Cost Estimation Analysis of Web Applications: Assessing Evidence from Freelancer Community

Kunal Sanjeevkumar Setpal , Palak Sharma , Priyanka Honnakasturi , Venkat Nitin Panaganti
MS Students in Software Engineering at Rochester Institute of Technology. Rochester, NY. 14623. USA

Abstract—In today's global commercial industry, small business entrepreneurs want to scale their business scope by exploring the power of internet [13]. A common hindrance in their path is little or no expertise in the field of web development [19]. There are no previous research studies supporting these software application seeking entrepreneurs. Our study focuses on providing guidance to these customers by analyzing data from freelancer developer community (freelancer.com [7]) and also to provide the software engineering research community with certain facts related to freelance web application development. In order to accomplish this, we have identified 5 research questions. Using the developer demographics and completed project metrics, the answers to these questions provides customers/clients suggestions on how to choose a freelancer.

I. INTRODUCTION

This paper aims at providing the statistical evidence of finding influential factors behind the cost of freelance projects and their developers, which would benefit the software research community and small business organizations. Time and again we see several small business trying to make it big by tapping into the power of the virtual world and using web applications to lure customers[13]. Not just for e-commerce but also for the sake of advertising and marketing and also in some situations to prove that they, in some fashion, have an edge over their competitors[14][15]. The growing number of businesses looking to build a web application is alarming[8]. However, this brings good news to all the freelance developers who make a living by developing such standalone projects. This is because most of the small organizations would turn to freelancers to get their job done partly because they don't have the money and the requirements to approach a large firm[18,20]. A problem that arises during this process is of the budget resources to be allocated for this purpose, since the customers/clients do not have any idea of how much they should be spending, they also do not have any idea if they are spending the required amount or more than required [19]. In our study we aim to bridge this gap by providing statistical evidence of influential factors that impact the development of web applications by freelancers in different domains and also to provide suggestions for a customer/client on choosing an economical option. Additionally our study also aims to provide the research community with facts apart from the ones which help the customer, we also look into what type of project i.e. either development or maintenance projects cost more.

In order to fetch the data, that would be relevant, in determining how to choose a freelancer, and to associate the cost for the web applications in various domains, freelance websites made the best fit. Among these, we choose to obtain our data from a renowned freelancer website www.freelancer.com [7] as they had relatively more information on the projects and the developers, when compared to other leading freelancing websites such as www.Guru.com [9] or www.Project4hire.com[10] and also because it (rebranded version of RentACoder platform) was used as a case study in [17,18].

We prescribe two models - one that helps in providing essential freelancer information and his cost and the other that takes data about projects and associates it with the total cost of the project. The first model is concerned with the developer details, we obtain our data by using a scraping tool called "web scraper", an extension provided in Google chrome. This extracts the required fields such as rate per hour of the freelancer developer, number of reviews given to the freelancer developer, ratings given to the freelancer developer, skills of the freelancer developer, number of recommendations of the freelancer developer, and the country of the freelancer. This is obtained in a CSV file with thousands of records. From there we perform the statistical analysis using R programming language and try to find a correlation between the developer demographics and the cost charged per hour.

For our second model, we manually fetch the data out of all the projects that the freelancers have worked upon and obtain attributes such as project cost, project completion time, project name, domain of the application, type of project, features of the project, date of completion of the project. This is obtained for approximately 500 records and then we perform the statistical analysis on the data to find the relation between the details of a project and its total cost.

II. MOTIVATION

We observed that, in the recent times, freelance developers have increased in number and most of the small business organizations turn to them for their web application requirements (development/maintenance)[15][16]. So what is the best way to select a freelancer? Would it be based on their skill set? Or their reputation? What would be cheaper yet effective? Such questions have been unanswered [19]. Since we could not find any standardized method or factor upon which the freelancers quote their prices, we thought that it would be interesting to consider this as our research focus.

