Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Comparative Study of Applying Different Classification Techniques to Enhance Customer's Segment Prediction

Raghda M. Nasr[*], Mohammed B. Senousy

Computers and Information Systems Department,
Sadat Academy for Management Sciences, Egypt

Abstract— now all the organizations appreciate the great value of the CRM and understand that customers are the most valuable assets. All of current researches focus on enhancing the aCRM "analytical CRM" category using data mining techniques. Most of the researches revolve around target customers according to their purchasing history or avoid churn, that's all good until dealing with real world. In the real world, it is not easy to have critical data about your current customers so what is about the target customers, collecting customers purchasing history is a very hard task if it is not impossible. Another point of view there is a common problem with all the real data which is imbalanced. This research based on predicting customers segments from their simple demographic data that can be easily collected. This research will present the results of a comparative study between two classification algorithms and their stacking through imbalanced data set then will apply the same comparative study after fixing the data using SMOTE.

Keywords— aCRM, RFM, Clustering, Classification, SMOTE, Metastacking

I. INTRODUCTION

This research revolves around the ability to use the classification techniques in enhancement the CRM "Customer Relationship Management" tasks. CRM has four CRM dimensions (identification, attraction, retention and development) each of them has several tasks such as: target customer, customer segmentation, direct marketing, and loyalty programs [1]. Most of these tasks need customer segments to be predefined. The first step in this research, data mining clustering techniques will use RFM analysis model to segment customer according to their purchasing behaviour. The Second phase will be the prediction of customer segments based on their demographic data using some different classification techniques to identify which is the suitable algorithm. This research aims to identify the common demographic characteristics in each segment, so it will be easy to target customers similar to those of specific segment. This research will be applied over a dataset contain the transactions of retail store during two years, also some of the customers demographic data will be used.

This paper is organized as follow: Section two reviews related researches, in section three clustering using RFM is discussed. Section four demonstrates problem of imbalanced data in predicting customer segment. Section five demonstrates the evaluation results of predicting customer data, and finally conclusion in section six.

II. RELATED RESEARCHES

Applying K-mean clustering algorithm on RFM analysis in order to segment customers based on their purchasing behaviours [2]. Another research used clustering techniques first to segments customers within the banking sector medical insurance then develop a model to recognize them [3].Another study used two classification models Naïve Bayes and Neural Networks to propose a framework for customer behaviour prediction [4].

III. CLUSTERING USING RFM

RFM model was proposed by Hughes 1994 [5]. RFM stands for Recency, Frequency, and Monetary. Recency measures the purchase last time; frequency measured by the number of purchases actions during the same period [1] and monetary express the total amount for units and items that were purchased during a specific period [6].

After the calculation step for all of the Recency, Frequency and Monetary for the customers during the same specific period, a score from 1 to 5 will be assigned to each of the R, F and M. Data should be split into 5 groups, 5 assigned to the top and so on from the most to the least. By the end there may be 125 score probabilities from (555) to (111). RFM considered as an indicator for the past behaviour that can be used to express the customer value, so the RFM scores will be used to divide the customers into convergent clusters that based on their similar behaviour and values.

Using K-mean algorithm, the number of clusters is predetermined to be eight clusters. There will be eight possibilities if each value of R, F, and M expressed by probabilities ↑ (above overall average) or ↓(less than overall average) [2]. Each of the 3 scores will compared individually to its overall average for the data, if it is more the average then ↑ will be assigned or ↓ if less than the average.

If the study data set gives different 8 probabilities it will be shown in Table 1:

TABLE 1 CUSTOMER'S TYPES AND RFM

Customer Type	RFM
Best	R↑ F↑ M↑
Shopper	R↑ F↑ M↓
Valuable	R↑ F↓ M↑
New Customer	R↑ F↓ M↓
Churn	R↓ F↑ M↑
Observer	R↓ F↑ M↓
Bargain Hunters	R↓ F↓ M↑
Zero Gain	R↓ F↓ M↓

Table 2 shows the resulted clusters for the study dataset. The sixth column RFM pattern assign each value a sign to illustrate its comparison with the overall average.

