International Journal of Engineering & Technology, *7 (4.39) (2018) 1047-1052*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Yet another Approach for Construction of Cost Sensitive Classifiers for E-Learning Datasets

Mr.C.S.Sasikumar[1]*, Dr.A.Kumaravel[2]

[1]Research Scholar, Department of CSE, Bharath Institute of Higher Education and Research, India
[2]Professor and Dean, School of Computing, Bharath Institute of Higher Education and Research, India
**Corresponding author E-mail: sasi_kumin@yahoo.com*

Abstract

Cost Sensitive classifiers assumes essential job in choices utilizing forecast in exceedingly imperative research field for information mining specialists. Be that as it may, the choice of classifiers for such process assumes an essential job in more precision and less expense in the basic situations. For most extreme precision and least mistake, cost delicate and Cost dazzle are known to its execution. In the situation of Student points of interest from two district right measurements must be connected to get correct minimal effort esteems. In this paper, we will think about the cost delicate classifiers and measure their execution by fluctuating the parameters that is False Positive and False Negative. Add up to cost for various reaches are investigated independently and the execution in the two situation of Student points of interest from those locales while modifying and perusing the parameters. These discoveries can bolster the choice of finding the more profitable choose with foruming or non-forming with more certainty.

Keywords: *Networks, Wireless, RFID, Localization, Received Signal Strength, Accuracy, Optimization.*

1. Introduction

We are picking two datasets from the accompanying URL taken forapproval, https://analyse.kmi.open.ac.uk/open_dataset. The required data is being downloaded and changed as single dataset as said in the segment 5. The characteristics are sifted as Foruming and Non-Foruming. The foruming set comprise of information about the learning through Online, Presentation, Computer-based preparing. The Non-foruming comprise of information about classroom educating. Presently with this dataset we will recognize **which model will be more cost effective** Recall and precision are expressed by the ratios TP/(TP+FN) and TP/(TP+FP) respectively.

The F-measure is described as a harmonic mean of precision (P) and recall(R): i.e F =2PR/ (P + R).

Sensitivity = TP / (TP + FN)

Specificity= TN / (FP + TN)

The same proportion for a affirmative result is = sensitivity/ (1-specificity).

The same proportion for a pessimistic result is = (1-sensitivity)/specificity.

The contributions for the characterization calculations are the cost networks of fluctuating proportions of false negative to

False positive. We separate the yield from the perplexity network. There are two difficulties in preparing of cost touchy classifier. The unclassified order costs assume a pivotal job in the development of a csl demonstrate for accomplishing expected arrangement results. On the off chance that C (I, j), where I, j take esteems either 1 or 2, be the expense of anticipating a case having a place with class I when in truth it has a place with class j, at that point we are keen on C (1,2)/C (2,1) or the opposite of this. Our primary target is to discover adequate proportion as it changes incredibly crosswise over various settings.

Table 1: Algorithm components based on Ratio formats

Ratio Pattern (#FP: #FN)	Uniform	Non-Uniform (Relatively Prime)
Normal	CSTMC-U	CSTMC-NU
Reverse	CSTMC-RU	CSTMC-NRU

2. Literature Review

2.1 Data Mining Technique for E-Learning

Knowledge Discovery in Databases (KDD) or Data Mining (DM) is an incredible new innovation with extraordinary potential to enable organizations to concentrate on the most critical data in the information they have gathered by means of exchanges. In the instruction field, the expectation of understudies learning execution, recognition of improper learning practices, and advancement of understudy profile might be viewed as e-learning issues where information mining can effectively fathom them. In this paper, the creator examinations the conceivable outcomes to apply information mining procedures in e-learning setting, to foresee the understudies' status alluding to their exercises and the enthusiasm for utilizing progressed mentoring instruments. The examinations were performed based on information given by an e-learning stage (Moodle) in regards to the logging parameters of understudies selected on Interactive Tutoring Systems discipline amid the second semester of current year.

