Assignment 1

DATA 630-9040

Association Rules Analysis

Prepared by Kierra Young

Prepared for Professor Bati

6/08/2021

**Introduction**

Objective

The goal of this analysis is to uncover the leading factors that contribute to nursery applicants being rejected frequently from nursery schools. I will be doing an explanatory analysis to uncover the key variables that determine whether a person’s child can be accepted into a given nursery school or not. The specific modeling types that will be used for this analysis are bar plot, histogram, and a faceted box plot. There will also be testing and more forms of detailed analysis to come to a significant explanation of why certain applications are continuously being rejected.

Problem Domain

Some interesting background information on nursery school rejections comes from an article based in New York . It reads the following:

**“NEW YORK, 9 April 2019 –**More than 175 million children - around half of pre-primary-age children globally - are not enrolled in pre-primary education, missing a critical investment opportunity and suffering deep inequalities from the start, UNICEF warned in a new[report](https://uni.cf/world-ready-to-learn-report) released today. In low-income countries, the picture is much bleaker, with only 1 in 5 young children enrolled in pre-primary education.”( unicef.org)

Even though this article is recent, the main problem is that early childhood educational rejections have been happening for years. It is an on-going procedure and is expected to continue happening to families that are low income as mentioned in the article above. The question at hand is “Why?”, and “Are there any other factors that could possibly be leading to so many rejections?”

Method Rationale

The method of rationale that I will be using is relevant to the nursery rejection application problem. The exploratory analysis will give a clean interpretation of what may be the problem with some many applications constantly being rejected by explaining a dataset relevant to the topic at hand. Likewise, the modeling techniques such as box plot, bar plot, and histogram will display a visual insight on the possible reasons why so many people are being rejected.

**Analysis**

Data

“This database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It was used during several years in 1980\'s when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation. The task is to rank applications for a given nursery school”( sci2s.ugr.es).

Furthermore, the source of this dataset comes from Knowledge of Extraction based on Evolutionary Learning. This is real world data and it has 9 variables and sub categories. The variables are the following: “Parents{pretentious, great\_pret, usual}, Has\_nurs{critical, very\_crit, less\_proper, proper, improper}, Form{incomplete, foster, completed, complete}, Children{1, 3, more, 2},Housing{critical, less\_conv, convenient},Finance{inconv, convenient},Social{problematic, nonprob, slightly\_prob},Health{priority, not\_recom, recommended}, and Class{recommend, very\_recom, spec\_prior, priority, not\_recom}”.( sci2s.ugr.es). Lastly there are no missing values and there are 12690 instances of data.

Exploratory Analysis

**Str**

The str function gives an overview of the nursing data framework. The output that is command gave is in the Appendix called figure 1. The output describes the type of variables in the data frame. All the variables were detected as factors and there were 12960 observations for 9 different variables. In addition, it also shows the number of variables and data in the data frame. Furthermore, it also details the main number of levels each variable has in their columns.

**Summary**

The summary portion gives the main statistics for each variable in the dataset.

The output for the nursing dataset was structured in a way where all the variables were separated, and sub categorized. The only interesting fact about this summary is that most of the numbers in the sub-categories were the same except for in the variable ‘class’. I do not believe they were duplicated, because the variable numbers on each section are different. However, if all the variable categories were the same number, I would be that duplication in the dataset could have been possible. I also ran a second command to figure the datatype of the variables and output was that they were all factors. This may be why some of the sub-category numbers are similar as shown in the str and summary report.

**Histogram**

The variable I decided to make the target was “class”. I chose class because class has multiple sub-categories that are very useful when it comes to determining if a person’s application has been placed in a recommended, non-recommended, spec priority, priority, or very recommended category when applying to a nursery school. The histogram in figure 3 shows the percentage of nursery applications by class status. From this histogram one can see that major of applications are placed in the not recommended category on a percentage scale of 0 to 30. Over 30% of applications are likely to be placed in non-recommended. Priority comes in second place because that category holds over 30% of applications as well. Third, spec\_prior is right at 30 % when it comes to the amount of nursery applications that are sent in for that category.

In contrast, the recommended and very recommended categories are more likely to have the least number of applicants placed in those categories. That is probably why so many people are constantly being rejected. The recommended section has an exceedingly small percentile that’s barely visible on the histogram. Furthermore, the very recommended section has a green percentile because those applications that are sent in are usually accepted rather than the rest, however very few applications are placed on that list.

