Abstract

Author: John Walsh

Title: Spatiotemporal Determinants of Football Stadium Incidents Using Apriori Association Rules Mining

Institution: Wilkes Honors College at Florida Atlantic University

Thesis Advisor: Dr. Bharat Verma

Degree: Bachelor of Arts in Interdisciplinary Mathematical Sciences

Concentration: Biochemistry

Year: 2023

Given that the government is rekindling public interest in sports by hosting the World Cup in the United States, it is vital to better understand security and safety incidents at stadiums. Predicting incidents at sports stadiums is valuable for ensuring safe sporting events. Examples of stadium incidents include fan misbehavior, security incursions, and maintenance requests. We used proprietary data from a football stadium to build our model. We chose the Apriori Association Rules Mining Algorithm (ARM) to model the data. We chose the Apriori ARM due to the co-linearity of features and the depth of the data. The findings elucidated that security incidents are likely to occur during the post-game of an event. Also, specific locations had higher security incidence likelihood. Another of the findings was accessibility requests most often occur in the beginning and at the end of events.

Table of Contents

1. Introduction

During a sporting event, incident types may be related to security, maintenance, fan code of conduct, accessibility requests, janitorial assistance, requests for venue staff, and parking. With all incidents simultaneously occurring in various locations throughout the events. Standard football events can be broken up into seven different event markers. These event markers are: “Parking Lots Open”, “Gates Open”, “1st Quarter”, “2nd Quarter”, “3rd Quarter”, “4th Quarter”, and “Post Game”. Given the structure of events and the nature of incidents, the initial analysis led to the development of one primary goal. To identify the spatiotemporal determinants of incident type.

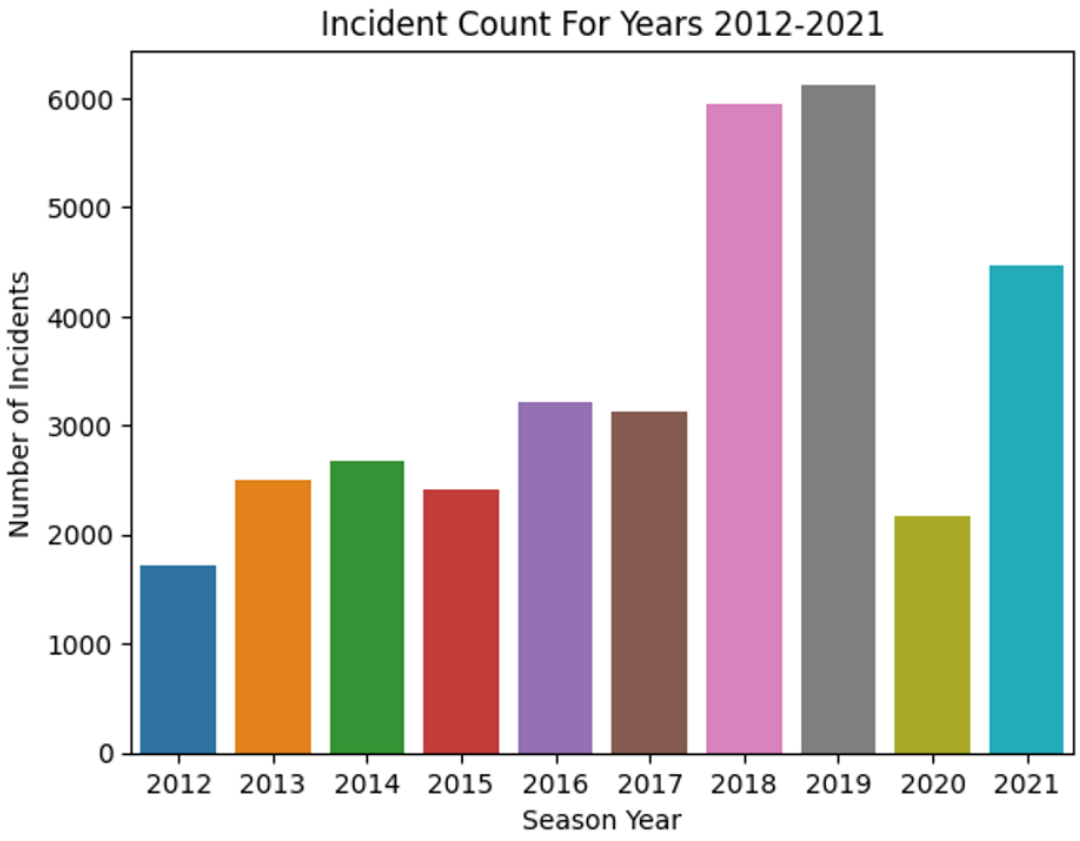

Figure 1A: Count of Total Incidents per Season Year

The general trend is incidents are increasing with respect to time. Even during the COVID impacted year 2021 the number of incidents exceeded the number of incidents in many previous years. Thus, the relevance of this work is notable because if incident likelihood is known, stadium managers can direct venue staff and security forces to be in places where specific incidents are likely to occur, hopefully before they happen. Providing stadiums with an unparalleled ability to improve safety for patrons and staff alike, while also enhancing the enjoyment of the fans that are paying to attend the event. This improvement of experience is important in the context of our modern society because stampedes, active shooters, and bomb threats hold a near constant place in news cycles, hence safety is a legitimate concern for event patrons. With events like the World Cup happening in the United States in the future, analyses on par with the work done here can be a force to cultivate peace of mind for event attendees, prevent potential tragedies, and improve efficiency in stadium operations.

