MIS272 – Predictive Analytics

T2 2023

Assignment 1 – Individual
Student name: Nguyen Anh Ngoc Le
Student number: 223142795

Executive summary

Executive problem statement:

Problem: Property reviews and details, aiming to enhance Airbnb operations in Denmark for better customer retention. Main goal: Analyze rental data, offer the best option, and increase revenue.

Objectives:

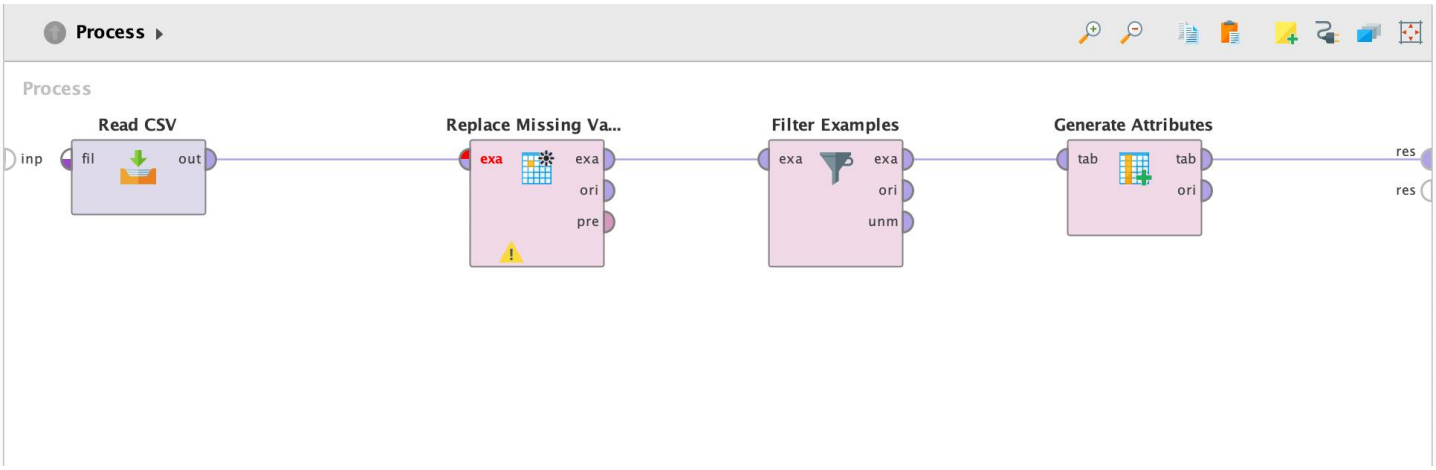
1. Geospatial View: Categorize Sealand and Denmark rentals.
2. Regional Differences: Probe price, single-person cost, and satisfaction differences between Sealand and other areas.
3. Affordable Stays: Build models to identify budget rentals in Sealand, aiding an accessibility-focused ad campaign.

Executive solution statement:

23,800 rentals are included in the dataset, including 14438 in Sealand and 9362 in other parts of Denmark. Prices per night, the cost of staying for a single person, and overall satisfaction vary significantly between Sealand and the rest of Denmark. The analysis suggests using a Decision Tree model to achieve successful categorization because of its accuracy in this situation.

Data exploration, pattern discovery, and preparation

23,941 rentals were analysed: 23,861 with full data and 80 had missing values.



Using the Replace Missing Value operator, each missing value for 80 rental properties is replaced with the average of that column. Filter Example with the class criterion "no_missing_attributes" can then be used to remove the last remaining missing values from the data. We can therefore have 23,800 samples with complete attribute data.

Task A: By using Generate attribute, we can do this:

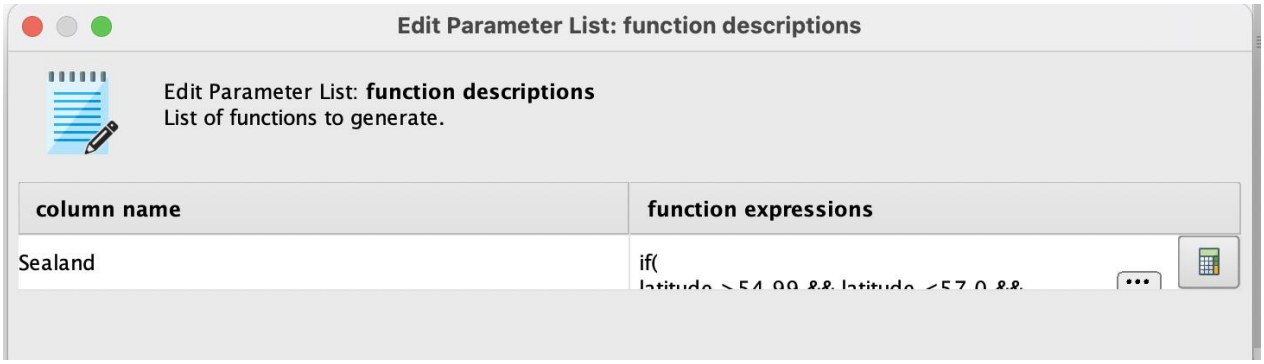


Figure 1 - Point Map of Sealand in Denmark versus the rest of Denmark

Task A: Task A: The map (Figure 1) shows property distribution in Sealand (blue) and another location (green). Per the Statistic report, properties in both regions:

Index	Nominal value	Absolute count	Fraction
1	Sealand	14438	0.607
2	Rest of Denmark	9362	0.393

Task B: Figure 2 compares price_per_night and person_per_night averages between Sealand and another Danish location. Sealand's prices are notably lower due to its 14438 rental properties, leading hosts to set competitive rates. Conversely, the other location with 9362 rental properties has higher prices due to limited options, making affordable housing harder to find.