2.2 Utilizing Data Mining methods to e-Learning Problems [11]

This area intends to give in the current style delineation of the force state of research and usages of Data Mining procedures in e-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

learning. The cross-readiness of the two districts is still in its most punctual stages, and even academic references are uncommon on the ground, but few driving preparing related conveyances are presently beginning to concentrate on this new field. With the ultimate objective to offer a sensible relationship of the open bibliographic information as demonstrated by different criteria, immediately, and from the Data Mining proficient point of view, references are dealt with according to the sort of showing frameworks used, which include: Neural Networks, Genetic Algorithms, Clustering and Visualization Methods, Fuzzy Logic, Intelligent administrators, and Inductive Reasoning, among others. From a comparative point of view, the information is made by the sort out of Data Mining issue oversaw: gathering, game plan, desire, etc.

2.3 Protected E-Learning by Data Mining Techniques [12]

There are various fields which are using data and advancement approaches to manage upgrade Education and Learning. Academic Analytics is a supplier of phenomenal, custom business information data and course of action. Learning Analytics is the estimation, social event, examination and uncovering of data about understudies and their factors, for reasons of perception and updating learning and the conditions in which it occurs. Data Mining methodologies have been associated in both, Learning and Administrative issues. In Learning, the methodology is isolated into student arranged and instructor arranged. In understudy orchestrated the consideration is on supporting the understudy to take in more effectively by proposing new substance and in educator arranged the goal is to give the instructor a mechanical assembly to allow structure so it can deal with the understudy even more sufficiently.

2.4 Effectiveness of Data Mining Approaches to E-Learning System: A Survey [13]

Presently days, internet learning frameworks increment understudy's capacity to learn without anyone else. The utilization of Data Mining in training framework has turned into a noteworthy research territory, and it is utilized to gather data effectively from electronic learning frameworks. The instructive frameworks are confronting different issues, for example, static conveyance of the material; distinguishing proof of understudy needs and checking the nature of understudy collaboration level. This paper reviews instructive information mining methodologies, for example, design mining, bunching, grouping, and man-made consciousness. The objective of this paper is to find effective information from electronic learning frameworks. This work gives specific electronic courses, surely understood versatile condition, and savvy learning frameworks. The examination of electronic learning frameworks and nitty gritty investigation empower understudies to enhance the learning knowledge. This paper exhibits the recently performed research related examinations, strategies that can be utilized to enhance the understudy information and scholastic advancement in an E-Learning framework.

2.5 Data Mining in E-Learning [4]

The web has upset the manner in which data is conveyed to individuals all through the world. It didn't take yearn for material to be conveyed through the Web, utilizing electronic course books. The utilization of hypertext joins gives the student a ton of opportunity to settle on the request in which to think about the material. This prompts issues in understanding electronic reading material, which can be fathomed utilizing versatile hypermedia strategy and system. In this section, we depict how the field of instructive hypermedia profits by client displaying and adaption. We additionally demonstrate that the data fathered about the students and their learning procedure can be utilized to enhance the nature of electronic reading material.

2.6 Secure E-learning using data mining techniques and concepts [14]

Current instruction framework swinging to the advanced learning method. E-learning is a standout amongst the most broadly utilized strategy for the instruction. E-learning gives understanding as individual learning whenever and in addition anyplace, so client get more premium, adaptability at learning. In this paper, we have presented e-learning stage which has partitioned into two essential part; initial segment is learning information must be anchored, for anchoring information we have utilized document encryption and unscrambling strategy, and second part comes utilization of information digging methods and ideas for enormous information stockpiling.

3. Methods & Materials

3.1 Cost-Sensitive Learning (CSL)

Most classifiers expect that the misclassification costs (false negative and false positive cost) are the identical. In most genuine applications, this doubt may false. Another point of reference is danger determination: misclassifying a sickness is significantly more certified than the false alarm since the patients could lose their life because of a late finding and treatment.

The parameters and conditions for cost calculations are changed with the accompanying terminology. The cost administering values are arranged in the cost network which has indistinguishable structure from perplexity grid as appeared in the accompanying table.

