Market Basket Analysis Report

Discovering Survival Patterns in Titanic Crash

Ruchir Malik

July 26, 2024

INTRODUCTION

Market Basket Analysis (MBA) is a data mining technique used to discover patterns or associations between items in large datasets, typically transactional data from retail settings. It aims to identify the likelihood of items being purchased together, enabling businesses to understand customer purchasing behaviours and improve decision-making processes.

Traditionally used in retail to discover associations between products purchased together, MBA can be applied to a variety of fields to uncover valuable insights. In this project, I will utilize MBA on the famous Titanic dataset to explore patterns and associations among the passengers' attributes and survival outcomes.

The Titanic dataset, provided by Kaggle, contains detailed information about the passengers on the ill-fated RMS Titanic. The dataset includes variables such as age, gender, ticket class, fare, and survival status. By applying MBA to this dataset, I aim to identify interesting relationships between these attributes that may offer deeper insights into the factors influencing survival.

This analysis will help me:

*1. **Uncover Hidden Patterns**: Identify associations between different passenger attributes and their survival rates, providing a deeper understanding of the factors affecting survival.*

*2. **Enhance Predictive Models**: Improve the accuracy of predictive models by incorporating discovered associations, leading to better predictions of survival based on passenger characteristics.*

*3. **Gain Analytical Skills**: Develop my data mining and analysis skills by applying MBA techniques to a non-traditional dataset, broadening my expertise in data science.*

By leveraging MBA on the Titanic dataset, I hope to demonstrate the versatility of this technique beyond its conventional retail applications and gain valuable insights into one of the most studied historical datasets.

OBJECTIVES

Primary Objective*: Identify patterns and associations between passenger attributes and their survival outcomes.*

Secondary Objectives*:*

- *Enhance understanding of factors influencing survival on the Titanic.*
- *Improve predictive models for survival based on identified associations.*
- *Develop and refine data mining skills through the application of MBA techniques.*

DATA COLLECTION

- ***Data Source****: The Titanic dataset from Kaggle.*

- ***Data Description****: The dataset is a representation of the famous Titanic Machine Learning dataset. It includes 891 records of passengers with variables such as FirstClass, SecondClass, Female, AgeMissing, Child, IsSolo, IsCouple, HasChild, etc. Each column is a characteristic of a Titanic passenger and has been engineered to hold boolean values i.e. 0 or 1. Each row represents a passenger whose attributes or characteristics are defined by different columns.*

METHODOLOGY

- ***Techniques Used****: The Apriori algorithm was used to identify frequent itemsets and generate association rules.*
- ***Tools and Software****: Excel was used for the analysis and Tableau for visualisations.*

ANALYSIS AND RESULTS

Frequent Itemsets and Association Rules

A comprehensive Market Basket Analysis (MBA) was conducted on the Titanic dataset to uncover significant patterns and associations among passenger attributes and their survival outcomes. The Apriori algorithm was employed with a minimum support of 0.05. Additionally, Excel Solver was utilized to determine the maximum and minimum lift values for specific associations.

Key Findings

Maximum Lift Association:

- ***Association:*** *First Class, Female, Survived*
- ***Lift Value:*** *2.522116461*
- ***Interpretation:*** *Female passengers in first class exhibited a significantly higher likelihood of survival compared to the general population. This compelling association underscores the substantial impact of both gender and socioeconomic status on survival chances.*

