DATA MINING THEORIES
1st Semester, 2023

CAUSAL ANALYSIS PROJECT REPORT

AKAR SIMAY 1114233

About the project:

For this project, the data named “Amazon sales data” obtained from Kaggle was selected and it was aimed to make a causal analysis project using this dataset. The variable named *discount_percentage* was chosen as the treatment variable, and the variable named *rating* was chosen as the outcome variable. In the project, causal discovery, which is one of the causal analysis methods that aim to uncover causal relationships and examine the relationships between variables, was preferred. It is aimed to examine the effect of the *discount_percentage* variable on the *rating* variable.

Steps:

1. **Data Collection:**

The dataset involves analyzing data from Amazon. It has information of 1000+ products sold by Amazon like their names, categories, prices, discount information, ratings, and reviews.

Features of dataset:

Feature Name	Feature Description
product_id	Product ID
product_name	Name of the product
category	Category of the product
discounted_price	The discounted price of the product
actual_price	The actual price of the product
discount_percentage ***(Treatment Variable)***	Percentage of discount for the product
rating ***(Outcome Variable)***	Rating of the product
rating_count	Number of people who voted for Amazon's rating

about_product	Description about the product
user_id	The ID of the user who wrote a review for the product
user_name	Name of the user who wrote a review for the product
review_id	The ID of the user review
review_title	Short review
review_content	Long review
img_link	Image link of the product
product_link	Official website link of the product

Table 1. Features of dataset

2. Data Preparation:

Cleaning the dataset and formatting properly. These steps are explained as comments on the code columns in the project.

- **Data Inspection & Cleaning:** The dataset is examined for missing values, duplicates, or inconsistent data. It is checked whether the data types are correct, and if they are incorrect, they are converted to the appropriate ones. The dataset is then cleaned by removing or correcting any errors, inconsistencies, or irrelevant information.
    - There were no duplicate values.
    - In the dataset, two missing rows were identified in the rating_count column and removed.
    - The data type of “discounted_price”, “actual_price”, and “discount_percentage” were changed to float and symbols were removed.
    - The incorrect character has been excluded from the values in the “rating” column.
    - The “rating” and “rating_column” data types were also changed to float.

3. Data Configuration:

After the data preparation was completed, some configurations were made to make the data suitable for analysis.

a) To be used in the analysis, the length of the text content in the "review_content" column was considered and calculated. The result was assigned to a new column named "review_content_length" on the dataset.

b) The “category” column was split into “main_category” and “sub_category” columns, thus creating two more columns with these two different names.

c) It was checked how many unique main categories there were, and then the category names were matched with numbers from 1 to 9 by mapping these unique categories. Based on the defined mapping, numeric values were assigned to the main categories, these values were transferred to the newly created column named “category_number”.

d) Main categories & subcategories were shown for getting more information.

4. Causal Analysis:

a) A graphical representation based on an adjacency matrix was created. For this, *the make_graph function* was defined.

b) Unnecessary columns were removed for analysis. These columns are:

```
[ 'product_id', 'product_name', 'category', 'about_product', 'user_id', 'user_name',
'review_id', 'review_title', 'review_content', 'img_link', 'product_link', 'sub_category',
'main_category' ]
```

c) The "rating" level was adjusted based on a threshold. If the "rating" value for it was greater than 4.1, it was considered a high grade, and a value of 1 was assigned. Otherwise, a value of 0 was assigned.

d) For causal discovery, prior knowledge was generated using *the make_prior_knowledge function.* Here Index = 3 refers to the "rating" column, which is the outcome variable.

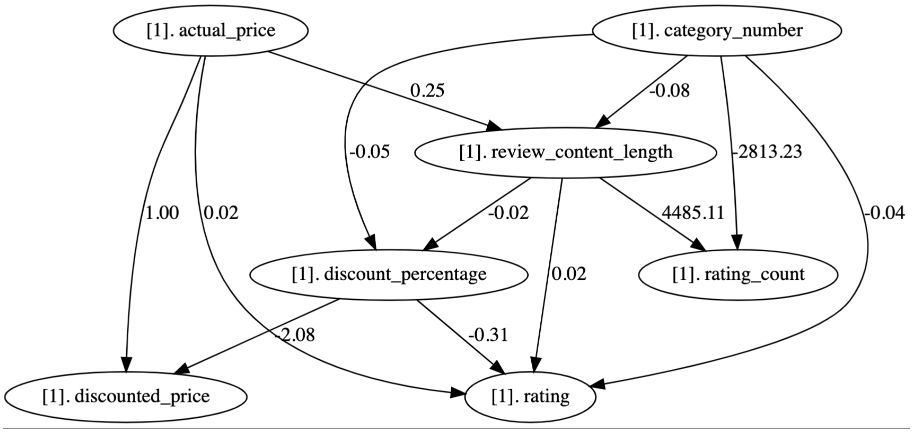
e) The "actual_price", "discounted_price" and "review_content_length" columns were transformed using the natural logarithm function(np.log). This transformation is commonly used to reduce the skewness of the data and make it more suitable for analyses that assume a more linear relationship.

