Rationalized Distributed Software Development (DSD) in Pakistan

Ghulam Samdani, Maria Latif, Jawwad Ibrahim
Department of Computer Science and IT, University of Lahore, Pakistan
ch_samdani@yahoo.com, maria.latif@cs.uol.edu.pk, jawad.ibrahim@cs.uol.edu.pk

Abstract— Distributed Software Development (DSD) is the backbone of software development. Pakistan is the best place for distributed software development (DSD) due to: Availability of low cost but the highly talented pool, Tax holiday for information technology industry, English Proficiency, Fast and reliable internet services are available. One plus point according to age structure is that the population of Pakistan between 15 and 64 years is 59.1%. On the other hand, some factors which effect DSD. In order to handle these issues, we use empirical analysis and the qualitative method (the quantitative method may also be used). A first comprehensive literature review carried out. After that, the risk factors related to DSD specifically for Pakistan's software industry, are generated empirically. In this paper some major factors identified, those badly affect DSD in Pakistan. We use Artificial Neural Network, a powerful tool for data analysis, to verify the risk factors.

Index Terms— Artificial Neural Network, Distributed Software Development, In-house Software Development, Multi-sourcing Risk Management, Software Offshore Outsourcing, Streamlined Distributed Software Development, Statistical Package for Social Sciences, Single-source Risk Management, Work Breakdown Structure

1 Introduction

IN this modern era of merging technologies, it is approximately impossible to live in isolation. Outsourcing is a trust game and "Trust is like a paper once it's crumpled it cannot be perfect", so how trust may be made on an organization which lies thousands of miles away having cultural, traditional and lingual differences. Trust is perceived as an influencing factor when it is situated at the inter-organizational level and not at an interpersonal level [1]. Due to advancement, the problems related to Distributed Software Development (DSD) are becoming more challenging. The risks occur in DSD are more different than the risks in In-house Software Development. Outsourcing is the backbone of global software development. In outsourcing you are bound to transfer, handover complete business process to the third party, so ball is directly transferred to the court of the third party (in case of nested outsourcing the ball goes to fourth or fifth parties court) Now very interesting question comes to mind "How third or fourth party takes pain up to the same intensity level of first one" the probability of catastrophic risks increases in global software development as compared to In-house Software Development. In DSD some tinny risk becomes a serious cataclysmic risk. Three types of risks are identified that exist in both DSD and ISD but were exacerbated in the offshore context; and those that were unique to the DSD context [2]. Infect risk is managerial issues which influence quality, cost, and schedule. Now a day distributed software development is becoming a need for Distributed Software Development, which eliminates Temporal Distance. In Pakistan, software houses are in few hundreds that lie in CMMI level 3-to-level 5[1]. Most of the software development centers have no proper documentation and planning, the absence of mitigation and contingency plans. There is no proper hierarchy among the departments. Mostly one department performs all major activities of the software development cycle. Prejudice, leg-pulling, and favoritism are some common practices. In spite of all these legal evils, Pakistani s/w industry is growing up rapidly. Young developers are fully energetic and have potential to lead the region, therefore, it is evolving as a powerhouse in the south Asian region, and other factor includes the obtainability of a number of English speaking proficient skilled professionals with affordable connectivity rate [3].

Section 2 describes the motivation and problem statement. The methodology is discussed in section 3, followed by results and conclusion in section 4 and 5 respectively.

2 Motivation and Problem Statement

DSD is considered as the backbone of modern software development, it is also called "SUN NEVER SETS DEVELOPMENT" [4]. DSD is a contemporary phenomenon, which helps both the business entities and countries in boosting their profit and enhances the quality of service [5].

According to the annual report of 2013 of Higher Education Commission (HEC) of Pakistan , Pakistani universities have been producing more than half a million graduates, including over 10,000 IT graduates every year since 2010. This number has increased in the past couple of years. According to the UNESCO's global education digest 2009, Pakistan is ahead of India and Malaysia in the rate of attainment of higher education among the adult population. However, according to the Consultancy UK, Pakistan is numbered 28 in the list of countries for business process outsourcing in 2016, with India and Malaysia at number 1 and 3 respectively.

