A Supervised Learning Approach to Assessing Accounts Receivable Risk in Small-to-Medium Enterprises

Jim Dewey, Calvin Ingram, and Reinaldo Sanchez-Arias
Department of Data Science & Business Analytics
Florida Polytechnic University
Lakeland, Florida, USA
jdewey@floridapoly.edu, calviningram2013@gmail.com, rsanchezarias@floridapoly.edu

Abstract

Information technology, e-commerce, and data analytics have drastically changed the way companies do business. Few data related practices are more important than assessing risk; for example, the risk inherent in a firm's accounts receivable (AR). While small-to-medium enterprises (SMEs) comprise a large portion of the global economy, they may be too small to devote extensive internal resources to risk analysis. At the same time, readily available risk related information may be ill suited to the needs of diverse SMEs. We explore the potential for readily available machine learning techniques, accessible to SMEs at modest cost, to assess the AR risk faced by a medium sized transportation brokerage, John J. Jerue Truck Broker. Of the models considered, linear discriminant analysis performed best when risk is modeled with three categories and decision trees performed best when there were two categories. Our results demonstrate the potential for SMEs to develop risk assessments to flexibly meet their own diverse needs.

Keywords
Risk Factors, Account Receivables, Supervised Learning, Data Mining

1. Introduction

According to the World Trade Organization (World Trade Organization 2016), small to medium enterprises (SMEs) (which employ between 10 and 250 people) and Micro firms (which employ 10 or fewer) account for approximately 95% of all enterprises and two-thirds of employment worldwide, and half of gross domestic product in developed countries. As for larger companies, the extension of credit to customers plays an important role in facilitating transactions for many SMEs. For example, transportation brokerage firms incur significant costs to procure carrier services well before collecting revenues from consignors.

Clearly, assessing and managing risk in accounts receivable (AR) is important for such firms. However, due to their small size and diverse nature, SMEs have limited resources to conduct risk analysis customized to their needs given their specific business, locale, and customer mix. They may purchase related measures from companies like Experian and Dun & Bradstreet; for example, measures intended to reflect the likelihood a company will face serious financial distress within a specified timeframe. The usefulness of such measures is limited for two reasons. First, they are only available at infrequently, typically annually or quarterly, while the conditions that influence AR risk for SMEs vary continuously. Second, such measures are not adapted to the SMEs particular blend of risk.

Traditional definitions of risk have focused on likelihood of failure from the perspectives of financial institutions (FIs). FIs are distinguished from other businesses by their use of lending as a primary function, rather than its use to facilitate transactions. Most definitions of failure refer to a ratio of debts to assets that is thought to indicate serious likelihood of nonpayment or financial distress (see Altman and Saunders (1997) for example). Data in this setting is often available on a quarterly or yearly basis and consists of information such as historic creditworthiness and current and historic earnings.

Following the 2008 financial crash, emphasis was placed on systemic risk. Systemic risk measures (SRMs) focus on failure as it relates both to individual institutions and the system in which they operate. Several different approaches to its measurement exist, see Danielsson et al. (2011) for an overview. Research in this area tends to focus on the failure of the financial system and its relationship to FIs. This type of assessment uses financial information to predict the probability of failure by grouping customers according to their similarity with customers known to have failed.

Outside of the risk of failure, many different definitions of risk are used depending on industry and context. Existing work directly related to AR focuses on key performance indicators (KPIs) such as days outstanding. Zeng et al. (2007) used a partial pruned decision tree model to predict the likelihood of a categorical AR delinquency variable (current, 31-60 days, 91+). Their dataset comprised four major firms including two Fortune 500 companies, not SMEs.

Wu et al. (2014) provide a decision support framework for assessing AR risk employing logistic regression to predict if an account will meet one of several negative collections criteria within five months. Their analysis is conducted using a database of bank and Credit Bureau information placing it more in the realm of FIs as opposed to SMEs. However, one of their findings is particularly relevant here. They find that as the population changes, the existing criteria and cutoffs tend to migrate and become less appropriate for modeling risk, demonstrating the need for an understanding of risk that adapts to changing business environments over time.

The differences between SMEs and FIs are well established (Altman and Sabato 2007). In addition to the difference in available data, the factors relevant for predicting failure also change (Edminster 1972). However, we know of no empirical work that focuses on assessing the risk faced by SMEs in a way that is adaptable to their varied needs, that may be updated frequently as circumstances change, and that is feasible for individual SMEs to implement.