The model employed to answer the questions was Apriori Association Rules Mining (ARM), details pertaining to how the decision to use this model were made as well as the subtleties of this specific use case are addressed in Section 3 “Modelling”. Due to the size of the dataset and variability of features, the model is extremely adaptable and can be adjusted to answer more specific questions dependent upon the desirability of output. As seen in (Figure 3B), lower support and confidence thresholds produce larger association rule sets which have benefits as well as considerable pitfalls which will be addressed in the "Discussion" Section. Due to the categorical nature of the question, the set theoretic foundations of the ARM model were perfectly suited to handle the data. However, various optimization techniques like dynamic itemset counting can be applied if working with a larger dataset. It should also be noted that this

data comes from a single football stadium, and if the model was generalized to all football teams the insight may be different and statistically more robust due to a larger sample size.

2. Literature Review

In the seminal paper "Mining Association Rules between Sets of Items in Large Databases", (Agrawal, Imielinski, & Swami, 1993), the authors introduced a novel data mining algorithm. The algorithm was originally developed to model shopping cart data which is also known as market basket data. This in turn gave rise to what is now known as Market Basket Analysis. The algorithm posed in the paper aims to generate association rules for all possible baskets or carts. The rules generated consist of a calculated likelihood or observed probability that if antecedent items are present there is an observed probability that consequent items will also be present in the basket. This observed conditional probability is called the confidence of the rule. Along with confidence, the authors introduced an additional measure called support which is simply the frequency of the items present. Support is the primary measure utilized during the algorithm where confidence is a measure computed to assess the probabilistic likelihood of the rules generated, effectively making confidence a measure of model performance, whilst support is a predetermined model parameter.

The algorithm consists of two constraints pertaining to rule generation. The first is a syntactic constraint which is synonymous with the technique known as pruning. The syntactic constraint is applied to the algorithm to only generate rules of interest. For instance, if analyzing shopping cart data, a syntactic constraint may be only keeping rules with milk as an antecedent item. Syntactic constraints are arbitrarily determined and are not necessary if the goal is to identify all rules present in the dataset. While the second constraint is quantitative in nature, this

is the support constraint. The support constraint pertains to the frequency of the rule appearing in the dataset. As discussed, the support measures the frequency of antecedents and consequents appearing together given the number of antecedents. Probalistically support can be represented as supp(A→C) = $P\,(x \in (A \cap C))$, where A is an antecedent itemset and C is the consequent itemset. Support helps determine which rules are important because rules appearing in 15% of all data entries are far more statistically significant than rules appearing in 0.2% of entries.

The general structure of the algorithm devised by the authors is iterative in nature. It views each row or entry of the dataset as an itemset. Then sequentially measures the relative frequency or support of each item from the itemset. If the support exceeds the minimum support constraint predefined by the user, then the rule is generated for that item. If the support is less than the defined support constraint, then the item is discarded, and a rule will not be generated. This process continues for all items of the original itemset until the itemsets have been exhausted iteratively. Ultimately producing rules consisting of antecedent itemsets associated with their respective consequent itemsets. A schematic of this algorithm is found in Figure 3C. However, the algorithm in Figure 3C represents the Apriori Association Rules Mining Algorithm (ARM) devised by Agrawal, Manilla, Srikant, Toivonnen, and Verkamo in 1996. The Apriori ARM is a more efficient and generalized rendition of the original ARM. The improvement stems from the allowance of antecedent and consequent itemsets being able to contain more than one item, where the original ARM restricted consequent itemsets to contain only one item.

3. Modeling

We observed Season Year had a Gini-index of 0.00044 thus season year was almost perfectly equally distributed throughout the dataset. Therefore, exclusion of season year from a predictive role in the model is appropriate due to the contribution of noise from the variable.

Given the distribution of season year in the dataset. Further investigation of the proportions of incidents during specific event markers throughout each season year was required. As shown in Figure 3A, from year 2016 onwards, the event markers were proportionately consistent throughout the season years. Thus, despite large discrepancies in the number of incidents occurring annually, the proportions of incidents per event marker were similar. From this distribution we were able to ascertain that if we are only considering seasons after 2015, we can effectively model when incidents are likely to occur during an event.

