Ayan Sarkar

ABC Private limited

customer purchase behaviour analysis

**Business Understanding**

This study involves analyzing the purchasing patterns of the customers of a retail company.

One of the best examples of how we both overestimate and underestimate changes in the future is the evolution of consumer behavior throughout this century. And a lot of research has gone in this area and is one of the top areas which capitalizes the use of Analytics, to understand this ever-changing trend.

The potential to better understand this industry domain and the exciting opportunities to study the retail customer purchasing trends drove us to pick up this project.

A retail company “ABC Private Limited” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from a particular month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

We took this a few steps forward, in addition to predicting the purchase amount in the sales, we also predicted sales in various price ranges: Low, Medium and High so that the targeted marketing can be more effective in producing results if the demographic is split based on their purchasing power. In addition to this, we also identify the niche products which yield us the highest mean purchase amounts in sales, so that special attention can be given in catering to such customers.

**Data Understanding/Preparation**

This business case and its data has been taken from an open practice problem for users on an analytics website: https://datahack.analyticsvidhya.com/contest/black-friday/

The data was provided in .csv files. The data contains 550068 records; the data set contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **User\_ID** | User ID |
| **Product\_ID** | Product ID |
| **Gender** | Sex of User |
| **Age** | Age in bins |
| **Occupation** | Occupation (Masked) |
| **City\_Category** | Category of the City (A,B,C) |
| **Stay\_In\_Current\_City\_Years** | Number of years stay in current city |
| **Marital\_Status** | Marital Status |
| **Product\_Category\_1** | Product Category (Masked) |
| **Product\_Category\_2** | Product may belong to another category also (Masked) |
| **Product\_Category\_3** | Product may belong to another category also (Masked) |
| **Purchase** | Purchase Amount (Target Variable) |

Exploratory Data Analysis

Our data set primarily consists of categorical variables with one continuous variable as the target. Please find below our findings from the data audit:

* There are 11 categorical variables, 2 out of them are ID variables and hence not usable
* The target variable is continuous, does not have any outliers.
* The distribution of values in the target variable is fairly normal with a longer tail on the right
* Missing Values were encountered in only 2 variables: Product Category 2 (30 % missing values) and Product Category 3 (70% missing values)
* The various levels in the categorical fields are: Gender-2, Age-7, Occupation-21,

City\_Category-3, Stay\_In\_Current\_City-5, Marital\_Status-2, Product\_Category\_1-20,

Product\_Category\_2-17 & Product\_Category\_3-15

Please refer to [**Section 1**](#Section1) of the Appendix to see detailed output graphs for the data audit done on the data set.

Data Preparation

1. Product 3 category field had 70% missing values; so, we dropped it from our analysis
2. The 32% missing values in Product Category 2 field has been imputed by the mode value: 8
3. The variable: Gender (M/F) has been converted to two numerical flag variables for its 2 levels
4. The variable: Age has been re-classified into 3 levels from the initial 7 levels by the below mapping:

0-17: 0-35

18-25: 0-35

26-35: 0-35

36-45: 36-55

46-50: 36-55

51-55: 36-55

55+: 56+

Then the transformed Age variable has been converted to 3 numerical flag variables for the 3 levels.

1. The variable: Occupation has been re-classified into 3 levels from the initial 20 levels by the below mapping:

0..6: 0-6

7..14: 7-14

15..20: 15-20

Then the transformed Occupation variable has been converted to 3 numerical flag variables for the 3 levels.

1. The variable: City Category (A/B/C) has been converted to three numerical flag variables for its 3 levels
2. The variable: Stay in Current City has been re-classified into 3 levels from the initial 5 levels by the below mapping:

0..1: 0-1

2..13: 2-3

>= 4: 4+

Then the transformed Stay in Current City variable has been converted to 3 numerical flag variables for the 3 levels.

1. The variable: Product Category 1 has been re-classified into 3 levels from the initial 20 levels by the below mapping:

1..7: 1-7

8..14: 8-14

15..20: 15-20

Then the transformed Product Category 1 variable has been converted to 3 numerical flag variables for the 3 levels.

1. The variable: Product Category 2 has been re-classified into 3 levels from the initial 17 levels by the below mapping:

2..7: 2-7

8..12: 8-12

13..18: 13-18

Then the transformed Product Category 2 variable has been converted to 3 numerical flag variables for the 3 levels.

1. A new variable: Purchase\_Range has been derived from the continuous variable: Purchase, having 3 levels: “Low”, “Medium” & “High”; the levels have been derived according to the below price range in the Purchase variable:

>=$0 & <=$5000 : **Low**

>=$5001 & <=$15000: **Medium**

>=$15001 : **High**

Then the new derived Purchase\_Range variable has been converted to 3 numerical flag variables for the 3 levels

**Modeling/Evaluation**

1. A **Linear regression model** has been built to predict the “Purchase” amount in the sales data. Please find below the details of the model:

* Adjusted R2: 64.2%
* Model Selection method: Forward Stepwise
* Predictors: Stay in Current City, Gender, City Category, Age, Occupation, Product\_Category\_1, Product\_Category\_2
* Target: Purchase
* The scatter plot for the Predicted Vs. Actual values indicate that the predictions have been moderately well.
* The residuals tend to follow a normal distribution.
* The residuals tend to be very close by the perfect regression line
* The most important predictor was Product\_Category\_1
* This was the best model we achieved for predicting the continuous target variable.