TABLE 2 RFM PATTERNS MAPPED TO CUSTOMER TYPES, 8 CLUSTERS

Cluster	R	F	M	RFM Pattern	Customer Type
C6	5	2.4848	3.4242	R↑ F↑ M↑	Loyal client
C7	5	2.2083	2	R↑ F↑ M↑	Profitable client
C3	5	1	3.0417	R↑ F↓ M↑	Valuable client
C4	5	1	2	R↑ F↓ M↑	Seasonal Client
C5	5	2	1	R↑ F↑ M↓	Casual Shopper
C1	5	1	1	R↑ F↓ M↓	New Customer
C2	4	1.0074	1	R↓ F↓ M↓	Uncertain
C0	2.281	1	1	R↓ F↓ M↓	Zero Gain
Total					
Overall Average	4.7793	1.0545	1.1914		

According to Table 2 the dataset customers divided into four patterns as shown may be 2 clusters have the same RFM pattern as in loyal client and profitable client classes both of them have above the average RFM (R↑ F↑ M↑). Also there may be some missing patterns such as such as (R↓ F↑ M↑: churn), (R↓ F↑ M↓: observer) and (Bargain Hunters: R↓ F↓ M↑). To the same pattern classes will be compared according to their scores to identify them such as in the valuable and seasonal client classes. Table 3 explains each cluster and presents all the RFM patterns in each cluster

TABLE 3 CLUSTERS PATTERNS

Clusters	RFM Patterns	Cluster Analysis
Loyal client	555 544 543 553 525 533 524 523	Customers with highest recency, frequency and monetary values, that kind of patterns ensure the achievement of high profit to the organization. In the K=5 clustering it include some patterns that their customers may be not loyal enough to deserve the same benefits of this class.
Profitable Client	522 532 552 542	Customer with high RFM vales, this cluster may need some researches to gain the customers in the loyal client cluster.
New Customer	511	This cluster show high recency with low frequency and monetary that means that these customers may be new who need something attractive to ensure that they will return.
Uncertain	411 421	Cluster with low RFM rates that the organization did not achieve too much profit from its customers.
Zero Gain	311 211 111	Cluster with very low RFM rates.
Valuable client	513 514	Customer with high recency, high monetary and low frequency. That kind of customer should be encouraged to visit the business more or it will be turn into churn.
Casual	521	The pattern of this cluster show high recency and moderate frequency with

Clusters	RFM Patterns	Cluster Analysis
Shopper		low monetary. This may means customers that have visit the organization in compelling conditions.
Seasonal client	512	Cluster with high recency and moderate monetary with low frequency. This pattern shows the customers that may attend during the seasonal discounts.

IV. CUSTOMER SEGMENTS PREDICTION

In this research two classification techniques will be applied to predict the customer segments, the two techniques will be used separately and combined (using Meta Stacking), in order to identify which of them the suitable for this case. The aim of the research is to find a method to enhance the classification results in the segmentation task. The chosen techniques will be the decision trees (J48) and k- nearest neighbour and will be combined together using Meta stacking. Meta stacking technique will be used to combine the two used algorithm into one algorithm to show how this combination will enhance the performance.

The Customer profile will be constructed to include the customer age, marital status, income range, property ownership, and the segment class. The split percentage will be 70% of the profiles will be training set and 30% testing set. The split will take place randomly. Table 4 shows the distribution of customer percentage in the eight clusters.

TABLE 4 INSTANCE DISTRIBUTION OVER THE EIGHT CLASSES

Class Name	Distribution percentage
New customer	69%
Valuable client	3%
Seasonal client	17%
Profitable client	3%
Uncertain	3%
Zero gain	2%
Loyal client	2%
Casual Shopper	1%

The majority class with 69% was about the new customers, followed by 17% for the seasonal client class and the rest 14% divided over 6 classes, which cause the neglecting the 6 classes with the distribution (3% - 1%) during the prediction phase within the three algorithms.

