Classification Methods for Predictive Lead-Scoring

Deeksha Doddahonnaiah, Sahil Gandhi, Donald Hamnett

February 27, 2018

Problem Description

An issue that often arises for companies in the retail and financial service industries is qualifying leads, information on potential customers. Companies receive leads via several channels, whether it be through marketing campaigns, cold-calling, or direct customer inquiries. One main problem that can arise is due to conflicts of interest, where those who are compensated for lead generation make as many leads as possible, regardless of the lead quality. Other factors include whether or not marketing campaigns reach the target audience, and whether customer inquiries are from those capable of making timely payments. This becomes a serious issue for the company providing the service, because a significant amount of time and resources go into processing each lead. When lead processing is a bottleneck in the company's pipeline, qualifying these leads efficiently has a direct effect on revenues, due to both direct expenditures and opportunity cost.

We see lead prioritization as a two-fold issue in relation to the customer relationship timeline: credit scoring and predictive lead scoring. The company will first decide which leads to act on initially, then qualify these leads by obtaining credit and other financial data. Credit scoring is an area where there has been considerable research and success[1], due to the convenience of having hard financial data off of which to make predictions. While credit-scoring is certainly a valuable metric, lead-scoring offers a higher potential to develop a novel model and discover underlying features not detectible in typical analysis. For these reasons, predictive lead-scoring, defined as the process of using limited data to prioritize the pursuit of potential customers, will be the focus of our project. Unlike credit scoring, this area has less publicly available research, and is a service offered as a proprietary solution by several CRM firms. The aim of this project is to assist a real-world company to develop a predictive lead-scoring model. The inputs will be data collected during the phases of customer processing, and the outputs will be classification of the potential customer into one of five categories, two tiers each for profitable and non-profitable, and a neutral class.

Data

For this project, we will use an dataset that includes anonymized information of 4,500 business loan applications. The dataset includes all the information that a normal loan application would need; including, but not limited to the company and individual demographic data, corporate bank transactions, credit information data, and detailed loan-lifecycle data.

Predictive lead scoring, customer churn prediction and other predictive metrics of a customer's behavior are extensively studied in various industries, especially the telecommunication industry. The majority of predictive research in the financial industry has been focused on credit scoring and finding the credit worthiness of an individual or a company. This project focuses on researching methods to optimize the resources a company spends on qualifying leads.

The dataset is a covers a breadth of raw data from various sub-domains (individual data, financial data, loan lifecycle data). Using this, we can distinctively select features that represent a certain behavioral or financial characteristic of the client. In addition to this, we also plan leverage existing credit-scoring research to generate uncorrelated features from the raw transactional and financial data.

Algorithms

Artificial Neural Networks

We plan to use a multi-layer, feed-forward artificial neural network (ANN) using back-propagation as described in the course texts[2, 3]. Neural networks are typically used when it is unclear what the key features are in the prediction model[4]. The choice of this method is based off of its ability to isolate hidden features through the training process, without explicitly selecting them. Depending on our results with an initial implementation, we may experiment with different depths, numbers of perceptrons in each level, and with unsupervised pre-training of the network. ANNs have well-documented used in credit scoring applications, and we believe that this success will translate into similar prediction accuracy for predictive-lead scoring. Though details are typically not publicly available, several proprietary solutions have reported positive results in applying ANNs to predictive-lead scoring[5], and we hope to replicate these results. Though we have not yet made a final decision on which Python framework to use, our current choice is the Keras Deep Learning Library.

Support Vector Machines

SVM is a machine learning method raised by Vapnik in the early 1990s, which arises from optimal linearly separable SVM classification surface. Optimal classification surface requires the separating line can not only separate two dimensions correctly (the training error rate is 0), but can also maximize the margin between the two classes. SVM aims to find a hyperplane which can meet classification requirements and make the trained points far from the classification surface, namely looking for a classification face to make the margin on its both sides maximum. The SVM model is fairly flexible and allows us to work with linear and non-linear. The models is usually used as a classifier and we will also be using it to classify the leads. A lot of research is done in using SVMs for lead predictive analysis[6]. Certain variations of SVMs in combination with Genetic Algorithms or Recurrent Neural Networks have been proposed.

Random Forests

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. The random forests method, introduced by Breiman [7], adds an additional layer of randomness to bootstrap aggregating (bagging) and is found to perform very well compared to many other classifiers. Random forests are seen to be robust to over-fitting. RFs de-correlate the decision trees in the forest via randomization of split attributes that leads to an improvement over traditionally bagged trees and reduces the variance when averaged over the trees [7].

Selection of the attributes at each level of the split, and choosing when to stop splitting, i.e., the depth of each tree in the forest are two main critical choices for an RF to perform optimally. Considering the case when the data is unbalanced, the standard random forests do not work well on such datasets. There are two ways to handle the imbalanced data classification problem of random forests. The first, weighted random forests, is based on cost-sensitive learning; and the second, balanced random forests, is based on a sampling technique. Both methods improve the prediction accuracy of the minority class and perform better than the existing algorithms. We plan to experiment the data with a couple variations to the standard random forests and choose the best performing.

Results

The results would illustrate the predicted priority levels on prospects, achieved using the three different models, Neural Networks, SVM and Random Forests, each, on two levels of lead processing. In the first phase of lead processing, with limited data, mainly on demographics, we are assigning each prospect a priority score ranging from 1-5, 1 being a weak lead, and 5 being a strong one. The priority scores indicate the likelihood of the prospect being a strong lead. Based on the priority given in this phase, we then filter out prospects which are more promising and re-prioritize them in the next level of application processing, which is credit scoring. In this phase of processing, with more concrete data, we will be able to prioritize the leads with less false positives.

Evaluation Metrics

We have a variety of options to evaluate our models, two of the most effective being, Root Mean Squared Error (RMSE), for regression models and Receiver operating characteristics(ROC) for classification models. ROC is seen to be a good metric for evaluation of such models as it produces a form of cost-benefit analysis[8]. The ROC curve is the plot between sensitivity and (1- specificity). The Area Under Receiver operating characteristics Curve(AUROC) being the ratio under the curve to total area, provides a good metric to evaluate TP and FP. A good model will produce an AUROC value > 0.5.

References

[1] Lyn C. Thomas. "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers". In: *International Journal of Forecasting* 16.2 (2000), 149?172. DOI: 10.1016/s0169-2070(00)00034-0. URL: https://www.sciencedirect.com/science/article/pii/S0169207000000340.

[2] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[4] *Deep Neural Networks: When, and When Not, to Use*. July 2017. URL: https://www.enterprisetech.com/2017/07/10/deep-neural-networks-not-use/.

[5] Brock. *Predicting Customer Behavior: How we use Machine Learning to Identify Paying Customers before they Subscribe*. URL: https://www.strong.io/blog/predicting-customer-behavior-machine-learning-to-identify-paying-customers.

[6] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. "Credit scoring with a data mining approach based on support vector machines". In: *Expert Systems with Applications* 33.4 (2007), 847?856. DOI: 10.1016/j.eswa.2006.07.007.

[7] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[8] Roger M. Stein. "The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing". In: *Journal of Banking & Finance* 29.5 (2005), 1213?1236. DOI: 10.1016/j.jbankfin.2004.04.008.