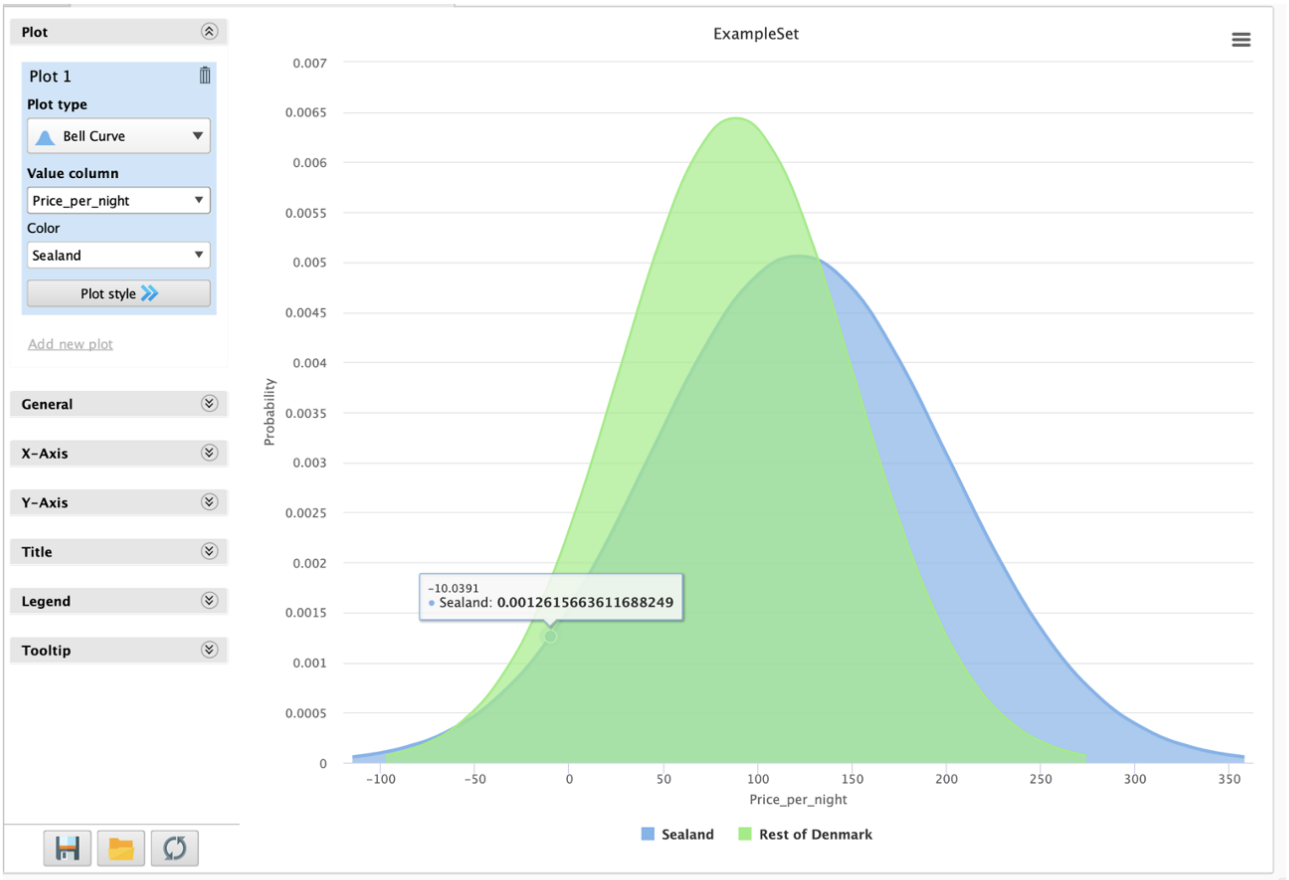


Figure 2: Bell Curve of Price for 1 night in Sealand and in the other place in Denmark

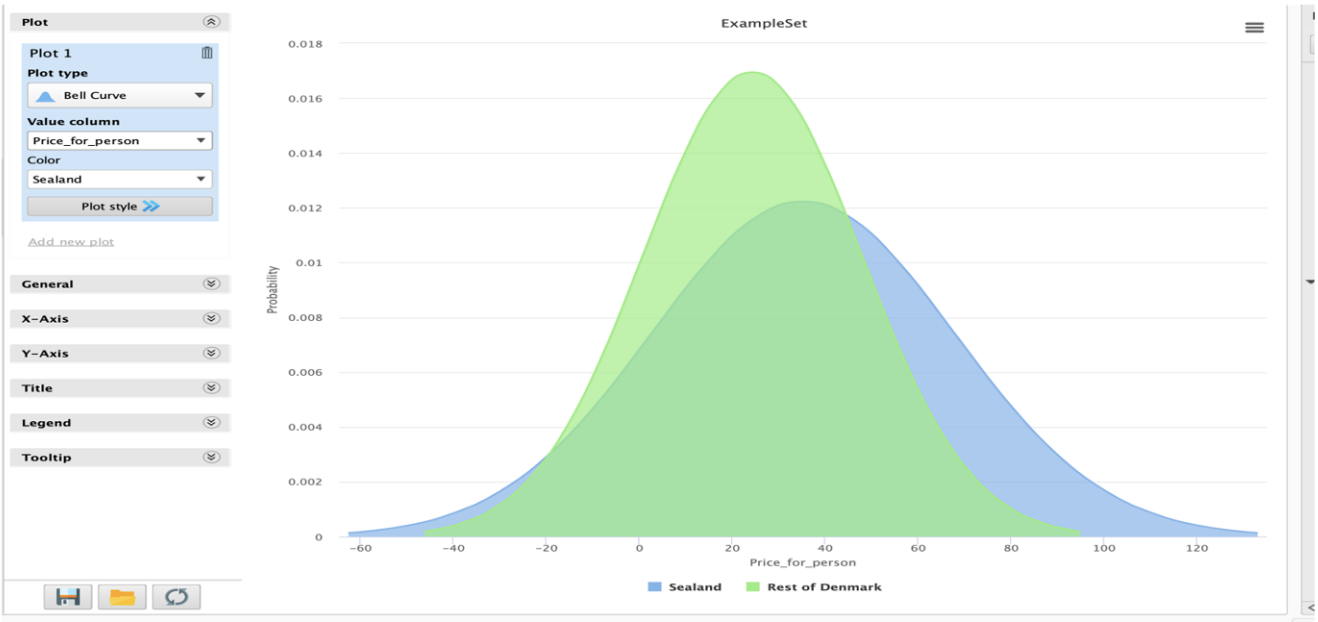


Figure 3: Bell Curve of Price for person in Sealand vs the other place in Denmark

Figure 4 shows the bar chart with Sealand's higher average overall satisfaction (2.7) compared to others (2.2), indicating more contentment with Sealand's rental houses. This highlights three key differences in Sealand's rentals versus the rest

of Denmark: lower prices, making them cost-effective options, and higher occupant satisfaction, solidifying its status as a preferred rental destination.

