Association Rule To fill missing values

Yoav Eliav

Abstract

Missing values is common problem when analyze an data set. The reasons behind missing values in datasets can be diverse, and the approach we take to address them can greatly impact machine learning outcomes.The approach suggested in this article is to fill missing values by utilizing association rule algorithms alongside conventional methods, such as using the most common value for categorical attributes and the mean for numerical attributes. The efficacy of the proposed approach was tested on four different datasets and showed only marginal improvement compared to traditional methods. Also the use of association rule algorithms was found to significantly increase the running time.

Problem description

Data preparation is a crucial step in the data science pipeline. One of the most common problems encountered during data preparation is missing values. missing values is defined as values that not present in the data for some attribute. The reason for this can arise from a variety of sources, for example:

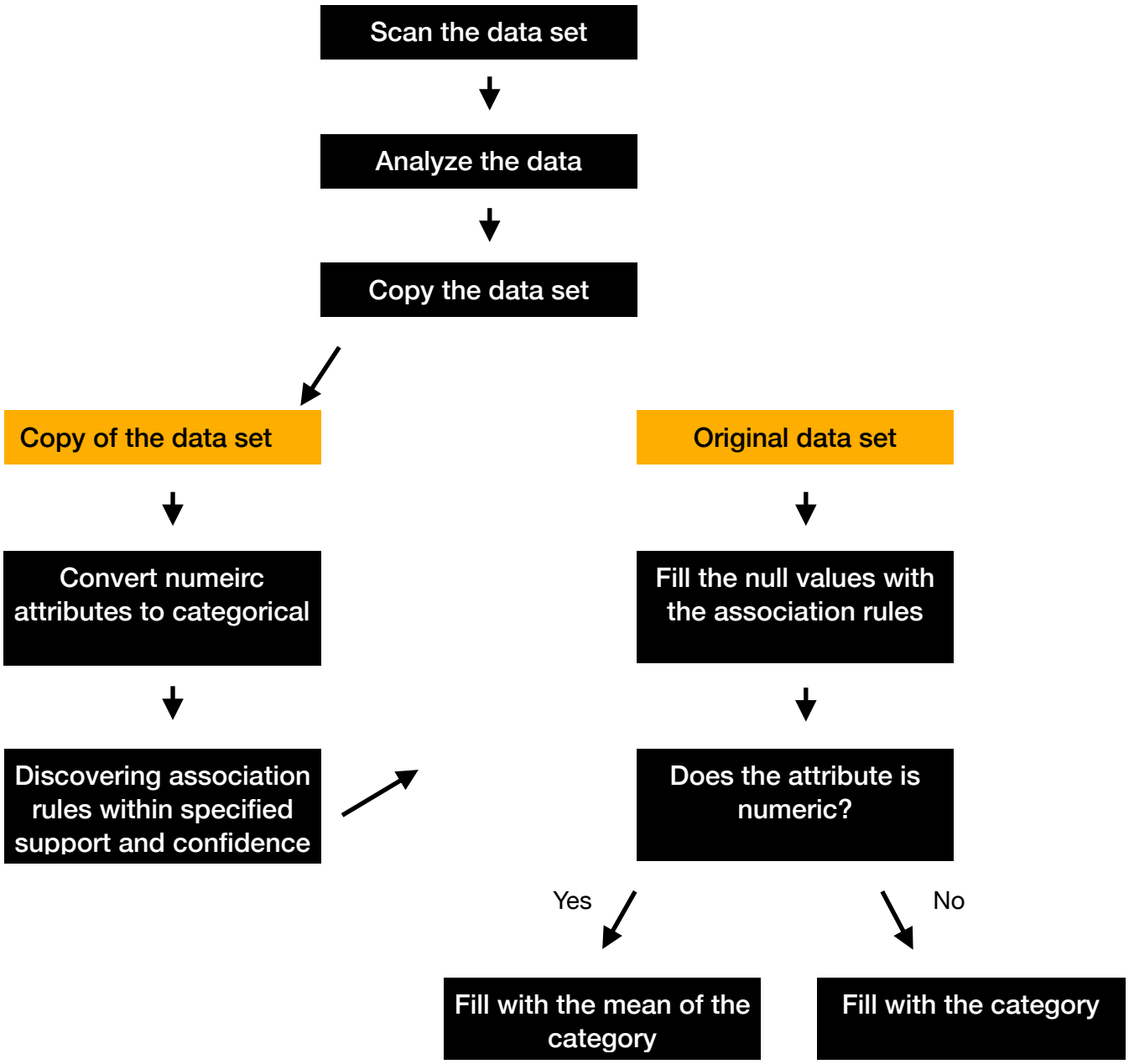
- **Sensitive data** - Certain data may be considered sensitive and the individual may opt to leave it blank.
- **data corruption** - The data may become damaged or degraded for various reasons like Technical issues, Cyberattacks, etc...
- **Hard-to-find** - some of the information may be hard-to-find and require extensive effort to secure.

missing values and how we deal with them can have big impact on machine learning models outcomes. The traditional methods for filling missing values, such as using the mean or median for numerical attributes and the most common value for categorical attributes, can be limited in their effectiveness and may lead to incorrect results.