This problem called imbalanced data. Data set is imbalanced when the classes are not approximately equal [7] that will be fixed by resampling the dataset using Synthetic Minority Over-sampling Technique “SMOTE”. SMOTE is a method to create a synthetic “Examples” from existing real dataset. This technique is used to over-sample the minority class [8].

Table 5 shows the distribution of instances within the 8 clusters after resampling the dataset using the SMOTE technique.

TABLE 5 INSTANCE DISTRIBUTION OVER THE CLASSES AFTER SMOTE

Class Name	Distribution percentage
New customer	17%
Valuable client	12%
Seasonal client	10%
Profitable client	11%
Uncertain	12%
Zero gain	13%
Loyal client	16%
Casual Shopper	9%

V. EVALUATION

For clear clarification the comparison between the three algorithms will be handled two times, the first within the imbalanced dataset and the second will be after applying the SMOTE to resample the dataset. The Classifiers will be evaluated based the detailed accuracy measures from the confusion matrix and ROC “Receiver Operating Characteristics” and its AUC “Area under Curve”.

A. Confusion Matrix accuracy Measures

For each cluster in the confusion matrix a confusion table consists of 4 values as shown in Table 6

TABLE 6 CONFUSION TABLE STRUCTURE

TP = True Positive	FP = False positive
FN = False Negative	TN = True Negative

The accuracy measures calculated from the 4 previous values. The detailed accuracy equations used to evaluate the confusion matrix are as follow [9]:

- Accuracy = TP + TN / (P+N)
- Sensitivity = TP/ (TP+FN)
- Specificity = TN/ (FP+TN)
- Precision = TP/ (TP+FP)

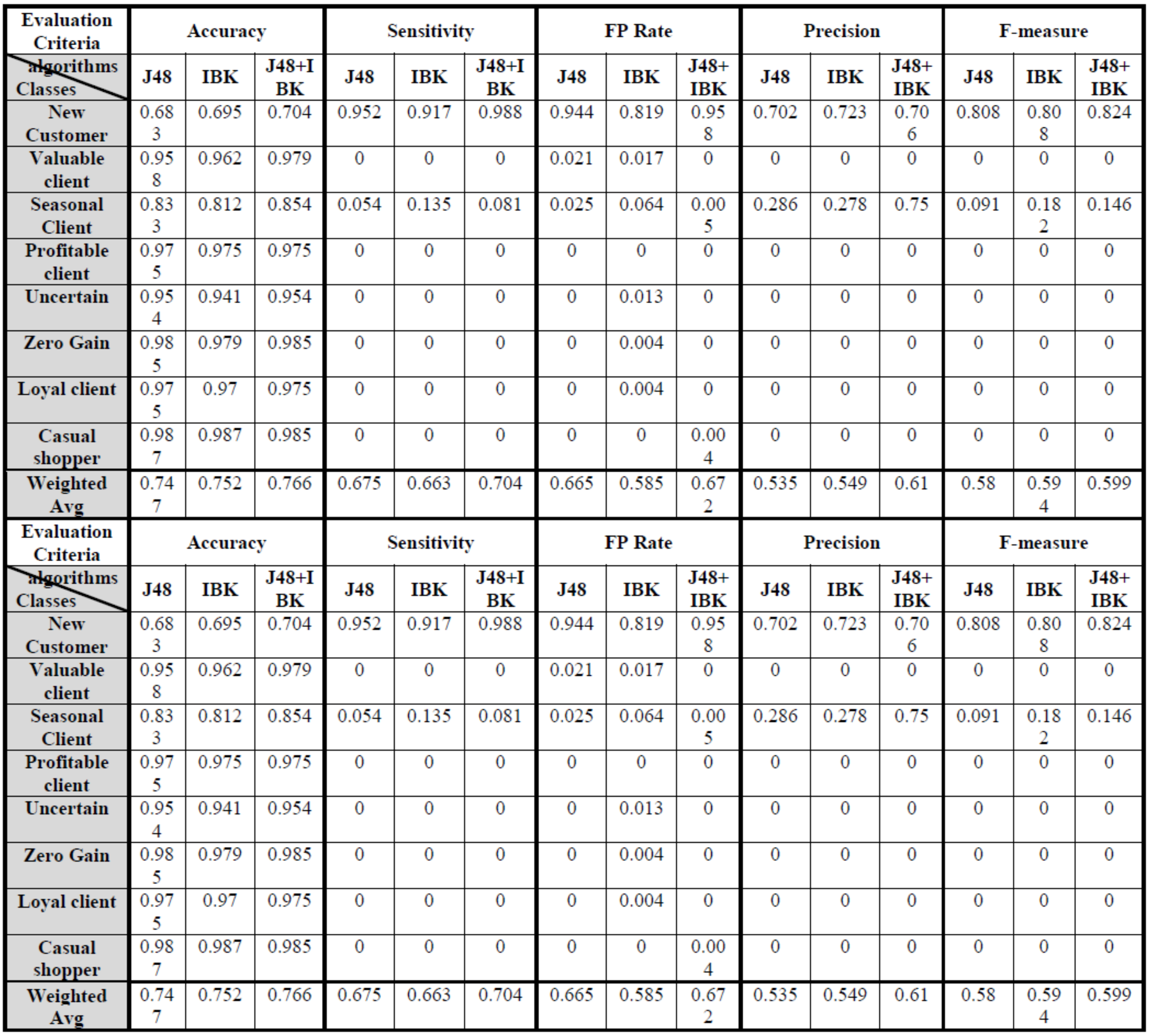
While F-measure is a combination between precision and recall10]**}**, and its equation as follow:

- F-measure = 2(Precision+ Recall)/ Precision+ Recall

F-measure is the way to measure the classifier prediction performance by how the actual set meets the classified set [11].

Table 7 shows the accuracy detailed measures (sensitivity (recall), FP Rate, Precision and F-measure) for the three used algorithms within the imbalanced dataset.

TABLE 7 DETAILED ACCURACY MEASURES FOR J48, IBK AND STACKING FOR IMBALANCED DATA FOR ALL CLASSES



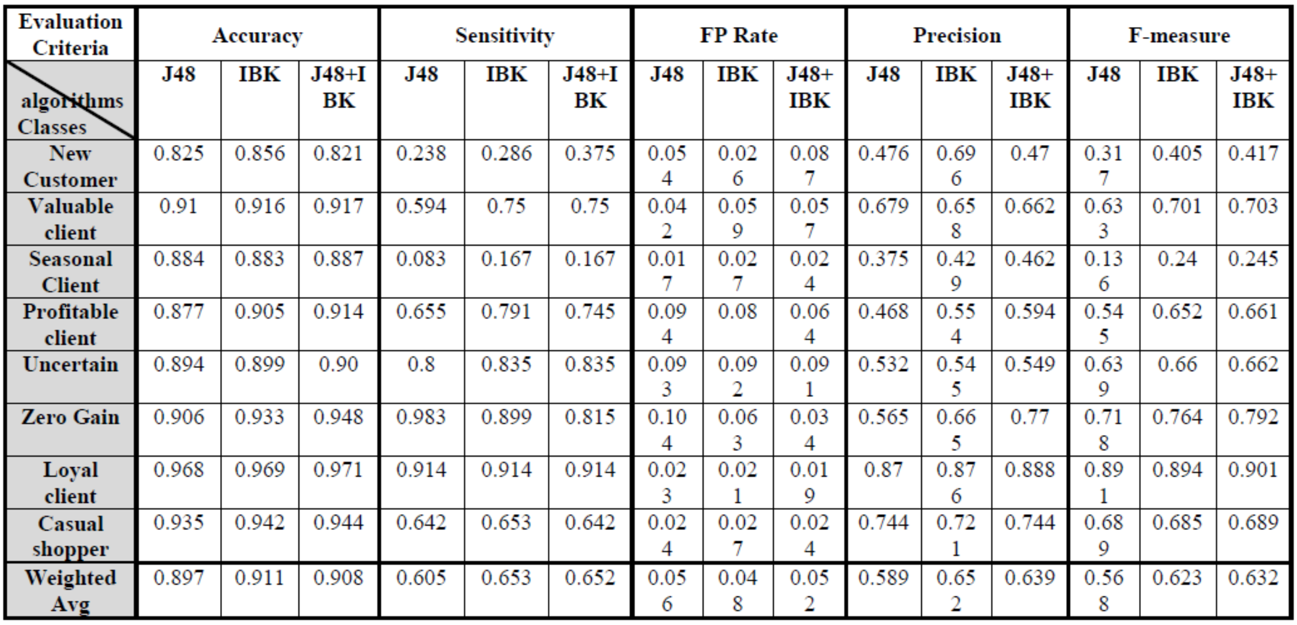
Evaluation Criteria	Accuracy			Sensitivity			FP Rate			Precision			F-measure		
algorithms Classes	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK
New Customer	0.683	0.695	0.704	0.952	0.917	0.988	0.944	0.819	0.958	0.702	0.723	0.706	0.808	0.808	0.824
Valuable client	0.958	0.962	0.979	0	0	0	0.021	0.017	0	0	0	0	0	0	0
Seasonal Client	0.833	0.812	0.854	0.054	0.135	0.081	0.025	0.064	0.005	0.286	0.278	0.75	0.091	0.182	0.146
Profitable client	0.975	0.975	0.975	0	0	0	0	0	0	0	0	0	0	0	0
Uncertain	0.954	0.941	0.954	0	0	0	0	0.013	0	0	0	0	0	0	0
Zero Gain	0.985	0.979	0.985	0	0	0	0	0.004	0	0	0	0	0	0	0
Loyal client	0.975	0.97	0.975	0	0	0	0	0.004	0	0	0	0	0	0	0
Casual shopper	0.987	0.987	0.985	0	0	0	0	0	0.004	0	0	0	0	0	0
Weighted Avg	0.747	0.752	0.766	0.675	0.663	0.704	0.665	0.585	0.672	0.535	0.549	0.61	0.58	0.594	0.599

Evaluation Criteria	Accuracy			Sensitivity			FP Rate			Precision			F-measure		