The issues of business process outsourcing and especially DSD in Pakistan are not properly addressed and a very little work of substance has been done in this field. The absence of a baseline, lack of mitigation framework and risk management guideline are the factors responsible for the present situation.

In this paper, a number of risk factor are identified, encoun-

[1] Pakistan IT and ITES Industry Survey 2014

International Journal of Scientific & Engineering Research, Volume 9, Issue 10, October-2018
ISSN 2229-5518

tered by the DSD industry of Pakistan. Further, the effectiveness and goodness of these factors have been proved by using artificial neural network techniques.

3 METHODOLOGY

In order to understand our work area in which we are going to do research, the research methodology is shown in figure 1. We initially performed a literature review and interviews to discuss different offshore outsourcing community methods based on different strategies. Then based on literature review and interviews we have identified the risk factors affecting the DSD in Pakistan. Later on, the proposed factors are classified and their importance is identified by analysis of data/feedback on the factors, collected through questionnaire, using the artificial neural network (ANN) approaches.

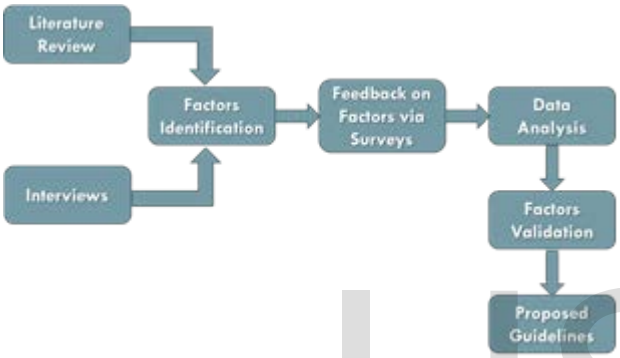

Fig. 1. Research Methodology

3.1 Data Collection Methodology

Various data collections methodologies are available. Our procedure of data collection begins as the "Research problem" is acknowledged. There are mainly two types of data, divided into two categories, namely primary data and secondary data. Primary data is collected through interview, surveys, questionnaires, observation etc., [6]. The data does not exist before and is gathered for the first time for the specific purpose of the study. Secondary data, on the contrary, is the one, which is already available in form of books, papers, magazines, internet etc. The researcher does not generate this data. He or she merely uses the data already available for study.

So as to find out the answers to the research questions and to fulfill the purpose of the research "Questionnaires" were used to collect the feedback on the identified factors. This activity was conducted from Feb 2017 to May 2017.

Questionnaire: Survey research is a technique usually used for collecting data or information about a population of concern. There are various types of surveys, several ways to administer them, and many techniques of sampling.

Through literature review and interviews, a number of factors were identified which need improvement. These factors were grouped into different categories and a questionnaire was created by consulting a few experts from the major software houses of Pakistan. Later, 250 professionals from different software houses were contacted to get feedback on the questionnaire to find out the importance of the factors. The designation and the experience of the consulted professionals are given in table 1.

TABLE 1. DATA COLLECTION RESOURCES

Sr. No.	Job Title	Experience in Years
1	CEO	10
2	Project Manager	8
3	Data Analyst	6
4	System Analyst	8
5	Subject Matter Experts	5
6	Team Lead	7
7	SQA Manager	6

3.2 Identification of Risks Specific to Pakistan

Based on a literature review and survey the major risks/issues with respect to DSD in Pakistan are shown in figure 2.