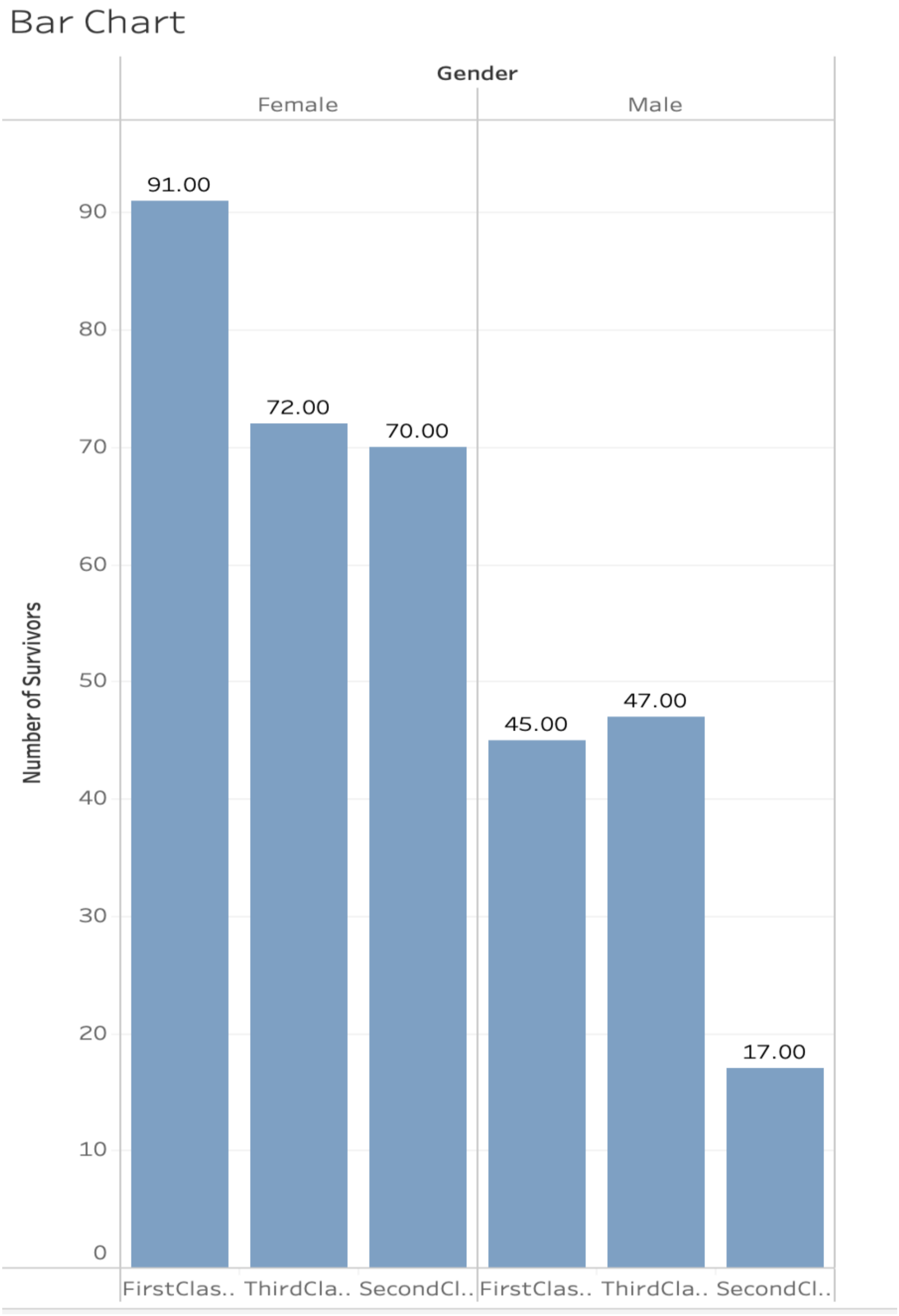
Minimum Lift Association:

- ***Association:*** *Third Class, Is Solo, Survived*
- ***Lift Value:*** *0.521052632*
- ***Interpretation:*** *Passengers traveling alone in third class demonstrated a notably lower probability of survival. This three-way association highlights the detrimental influence of both ticket class and travel status on survival outcomes.*