Our purpose is to demonstrate the potential for SMEs to analyze risk within the context of their specific market and with respect to their own qualitative and quantitative understanding of risk using readily available machine learning tools which they could deploy at low incremental cost. To this end, we develop and test a method to track AR risk for John J. Jerue Truck Broker Inc. We show it is possible to develop a tailored approach to assessing AR risk with only modest resources. While even this may be beyond the capability of the smallest SMEs, many have both accounts receivable and information technology departments and employ both business analysts and information professionals, and so have on hand the resources to develop and deploy such tools.

2. Case and Data Description

John J. Jerue Truck Broker is a Third-Party Logistics Provider (3PL) whose primary contribution is matching customers shipping a product to carriers with equipment to move the goods. One feature of their business practice is same-day-pay for carriers. Any carrier who delivers a load and turns in all appropriate paperwork is paid the same day. This makes careful management of their AR portfolio integral to their success in the highly competitive industry. For our study, data for different customer characteristics was provided on a weekly basis. The dataset showed typical SME data issues such as missing data points and significant disparities between customers of different volumes.

Before any of the learning process had taken place customers below a predefined threshold were removed from consideration in the learning process. This threshold pertained to missing data points in certain relevant fields. Values were used later in a reduced rank version of the model so that these could still be classified. Classification of customers was originally developed on a categorical system with 3 representing customer who represent a financial risk, 2 representing customers who could represent an eventual financial risk, and 1 for customers without signs of serious financial issue. A binary case was also considered where customers originally classified as 3 were marked as 1 and all other customers were given 0. Several subsets of customers were selected to be used for initializing the model. These were presented to the experts at Jerue's AR department who ranked them into categories. It was explained and should be noted that this ranking is in no way meant to represent the true financial state of the customer. Rather, it represents the risk associated with Jerue's current association with that customer. A rank 3 customer is not necessarily likely to default or have some number of unpaid invoices older than 60 days; but, given the company's current standing they present some type of financial risk to the company. been able to directly voice opinions as opposed to a single decision maker who may or may not represent their associates in decisions of credit worthiness.

Table 1. Training and Validation Datasets

Rank	Training Data	Survey Test Data	First Week Validation
1	77	11	853
2	53	13	103
3	32	9	17
Total	162	33	973

The general distribution of the data is compiled in Table 1. The training data was collected only for customers above the specified limit and represents the survey response for customers. The survey test data was a subset of the training data (not included in the training data) that was used for initial model selection. The process of selected survey test sets from the training data took place many times in a process called k-fold cross validation. Finally, after an initial model was selected, it was used to generate a test report that was given as a survey to experts who reviewed the model's responses and made corrections as they felt necessary.

Once this training set was collected, it underwent many changes (called basis transformations in machine learning literature) that standardized business according to significant individual and companywide statistics. Once collected this dataset was divided randomly into three parts. One part (70% of the data) was used for fitting and training the model (and is referred to as the training set). The remaining 30% was used as a test set (one example is noted here as survey test data; however, the process of randomly selecting survey and training data was repeated many times). Once a model had been preliminarily selected, it was used to generate a full test report that was given back to AR experts who made corrections to the model. The result was used as a validation set to estimate the approximate errors of all different model types.

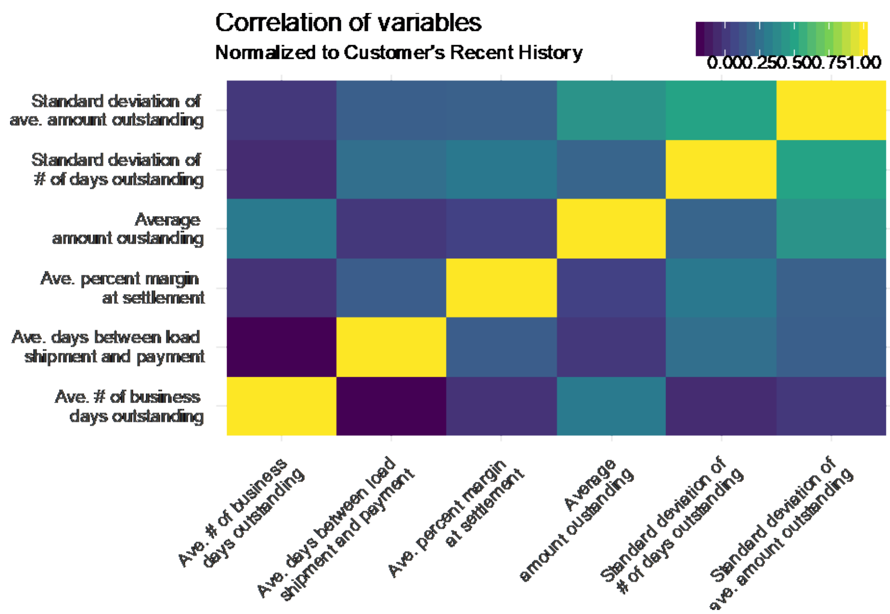

Figure 1. Correlation plot.

