Implications of Data Sources and Variables in Predicting Fraud at Application

Michiko I. Wolcott
Technology and Analytical Services
Equifax

Predictive modeling for fraud has received much attention in the recent years. While it can present challenges that are fundamentally different from predictive modeling for credit risk, discussions about fraud prediction tend to focus on the methodological aspects of modeling. There are key pragmatic questions in building analytical fraud solutions, such as value of data and variables, implications and interpretations of modeling results, design challenges specific to fraud modeling, among others. While the specific interpretation of the resulting models and the legal framework under which fraud models and the associated data can be applied vary widely from country to country, the fundamental ideas can be leveraged across markets.

Past studies by Equifax in application fraud modeling have led to some practical observations. First, the definitions of fraud and of fraud types may not always be consistent from lender to lender which can present a challenge. For example, substantial portion of charge-offs are sometimes intuitively categorized as fraud without thorough investigation and confirmation that a fraud was actually perpetuated, especially if no payments were ever made against the obligation. Lack of proper tools and processes for fraud mitigation often leads to equating the inability to verify with perpetuation of fraud, even when a vast majority of these cases are not truly fraudulent. When predictive modeling was attempted in such cases, the model not surprisingly resulted in predicting the outcome of the verification rules rather than the event for which the verification process exists. It has also been shown that there are significant differences in behavior among fraud types, especially between first-party and third-party fraud types. The variables that are most predictive are different, and the importance of variables varies by the type of fraud.

In addition, the value of the credit data in the context of application fraud modeling has some broad implications. It is intuitively tempting to think of credit bureau data as being predominantly about credit activity in the context of fraud. For example, one might expect that potentially fraudulent activities are reflected on the credit file in the form of multiple accounts opened in a short period of time, or in the form of increased balance within a relatively short period of time. However, it is equally important that the credit history of a person provides an objective characterization—credit or otherwise—of the applicant and/or the fraud victim, and this can be more important than the recent credit activity on file depending on the circumstances. Not surprisingly, the reporting frequency of the credit data by the lenders as well as the availability of the exact time stamp in historical data used for modeling have an impact on the sensitivity of the credit activity information in fraud prediction; therefore, one must understand the timing aspect of credit information and how that must be accounted for in model development.

Furthermore, while the types of data and variables most useful is often discussed in context of predictive modeling for fraud, it has been observed that a fraud score built primarily on inconsistencies and negative associations with past applications is not very well correlated with a fraud score built primarily on credit data. While both models individually had the same predictive power, there was a substantial lift in combining the two.

General Challenges in Fraud Modeling

While the methodological process of model estimation is generally the same as most predictive model building efforts, there are several practical challenges that are specific to fraud prediction. Some of these include the following.

1. Definition of the Dependent Variable

   One of the more important challenges in fraud modeling is directly related to the process of identifying the fraud itself, since frauds are used as the target events. The issue of consistency in the definition of fraud and fraud types has already been mentioned above.

   The fraud is perpetuated at the moment of application by definition; however, the fact that a fraud has been perpetrated is often not discovered until later, sometimes never. With that said, the length of time until the fraud is discovered and confirmed can vary widely depending on the fraud process of the lender. The amount of time a suspected fraud is worked and confirmed depends on a number of factors, such as the fraud policy of the lender and the type of fraud as some fraud types are less complicated to verify than others. Often six months or more pass before a fraud is identified and confirmed, and not all frauds are identified and captured in a timely manner, if ever. Furthermore, practices in classification and confirmation of fraud can vary widely from lender to lender, and many frauds go undetected and the loss simply written off, often indistinguishable from other credit losses.

