Proceedings of Machine Learning Research 71:121–125, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

Sleuthing for adverse outcomes: Using anomaly detection to identify unusual behaviors of third-party agents

Michelle R. Miller MICHELLE.MILLER2@CAPITALONE.COM
and **Robert Cezeaux** ROBERT.CEZEAUX@CAPITALONE.COM
8000 Dominion Pkwy, Plano, TX 75024

Abstract

Business transactions between customers and financing entities are often governed by intermediary agents. In this scenario, actions taken by these agents can affect the likelihood of adverse outcomes for both the customers and the financial institution. Our goal is to establish a general framework that identifies these types of anomalous agents. In this paper, we demonstrate a novel application of anomaly detection using isolation forests to identify which agents may be associated with adverse outcomes. We apply a genetic algorithm to understand which features were key to the performance of anomaly detection and and suggest a general framework for problems that similarly concern the behaviors of third-party agents.

Keywords: Anomaly detection, isolation forest, genetic algorithm

1. Introduction

Predicting a customer's likely outcome state is fundamental to many operations within financial institutions. Decisions regarding customership frequently rely upon predictions based only on the behaviors or attributes of the individual customer. However, like many industries, financial institutions also engage with intermediary third parties to manage different facets of their relationships with customers. This is common, for example, in the mortgage industry (mortgage brokers), in banking (tellers), or in investing (investment brokers).

Third party agents introduce an additional layer of complexity to interactions between customer and financial institution, as shown in Figure 1. It is known that actions taken by these agents can affect the likelihood of adverse outcomes for both the customers and the business. However, the incidence rate of adverse outcomes is typically low and their nature diverse. We seek to identify these agents as anomalies to typical third-party agent patterns of behaviors, and consequently investigate the use of anomaly detection algorithms (Chandola et al., 2009; Emmott et al., 2013, 2015). Since we address a large and noisy feature space, and since some third party agents who cause adverse outcomes may themselves cluster within our feature space, we apply isolation forests to approach this problem (Liu et al., 2008, 2012; Guha et al., 2016).

In this paper, we explore the use of anomaly detection techniques to establish a general framework for identifying agents whose actions correlate with adverse outcomes. We demonstrate the importance of feature engineering and selection in three different time samples of historical customer outcomes. Even with a large required feature set, isolation

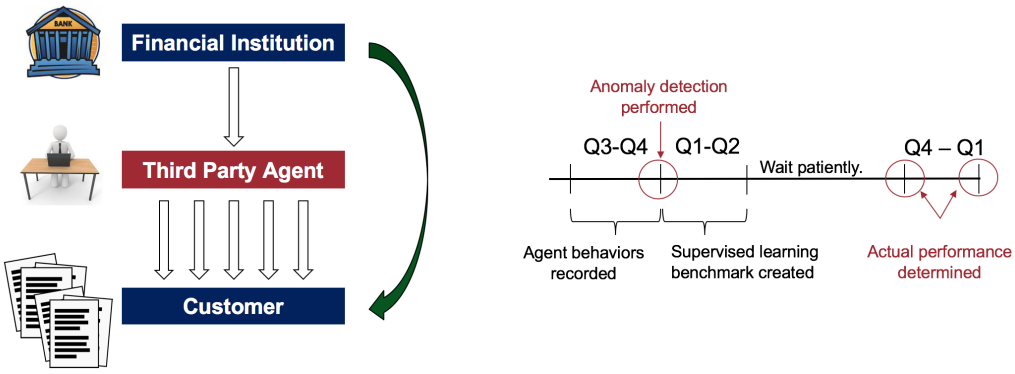

Figure 1: (Left) A schematic diagram of the relationship between financial institution, third party agent, and customer. This diagram shows how third party agents mediate the relationship between FI and customer, introducing complexities that may not be captured by decisioning systems that exclusively consider behaviors of the customer (green arrow). (Right) A time line describing the collection of data and benchmarking of customer performance.

forests are shown to successfully distinguish agents who yield large numbers of unexpected adverse outcomes.

2. Methods

In this analysis, we consider a scenario in which an industry-standard supervised learning model exists to predict the outcome of a customer relationship. As seen in Figure 1 (a), we assume that this model exclusively considers properties of the customer. However, we consider that third-party agent behaviors may also influence outcomes in a way that a predictive customer model will not anticipate. Our goal is to use anomaly detection to find agents whose unusual behaviors result in higher numbers of adverse outcomes than predicted. To accomplish this, we follow the process outlined in Figure 1 (b). Over the course of two quarters, we record a sample set of customers and agent activity. We perform anomaly detection on this sample and rank approximately 12000 agents according to their anomalousness. Over the next two quarters, we record the predicted outcomes our customers. After time elapses, we record actual customer outcomes. We then assess the success of anomaly detection by computing the number of unexpected outcomes as the difference between actual and predicted outcomes for each agent, and judge whether this quantity correlates with the anomalousness of the agent. We repeat this analysis for two additional time periods of customer interactions for out-of-time verification.