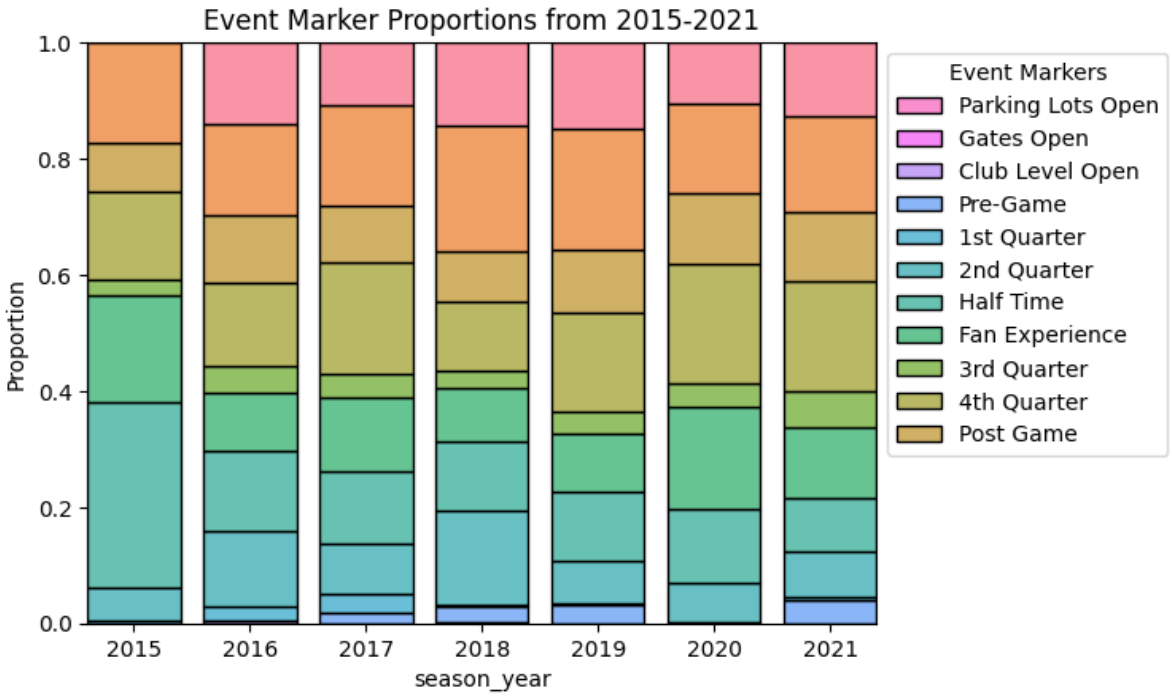

Figure 3A: Proportions of incidents per event marker by season year.

The next step was to analyze if there is a correlation between incident types and event markers. To accomplish this, we first needed to generalize the incident types into incident categories because the dataset contained over 200 unique incident types. Even though there was an abundance of incident types, all incidents fell into distinct categories which were defined as Code of Conduct (C), Security (S), Maintenance (M), Janitorial (J), Parking (P), Accessibility (ADA), and Venue Staff (V). This step of mapping the incident types into incident categories greatly reduced the dimensionality of the incident variable. Once reduced, we computed the ϕ_k correlation of the variables of interest (location, event marker, and incident category).

In Figure 3B the ϕ_k correlation matrix shows a correlation of 0.43 between the incident categories and event makers. While in a traditional linear regression model this correlation may not be significant. The correlation was acceptably high because it indicates that across more than seven event markers and seven incident categories the variables are still correlated. Other notable correlations that contributed to model development are the correlation between locations and event markers, also locations and incident categories.

Particularly important of the two aforementioned correlations was the relationship between location and incident category. A correlation of 0.7 indicates that specific incident categories are correlated with specific locations. As seen in the model results in Figures 4A and 4B, only location, event marker, and incident categories are used in the model. However, from the correlation matrix, a high correlation between season year and location is found. This indicates that in different years various locations have had more or less incidents; therefore, further analysis can be done to investigate this relationship but that is not within the scope of this project.

