Assignment 9: Dictionary Customization

Predict 453

Section 55

Spring Quarter

School of Continuing Studies

Northwestern University

Janki Vora	&	Daniel Prusinski
Software Engineer		Data Analyst
IBM		US Bank
Dallas/Fort Worth		Minneapolis
TX		MN

In Compliance with Master of Science Predictive Analytics

Deployment Strategy

Everyday millions of dollars are lost in fraudulent transactions, of which few are ever returned to the rightful owners. The most recent trends as of 2013 incorporate grocery stores, bodega shops, and dry cleaners as common businesses that crook's hack into and steal confidential customer electronic data. Given the sheer number of potential compromise points, it is near impossible to individually monitor potential points of compromise. Transactions happen in real-time and fraud trends are based on these transactions. From here, analysis and communication is done post-time from the actual transaction. This is a literal cat and mouse game involving billions of dollars.

Fraud analytics is a sub-field of analytics that is in a constantly evolving landscape. As crooks find new methods and techniques to execute fraud schemes, banks and law enforcement are tasked with utilizing new tools and strategies to prevent, fight, and mitigate fraud losses. At the base level, any electronic fund transaction in the US marketplace leaves a trail of semi-structured data. Listed below is an example of this data:

05/13/13	CHECK CRD PURCHASE 05/09 CARIBOU COFFEE#123 MINNETONKA MN 434257XXXXXX3726 283129798384311 ?MCC=5814	$4.48

As seen above, the date, cost, date posted, MCC, merchants name, location, and random numbers are found in each transaction.

Within this trail, a Merchant Category Code (MCC) is almost always found along with the name of the merchant. Electronic fund transfer (EFT) businesses, such as VISA, MasterCard, and American Express, create and facilitate the electronic records. The bank literally has millions of these transaction trails accumulating every day. They are stored in massive data warehouses and can be manipulated and extracted on a daily basis with relative ease in an effort to mine the text.

Given that this class has taught the SPSS Modeler software, the extraction process would utilize the 'File List' node found within the Text Analytics component. The data is stored as a '.txt' file and is an acceptable format for SPSS analysis. Utilizing a text mining capability to focus on the individual merchant would allow institutions to find fraudulent merchants at the micro level and block specific vendors before a trend develops at the MCC level. As transactions would be mined, a dictionary of merchants would be built and a simple count function based on valid and fraudulent transactions would be established for each merchant. As the count variable for each merchant is established, the foundation for a predictive model is being built based on predicting a fraud trend.

The basis for the predictive model is the accumulation of merchants, which are the variables for the predictive model, and their transaction records. A requirement for regression analysis is that the variables must be continuous or linear in nature. Based on the data, the variables are categorical in nature having a binary response, fraud or not fraud, and do not conform to the basic linear regression assumptions. Bruce Ratner has an Exploratory Data Analysis (EDA) paradigm for building predictive models.

Statistical and Machine-Learning Data Mining:

Problem/Objective: Explore the relationship between Y (Merchants). The variable (s) takes a value of 1 if the transaction was fraudulent and 0 if not fraudulent. Logistic regression, which is based on maximum likelihood and utilizes categorical data, will be used to build the model for this data set.

Data: The data has been aggregated and has been supplied from the banks data warehouse. Please note that there is a binary response variable and binary explanatory variables. The data would be exported into SAS as a data step for each variable.

Analysis: Once the count variables have been assigned to each merchant from the SPSS Text Mining, a PROC Logistic statement with a class statement for each merchant would be used to assess the predictive accuracy of individual merchants with the response variable, for each merchant, which is the probability that the merchant has an emerging fraud trend. At this point in the model building, a fraud threshold will have to be established to consider a merchant fraudulent. For example, if Bill's Deli has had 100 transactions in the past day and 35 were fraudulent, the fraud threshold would be set at a .35 probability to then block that merchant, and management would have to agree on this threshold. Management would probably want different thresholds for different industries.

Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the variables predict emerging fraud trends. This can easily be assessed by looking at the transactions that have been declined and analyzing whether they were fraud or valid.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best Logistic regression model, and the analyst's personal bias is mitigated.

Proc Logistic produces quite a bit of convoluted output and finding a few of the key goodness-of-fit statistics are all that is needed to make an informed decision regarding fraudulent merchants. Analyzing the inferential statistics AIC, SC, and -2 Log are informal methods to assess the model fit. All of these statistics can be used to compare different sets of variables, merchants in this instance. Higher values for these statistics mean a worse fit to the data. Of the three statistics, the -2 Log is the most important. This statistic is the maximized value of the logarithm, which is derived from the likelihood function multiplied by -2 (Allison 2012). The benefactors of this model will most likely not have a background in predictive analytics, and keeping the interpretation of the output simple is essential for the model performing for its intended use. Each variable produces a coefficient value, which is its predictive importance, but in logistic regression it is hard to interpret the coefficients value. The odds ratio is simply computed by taking the natural log of e^estimate. In addition to the odds, this calculation is also adjusted since it controls for other variables. Once the odds ratio has been calculated, the output is easier to understand. This function is a part of the Logistic output, and the greater the odds for a particular variable the greater the probability that the variable is fraudulent. In order to implement this model, a goodness-of-fit threshold would need to be set, and an odds threshold would need to be set in order for the average analyst to utilize the model. For example, if 70 merchants in a model had an overall AIC score of less than 50 and an odds threshold score of greater than 8, an analyst should block the specific merchants.

Banks have specific analysts that monitor merchants or portfolios made up of multiple merchants. The Logistic regression model could be updated daily or hourly depending on the vigilance of the analyst. Creating an automated system for the text extraction and modeling component would be the next step in creating an ease of use for the average analyst. Bear in mind that the goal is simply to have the average analyst interpret the model output and make the necessary changes to the portfolio in regard to blocking specific merchants. Models need iterations and maintenance as new fraud trends emerge. This model is no exception to updating, but the competitive advantage lies in the models ability to analyze and predict individual merchants at the micro level.

Utilizing text mining capabilities would allow institutions to mine merchants at the micro level and block specific vendors before a trend develops at the million dollar level. This would have the cost savings potential of billions annually for the banking industry. Given the customized data dictionaries for the bank, an organization would mine real-time transactions through the dictionary and as specific merchants were mined they could in-turn be blocked at the individual level saving the need to shutdown other merchants. Pairing Logistic regression and text mining strategies would result in stronger predictive fraud models that in turn would save financial institutions billions. The technology exists and it now boils down to building the infrastructure and due diligence for the necessary iterations needed to build a great model.