1. A **K-Means model** was built to identify the niche products in the two product categories: Product Category Groups: 1 & 2, which yield us the highest mean purchase amounts in sales, so that special attention can be given in catering to such customers. This model was built for the price range > 20000.

Please find below the details of the model:

Model for Product Category Group 1

* Inputs: Purchase, Product\_Category\_1
* K=5
* Cluster Quality: 0.6
* Product Category 10 has highest mean Purchase Price of $23495.43

Model for Product Category Group 2

* Inputs: Purchase, Product\_Category\_2
* K=5
* Cluster Quality: 0.4
* Product Category 14 has highest mean Purchase Price of $23524.11

1. **Naïve Rule** was used to calculate Error rates for Predicting price ranges of “High”, “Medium” & “Low” in the Testing Data set.

Error Rate for “High” price range: 25.20%

Error Rate for “Medium” price range: 59.74%

Error Rate for “Low” price range: 20.87%

1. An **Apriori algorithm** was used to explore the pattern in the demographics participating in the Purchase Ranges: “High”, “Medium” & “Low”.

Model for Purchase Ranges: “High”:

* If a sales transaction contains product categories: 2 to 7 from Product Category Group 2 and the customer is a male, then it is 1.74 times more likely than a general guess, that this is a transaction of a high range of purchase price
* If a sales transaction contains product categories: 2 to 7 from Product Category Group 2 & product categories: 1 to 7 from Product Category Group 1 and the customer is a male, then it is 1.74 times more likely than a general guess, that this is a transaction of a high range of purchase price

Model for Purchase Ranges: “Medium”:

* If a sales transaction contains product categories: 8 to 14 from Product Category Group 1& product categories: 8 to 12 from Product Category Group 2, then it is 1.243 times more likely than a general guess, that this is a transaction of a medium range of purchase price
* If a sales transaction contains product categories: 8 to 14 from Product Category Group 1& product categories: 8 to 12 from Product Category Group 2 and the customer is a male, then it is 1.23 times more likely than a general guess, that this is a transaction of a medium range of purchase price

Model for Purchase Ranges: “Low”:

* If a sales transaction contains product categories: 8 to 14 from Product Category Group 1& product categories: 13 to 18 from Product Category Group 2, then it is 1.562 times more likely than a general guess, that this is a transaction of a low range of purchase price
* If a sales transaction contains product categories: 8 to 14 from Product Category Group 1 and the customer is a male aged up to 35, then it is 1.503 times more likely than a general guess, that this is a transaction of a low range of purchase price

1. **Logistic Regression** was used to predict the Purchase Ranges: “High”, “Medium” & “Low” in the sales data which would prove to be of great use for targeted marketing.

Model for Purchase Ranges: “High”: Error Rate in Testing Data Set: 20.15%, lesser than Naïve Rule error rate of 25.20%; thus, proving to be an effective model

Model for Purchase Ranges: “Medium”: Error Rate in Testing Data Set: 35.52%, lesser than Naïve Rule error rate of 59.74%; thus, proving to be an effective model

Model for Purchase Ranges: “Low”: Error Rate in Testing Data Set: 17.27%, lesser than Naïve Rule error rate of 20.87%; thus, proving to be an effective model

1. **Neural Nets** was used to predict the Purchase Ranges: “High”, “Medium” & “Low” in the sales data which would prove to be of great use for targeted marketing.

Model for Purchase Ranges: “High”: Error Rate in Testing Data Set: 20.15%, lesser than Naïve Rule error rate of 25.20%; thus, proving to be an effective model

Model for Purchase Ranges: “Medium”: Error Rate in Testing Data Set: 35.52%, lesser than Naïve Rule error rate of 59.74%; thus, proving to be an effective model

Model for Purchase Ranges: “Low”: Error Rate in Testing Data Set: 17.27%, lesser than Naïve Rule error rate of 20.87%; thus, proving to be an effective model

1. So, in principle, Logistic Regression & the Neural Nets gives us the same predictive power in terms of the Error rates discussed above.

Please refer to [**Section 2**](#Section2) of the Appendix to see detailed output graphs for all the models discussed on the data set.