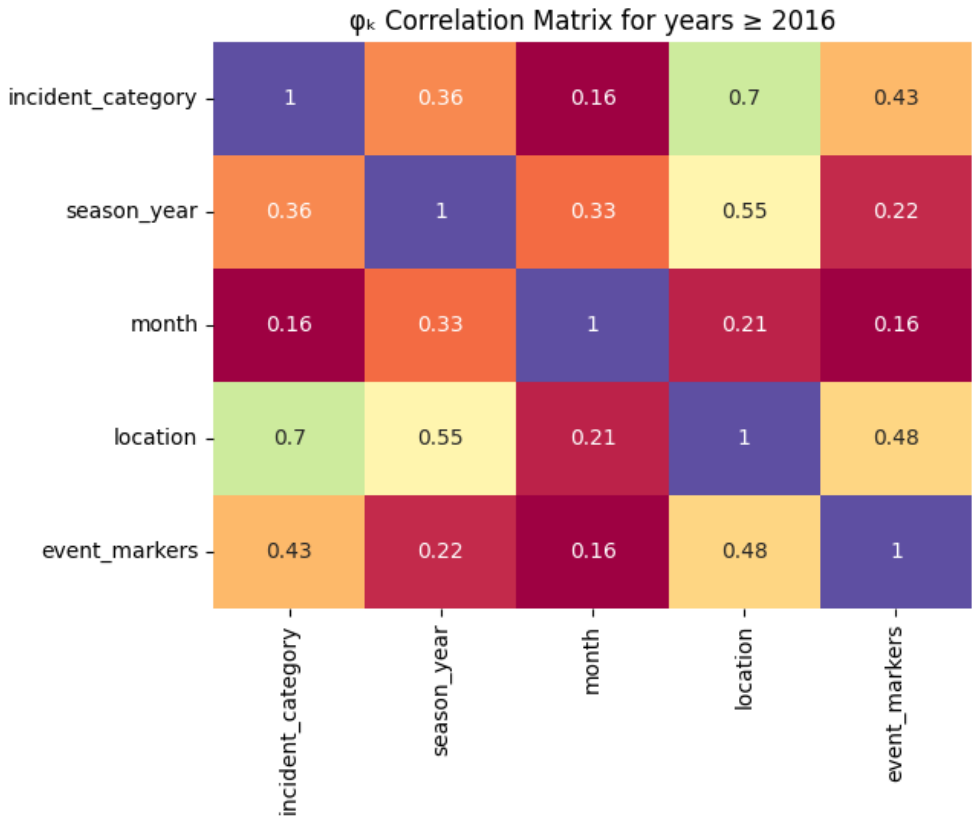

Figure 3B: ϕ_k Correlation Matrix displaying the correlation between variables.

To understand the relationship between location and incident type observed in the correlation matrix, a χ^2analysis was performed on the incident category variable to verify the extent of correlation between incidents and location. χ^2 was chosen due to the categorical nature of each variable. From the χ^2, sufficient correlation was observed between the incidents and location which confirmed the ability to model incident type using location. A similar analysis was then performed on incident category with respect to event markers and sufficient correlation was also observed. Thus, the correlations observed in Figure 3B were statistically verified using a categorical analysis. Ultimately elucidating that it is viable to model the data using incident category, location, and event markers.

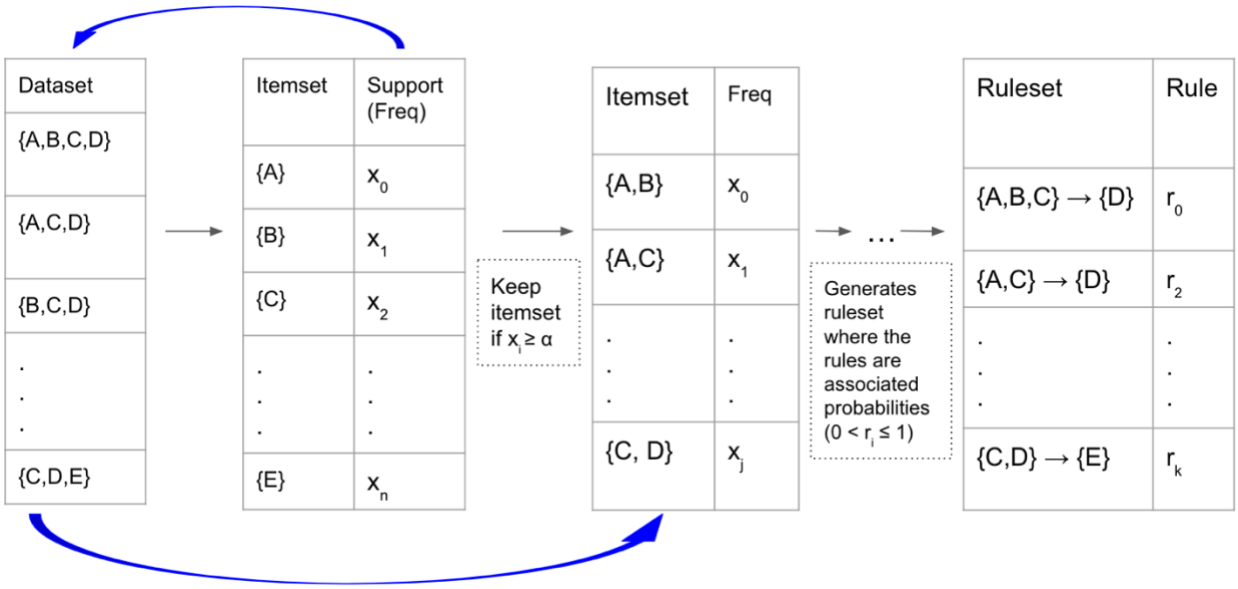

Figure 3C: ARM Schematic Diagram for min_support = α

The Schematic diagram above illustrates the Apriori ARM algorithm used to model the dataset. As discussed in the Literature Review, the algorithm generates association rules based on the mutual cooccurrence of items in the dataset. Each incident report is viewed as an itemset, and the algorithm iterates through all the items in each itemset, computing the support and keeping the itemset if the support exceeds the predefined min_support. Then the probability of an additional item being present in an itemset is computed. From the diagram above, an example of a rule (r_0) is the associated conditional probability of an item (D) also being present in an itemset containing other items ({A, B, C}).