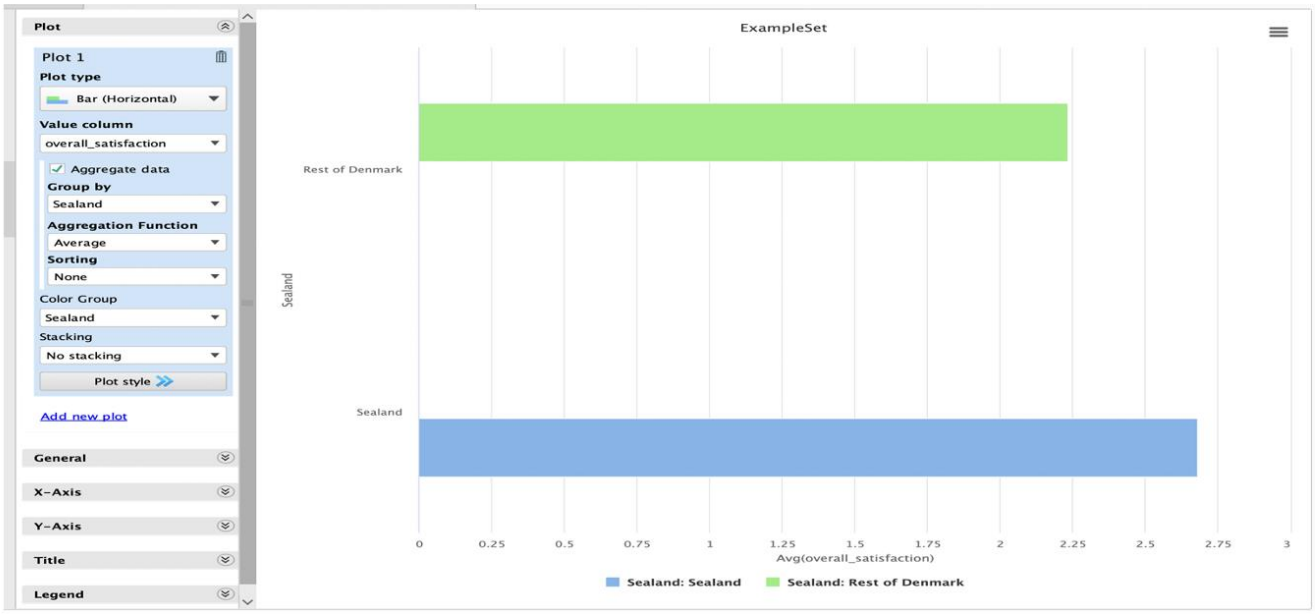
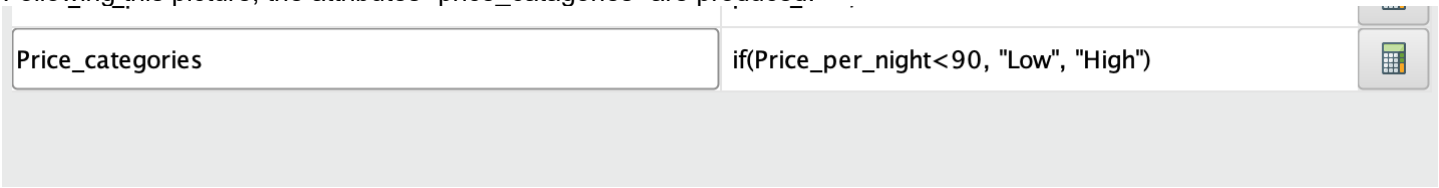


Figure 4: Bar Chart of Overall satisfaction in Sealand vs the rest of Denmark

Predictive modelling

Following this picture, the attributes "price_catagories" are produced.