Raw data was collected in weekly intervals of relevant statistics (Figure 1). The information presented several key challenges likely common to SMEs in this setting. First is the prevalence of small customers. These are customers who do not represent enough internal information to make confident assertions as to the level of risk they present. As previously mentioned, these were removed for use in a reduced rank model and not considered in the larger model. Next is the extremely limited pool of rank 3 customers to train from. These represent less than 2% of the total data collected for this study. Removing the low volume customers changes this (since by their nature low volume customers tend to be low risk) and was adequate for training/testing of the categorization model. The binary case required additional changes that used advanced sampling techniques to achieve a more balanced distribution. The data was normalized both to company standards and to each company's recent history. This was done in order to develop a sense of continuity between customers in different periods, and the customers in relation to the overall AR portfolio. The exact nature of these normalization differed, but for the purpose of this study represents the maximum value either of the customer within the past 5 weeks, or the company for the given period.

Many of the variables were interrelated such as the amount and time outstanding. These correlations were considered in interpretation of the model but not directly used in conjunction with subset selection. (Figure 2)

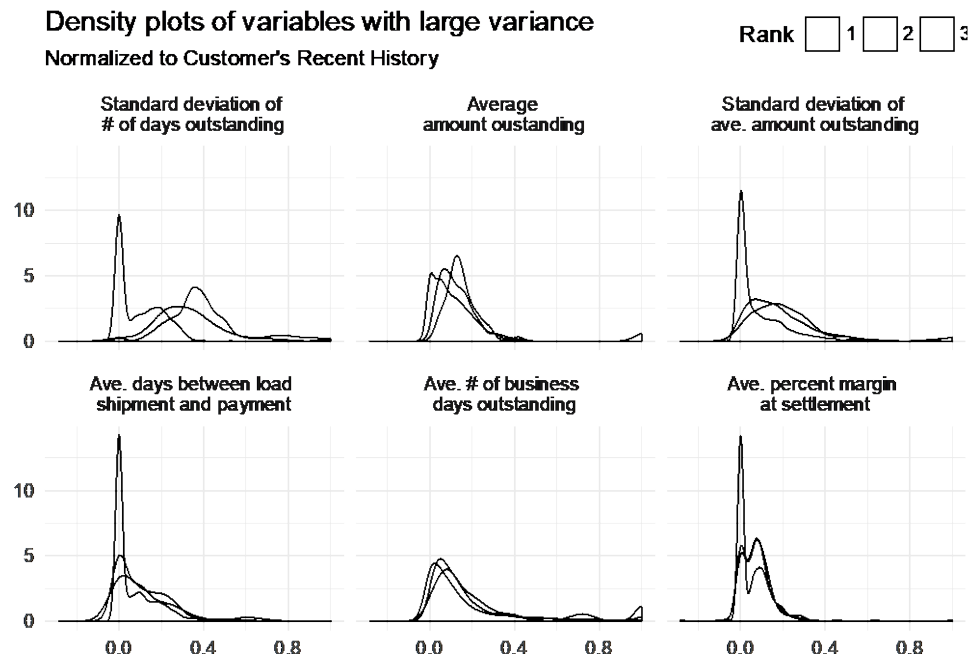

Figure 2. Density plots of features per rank

One of the key advantages to the survey method of model initialization was that it prevented a seniority bias and allowed for multiple stakeholders to have input on the definition of risk at the company. With this flexibility comes the opportunity for disagreement between stakeholders. Disagreement was settled by a majority of experts, although due to limited scope of the survey random selection compared to the available data-pool few instances of disagreement exist.