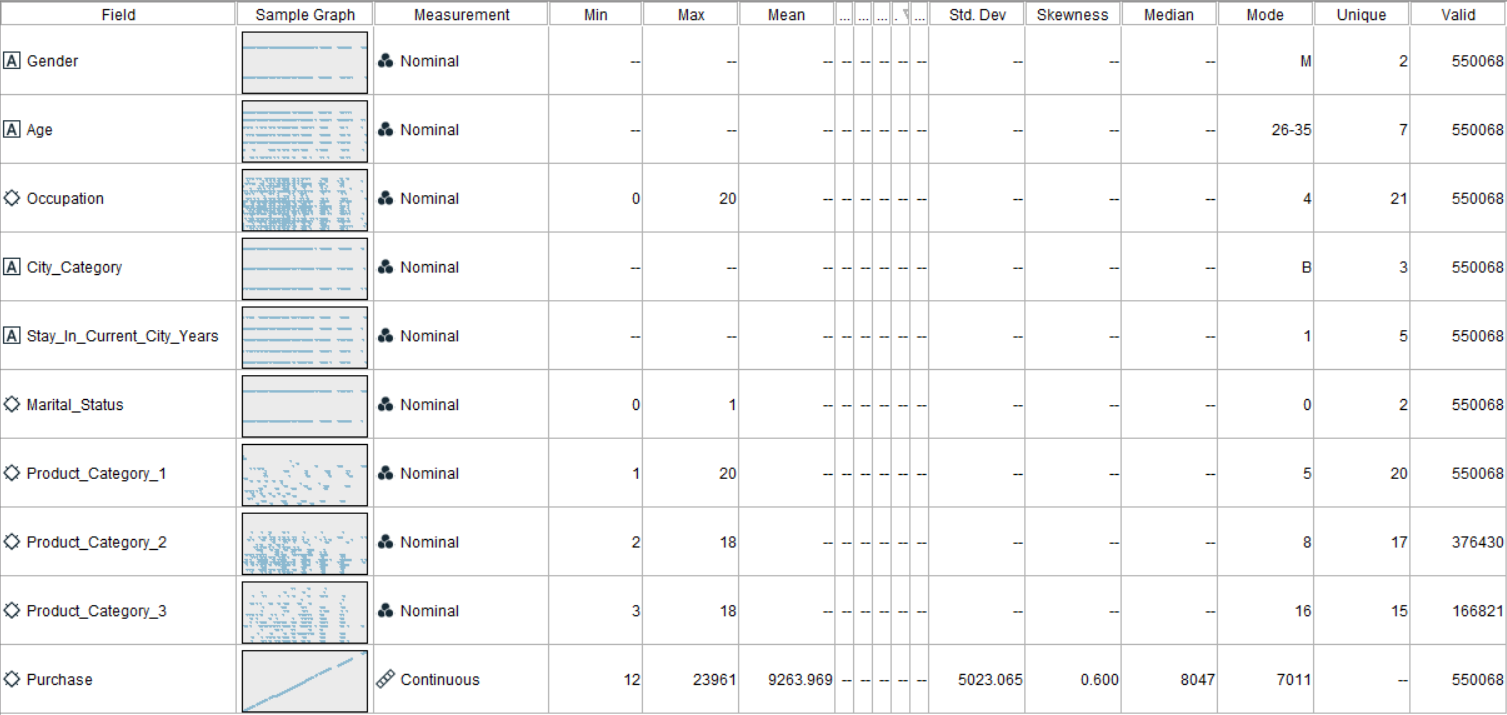
**Conclusions**

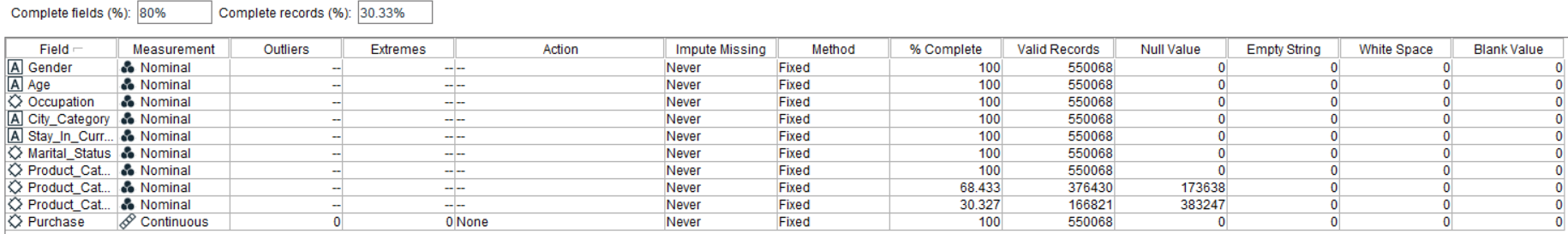
1. Predicting a continuous target variable with only categorical variables as the predictors, seems to hinder the predictive power of a linear regression model. In our case, the model could only explain 62% of the variability in the data
2. Product Category 10 is a niche product belonging to Product Category Group 1 demanding an average Purchase Price of $23495.43
3. Product Category 14 is a niche product belonging to Product Category Group 2 demanding an average Purchase Price of $23524.11
4. Male customers exhibit a trend of shopping in the high and medium purchase price range.
5. Logistic Regression & the Neural Nets gives us the same predictive power in terms of the Error /Accuracy.
6. We can predict transactions with the High range of purchase price with an accuracy of 79.85%
7. We can predict the Medium range of purchase price with an accuracy of 64.48%
8. We can predict the Low range of purchase price with an accuracy of 82.73%
9. The Apriori algorithm does not give us significant trends in the customer demographic as far as their purchasing habits go.