f) A causal discovery model was set up using Lingam. To visualize the causal relationships identified by the model, the variable labels were created using the column names from the dataset.

g) By using *the make_graph function* the graph was represented in the DOT language format, which is a plain text graph description language used by Graphviz.

h) A graph visualization was generated with the make_dot function. After dropping the unnecessary columns for the analysis as described in the previous steps, the features used in the analysis are as follows:

actual_price, discounted_price, **discount_percentage**, review_content_length, category_number, rating_count, **rating**

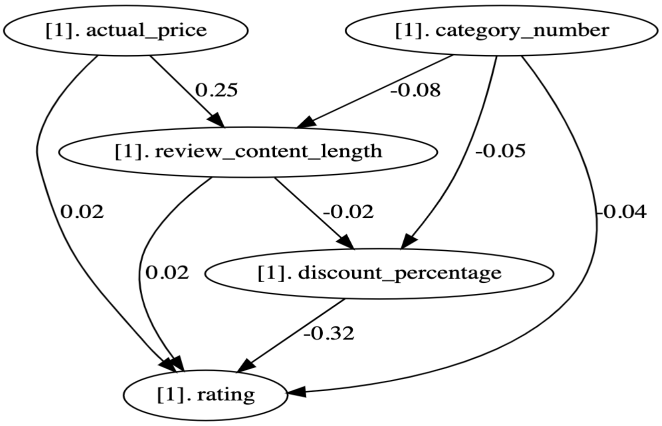


Graph 1. Causal Graph

i) Dead-end variables in a causal graph are those that have no causal effect on other variables in the system. The "discounted_price" and "rating_count" columns were removed because they are defined as dead-end variables in the chart.

j) The analysis was narrowed down to focus on variables that have a direct causal effect on the sink variable, which is the "rating", leaving the dead-end variables.

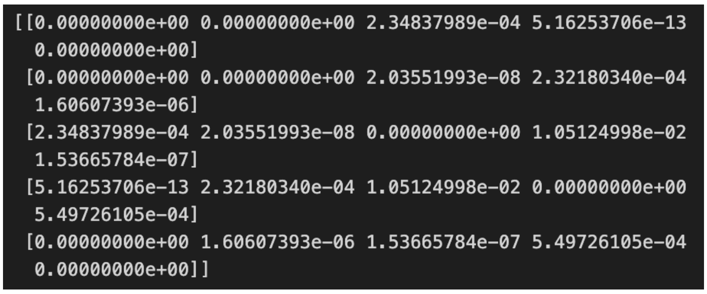
k) After removing the dead-end variables, the model was recreated as updated_model. Causal relationships were visualized again graphically with this updated model.



Graph 2. Causal Graph with Updated Model

l) For independence tests of causal relationships, p values were obtained with the get_error_independence_p_values function.

- If a p-value is close to zero, it suggests strong evidence to reject the null hypothesis of independence. If a p-value is relatively high, it suggests weak evidence to reject the null-independence hypothesis.



```
[[0.00000000e+00 0.00000000e+00 2.34837989e-04 5.16253706e-13
  0.00000000e+00]
 [0.00000000e+00 0.00000000e+00 2.03551993e-08 2.32180340e-04
  1.60607393e-06]
 [2.34837989e-04 2.03551993e-08 0.00000000e+00 1.05124998e-02
  1.53665784e-07]
 [5.16253706e-13 2.32180340e-04 1.05124998e-02 0.00000000e+00
  5.49726105e-04]
 [0.00000000e+00 1.60607393e-06 1.53665784e-07 5.49726105e-04
  0.00000000e+00]]
```

Interpretation of these p_values:

Row 2: “discount_percentage”

- The p-value between the error term of “discount_percentage” and “actual_price” is 2.38821306e-04, suggesting a significant relationship between these variables' error terms.

Row 3: “rating”

- The p-value between the error term of “rating” and “actual_price” is 2.00421854e-08, suggesting a significant relationship between these variables' error terms.

- **The p-value between the error term of “rating” and “discount_percentage” is 2.55761040e-04, suggesting a significant relationship between these variables' error terms.**

Row 4: “review_content_length”

- The p-value between the error term “review_content_length” and “actual_price” is 1.27989955e-06, suggesting a significant relationship between these variables' error terms.

