decisions. Vliet and Poort demonstrated how risk and cost affect making a decision about the software architect by their model RCDA (The Risk and Cost Driven Architecture) [8].

## III. RISK ASSESSMENT MODEL FOR SOFTWARE PROJECTS

Several methods, including Monte Carlo simulation [9][10][11], COCOMO (Constructive Cost Model) [12][13] and data mining techniques [14][15], have been proposed to create a risk assessment model. In our future work, we are planning to create a new risk assessment model that is able to predict possible risks of a given project in the early stages. Using this model, we also plan to predict whether a given project will succeed or fail.

In order to create this model, we plan to use a classification algorithm that is able to dig out the most similar projects from a given project pool. Naïve Bayes classifier [16] and K-Means classification algorithms [17] are possible candidates for our model. We prefer The Naïve Bayes classifier because it is among the most effective at learning algorithms known and its accuracy is higher than the other learning algorithms [16]. The vectors in the K-Means classification algorithm can be replaced during the procedure and it always sets an algorithm that converges to a local optimum. The K-Means algorithm is faster and effective for most applications as the K-Means procedure is easily programmed [17].

A large dataset that contains a company’s past projects is needed in order to create such a risk assessment model. However, such a dataset might contain irrelevant features and missing information. In this study, we collected a real life dataset from a telecommunication company and applied several preprocessing steps. Then, we validated this dataset using statistical features and clustering algorithms.

## IV. EXPERIMENT AND RESULT

In this study, we used software projects, which were developed between 2010 and 2012 by a company operating in the Turkish telecommunication sector. Unutulmaz, Cingiz, and Kalipsiz worked on the same company’s project data to examine the risk factors of the projects before the initiation phase [18][19]. This study discussed the whole risks during the project life-cycle are discussed. Risk data included the technical feasibility studies and the software projects that were developed according to the Waterfall methodology [20]. Data features of the risks involved in the project management database are shown in Table I.

TABLE I. RISK DATA

| Feature Name | Description | Type |
|---|---|---|
| Risk No | Number generated by the system | Integer |
| Risk Status | Last status of risk | Multiple Choice |
| Project | Project name to which risk was belonged | Text |
| Assigned To | Responsible name of the risk | Multiple Choice |
| Risk Level | Risk level | Multiple Choice |
| Created By | Who created the risk record in the system | Multiple Choice |
| Created On | Risk created date | Date |
| Date Identified | Risk identified date | Date |
| Description | Description of the risk | Text |
| Probability | Risk probability | Multiple Choice |
| Risk Category | Risk category | Multiple Choice |
| Last Updated | The date of the risk register was last updated | Date |
| Detailed Description | Detailed description of the risk | Text |
| Action Plan | Plan for preventing the risk | Text |
| Closure Criteria | Risk criteria for closing | Text |
| Inform To | Person shall be informed in case of realization of the risk | Multiple Choice |
| Negative Impact | The magnitude of the impact of the risk | Multiple Choice |
| Phase Identified | Phase of the project when the risk is identified | Multiple Choice |
| Response | Action taken to risk | Text |
| Risk Factors | Risk factors affecting | Multiple Choice |

### A. *Feature Selection*

Dataset acquired from the company includes 19 features. Features which had test type were removed from the dataset as free text format information could not be formalized in an assessment model.

Features related to date (“Created On”, “Date Identified”, “Last Updated”) and person name (“Assigned To”, “Created By”,”Inform To”) were also removed from the dataset since this information is not useful for assessment risks.

However, “Risk Factors” and “Risk Category” will be used in our future risk assessment model. These features are not used in this study because they can not be used for verification of the dataset (which is the target of this study). As a result, the following four data variables, shown in Table II, were chosen to be used in this study.

TABLE II. VALUES OF RISK FEATURES

| Feature Name | Values |
|---|---|
| Risk Level | Critical, High, Normal, Low |
| Probability | Very High, High, Medium, Low, Very Low |
| Negative Impact | High, Very High, Tolerable ,Very Low, Low |
| Phase Identified | Test, Deployment, Planning, Analysis, Development, Closing |

### B. Preprocessing and Statistical Distribution

1658 risk records were extracted from the company's software projects between 2010 and 2012. Records that had any value were cleaned from the dataset; so, 434 records are remaining.

The statistical distribution of risk data according to the phases of software development life cycle is given in Figure 1. The number of risk records reduction from the analysis phase is to be expected. If this reduction will begin from the planning phase, our risk assessment model will reach the goal in the future because the risks will be estimated from the first phase of the project.