4. Analysis and Model Comparison

4.1 Subset and Model Selection

Model selection took place as an iterative process comprised of two steps: subset selection and model selection. Subset selection is the process of reducing the full dataset (consisting of all initial variables and their basis transformations) into a smaller but significant group of data features. The test dataset was used in every case for this selection process to avoid any 'unfair' predisposition of the model towards arbitrary groupings of specific variables such as might occur if a user could select variables that fit best for a testing iteration. Subset selection can be done in many ways that range from seemingly arbitrary to state-of-the-art, but perhaps most important to the specific goal of this model was the expert input on what variables were most important to them (Figure 3). Although few of the direct suggestions made at this point made it into the final model, the general categories of variables they selected were helpful in reducing the time it took to select the different models.

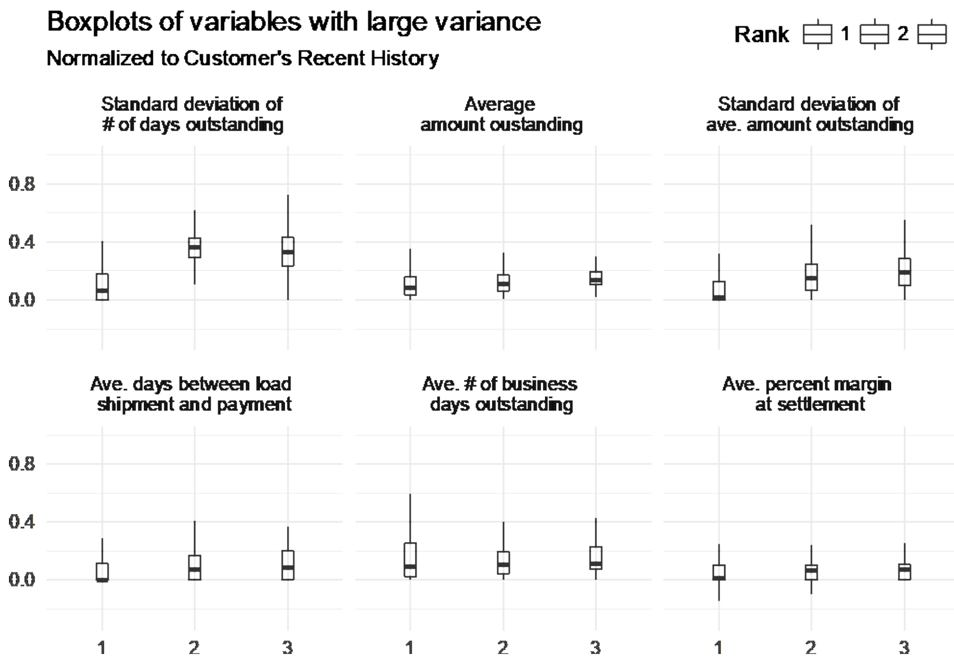

Figure 3. Boxplots of features per rank

Other methods of subset selection included forward and backward stepwise subset selection, principal component analysis (PCA) by Kaiser-Meyer-Olkin (KMO) maximization, and analysis of variance (ANOVA). Each of these methods provided different subsets with different tests groups. After these methods had been used over many different random subsets a consensus was taken and priority given to variables that appeared most often. These variables then went through further basis transformations that looked at possible relevant ratios between them. The ratio process followed earlier work by Edmister (Edmister 1972) which looked specifically at the important of certain financial ratios in SME risk assessment.

Once a subset had been selected several different models were tested. An interesting conclusion from this stage of study was that the more stringent assumptions placed on the data in discriminant analysis tended to be helpful in improving the results of predictions. The more flexible models tended to overfit to the test information, capturing noise rather than true trends in the data.

4.2 Three-Tier Case

Three model types are discussed in the context of a three-tier risk classification model. The first, called discriminant analysis, seeks to find some division in the criteria between customers who present high risk and those who do not. This division is in some way predefined by constraints on its equation type such as a linear relationship (linear discriminant analysis or LDA), quadratic relationship (quadratic discriminant analysis or QDA), or a logistic relationship. Second, the least absolute shrinkage and selection operator (LASSO) is a more flexible linear method that penalizes the size of terms thereby 'shrinking' any extraneous variables toward zero. Finally, a Random Forest is considered which looks for divisions in the predictor variables that provide the best prediction of the response variable. An extension of this work then considers a binary classification in which an additionally decision tree model is considered.

