Predicting and Defining B2B Sales Success with Machine Learning

Stephen Mortensen, Michael Christison, BoChao Li, AiLun Zhu, Rajkumar Venkatesan
sam8sp@virginia.edu, mjc7nz@virginia.edu, bl2fh@virginia.edu, az8ec@virginia.edu, venkatesanr@darden.virginia.edu

Abstract - **The objectives of this project are two-fold: 1) to use statistical modeling techniques to help a Fortune 500 paper and packaging company codify what drives sales success and 2) to develop a model that can predict sales success with a reasonable degree of accuracy. The desired long-run result is to enable the company to improve both top-line revenue and bottom-line profits by increasing sales close rates, shortening sales cycles, and decreasing the cost of sales. The research team generated several models to predict win propensities for individual sales opportunities, choosing the model with the greatest predictive power and ability to generate insights to use as the backbone for a client tool. To accomplish this, the team leveraged structured and unstructured data from the company's Salesforce.com customer relationship management system. The team experimented with several techniques including binomial logit and various decision tree methods, including boosting with gradient boost and random forest. Individual attributes of customers, opportunities, and internal documentation methods that have the greatest influence on sales success were identified. The best model predicted win propensity with an accuracy of 80%, with precision and recall of 86% and 77%, respectively, which proved to be an improvement over current sales forecast accuracy.**

Index Terms - statistical modeling; decision tree; machine learning; process improvement

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 [1]—a company recently acquired by Salesforce.com that focuses on data automation and predictive modeling—and InsightSquared [2], 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 [3]. Other relevant research has centered around applying highly accurate machine learning algorithms based on sales pipeline data to integrate the insights they produce into an organization's practices [4], and explaining the output of black-box machine learning models [5].

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.

DATA SOURCE AND PROCESSING

The data for this project were sourced from the company's Salesforce.com customer relationship management system (SFDC). SFDC is a software-as-a-service 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

978-1-7281-0998-5/19/$31.00

launched for the company[1]. 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.

MODELS AND METHODOLOGY

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 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.

[1] Prior to this date, portions of the company used the system, but it had not been rolled out companywide.

Based on an initial analysis, the tested models produced the following results:

TABLE I
MODEL COMPARISON

Model	Accuracy	Precision	Recall	Run Time (seconds)	Importance
Logistic Regression	63.78%	68.80%	51.84%	2.35	NO
Decision Tree	49.27%	52.62%	39.69%	**0.20**	NO
Random Forest	**82.39%**	**77.14%**	**79.07%**	1,965.43 (77.87*)	**YES (VARIABLES)**
XGBoost	48.85%	53.11%	47.51%	95.92	YES (ALL VALUES)

Accuracy for all models are based on initial runs with limited parameter tuning. Additional tuning may have improved the accuracy of some models, such as XGBoost and decision trees. However, the random forest model not only exhibited exceptional accuracy, but also provided importances 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 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 "meta-variables"—or variables describing the data itself more than the sales opportunity[2]—and one excluding them. This resulted in models with very different accuracies and variable importances, but allowed for the isolation of variables useful for prediction in contrast to those more informative of how the system is used.

Predictions were made with a conservative win threshold of 55% confidence, which limits type 1 errors, or false positives. Accuracy and importance figures were then computed from these predictions and the labels, which helped to iteratively select variables and inform insights.

[2] Meta-variables include: Fields Completed (a measure of how thoroughly an opportunity owner input data into the system), Open Time (a measure of how long an opportunity was open before it was designated "won" or "lost"), Last Action Time (a measure of the time between the last action in the system and the close date of an opportunity), Tasks Completed (a measure of the number of task objects created and attached to an opportunity's account object), and Valid Open Time (a binary variable indicating whether the open time was positive or negative: negative open times are possible when an opportunity was created after it had already been won or lost).

RESULTS AND ANALYSIS

After performing the full division-level random forest analysis, the following results were assessed for each division:

TABLE II
RANDOM FOREST MODEL WITH META-VARIABLES

Division	Records	Precision	Recall	Accuracy
1	1,061	88.89%	56.80%	83.96%
2	3,750	90.77%	72.72%	86.24%
3	11,622	81.38%	74.76%	76.91%
4	6,401	87.79%	79.01%	84.91%
5	5,366	90.14%	87.45%	86.73%
6	5,092	85.69%	75.79%	81.58%
7	325	78.57%	40.74%	88.27%
8	2,843	85.54%	61.85%	81.55%
9	4,004	91.96%	82.56%	85.66%

TABLE III
RANDOM FOREST MODEL WITHOUT META-VARIABLES

Division	Records	Precision	Recall	Accuracy
1	1,061	80.00%	11.83%	70.94%
2	3,750	66.94%	11.16%	62.56%
3	11,622	69.82%	62.35%	64.77%
4	6,401	77.82%	68.54%	75.94%
5	5,366	79.58%	83.98%	77.45%
6	5,092	73.11%	59.20%	68.74%
7	325	0.00%	0.00%	83.33%
8	2,843	56.78%	11.67%	60.73%
9	4,004	77.44%	67.53%	69.68%