Table 3.1: Template for Cost Matrix based on confusion matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	C_{11}	C_{12}
	Negative	C_{21}	C_{22}

At the point when misclassification costs are known, the best measurement for assessing classifier execution is add up to cost. Mean expense is the fundamental appraisal metric used in this paper and is moreover used to register every one of the three cost-sensitive learning strategies. The formula for total cost is showed up in condition.

Total Cost = (FN × CFN) + (FP × CFP) where CFN is cost of false negative and CFP is cost of false positive values denoted by C_{21} , C_{12} respectively.

Recall and precision are expressed by the ratios TP/(TP+FN) and TP/(TP+FP) respectively.

The F-measure is described as a harmonic mean of precision (P) and recall(R): i.e F =2PR/(P + R).

Sensitivity = TP / (TP + FN)

Specificity= TN / (FP + TN)

The likelihood ratio for a optimistic result is = sensitivity/ (1-specificity).

The likelihood ratio for a pessimistic result is = (1-sensitivity)/specificity.

The accompanying area portrays the calculations to create false negatives and false positives which anticipated that would be in the base dimension. The contributions for these calculations are the cost grids of shifting proportions of false negative to false positive. We separate the yield from the perplexity grid. In the event that C(i, j), where i,j take esteems either 1 or 2, be the expense of foreseeing an example having a place with class I when in actuality it has a place with class j, at that point we are occupied with C(1,2)/C(2,1) or the reverse of this. Our primary target is to discover satisfactory proportion as it changes enormously crosswise over various settings.

3.2 Meta Classifier

3.2.1 Cost Sensitive Classifier

Name

CostSensitiveClassifier

Synopsis

A met classifier that makes its base classifier cost-fragile. Two systems can be used to introduce cost-affectability: reweighting getting ready events as demonstrated by the total expense doled out to each class; or envisioning the class with minimum expected misclassification cost (rather than the more then likely class). Execution can as often as possible be improved by using a Bagged classifier to upgrade the probability examinations of the base classifier.

3.3 Base Classifiers

3.3.1 J48 Decision Tree

Description

Class for producing a pruned or unpruned C4.5 decision tree.

3.3.2HoeffdingTree

Description

A Hoeffding tree (VFDT) is a steady, whenever choice tree acceptance calculation that is equipped for gaining from gigantic information streams, expecting that the dissemination creating precedents does not change after some time. Hoeffding trees abuse the manner in which that a little precedent can consistently be adequate to pick a perfect part trademark. This thought is upheld scientifically by the Hoeffding bound, which measures the quantity of perceptions (for our situation, models) expected to assess a few insights inside a recommended exactness (for our situation, the integrity of a property).

A speculatively captivating segment of Hoeffding Trees not shared by other enduring decision tree understudies is that it has sound confirmations of execution. Using the Hoeffding bound one can show that its yield is asymptotically about vague to that of a non-slow understudy using unendingly various points of reference.

3.3.3 LMT (Logical Model Tree)

Description

Classifier for building 'determined model trees', which are gathering trees with vital backslide limits at the takes off. The figuring can oversee twofold and multi-class target factors, numeric and apparent attributes and missing characteristics.

3.3.4 REPTree

Description

Quick decision tree learner. Fabricates a choice/relapse tree utilizing data gain/change and prunes it utilizing diminished blunder pruning (with backfitting). Just sorts esteems for numeric qualities once. Missing qualities are managed by part the relating occasions into pieces.

4. Dataset Description

These Datasets are about understudy points of interest from the area of East London and from the locale of Yorkshire and the accompanying properties and its portrayal which are in the tables previously pre-handling.

4.1 Methodology Proposed

Cost delicate learning is tied in with contrasting the False Negative (FN) values and False Positive (FP) values. The datasets which I'm utilizing here is constant information. This dataset comprises of 15 properties and it is separated to 9 qualities (counting 1 class) after pre-handling process in particular id_student, date_registred, sex, date, sum_click, activity type, date summit, center, weight. The Class is an ostensible class quality which is having two kind of qualities forming (forumng) and n-foruming (nforumng).