All models were built using the R statistical programming language (R Core Team 2017) and the `caret` package (Kuhn 2018) (short for `C`lassification `A`nd `RE`gression `T`raining) which includes a set of functions that attempt to streamline the process for creating predictive models. We use this package here for pre-processing, model building, and performance metrics of our proposed predictive models. Full documentation can be found at http://topepo.github.io/caret/index.html

The selected model based on the validation errors was the linear discriminant analysis (LDA). Although the logistic model performs similarly, LDA was given preference because of its structure. Logit models consider more predominantly values lying on the division between classes. Given the somewhat subjective nature of survey initialization in this context, this feature could overfit to specific datapoints and miss some of the underlying trend. A mathematical justification for the merit of LDA in a risk setting is well laid out in (Fantazzini and Figini 2009). Since the test dataset is not balanced, we report the balanced accuracy for each class (Figure 4), defined as the average of sensitivity and specificity (Brodersen et al. 2010). Table 2 shows the dominance of LDA in this study at all risk levels, followed consistently by the Random Forest method.

Table 2. Comparison of methods for multi-class problem in testing dataset

	Class: Low			**Class: Medium**			**Class: High**		
Method	Sensitivity	Specificity	Bal. Acc.	Sensitivity	Specificity	Bal. Acc.	Sensitivity	Specificity	Bal. Acc.
LASSO regularized	0.9915	0.6638	0.8276	0.6782	0.9847	0.8314	0.3448	0.9978	0.6713
Random Forest	0.9963	0.6638	0.8301	0.6897	0.9906	0.8401	0.3793	0.9989	0.6891
LDA	0.9890	0.7241	0.8566	0.7126	0.9870	0.8498	0.4482	0.9923	0.7203

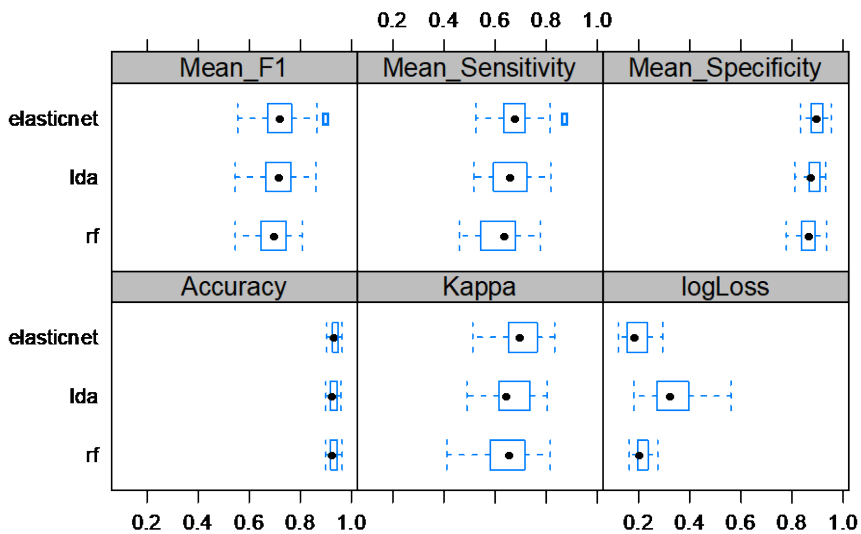

Figure 4. Comparison of methods.

4.3 Variable Importance

This done, a similar process was conducted for lower-volume customers, albeit with a markedly simpler model. Several key assumptions that went into the selection of this subset and model included that the low-volume customers tended to be inherently lower risk. Overall, they did not tend to include many rank 3 customers except in certain well-defined situations (such as might directly suggest likely nonpayment). The final reduced-rank model was built with a logical structure of the form:

$$C_i \in \quad R_3 \text{ iff } (\text{condition}_1 \text{ or condition}_2 \text{ or} \ldots \text{condition}_N)$$

where C_i is the customer with limited data, R_3 is the class of rank 3 customers, and C_i is given a rank of 3 if and only if it meets one of the specific conditions of a rank 3 customer (table 3 and figure 5).