The most important insights observable in these results are as follows:

1. The model without meta-variables performed worse, in most cases, than the model with meta-variables. More detail can be found below as to the significance of these meta-variables, but they may demonstrate either the effect of system use on opportunity success, or the effect of opportunity success on system use.
2. There was a strong correlation between the size of a division's dataset and recall, which was more pronounced in the models without meta-variables. Smaller datasets provided fewer training samples, and thus less information for the model to learn from and make predictions with. In this case, smaller divisions also tended to be less balanced in terms of true wins vs. losses; for example, division 7 only won 16.7% of opportunities in our data, and divisions 1 and 2 both won under 40%. This imbalance further reduced the information available to the model for predicting wins.
3. Most models were biased toward "loss," reflecting two facts: first, that in making predictions, the threshold for a win was 55% certainty, in order to limit type 1 errors; and second, that true win percentages tended to be less than 50%, broadly. The exceptions to this would be

divisions 5 and 9, where win percentages exceeded 50%; these models were likewise biased toward "win."

Again, as a result of this analysis, the following variable importances were identified for each division and variable:

TABLE IV
VARIABLE IMPORTANCES BY DIVISION

Variable	Div. 1	Div. 2	Div. 3	Div. 4	Div. 5
Open Time	37.78	50.85	133.61	68.70	64.62
Fields Completed	50.24	23.69	91.76	57.88	57.95
Lasty Action Time	22.49	55.29	40.23	54.81	50.24
Industry	5.68	0.00	52.75	86.91	41.78
Type	1.65	22.59	23.82	76.68	65.75
Amount	13.98	27.24	45.70	47.63	46.09
Task Count	5.56	5.48	47.71	56.22	45.26
Complexity	21.68	15.62	31.45	12.37	41.89
Calid Open Time	15.29	19.85	11.92	15.75	23.20
Industry Code 2	5.83	0.00	53.71	6.12	23.14
Account Tier	6.34	11.30	20.83	26.78	25.92
Account Type	4.17	0.00	30.11	19.06	49.35
Enterprise Account	4.16	11.54	13.21	11.12	28.38
Customer Classification	4.82	0.00	24.83	10.43	29.13
Industry Code 1	6.21	0.00	18.12	4.50	13.45

TABLE IV
VARIABLE IMPORTANCES BY DIVISION (CONTINUED)

Variable	Div. 6	Div. 7	Div. 8	Div. 9	**Average**
Open Time	60.92	15.87	51.59	75.66	**62.18**
Fields Completed	90.50	12.14	122.17	39.25	**60.62**
Lasty Action Time	57.44	16.54	37.47	28.20	**40.30**
Industry	50.44	1.52	14.02	45.24	**33.15**
Type	56.31	6.91	10.68	23.57	**32.00**
Amount	45.31	-0.70	5.57	23.45	**28.25**
Task Count	39.53	13.94	16.59	18.71	**27.67**
Complexity	15.43	8.79	27.85	12.10	**20.80**
Calid Open Time	15.55	8.18	17.70	27.50	**17.21**
Industry Code 2	9.74	-0.35	5.28	43.14	**16.29**
Account Tier	22.02	9.46	6.06	16.21	**16.10**
Account Type	12.33	4.53	4.63	11.83	**15.11**
Enterprise Account	13.18	3.55	23.27	1.14	**12.17**
Customer Classification	8.89	2.21	4.51	7.50	**10.26**
Industry Code 1	8.32	1.84	6.26	15.97	**8.30**

These variables' importances were assessed in terms of the %IncMSE metric.

Open Time, on average, contributed 60% MSE when removed from the model, the highest average of any one variable. Upon inspection of the data itself, it became clear that longer open times generally led to a lower likelihood of winning. This was logical from two perspectives: first, longer open times likely reflected a less pressing need on the part of the customer; second, negative open times (of which there were many), implied that an opportunity was entered into the system after its close. This may have been the result of a migration from a legacy system to a new implementation, or it may have simply been a laxer approach to maintaining data fidelity. Regardless, if entering data post hoc, it seemed likely that opportunity owners would more readily remember those opportunities they won versus those they did not. In terms of usefulness to the company, this data could also be visualized to help identify which business units and teams need training and provide reminders to maintain the input of data.

Field Completed, on average, fell just beneath Open Time with a contribution of 58%. Again, there were two ways this could be interpreted: the first was that use of the SFDC system was facilitating success in an opportunity; the second was that success in an opportunity causes the opportunity owner to input more details into the system. It would seem that the second explanation was more likely than the first, but both are plausible. Causality, in this case, would be almost impossible to establish one way or the other without further investigation by the company itself.