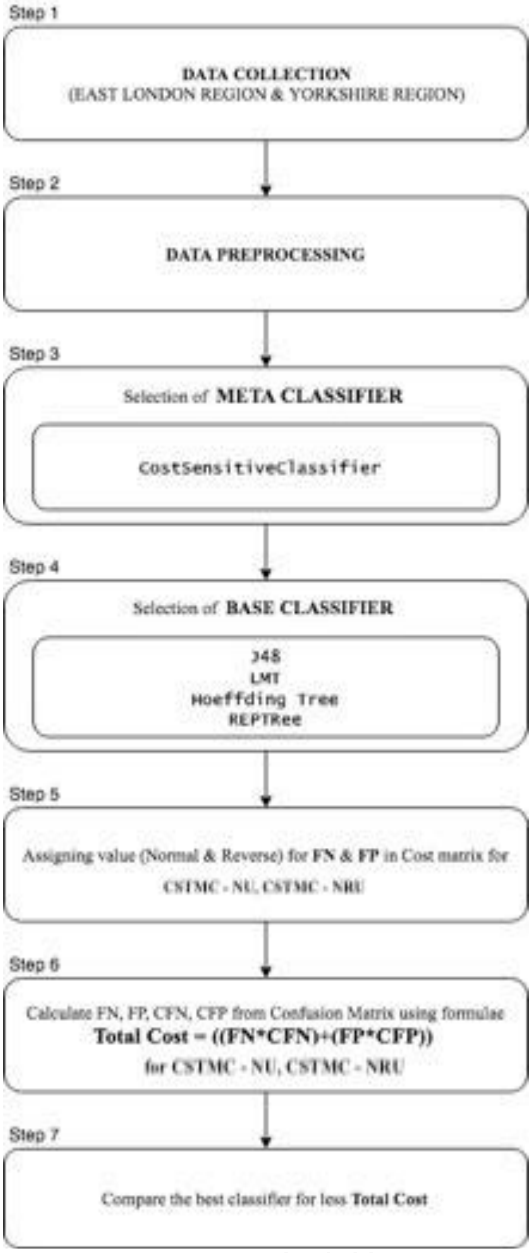
Step 1
DATA COLLECTION
(EAST LONDON REGION & YORKSHIRE REGION)
Step 2
DATA PREPROCESSING
Step 3
Selection of META CLASSIFIER
CostSensitiveClassifier
Step 4
Selection of BASE CLASSIFIER
J48
LMT
Hoeffding Tree
REPTRee
Step 5
Assigning value (Normal & Reverse) for FN & FP in Cost matrix for
CSTMC - NU, CSTMC - NRU
Step 6
Calculate FN, FP, CFN, CFP from Confusion Matrix using formulae
Total Cost = ((FN*CFN)+(FP*CFP))
for CSTMC - NU, CSTMC - NRU
Step 7
Compare the best classifier for less Total Cost

Figure 1: Proposed methodology

In Weka Application, information pre-processes are introduced, that is transferring the whole dataset into the application. The steps are

- Choose the dataset **East London Region** with17 Attributes and 83732 Instances& **London Region** with17 Attributes and 69829 Instances.
- In both these datasets delete the attribute ‘id_site’ before starting pre-process.
- Now, choose ‘Resample’ under Filter (unsupervised).

- In Resample set the ‘SampleSizePercentage’ as 10.0 then Apply.
- Now the attributes will 16 and the instances will be reduced to 8373.
- Under select trait tab and pick 'CFsSubsetEval' for characteristic Evaluation and 'BestFirst' for inquiry technique and pick Activity_Type quality before beginning pre-process. Activity_Type quality is chosen as a class characteristic since this trait having bi-esteemed information which resembles Boolean.