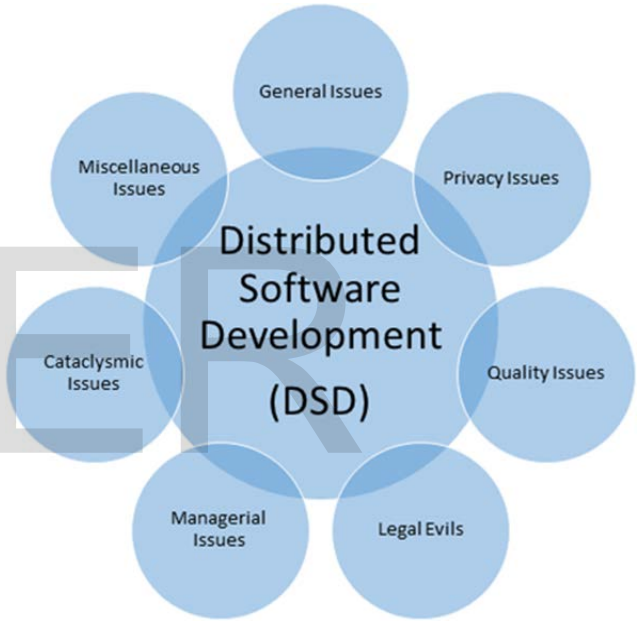

Fig. 2. Risk Categorization

Each issue/risk factor mentioned in figure 2 has further a number of latent variables. These latent variables along with the major risk factors are shown in figure 3.

4 EVALUATION AND RESULTS

In this study, seven major independent and four dependent variables from the literature review and interviews has been identified. These factors affect Distributed Software Development seriously, especially with respect to Pakistan. Our independent factors are based on 36 latent variables. The use of the Artificial neural network (ANN) help in analyzing the relative importance of different factors identified [7]. It also helps us to analyze the extent of correlation between latent variable and factors. It is different from conventional technique inasmuch as it is assumption free and analytical. Its use is mainly common in forecasting, in which quantitative and nominal data is used to foretell results; classification, wherein data is classified

International Journal of Scientific & Engineering Research, Volume 9, Issue 10, October-2018
ISSN 2229-5518

