NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal
ISSN No:2581-4230
VOLUME 4, ISSUE 4, Apr. -2018

RECOMMENDATION ENGINE FOR B2B CUSTOMERS IN TELECOM BY CUSTOMIZING KNN ALGORITHM

MAYANK GOEL
Manager, Tata Consultancy Services
Mayankgoel_ymca@yahoo.com

ABHISHEK KUMAR TIWARI
Manager, Tata Consultancy Services
tiwariabhishek1120@gmail.com

HARSHAL SURESH PATIL
M.Sc (Computer Science)
harshalpatil.nmu@gmail.com

ABSTRACT:

K-Nearest Neighbours (KNN) is one of the most popular algorithms for classification; however traditional KNN algorithm has two limitations:

1. There is no weight difference between closest training examples
2. Revenue Prediction is not feasible

In this paper, we propose a KNN type method for classification that is focussed at overcoming above shortcomings. Our method constructs a cross-sell penetration model using Revenue, Usage, and Firm graphics data for targeting telecom Enterprise Customers. Value of K is varied for different data, and is optimally chosen based on classification accuracy. After Propensity of an account is determined from traditional algorithm, weights are assigned to nearest neighbours and Revenue is determined.

BUSINESS PROBLEM/OBJECTIVE:

Business is currently offering a total 36 mobility (Red, Connect, Tablet etc.) as well as fixed line products (Toll free services, Internet leased lines etc.) to 80K accounts across different geographic circles

However, the depth of penetration of certain products to these accounts is very low. Considering internet leased line (ILL) as the product for which experiment was carried out, penetration was as low as 6% which is where business is looking to derive revenue opportunity by identifying right set of customers having appetite to buy this product. With the available annualized product revenue information of 36 products, connections & usage data, the objective is to provide business with account that have high propensity to buy ILL product and potential revenue opportunity that can be derived. To achieve this, K-Nearest Neighbours algorithm was implemented

Assumption: Audience reading this paper is already aware of K-NN Algorithm or same can be referred in the Appendix section.

APPROACH:

a) WHY KNN OVER LOGISTIC REGRESSION:

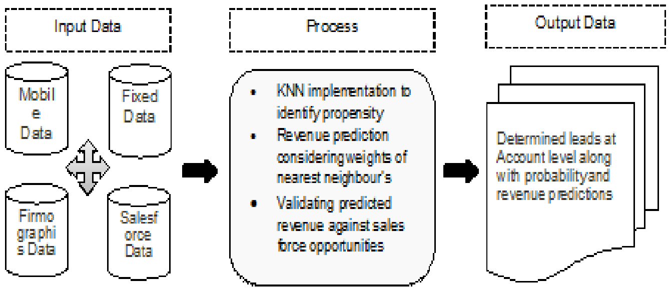
The distribution of revenue data being quite skewed/volatile and not linearly separable made us to approach the problem using KNN which is a non-parametric algorithm. Even the difference was observed at implementation stage wherein accuracy with KNN model was observed to be higher (75%) as against Logistic model where it was 62%

Another related reason to use KNN is large number of predictor revenue, connections, usage variables which would have given high dimensionality problem with Logistic but still better results with KNN. The reason SVM was not implemented being that it can be painfully inefficient to train and calculation of revenue based on similar nearest accounts was not feasible.

b) CHALLENGE WITH KNN- PREDICTION OF REVENUE USING SAME ALGORITHM AND SAME MODEL:

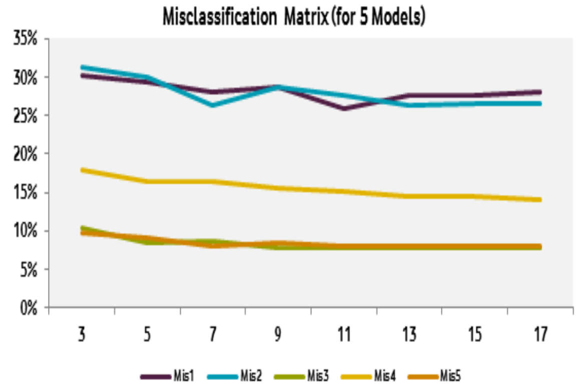
We know that traditional KNN classification finds the K closest observations based on Euclidean distance and then classifies the test point to the majority. Similar logic holds true for KNN regression for interpreting the revenue numbers. However thinking logically, highest Weightage should be given to closest observation (due to more similarity of attributes)and least to farthest observation from test data point in terms of Euclidean distance which at this point KNN algorithm doesn’t takes into consideration. Keeping above in mind, leveraging the capability of KNN algorithm and calculation of revenue as part of same model which would give more accurate numbers to business is a challenge.