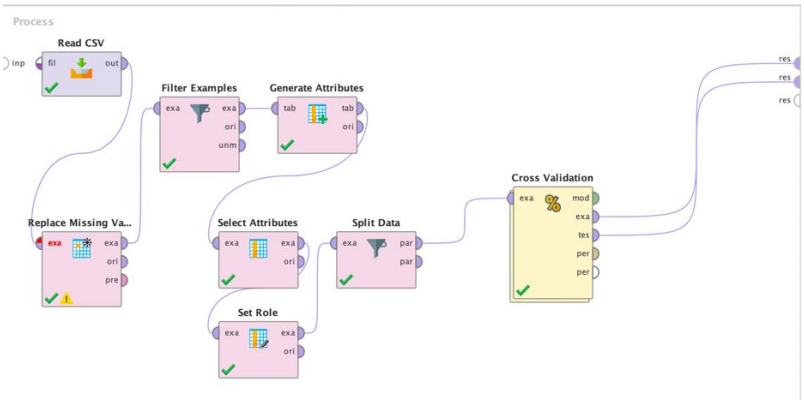
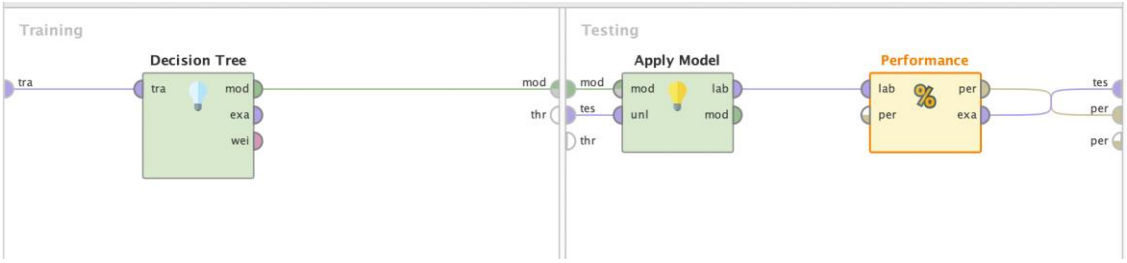
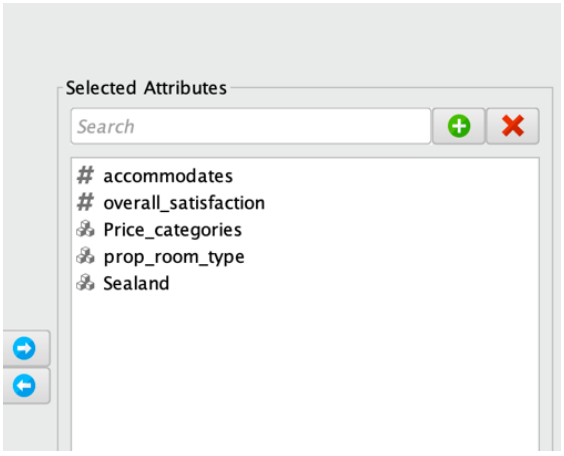


Figure 1



Similar to what was previously described, reading a CSV file produces attributes. The "Select Attributes" phase is used to selectively select pertinent attributes by referring to this diagram:



The "Price_catogories" are categorized as "Low" or "High." Initially, a "Set Role" operation assigns a special role "label" to the new attribute "Price_catagories." The data is split using a 0.7/0.3 ratio for training and testing. Statistical performance

is evaluated with 10-fold Cross-validation. Despite using a Decision Tree Model with Cross Validation for prediction, results were unreliable. Switching to a Vote Operator for majority voting from three models also proved ineffective.