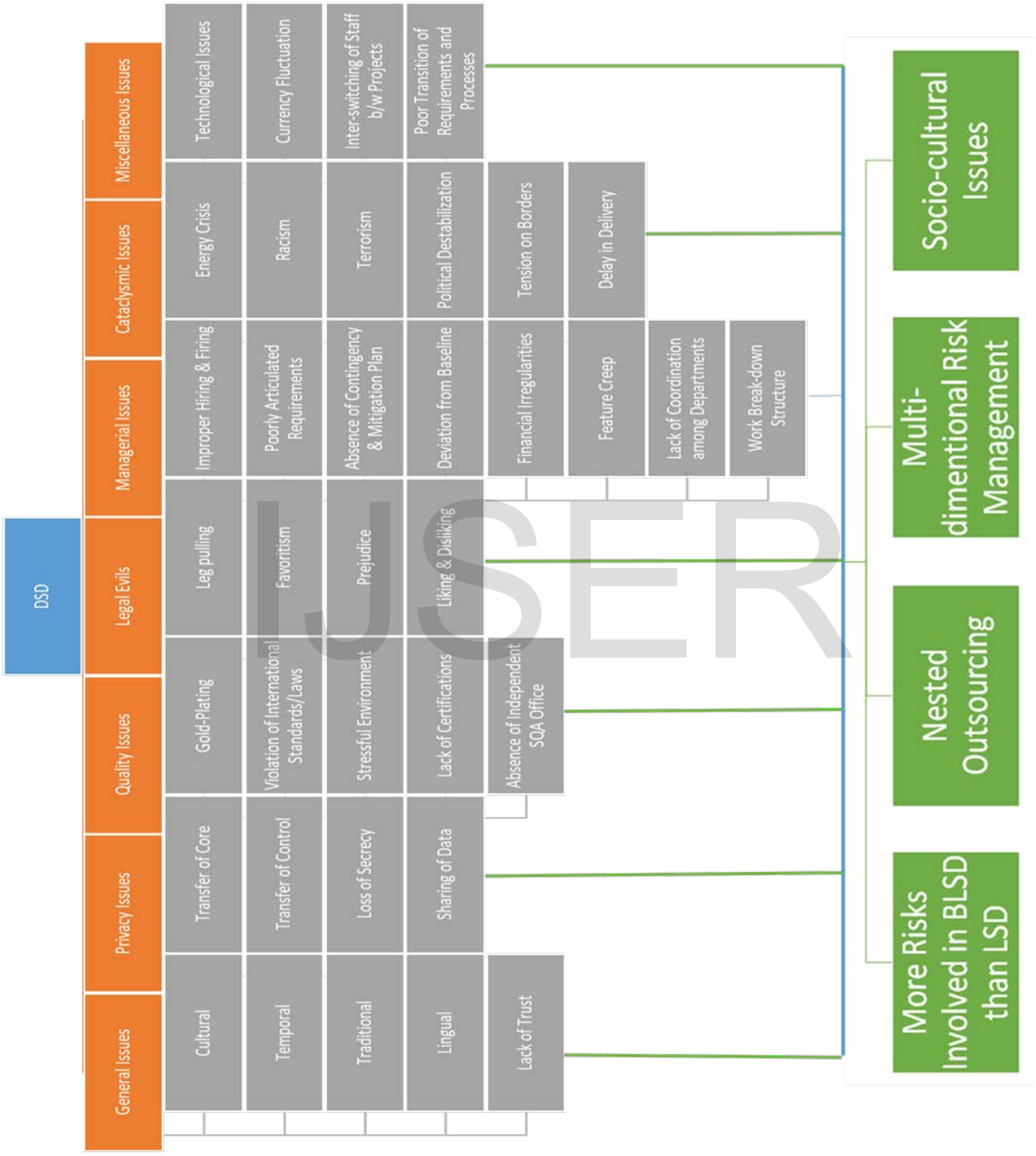

Fig. 3. Risk Factors and Variables

International Journal of Scientific & Engineering Research, Volume 9, Issue 10, October-2018
ISSN 2229-5518

into two or more than two categories, and finally to identify and reveal different statistical patterns.

In this paper, Multilayer Perceptron (MLP) – an ANN-based SPSS technique – is used to verify our proposed factors for DSD in Pakistan. The MLP yields a predictive model for one or more dependent variables based on the values of independent variables.

Table 2, represents the case processing summary for Multilayer Perception Model. Out of 250, 174 samples or individuals are chosen as Training sample and the Artificial Neural Network takes 76 samples as testing samples, and there is no excluded case in the model.

TABLE 2. CASE PROCESSING SUMMARY

		N	Percent
Sample	Training	174	69.6%
	Testing	76	30.4%
Valid		250	100.0%
Excluded		0	
Total		250	

Table 3 demonstrates the information about the Artificial Neural Network. It is used to make sure whether the specifications are accurate or not. It provides information about the Artificial Neural Network (ANN) used in our scenario. This ANN consists of Input, Hidden, and Output layers. It also shows that for data analysis with different activation functions, rescaling methods, and error function are used. It shows that the ANN for our proposed scenario contained 7 covariates (Units) in the Input layer and a Standardized method is used for the rescaling of units. It also gives us the information that the Artificial Neural Network has one hidden layer with 4 units. For the hidden layer, to link the weighted sum of value with the next layer the hyperbolic tangent activation, the function is automatically selected by the Artificial Neural Network. The output layer contains one dependent variable with 2 units and Softmax activation function is used to determine the output of a processing unit for the output layer.

Figure 4 shows the Artificial Neural Network diagram. Grey and blue lines are used to describe the relationship between input, hidden and output layers. The gray lines confirm the positive weights and the blue lines confirm negative weights. The dark blue and dark gray lines demonstrate the strong relationship between the units, whereas, the light blue and light gray lines illustrate the weak relationship between the units.

TABLE 3. NEURAL NETWORK INFORMATION

Input Layer	Covariates	1	General Issues
		2	Privacy Issues
		3	Quality Issues
		4	Managerial Issus
		5	Cataclysmic Risk/Issus
		6	Miscellaneous Issues
		7	Legal evils
	Number of Units[a]		7
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1a		4
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Dependent
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy
a. Excluding the bias unit			