**Appendix**

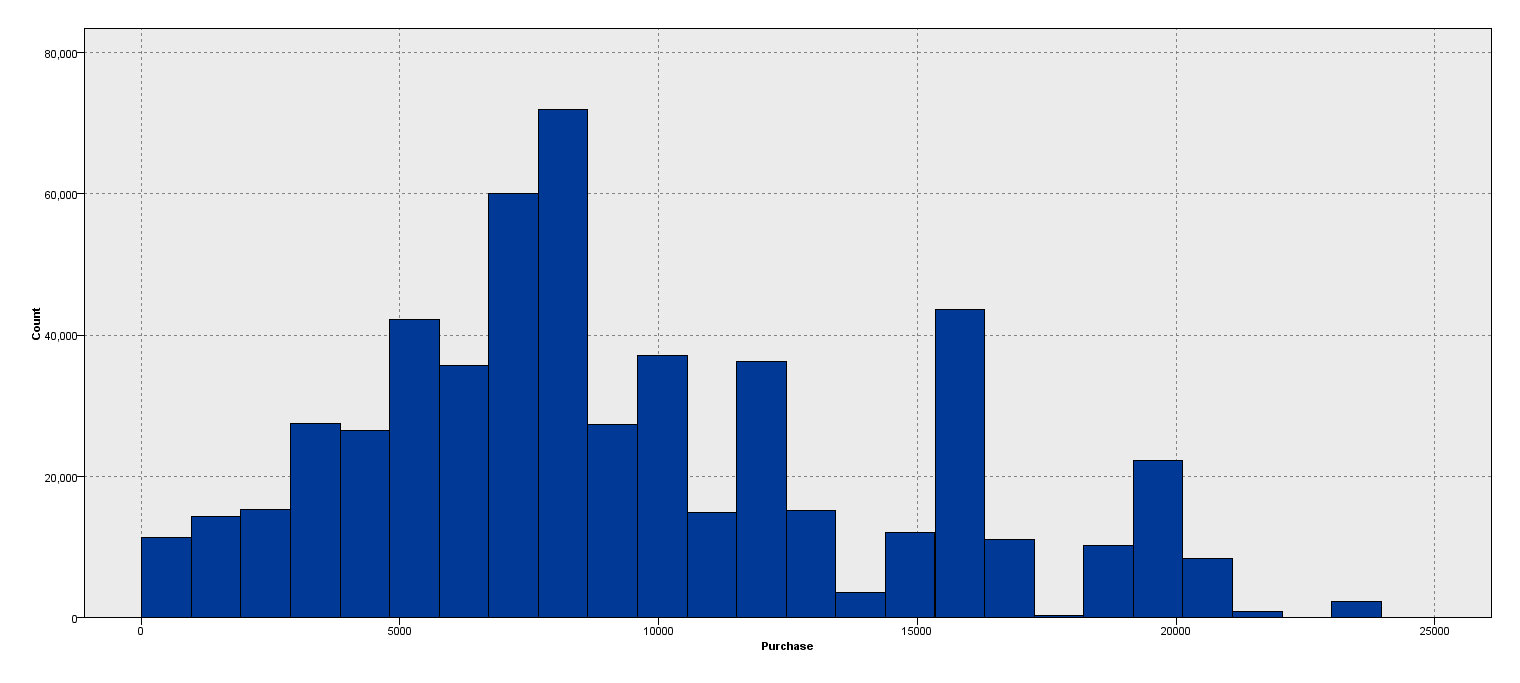
Section 1: Exploratory Data Analysis

Overall Data Audit Findings:

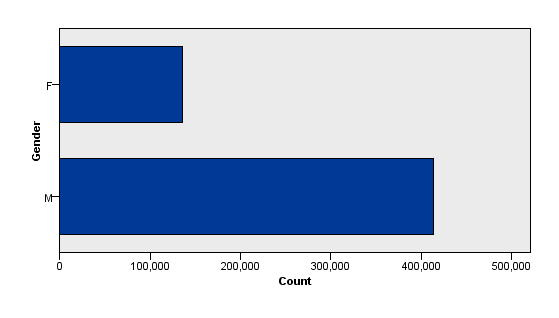


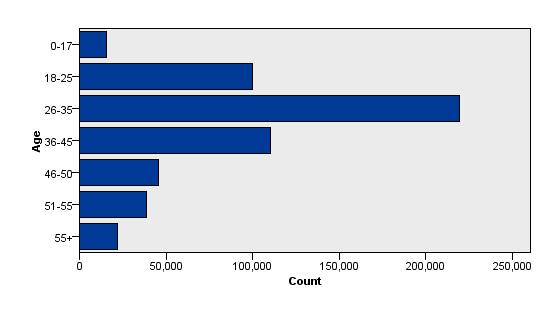


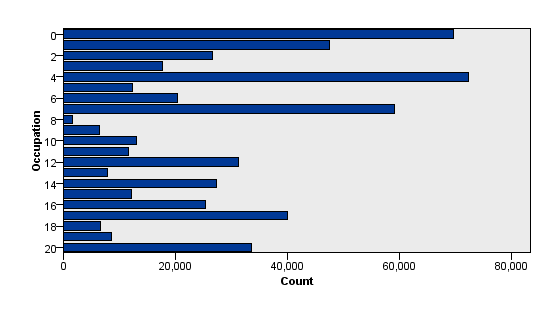
Continuous Variable Analysis:

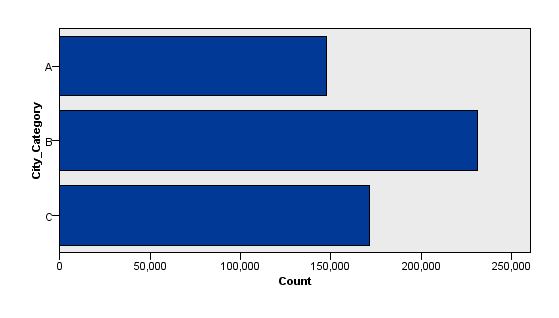


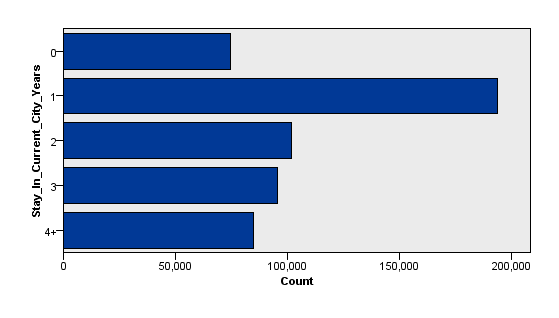
Categorical Variable Analysis:

