Predictive Analysis of ECB Violations

Darius Mehri, Lior Shahverdi

Summary

Predictive and exploratory data analysis was conducted on 5 years of ECB violations. The objective of this study was to develop a predictive model for ECB violations. We utilized two methods – market basket and social network analysis. The market basket analysis was used to predict violations and the social network analysis was used as an exploratory visualization tool to complement the market basket analysis. The research revealed mixed results - violations are issued to buildings around a specific initial incident and are therefore highly connected in content. The lack of domain knowledge among the researchers led to incomplete interpretation of the results. The conclusion in this memo briefly recommends a research design for future ECB predictive analysis.

Data and Methods

The objective of market basket analysis is to determine what items appear in any given transaction. A transaction is composed of a group of one or more items (an itemset). The result of a market basket analysis is a list of *association* rules that show patterns found in the relationships among items in an itemset. Association rules are denoted by relating one itemset on the left-hand side of the rule to another itemset on the right-hand side of the rule. The left-hand side is the condition that needs to be met in order to trigger the rule, and the right-hand side is the result (or the outcome) of meeting that condition. For example, a classic grocery store association rule formed from customers who purchase lunch related items is: {peanut butter, jelly} -> {bread}. This association rule states that if peanut butter and jelly are purchased together, bread is also highly likely to be purchased.

Whether an association rule is deemed to be useful is determined by two measures – support and confidence. Support measures the frequency the itemset occurs in the data. An association rule's confidence is a measure of accuracy. Typically, in an analysis, the researcher would like both of these measures to be high because for an association rule to be useful, it must occur frequently and with high accuracy.

For the ECB violation analysis, a transaction is an itemset of violation infraction codes for a single building. For example, BIN number 2065945's itemset includes infraction codes 101, 103 and 110; BIN number 1000038's itemset includes infraction codes 22, 770, 80, 64, 110, and so on. The dataset for this analysis included 2010-2015 ECB violations which resulted in 324,333 total observations (rows of data).

The social network analysis had two objectives. The first was to visualize the strength of ties among the violations. Violations with strong ties are more likely to co-occur together (i.e. if one violation is given, another is automatically given). The second was to explore clustering of violations – high clustering (nodes that are close together) means that the violations are similar to each other in the network. Since the dataset was very large, network graphs were created for each month in 2015. For each of these infraction codes, the itemsets generated an undirected set of ties. Once the ties were generated for each BIN we counted how many times each distinct association appeared across all BINs and assigned these counts as weights to the corresponding tie (or edge).

Results

Market Basket Analysis

The analysis produced over 200 association rules, below are two that were most interesting:

Rule 1: Confidence: 0.92 Support: 0.00347
155: FAIL MAINTAIN BLDG IN COMPLIANT MANNER:LACK OF AUTOMATIC SPRINKLERS
189: FAIL TO MAIMTAIN REQUIRED NUMBER OF MEANS OF EGRESS FOR EVERY FLOOR
=>106k:ILLEGAL ACTIVITY

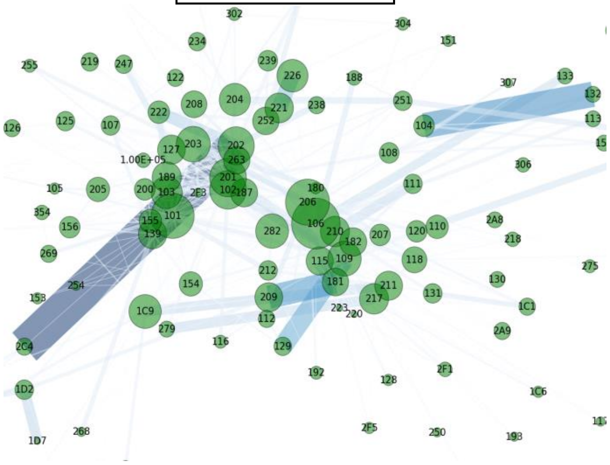
Rule 2: Confidence: 0.91 Support: 0.0065
105: 1OR2 FAMILY CONVERTED/MAINTAINED AS DWELLING FOR 4 OR MORE FAMILIES
187: UNLAWFUL ACTS.FAILURE TO COMPLY WITH AN ORDER OF THE COMMISSIONER
=>101: WORK WITHOUT A PERMIT

Rule 1 states that a failure to maintain automatic sprinklers and an egress for every door predicts illegal activity. This rule has an accuracy of 92% and frequency of 0.347%. Rule 2 states that the conversion of a 1 or 2 family home and unlawful acts predicts work without a permit.

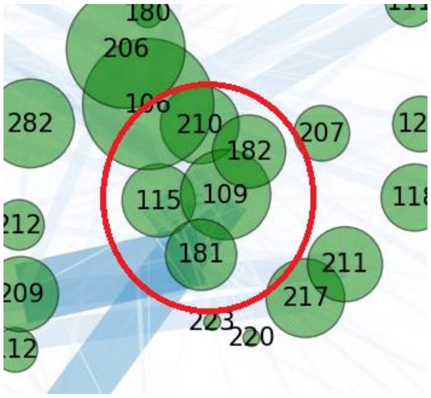
Social Network Analysis

The below graphs shows the network with edge thickness displayed and a cluster related to safety for January 2015 :

January 2015



January 2015 Safety Cluster



The January results show strong ties between 4 pairs of violations - 202: Failure to Maintain Compliance and 2C4: Electrical Work Without Permit; 181:Failure to Maintain Housekeeping and 209: Failure to Safeguard Persons; 109: Failure to Safeguard Persons and 129: Unqualified Supervisor on Scaffold; 104: Failure to Maintain Walls and 132: Failure to Maintain Exits. These strong ties indicate that they co-occur regularly – perhaps when one violation is given another related violation is automatically issued.

Three clusters can be identified in the network – one in the center and two on the left. The cluster pictured on the above right is a sub-cluster of the center cluster and is of interest because it is related to safety. It is composed of violations 210: Failure to Provide Documents at Construction Site, 115: Failure to Maintain Safety Equipment, 182: Work Not Conforming to Documents, 109: Failure to Safeguard Persons and 181: Failure to Maintain Housekeeping.

Conclusion

To further develop a model to predict ECB violations, the researchers recommend abandoning the bottom-up, inductive analysis in favor of a deductive, top-down approach. A top-down approach will involve working together with experts in the department with ECB domain knowledge to develop a robust research question, narrowing the scope of the analysis, hypotheses testing and validating the results.