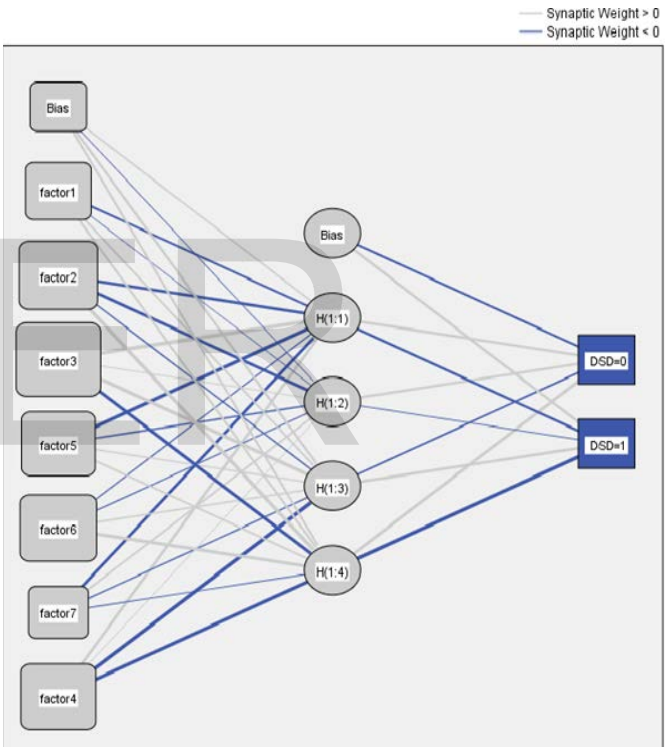

Fig. 4. Neural Network Structure Diagram

For both training and testing sample, Cross Entropy Error is provided. The cross entropy error is used so that the concluding layer i.e. the output layer uses the softmax activation function. The Network tries to minimize this error function during training. The smaller the value of this error the better will be the model for the prediction.

The cross entropy error for the testing sample (38.05), as shown in table 4, is less than the training sample (66.62). It reveals that the model performance is good for the testing sample as compared to the training sample. Therefore, we can use this model for making predictions. The model summary table also includes the information about the incorrect percentage of testing sample and training sample, with an incorrect prediction for training sample 17.8% and testing sample

International Journal of Scientific & Engineering Research, Volume 9, Issue 10, October-2018
ISSN 2229-5518

22.4%. These incorrect percentages show that our model is good for prediction.

TABLE 4. MODEL SUMMARY

Training	Cross-Entropy Error	66.619
	Percent Incorrect Predictions	17.8%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	00:00:00.101
Testing	Cross-Entropy Error	38.050
	Percent Incorrect Predictions	22.4%
Dependent Variable: dependent		
a. Error computations are based on the testing sample.		

Table 5 shows the parameter (synaptic weights, the relationship between the units of layers) estimations for different factors in the input layer and also for the units in the hidden layer.

TABLE 5. PARAMETER ESTIMATES

Predictor		Predicted					
		Hidden Layer 1				Output Layer	
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	[DSD=0]	[DSD=1]
Input Layer	**(Bias)**	.484	-.073	1.025	.807		
	factor1	-.755	-.052	.714	.791		
	factor2	-1.234	-1.332	-.359	1.903		
	factor3	2.719	.329	1.502	-1.458		
	factor5	-1.767	-.400	.386	.629		
	factor6	-.354	-.348	.427	1.448		
	factor7	-1.347	.698	-.359	-.327		
	factor4	1.251	.218	-2.289	-1.768		
Hidden Layer 1	**(Bias)**					**-.769**	**1.085**
	H(1:1)					**1.061**	**-.964**
	H(1:2)					**.828**	**-.255**
	H(1:3)					**-.486**	**1.107**
	H(1:4)					**1.394**	**-2.046**

Table 6 illustrates the realistic results of the Artificial Neural Network. For every sample, the predicted answer is Yes if that sample's predicted pseudo-probability is greater than 0.5. For every case

- The samples on the main diagonal of the cross-classification table are correct predictions.
- The off-diagonal samples of the cross-classification table are incorrect predictions.

Of the cases used to create the model, 18 out of 50 peoples for observed value "NO" are correctly classified. 111out of 124 peoples who have answered "YES" are classified correctly. Overall, 82.2% of training samples are classified correctly, with 17.8% incorrectly identified samples as shown in the model summary table. A good model must correctly classify a higher proportion of the samples.
In addition, the model correctly classified 77.6% of the testing samples. This indicates that the ANN model is capable of correctly classifying input samples.