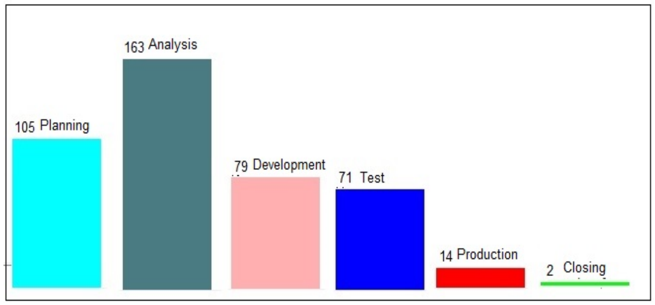

Figure 1. Risk Distribution according to the Project's Phases

The statistical distribution of the risk data according to the risk level is shown in Figure 2. This distribution will be used to compare with the results of the K-Means Clustering of the dataset in Part C.

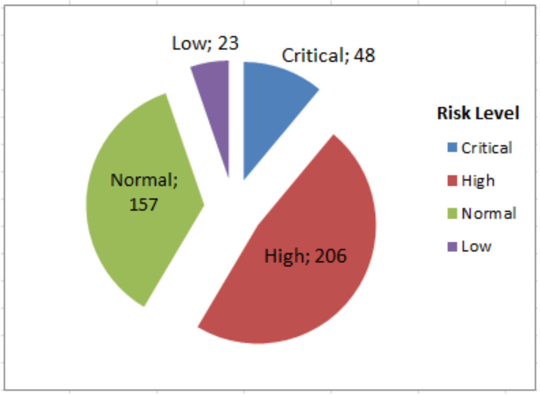

Figure 2. Statistical Distribution of Risks According to the Risk Levels

We examined the distribution of risk probability according to the risk level. This distribution didn't show any non-normal result, as shown in Figure 2. For example, the probability of high level risks is critical or high. If there will be a low risk level in this very high probability class, this result must be investigated.

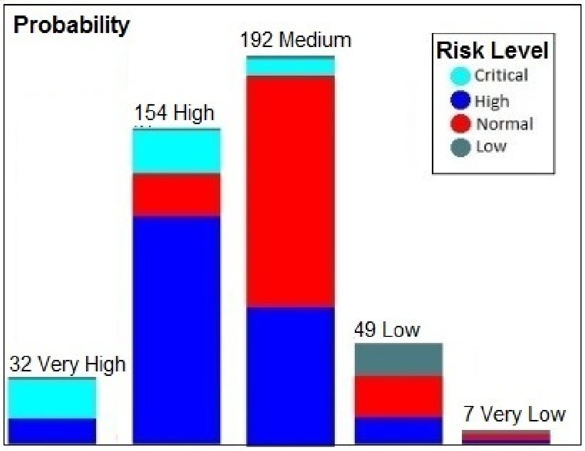

Figure 3. Statistical Distribution of Probability According to the Risk Levels

The relation between the negative impact of risk and level of risk was also examined in our study. The distribution of 58 records that had very high negative impact seems normal because the risk level of these records is critical, high or normal.

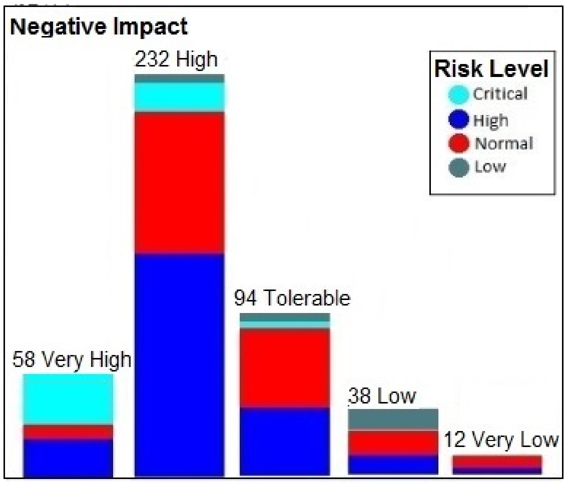

Figure 4. Statistical Distribution of Negative Impact According to the Risk Levels

### C. Preliminary Validation Analysis on the Data Set

The dataset was formed by the project managers. There was not any automatic calculation or any information received directly from the project management process. Therefore, the data were open to human error. For this purpose, the risk probability and the negative impact, according to the four levels of risk, were tried to be considered by using the K-Means Clustering method. Table III showing the result of 434 records was obtained by applying K-Means Clustering to "probability" and "negative impact" features.