As an outcome, we get an arrangement of ascribes to be considered for pre-process and rest of the qualities will be expelled from the table and with the chose properties activity_type characteristic ought to likewise be chosen.

- For East_London Region dataset we got the accompanying traits: For East_London Region dataset we got the following attributes:

-id_student - date_registration
- Gender - date
- sum_click - activity_type
- Score - weight

After the pre-forms, we have to choose the meta classifier. Information grouping is a procedure that helps for proficient expectation utilizing the informational index. In information order, there are numerous meta classifiers are available and we utilize especially a classifier and that is CostSensitiveClassifier. Inside these meta classifiers, there are various base classifiers to be specific J48, Hoeffding Tree, LMT, REPTree. In the wake of choosing each base classifier, we have to pick the cost lattice 2:2 network (1.0,1.0,1.0,1.0). In that Matrix, we need to fix and change the False Positive (FP) values and False Negative (FN) values. Here we have to discover the Cost Sensitive incentive for CSTMC - NU, CSTMC - NRU utilizing the formulae f(x) = ((FP*CFP) +(FN*CFN)). In the wake of finding the qualities think about the qualities and locate the less cost esteem. The various steps in proposed methodology is shown in figure 1.

5. Experiment Results

We consider the execution of base tree classifiers for cost touchy meta students in Weak stage. The best such classifiers are J48, LMT, Hoeffding Tree, REPTree. For this reason, we receive the information from clinical records for East London Region and Yorkshire Region understudies detail. Table3 shows the total cost.

Table 3: Total cost for Variation of false positive and false negative in Cost sensitive Meta Classifiers for East London Region dataset using CSTMC-NU and CSTMC-NRU

Total Cost for East London Region NORMAL					Total Cost for East London Region REVERSE				
(Cost Ratio)	J48	LMT	Hoeff	REPTree	(Cost Ratio)	J48	LMT	Hoeff	REPTree
2:2	5222	5626	6138	5192	2:2	5626	5626	6138	5192
2:3	6648	7003	8545	6748	3:2	6468	6468	6069	5926
3:3	7833	8439	9207	7788	3:3	8439	8439	9207	7788
3:4	9199	9893	11863	9580	4:3	9307	9307	9026	8623
4:4	10444	11252	12276	10384	4:4	11252	11252	12276	10384
4:5	11787	12897	14921	12156	5:4	12396	12396	12542	11292
5:5	13055	14065	15345	12980	5:5	14065	14065	15345	12980
5:6	14528	15577	18185	14878	6:5	15029	15029	15563	14013
6:6	15666	16878	18414	15576	6:6	16878	16878	18414	15576
6:7	17105	18514	21033	17544	7:6	17922	17922	19224	16608