A. Online Portal

As seen in figure 1, there are many entrepreneurs and individuals who seek out answers to questions that help them choose a better freelancer for their project. Giving some direction to these questions is the focus of this research.
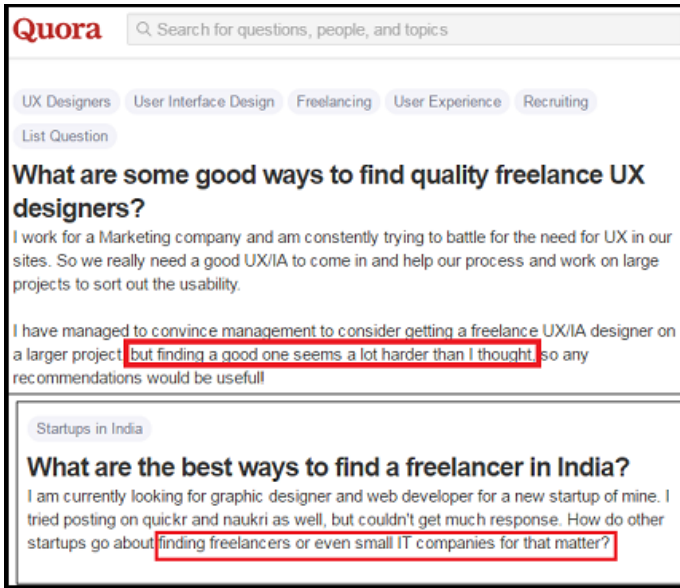
Fig. 1. Customer seeking freelancer developer related questions on Quora

Fig. 2. Flowchart showing our methodology

B. Survey

Based on the survey conducted by Google's get your business online program in 2012 [8] ; certain amazing facts were uncovered.

- In California, USA 62% of small business do not own an official website.
- 53% of small business in New York, USA do not own a website.
- 69% of small business in North Carolina, USA do not have a website.
- In most states, somewhere between 50% and 60% of small businesses do not have websites.
- 97% of consumers look online for local products and services.

Figure 1 and this survey gives an overview that many small business are still not adopting modern technology and marketing trends. Surprisingly, it indicates a sharp difference between the customer expectations and the current market scenario and also revels that these businesses would do better if they can tap into even a small portion of the 97% percent of online customers. Our goal is to provide facts which act as suggestions while finding a freelancer to do the customer/client's project.

III. METHODOLOGY

Figure 2 pictorially represents our research methodology. In brief, we begin with extraction of the required metrics by using web scrapper tool as well as manual extraction. Then the collected data is scanned for duplicate values and values which have to be analyzed manually (project description,project features), the duplicate values are deleted using MS excels internal functionality where we select a primary key to check all the records for duplication and any duplicate values found will be deleted along with the whole row. From the data we
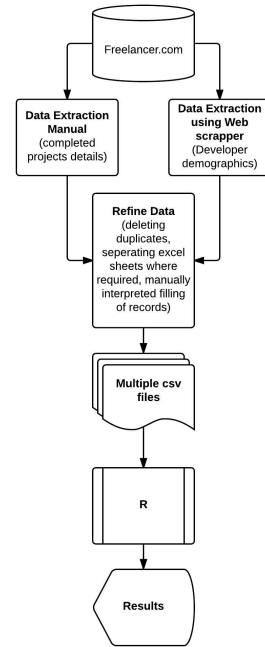
now obtained after refining, we formed multiple Excel sheets by re-arranging the columns of data(as per the requirement of the test) which are sub-sets of the two main data sets i.e. project details and developer demographics. These excel sheets were analyzed using R (R Studio) based upon the research questions under consideration to obtain the results.

A. Data Extraction

For the purpose of this cost estimation analysis, we looked into details at two levels: (a) Freelancer Developer Demographics (b) Freelancer Completed Project Metrics.

(a) Freelancer Developer Demographics : We collected around 2000 freelancer developer records in "Microsoft Excel 2013" by making use of automated web scraper tool - Google chrome plugin [12] - from popular freelancer website "freelancer.com" [7]. On merging all the records, duplicate records and records with missing information were deleted and we were left with 1359 records. We scraped these details for each developer: Name,Link,Hourly rate,Number of Reviews,Number of Recommendation and Top 5 Skills.