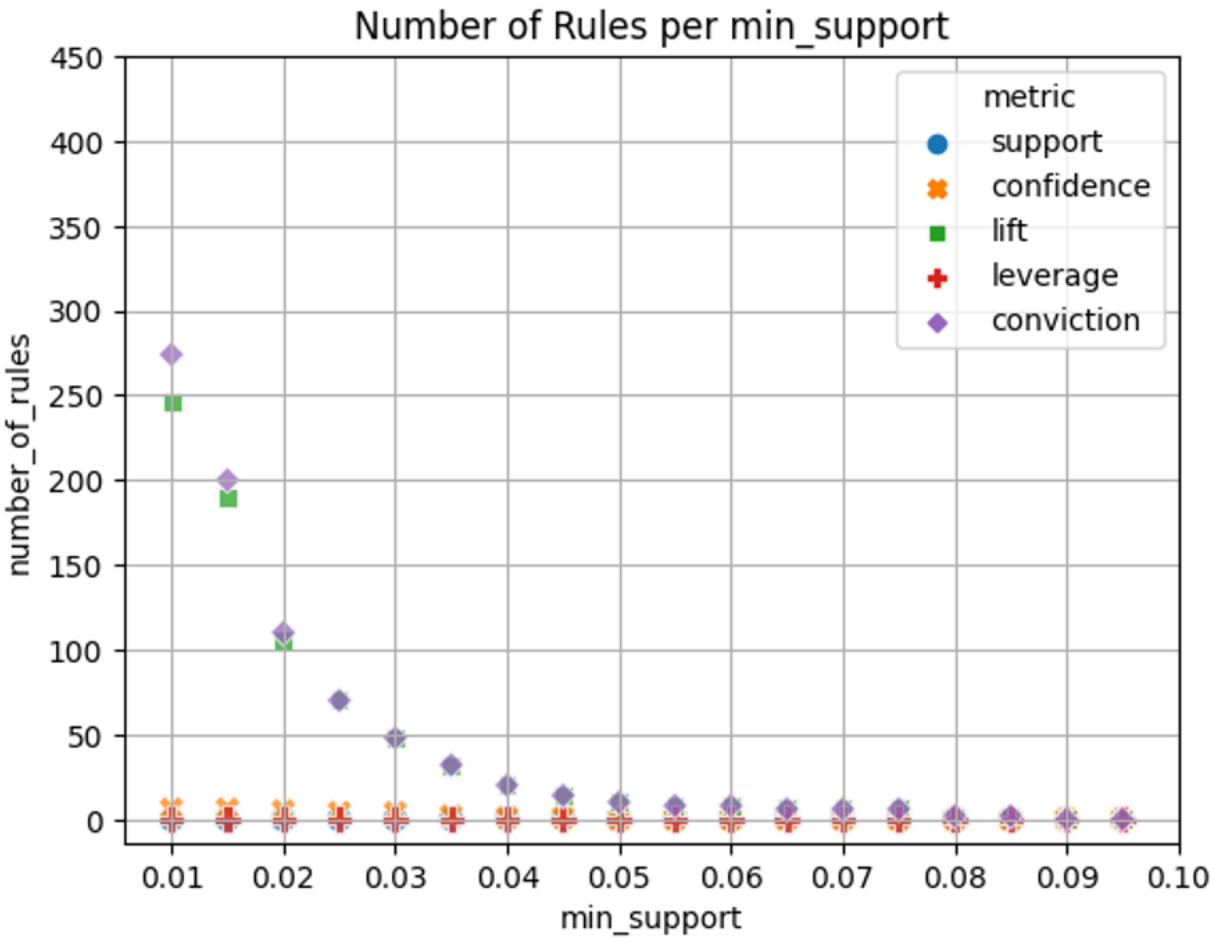

Figure 3D: Model Performance for min_support ranging from 0.01 to 0.1

In the ARM model five equations are considered for choosing how the itemsets are filtered (See Definitions below) in the algorithm. A minimum support threshold of 0.02 was decided after assessing the performance of the algorithm using a range of minimum support values see Figure 3D. The minimum support values were set for models using the various filtering metrics. Lift was identified as the optimal metric for our dataset. Lift determines if the observed occurrence of a rule is greater than the expected occurrence of the rule. Hence, if Lift(A→ C) > 1 then the rule (A→C) is more likely than statistically expected.