algorithms Classes	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK	J48	IBK	J48+IBK
New Customer	0.683	0.695	0.704	0.952	0.917	0.988	0.944	0.819	0.958	0.702	0.723	0.706	0.808	0.808	0.824
Valuable client	0.958	0.962	0.979	0	0	0	0.021	0.017	0	0	0	0	0	0	0
Seasonal Client	0.833	0.812	0.854	0.054	0.135	0.081	0.025	0.064	0.005	0.286	0.278	0.75	0.091	0.182	0.146
Profitable client	0.975	0.975	0.975	0	0	0	0	0	0	0	0	0	0	0	0
Uncertain	0.954	0.941	0.954	0	0	0	0	0.013	0	0	0	0	0	0	0
Zero Gain	0.985	0.979	0.985	0	0	0	0	0.004	0	0	0	0	0	0	0
Loyal client	0.975	0.97	0.975	0	0	0	0	0.004	0	0	0	0	0	0	0
Casual shopper	0.987	0.987	0.985	0	0	0	0	0	0.004	0	0	0	0	0	0
Weighted Avg	0.747	0.752	0.766	0.675	0.663	0.704	0.665	0.585	0.672	0.535	0.549	0.61	0.58	0.594	0.599

According to Table 7 the stacking enhanced all the results for accuracy, sensitivity, precision and F-measure. Stacking predicted about 70% correct positives from the actual set, which is more than (J48 by 3% and IBK by 4%). While the percentage of the correct positive predicted from the predicted set is about 61%, which is high than (J48 by 8% and IBK by 6%). According to the F-measure the predicted set matched the classified set by about 60%, which is higher than (J48 by 2% and IBK by 1%). On the other side the percentage of the false prediction to total negatives is by 67% which is not acceptable rate.