(b) Freelancer Completed Project Metrics : We manually collected 500 freelancer completed project records. After removing the duplicate and incomplete records, we were left with 459 records.We recorded following project specific details : Project Name, Project Description, Project Cost, Project Time, Project Completion Date, Category, skils required for the project, Project Link, Developer's Name, Developer's Cost, Developer's Country. We manually analyzed the project description and classified the projects under "Development" or "Maintenance" category. Overall there were 249 Development

projects and 207 Maintenance projects and 3 projects were classified as N/A due to data insufficiency.

B. Data Analysis

In order to analyze the cost estimation factors from two perspectives specifically the developer perspective and the project perspective and with the help of many metrics mentioned in the above data extraction section, we came up with 5 research questions that help to provide insight on the dependency of the cost on one or more of these metrics. Each of the research questions help to cumulatively account for the overall cost estimation of freelancers and the projects completed by them. Each research question was carried out using separate analysis techniques.

The first two research questions are aimed at determining if the developer demographics played a role in his/her hourly cost. The third and fifth research questions take into account the project details that the various developers have completed. The third question shed's light on the dependency of the type of work that the project calls for - development or maintenance and the cost. Question 5 infers if the cost of the project is higher, if it requires to be built using a certain technology or by a developer with a certain skill. The fourth research question helps to answer an essential question from the customer's point of view, i.e whether to approach the developer directly for a given project, or to put the project up for bidding on the freelance website. This question will help reveal if the developers bid lesser than their hourly charge as displayed on their website. The following are the research questions identified precisely :-

RQ1) Does the cost charged by a developer depend on his/her country of origin?

RQ2) Does the number of recommendations given to a developer and the skills have any effect on his/her hourly rate?

RQ3) Is there a difference in cost for projects that involve maintenance vs projects that call for development work?

RQ4) Does the developer underbid or overbid for a project relative to his per hour rate?

RQ5) What are the factors that influence the project cost?

For RQ1, since we're trying to find the effect of the developer's country on his/her hourly rate, we first found out which were the countries that had maximum number of developers. The top 5 countries were thus chosen and then we found the mean hourly cost of all the developers belonging to one country. This was obtained for all the top 5 countries i.e. countries with most developers in the context of the data collected. To visually communicate our findings we created a box plot showing all the countries hourly rate distribution along with their mean values.

For RQ2 and RQ5, we performed a linear regression analysis with cost and skills of developer and skills required by the project respectively to see if the developer skills, number of reviews, recommendations and project cost are related to each other or not, similarly with the skills required by the project. While doing this we also took into account few of

the details related to the developer i.e. specifically number of reviews, number of recommendations and few details related to the project i.e. specifically hourly rate of the developer who completed the project, total time of completion of the project respectively to see if they effect the cost. We also proceeded to see if there is any metric which influences the cost a lot by using the random forest graph technique.

For RQ3 and RQ4 we did a Wilcox test. For RQ3, we began with box plots of the development and maintenance projects against cost and found their mean.

Then we examined our null hypothesis (H_{01}) - The cost of projects involving development work and the cost of the project involving maintenance work is the same.

As for RQ4, to know if the developers underbid or overbid for the projects, we limited our study to the projects completed by the developers from India as they were maximum in number. Since the rate of the developers was in hours and the project cost was in the terms of days, in order to standardize the cost we multiplied the hourly rate of the developer by 8(normal work hours for a corporate person) and for the project's per day cost we divided the total cost of the project by the number of days taken to complete the project. We then obtained the mean value of the project's cost per day and developer per day cost.

Wilcox test was carried out to check if there's a relation in their distribution using the null hypothesis (H_{02}) - Cost of the project per day and the developer's per day cost is the same.