4. Definitions

Support: $supp(A_j \rightarrow C_j) = \frac{frequency(A_j, C_j)}{N} = P(x \in (A \cap C))$

Confidence: $conf(A_j \rightarrow C_j) = \frac{frequency(A_j, C_j)}{frequency(A_j)} = P(x \in (C|A))$

Lift: $lift(A_j \rightarrow C_j) = \frac{supp(A_j \rightarrow C_j)}{supp(A_j) \times supp(C_j)} = \frac{conf(A_j \rightarrow C_j)}{frequency(C_j)}$

Leverage: $leverage(A_j \rightarrow C_j) = supp(A_j \rightarrow C_j) - supp(A_j) \times supp(C_j)$

Conviction: $conviction(A_j \rightarrow C_j) = \frac{1 - supp(C_j)}{1 - conf(A_j \rightarrow C_j)}$

5. Results and Insights

Association Rules Mining is colloquially known as Market Basket Analysis due to its prevalent use in the commercial sector. If viewing the dataset in this project through the lens of Market Basket Analysis, there is a poignant analogy. In Market Basket Analysis a common result may be that grocery shoppers often buy eggs if they also buy bread and milk. In stadiums, a code of conduct incident may occur in a specific location if it is the second quarter of the game. Henceforth, the same way a grocery store may then choose to place eggs in the section near the bread and milk. A stadium manager may also choose to place security personnel in that specific location during the second quarter to prevent more code of conduct violations from occurring.