We consider two different feature engineering strategies in this application of anomaly detection. As a first approach, we use information about customers to perform anomaly detection. The variables used comprise both categorical and numerical values that describe the relevant characteristics of the customer for the business decision being made. We then

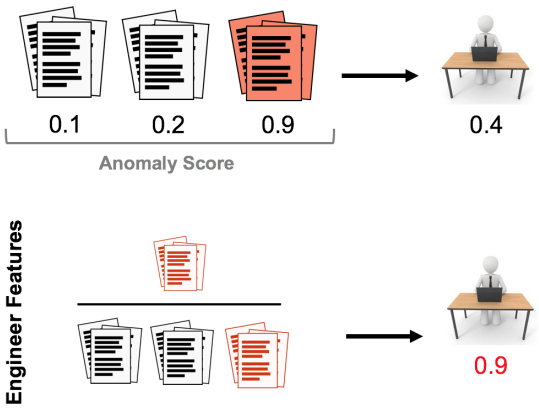

Figure 2: Two different methods of engineering features for the detection of anomalous third-party agents are shown. (Top) Individual customer records are scored separately, and agents are ranked by the average anomaly score of their customers. (Bottom) A large number of features are generated by computing the fraction of times agents engage in behaviors hypothesized to relate to adverse outcomes. Agents are scored directly using anomaly detection and rank-ordered accordingly. Both methods of feature engineering result in different agent orderings.

score agents by averaging the anomaly scores of the customers belonging to each agent. This approach seeks to flag agents with disproportionate numbers of customers who are reported with unusual characteristics. In a second approach, we consult with a subject matter expert to form a long list of agent behaviors that are suspected to correlate with adverse outcomes. For each agent, we count the number of times one of these behaviors occurred and divide it by the number of total customers interacted with. Each feature is numerical, representing the fractional incidence of each behavior. Both of these scenarios are depicted in Figure 2.

In each case, we are trying to detect anomalies using approximately 100 features. Many of these features may not be relevant for detecting unexpected numbers of adverse outcomes, but we have no way of discerning this prior to performing anomaly detection. Since we have a large number of features that are possibly unrelated to the anomalous event of interest, we use isolation forests (Liu et al., 2008) and robust random cut forests (Guha et al., 2016) to perform anomaly detection. These algorithms are selected over other methods due to their known good performance when dealing with large feature sets. We also want to understand whether there is a subset of all features that suffice to detect anomalous agents. To do this, we use a genetic algorithm to optimize the selection of subset of features that capture the largest number of unexpected adverse outcomes among the top 10% of agents. We use an algorithm that allows for mutations and cross-overs when creating the next generation, and run the algorithm through 60 generations of individuals until we achieve stable results for at least 10 generations. Repeating this implementation of the algorithm with different seed

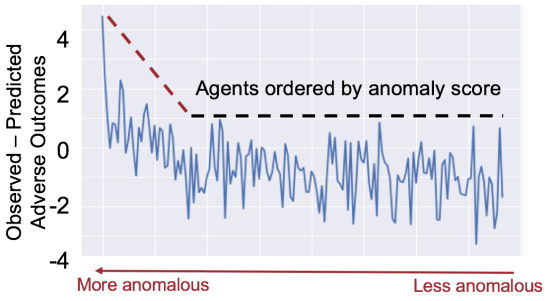

Figure 3: Typical results for anomaly detection on third party agent behaviors are shown. In this figure, agents are ordered by decreasing anomaly score and binned to reduce noise. We plot the number of unexpected adverse outcomes according to this ordering. Red and black dashed trend lines indicate two distinct regimes of behavior: an initial steeply-sloped interval demonstrating a strong relationship between anomaly score and unexpected adverse outcome, and a flat interval showing normal population behavior.

populations shows generally convergent results, which is sufficient to gain insight into which features are persistently important for detecting the largest number of adverse outcomes.

3. Results

A key challenge of using anomaly detection to address a business problem is producing actionable results. Our first feature engineering strategy demonstrates this problem. Detecting anomalies within customer behaviors (the first strategy shown in Figure 2) successfully found customers who interacted with the business under unusual circumstances. However, these interactions did not correlate with adverse outcomes for the customer. Consequently, averaging customer anomalousness to score third-party agents could not be used as a business-actionable metric.

On the other hand, anomaly detection within third-party agent behaviors (bottom of figure 2) was highly successful at flagging agents who caused unexpected outcomes. The effectiveness of anomaly detection is seen in Figure 3, in which agents are ordered by anomaly score and the number of adverse outcomes are plotted. This graph shows two distinct trends. First, as indicated by the red dashed line, we see a steeply sloped trend relating increasing numbers of unexpected outcomes with increasing agent anomalousness. Second, as marked by the black dashed line, we observe a regime in which anomaly score has no association with the rate of unexpected adverse outcomes. We attribute this second behavior to the presence of a 'normal' population whose outcomes result through a random and homogeneous mechanism. In contrast, the sloping of the initial trend is indicative of the differentiating power of anomaly detection. Even without the confirmation of a predictive model, as would be the case if anomaly detection was enacted in a real-time framework,

the consistency of the relationship between anomaly score and outcome in the first regime points to an actionable business strategy for managing the complexities of third-party agent interactions.

Following the success of this anomaly detection application, we considered whether the full 100 member set of behavioral features was necessary to achieve good performance. We used a genetic algorithm to permute the membership of smaller subset of features and optimized the feature set to flag the largest number of adverse outcomes as anomalous. When we restricted the subset membership to fifteen or fewer elements, the genetic algorithm was found to prioritize a common set of seven features between at least two of the three considered time lines, indicating that a core set of behaviors commonly correlates with anomalousness. However, we note that using the full feature set to detect anomalies performs nearly as well, and presents less risk of over fitting the results of a particular time line.

4. Conclusions

Isolation forest models constructed on features that characterize agent behaviors can successfully identify agents who are correlated with adverse outcomes. The problem we have explored here generalizes to many related scenarios in the financial sector and beyond. Intermediary parties who interact with customers can influence outcomes in a way that is difficult to capture through traditional supervised learning techniques. As we have seen here, appropriately engineered features enable anomaly detection to act as an exceptional solution to this problem. (Guha et al., 2016)

References

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*, 2015.

Andrew F Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 16–21. ACM, 2013.

Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In *International Conference on Machine Learning*, pages 2712–2721, 2016.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012.