TABLE III. THE POSSIBILITY OF NEGATIVE EFFECTS, ACCORDING TO THE K-MEAN DISTRIBUTION

| | **Total Data** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|---|---|
| Number of Record | 434 | 99 | 232 | 53 | 50 |
| **Features** | | | | | |
| Probability | Medium | Medium | Medium | High | Medium |
| Negative Impact | High | Very High | High | Tolerable | Tolerable |

Table IV shows the results of the risk levels clustering. The critical risks were assigned to the cluster 0. This result seems to make sense. However, the low level risks were assigned to the cluster 2 instead of the cluster 3. This result is needed to be investigated.

TABLE IV. THE LEVEL OF RISK ACCORDING TO THE DISTRIBUTION OF THE CLUSTERS

| **Original Risk Level** | **Cluster 0 (CRITICAL)** | **Clsuter 1 (HIGH)** | **Cluster 2 (LOW)** | **Cluster 3 (NORMAL)** |
|---|---|---|---|---|
| Critical | **27** | 16 | 4 | 1 |
| High | 33 | **129** | 29 | 15 |
| Normal | 26 | 82 | 16 | **33** |
| Low | 13 | 5 | **4** | 1 |

Risks, according to the only negative impact obtained clustering results, are in Table V; the results obtained by the level of risk we ran are shown in Table VI.

TABLE V. CLUSTERING RESULTS BY NEGATIVE EFFECTS

| | **Total Data** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|---|---|
| Number of Record | 434 | 70 | 232 | 94 | 38 |
| Negative Impact | High | Very High | High | Tolerable | Low |

TABLE VI. THE LEVEL OF RISK ACCORDING TO THE DISTRIBUTION OF THE CLASS – THE NEGATIVE IMPACT

| **Original Risk Level** | **Cluster 0 (CRITICAL)** | **Cluster 1 (HIGH)** | **Cluster 2 (LOW)** | **Cluster 3 (NORMAL)** |
|---|---|---|---|---|
| High | 27 | **129** | 39 | 11 |
| Normal | 15 | 82 | **46** | 14 |
| Critical | **27** | 16 | 4 | 1 |
| Low | 1 | 5 | 5 | **12** |

Table VII shows the clustering results using the method only distributed by probability. Table VIII shows the distribution of the risk level is much more accurate.

TABLE VII. CLUSTERING RESULTS BY PROBABILITY

| | **Total Data** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|---|---|
| Number of Record | 434 | 39 | 49 | 154 | 192 |
| Probability | Medium | Very High | Low | High | Medium |

TABLE VIII. THE LEVEL OF RISK ACCORDING TO THE DISTRIBUTION OF THE CLASS –PROBABILITY

| **Original Risk Level** | **Cluster 0 (CRITICAL)** | **Clsuter 1 (LOW)** | **cluster 2 (HIGH)** | **cluster 3 (NORMAL)** |
|---|---|---|---|---|
| High | 14 | 13 | **111** | 68 |
| Normal | 3 | 20 | 21 | **113** |
| Critical | **19** | 0 | 12 | 8 |
| Low | 3 | **16** | 1 | 12 |

When we obtained all these results in Table IX, we recognize that it will be more useful to use the results by the K-Means Clustering according to two features (negative impact and probability) in our future risk assessment model.

TABLE IX. THE RISK LEVEL DISTRIBUTION

| | CRITICAL | HIGH | NORMAL | LOW |
|---|---|---|---|---|
| Original Data | 48 | 206 | 157 | 23 |
| Results by K-Means Clustering with 2 Feature | 99 | 232 | 50 | 53 |
| Results by K-Means Clustering with Negative Impact Feature | 70 | 232 | 94 | 38 |
| Results by K-Means Clustering with Probability Feature | 39 | 154 | 201 | 49 |

## V. CONCLUSION AND FUTURE WORK

The statistical distribution of the risk dataset and K-Means Clustering's results proved that our risk dataset is appropriate for our risk assessment model.