**Stacked Bar Plot**

To further this topic, my next question was what might be causing people’s applications to be placed into these different categories? The stacked bar plot in the Appendix Figure 4 displays the number of applications by children and by the target variable ‘class’ with the sub divisions of the non-recommended, prioritiy, recommended, spec\_priori, and very recommended. The number of children are highlighted in different colors with a legend for viewing. The color black is for people with one of kid, pink is for people with two kids, green is for people with three kids, and blue is for people with more than 3 kids.

Furthermore, on the stacked bar chart it shows the running amount of applications from 0 to 4000. It appears people with only one child are more likely to be placed into the recommended or very recommended class than people with more than one child. In addition, people with three or more children are not usually put into the recommended or very recommended class. This is remarkably interesting because that means that if a person has more than three children it’s not likely that they will have a good chance at being accepted into the nursery school.

**Faceted Box Plot Part 1**

The previous findings brought to my third question “If people that have three or more kids are not usually placed in the recommended or very recommended classes, can parent status’s make a difference?” Figure 5 in the Appendix is a faceted box plot that shows the number of children, the category of the parents, and the class categories that they are all put into when filing out applications. The figure displays the number of children parents have and what the parent was labeled as. The goal is to find that if a parent is placed into usual, pretentious, or great\_pret category if that has something to do with them being on the recommended or very recommended class.

Furthermore, from looking at the information on the box plot, you can see that parents that are in the pretentious great\_pret categories are not likely to have their applications placed in the recommended category. While parents that are in the usual category are more likely to have their application put on the recommended category. It is very hard for an individual with more than 1 child to be placed in that list unless they are a usual.

In addition, the ones that do get placed on the recommended only have one child filling out an application. Also, if you look at the very recommended class it is possible for a parent with 1 or more children in the usual or pretentious group to place in that class.

**Faceted Box Plot Part 2**

Out of curiosity, the previous section made me ask one last question. That question was “If a parent category plays a major role in determining if a child is placed on the recommended or very recommended list, then what about finances? What role does that variable play when being placed into a class?” I made another faceted box plot labeled as Figure 6 just to view this information. The finance variable is categorized into convenient and inconvenient. Surprisingly, it is shown that a person who has one or two kids and their finances are inconvenient they might still be placed into the very recommended application section. This may be because they know a facility member and not out of the random since people are with two kids are more than likely placed in categories that are not recommended or very recommended.

However, if they have three or more and their finances are inconvenient, they would not be placed into the very recommended section. For people with three or more kids with convenient finances they are placed into the very recommended section. It is also notable that for recommended, which is the hardest category to be placed in, a person with inconvenient finances regardless of the number of children will not be placed in the recommended class. If a person with convenient finances and one child applies, they are highly likely to be placed into the very recommended section.

Preprocessing

For pre-processing this dataset was very convenient because all the values were already set as nominal. There were no duplicated entries. Likewise, no missing values, outliers, or discrepancies in the values as the Data Pre-processing using R article talks about. I did not have to do any data reduction ,because all the data that was there was needed to make the histograms, box plots, and bar plots for the explanatory analysis. All the data was needed to make the insights as detailed as possible. As far as data discretization all the values according to the dataset description were categorized as factors so no discretization was needed to make the rules that will mention in the sections to come. There were some parameters during the rules process that were used for further testing. However, the dataset was presented in a very manageable format.

Algorithm Intuition

“The Apriori rules method requires all variables in the dataset to be discrete, or factor”(Association Rules). All the variables in the dataset had factors as listed in the data frame so the logic of finding an algorithm was simple. Since my dataset were already factors, I did not have to use a regular discretization method instead I used a factor function discretization method. This method was only help for the variable ‘class’.

The logic of me using the factor function on the variable ‘class’ was because all of the other variables rely on that one variable for application determination. With the equation I was able to summarize the amount of people non-recommended=4320, priority= 4266, recommend= 2, spec\_prior= 4044, and very\_recom 328. This information is helpful because it proves that many people’s nursery applications make it to the recommended list and majority are non-recommended.

Likewise, for the Apriori algorithm, I used two key parameters based off the number of rules that were given after the rules generation. The key parameters of the algorithm were set to support 0.1, confidence 0.9 and support 0.2 , confidence 0.8. The reason I set the parameters like this is because only 24 rules were generated in the output after the rule’s inspection command was completed. I wanted to test the confidence and support high and low to see if there was a noticeable difference when the key parameters of the algorithm changes.

Rules Generation

Instead of just jumping to the results there were key steps and activities that were performed to extract the rules are directed in the following information below. The step I took was to create a command that could inspect all the rules. The command that could inspect the rules is shown below

rules<-apriori(nursery)

inspect(rules)

The second step that I took to extract the rules was pruning the information at hand. I used the code below to prune the dataset and ask for the redundant information.