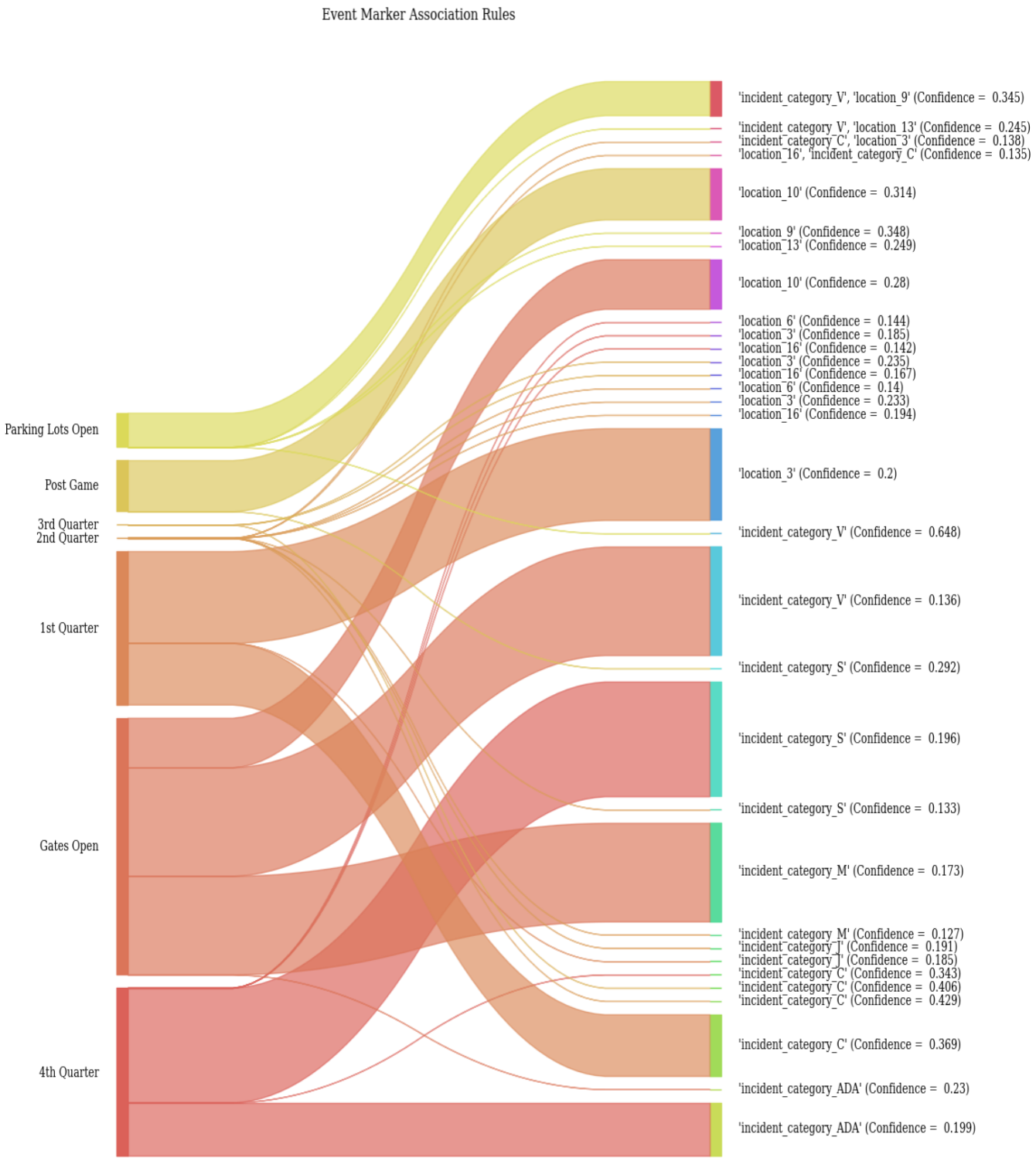

Figure 5A: Event Marker Association Rules with min_support = 0.02

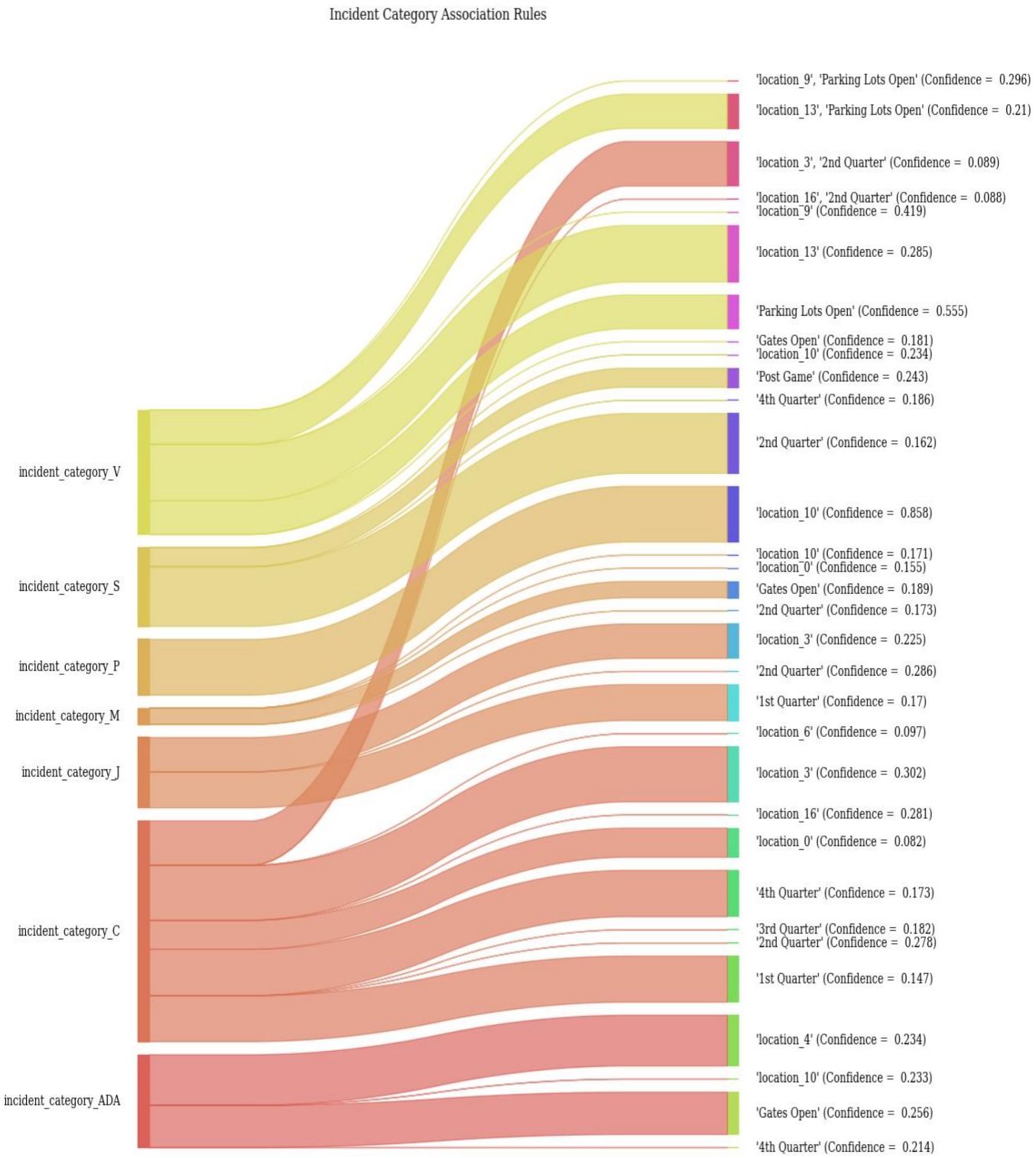

Figure 5B: Incident Category Association Rules with min_support = 0.02

6. Discussion

The rulesets generated from the ARM model provide extensive insight as to how stadium incidents occur during events. It is important to reiterate that the rules do not infer causation, instead each ruleset is a representation of mutual cooccurrence. The rules were visualized using Sankey Diagrams. The diagrams display the antecedents on the left column with the corresponding consequents displayed on the right column along with the associated confidence value of each rule. Confidence is analogous to the conditional probability of the consequent appearing given the antecedent is present in an incident report.

Therefore, Figure 5B displays the conditional probabilities of an incident occurring during an event marker and at a location if given a specific incident category. While Figure 4A displays the conditional probabilities of incident category and location if given a specific event marker. In essence the results model how incidents are likely to occur throughout an event. This analysis provides an unparalleled overview of the distribution of incidents for the given stadium.
The results also serve to highlight operational shortcomings of the stadium. The 'incident_category_ADA' rulesets observed in Figure 5B describe many accessibility requests occur when the gates open and during the 4th quarter. The significance of the rule pertains to the nature of how this incident is reported. The report only occurs when a wheelchair or similar service has been requested, hence all requests occurred from patrons that did not have the services needed present when they initially needed them. This indicates that the stadium could improve operations by having more staff members ready during these two times during events.