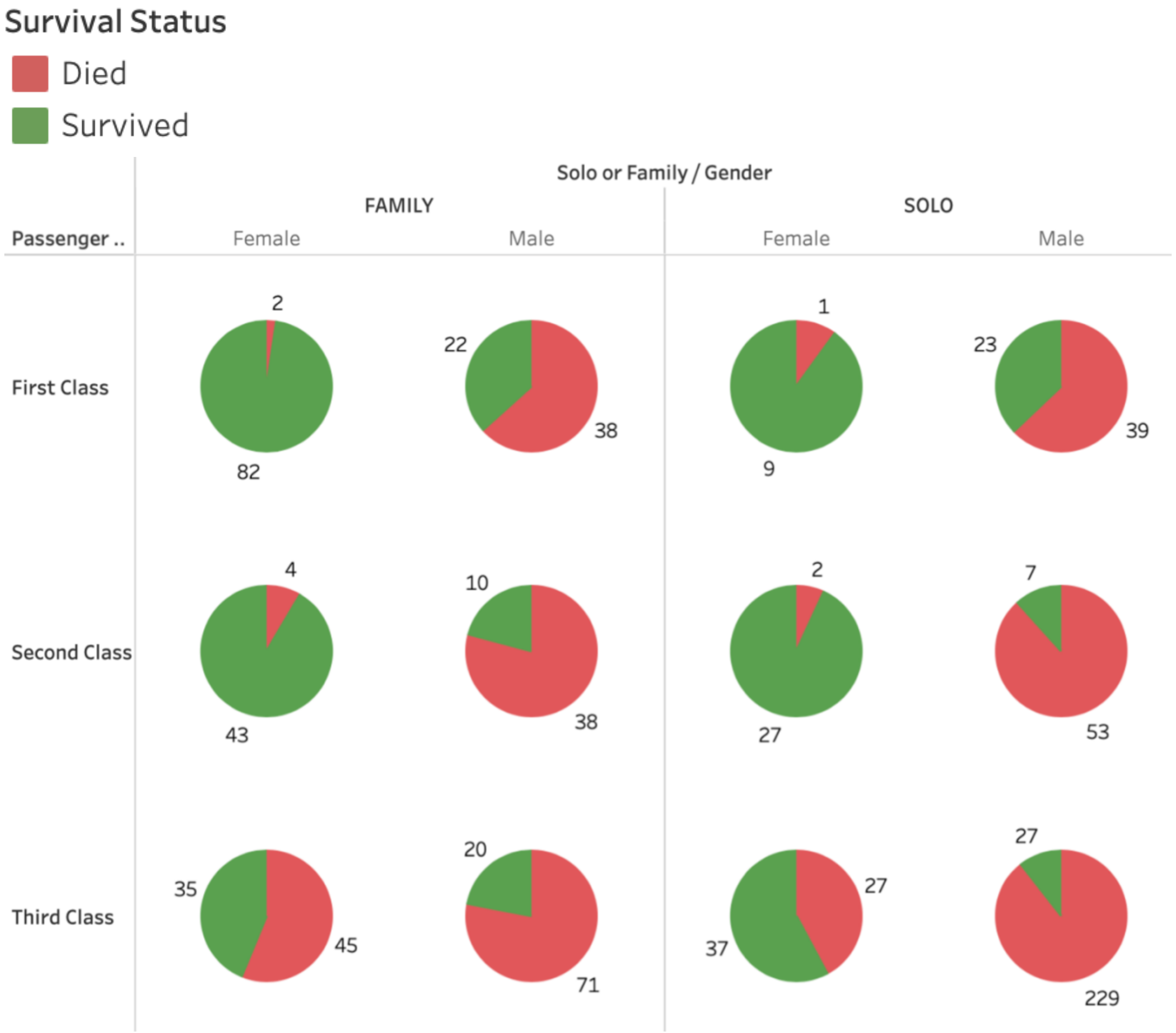
Visualizations

To enhance understanding and communication of the results, various visualizations were employed:

- ***Bar charts:*** *Depicting survival rates across different passenger classes and genders.*



- ***Pie Charts:*** *Depicting number of female and male survivors travelling solo vs with family in each class*



These visualizations effectively convey the complex interplay of factors influencing survival on the Titanic.