IV. RESULTS

A. RQ1 : Does the cost charged by a developer depend on his/her country of origin?

We plotted box plots (figure 3) for top 5 countries with maximum number of developers and observed the per hour rate of developers. It is noticeable from the mean values shown in table 1 and figure 3 that Indian developers charge the least per hour rate and developers from the United States charge the most per hour rate.

SNo.	Country Name	Number of Developers	Mean (Per Hour Developer rate in USD)
1	India	368	14.89
2	Bangladesh	113	17.82
3	Pakistan	152	19.08
4	USA	88	42.36
5	Russia	37	20.81

TABLE I
TOP 5 COUNTRIES AND THEIR MEAN DEVELOPER RATE

B. RQ2 : Does the number of recommendations given to a developer and the skills have any effect on his/her hourly rate?

We calculated the spearman's correlation for the different metrics (Website Design, Graphic Design, HTML, PHP, WordPress, Joomla, Shopping Carts, Logo Design, eCommerce,
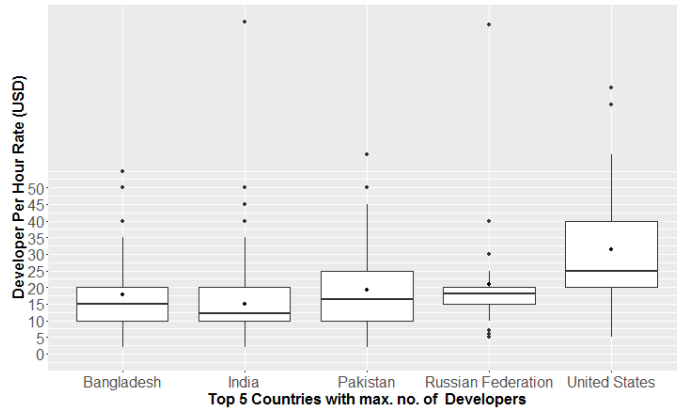
3

Fig. 3. Box Plot of Top 5 Countries' Developers and their Hourly Rate

Fig. 4. Random Forest Test Result for RQ2

numRecommendations, numReviews), but we could not find any significant r-values(r-value ¿ 0.5). Table 2 depicts the linear regression model we created for analyzing the data set for this research question. We could not find any significant results from the model. On carrying out a random forest test for the data set (Figure 4), we found out that "number of reviews" was the best predictor metric of cost followed by "numRecommendation" which is the number of recommendations the developer has received from the customers/clients .

table 3. So we cannot reject our first null hypothesis (H_{01} - The cost of projects involving development work and the cost of the project involving maintenance work is the same.)

Metric	Development	Maintenance
Mean (Per Day Cost)	63.27	52.82
Total Number of Projects	249	207
Formula	Total Project Cost/Total Project Time	Total Project Cost/Total Project Time
Minimum	5.00	3.30
1st Quartile	28.93	26.48
Median	43.83	40.00
Mean	63.27	56.40
3rd Quartile	65.75	62.36
Maximum	1313.00	792.56

Test	p-value (Wilcox Test)	Comments
Development Per Day Cost = Maintenance Per Day Cost	p-value = 0.1193	Cannot say anything about the per day cost for development and maintenance as p values are not less than 0.05 for any of the null hypothesis.
Development Per Day Cost < Maintenance Per Day Cost	p-value = 0.9404	
Development Per Day Cost > Maintenance Per Day Cost	p-value = 0.05966	

TABLE III
WILCOX TEST FOR DEVELOPMENT VS MAINTENANCE

SNo.	Skill Set	Variance (R~2) %
1.	Top 4 Skills : Website Design + Graphic Design + HTML + PHP	0.3598%
2.	E Commerce Skills : eCommerce + Shopping Carts	0.02038%
3.	Top 4 Skills + Ecommerce : Website Design + Graphic Design + HTML + PHP + eCommerce + Shopping Carts	0.3707%
4.	CMS : Joomla + Wordpress	0.2065%
5.	CMS + Ecommerce : Joomla + Wordpress + eCommerce + Shopping Carts	0.2236%
6.	All skills	0.4771%
7.	All Skills + numRecommendations	0.4836%
8.	All Skills+ numRecommendations + no.of.reviews	0.4855%

