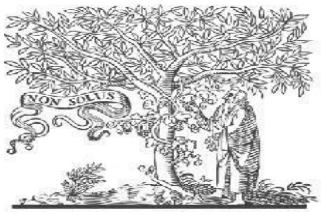
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 1st Jan 2021. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-12

DOI: 10.48047/IJIEMR/V09/I12/140

Title: **PREDICTING AND DEFINING B2B SALES SUCCESS WITH MACHINE LEARNING**
Volume 09, Issue 12, Pages: 823-828
Paper Authors

JENNY MANEESHA, GUVVA SHIREESHA, ATHMAKUR DIVYA, E.LAXMAN

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

PREDICTING AND DEFINING B2B SALES SUCCESS WITH MACHINE LEARNING

JENNY MANEESHA[1], GUVVA SHIREESHA[2], ATHMAKUR DIVYA[3], E.LAXMAN[4]

[1,2,3,] B TECH Students, Department of CSE, Princeton Institute of Engineering & Technology For Women, Hyderabad, Telangana, India.

[4] Assistant Professor, Department of CSE, Princeton Institute of Engineering & Technology For Women, Hyderabad, Telangana, India.

ABSTRACT:
Business to Business (B2B) sales forecast can be described as a decision-making process, which is based on past data (internal and external), formalized rules, subjective judgment, and tacit organizational knowledge. Its consequences are measured in profit and loss. The research focus of this paper is aimed to narrow the gap between planned and realized performance, introducing a novel model based on machine learning techniques. Preliminary results of machine learning model performance are presented, with focus on distilled visualizations that create powerful, yet human comprehensible and actionable insights, enabling positive climate for reflection and contributing to continuous organizational learning.
Keywords: B2B, internal external storage, reflections, security services.

1. INTRODUCTION

The paper and packaging company that provided the data for this research has a long history of sales expertise. This expertise is captured predominantly in the intuition of sales representatives, many of whom have worked in the industry for 20 years or more. Intuition is not easy to record and disseminate across an entire sales force, however, and thus one of the company's most valuable resources is inaccessible to the broader organization. As a result, the company tasked this team with extracting the most important factors in driving sales success and modeling win propensities using data from their customer relationship management (CRM) system. Most prior work in this space has been performed by private companies, both those that have developed proprietary technologies for internal use and those that sell B2B services related to predictive sales modeling. As a result, research in the field is typically unavailable to the public. Some examples include Implisit a company recently acquired by Salesforce.com that focuses on data automation and predictive modelling and Insight Squared, which sells software that includes a capability to forecast sales outcomes. The academic work that does exist either is related to forecasting aggregate sales instead of scoring opportunity level propensity, or is based on custom algorithms that fall outside the standard tools used by data scientists in industry. The earliest relevant publication dates only to 2015, in which a joint team from Chinese and US universities employed a two-dimensional Hawkes Process model on seller-lead interactions to score win propensity. Other relevant research has cantered on applying highly accurate machine learning algorithms based on sales pipeline data to integrate the insights they produce into an organization's

practices, and explaining the output of black-box machine learning models. Considering the lack of visibility into work predicting sales outcome propensity, this research serves to create an initial baseline of understanding on the subject. This project applies and compares several well-known methods for classifying and scoring propensities, a majority of which fall into the category of decision tree modeling.

2. LITERATURE SURVEY

The learning is characterized by the change of behavior as a result of an individual and/or group exposure to experience (Kljajić Borštnar et al., 2011). Two types of learning are distinguished: the single-loop and the double-loop learning (Argyris, 1996; DiBella and Nevis, 1998; Gephart, Marsick, Mark, VanBuren and Spiro, 1996, Nonaka and Takeuchi, 1995). The double-loop learning refers to not just changing the behavior in order to achieve the stated goal (single loop), but changing mental models, visions and beliefs, and therefore organizational knowledge. With the proposed approach we build a foundation to achieve the double-loop learning – as a basis to establish new premises (i.e. paradigms, schemes, mental models or perspectives), with potential to override existing ones (Nonaka and Takeuchi, 1995). Same authors are fully aware that an effort to question and rebuild existing perspectives, interpretation of frameworks or decision premises can be very difficult to implement in an organization; it requires persistent activities. Organizational learning presents ongoing effort of creating organizational knowledge. Team learning, personal mastery and mental models principles (Senge, 1990) are built-in into organizational knowledge. In this paper we propose a classification model, which builds on insights from B2B sales professionals. Insights are presented in a form of sales history described with features reflecting attributes of sales process and B2B relationships (Bohanec et al., 2015). Machine learning techniques are applied to build the classification model, which is capable to classify future, unseen sales opportunities. The classification model represents the organizational knowledge which is presented and visualized in a human comprehensible form to support the double-loop learning process within an organization. Our aim is to investigate whether it is possible to develop such a model, based on B2B sales history, which supports process of forecasting and transparent reasoning.