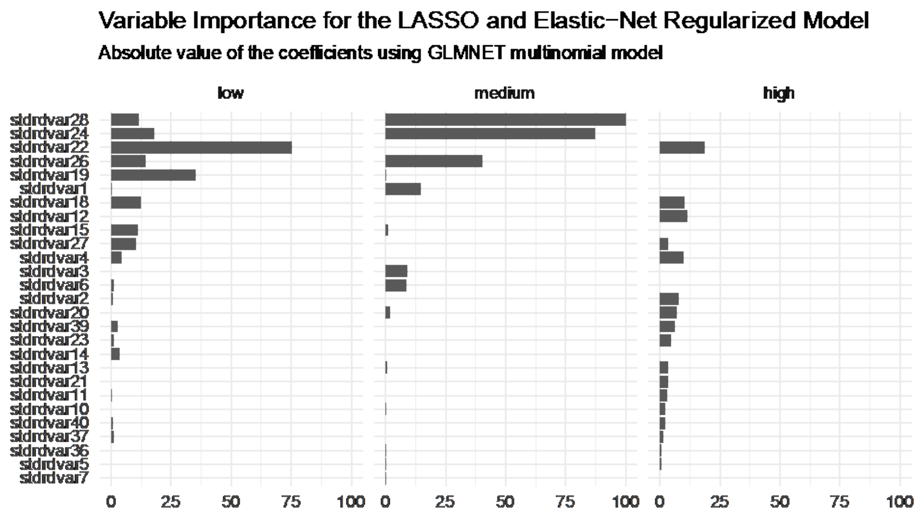

Figure 5. Variable importance plot.

Table 3. Confusion matrix for LASSO, Random Forest, and Linear Discriminant Analysis (LDA)

		LASSO			Random Forest			LDA		
		Reference			Reference			Reference		
		Low	Medium	High	Low	Medium	High	Low	Medium	High
Prediction	Low	812	26	13	816	26	13	810	21	11
	Medium	7	59	6	3	60	5	6	62	5
	High	0	2	10	0	1	11	1	4	13

4.4 Binary Case

We implemented a binary classifier with the same dataset, this time grouping customers ranked as 1 or 2, in category not-at-risk, and the remaining customers ranked 3 in category at-risk. To tackle the issue of imbalance classes, besides over and under-sampling, there are hybrid methods that combine under-sampling with the generation of additional data. One popular technique known as ROSE (Random Over-Sampling Examples) introduce in (Menardi and Torelli 2014) is used here. The `ROSE` R package (Lunardon, Menardi, and Torelli 2014) provides functions to deal with binary classification problems in the presence of imbalanced classes. Artificial balanced samples are generated according to a smoothed bootstrap approach that allows for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class. Removal of near-zero variance predictors and standardization was implemented as part of the pre-processing stage.

We fit a Random Forest using the `ranger` package (Wright and Ziegler 2017), which is a fast C++ implementation of the original algorithm in R. The tuning parameter used in our implementation, is the number of randomly selected predictors at each cut in the tree. Using the implementation followed by the `caret` package, when using the `rpart` method, the complexity parameter is used to automatically prune the tree (Table 4).

Table 4. Binary classification results

Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
Decision Trees	0.9701	0.5517	0.9835	0.7676	0.5180
Random Forest	0.9744	0.2413	0.9977	0.6195	0.3590
Linear Discriminant Analysis	0.9669	0.3448	0.9867	0.6658	0.3755

The kappa value is a metric that compares an observed accuracy with an expected accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. It helps to have an indicator of how much better the classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Following the interpretation from (Landis and Koch 1977), kappa values between 0.21 and 0.40 indicate a fair agreement, while values between 0.41 and 0.60 indicate moderate agreement. In the binary classification framework, the decision trees approach produced better results compared with the random forest and linear discriminant approaches (Table 5 and figure 6).

Table 5. Confusion matrix for Decision Trees, Random Forest, and Linear Discriminant Analysis (LDA)

		LASSO		Random Forest		LDA	
		Reference		Reference		Reference	
		At risk	Not at risk	At risk	Not at risk	At risk	Not at risk
Prediction	At risk	16	15	7	2	10	12
	Not at risk	13	892	22	905	19	895

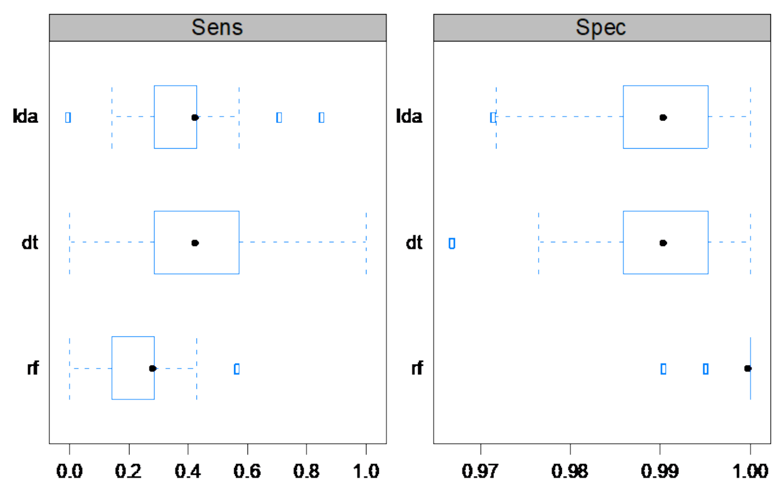

Figure 6. Binary classification results.