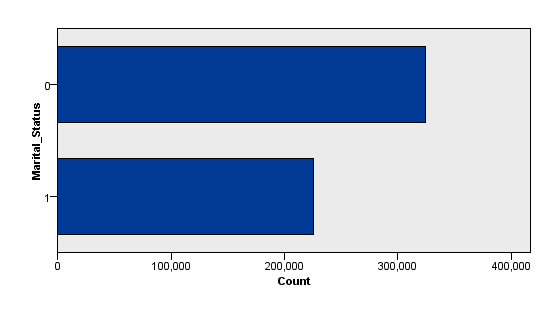


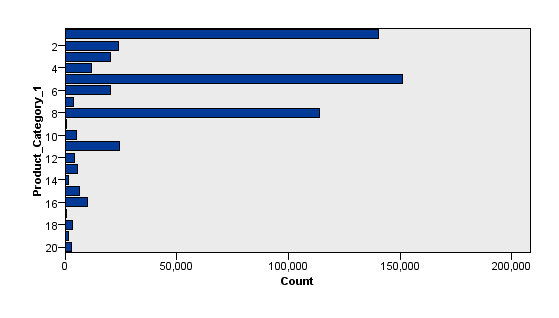


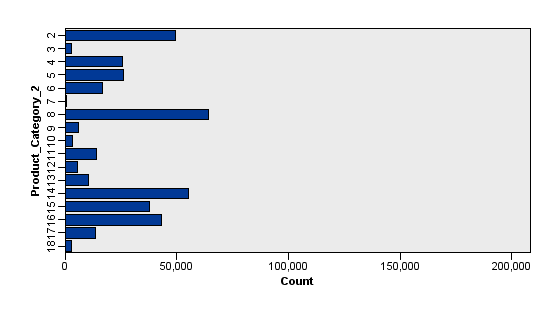


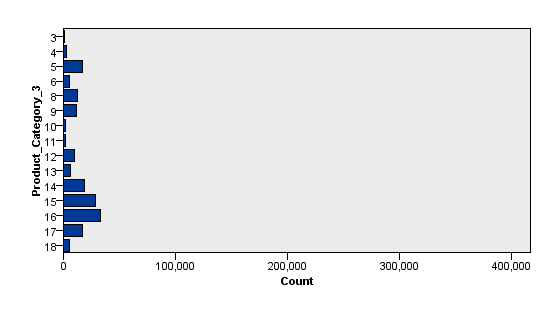








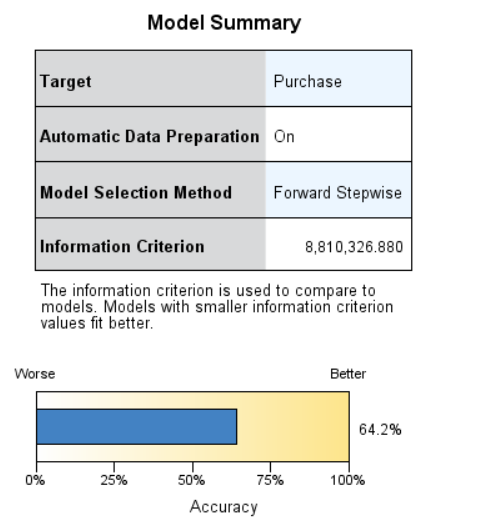


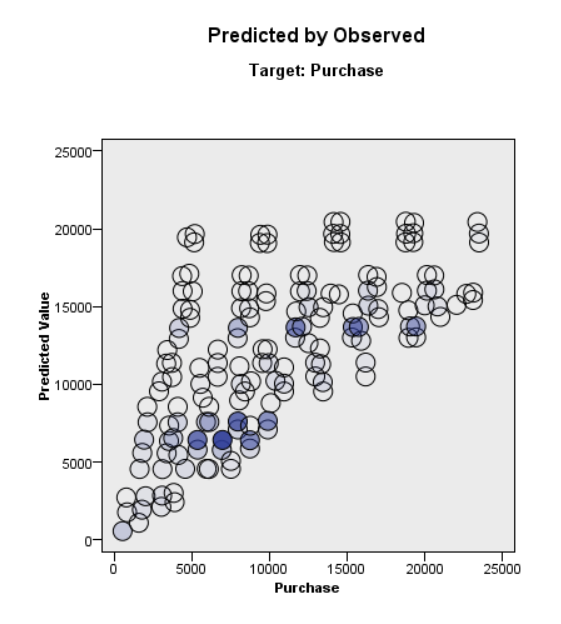


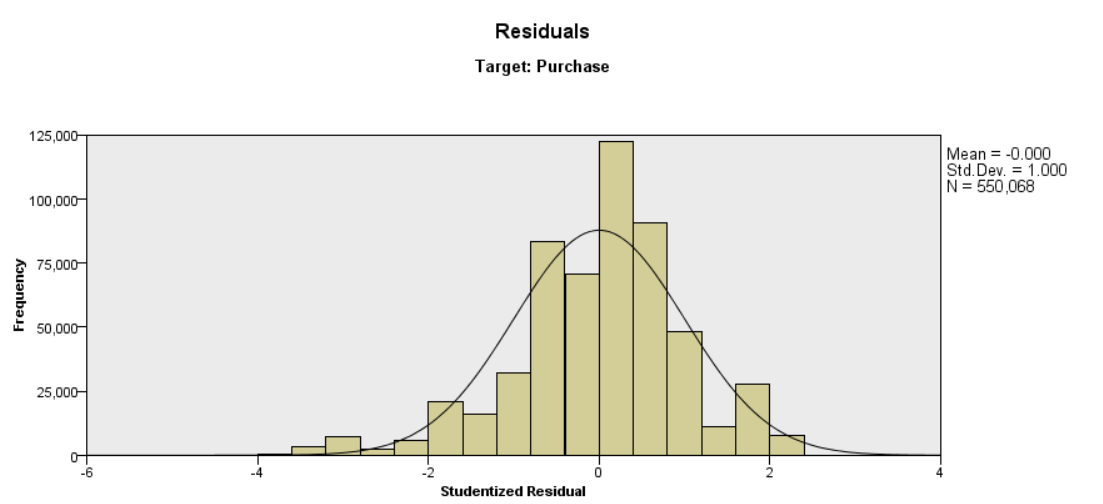
Continue to [**Data Preparation**](#DataPrep)