Machine learning (ML) in our context is interpreted as an acquisition of structural descriptions from examples (Witten, Eibe and Hall, 2011). The fact that it leverages different models and algorithms to approximate complex theories which are difficult to be exactly represented with other mathematical tools, connects it to the field of artificial intelligence. ML has been successfully applied in different fields, e.g. medical diagnostics, spam filtering, OCR, internet browsers etc. (Liao, Chu and Hsiao, 2012; Ngai et al., 2009; Bose and Mahapatra, 2001). ML techniques take training data set to learn relationships needed to categorize new, yet unseen, objects to target categories (Witten et al., 2011; Robnik-Šikonja and Kononenko, 2008). Some classification models produced are able to explain their decisions, which can help in better adoption of ML techniques in practice due to participant's faster understanding of ML insights (Robnik-

Šikonja and Kononenko, 2008; Collopy, Adya and Armstrong, 2006).

3. RELATED STUDY

The data for this project were sourced from the company's Salesforce.com customer relationship management system (SFDC). SFDC is a software-as-aservice application that allows sales teams to record details about customer relationships and sales opportunities as they move through the sales pipeline. The data included a static snapshot of details on sales employees, customer accounts and account histories, individual customer opportunities, sales representative activities, and contact information. Some inputs in the system were automatically generated and easily readable by machine. For others, sales representatives entered customer information manually, either via restrictive forms of entry such as a drop-down list or numeric field, or freeform, in a text field or uploaded as an attachment. To clean the data and cut out inessential information prior to modeling, the team first filtered out all entries created before Apr. 1, 2016 when the system was formally launched for the company1 . Variables with a high percentage of null values were then excluded to ensure a sufficient sample size. The remaining variables were further screened based on potential importance determined by conversations between the team and key company stakeholders. Additionally, data exploration resulted in several opportunities for feature engineering and custom variables to capture potential influence not captured in the default fields. The following are several examples of custom fields generated:

1. Fields Completed — count of the number of fields completed in one record.

2. Task Count — count of the number of tasks for the customer account associated with an opportunity.

3. Age-related variables — analyzes the impact from the age of opportunities. a. Open Time — the duration that an opportunity remained open in the system. b. Last Action time — the duration from when an opportunity was created to the time of last activity on that opportunity c. Valid Open Time — a Boolean variable that equals 1 for opportunities with positive Open Time and 0 for the remaining opportunities. After a number of iterations between modeling and feature engineering, the final master table used in this analysis included 15 variables and was built on the opportunity-level. Account information related to each customer and custom variables from other tables were also merged into this set. Each observation on this master table and on previous table iterations were considered to be individual sales opportunities described by a number of features and associated variable values. Opportunities could be considered synonymous with sales "deals" and originally included both open and closed opportunities before being filtered to maintain only closed. Each variable corresponded to a filled or calculated field in the SFDC system, characterizing the opportunity's duration, type, amount, or any other information.

4. PROPOSED SYSTEM

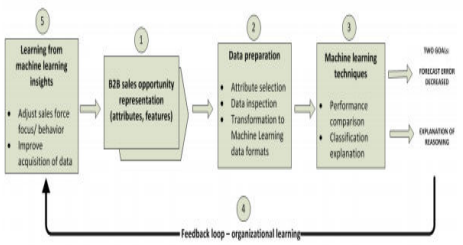
The research team employed several well-known classification models to extract important features from the data, in addition to calculating the win/loss propensity for each opportunity record. With the goal of modeling probability, the team chose different supervised machine

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

learning algorithms that fit these criteria: Logistic Regression, Decision Tree, Random Forest, and XGBoost. In each of these supervised algorithms, the classifier was pre-defined with an iterative variable selection process. A classification model was then built with a training set split from the master table and used to predict win propensities examined by the actual win or loss of the opportunities in the testing set built from the remainder of observations. Variable selection was a critical component of this project. As previously stated, variables came directly from the SFDC system and went through a series of data processing steps. The main purpose of this research was to interpret features that gave the most useful information in terms of win propensity prediction accuracy. Both the quality and quantity of variables significantly affected the accuracy and efficiency of all algorithms. An important consideration about the current data was the widely varying quality of variable inputs. This issue created constraints on the algorithm-generated selection results. Therefore, the variable selection process also involved constant communication and validation between the team and company. The four algorithms used in this research are briefly described below:

- Multiple Logistic Regression — a generalized linear model (GLM) that describes the relationship between a binary dependent variable and more than one predictor.
- Decision Tree — a non-parametric algorithm that makes sequential, hierarchical decisions about the outcomes based on the predictors.
- Random Forest — an ensemble algorithm that constructs a multitude of decision trees and outputs the mode of the classes, correcting the overfitting habit of decision trees.
- XGBoost — an implementation of gradient boosted decision trees that minimize the loss when producing an ensemble of weak decision trees. The metrics for evaluating the models comprised the following:

1. Accuracy—the percentage of correctly predicted opportunities over the total number of opportunities. Outputs were given in confusion matrices that illustrated a more detailed level of accuracies: a. Precision — the percentage of correctly predicted won opportunities over the total number of predicted won opportunities. b. Recall — the percentage of correctly predicted won opportunities over the total number of actual won opportunities.
2. Access to variable importance — certain algorithms provided information to evaluate the importance of variables included in the model. The metric used was “percentage increased Mean-squared-error (%IncMSE)”, which implied the loss of accuracy if a certain variable was missing in the model.
3. Efficiency — resources used to build the model including time, memory, and complexity.