TABLE II
LINEAR REGRESSION MODEL FOR RQ2

C. RQ3 : Is there a difference in cost for projects that involve maintenance vs projects that call for development work?

As depicted by box plot in figure 5,there was no significant difference between the maintenance and development project cost. When we performed the Wilcox test on the data set, we could not get any concrete results about how the project price differs for development and maintenance tasks. The data set summary parameters and Wilcox tests results are reported in
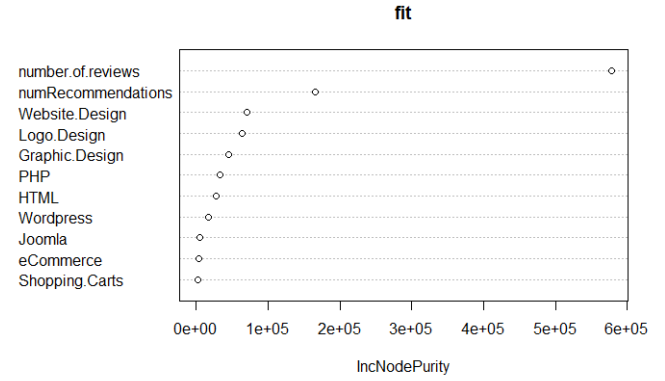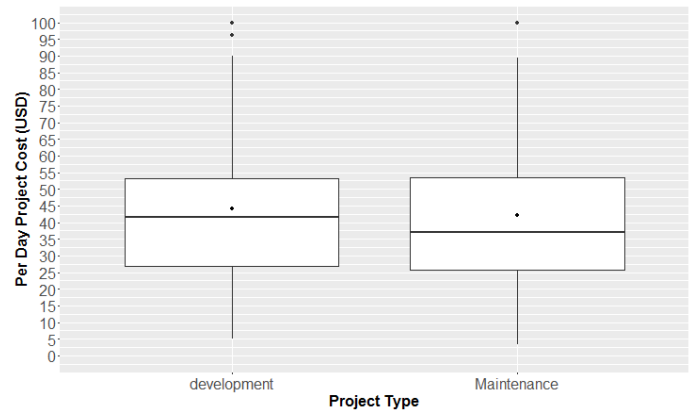

Fig. 5. Box Plot for Development vs Maintenance

D. RQ4 : Does the developer underbid or overbid for a project relative to his per hour rate?

On running Wilcox test for analyzing the trend in "per day project cost" and "per day developer cost",we found out that the developers underbid the amount they charge for a project. Table 4 shows the results of the Wilcox test and the other related metrics. So we can reject our second null hypothesis (H_{02} - Cost of the project per day and the developer's per day cost is the same.)

In accordance with our results, the histogram in Figure 6 shows that most projects have per day cost in range of (0-200)USD while there are some developers who charge between (0-100)USD and others charge between (100-200) USD.

Metric	Project Per Day Cost	Developer Per Day Cost
Country	India	India
Mean	58.26	135.83
Total Number of Records	273	146
Formula	Project Total Cost/ Project Complete Time	Developer per hour cost * 8 hours
Minimum	3.30	24.0
1st Quartile	27.00	80.0
Median	42.00	120.0
Mean	58.27	135.8
3rd Quartile	56.00	160.0
Maximum	1313.00	800.0