TABLE 6. CLASSIFICATION

Sample	Observed	Predicted		
		No	Yes	Percent Correct
Training	**No**	**32**	**18**	**64.0%**
	Yes	**13**	**111**	**89.5%**
	Overall Percent	**25.9%**	**74.1%**	**82.2%**
Testing	**No**	**10**	**12**	**45.5%**
	Yes	**5**	**49**	**90.7%**
	Overall Percent	**19.7%**	**80.3%**	**77.6%**
Dependent Variable: dependent				

ROC curve, as shown in figure 4, provides us a visual representation of the sensitivity and specificity for all probable cut-offs in a single plot, which is more powerful and cleaner than a series of tables. The ROC curve presented here show two curves, each for the category No (blue curve) and Yes (green curve).
A Central line is a criterion in this figure 4. Points below the central line depict the higher percentage of incorrect classifications of the cases and above the central line show the high percentage of accurate classification of the cases.
Figure 5 shows that both the curves are towards the upper side of the central line, which concludes that the curve shows a large probability of correct results, i.e. the model is good.

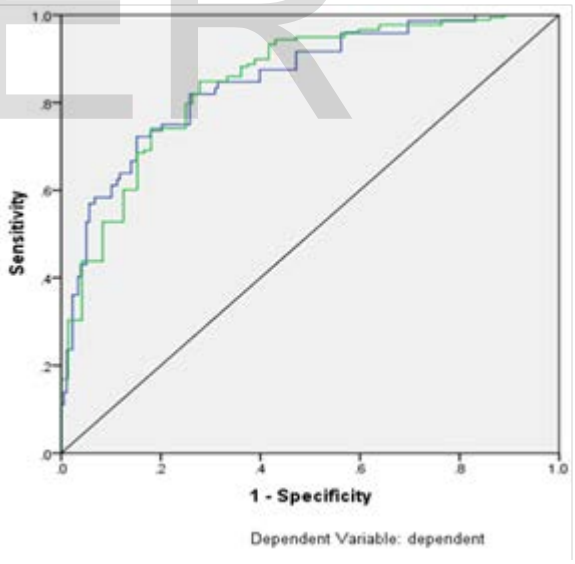

Fig. 5. ROC Curves

Table 7 is a very helpful statistic summary, which is used to identify the correctness of any multiple perception models. If the area under the curve is close to 0.5, the model is less accurate and if it is close to 1 then it is the best model. As in this model, the area under the curve is 0.850, which is much greater than 0.5 and close to 1, hence this model is good for analysis.

TABLE 7. THE AREA UNDER THE CURVE

		Area
dependent	No	.850
	Yes	.850

Table 8 represents the importance of independent variables. The importance of an independent variable is measured in term of changes in values of a predicted network model for different values of the independent variable. Normalized importance is simply the quotient of important values and the largest importance values expressed as percentages. This table shows that factor 3 is most important and factor seven is least significant.

TABLE 8. INDEPENDENT VARIABLE IMPORTANCE

	Importance	Normalized Importance
General Issues	.119	66.5%
Privacy Issues	.159	89.0%
Quality Issues	.178	100.0%
Managerial Issus	.144	80.5%
Cataclysmic Risk/Issus	.153	85.6%
Miscellaneous Issues	.100	55.9%
Legal evils	.148	82.9%

The normalized importance chart is a bar chart of the values given in the Area under the curve table. These values are arranged in descending order of significance. Figure 6 shows that variables related to Factor 3 (Quality Issues) and Factor 2 (Privacy Issues) have the 100% and 89.0% effect respectively on how the network classifies the DSD of the respondents.