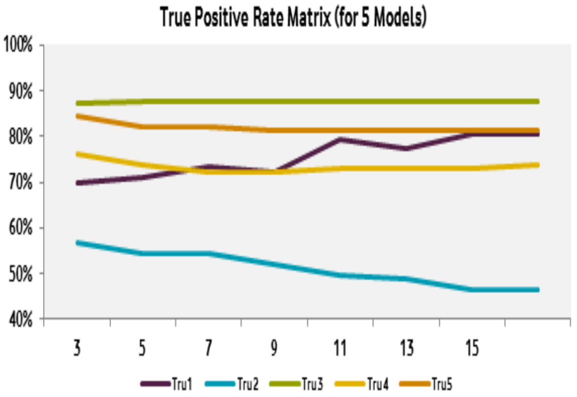
c) MODELLING STEPS:


Input Data
Mobil e Data
Fixed Data
Firmo graphi s Data
Salesf orce Data
Process
KNN implementation to identify propensity
Revenue prediction considering weights of nearest neighbour's
Validating predicted revenue against sales force opportunities
Output Data
Determined leads at Account level along with probability and revenue predictions

d) SELECTING OPTIMISED VALUE OF K:

Value of K being the only parameter that needs to be optimised should be carefully selected. It should be large enough that noise in the data is minimized and small enough so the samples of other classes are not included. For our modelling, 5-fold cross validation was done by varying K from 3 to 17 and studying Misclassification True positive rate Matrix. Below charts were obtained:

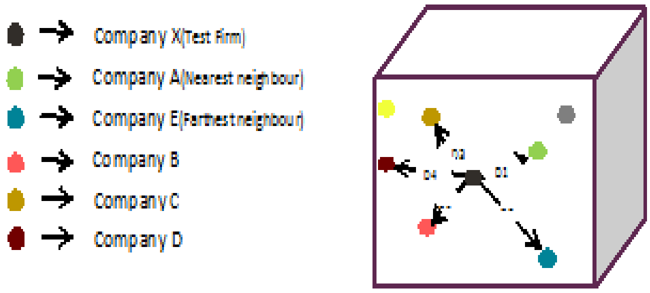




From above, the graphs corresponding to iteration 1& 2 were showing substantial variation and were picked for consideration while graphs from iterations 3,4,5 were showing flat pattern and hence neglected. Second iteration was further rejected with Tru2 quite low, it can be interpreted from first iteration that at K=11, Misclassification (Mis1) is least whereas True positive rate (Tru1) is maximum and thus, K=11 was selected as final optimised value.

e) ASSIGNING WEIGHTS TO NEAREST NEIGHBOUR'S:

For convenience purpose, 5 nearest neighbours are shown in figure as against actual 11 for easy visualization and understanding



Let:

D1:D5 - Euclidean distance of nearest neighbours from Company X

W1:W5 - Weights assigned to 5 nearest neighbour's

R1:R5 - Current Revenue of nearest neighbours

Keeping above in mind, below method has been devised to assign weights.

Wi = Dk -
Di

Where: K = 5, i= 1: K

With this approach, below weights gets assigned in our scenario:

W1= D5 -
D1 ,

W2= D5 -
D2,

W3= D5 -
D3,

W4= D5 -
D4,

It is evident from here that weights will be assigned in descending order from nearest to farthest.

f) REVENUE CALCULATION:

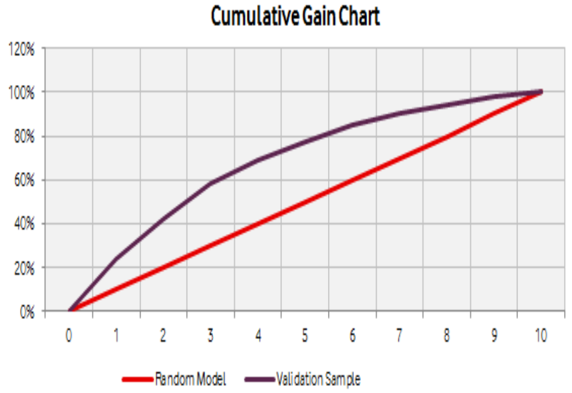
Revenue numbers of nearest neighbour's buying the product are already known. Having calculated

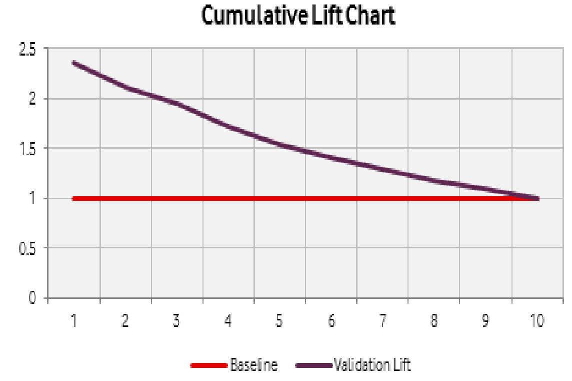
Revenue (R) = W1 * R1 + W2 * R2 + W3 * R3 + W4*R4 + W5*R5 the weights, unknown revenue R of company X can be computed as below:

The root mean square error computed with customized Weightage approach came out to be 283 whereas with KNN regression approach same turns out to be 678 which show significant improvement with our suggested methodology.

g) RESULTS:

Considering the Event rate of ILL product being too low (6%), oversampling of events was carried and 5 KNN models were built which was finally ensemble to reduce the noise. Below were the cumulative Gain and Lift chart of the Ensemble model:





From above, it can be inferred that first three declined captures close to 60% of the positive responses with accuracy of model being 75%.

h) BENEFITS FOR BUSINESS:

Total 20% leads from overall accounts (80K) have been provided to business which is potential customers for buying ILL product.

Incremental revenue opportunity for ILL product considering conversion of top 100 accounts stands out to be INR 22M.

Likelihood numbers provided above had been a great help for business if intention is to target few handful of accounts from top 100.

CONCLUSION:

Rather than adopting the traditional KNN algorithm for predicting likelihood and revenue based on equal Weightage of nearest neighbour's, we propose customising KNN which deploys novel weights taking a significant step to produce output with better classification accuracy. Extensive experiments and comparisons using standard datasets show that our method is competitive among the state-of-the-arts and the methodology can be extended across Financial and E-commerce domains. Future work includes investigating and implementing other Ensemble algorithms i.e. Gradient Boosting, Random Forest and selecting the appropriate method based on product.

REFERENCES:

I. D.T. Green and J. M. Pearson, "The examination of two web site usability instruments for use in B2C Online Libraries organizations," *Journal of Computer Information Systems*, Vol. 49, No. 4, 2009, pp. 19-32

II. Neha, Dheeraj Malhotra, Monica Malhotra, and Jatinder Singh. "Online Libraries Website Recommendation Using Semantic Web Mining and Neural Computing." *Procedia Computer Science 45 (2015): pp. 42-51, ELSEVIER.*

III. K. Abbas, and Y. Niloofar, "A Proposed Classification of Data Mining Techniques in Credit Scoring", in *Proceedings of the 2011International Conference on Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia*, 2011, p. 416-424.

IV. N.C. Hsieh, and L.P. Hung, "A data driven ensemble classifier for credit scoring analysis", *Expert Systems with Applications*, vol. 37,pp. 534–545, 2010.

V. T. Wang and Y. Lin, "Accurately predicting the success of B2B ecommerce in small and medium enterprises," *Expert Systems with Applications, Vol. 36, No. 2, published by Elsevier, 2009, pp. 2750–2758.*

VI. D. Jannach, M. Zanker, A.Felfernig and G. Friedrich, Recommender Systems – *an introduction Cambridge University Press, 2010*

VII. Levent Ertoz, Michael Steinbach, Vipin Kumar. Finding Clusters of Different Sizes, Shapes ,and Densities in Noisy, *High Dimensional Data Proceedings of the third SIAM International conference on Data Mining 2003.1.P47-58*