In the next stage of our study, we decide to create a predictive risk model to discover and analyze risks of a software project by using fuzzy logic methods [21] and other intelligent methods [16][22]. Inspired by the study on risk assessment model in 2010 [23], we decided to use fuzzy logic. In this study, they created the software project risk assessment model was based on fuzzy theory, then the domain experts used fuzzy language to evaluate and calculate the probability and impact of risks [23]. We will use these methods to improve the relationships between risk and other project features in our model. For this purpose, we will collect other project features data such as "project category", "project size" or even "the experience of project managers". We also see that the project manager's experience is needed to be taken into account in our model.

## ACKNOWLEDGEMENT

Thanks to Turkcell Technology for allowing us to use the data in our study in accordance with their information security policy.

## REFERENCES

[1] The Standish Group. Chaos Manifesto 2013 Report. The Standish Group International, Inc., 2013.

[2] J.M. Verner, O.P. Brereton, B.A., Kitchenham, M. Turner, and M. Niazi, “Risks and Risk Mitigation in Global Software Development: A Tertiary Study”, Information and Software Technology, vol. 56, 2014, pp. 54–78.

[3] C. Ravindranath Pandian, “Applied Software Risk Management – A Guide for Software Project Managers”, Auerbach Publications, 2007, p. 128.

[4] B.E. Gayet and L.C. Briand, “METRIX: A Tool for Software-Risk Analysis and Management”, Annual Reliability and Maintainability Symposium, 1994, pp. 310–314.

[5] S.W. Foo and A. Muruganantham, “Software Risk Assessment Model”, IEEE International Conference on Management of Innovation and Technology (ICMIT), IEEE Press, vol. 2, 2000, pp. 536-544, doi: 10.1109/ICMIT.2000.916747

[6] W. Jiamthubthugsin and D. Sutivong, “Resource Decisions in Software Development Using Risk Assessment Model”, Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[7] D. Gupta and M. Sadiq, “Software Risk Assessment and Estimation Model”, International Conference on Computer Science and Information Technology, 2008.

[8] E.R. Poort and H. Vliet, “RCDA: ‘Architecting As A Risk- And Cost Management Discipline”, The Journal of Systems and Software, vol. 85, 2012, pp. 1995-2013.

[9] G.S. Fishman, “Monte Carlo Concepts, Algorithms, and Applications”, 3rd ed, Springer, 1999, p. 1.

[10] Electronic Publication: http://www.oracle.com/us/products/applications/crystalball/risk-analysis-overview-404902.pdf, [retrieved: February, 2015].

[11] Electronic Publication: http://www.palisade.com/risk/monte_carlo_simulation.asp, [retrieved: February, 2015].

[12] B.W. Boehm, Software Engineering Economics, Prentice-Hall Inc., 1981, pp. 329-342.

[13] R. Fairley, “Risk Management for Software Projects”, IEEE Software, vol. 11, May 1994, pp. 536-544.

[14] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann. 2006, p. 5.

[15] H. Jiang, C.K. Chang, J. Xia, and S. Cheng, “A History-Based Automatic Scheduling Model for Personnel Risk Management”, Computer Software and Applications Conference, COMPSAC 2007, 31st Annual International, IEEE Press, vol. 2, July 2007, pp. 361-366, doi:10.1109/COMPSAC.2007.25.

[16] T.M. Mitchell, Machine Learning, McGraww-Hill Science, 1997, pp. 177-178.

[17] J.B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press, 1967, pp. 281–297.

[18] A. Unudulmaz, O. Kalıpsız, and M.Ö. Cingiz, “Risk Faktörleri ve Risk Değerlendirme Modellerinin Farklı Veri Setleri Üzerinde Gerçeklenmesi”. UYMS, 2013.

[19] M.Ö. Cingiz, A. Unudulmaz A., and O. Kalıpsız, “Yazılım Projelerindeki Problem Etkilerinin Yazılım Mimarisi ile İlişkilendirilmesi”, UYMK, 2012.

[20] W.W. Royce, “Managing The Development of Large Software Systems”, Proceedings of IEEE WESCON 26 (August), 1970, pp. 1–9.

[21] F.M. McNeill and E. Thro, Fuzzy Logic a Practical Approach, AP Professional, 1994, pp. 13–14.

[22] W. Elmenreich, “Intelligent Methods for Embedded Systems”, Proceedings of 1st Workshop on Intelligent Solutions in Embedded Systems(WISES03), June 2003, pp. 3–5.

[23] A. Tang and R. Wang, “Software Project Risk Assessment Model Based on Fuzzy Theory”, International Conference on Computer and Communication Technologies in Agriculture Engineering, 2010, pp. 328–330