Solution overview

In this article, I propose a solution to address missing values in the data set. The solution involves utilizing the Apriori association rule algorithm to fill in the missing values. In order to find the appropriate rules, some numerical attributes in the dataset are first converted into categorical attributes. The missing values are then filled with the appropriate category based on the rules found using Apriori. In cases where the attribute is numeric, the missing values are filled with the mean of the category. However, as Apriori cannot guarantee the complete filling of all missing values, i combined the Apriori algorithm with conventional methods, such as using the mean for numerical attributes and the most common value for categorical attributes.

Solution Process



Experimental evaluation

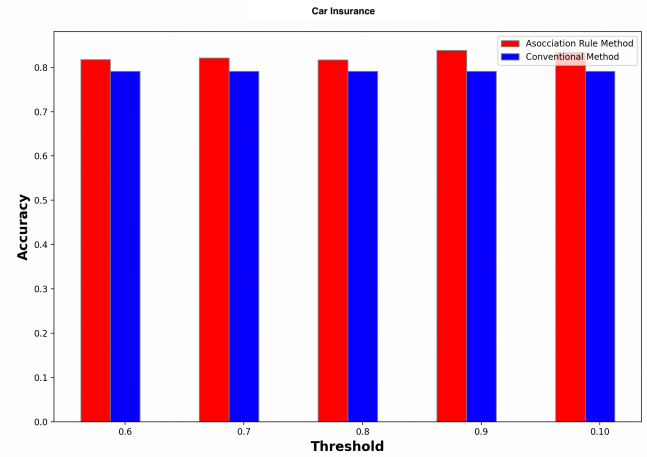
For the experimental evaluation, i compared the performance of the association rule method against conventional methods using 4 diverse datasets. To measure the effectiveness, i varied the thresholds and evaluated the accuracy of a Logistic Regression Model in three datasets and the r2 score of a Linear Regression Model in one dataset. i run the two methods multiple times and checked the mean of their performance.

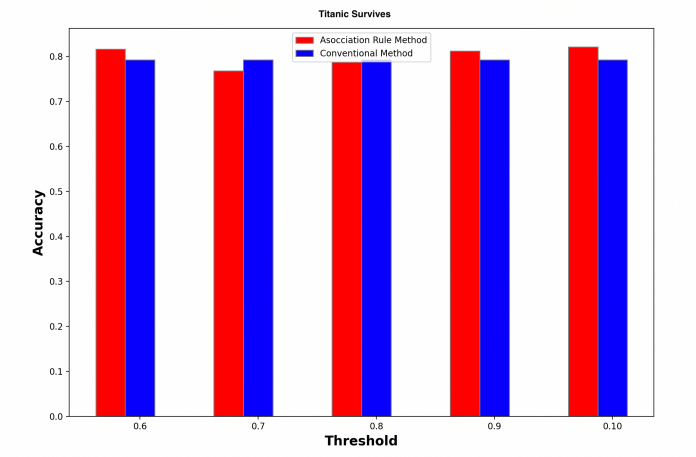
the four data sets are:

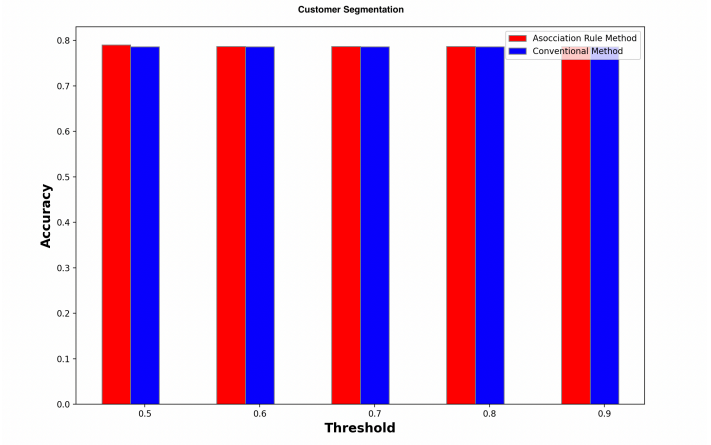
- **Car insurance:** contains information about customers and aims to predict if a customer has made a claim on their insurance policy (represented by a target of 1) or not (represented by a target of 0).
- **Used car price:** the used car pricing dataset provides information on the sales and condition of previously owned vehicles. The aim was to predict the selling price of the cars
- **Titanic Dataset:** features information on individuals aboard the Titanic, with the objective of predicting the survival of passengers.
- **Customer Segmentation:** contains information about customers and aims to predict their spending score.