Other results pertinent to improving stadium operations are the associations of 'incident_category_C' found in Figure 5B. These are fan code of conduct violations, and the model reports code of conduct violations are occurring primarily in only four stadium locations.

Therefore, security forces can be directed to these locations in larger numbers to deal with and prevent situations from escalating. Particular to code of conduct incidents is the rule that states if the incident category is code of conduct, there are 8.9% and 8.8% chances that the incident occurs during the 2nd quarter in 'location 3' or 'location 16' respectively.

A beneficial facet of the ARM algorithm is that it computes associations independently. This independence stems from the fundamental equations used to quantify associations. All five metrics defined are unique in the sense that a rule (A→C) is different from rule (C→A). This difference arises due to the proportions or frequencies computed in the algorithm. For instance, if conf(A→C) is 0.088 this means $\frac{freq(A,C)}{freq(A)} = 0.088$; however, this does not imply conf (C→A) will also be 0.088 due to the freq(A) component on the denominator. This incongruency present in the metrics enhances the robustness of the model because rule (A→C) may provide entirely different insight compared to rule (C→A). An example of this is found in Figure 5A. As discussed above, the rule (code of conduct, 2nd quarter → location 16) has a confidence of 8.8%, yet the rule (2nd quarter, code of conduct → location 16) has a confidence of 13.5%. This rule tells a different story than that told in Figure 5B. The rule is stating that 13.5% of all incidents in the 2nd quarter were code of conduct violations in location 16.

Ultimately this rule states that during the second quarter, many code of conduct violations occur in location 16. From the perspective of a stadium manager attempting to improve response times to incidents, this rule is more beneficial than the rule found in Figure 5B because it addresses the incidence likelihood during a specific timeframe. Whereas the similar rule generated in Figure 4B provides a likelihood with respect to all code of conduct incidents. So, if a stadium manager desires to know when specific incidents are most likely to occur Figure 5A is more powerful.

Another important note when addressing the results generated by the ARM model is that the model can be iteratively performed on subsets of the dataset and generate entire rule sets for each incident category to be analyzed separately. Then by analyzing model performance with respect to ruleset size, model parameters can be tuned to produce rulesets specific to each incident category. For industrial applications this could be beneficial since if each incident category is analyzed on an individual basis, optimal strategies could be devised in terms of staff placement during various timeframes throughout events relative to the required staff for each incident type.

Overall, the ARM was a robust and well-suited model for determining stadium incident likelihood. ARM could handle the extensive depth and collinearity present in the dataset while still producing valuable insight. The flexibility of the model to capture different antecedents and consequents allows for the ruleset to be pruned to only include associations of interest if desired. This is beneficial considering it can empower stadium managers to view rules that pertain to any variable of interest.

7. Conclusion

As shown in this analysis, the ARM algorithm first introduced by Agrawal, Imielinski, and Swami in 1993 can be a useful tool to manage incident response. Given adequate correlations and consistency with the distribution of events through time as shown in Figure 3A the model was suitable for the analysis. However, due the presence of user error in incident reporting, extensive cleaning and standardization was required for the dataset to ensure proper model results.

During the exploratory data analysis variables such as month, time of day, part of season, and away team were phased out of the analysis. This does not imply that they are not of potential use for future ARM models, especially considering that Figure 3B displays a correlation between both month and season year with incident location. These relationships can be studied further to potentially improve future model performance. Although season year may not provide great operational insight given the low Gini Index, month is certainly a viable candidate for further investigation.

From football stadiums to concert venues, ARM can be a contributing factor to improving stadium safety around the world. Not only stadium safety but also increased operational efficiency as well as improved patron experience. The improvement of experience because incidents whether assault, a missing child, or wheelchair requests can be statistically modeled to hopefully have the proper personnel at the incident location before it happens.

8. Acknowledgements

Special thanks to Michael Karp and 24/7 Software for granting access to an anonymized stadium incident dataset and assisting in understanding the structure and meaning of various data points. Also thank you to Dr. Bharat Verma for guiding me during this project, his insight was critical in the development of the model.

It is imperative to note that the open-source nature of the python programming language made this project possible. The python packages used in this work were pandas, numpy, seaborn, pySankey, mlxtend, regex, matplotlib, and phik.

9. Citations

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H.(., & Verkamo, A.I. (1996). Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining.

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2018). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. http://arxiv.org/abs/1811.11440

Raschka, (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of Open Source Software, 3(24), 638, https://doi.org/10.21105/joss.00638

Anazalea (2018) pySankey [Source Code].https://github.com/anazalea/pySankey

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

Van Rossum, G. (2020). The Python Library Reference, release 3.8.2. Python Software Foundation.

IBM. (n.d.). Lift in an association rule. Data mining –. Retrieved February 2, 2023, from https://www.ibm.com/docs/en/db2/10.5?topic=SSEPGG_10.5.0%2Fcom.ibm.im.model.doc%2Fc_lift_in_an_association_rule.htm