Report on observation of wholesale customer by Data Modeling

COS60008 Introduction to Data Science
Year, 2024 Semester 1
Assignment 2 Report
Name: Pornkanok Tantewee
Student ID: 103503476
Email: 103503476@student.swin.edu.au
10 May 2024

Abstract

The report analysis a spending of customer through wholesale to gain insight into customer behaviour. The dataset contains information about annual spending of customer who purchase different products through this wholesaler. The dataset was cleaning and transformation including addressing missing value to ensure the data is ready to model performing. Exploratory data analysis revealed that correlations positive correlation and negative correlation and encoding numerical data to categorical data, the dataset is prepared for analysis. The dataset exploratory data analysis revealed positive correlations between spending and related factors such as customer region. Moreover, negative correlation is revealed. This report investigates insight into customer spending annually. The dataset is split into three suites before model building. Two algorithms are chosen to build model are K-nearest neighbors algorithm and Decision Tree algorithm build model for this dataset. The models were trained and evaluated on the dataset to predict customer region. The Decision Tree model tends to have higher accuracy whereas K-nearest neighbors model had a better precision. The finding

Table of Content

Introduction

The report aims to communicate the findings of Task 1 to 3 of the assignment, which involved analysis an annually customer spending which expend to wholesaler. This dataset is explored to understand spending behavior of customer per year through key relationships between attributes and region were identified and type of customer can be exploring impact to another feature. In addition, two models is train and evaluated to predict the region of customers.

Task 1: Data Acquisition & Preparation

Dataset contains information about customer spending behaviour to a wholesaler, including a vary attributes related to customers and earn outcome. The following detail is overview of the dataset.

- There are seven features. The 7 features contain six numerical features and one categorical feature.
- The Region feature is label for this dataset.
- There are 439 rows.
- Fresh is a numerical attribute that the customer spending in fresh product per year.
- Milk is a numerical attribute that the customer spending in milk product per year.
- Grocery is a numerical attribute that the customer spending in grocery product per year.
- Frozen is a numerical attribute that the customer spending in frozen product per year.
- Detergents Paper is a numerical attribute that the customer spending in detergents and paper per year.
- Delicatessen is a numerical attribute that the customer spending in Delicatessen per year.
- Channel is a categorical attribute that detail the type of customer (Retail and)
- Retail and Horeca is value in Channel attribute.
- Region contain three value like Lisbon, Oporto and Other region.

Next, conduct essential data preparation to support the ensuring the dataset is ready to be explored and built model efficiently.

1.1. Import Libraries

- numpy is used to numerical operation library for array and mathematic functions.
- Matplotlib is used to visual graph.
- Seaborn is statistical data and visualization library which is built on Matplotlib
- Pandas is used to manipulate data library for working with DataFrame.
- Sklearn.model_selection.train_test_split is used to split dataset into training and testing sets.
- sklearn.metrics is used to provide a variety of metric for evaluating machine learning.
- sklearn.neighbors import KNeighborsClassifier is used to KNeighbors model creating
- sklearn.tree provide Decision tree model to train and test dataset.

1.2. Data Loading and Cleaning

- The dataset is checked to ensure that no missing value by isna()
- The data is generally looked to observe overview of dataset by .info()
- The data is check statistic value in this dataset by .describe
- The categorical such as Channel is changed from int to categorical data. For example, dfT['Channel'].replace(1, 'Horeca') 1 is change from 1 to Horeca which is comfortable for reader to read clearly.
- Unique value is checked to ensure that not found missing categorical data.

Task 2: Data Exploration

2.1. Explore key features

1.Channel

2.Region

3. Milk

4.Grocery

5.Frozen

6.Fresh

7.Detergen_Paper

8. Delicassen

2.1.1. Channel

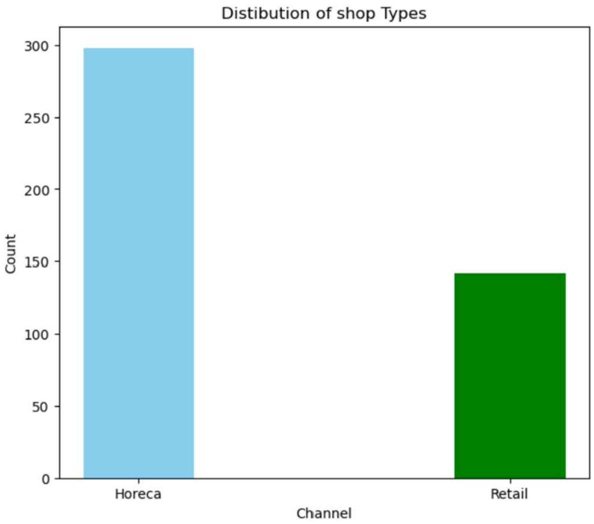


Figure 1-Distribution of shop Type

- Customer in Horeca spend much more than retail to buy products from wholesaler.
- It possible to focus on Horeca because of higher shop number in case offer promotion for them.