Regularized LASSO, using the `GLMNET` package (LASSO and Elastic-Net Regularized Generalized Linear Models), performed similarly with slightly better results for sensitivity compared to the random forest approach. A ROC metric (the area under the receiver operating characteristic curve) was used to select the optimal model using the largest value (figure 7). It represents a model's ability to discriminate between positive and negative classes. The final optimal values used for the model were $\alpha = 0$ and $\lambda = 10^{-4}$.

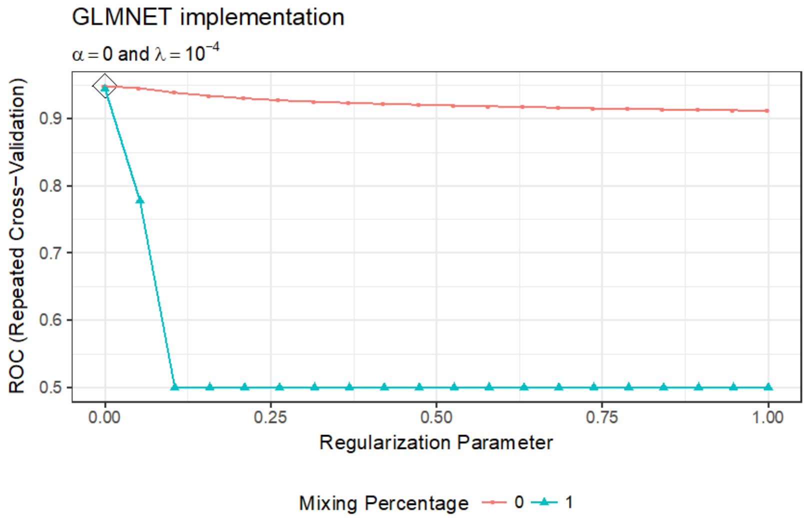

Figure 7. Cross-validation results for parameter tuning (Regularized LASSO)

5. Conclusion

The final model was implemented in 2018 by the AR department at John J. Jerue Trucking. It is used to provide weekly updates on customers needing additional attention for collections purposes. The model is refreshed with new data based on corrections from AR experts over time and occasional resurveys. This prevents the model from stagnating and allows the company to adapt (nearly) continuously to both internal and external changes.

Approaches to tailored risk modeling like the one presented have the potential to increase efficiency and profitability by focusing limited resources in an AR department where they are most needed. Although the subset and model selection used here should not be generalized to other businesses, the concept of tailored SME risk assessment can fill the gap between SMEs and large enterprises and FIs in a way that is suited to the diverse needs of SMEs.

Despite its success, our approach has limitations. First, all information driving the model comes from within the company itself, posing an inherent risk of customer misclassification or potential bias. Future iterations might be improved by including other measures specific to the business, such as their credit score, obtained from other sources. This initial implementation also lacks additional functions that might be useful. For example, later designs might include automatically setting credit limits, credit blocking, and other functions where a speedy response might be crucial to prevent a negative outcome.

A limited range of models was tested given the specific needs and situation of our particular case, pointing to opportunities for future work. Crook, Edelman, and Thomas (2007) provide a survey of major risk-of-failure studies that finds neural networks (NNs) perform best in a number of cases. However, we did not consider such models because the end users at John J. Jerue Truck Broker require an explanation of risk scores that can be impossible to discern from NNs due to their black box nature. However, NNs may prove useful in other SME contexts. Tsai (2008) models supply chain cash flow risks employing a dynamic stochastic general equilibrium (DSGE) simulation approach to modeling demand distribution, production, inventory, and sales. This approach might also prove useful in some SME settings. Michalski (2008) proposed a portfolio management approach to tracking and predicting operational risk that focuses on groups of customers and not individual customers, lending itself to optimizing a portfolio of AR as one might any group of financial assets. Development of such an approach would likely be useful in many SME contexts, including ours.

References

Altman, Edward I, and Gabriele Sabato. Modelling Credit Risk for SMEs: Evidence from the US Market. Abacus 43 (3): 332–57. 2007.