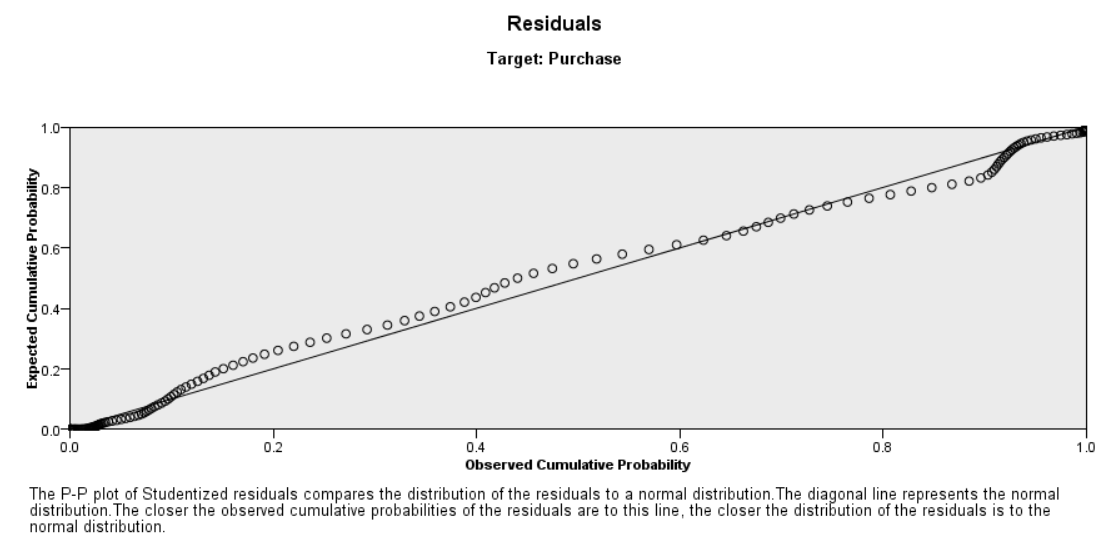
Section 2: Model Analysis/Results

**Linear Regression Model:**

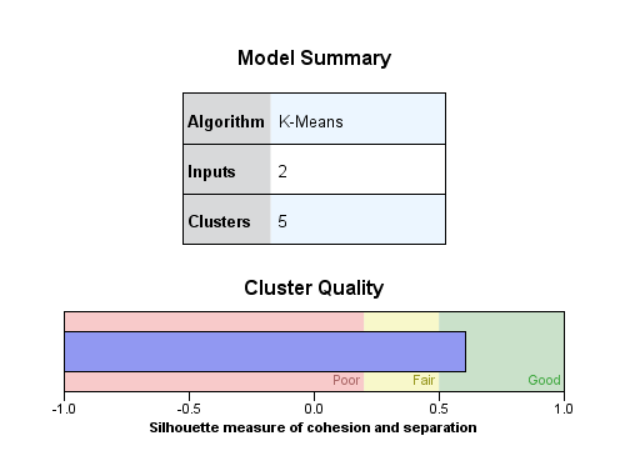


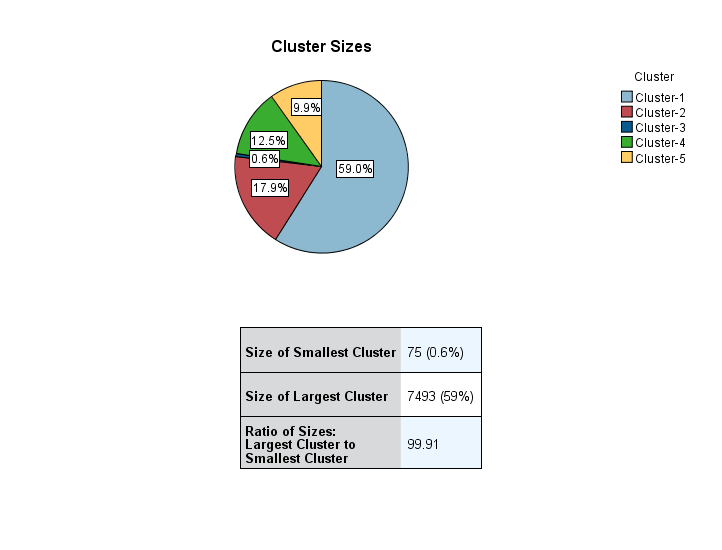


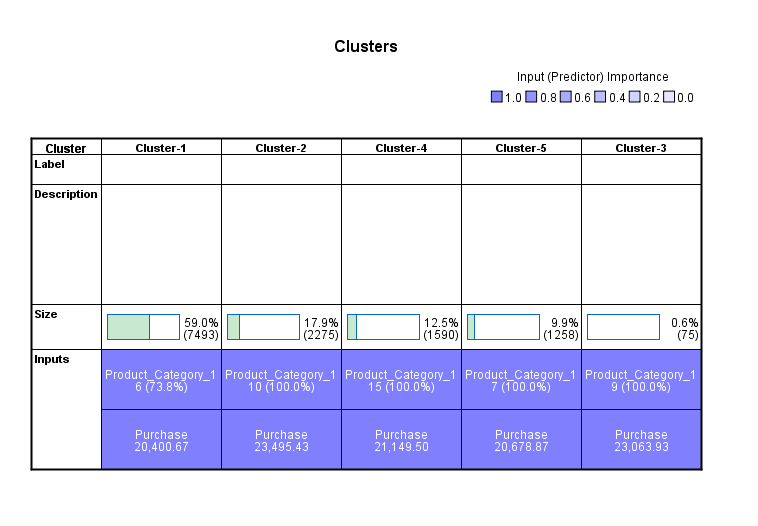




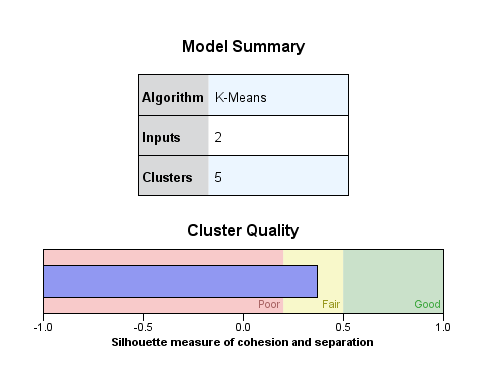
**K-Means Model for Product Category Group 1:**

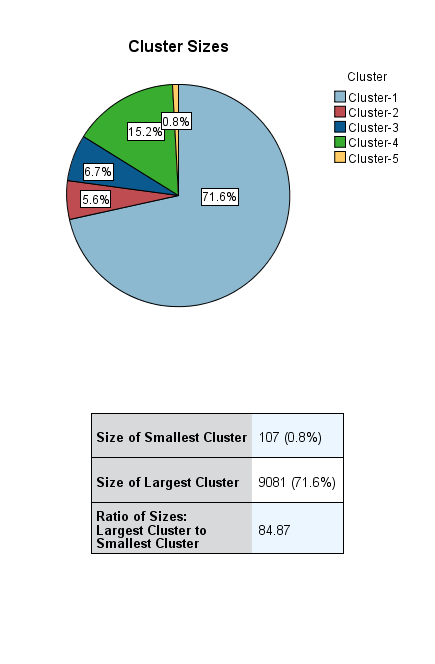