   These issues have a direct impact on the definition of the dependent variable in fraud modeling. First, the concept of "performance period" as commonly understood in typical predictive modeling does not apply in application fraud modeling, since the fraud itself is perpetuated at the point of application; it is the *discovery* of the fact that a fraud has been perpetuated that takes time. This impacts how the analysis periods should be designed, yet there is a need to account for time periods that are as recent as possible due to the changing nature of fraud. Second, frauds that are not identified in a timely manner as well as frauds that are never identified must be treated appropriately. One may try to capture some fraudulent phenomena by examining credit files for the several months following application, and use other information to establish patterns that are otherwise undetectable. Understanding the length of time it takes for most of the frauds to be identified, understanding the process in which these frauds are identified, and the proper identification and treatment of potential frauds that have yet to be confirmed, are some of the keys to effective analysis.

2. Modeling Methodology for Fraud

   From the methodological perspective, one of the issues that often arise in fraud modeling is that it involves prediction of rare events. Common approaches to this problem include the use of modeling methodologies that are more conducive to rare event prediction and use of design techniques such as oversampling. However, the importance of understanding the nature of the fraud and the fraud characteristics can often outweigh the importance of the methodologies employed; head-to-head comparisons of modeling methodologies have shown that logistic regression is still a very viable methodology and often outperforms models built using other single statistical methodologies as long as the insight into the fraud behaviors are properly captured by the model builder.

3. Model Performance Measures

   Measuring model performance for fraud can present a unique challenge. Traditional model performance statistics such as KS, AUROC, and Gini are generally useful only as a very general indicator of predictive power and can be inadequate when comparing model performance in specific cases from a practical perspective. This is due to the fact that typical operating ranges for fraud is usually in the single-digit percentiles, while these model performance statistics is dependent on ranges which are either vastly different or much wider than the typical operating range. For example, the fact that the maximum separation in cumulative distribution of goods and bads occur at the twenty-fifth percentile is practically irrelevant when the lender's policy only allows the worst 5% of the applications to be investigated, and a model with a lower KS or AUROC can actually have a larger impact with respect to the lender's fraud process.

   Other measures widely used such as false positive rates depend on the volume examined, and therefore is only valid for comparison for a given volume for a given population.

Case Study: Generic Fraud Score Development for Canada

As a case study, the development of a generic application fraud score for the Canadian market is examined. In this development, the current application information, the data from an application fraud management system, and the credit bureau data were used among other information to predict both the first-party and third-party frauds across various product types and lending institutions.

Financial institutions in Canada are relatively advanced in the management of application fraud, each with its own set of well-developed rules and policies. In response to the need of the Canadian lenders for additional tools, a generic application fraud score was developed for the market, leveraging Equifax's previous fraud modeling experience in the Canadian and other markets. The data from a comprehensive application fraud prevention and management system was incorporated in this development.

The population of interest for this development consisted of all applications for consumer credit in Canada. The event of interest was credit application fraud of all types, which were classified into the following four (4) general categories: synthetic identity, other first-party manipulations, true-name fraud, and other third-party manipulations.

The sample consisted of applications for various consumer credit products in the Canadian market, primarily from the last quarter of 2009, with an overall known fraud rate of 0.07%.

1. Data Analyzed

   For this analysis, the following were considered as the primary sources of data:

   - Current application data. Variables include personally identifiable information such as name, address, phone number, date of birth, etc.; application characteristics such as product type applied for, channel of application, individual vs. joint application, etc.; as well as employment and applicable security information. They can also be run through a verification or matching process to other data sources for validation.
   - Credit data. They include tradeline information, inquiries, public records, and other miscellaneous financial information available on the Canadian credit file, as well as header data which contain personally identifiable information among others. Variables aggregated from the credit file can represent the characteristics (credit or otherwise) of a person and/or the credit activities of a person, and the information in the credit file also serves as additional source of verification of the current application data.
   - Data from an application fraud management system. This generally consists of past applications and their fraud status, allowing the evaluation of new applications with respect to the validity of the information submitted.

   Other sources include hot word lists and alerts based on additional public and private information. The variables analyzed generally fall into one of the following categories:

   - Characteristics and behaviors (demographic characteristics, credit characteristics, etc.)
   - Activity (application activity, credit activity, etc.)
   - Alerts and warnings
   - Inconsistency/verification attributes
   - Negative matches to past fraudulent activities

   The data sources can further be used in cross-source comparisons and combinations.