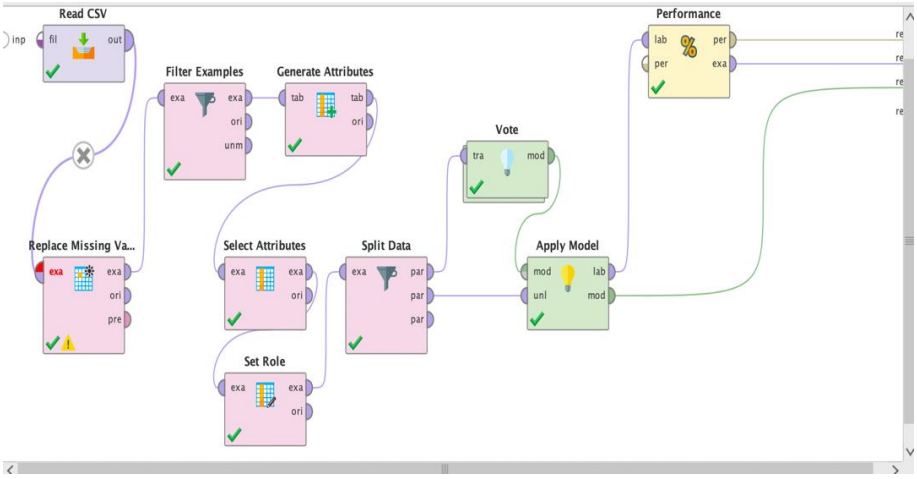


Figure 6

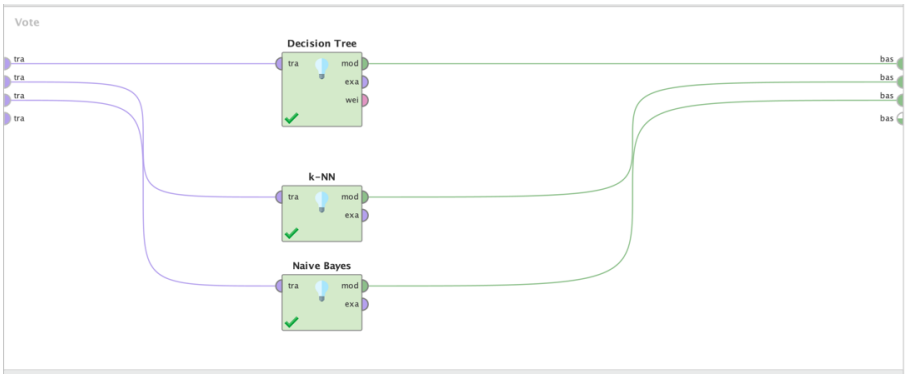


Figure 7

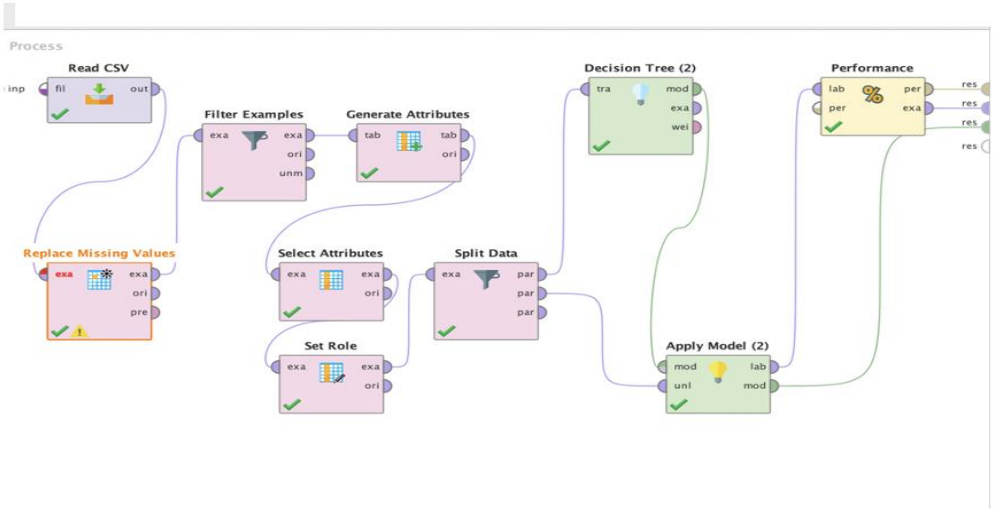
PerformanceVector

```
PerformanceVector:
accuracy: 74.66%
ConfusionMatrix:
True:    Low     High
Low:     2617    860
High:    949     2714
kappa: 0.493
ConfusionMatrix:
True:    Low     High
Low:     2617    860
High:    949     2714
```

Figure 8

I used the Decision Tree Model (Figure 9) without cross-validation to test its applicability to both continuous and categorical data, and it proved effective (Figure 10)

Figure 9



PerformanceVector

```
PerformanceVector:
accuracy: 74.83%
ConfusionMatrix:
True:    Low      High
Low:     2617     848
High:    949      2726
kappa: 0.497
ConfusionMatrix:
True:    Low      High
Low:     2617     848
High:    949      2726
```

Figure 10

Model evaluation and improvement

	With Cross Validation	No Cross Validation
Decision Tree	-Accuracy: 74.71% -Kappa: 0.494	-Accuracy:74.83% -Kappa:0.497
K-NN	-Accuracy: 71.57% -Kappa: 0.431	-Accuracy:70.95% -Kappa: 0.419

Regarding accuracy and kappa, the Decision Tree model surpasses the K-NN model, which is 74% in accuracy. It is suggested that the business utilize the Decision Tree model to categorize the budget because it has shown to be the most effective predictive model.

When using a cross-validation k=10 combined with a 70:30 data split, reliability is improved since each time the cross-validation process is run, it uses a different 70:30 sample.