**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**:
  + [List of identified frequent itemsets]
* **Association Rules**:
  + [List of generated association rules with their support, confidence, and lift values]
* **Visualizations**:
  + [Include bar charts, heatmaps, network graphs, etc., to illustrate findings]

**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.

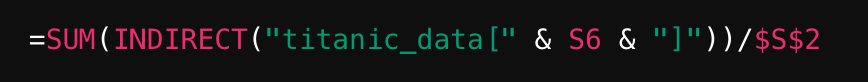


* + Second, there’s a feature frequency lookup table that holds the frequency values for each feature/column in the dataset. There’s 16 features in this particular version of the Titanic dataset.

A table with numbers and text

Description automatically generated

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



*Components of the formula:*

1. 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.
2. 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.
3. /$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.
   * **/ $S$2**: This divides the sum obtained in the previous step by the value in cell S2.

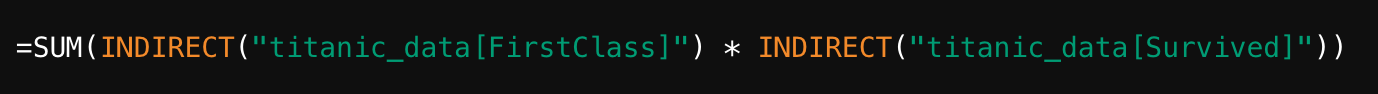
Similarly, the frequency percentages for the remaining 15 features can be calculated by replacing S6 with their respective cell references. I decided to use Excel’s fill function to speed up the process.

* **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:

1. ***LHS (Left-Hand Side)***: This column lists the different features or conditions (e.g., FirstClass, SecondClass) whose association with survival is being analysed.

2. ***RHS (Right-Hand Side)****:* This column lists the outcome "Survived," which is the target of the analysis.

3. ***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):



4. ***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:  
  


5. ***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:



6. ***Trans (Transactions)****:* The total number of transactions (891 in this dataset).

7. ***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:



8. ***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:

A black and white sign with white text

Description automatically generated

**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.

* **Implications for Predictive Modeling**:
  + [Explain how the discovered associations can enhance predictive models]
* **Limitations**:
  + [Acknowledge any limitations of the analysis, such as data quality or scope of the dataset]