While Table 8 shows the accuracy detailed measures (sensitivity (recall), FP Rate, Precision and F-measure) for the three used algorithms within the resampled dataset with SMOTE.

TABLE 8 DETAILED ACCURACY MEASURES FOR J48, IBK AND STACKING FOR ALL CLASSES AFTER APPLY SMOTE



Evaluation Criteria	**Accuracy**			**Sensitivity**			**FP Rate**			**Precision**			**F-measure**		
algorithms Classes	**J48**	**IBK**	**J48+IBK**	**J48**	**IBK**	**J48+IBK**	**J48**	**IBK**	**J48+IBK**	**J48**	**IBK**	**J48+IBK**	**J48**	**IBK**	**J48+IBK**
New Customer	0.825	0.856	0.821	0.238	0.286	0.375	0.054	0.026	0.087	0.476	0.696	0.47	0.317	0.405	0.417
Valuable client	0.91	0.916	0.917	0.594	0.75	0.75	0.042	0.059	0.057	0.679	0.658	0.662	0.633	0.701	0.703
Seasonal Client	0.884	0.883	0.887	0.083	0.167	0.167	0.017	0.027	0.024	0.375	0.429	0.462	0.136	0.24	0.245
Profitable client	0.877	0.905	0.914	0.655	0.791	0.745	0.094	0.08	0.064	0.468	0.554	0.594	0.545	0.652	0.661
Uncertain	0.894	0.899	0.90	0.8	0.835	0.835	0.093	0.092	0.091	0.532	0.545	0.549	0.639	0.66	0.662
Zero Gain	0.906	0.933	0.948	0.983	0.899	0.815	0.104	0.063	0.034	0.565	0.665	0.77	0.718	0.764	0.792
Loyal client	0.968	0.969	0.971	0.914	0.914	0.914	0.023	0.021	0.019	0.87	0.876	0.888	0.891	0.894	0.901
Casual shopper	0.935	0.942	0.944	0.642	0.653	0.642	0.024	0.027	0.024	0.744	0.721	0.744	0.689	0.685	0.689
Weighted Avg	0.897	0.911	0.908	0.605	0.653	0.652	0.056	0.048	0.052	0.589	0.652	0.639	0.568	0.623	0.632

Table 8 shows the results for the same three algorithms after SMOTE the dataset. IBK and Stacking show the same sensitivity percentage about 65 % higher that J48 by 5%. IBK shows highest precision with 65% followed by Stacking 64% then J48 59%. According to the F-measure the predicted set matched the classified set in the Stacking by about 63%, followed by IBK with 62% then J48 with 56%. After resampling the dataset with SMOTE the FPR results get better with 4.8% for IBK followed by Stacking with 5.2% then J48 with 5.6%.

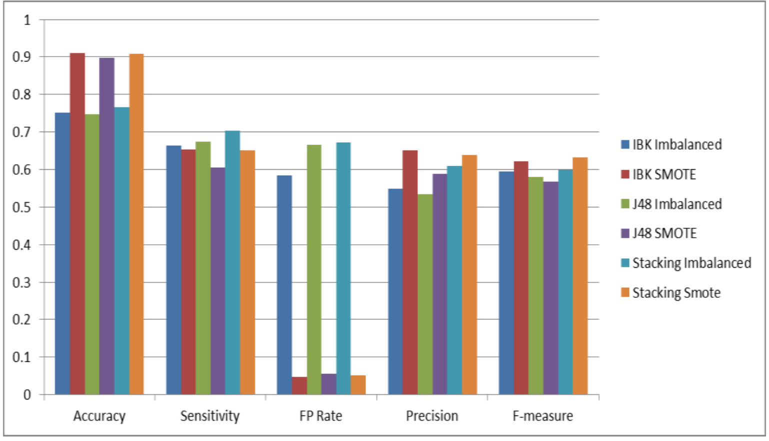


Figure 1 Performance Measures for J48, IBK and Stacking for Imbalanced and SMOTE Data