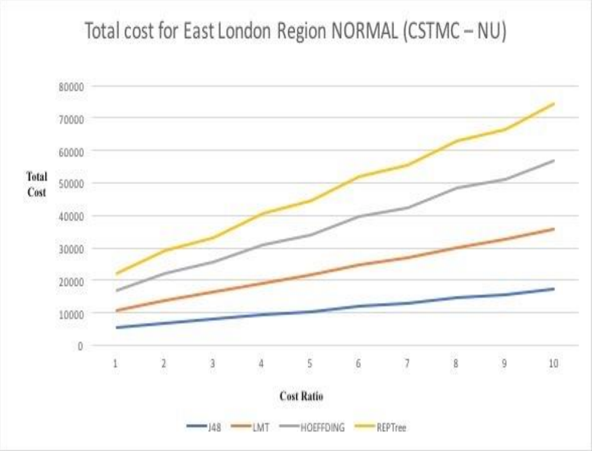

Figure 2: Total cost for East London region

Dataset is being arranged utilizing meta classifier CostSensitiveClassifier. Under meta classifier we have chosen four base classifiers specifically J48, Logical Model Tree (LMT), Hoeffding Tree, REPTree. Subsequent to choosing the base classifiers, we have to pick the cost network esteems False Negative and False Positive to discover which is delivering the less cost esteem. The esteem which is settled in the cost network must be in two kind, they are ordinary frame and turn around form. After characterization, perplexity framework esteem (FN and FP) needs to determined with CFP and CFN utilizing the formulae Total Cost = (FN × CFN) + (FP × CFP) where CFN is rate of false pessimistic and CFP is rate of false optimistic qualities signified by C21, C12respectively. Fig 2 and 3 demonstrates the aggregate expense for East London locale and its turn around. Dataset is being characterized utilizing meta classifier CostSensitiveClassifier. Under meta classifier we have chosen four base classifiers After choosing the base classifiers, we have to pick the cost lattice esteems False Negative and False Positive to discover which is delivering the less cost esteem. The esteem which is settled in the cost grid must be in two sort, they are ordinary shape and invert frame. After characterization, disarray lattice esteem (FN and FP) needs to determined with CFP and CFN utilizing the formulae Total Cost = (FN × CFN) + (FP × CFP) where CFN is rate of false pessimistic andCFP is rate offalseaffirmative qualities signified by C21, C12 individually. With the aggregate cost got from the formulae the esteem must be thought about and the less aggregate cost will be picked. Table 4 shows the total cost for Yorkshire region and fig. 4 and 5 shows the graphical representation of total cost for Yorkshire region and its reverse.With the aggregate cost got from the formulae the esteem must be looked at and the less aggregate cost will be picked. Table 5 demonstrates the variety for the two locales and fig 6 and 7 demonstrates the graphical portrayal of the equivalent and its switch. Add up to cost is expanding more for false negative than that of false positive in the both East London locale and Yorkshire area understudy points of interest trial setups. Besides, the size of aggregate expense is more than that in the East London Region than the Yorkshire Region.

We present the outcomes for the four portions of the primary calculation as appeared in the above charts. These realities are displayed as patterns and conduct in each fragment of ρ values in addition the needs of impacts of these sections likewise introduced for probability proportion.

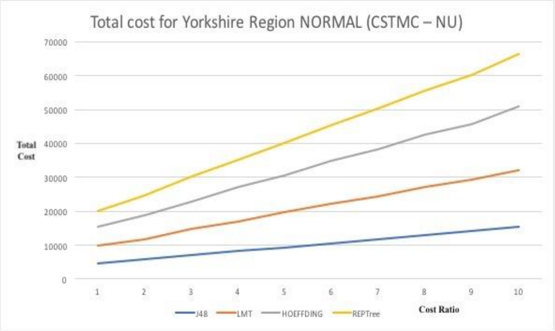

Figure 4: Total cost for Yorkshire region

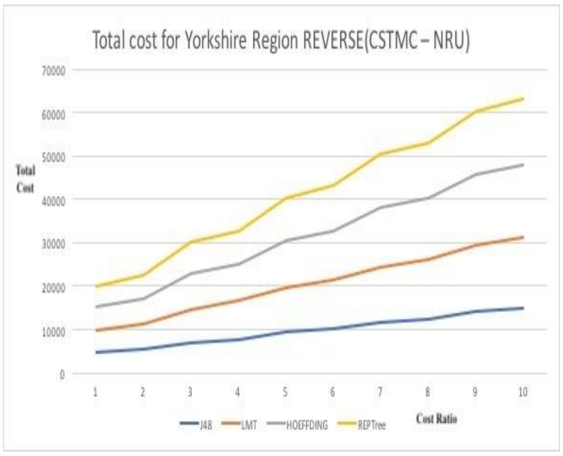

Figure 5: Total cost for Yorkshire region (reverse)

Table 4: cost for Variation of false positive and false negative in Cost sensitive Meta Classifiers for Yorkshire Region dataset using CSTMC-NU and CSTMC-NRU