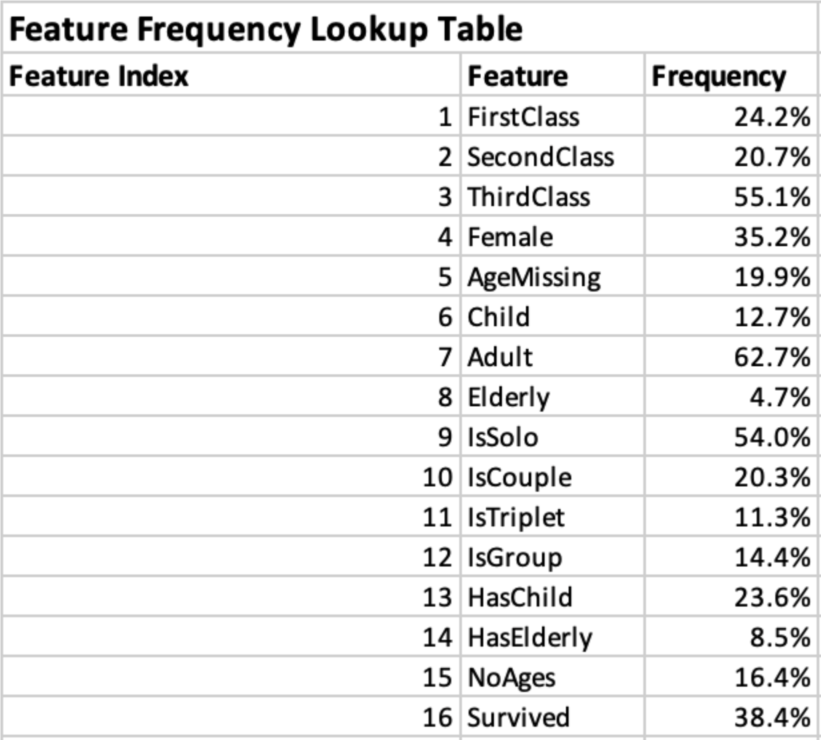
DISCUSSION

Interpretation of Results

To make our calculations easier, I first created a few variables and tables. For instance, there's a cell (S2) that contains the total number of rows (891), which is basically the total number of passengers in the dataset. This will come in handy in future calculations.

Total # of Passengers:	891

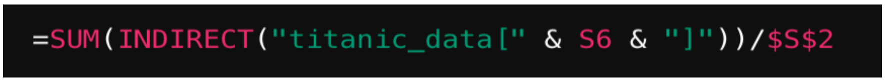
Second, there's a feature frequency lookup table that holds the frequency values for each feature/column in the dataset. There's 16 features in this particular version of the Titanic dataset.



Feature Frequency Lookup Table

Feature Index	Feature	Frequency
1	FirstClass	24.2%
2	SecondClass	20.7%
3	ThirdClass	55.1%
4	Female	35.2%
5	AgeMissing	19.9%
6	Child	12.7%
7	Adult	62.7%
8	Elderly	4.7%
9	IsSolo	54.0%
10	IsCouple	20.3%
11	IsTriplet	11.3%
12	IsGroup	14.4%
13	HasChild	23.6%
14	HasElderly	8.5%
15	NoAges	16.4%
16	Survived	38.4%

The frequency in the above table has been calculated using the following general formula:



```
=SUM(INDIRECT("titanic_data[" & S6 & "]"))/$S$2
```

Components of the formula:

INDIRECT("titanic_data[" & S6 & "]")

- ***S6**: This is a cell reference. The cell S6 contains a string value, which represents the name of a column in the titanic_data table. For example, 'S6' contains the string 'FirstClass' and therefore the formula will reference the 'FirstClass' column in the 'titanic_data' table.*
- ***"titanic_data[" & S6 & "]"**: This concatenates the table name titanic_data, the column name stored in S6, and the closing bracket] to form a reference to a specific column in the titanic_data table.*
- ***INDIRECT**: This function takes a text string and converts it into a valid cell or range reference. So, INDIRECT("titanic_data[" & S6 & "]") will convert the string into a reference to the column in titanic_data whose name is specified in cell S6.*

SUM(INDIRECT("titanic_data[" & S6 & "]"))

- ***SUM**: This function calculates the sum of the numbers in a specified range. Here, the range is determined by the INDIRECT function. So, it sums all the values in the column of the titanic_data table referenced by the INDIRECT function.*

- ***/$S$2***

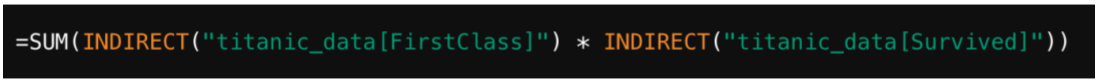
  ***$S$2**: This is an absolute cell reference to cell S2. The dollar signs ($) indicate that the reference will not change if the formula is copied to other cells.*

Similarly, the frequency percentages for the remaining 15 features can be calculated by replacing S6 with their respective cell references. I decided to use Excel's fill function to speed up the process.