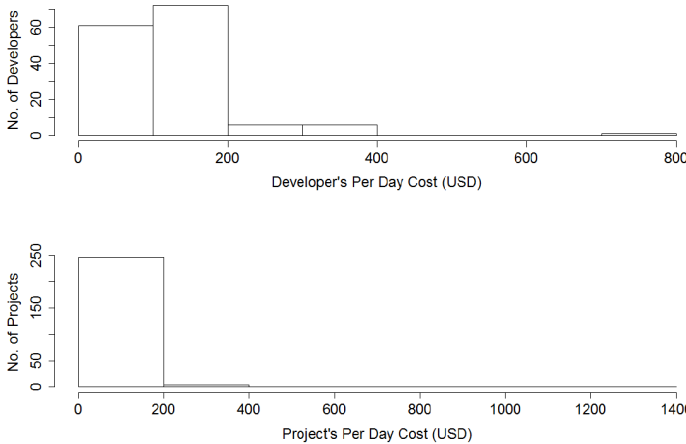| Test	p-value (Wilcox Test)	Comments
Project Per Day Cost = Developer Per Day Cost	p-value < 2.2e-16	Project Per Day Cost not equal to Developer Per Day Cost.
Project Per Day Cost < Developer Per Day Cost	p-value = 1	
Project Per Day Cost > Developer Per Day Cost	p-value < 2.2e-16	Project Per Day Cost greater than Developer Per Day Cost.

TABLE IV
WILCOX TEST RESULTS FOR RQ4

Fig. 6. Histogram for RQ4

E. RQ5 : What are the factors that influence the project cost?

Table 5 shows the results of calculating the spearman correlation for different metrics for the project cost. There were two important findings from this test: (1) "Developer's

Hourly Cost" did not have any correlation with the project cost, which is inline with our findings from research question 4, that the developers' hourly cost for projects is lesser than they quote on their profiles.

(2) The total time taken to complete the project had high correlation with the project cost. We dint find any other metric impacting the project cost.As depicted in table 6, we created a linear regression model to analyze the impact of the metrics,we found that the "Total time" metric led to a significant (39 percent) increase in the model. On running a random forest test (Figure 7), we found the same result that the "Project Time" is the best predictor metric for the project cost.

S No.	Category	Metric	Spearman correlation Value
1		Website Design	0.15
2		Graphic Design	0.1
3	Top 4 Skills	HTML	-0.03
4		PHP	0.04
5		WordPress	-0.18
6	Content Management Skills	Joomla	-0.06
7		Shopping Carts	0.13
8	eCommerce	Logo Design	-0.03
9		eCommerce	0.11
10			**0.81**
		Project Time	
11		Developers Hourly Cost	0.03

TABLE V
CORRELATION VALUES FOR RQ5

S NO.	Skill Set	Variance (R~2)
1.	Top 4 Skills : Website Design + Graphic Design + HTML + PHP	0. 58 %
2.	Ecommerce Skills : eCommerce + Shopping Carts	0.72 %
3.	Top 4 + Ecommerce : Website Design + Graphic Design + HTML + PHP + eCommerce + Shopping Carts	1.54 %
4.	CMS : Joomla + Wordpress	1.58 %
5.	CMS + Ecommerce : Joomla + Wordpress + eCommerce + Shopping Carts	2.23 %
6.	All Skills+ Developer rate	3.07%
7.	All Skills+ Developer rate + Total time	**40.06%**

TABLE VI
VARIANCE VALUES FOR RQ5

V. THREATS TO VALIDITY

External Validity: In our study, we have considered records only from one website (freelancer.com [7]). So the results analyzed and the conclusions derived are based on the data we retrieved from this site alone. The results may vary for other sites. In order to mitigate this risk, we have tried to collect more attributes both manually and also by using web scrapper tool, so that we can have many variables to consider to perform
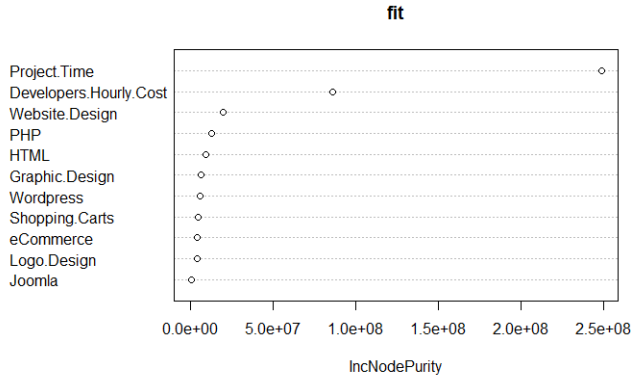
Fig. 7. Random Forest Test Result for RQ5

this study. But we have limited set of records for this study. We have analyzed 500 manual records and 2000 records scraped by the web scrapper tool. If the number of records were more, then the results might have varied. We tried to mitigate this risk by choosing a widely used website (freelancer.com [7]) which was also used in the previous studies on freelancer developers [7,17,18].