2.1.2. Region

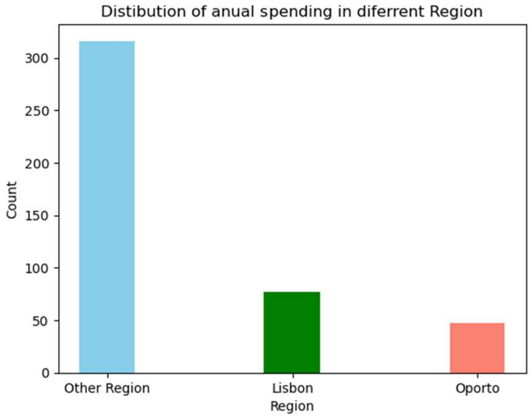


Figure 2- Distribution of annual spending in different Region

- There is the most spending in other regions, following by Lisbon and Oporto.
- More than 300 customers spend money for wholesaler per year in another region.
- Few customers around 50 spend money per year to wholesaler in Oporto

2.1.3. Milk

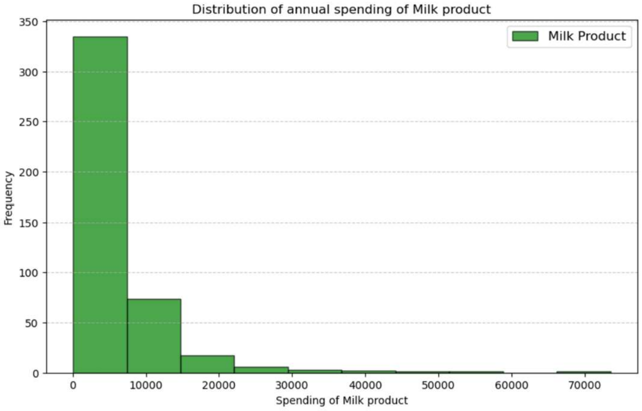


Figure 3- Distribution of annual spending of milk product

- The milk product is bough in wild rang of money. However, most customers spend in range is not over 8000 per year for milk product.
- Many customer spend in range 0-8000 this can be indicate that

2.1.4. Grocery

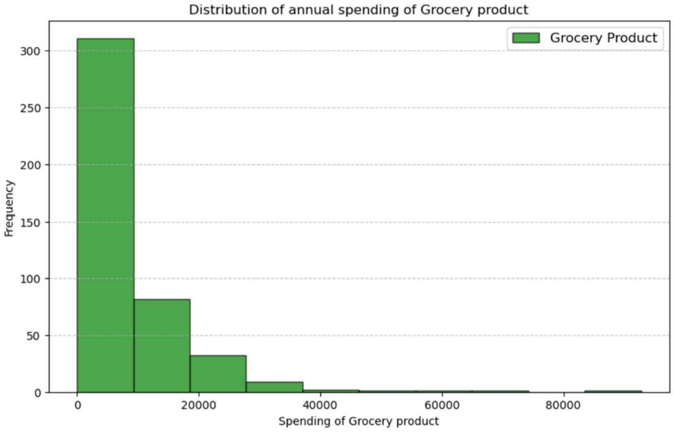
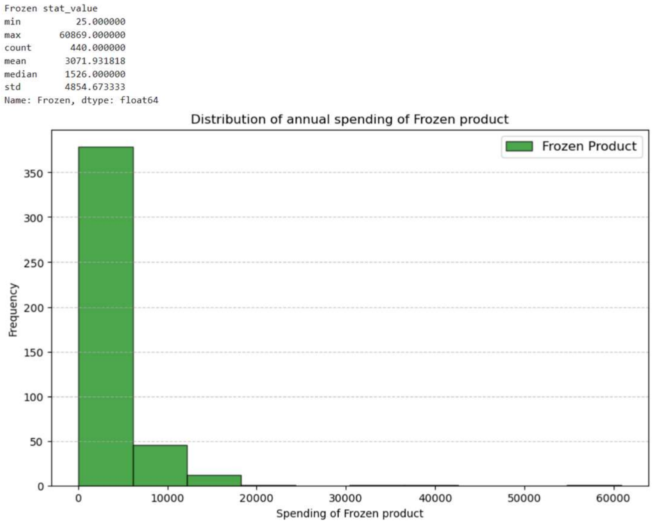