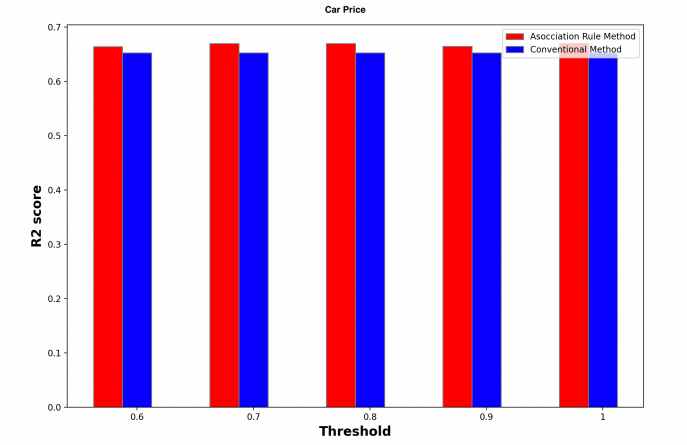
experimental visualization

Accuracy

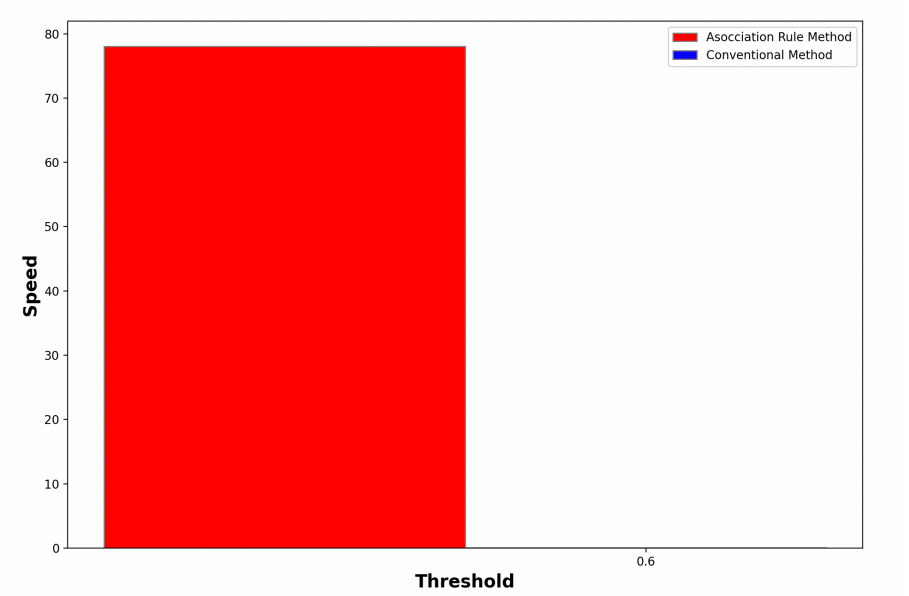








Processing Time



Experimental Results

As we can see from the graphs, using association rule algorithm give minor improvement compared to traditional methods such using the mean or the must common , Also the efficciency of the algorithm depends on the treshold, In datasets like the Customer Sementaion, threshold values above 0.5 showed little to no effect on the prediction model. furthermore, in the Titanic data set, If the threshold was not set correctly, the accuracy was negatively impacted and even caused to the association rule method to less accurate compare to the traditional methods. in my research, i found that confidence threshold around 0.6 generates good results.

Another contributing factor to the effeciency of the algorithm was how to convert the numeric attributes to categorical, the method of dividing the numeric attributes into bins led to the discovery of diffrent rules casing variations in the filled valued for the missing values. in attributes like "age" the problem was less noticeable since its easy to divide age attribute into categories, but in numeric attribute without established guidelines, the way of dividing into bins greatly influenced the results of the algorithm.

Another issue is that the apriori algorithm is slow, hindering the ability to conduct more precise and comprehensive tests. furthermore in data sets like "Car price" with large value ranges in certain attributes, the only way to effectively run the apriori algorithm was to remove some of the attributes, this type of problem is much less noticeable with the traditional methods.

The slow speed of the algorithm also impacts the amount of data that must be processed from the dataset, processing too much data can be time-consuming, while processing too little may limit the ability of the algorithm to uncover rules effectively.

Related work

Except from the traditional methods like using the mean or the must often value, other method to deal with null values is the "K-nearest nighbors" algorithm, the knn algorithm is a supervised machine learning technique that is used for both classification and regression tasks, in the context of handling missing values, the knn fill the null value by finding the k-closest data point, The value of 'k' is specified by the user and can have big impact on the accuracy of the model.

The diffrence between the association rule method and knn is that association rule using patterns in the data to fill the null values while knn use the concept of proximity. Knn has the advantage of being quicker and more efficient with large datasets compared to Association Rules. Further more, knn can hadle with numeric and categorical attributes which as we saw in the Experimental Results the converting from numeric to categorical can cause negative impact. Despite this, Association Rules may be a more suitable option for datasets with discernible patterns.

Conclusion

In conclusion, it seems that to the association rule method have potential in addressing missing values in datasets. However its efficacy depends on the data set, although in the research i did not find the association rule method to be inferior compare to the traditinional methods. Its computational intensity and prolonged processing time could be considered a drawback.

Also i found that the way of pre processing the train data before finding the rules i very important and could have a drastic effect.