Internal Validity: For answering research question 3 and 4, we analyzed 500 records. All these records were collected manually. So, they are prone to human error. We tried to reduce human error by using only copy and paste instead of manually typing anything, except for the column which indicates the type of the project - "maintenance" or "development" which we have deduced from the name of the project and the description provided. We also removed redundant records by using Excel tools.

Construct Validity: To ensure construct validity, we performed non-parametric statistical hypothesis test called Wilcox test, which does not assume the data to be normally distributed.

Content Validity: In our study, we have taken into account some factors like developers country, developers hourly rate, project rate, etc, that might effect the cost of web application development. There may be other aspects which we have not considered in our study like project descriptions and features, developer's experience and expertise, which may impact the cost estimation analysis.

VI. RELATED WORK

Cost Estimation is a crucial factor to the success of a software project. Since 1960's, there are numerous empirical research studies based on different algorithmic models to propose the cost estimation models.

The most conventional and popular algorithmic model is COCOMO - Constructive Cost Model, which was introduced in 1981 by Barry W. Boehm[1]. This model makes use of regression formula along with details extracted from historical project records, current and future traits.

SEER-SEM is an algorithmic software application proposed by Galorath and Evans which relies on parametric algorithms,

prior knowledge and simulation based probability [2].

In the recent times, new techniques have come into picture, which make use of machine learning algorithms for proposing cost estimation models. One such study is [3], which focuses on improving cost estimation models by developing a neuro-fuzzy model - Combining a neural network model integrated with fuzzy model.

There are studies related to tailoring the cost estimation models as per the domain [4] that focuses on how to modify COBRA method(Cost Estimation, Bench marking, and Risk Assessment) for estimation of the web application for small organizations.

Research paper [5] compares three Case-based Reasoning (CBR) techniques and the prediction accuracy of the best technique is evaluated against three commonly used prediction models, namely (i) multiple linear regression, (ii) step wise regression, and (iii) regression trees.

Research paper [6] is inclined towards developing cost estimation models utilizing data set structure. One of the interesting findings are that as the project keeps on evolving, the cost estimation parameters also change. Also no significant relation was found between line of code and cost estimation.

A common characteristic of all these research studies is their focus on commercial software systems. The goal of our research work is to provide cost estimate for software application development for small businesses. Our main focus is on providing cost estimation model for small and mid-sized business entrepreneurs who want to build a website for promoting their business. Currently there is no empirical study which addresses the cost estimation of website from a small business perspective.

As most small enterprises look for freelancers to develop their software application [17], we are looking at the freelancer developer profile and their project's metric to provide a cost and time estimate required for creating a web application.

VII. CONCLUSION

From the five research questions that we proposed and gathered results for, we have four major conclusions. 1) Indian developers are more in number and charge less per hour rate in the context of the website from which we extracted our data. 2) From our data, we could not find any metrics which strongly influence the developer hourly rate, the strongest predictor for the "per hour rate of a developer" is the "number of recommendations" and the metric that most influences the "project's cost" is the "time span of the project". 3) We did not get any significant results when analyzing cost estimation difference between development and maintenance projects. 4) When comparing the "price per hour" quoted on the developers profile and the price the developers bid for the completed projects, we found out that they bid lower amount for the project than the price mentioned on their profile. This suggests that going for a bidding process is better for a customer who is looking for a freelance developer on this particular website rather than contacting a developer and getting charged according to his/her per hour rate.