Figure 4- Distribution of annual spending of Grocery product

- The grocery product is bough in wild rang of money. However, most customers spend in range is not over 5000 per year for milk product.

2.1.5. Frozen

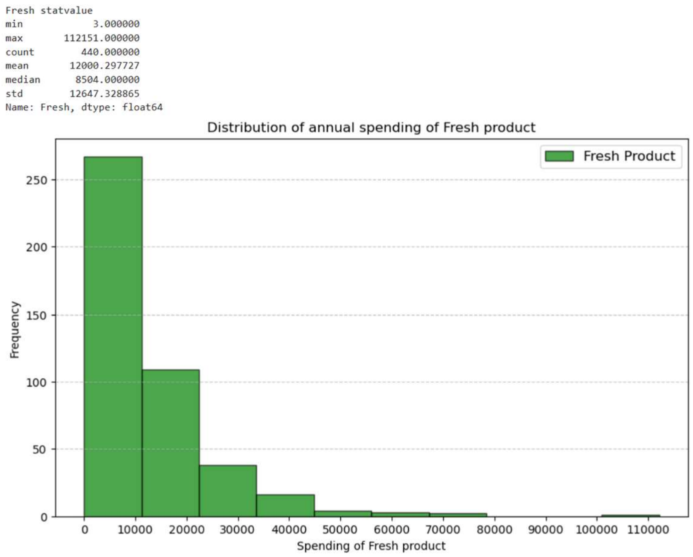


```
Frozen stat_value
min          25.000000
max       60869.000000
count       440.000000
mean       3071.931818
median     1526.000000
std        4854.673333
Name: Frozen, dtype: float64
```

Figure 5-- Distribution of annual spending of frozen product

- Most customer buy frozen product per year less than 5000.
- The frozen product is bought by customer less than 65000.

2.1.6. Fresh

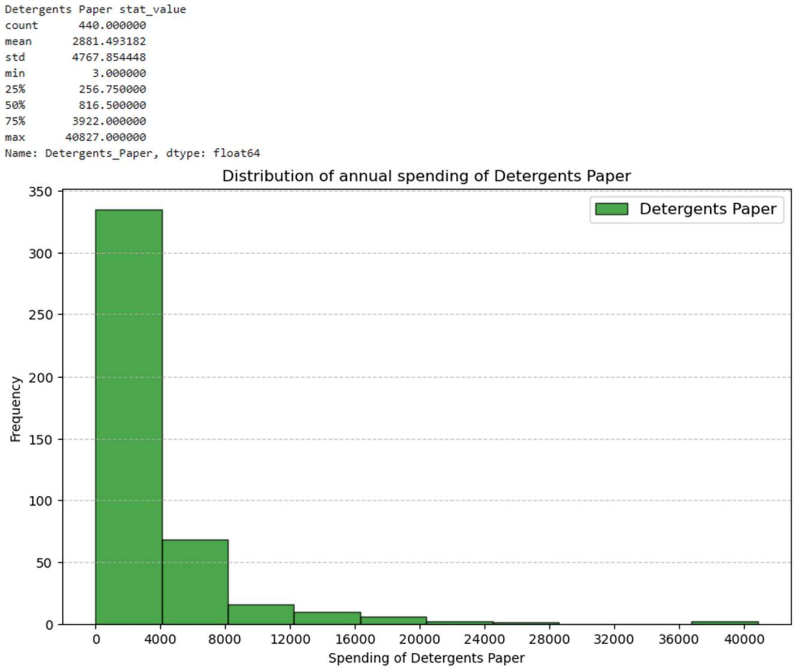


```
Fresh statvalue
min            3.000000
max       112151.000000
count        440.000000
mean       12000.297727
median      8504.000000
std        12647.328865
Name: Fresh, dtype: float64
```

Figure 6- - Distribution of annual spending of Fresh product

- Fresh product is spent by customer up to 120000 per year.
- Most customer buy fresh product the most in range less than 15000.