Figure 1 visualizes the three classifiers performance within the imbalanced data and after the SMOTE from accuracy, sensitivity, FP rate, precision and F-measure point of view. In F-measure stacking smote achieve the highest result followed by IBK smote while J48 smote got the lowest average. The Smote improve the FP rate results with a very high difference, also improve the precision results with the highest percentage for IBK followed by Stacking. In the sensitivity Smote archived lower rates than those within the imbalanced data, but not with high differences. Stacking sensitivity decreased by 5% while J48 decreased by 7% and IBK decreased by 1%.

B. ROC and AUC

ROC curves will be used to represent the relation between sensitivity (TPR) in Y-axis and the FPR (1-specificity) on X-axis. AUC is a single value to facilitate the comparison between the classifiers.

Table 9 presents the weighted AUC average for the 3 classifiers within the 2 cases (imbalanced and smote). It can be mentioned that SMOTE the dataset provides better results than the imbalanced case. In the SMOTE case, stacking achieved high results (90.1%) followed by IBK with (90%) then J48 with (87%).

TABLE 9 AUC RESULTS

	J48 Imbalanced	**Imbalanced IBK**	**Imbalanced stacking**	**SMOTE J48**	**SMOTE IBK**	**SMOTE stacking**
New customer	0.605	0.575	0.468	0.682	0.725	0.761

Valuable client	0.496	0.271	0.361	0.948	0.956	0.953
Seasonal client	0.645	0.532	0.567	0.717	0.756	0.739
Profitable client	0.53	0.479	0.323	0.9	0.932	0.924
Uncertain	0.451	0.385	0.439	0.939	0.947	0.952
Zero gain	0.674	0.701	0.743	0.963	0.973	0.971
Loyal client	0.678	0.603	0.762	0.987	0.991	0.988
Casual shopper	0.404	0.79	0.205	0.96	0.978	0.957
Weighted Average	0.6	0.557	0.485	0.879	0.9	0.901

Within the imbalanced data J48 gives the highest AUC results for new customer, valuable client, seasonal client, profitable client and uncertain classes. While for both zero gain and loyal client classes, the imbalanced stacking gives the largest AUC. IBK provide the largest AUC for the casual shopper class.

For the resampled dataset with SMOTE; IBK provides the highest AUC for valuable, seasonal, profitable, zero gain, loyal client and casual shopper classes. Stacking followed IBK with small differences and gives highest AUC for new customer and uncertain classes. While J48 gives the lowest AUC for all classes.

From the previous results, it can be concluded that SMOTE the data set clearly affected the sensitivity and specificity in a positive way for all the classes.

Another important measure for the classifiers performance other than their AUC is the classifiers location in the ROC space. Table 10 shows the FPR and TPR rate for each classifier, to be used in Fig 2 to display how away their location from the line of no-discrimination (diagonal line or random classifier).

TABLE 10 FPR AND TPR FOR J48, IBK AND STACKING

Imbalanced J48		**SMOTE J48**		**Imbalanced IBK**		**SMOTE IBK**		**Imbalanced Stacking**		**SMOTE stacking**	
FPR	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**
0.665	0.675	0.056	0.605	0.585	0.663	0.048	0.653	0.672	0.704	0.052	0.652