Based on a literature review and survey the major risks/issues with respect to DSD in Pakistan are as under:

1. Quality Issues
2. Privacy Issues
3. Cataclysmic Issues
4. Legal Issues
5. Managerial Issues
6. General Issues
7. Miscellaneous Issues

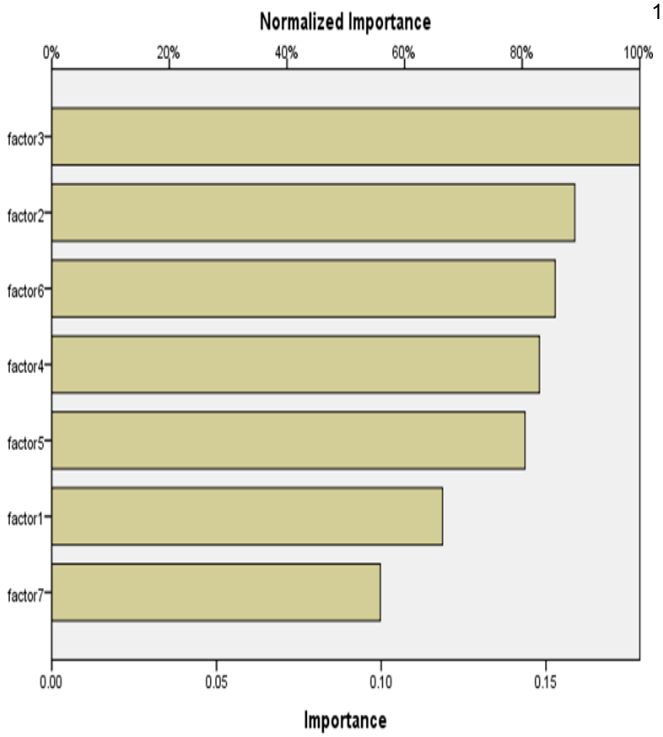

Fig. 6. Independent Variable Importance Graph

5 CONCLUSIONS

DSD is the desire of the time. Price savings is the chief catalyst driving the outsourcing decisions of these SMEs, just as it is with large firms [8]. The risk is the possibility of loss [9]. There are so many risks befell in DSD like other disciplines. In literature, there are many risks identified regarding DSD in wide-range. Some risk management frameworks are also proposed. However, there is no work done with respect to the DSD in Pakistan.

In this paper, we have identified different risk factors and grouped them into different categories as mentioned in section IV. These risk factors were finalized after a consultation with professionals from different organizations.

To verify the correctness of the identified risk factors a questionnaire-based survey was conducted involving 250 professionals from different software development organizations. The data collected through the surveys were analyzed using the artificial neural network (multilayer perceptron) based approaches to check the correctness of the risk factors.

The results obtained through the ANN-based analysis approaches have shown that the identified risk factors are the major cause of failure of DSD in Pakistan.

In the future, the guidelines will be proposed against each risk factor in order to mitigate the effect of risk factors which helps Pakistan's software organizations to improve their offshore services.

International Journal of Scientific & Engineering Research, Volume 9, Issue 10, October-2018
ISSN 2229-5518

REFERENCES

[1] Alya M., & Hajer K. (2008). "TRUST AS AN ORGANIZING PRINCIPLE IN OFFSHORING INTERCULTURAL RELATIONSHIPS. MCIS 2008 Proceedings. Paper 15.

[2] Robbie T. Nakatsu & Charalambos L. Iocavou. (2009). A comparative study of important risk factors involved in the offshore and domestic outsourcing of software development projects: A two-panel Delphi study. Information & Management, 2009.

[3] Haider S. A., Samdani G., Ali M. & Kamran M. (2016). A comparative analysis of In-house and outsourced software development in software Industry. IJCA.

[4] Khan, Q., & Ghayyur, S. (2010). Software Risks and Mitigation in Global Software Development, Journal of theoretical and applied information Technology.

[5] Ahmed R. R., Javed S. H. Business process outsourcing: A Case Study on Pakistan Outbound call centers.

[6] Introduction to Research, University of Bradford, School of Management (October 2013). Retrieved from www.brad.ac.uk/.../Introduction-to-Research-and-Research-Methods.

[7] G. Zhang, B.Eddy., P. Michael., & Y. Hu. (1998). "Forecasting with artificial neural networks:: The state of the art " International Journal of Forecasting.

[8] Coward C.T. (2003). Looking Beyond India: Factors that shape the Global Outsourcing Decision of Small and Medium Sized Companies in America. EJISDC.

[9] Chowdhury A.M. & Arefeen S. (2011). Software Risk Management: Importance and practices. IJCIT.