#pruning

rules.sorted <- sort(rules, by="lift")

inspect(rules.sorted)

subset.matrix <- is.subset(rules.sorted, rules.sorted)

subset.matrix[lower.tri(subset.matrix, diag=T)] <- F

redundant <- colSums(subset.matrix, na.rm=T) >= 1

which(redundant)

The third and last step that I took was to use a code that could remove all of the redundant information and then to inspect the information that was left. Then I used a summary command to make sure the rules pruned were not visible anymore

#remove redundant rules

rules.pruned <- rules.sorted[!redundant]

inspect(rules.pruned)

#rules pruned summary

summary(rules.pruned)

For experimental reasons and to make sure the model preforms highly, I used two parameter tuners. The code that I used for both of them are below:

#Apriori method 1

rules.pruned <- apriori(nursery, parameter= list(supp=0.1, conf=0.9))

inspect(rules.pruned)

#Apriori method 2

rules.pruned <- apriori(nursery, parameter= list(supp=0.2, conf=0.8))

inspect(rules.pruned)

**Result**:

Output

After inspection, pruning, removing redundancy, and using two key parameters for Apriori algorithm equations for the rules the output was rules were proven significant. The output after viewing the Apriori rules on the R script displayed that their were 24 rules in the dataset given. The beginning confidence was 0.8 and support was 0.1. The support accounted for 1292 items in the dataset. That means that the rules given are more than likely accurate. The next step was inspecting the rules in the data frame. From this the output on the left-hand side stated that the variable class = not recommended and the right hand said stated that health=not recommended also. The support interval was 0.33, confidence 1, and the lift was at 3. With all three of the main metrics high that means my original objective of certain factors was the cause of many applications being rejected is true. It was later discussed in my data analysis that the variable ‘class’ is the leading target variable that will lead to someone being rejected or not. This information checks my objective because class=not-recommended was the strongest rule to show on the rules output.

Next, the pruning and removing the redundancies showed that only one rule in the data was not redundant and could be used for great analysis. That rule was the very same one as mentioned above. Only the first rule once again was the main determining rule for if one would be accepted or rejected when it comes to submitting nursery applications. In addition, I also made a summary for rules pruned and the summary also mentioned the same confidence, lift, and support intervals as before. Furthermore, the key parameters with confidence 0.1, support 0.9 on rules pruned showed that 24 rules were found. I then decided to scale the parameters down to confidence 0.2, support 0.8. Two rules were found and the first was also the same as mentioned above and the second one was a mirror of the first. My objective to find the main variables help with determining whether a person will be rejected or not was found, and it is the variables class and health when placed on not recommended.

Rules Properties

The characteristics of my rule’s properties will be explained by using three different models. The first model that I built was a graph, second one was a parallel graph, and third one was a scatterplot. The leverage of using these three functions in model form is to further verify the confidence and support levels of my previously stated rules. To prove that there are two main rules that are similar and to demonstrate that the variables class and health correlate with one being not recommended for approval in the nursery school.

In figure 7, there is a graph to leverage the rules. The graph shows that the confidence between the two plots is linear. Which means there is a strong correlation between class and health being not recommended for application approval. In addition, the lift is also at three and since the shade is deep orange that means that the lift is very proven to be accurate. Also, support is at 0.33 while confidence is at one so that means all the measurements are fairly accurate. On the diagram it shows a square and the square demonstrates how far or connected each variable’s relationship is to not recommend. The results are category not recommended is more than likely to happen in case due to class or health according to the graph plot.

In figure 8, there is a parallel diagram that explains the correlation between the categories. In the parallel diagram there is two different arrows. On the bottom of the diagram the left side shows number one and the right side shows rhs. You can also see that health equals not recommended on the top corner, however the arrow does point down and you can also see that class equals not recommended is on the bottom, however the confidence arrow is rising. From this diagram you can see that the coordinates are parallel between both rules. While observing this diagram it can also be said that if health or class is being affected in a certain way, that determines whether a person will be placed in recommended or not recommended.

In figure 9, there is a scatterplot for the top two rules. The plots as you can see are linear to each other. It appears one of the supports are rising above the confidence and the 0.331 confidence level, while the other rule has an extremely high confidence level, but not as a high support level. The scatterplot to the right that is in the top corner is the class equals non-recommended rule and the scatterplot to the left at the bottom is the help non- recommend it rule. This leveraged function is greatly beneficial, because in the previous diagram you could see that the correlations were parallel to each other. Health equals not recommended was above class equal non recommended and the arrow is pointing downward while classes arrow was pointing upward. So, this is a particularly good representation of why class equals non-recommended is the main determining factor for if a person’s application gets rejected or not.