However, the random forest model not only exhibited exceptional accuracy, but also provided importance’s at the variable level. Because of a requirement for dummy variables, the XGBoost model output importances for every possible value of all categorical variables, producing a very high number of importances that was much less easy to read and act on for the company. The random forest proved best in

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

every metric except run time, which was over 30 minutes for the full model. By creating individual models at the division level, however, this was improved to a manageable 77.87 seconds for all divisions combined. Based on these results, random forest was selected as the optimal model to provide insights to the company. A division-level model not only improved model performance, but was critically important in deriving insights for the company. Within the organization, different divisions exhibit significant differences in client profiles, processes, and use of the SFDC system. By creating a model for each division, recommendations could be tailored to each business unit individually. Additionally, it was determined that two models should be created for each division, one incorporating "met variables"—or variables describing the data itself more than the sales opportunity2 —and one excluding them. This resulted in models with very different accuracies and variable importance's, but allowed for the isolation of variables useful for prediction in contrast to those more informative of how the system is used.

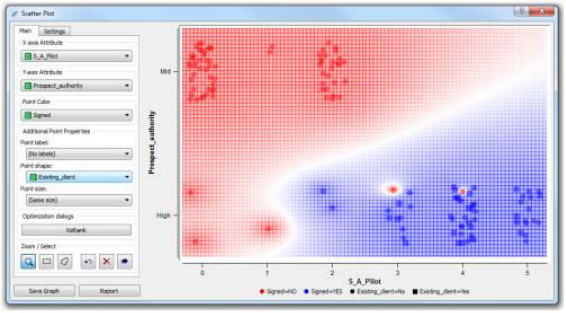


Fig.4.1. Variable importances.

5. CONCLUSION

Despite initial inconsistencies in the data, overall accuracy appeared promising and indicated further improvements could be made with better data quality and quantity, more feature related investigation and tuning, or perhaps different methods such as neural nets. The analysis also uncovered new insights into what is important regarding sales success. But new insights are often accompanied by new questions: For instance, what kinds of data need to be captured to improve the model's predictive capabilities? How does the culture need to change to improve data capture? This cascade is to be expected, as the broader project lends itself to being a heavily iterative process. There may appear to be a seemingly infinite pool of potential next steps to take in this case. With this in mind, there are a few the team would recommend as the most prudent to consider. Currently, the company could feasibly use the non-meta-variable model to attempt prediction on opportunities in progress for those divisions where accuracy is adequate. To better achieve the objective of predicting open opportunities, it would be prudent to capture and model how opportunity fields change over time, perhaps via periodic snapshots. This way, the company would be able to make predictions at different stages in the opportunity lifecycle. Another important application of these kinds of prediction models is to assist in determining where to invest sales time and resources for business planning optimization. Predictions from accurate models are also worth rolling up into aggregate sales forecasts and adjusting existing "bottom-up" methods.

REFERENCES

[1] Miller, G.A. (1956): "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information", Psychological Review, Vol. 63, pp. 81-97.

[2] Monat, J. P. (2011): "Industrial sales lead conversion modeling", Marketing Intelligence & Planning, Vol. 29, Iss: 2, pp.178 – 194.
[3] Ngai E.W.T., Xiu Li, Chau D.C.K. (2009): "Application of data mining techniques in CRM: a literature review and classification", Expert Systems with Applications, Vol. 36, pp. 2592– 2602.
[4] Nonaka I., Takeuchi H. (1995): "The knowledge creating organization", Oxford University Press, New York. Senge P. (1990): "The Fifth discipline: The Art & Practice of the Learning Organization", Doubleday Currency, New York. Sein M.K., Henfridsson O., Purao S., Rossi M., Lindgreen R. (2011): "Action Design Research", MIS Quarterly, Vol 35, pp. 37-56.
[5] Shmueli G., Patel N.R., Bruce P.C. (2007): "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner ", John Wiley & Sons.
[6] Simoff S. J., Böhlen M. H., Mazeika A. (eds.) (2008): "Visual data mining", Lecture Notes in Computer Science, Springer. Witten, I.H., Eibe F., Hall M.A. (2011): "Data mining – Practical Machine Learning Tools and Techniques", third edition, Elsevier.