2.1.7.Detergen_Paper

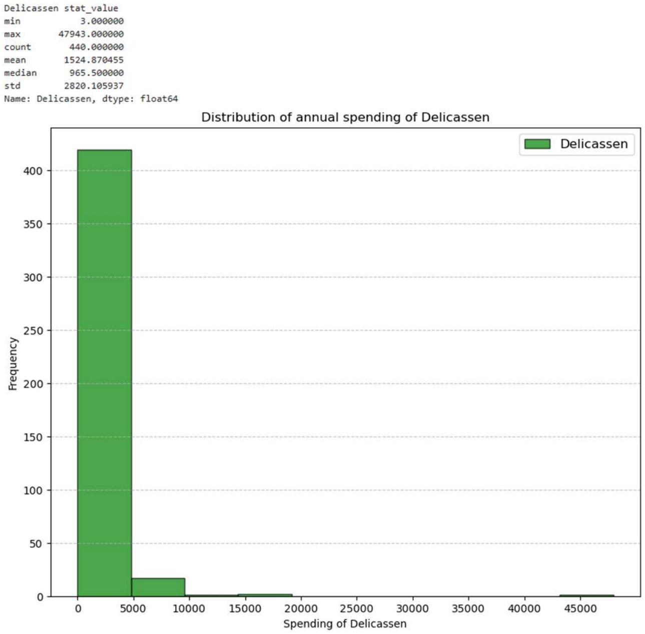


```
Detergents Paper stat_value
count      440.000000
mean      2881.493182
std       4767.854448
min          3.000000
25%        256.750000
50%        816.500000
75%       3922.000000
max      40827.000000
Name: Detergents_Paper, dtype: float64
```

Figure 7-- Distribution of annual spending of Detergents Paper product

- The detergents and paper are the most spending in money range less than 4000 per year.
- A few customers pay more than 40000 per year to buy Detergents paper.

2.1.8. Delicassen



```
Delicassen stat_value
min           3.000000
max       47943.000000
count       440.000000
mean       1524.870455
median      965.500000
std        2820.105937
Name: Delicassen, dtype: float64
```

Figure 8-- Distribution of annual spending of Delicassen product

- The Delicassen are the most spending in money range less than 5000 per year.
- A few customers pay more than 48000 per year to buy Delicassen product.

2.2. Explore relationship between pair columns.

2.2.1. Frozen - channel

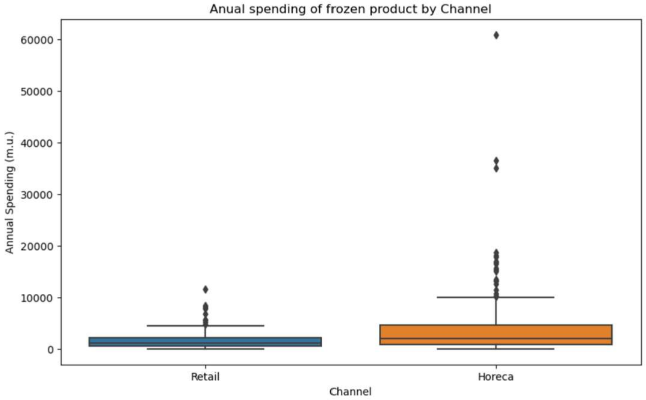


Figure 9 boxplot Frozenfood by Channel

- Horeca have more outlier that fall outside the range of the whiskers than Retail.
- Annual spending of frozen product is expended in low range.
- Both Retail and Horeca have similar spending of frozen product.

2.1.2. Grocery – Detergents Paper.

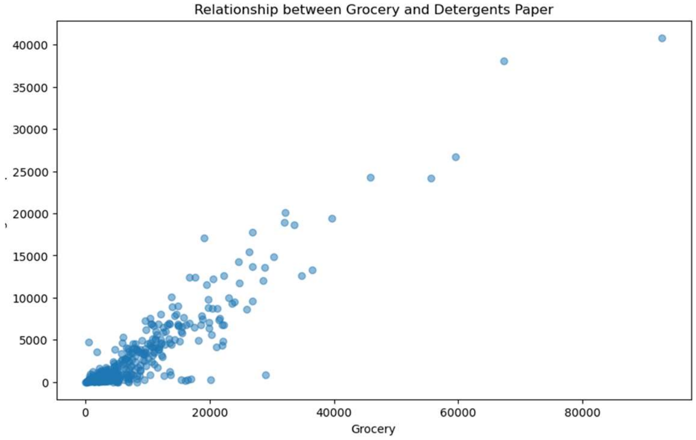


Figure 10- Relationship spending Grocery and Detergents Paper

- There is positive correlation between Grocery and Detergents Paper.
- Customers are more like to buy grocery product and Detergents Paper the same time because they are product that is important for daily life.