2. Dependent Variable Definition

   The primary source of confirmed frauds and their types came from a universe of known fraud data. As discussed earlier, since identification and confirmation of fraud can vary widely, first the time elapsed from the point of application until the confirmation of the fraud was

analyzed by fraud type. Figure 1 shows the cumulative proportion of frauds identified against the time elapsed.

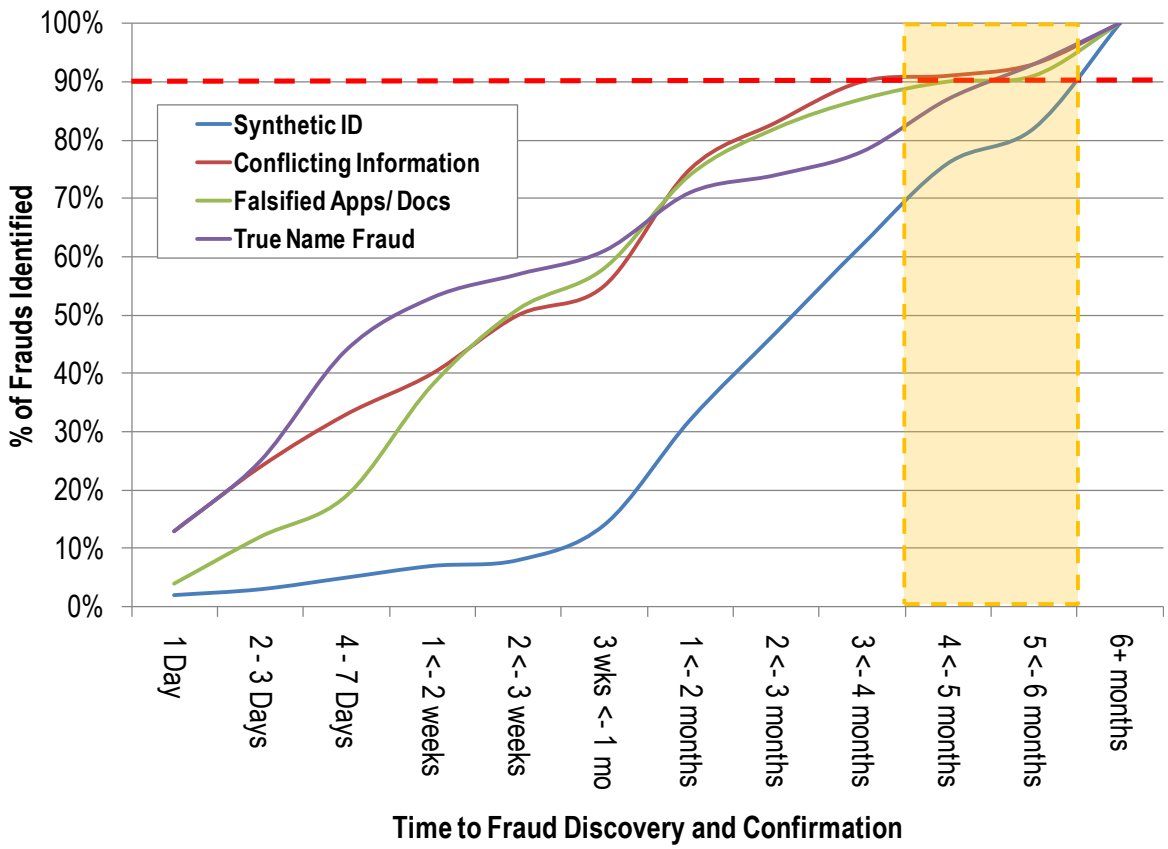

Figure 1. Frauds Identified by Time Elapsed

In this case, 90% of the frauds are confirmed by the sixth month; however, it takes weeks for a simple majority of the frauds to be confirmed, and a few months before 75% of the frauds are confirmed. This implies that the analysis must take into account the fact that it takes up to six months for frauds to be identified and confirmed under the current scenario.