Altman, Edward I, and Anthony Saunders. Credit Risk Measurement: Developments over the Last 20 Years. Journal of Banking & Finance, vol. 21, no. (11-12), pp. 1721–42. 1997.

Brodersen, K. H., C. S. Ong, K. E. Stephan, and J. M. Buhmann. The Balanced Accuracy and Its Posterior Distribution. In 2010 20th International Conference on Pattern Recognition, 3121–24. 2010.https://doi.org/10.1109/ICPR.2010.764.

Crook, Jonathan N, David B Edelman, and Lyn C Thomas. Recent Developments in Consumer Credit Risk Assessment. European Journal of Operational Research vol. 183, no.3, pp. 1447–65. 2007.

Danielsson, Jon, Kevin R James, Marcela Valenzuela, and Ilknur Zer. Model Risk of Systemic Risk Models. London School of Economics (Mimeo). 2011.

Edminster, RO. An Empirical Test of Financial Ratio Analysis for Small Business Failure. Journal of Financial and Quantitative Analysis. 1972.

Edmister, Robert O. An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. Journal of Financial and Quantitative Analysis , vol. 7 ,no.2, pp. 1477–93, 1972.

Fantazzini, Dean, and Silvia Figini. Random Survival Forests Models for SME Credit Risk Measurement. Methodology and Computing in Applied Probability, vol. 11, no.1, pp. 29–45, 2009.

Hendricks, Darryll. Evaluation of Value-at-Risk Models Using Historical Data. 1996.

Kuhn, Max. Caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret. 2018.

Landis, J Richard, and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. Biometrics, 159–74. 1977.

Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. ROSE: A Package for Binary Imbalanced Learning. R Journal, vol. 6,no.1, 2014.

Menardi, Giovanna, and Nicola Torelli. Training and Assessing Classification Rules with Imbalanced Data. Data Mining and Knowledge Discovery, vol. 28, no.1, pp. 92–122, 2014.

Michalski, Grzegorz. Operational Risk in Current Assets Investment Decisions: Portfolio Management Approach in Accounts Receivable (Agro Econ-Czech: Operační Risk v Rozhodování o běžných Aktivech: Management Portfolia Pohledávek). 2008.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.2017.

Tsai, Chih-Yang. On Supply Chain Cash Flow Risks. Decis. Support Syst. 44 (4): 1031–42. https://doi.org/10.1016/j.dss.2007.12.006. 2008.

World Trade Organization. World Trade Report 2016: Levelling the Field for SMEs. Available at https://www.wto.org/english/res_e/booksp_e/world_trade_report16_e.pdf, Accessed on April 25, 2022. 2016.

Wright, Marvin N., and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77 (1): 1–17. 2017.https://doi.org/10.18637/jss.v077.i01.

Wu, Desheng D, David L Olson, and Cuicui Luo A Decision Support Approach for Accounts Receivable Risk Management. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 44, no.12, pp. 1624–32. . 2014.

Zeng, Sai, Ioana Boier-Martin, Prem Melville, Conrad Murphy, and Christian A Lang Predictive Modeling for Collections of Accounts Receivable. In Proceedings of the 2007 International Workshop on Domain Driven Data Mining, 43–48., 2007.

Acknowledgements

The authors thank Jerue Logistics Solutions and Florida Polytechnic University for supporting this collaborative effort.

Biographies

James Dewey is an assistant professor in the Department of Data Science and Business Analytics and Director of Economic Analysis at Florida Polytechnic University. He has a Ph.D. in economics from the University of Florida. His master's degree is in Economics from the University of South Florida, where he also received his bachelor's degree in Economics and Political Science. His research interests are in the areas of applied microeconomics, policy analysis and evaluation, education finance, labor economics, and urban and regional economics. He is a member of the Regional Science Association International and the National Education Finance Academy.

Calvin Ingram is a Marketing Data Analyst for Ramsey Solutions. He has a B.S. in Mechanical Engineering from Florida Polytechnic University. He was a Data Scientist at John J. Jerue Truck Broker when the bulk of the analysis described herein was conducted and he continues to consult with them on projects of mutual interest.

Reinaldo (Rei) Sanchez-Arias is an assistant professor in the Department of Data Science and Business Analytics at Florida Polytechnic University. His research interests include data mining and statistical learning, numerical optimization, operations research, and data science education. Sanchez-Arias received a Ph.D. in Computational Science from The University of Texas at El Paso, and a B.S in Mathematics from Universidad del Valle in Cali, Colombia. He is a member of SIAM, INFORMS and IEEE.