2.2.3.Grocery-Milk

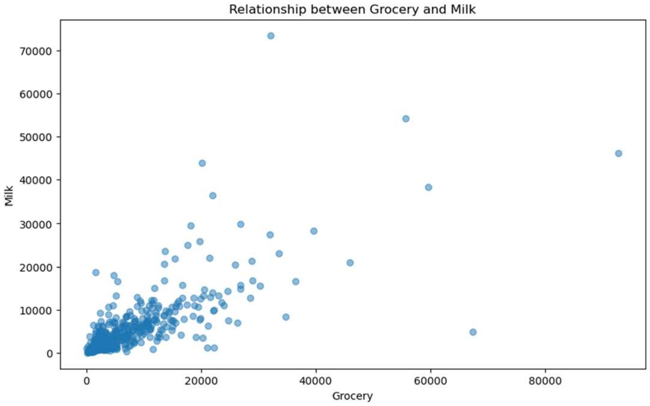


Figure 11Relationship spending Grocery and Milk

- There is positive correlation between Grocery and Milk.
- The finding is found that milk spending increase while Grocery product is also increasingly bough as well.

2.2.4. Channel - Fresh

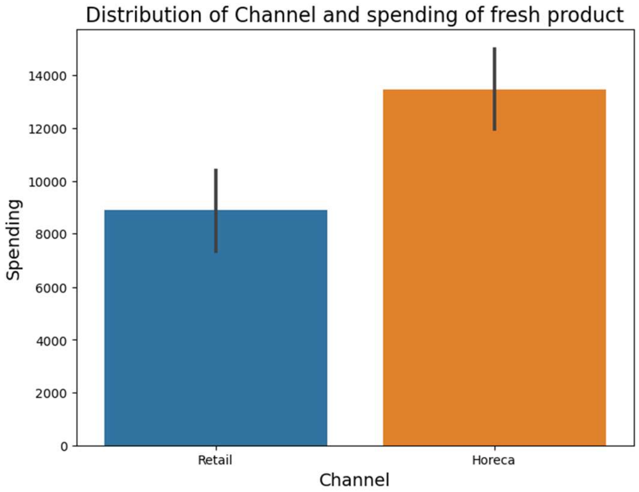


Figure 12-Distribution of channel and spending of fresh product

- Spending of fresh product is higher than retail.
- The finding found that restaurant tend to buy fresh product for customer than retail.
- The retail shop can spend money less than 8500 while Horeca can spend high level of money around 13000 per year.

2.2.5. Milk – Detergent and Paper

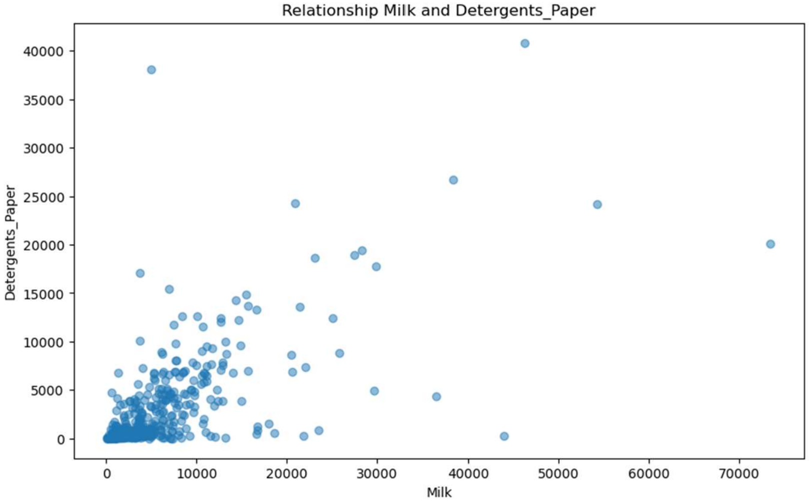


Figure 13- Relationship of annual spending between Milk and detergentd paper

- There is positive correlation between Detergent and Milk product in figure
- Both of them are bough in low range of spending.