Evaluation

Some additional metrics measures were to use the interest measures command. With this command I was able to see the confidence, support, and lift rule data on a more statistical scale. This this was also evaluated as equivalent to the rule’s summary as well. Likewise, I also used two more models for final evaluation.

The grouping model shown as figure 10 in the Appendix groups the rules into two separate categories and two separate plots on the diagram. In a group diagram the larger the circle means the higher the support. However, in this diagram you can see that both two different rules are basically parallel to each other. The only difference is that class equals not recommend it with the right-hand result of health equals not recommended is on the more positive and higher spectrum then the other plot is.

Furthermore, the matrix model shown as Figure 11 in the Appendix, the top hand side or all the items sets for the left-hand side of the rules which is class equals not recommended to health equals not recommended. Whereas on the bottom right-hand side that is where it is health equals not recommended to class equals recommended. The lift is shaded orange, and it is three so it is very probable and more than likely very accurate.

**Conclusion**

Summary

To close, some of the key findings here that relate to the stated objective is that the main variables class and health are the number two determinist for if a nursery application is rejected or not. Likewise, we were also able to view knowledge on why people are placed into certain classes. From some of the models in the diagram we could conclude that the type of parent, number of children, and finance can have an impact on if a person is placed into a recommended or non-recommended class for nursery school acceptance. It was also found that health is a major determining factor when being placed into a class category as well.

Limitations

The limitations of the analysis were the data. The data was all factorable values and when trying to make simple box plot, histograms, or stacked bar charts it was difficult because the factor items can not be placed numerically unless certain packages are installed first. Another limitation was doing the evaluation metrics of the model. I believe I need more detailed information besides the information that the grouping and metrics model showed. Maybe if there were other models that could give more detailed insights that would help it to have been more explanatory.

Improvement Areas

Some areas for potential future improvements would be finding a dataset with a variety of different type of values besides just factors to challenge myself more. I could improve on the explanation of rule properties by manipulating rules into more advanced models. Likewise. with the evaluation section I could find more detailed models to evaluate the data with.

Appendix

**Output Str Figure 1.**

str(nursery)

'data.frame': 12960 obs. of 9 variables:

$ parents : Factor w/ 3 levels "great\_pret","pretentious",..: 3 3 3 3 3 3 3 3 3 3 ...

$ has\_nurs: Factor w/ 5 levels "critical","improper",..: 4 4 4 4 4 4 4 4 4 4 ...

$ form : Factor w/ 4 levels "complete","completed",..: 1 1 1 1 1 1 1 1 1 1 ...

$ children: Factor w/ 4 levels "1","2","3","more": 1 1 1 1 1 1 1 1 1 1 ...

$ housing : Factor w/ 3 levels "convenient","critical",..: 1 1 1 1 1 1 1 1 1 1 ...

$ finance : Factor w/ 2 levels "convenient","inconv": 1 1 1 1 1 1 1 1 1 2 ...

$ social : Factor w/ 3 levels "nonprob","problematic",..: 1 1 1 3 3 3 2 2 2 1 ...

$ health : Factor w/ 3 levels "not\_recom","priority",..: 3 2 1 3 2 1 3 2 1 3 ...

$ class : Factor w/ 5 levels "not\_recom","priority",..: 3 2 1 3 2 1 2 2 1 5 ...

**Summary Output Figure 2.**

summary(nursery)

parents has\_nurs form

great\_pret :4320 critical :2592 complete :3240

pretentious:4320 improper :2592 completed :3240

usual :4320 less\_proper:2592 foster :3240

proper :2592 incomplete:3240

very\_crit :2592

children housing finance

1 :3240 convenient:4320 convenient:6480

2 :3240 critical :4320 inconv :6480

3 :3240 less\_conv :4320

more:3240

social health class

nonprob :4320 not\_recom :4320 not\_recom :4320

problematic :4320 priority :4320 priority :4266

slightly\_prob:4320 recommended:4320 recommend : 2

spec\_prior:4044

very\_recom: 328

**Figure 3.**

**Histogram:** Percentage of Nursery Applications by Class Status

**Chart, bar chart, waterfall chart

Description automatically generated**

**Figure 4. Stacked Bar Plot**

**Chart, bar chart

Description automatically generated**

**Figure 5. Box Plot – Faceted**

Chart, calendar

Description automatically generated

Figure 6: Box Plot Faceted – Finances Categorized by Class and Amount of Children