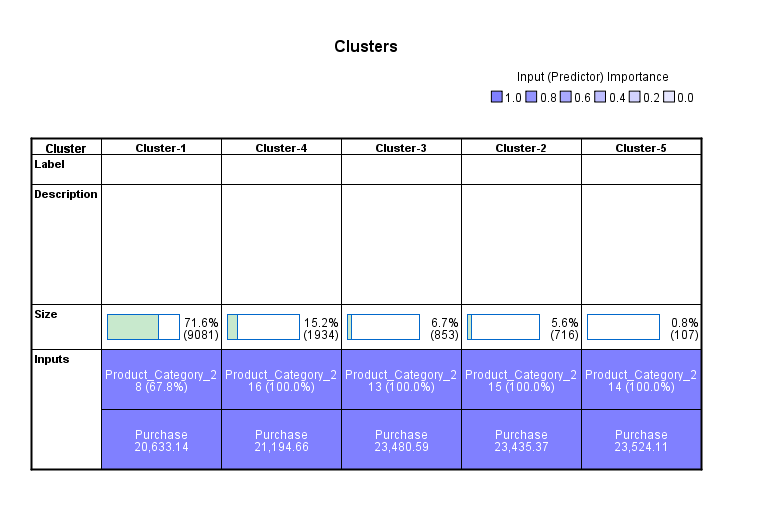




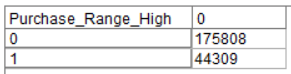
**K-Means Model for Product Category Group 2:**

****

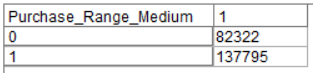
****

****

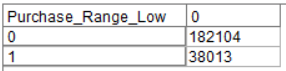
**Naïve Rule for Price Range - High:**

 Error Rate: 25.20%

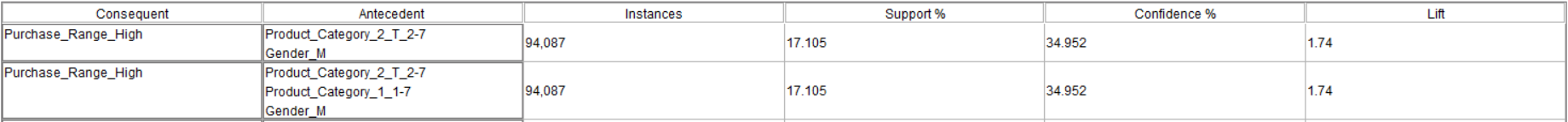
**Naïve Rule for Price Range - Medium:**