Explanation of the 2-way Lift Table

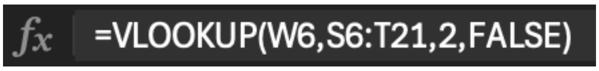
This table is used in Market Basket Analysis to understand the relationship between various features (LHS) and the outcome of "Survived" (RHS). Here's a breakdown of the columns and what they signify:

- ***LHS (Left-Hand Side)**: This column lists the different features or conditions (e.g., FirstClass, SecondClass) whose association with survival is being analysed.*
- ***RHS (Right-Hand Side)**: This column lists the outcome "Survived," which is the target of the analysis.*
- ***Occurrences**: The number of times the LHS condition and the RHS outcome both occur in the dataset. I used the following formula to calculate occurrences for each feature (LHS) with Survived (RHS):*



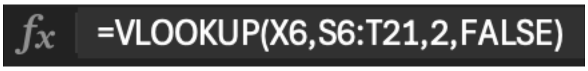
```
=SUM(INDIRECT("titanic_data[FirstClass]") * INDIRECT("titanic_data[Survived]"))
```

- ***LHS Freq (Left-Hand Side Frequency)**: The proportion of the dataset that has the LHS condition. This is calculated as the number of occurrences of the LHS condition divided by the total number of transactions (891 in this case). I simply used the following VLOOKUP function to extract values from the Frequency Lookup Table:*



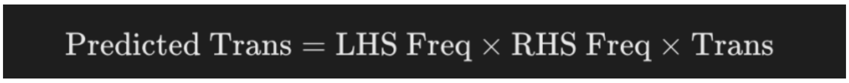
```
=VLOOKUP(W6,S6:T21,2,FALSE)
```

- ***RHS Freq (Right-Hand Side Frequency):** The proportion of the dataset that has the RHS outcome ("Survived"). This is calculated as the number of occurrences of "Survived" divided by the total number of transactions. A similar VLOOKUP function was used to calculate these values:*



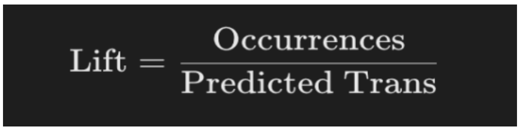
```
=VLOOKUP(X6,S6:T21,2,FALSE)
```

- ***Trans (Transactions)**: The total number of transactions (891 in this dataset).*
- ***Predicted Trans (Predicted Transactions)**: The expected number of transactions where both the LHS condition and RHS outcome occur, assuming they are independent. It is calculated as:*



$$\text{Predicted Trans} = \text{LHS Freq} \times \text{RHS Freq} \times \text{Trans}$$

- ***Lift**: This measures how much more likely the LHS condition and RHS outcome occur together compared to if they were independent. It is calculated as:*



$$\text{Lift} = \frac{\text{Occurrences}}{\text{Predicted Trans}}$$

Interpretation of Lift Values

- ***Lift > 1***: *Indicates a positive association, meaning the LHS condition increases the likelihood of the RHS outcome.*
- ***Lift = 1***: *Indicates no association, meaning the LHS condition does not affect the likelihood of the RHS outcome.*
- ***Lift < 1***: *Indicates a negative association, meaning the LHS condition decreases the likelihood of the RHS outcome.*

Examples from the Table

1. FirstClass:

- Occurrences: 136

- LHS Freq: 0.242424242

- RHS Freq: 0.383838384

- Predicted Trans: 82.90909091

- Lift: 1.640350877

- Interpretation: FirstClass passengers are 1.64 times more likely to survive compared to if the class and survival were independent.

2. Female:

- Occurrences: 233

- LHS Freq: 0.352413019

- RHS Freq: 0.383838384

- Predicted Trans: 120.5252525

- Lift: 1.933204827

- Interpretation: Female passengers are 1.93 times more likely to survive compared to if gender and survival were independent.

Additional Notable Entries:

3. IsGroup:

- Lift: 1.078741776

- Interpretation: Passengers traveling in a group are slightly more likely to survive compared to solo travellers.