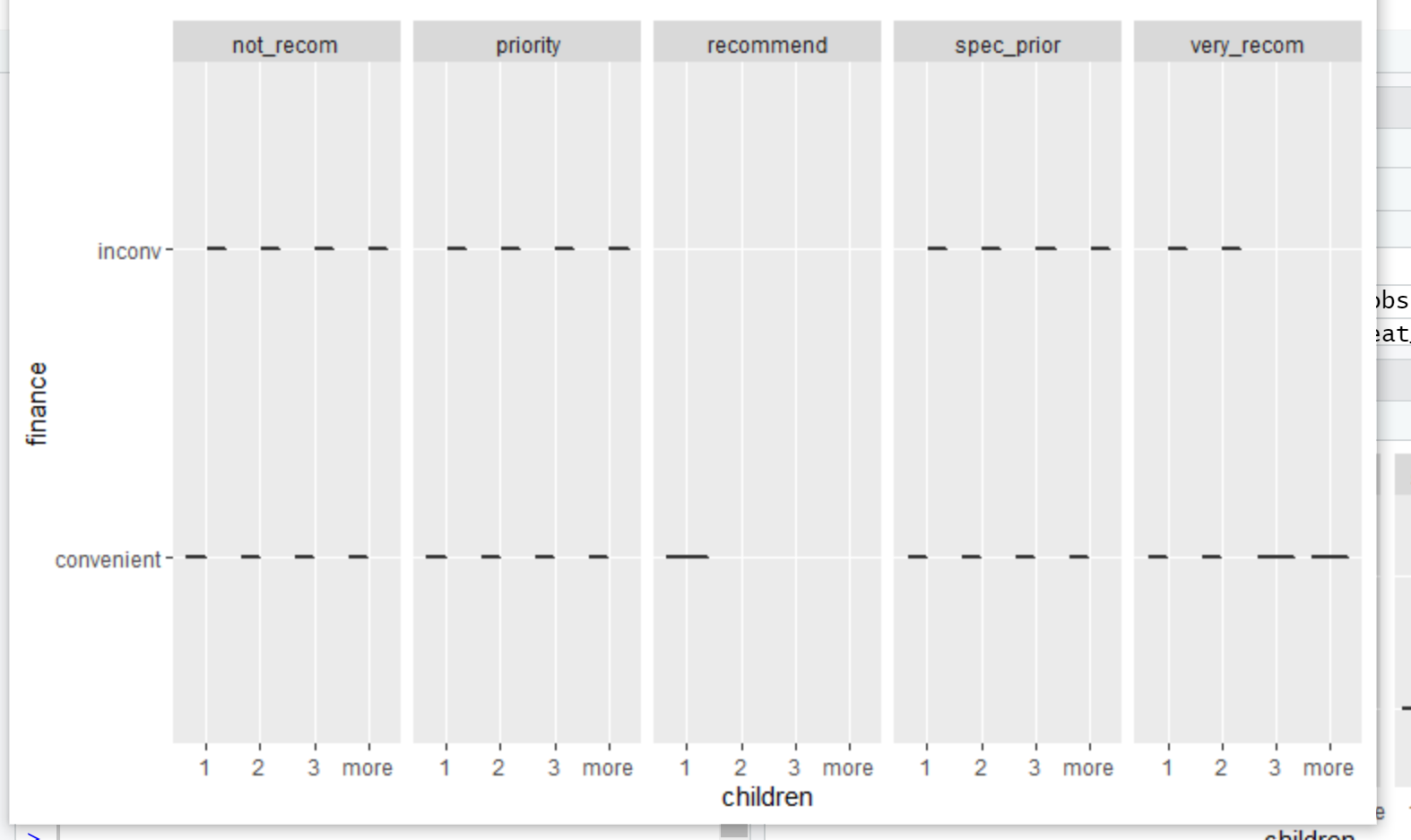


Figure 7. Graph

A picture containing diagram

Description automatically generated

Figure 8. Parallel Graph

Chart, line chart

Description automatically generated

Figure 9. Scatter Plot

Table

Description automatically generated with medium confidence

Figure 10 . Grouping Model

Chart, box and whisker chart

Description automatically generated

Figure 11. Matrix Model

Chart, bar chart

Description automatically generated

References

KEEL: Software tool. Evolutionary algorithms for Data Mining. Welcome to the SCI2S web site. (n.d.). https://sci2s.ugr.es/keel/dataset\_smja.php?cod=1376.

UMGC. (n.d.). Association Rules R. Aldephi.

175 million children are not enrolled in pre-primary education. UNICEF. (2021, June 8). https://www.unicef.org/press-releases/175-million-children-are-not-enrolled-pre-primary-education-unicef.