Despite all our assumptions and the small data set, we believe that we have answered a few questions that would aid customers while choosing freelance developers for developing their projects. We tried to find the factors which affected the rate that the developers are charging, but we did not find any factor to be significant enough, though we report that the most significant factors for both, "per hour rate of a developer" and also "the rate of a project" on the whole. The results from the research questions can be suggestive to customers, as they give insight into finding economical options for completing their project. Justifying our two part motivation of helping the customers as well as giving some valuable information to the software development research community.

VIII. REFERENCES

[1]B.W. Boehm, P.N. Papaccio:"Understanding and Controlling Software Costs", IEEE Transactions on Software Engineering, vol. 14, no. 10, pp. 1462-1477, 1988

[2]Galorath, D D, Evans, M.W., 2006. Software Sizing, Estimation, and Risk Management. Auerbach Publications, Boston, MA, USA. ISBN: 0849335930.

[3]Du W. L, Capretz, L. F. Nassif, A. B., HoD (2013). A Hybrid Intelligent Model for Software Cost Estimation. Journal of Computer Science, 9(11), 1506.http://thescipub.com/abstract/10.3844/jcssp.2013.1506.1513

[4]M.Ruhe, R. Jeffery, I. Wieczorek : Cost Estimation of Web Applications. Software Engineering, 2003, Proceedings, 25th International Conference

[5]E. Mendes, I. Watson, C. Triggs, N. Mosley, S. Counsell: A Comparison of Development Effort Estimation Techniques for Web Hypermedia Applications. Proceedings of the IEEE Metrics Symposium, 2002

[6]Sayed Mohsin Reza, et.al: Innovative approach in web application effort and cost estimation using functional measurement type. Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference

[7]Freelancers: A Global Software Engineering Approach for Small Projects Abhinandan Bhadauria Corporate Development Center India Siemens Technology and Services Private Limited Bangalore, India abhinandan.bhadauria@siemens.com

[8]Google's get your business online program (2016, March 10). Retrieved from http://onepagerapp.com/blog/shocking-stats-on-small-business-websites/

[9]Guru. (2016, March 10). Retrieved from http://www.guru.com/

[10]Project4Hire. (2016, March 10). Retrieved from https://www.project4hire.com/

[11]Karl Rexer, Heather Allen, and Paul Gearan (2011); 2011 Data Miner Survey Summary, presented at Predictive Analytics World, Oct. 2011.

[12]Webscraper. (2016, May 03). Retrieved from http://webscraper.io/

[13]Small Business Online. (2005, March). Retrieved May 15, 2016, from https://www.entrepreneur.com/page/216024

[14]5 Reasons to Have a Web Presence Now. (April 2). Retrieved May 15, 2016, from https://www.entrepreneur.com/article/192174

[15]Why You Need a Website. (2004, October 27). Retrieved May 15, 2016, from https://www.entrepreneur.com/article/65204

[16]SCHRADER, B. (2015, August 10). Here's Why The Freelancer Economy Is On The Rise. Retrieved May 15, 2016, from http://www.fastcompany.com/3049532/the-future-of-work/heres-why-the-freelancer-economy-is-on-the-rise

[17]Yamashita, Atsushi, and Leon Moonen. "Surveying developer knowledge and interest in code smells through online freelance marketplaces." User Evaluations for Software Engineering Researchers (USER), 2013 2nd International Workshop on. IEEE, 2013.

[18]D. F. Bacon, Y. Chen, D. Parkes, and M. Rao, A market-based approach to software evolution, in ACM SIGPLAN Conf. Object-oriented programming, systems, languages, and applications (OOPSLA). ACM, 2009, p. 973.

[19] What-are-the-top-problems-faced-by-small-business-owners-in-building-a-business-website (April 15, 2013). Retrieved May 17, 2016, from https://www.quora.com/What-are-the-top-problems-faced-by-small-business-owners-in-building-a-business-website

[20] Why Hiring Freelancers Makes Sense for Small Businesses - (July 16, 2014). Retrieved May 17, 2016, from http://www.businessnewsdaily.com/6805-freelance-hiring.html