**Project on BussinessAnalytics**

**Fry’s Marketplace**

**Describe the business process and/or rule related with the business.**

Fry's Marketplace is a chain of [supermarkets](https://en.wikipedia.org/wiki/Supermarket) that has a major presence in the [U.S. state](https://en.wikipedia.org/wiki/U.S._state) of [Arizona](https://en.wikipedia.org/wiki/Arizona). It is a multi-department store that offers a combination of groceries and general merchandise comparable to a Walmart Supercenter. In this project, we will visualize the data on how the sales are going for Fry’s stores in each city and analyze and examine the following data models that best fits the Fry’s Marketplace.

* Association Analysis
* Cluster Analysis
* Decision Tree
* Logistic Regression
* Neural Network
* Text Mining
* Social Network Analysis

**Define the objective/goals for this business (e.g., provide decision-making/related questions that you would work with through this project)**

After data visualization, we can identify which city’s Fry’s store has fewer profits & sales. Then the goal of this project is to increase the sales and attract customers for a store with fewer profits and to decide which data model best fits the Fry’s Marketplace.

**Explain data that will be collected in this business**

For the data visualization, I have found a sample data set online and I have made some changes according to my project and used that for my analysis.

I have analyzed sales and profits by each city and by different categories in each city. I have also analyzed sales vs profit and overall sales and profit for different categories.

Chart, bubble chart

Description automatically generated

Chart

Description automatically generated with medium confidence

**Identify the decision that needs to be made and select model(s) to help with decision  
making.**

After data visualization, I have analyzed that Fry’s store at chandler has fewer sales and profits. In order to increase the sales and profits of Fry’s store at chandler we will examine the following data models that helps in decision making.

* Association Analysis
* Cluster Analysis
* Decision Tree
* Logistic Regression
* Neural Network
* Text Mining
* Social Network Analysis

**Explain how each model we learn in the class can be embedded in your project**

**Association Analysis:**

Association analysis helps us to understand identifying the patterns, relations in existing data sets. Considering the data set available for Fry’s Marketplace, we need to figure out the relationship between products.

To find the patterns or relations in existing dataset we must define Rule. Let us consider a dataset of 5 transactions

|  |  |
| --- | --- |
| **Transaction** | **Products purchased** |
| 1 | Bread |
| 2 | Bread, Jelly, Nutella |
| 3 | Bread, Nutella, Eggs |
| 4 | Bread, Eggs, Orange, Nutella |
| 5 | Bread, Nutella, Onion, Milk |

A Rule can be derived from here as a customer who purchases {Bread, Eggs} will purchase {Nutella} because based on the transaction dataset, it shows that most of the customers who ever purchased Bread and Eggs also purchased Nutella.

{Bread, Eggs} 🡪 {Nutella}

The strength of this rule can be measured by using the Support and Confidence. **Support** tells that for a given rule how many transactions exist in the dataset that supports the given rule. **Confidence** tells us how often the rule is found to be true.

In our scenario,

Support = No.of transactions {Bread, Eggs, Nutella}/ Total no. of Transactions

=2/5 =0.4

Which shows that 40% support, the customers who ever purchased {Bread, Eggs} also bought {Nutella} and vice versa. It is symmetric

Confidence = Support for {Bread, Eggs, Nutella} / Support for {Bread, Eggs}

= (2/5)/ (2/5) =1

This indicates that 100% confidence towards support which shows that confidence for customers buying {Bread, Eggs} also bought {Nutella} is 100%. It is asymmetric.

**Lift** tells us about the support for {Bread, Eggs, Nutella} if antecedents and consequents were independent

Lift= (support for {Bread, Eggs, Nutella})/ (support for{Bread, Eggs}X support for{Nutella})

= (2/5)/(2/5 X 4/5) = 1.25

Here lift is >1 which indicates that there is a high possibility of an association between customer buying {Bread, Eggs} also bought {Nutella}

**Business suggestion**: To increase the sales and profitability of the company, I would suggest every customer who ever purchases 2 {Bread, Eggs} gets 30% discount on {Nutella}. In addition to that they can attract the customers by placing all the products together.

Graphical user interface, text

Description automatically generated

**Cluster Analysis:**

Cluster analysis is the task of grouping a set of products in such a way that products in the same group are more similar to each other than to those in other groups.