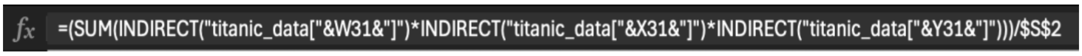
Conclusion

The lift values in this table provide insights into the strength of association between various features and the likelihood of survival. These insights can be used to identify patterns and make informed decisions based on the data.

Explanation of the 3-way Lift Table

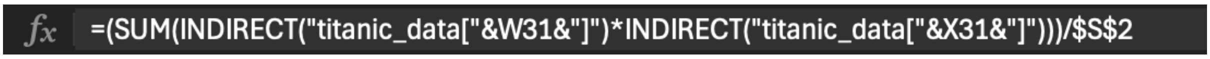
This table is used in Market Basket Analysis to understand the relationship between combinations of two features (LHS) and the outcome of "Survived" (RHS). Here's a breakdown of the columns and what they signify:

- ***LHS (Left-Hand Side):***
  Feature 1 Index: Index number of the first feature in the dataset.
  Feature 2 Index: Index number of second feature in the dataset.
- ***RHS (Right-Hand Side)**:*
  Feature 3: The outcome "Survived," which is the target of the analysis.
- ***Occurrences Freq**:*
  The joint frequency of occurrences where both LHS features and the RHS outcome occur together. Calculated as the proportion of the dataset where both conditions (Feature 1 and Feature 2) are present along with the outcome "Survived." Here's the formula I used to calculate occurrences frequency:



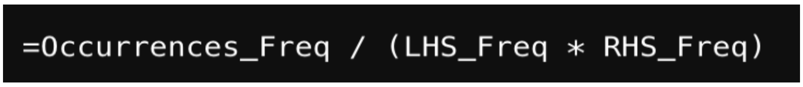
```
=(SUM(INDIRECT("titanic_data["&W31&"]")*INDIRECT("titanic_data["&X31&"]")*INDIRECT("titanic_data["&Y31&"]")))/$S$2
```

- ***LHS Freq (Left-Hand Side Frequency):***
  The proportion of the dataset that has both LHS conditions. This is calculated as the product of the individual frequencies of Feature 1 and Feature 2, divided by the total number of transactions.



```
=(SUM(INDIRECT("titanic_data["&W31&"]")*INDIRECT("titanic_data["&X31&"]")))/$S$2
```

- ***RHS Freq (Right-Hand Side Frequency):***
  The proportion of the dataset that has the RHS outcome ("Survived"). This is calculated as the number of occurrences of "Survived" divided by the total number of transactions. Similar to 2-way lift, I used the VLOOKUP function to get this value.
- ***Lift**:*
  This measures how much more likely the combination of the two LHS features and the RHS outcome occur together compared to if they were independent. It is calculated as:



```
=Occurrences_Freq / (LHS_Freq * RHS_Freq)
```

Examples from the Table:

1. Female & Elderly:

- Occurrences Freq: 0.01010101

- LHS Freq: 0.011223345

- RHS Freq: 0.383838384

- Lift: 2.344736842

- Interpretation: The combination of Female and Elderly passengers is 2.34 times more likely to survive compared to if gender, age, and survival were independent.

2. Other Combinations:

- LHS Freq and RHS Freq: Calculated similarly using the individual frequencies of the respective features and the survival rate.

- Lift: Varies based on the specific combination of features, providing insights into which feature combinations have a significant association with survival.

Conclusion:

The lift values in this 3-Way Lift table provide insights into the strength of association between various combinations of two features and the likelihood of survival. These insights can be used to identify patterns and make informed decisions based on the data, highlighting which combinations of features are most influential in determining survival.