Total Cost for Yorkshire NORMAL					Total Cost for Yorkshire REVERSE				
(Cost Ratio)	J48	LMT	Hoeff	REPTree	(Cost Ratio)	J48	LMT	Hoeff	REPTree
2:2	4690	5112	5482	4826	2:2	4690	5112	5482	4826
2:3	5723	6109	6958	5814	3:2	5324	6071	5562	5428
3:3	7035	7668	8223	7239	3:3	7035	7668	8223	7239
3:4	8155	8794	10097	8292	4:3	7821	8796	8372	7846
4:4	9380	10224	10964	9652	4:4	9380	10224	10964	9652
4:5	10586	11481	12772	10660	5:4	10191	11292	11315	10225
5:5	11725	12780	13705	12065	5:5	11725	12780	13705	12065
5:6	12987	14080	15469	13092	6:5	12511	13726	14104	12603
6:6	14070	15336	16446	14478	6:6	14070	15336	16446	14478
6:7	15367	16781	18688	15664	7:6	15009	16259	16762	14957

We have thought about all the table and found certain qualities in CSTMC-U, CSTMC-RU. In the wake of breaking down the table qualities J48 and REPTree are the bases classifiers creating minimum cost touchy qualities. Among these two base classifiers J48 tree is the minimum touchy esteem creating classifier.

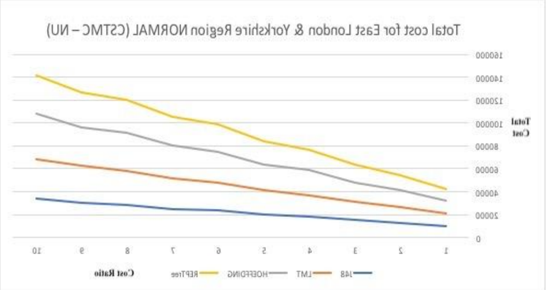

Figure 6: Total cost for both regions

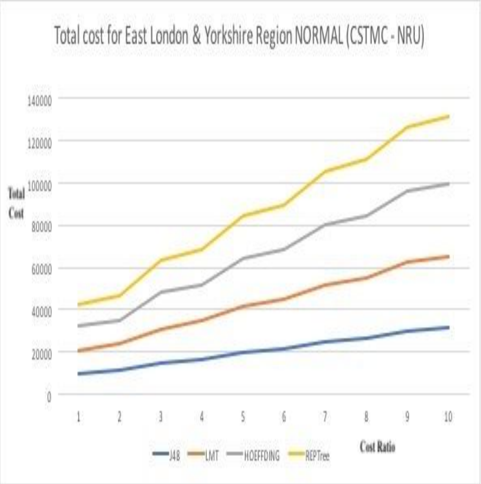

Figure 7: Total cost for both regions (reverse)

Table 5: Total cost for Variation of false positive and false negative in Cost sensitive Meta Classifiers for East London & Yorkshire Region dataset using CSTMC-NU and CSTMC-NRU

Total Cost for East London & Yorkshire Region NORMAL					Total Cost for East London & Yorkshire Region REVERSE				
(Cost Ratio)	J48	LMT	Hoeff	REPTree	(Cost Ratio)	J48	LMT	Hoeff	REPTree
2:2	9986	10782	11294	10164	2:2	9986	10782	11294	10164
2:3	12928	13244	15366	12965	3:2	11268	12519	11474	11615
3:3	14979	16173	16941	15246	3:3	14979	16173	16941	15246
3:4	18267	18798	21769	18184	4:3	16519	18012	17191	16668
4:4	19972	21564	22588	20328	4:4	19972	21564	22588	20328
4:5	23562	24249	27451	23382	5:4	21741	23314	23095	21730
5:5	24965	26955	28235	25410	5:5	24965	26955	28235	25410
5:6	28715	29571	33165	28513	6:5	26638	28702	29137	26817
6:6	29958	32346	33882	30492	6:6	29958	32346	33882	30492
6:7	33661	34787	39789	33607	7:6	31458	34042	34425	31971