As mentioned before, Fry’s Marketplace offers its customers online & in-store shopping experience for everything from groceries and essentials to clothing and furniture. Suppose a customer wants to purchase a product and if it is out of stock, then the online system should be able to provide a suggestion to the customer with similar products available. The system must compare the different product specifications before displaying the suggestions. Instead of losing the customer by saying the product is not available, instead, it can help the customer to buy a similar product which increases sales.

Having said that the system should be able to store all the product specifications into a dataset and form clusters considering the 2 or 3 key product specifications. For example, if you are trying to buy dairy milk chocolate, then the Fry’s Marketplace online system should be able to compare the product specifications such as calories, nutritional values, and price etc with other similar products and should show recommendations such as a cluster of products.

The cluster analysis could also be applied to instore shopping customers to improve the sales and instore shopping experience. The results from cluster analysis are a group of products that can be stacked together in one place at the store.

There are multiple clustering tools available in the market. In this course, we have learned to perform K-means clustering in the JMP tool.

Graphical user interface, text, application

Description automatically generated

Showing similar products based on the nutrition rating

Graphical user interface, application

Description automatically generated

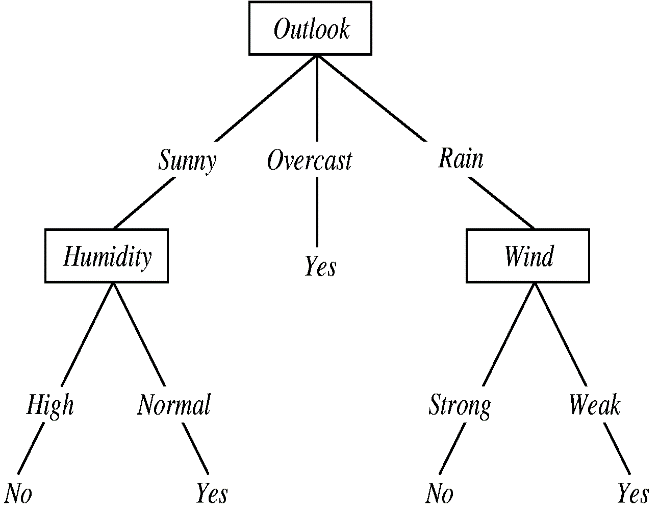
**Decision Tree:**

Decision tree is used to show the effect of categorical input on binary output. For our project, based on the sales data, the categorical inputs may be product, price, brand, type. The binary output may be customer buying the product or not.

Fry’s Marketplace should always keep the products available to the customers which directly impacts or increases the sales. This would be effectively managed by predicting the quantity needed for the future and reordering them from vendors on a routine basis.

In JMP we would do this decision tree analysis using the “Partition”. The CHAID will show the values of G^2 and LogWorth. LogWorth is the P value in the Log. We need to keep the G^2 value lower and P value higher.

The reliability of the prediction is determined using the Confusion Matrix. We have to determine the prediction accuracy of the purchase and make a reliable decision.

Diagram

Description automatically generated

**Logistic Regression:**

Logistic regression is used when we need to show the effect of a continuous independent variable in dichotomous/binary output whereas in decision tree though the output is binary the input is a clusters or group.

In logistic regression, Odds ratio is used to determine the rate of change for every change in input. For every change in input, the outcome(slope) changes by e^(intercept).

From the confusion matrix of each model decision tree and logistic regression, the reliability of the model is calculated independently which shows how many predictions were predicted correctly. By comparing the reliability of those models, we come up with a preference.

Result from Logistic Regression Result from Decision Tree



From the given Confusion matrix for the first given model logistic regression,

The reliability calculated as overall out of 303 it predicted 212 correctly.

Reliability= 212/303 =0.699

From the given Confusion matrix for the second given model decision tree,

The reliability calculated as overall out of 303 it predicted 226 correctly.