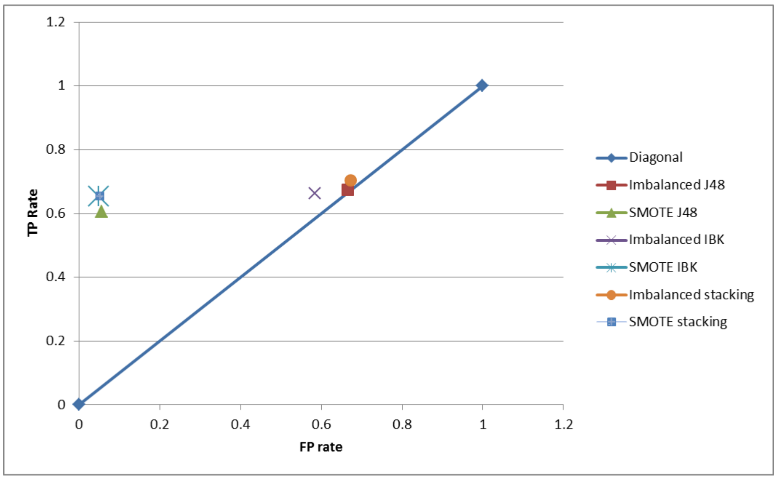


Figure 2 Classifiers Location on ROC Space

Stacking and IBK in the smote case placed in the best locations as they show high TP rates and very low FP rates, with close rates as it seems that both of them on the same location.

VI. CONCLUSION

This research runs within the aCRM (analytical CRM) category, which focuses on analysis and interpretation of the given dataset. Following this idea, the research placed the analytical phases using data mining techniques for clustering the customers, and then developed a prediction model.

The clustering has been performed based on the RFM model analysis which is basically segment the customer according to their purchasing behaviours. The objective of the classification phase was to identify which is the suitable classifier to predict customer segment based on the demographic data for the study dataset.

For the study dataset two algorithms have been chosen to test how they will perform (J48 as a decision tree and IBK from the lazy algorithms) then a stacking from both of them has been evaluated as the third classifier. During the first stage the result showed the dataset to be imbalanced so a SMOTE technique has been used to resample the dataset.

The results of the three classifiers within the two cases (imbalanced and after resampling) have been evaluated in order to select the suitable classifier for this dataset. The overall average for the accuracy measures within the imbalanced case was acceptable, but there were about 6 customers segments were completely neglected during the predictions. Sensitivity, Precession and F-measures were zeroes for loyal client, valuable client, profitable client, casual shopper, uncertain and zero gain classes. Although these results from the overall weighted average the stacking achieved the highest results. While according to the ROC curves results, J48 achieved better results.

After resampling the dataset better accuracy measures achieved for each algorithms from the weighted overall and the classes point of view. The same enhancement happens for the ROC curve and AUC results. In the resampling case, IBK was the better algorithm followed by Stacking with very limit difference.

REFERENCES

[1] Ed Peelen, *Customer Relationship management*.: FT-Prentice Hall, 2005.

[2] Derya Birant, *Data Mining Using RFM Analysis*.: INTECH Open Access Publisher, 2011.

[3] Pratik Biswas and ParthaSarathi Bishnu, "Application of Data Mining and CRM in Banking Sector Medical Insurance ," *International Journal of Innovative Research in Computerand Communication Engineering*, 2015.

[4] T. Femina Baharia and M. Sudheep Elayidom, "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour," in *Proceedings of the International Conference on Information and Communication Technologies*, India, 2015.

[5] Ching-Hsue Cheng and You-Shyang Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Systems with Applications*, pp. 4176-4184, 2009.

[6] Randy Collica, *Customer Segmentation and Clustring Using SAS Enterprise Miner*, 2nd ed.: SAS Institute, 2011.

[7] Nitesh V. Chawla, "DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW," in *Data Mining and Knowledge Discovery Handbook*., 2005, pp. 853-876.

[8] Nitesh V Chawla, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 2002.

[9] Malay Kumar Kundu, "Advanced Computing, Networking and Informatics, Volume 2," in *Wireless Networks and Security Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (Icacni-2014)*, 2014.

[10] Zhenhua Li, "Computational Intelligence and Intelligent Systems," in *4th International Symposium on Intelligence Computation and Applications, ISICA 2009*, Huangshi, 2009.

[11] Lior Rokach and Oded Maimon, *Data Mining with Decision Trees Theory and Applications*.: World Scientific Publishing Company Pte Limited, 2014.