Industry, in this case, was the most important non-"meta-variable" with an average percent increased MSE of 33, and denotes the industry in which a potential customer operates. This field was selected by opportunity owners from a list of 40 possible descriptors. This variable was tremendously helpful, both in terms of building a predictive model and in terms of learning for the customer. A better understanding of industry success can help inform investment, whether the company chooses to spend less time in less fruitful industries, or to spend more on training and hiring to shore up strategic areas of weakness.

Type—indicating whether an opportunity was net new, incremental, or a renewal—demonstrated an average increase of 32 percent. Intuitively, net new opportunities had lower win rates than incremental (or upsell) opportunities, which had lower win rates than renewals. It would appear logical that existing customers would be much more likely to renew than brand new customers would be to onboard a new vendor, based on the cost to each of doing so.

Amount indicated the size of an opportunity, or the amount the company could expect a customer to pay, and accounted for a 28% increase in MSE. As one might assume, larger opportunities are less likely to close than smaller opportunities.

Task Count was calculated by tallying the total number of tasks created and linked to an opportunity's parent account. This particular variable on average contributed 26% MSE. While this count was tallied at the account level, and could therefore include tasks undertaken for opportunities beyond those being analyzed, it nonetheless was indicative of broader engagement with a potential customer. Unsurprisingly, a larger task count was associated with a higher win rate.

Complexity referred to the complexity of a product being offered to a customer. This variable contributed 22% MSE, on average. The more complex products were associated with a lower win rate, in this case.

The other variables used exhibited less importance, and were excluded entirely from some models. These include:

- Valid Open Time, which indicated whether an opportunity's Open Time was positive or negative, in the form of a 0 for negative and 1 for positive.
- Industry Code 1 and 2. These were additional industry definitions under a separate categorization system than "Industry" mentioned above.

- Account Tier, which indicated the priority assigned to a particular customer by the company
- Account Type, which indicated whether a particular opportunity was limited to a single division or spanned multiple divisions within the company
- Enterprise Account, which indicated the size of the customer
- Customer Classification, which indicated the regional scope of an opportunity

While the full set of variables included in the importance table above was examined for each division initially, variables of low importance were then eliminated on a division-by-division basis to produce the most accurate model possible.

CONCLUSION

This research served as a first step in the development of a broader initiative for a Fortune 500 paper and packaging company to operationalize predictive modeling on sales success. As such, the challenges with any large company often include requiring the building of deep local knowledge of the data, in addition to corralling a large organization to assist with accurate data collection. 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.

Before these applications would be addressed however, data ops resources would be required to perform a number of critical tasks: continue building and tuning the model for better accuracy, establish a cadence around maintaining the models and incorporating new kinds of information, and connecting with the other business units to understand strategic priorities for operationalization.

REFERENCES

[1] Implisit (Sales Cloud by Salesforce.com). [Online]. Available: https://www.salesforce.com/blog/2014/08/infographic-7-powerful-predictors-closed-won-opportunity-gp.html

[2] Insight Squared. [Online]. Available: https://www.insightsquared.com/features/sales-forecasting/

[3] J. Yan, C. Zhang, H. Zha, et all, "On Machine Learning towards Predictive Sales Pipeline Analytics." *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1945-1951, 2015. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9444/9488 [Accessed: Mon. 24 Sept. 2018].

[4] M. Bohaneca, M.K. Borstnarb, M. Robnik-Sikonja, "Integration of machine learning insights into organizational learning: A case of B2B sales forecasting." *28th Bled eConference*, June 7-10, 2015. [Online]. Available: https://domino.fov.unimb.si/proceedings.nsf/Proceedings/B12ECF2381AB59EEC1257E5B004B39B7/$File/2_Bohanec.pdf [Accessed: Tue. 25 Sept. 2018].

[5] M. Bohaneca, M.K. Borstnarb, M Robnik-Sikonja, "Explaining machine learning models in sales predictions." *Expert Systems with Applications*, no. 71, pp. 416-428, 2017. [Online]. Available: http://lkm.fri.uni-lj.si/rmarko/papers/Bohanec17-ESwA-preprint.pdf [Accessed: Tue. 25 Sept. 2018].

AUTHOR INFORMATION

Michael Christison, MBA and Data Science Dual Master's Student '19, Darden School of Business, Data Science Institute, University of Virginia

Bochao Li (Karen), MBA and Data Science Dual Master's Student '19, Darden School of Business, Data Science Institute, University of Virginia

Stephen Mortensen, MBA and Data Science Dual Master's Student '19, Darden School of Business, Data Science Institute, University of Virginia

Ailun Zhu (Allyn), MBA and Data Science Dual Master's Student '19, Darden School of Business, Data Science Institute, University of Virginia

Rajkumar Venkatesan, Professor of Business Administration, Darden School of Business, University of Virginia