5.2 Final Experiment Comparing Uniform & Non-Uniform Cost Sensitive Analysis of East London Region and Yorkshire Region.

In the wake of looking at and assessing the consequence of all the meta classifiers J48, LMT, Hoeffding Tree, REP Tree under the Cost Sensitive Classifier Base Classifier from Normal and Reverse of Uniform and Non-Uniform utilizing the dataset East London Region and Yorkshire, in typical cycle of Uniform and Non-Uniform from both the datasets, most astounding quality got from Hoeffding Tree and the least esteem is gotten from J48. In turn around emphasis of Uniform and Non-Uniform from both the dataset the most elevated esteem is gotten from the Hoeffding Tree and the least esteem is gotten from the J48. Along these lines, with this outcome as end the J48 is the main Classifier delivers least expense among every single other classifier.

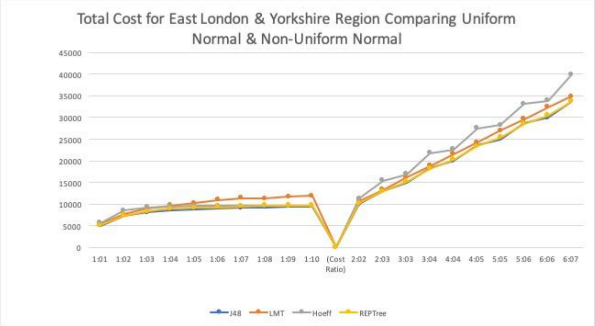

Figure 8: Total Cost for both regions (Uniform & Non-Uniform

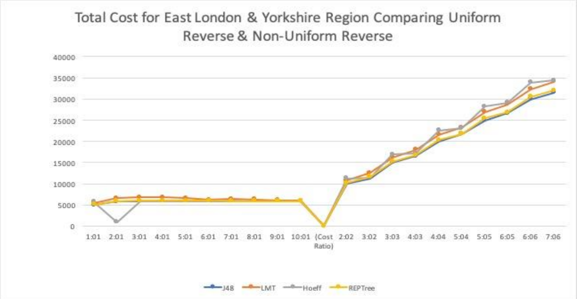

Figure 9: Total Cost for both regions (Uniform & Non-Uniform

6. Conclusion

The cost touchy models for East London locale and Yorkshire district understudy points of interest informational indexes are built and demonstrated the conduct for four scopes of cost proportion ρ. In this cost delicate process unmistakably demonstrates the requirement for isolated principles in basic leadership distinctively relying upon the cost proportion in various setting like the datasets we utilized. With these two datasets, East London locale and Yorkshire district we have discovered two different ways of training framework foruming and non-foruming and we characterized which method for framework is reacting better positive reaction. More over the greatness of aggregate expense is more in Yorkshire district than East London area. The future work can be reached out with the investigation for different sorts of cost touchy meta classifiers to gauge the mistake cost as talked about in this paper.

References

[1] Baker, R. S. J. d. 2011. "Data Mining for Education." In International Encyclopedia of Education, 3rded., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.

[2] Baker, R. S. J. D., and K. Yacef. 2009. "The State of Educational Data Mining in 2009: A Review and Future Visions." Journal of Educational Data Mining 1 (1): 3–17.

[3] Hamilton, L., R. Halverson, S. Jackson, E. Mandinach, J. Supovitz, and J. Wayman. 2009. UsingStudent Achievement Data to Support Instructional Decision Making (NCEE 2009-4067).Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

[4] C. Romero[1], S. Ventura [2]Data Mining in E-Learning, ISBN:1845641523, ISSN:17420172

[5] ZacharoulaPapamitsiou, & Anastasios A. Economides. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4), 49-64. Retrieved from http://www.jstor.org/stable/jeductechsoci.17.4.49

[6] Cios, K.J., Pedrycz W., Swiniarski, R.W. & Kurgan, L.A. (2007), Data Mining: A Knowledge Discovery Approach, Springer, New York.

[7] Klosgen, W. &Zytkow, J. (2002), Handbook of data mining and knowledge discovery, Oxford University Press, New York.