Assignment 6: Dictionary Customization

Predict 453

Section 55

Spring Quarter

School of Continuing Studies

Northwestern University

Janki Vora
Software Engineer
IBM
Dallas/Fort Worth
TX

&

Daniel Prusinski
Data Analyst
US Bank
Minneapolis
MN

In Compliance with Master of Science Predictive Analytics

Daniel Prusinski
4/22/2013

Dictionary Customization

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 charged with utilizing new tools and strategies to prevent, fight, and mitigate fraud losses. At the base level, any electronic transaction in the US marketplace leaves a trail of semi-structured data. 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 MCC's provide a major strategic stronghold for honing in on fraud trends and quarantining fraudulent common points of purchase. Utilizing the MCC's in text mining is another strategic approach to combating the ever changing fraud landscape.

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. In essence, fraud analytics seeks to predict tomorrow's fraud with yesterday's data. Honing this process can be the difference in stopping billions of dollars in fraud a year.

Time series analysis reveals that the more recent the data the better it is in quality. Traditionally, once a fraud trend is established financial institutions put a block on the MCC or even the actual merchant, but this is often after millions have been lost. In addition, it is a highly political process to block certain vendors from receiving EFTs. In the customized data dictionary for this assignment, each major MCC has its own category. Within each category, there many different vendors and businesses. Currently, when a trend is established it often takes time before the corresponding MCC can be entirely blocked. This is because numerical values are used for data mining not text. This difference equates to hours and even days to identify a trend big enough to warrant blocking an MCC. Utilizing a text mining capability would allow institutions to mine merchants at the micro level and block specific vendors before a trend develops at the MCC level. This would have the cost savings potential of millions. Given the customized data dictionary for this assignment, 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 entire MCCs.

Pairing numerical 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.

☒Select a domain, your own industry or one of interest to you, and create four new categories relevant to the domain.

☒Identify at least 40 concepts/terms that have special meaning in this domain and map them into the categories you selected.

☒Use a spreadsheet to show the mapping with a column for each category and the terms as row entries.

☒Provide a one- to two-page summary describing the logic behind the categories you created.

☒Describe also how they relate to important outcomes for which you might use predictive analytics.