Reliability= 226/303 =0.745

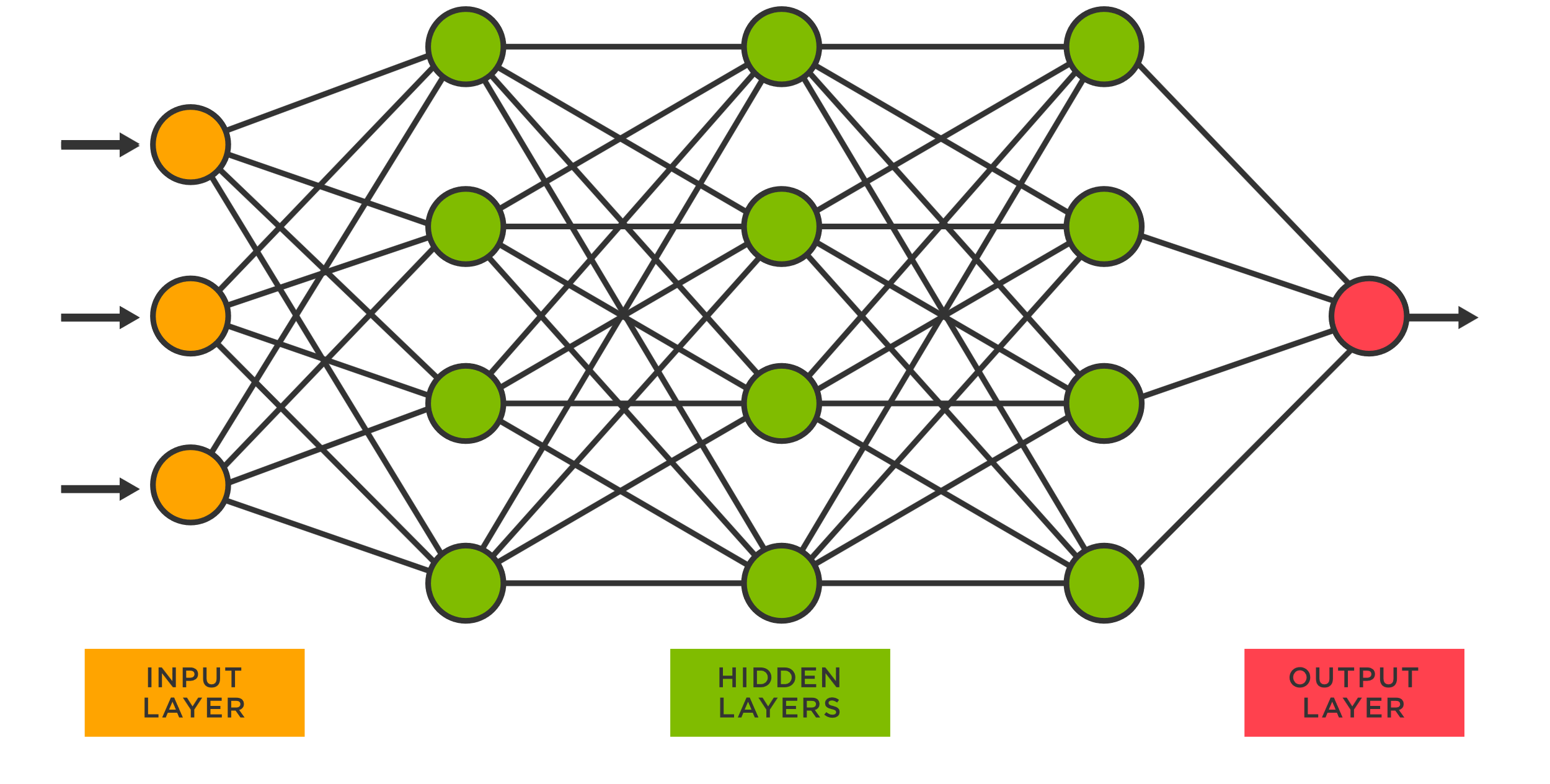
On comparing reliability of the two models, I would recommend the “**Decision tree**” model over the “Logistic Regression”.

**Neural Network:**

Neural network is a flexible model used to calculate the effect any input on an output. Neural Network provides better reliability calculations when compared to others using the multiple models in a single model in a hidden network.

The Rsquare values for any type of input data set are so accurate to its best and determines which of the input values has the most significance on output.

The neural network analysis can be applied in the event of store shelves organization. Considering the sales data as input, calculating the values of RSquare i.e., the likelihood of customer purchasing product ‘A’ will also purchase Product B, both the products should be placed together in the nearby shelves so that it provides better convenience for the customers.