Furthermore, in order to identify frauds that may not have been captured by the investigative process of the lender, the credit file associated with the applicant was examined for charge-offs, lost or stolen trades, etc., in attempt to establish patterns that can be incorporated in the analysis.

3. Value of Data Sources

The contribution of each main data source was evaluated in three ways:

(1) Determine the lift in predictive power attributed to each data source;
(2) Determine the impact of each data source in the score with respect to different fraud types; and

(3) Interpret what the model generally says about fraud based on variables in the final scorecards.

In order to measure the lift in predictive power attributed to each data source, a model was developed for each data source alone, for each pairwise combination of the main data sources, as well as for the combination of all three main data sources. The model with only the current application data serves as the starting point since all application will naturally have at least the application data available. Table 1 shows the results of this analysis.

Table 1. Predictive Power Attributed to Data Sources

Data Source	KS	Bad Rate Worst 1%
Current Application Only	48.0	1.75%
Historical Applications Only	38.4	1.20%
Credit Only	44.3	0.97%
Current Application + Historical Applications	59.5	1.91%
Lift vs. Current Application Only	*24%*	*9%*
Current Application + Credit	54.1	2.13%
Lift vs. Current Application Only	*13%*	*22%*
Current Application + Historical Applications + Credit	60.6	2.39%
Lift vs. Current Application Only	*26%*	*36%*

In this analysis, for convenience the KS is used as a general indicator of model performance, and the known fraud rate (of any type in the table above) in the worst 1% is used as an indicator of its predictive ability in the extremes (the overall known fraud rate is 0.07%). It is evident that single data source models are largely adequate, each with the KS in the range of high 30s to 40s. However, as data sources are combined, there is a substantial increase in the predictive power, especially in the worst extremes of the score distribution. Adding either historical application data or Credit data to the model results in a lift of 9% to 24%; the results also suggest that the contribution of historical application data lies more in the overall separation throughout the population, whereas the contribution of the credit data lies more in isolating the frauds in the extremes. Furthermore, the lift obtained by incorporating both historical application data and credit data is not so much in separating the frauds from non-frauds overall but in isolating the frauds in the worst extremes of the score distribution. In fact, this lift is greater than the sum of the lifts obtained by adding one data sources at a time, suggesting that there is an interaction between these data sources that cannot be captured by each data source alone. Adding both historical application data and credit data sources to the current application data is clearly superior to all other combinations presented.

Next, the impact of each data source in the model was evaluated by fraud type. Figure 2 shows the impact of each of the three data sources to the model by fraud type.