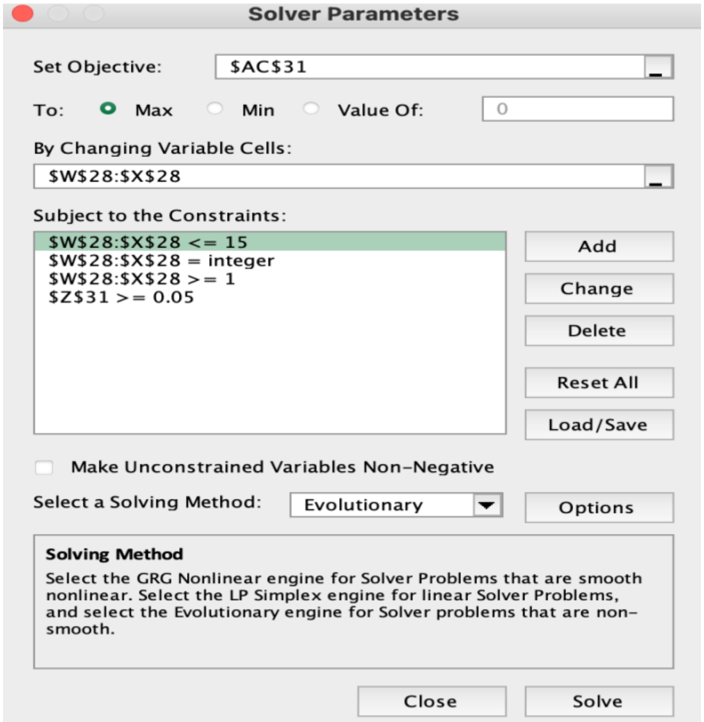
These calculations and interpretations can provide a detailed understanding of how various feature combinations are associated with the likelihood of survival in the Titanic dataset.

Alternate Method to calculate 3-Way lift:

1. ***Calculate occurrences:*** *Number of instances where all 3 features were true i.e. 1*
2. ***Calculate Predicted Transactions:*** *LHS Freq * RHS Freq * 891*
3. ***Calculate Lift:*** *occurrences/predicted transactions*

Either way, you'll get identical values for 3-way lift and is only a matter of personal choice.

I used Excel solver to iterate through all of the possible combinations. It enabled me to dynamically change LHS features based on their indexes and calculate updated lift values.



Solver Parameters

Set Objective: $AC$31

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:

$W$28:$X$28

Subject to the Constraints:

```
$W$28:$X$28 <= 15
$W$28:$X$28 = integer
$W$28:$X$28 >= 1
$Z$31 >= 0.05
```

□ Make Unconstrained Variables Non-Negative

Select a Solving Method: Evolutionary

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

CONCLUSION

Summary of Findings

The Market Basket Analysis (MBA) conducted on the Titanic dataset revealed several key associations between passenger attributes and survival outcomes:

1. **Strong Positive Associations:**
   - **First Class & Female:** Female passengers in the first class were significantly more likely to survive, with a lift value of 2.52. This highlights the strong influence of both gender and socioeconomic status on survival chances.
   - **Female & Elderly:** Elderly female passengers were 2.34 times more likely to survive compared to if gender, age, and survival were independent. This suggests a potential prioritization of elderly women during rescue operations.

2. **Negative Associations:**
   - **Third Class & Solo Travelers:** Passengers traveling alone in third class had a much lower likelihood of survival, with a lift value of 0.52. This points to the detrimental impact of both ticket class and lack of companionship on survival chances.
   - **Solo Travelers:** Passengers traveling alone, regardless of class, were less likely to survive, indicating the importance of group or family travel during the disaster.

3. **Other Notable Associations:**
   - **Group Travel:** Passengers traveling in groups had a slightly higher survival rate, suggesting that being part of a group may have provided some survival advantage.
   - **Female Passengers:** Across the board, female passengers were nearly twice as likely to survive compared to males, reinforcing the well-known "women and children first" protocol.

These findings underscore the significant influence of social and demographic factors, such as gender, age, class, and travel companionship, on survival outcomes during the Titanic disaster.

Limitations

It is essential to acknowledge the limitations of the study. The analysis is constrained by the dataset's scope and potential biases, such as missing values and inherent sampling limitations. Additionally, the MBA focuses on co-occurrence patterns and may not capture all causal relationships.

1. **Use of Excel for Data Analysis:** While Excel is user-friendly, it presents challenges when analysing complex datasets. Specifically, identifying maximum and minimum lift values across all possible combinations (2-way, 3-way, or 4-way lifts) requires substantial manual effort. A more efficient approach would involve using R, which can automate these calculations and handle larger datasets more effectively.

2. **Scope and Data Quality:** The analysis is limited by the dataset's scope, which may not capture all relevant factors influencing survival. Additionally, potential biases due to missing values and the dataset's inherent sampling limitations may affect the results.