**Text Mining:**

Text mining is the process of transforming unstructured text data into meaningful and actionable information. In the current business world, a large amount of semi-structured and unstructured textual data is encountered by business during its daily operations. Major portion of data an organization possesses is not in the structured form and it is difficult to process this data to get useful and desired insight. Text mining helps in such business scenarios to uncover hidden value from this data.

Text mining concept can be applied to surveys. The survey captures different types of unstructured data about products, store experience, and customer experience. Analyzing this data using text mining helps to identify the hidden truths from customers about their shopping experience. The kinds of ideas that may be extracted from such surveys about products are quite important, and they can assist organizations in developing products that customers would like the most. Maintaining a low-quality product at stores may lead to a bad reputation for the company which directly impacts the sales and reputation.

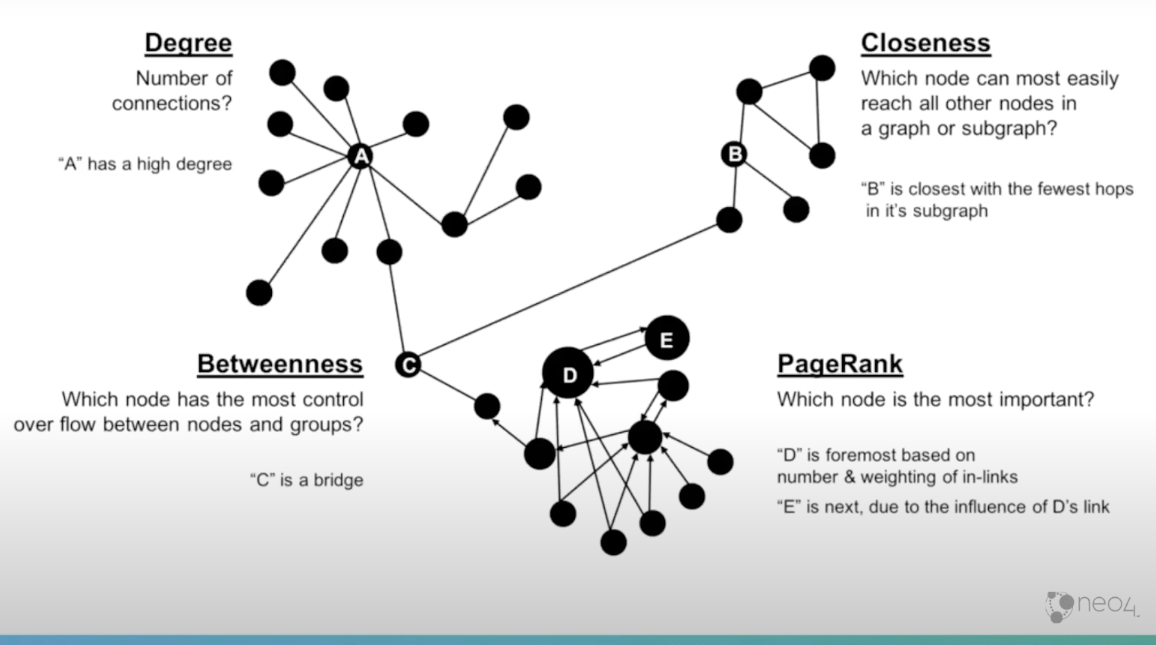
So, Text mining is always crucial in extracting such hidden truths within the unstructured data.

**Social Networking Analysis:**

Social network analysis (SNA) is used to analyze patterns of relationships among people in groups. SNA help in marketing identifying how to get product to the customers, who do we reach out that has higher influence on others, who has the closer access to others.

In networking, degree centrality can be used in identifying the popular individuals who can quickly connect with the wider community or groups. Betweenness centrality can be used for finding the individuals who has influence on others. Closeness centrality is finding the individuals who are in a closer path to influence the entire group quickly.

In marketing a product, we use the SNA to identify an “individual” with greater degree centrality and greater betweenness centrality, and smaller closeness centrality for marketing a product since they make more influence when compared to others. Though there is a high cost involved in hiring such an individual, he/she can create greater influence on the product which in turn increases the sales and profits.

 **Finalize the models that you prefer to use for your final project, explain why you chose  
those models and the information provided by the model(s) that will help with decision  
making.**

Out of the 7 models we discussed above, each model has its own advantages. I truly believe that each model has its own significance in making the right decision making.