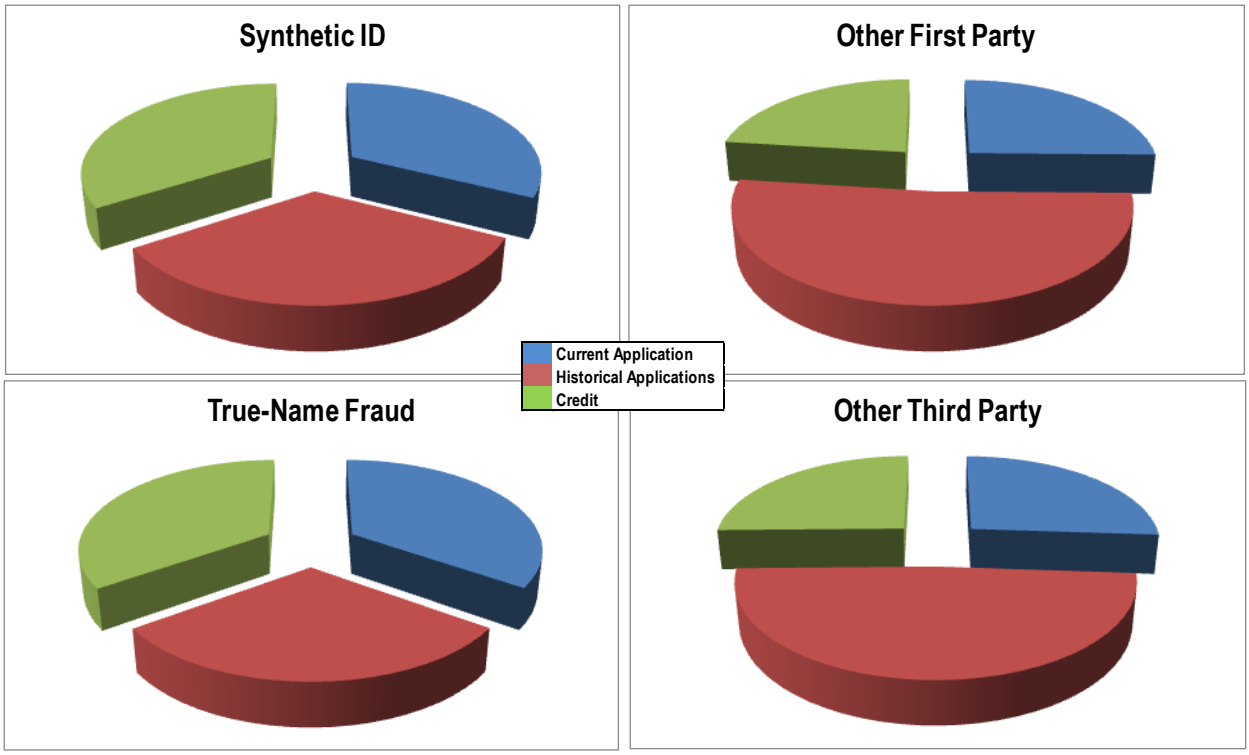

Figure 2. Impact of Data Sources by Fraud Type

The following points are observed:

- The current application data and the credit data are perhaps more important for synthetic identity frauds and for true-name frauds than for other miscellaneous manipulations, be it first-party or third-party manipulations.
- The historical application data have a special importance in predicting other miscellaneous manipulations. This is likely due to its ability to identify abused and/or misused information or inconsistencies based on historical applications, through negative match or inconsistency.

Finally, through the credit variables in the scorecards, the model generally says the following about application frauds in the Canadian market:

- For synthetic identity and true-name frauds, the credit characteristics are more important than the credit activity. For synthetic identity, it is likely that the associated credit file does not show rampant credit activity by the very nature of synthetic identity; in the case of true-name fraud, it may not always be the case that the same identity is used repeatedly to commit fraud.
- Credit risk tends to be negatively correlated with probability of true-name fraud. While there are some exceptions, this is a relatively common phenomenon.

4. General Model Results

The final score is a single score that predicts application fraud of all types. The general results of the model are shown below in Table 2.

Table 2. Summary of Fraud Model Results

Fraud Type	Known Fraud Rate	Fraud Rate Worst 1% by Score		KS	%Frauds Captured in the worst:			
		Known	True (est'd)		1%	2%	3%	5%
Overall	0.07%	2.39%	4.89%	60.6	34%	43%	48%	55%
All 1st Party	0.05%	1.08%	2.28%	54.2	23%	33%	38%	46%
All 3rd Party	0.02%	1.29%	2.69%	73.5	53%	62%	67%	73%
Synthetic ID	0.01%	0.24%	0.52%	62.3	24%	32%	39%	50%
Other 1st Party	0.04%	0.85%	1.79%	52.6	23%	33%	38%	45%
True-Name	0.01%	0.22%	0.46%	62.6	30%	39%	43%	56%
Other 3rd Party	0.02%	1.07%	2.25%	80.3	64%	72%	77%	81%

As can be seen, the model works well on all fraud types, and captures a substantial portion of the frauds in the worst extremes of the score distribution. The known fraud rate of the worst 1% by score is substantially higher than the fraud rate of the population in general.

5. Conclusion

As evidenced in this case study, there are several key challenges in predictive modeling for application fraud that are different from predictive modeling for credit risk. One must account for the particularities in fraud data and fraud mitigation processes, differences in fraud behavior, and their implications in the model design. It is critical that the fraud behaviors are correctly understood in the context of each data sources and variables for an effective prediction of application fraud.