- The p-value between the error term of “review_content_length” and “discount_percentage” is 1.15641825e-02, suggesting a significant relationship between these variables' error terms.

- The p-value between the error term of “review_content_length” and “rating” is 6.16531377e-03, suggesting a significant relationship between these variables' error terms.

Row 5: “category_number”

- The p-value between the error term of "category_number'" and "actual_price" is 0.0, suggesting a highly significant relationship between these variables' error terms.

- The p-value between the error term of "category_number" and "discount_percentage" is 1.60013762e-07, suggesting a significant relationship between these variables' error terms.

- The p-value between the error term of 'category_number' and 'rating' is 6.16531377e-03, suggesting a significant relationship between these variables' error terms.

- The p-value between the error term of "category_number" and "review_content_length" is 0.0, indicating a highly significant relationship between these variables' error terms.

- The p-values indicate the statistical significance of the relationships between the error terms of the variables in the model. Lower p-values suggest stronger evidence of dependence or causal relationships between the corresponding variables' error terms.

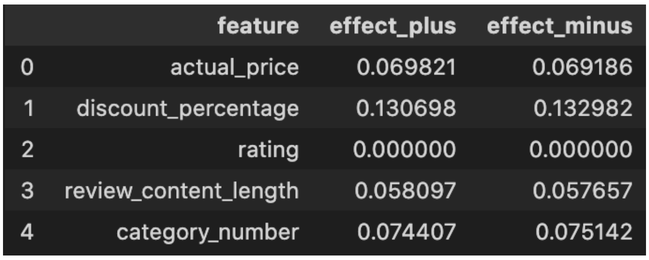
- What we are interested in here is the item marked in yellow, emphasizing that there is a significant relationship between the error terms of the "discount_percentage" and "rating" variables.

m) Logistic Regression:

After looking at the p-values, the scikit-learn library was used to perform logistic regression. Logistic regression is a classification method used in causal discovery to determine the cause-effect relationship between variables. Using logistic regression, a predictive model can be constructed to estimate the causal effect of one variable on another. It is used in this project to estimate the effect of "discount_percentage" on "rating". Logistic regression is commonly used when the dependent variable is binary, that is why the "rating" variable was set as 1 or 0 in the project.

Using logistic regression in causal discovery is advantageous as it has the ability to infer causal relationships based on the statistical properties of the predictive model. This allows us to directly estimate the effects of variables on each other.

n) The causal effects of features on prediction were calculated using the CausalEffect class from the Lingam library.



	feature	effect_plus	effect_minus
0	actual_price	0.069821	0.069186
1	discount_percentage	0.130698	0.132982
2	rating	0.000000	0.000000
3	review_content_length	0.058097	0.057657
4	category_number	0.074407	0.075142

Table 2. Causal Effects of Features

- *The "effect_plus" column* indicates the effect of each feature on the likelihood of a positive rating. Higher positive values suggest that increasing the corresponding feature tends to increase the probability of a positive rating.

- *The "effect_minus" column* represents the effect of each feature on the likelihood of a negative rating. Higher positive values imply that increasing the corresponding feature tends to increase the probability of a negative rating.

o) The feature that has the maximum effect on the prediction was found, which turned out to be *"discount_percentage"*.

5. Findings and Conclusion

In this project, it was aimed to investigate the causal effect of “discount_percentage” on “rating” and causal discovery analysis was performed using the Amazon sales dataset. By using logistic regression, the effect of the variable "discount_percentage" on the variable "rating" was estimated. In addition, the resulting model provided information about how the variables in the analysis affect the target variable ("rating") and the direction of this effect.

The interpretation of each feature in the table resulting from the analysis is explained in the comments section of the project code. In this section, the results of the “discount_percentage” and “rating” relationship, which this project is interested in, will be examined. Looking at the table, the effect_plus value for the "discount_percentage" feature is 0.130698, indicating a positive causal effect on the result. That is, the higher the discount_percentage, the more likely it is to have a positive rating. The effect_minus value is 0.132982, showing a negative causal effect on the result. That is, the higher the discount_percentage, the more likely it is to have a negative rating. These two effect values were found to be relatively close to each other. These findings suggest that increasing or decreasing the discount percentage has a causal impact on the outcome variable ("rating") but in opposite directions. According to the statistical properties of the predictive model table resulting from the logistic regression, the output "max index" represents the name of the variable with the highest effect value in the effects array. In this case, the output of the "max index" value which is "discount_percenatage" refers to the variable specified as "rating", meaning it has the highest effect on the "rating" variable.

As a result of this project, it was found that “discount_percentage” has the positive and the highest effect on “rating” by causal discovery analysis. These findings suggest a significant relationship between '"discount_percentage" and "rating", emphasizing that it is worth further research to determine the exact nature of this relationship.