 Error Rate: 59.74%

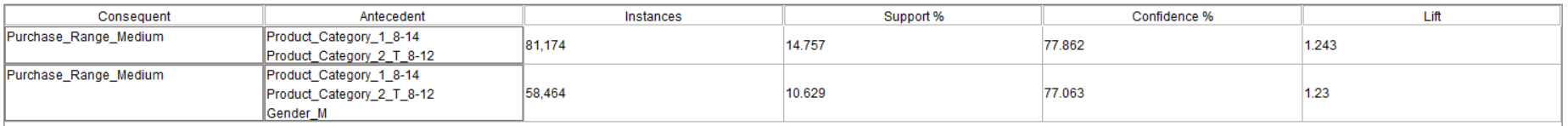
**Naïve Rule for Price Range - Low:**

 Error Rate: 20.87%

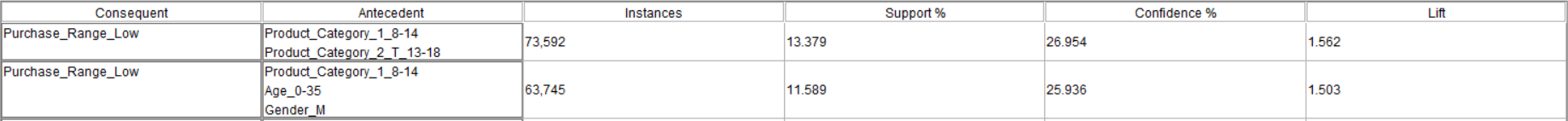
**Apriori Model for Price Range - High:**



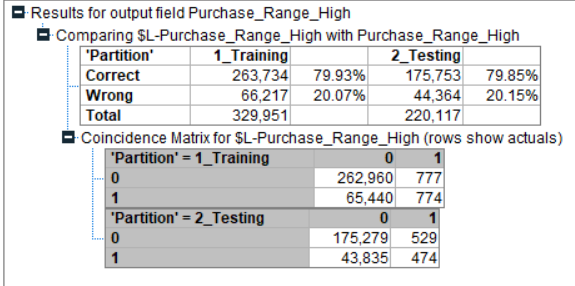
**Apriori Model for Price Range - Medium:**



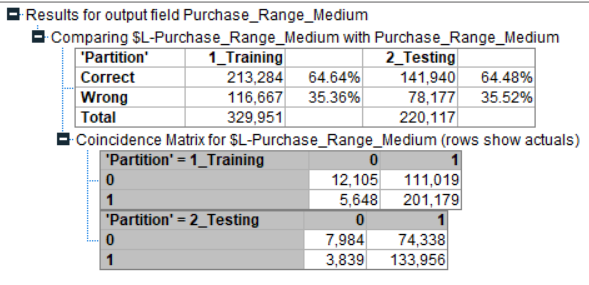
**Apriori Model for Price Range - Low:**



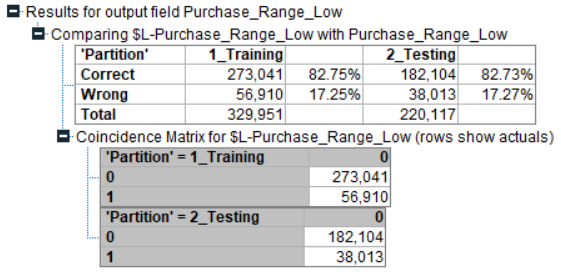
**Logistic Regression Model for Price Range - High:**



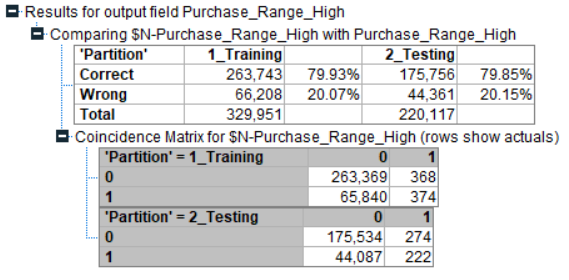
**Logistic Regression Model for Price Range - Medium:**



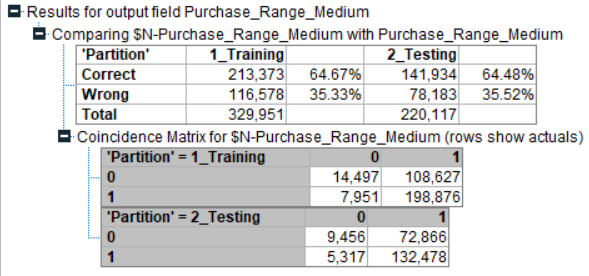
**Logistic Regression Model for Price Range - Low:**



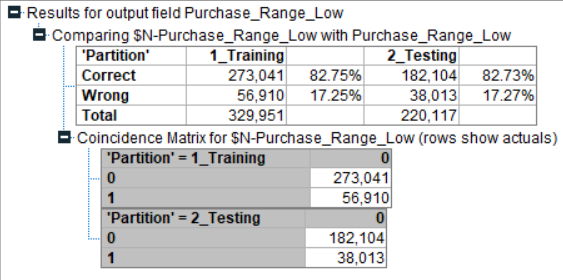
**Neural Nets Model for Price Range - High:**



**Neural Nets Model for Price Range - Medium:**



**Neural Nets Model for Price Range - Low:**



Continue to [**Conclusions**](#Conclusion)

Back to the [**Top**](#Top)