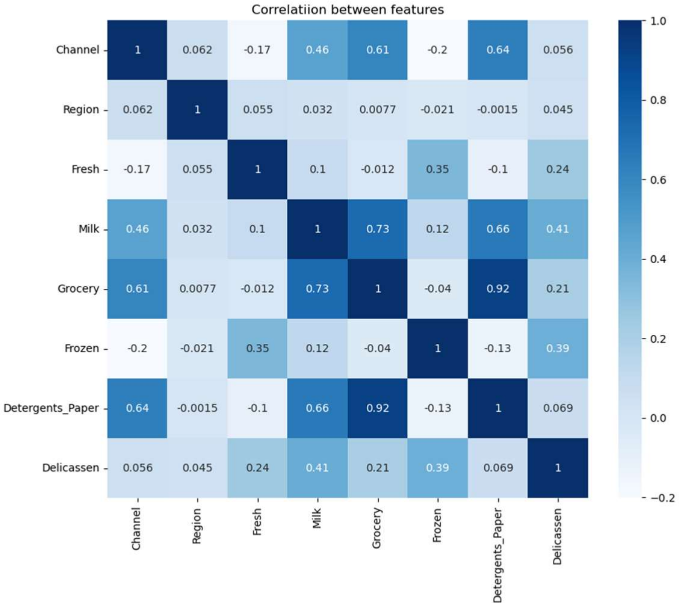


Figure 14- Correlation between feature

Figure14 show correlation between feature can be analysis to find spending trend of customer in different product.

2.3. How do differences in each product spending behavior vary among customers from different channels?

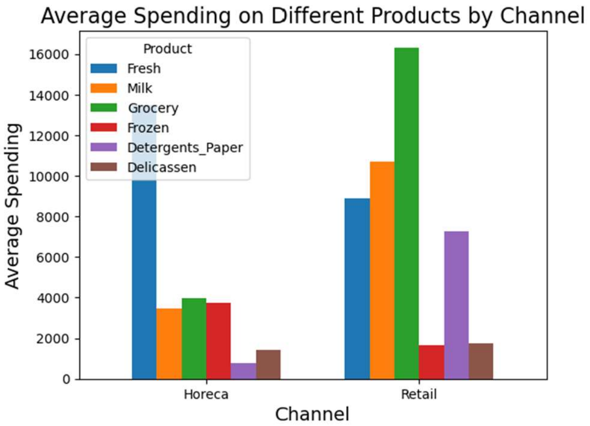


Figure 15-average Spending on different product by channel

- The retail spend high level of money to Grocery product which is particularly true because the retail shop is important to stock grocer product for their customers while Horace(restaurant) spend high level of money for fresh product.
- Retail channels spend money to buy wholesaler product more than Horace(restaurant).
- The Horace(restaurant) do not spend another except fresh product.
- Frozen product and Delicassen are bough in low spending which is not over 2000 per year.
- Detergent_paper product and Dellicassen are bough in low spending by Horace(restaurant).

Task 3
3.1. Data Splitting
X contain seven features while y is label that is region.

3.1.1. Suite1: 50% for training and 50% for testing

- The raw data or full data is split into 2 part from 440 datapoints to training data 220 and testing data 220 datapoints
- 50% for training data for model building and another part 50% for model testing.

3.1.2. 60% for training and 40% for testing

- The raw data or full data is split into 2 parts from 440 datapoints to 264 datapoints that is training data while 176 datapoints for testing data.
- 60% for training data for model building and another part 40% for model testing.
- Model should not see the testing data before testing or prediction.

3.1.3. 80% for training and 20% for testing

- The raw data or full data is split into 2 parts from 440 datapoints to 352 datapoints that is training data while 88 datapoints for testing data.
- 80% for training data for model building and another part 200% for model testing.

The key observations from these training data and testing data

- The split ration are accurately implemented with the training and test set sizes matching the specified 50/50, 60/40, 80/20 ration
- The random state (1) is used to ensure that when the code run the value is not change.
- The split is performed on full dataset including feature(X) and the target variable(y), ensuring that the distribution in both training data and testing data are representative overall data.

3.2. Model Performing
3.2.1. K-Nearest Neighbors (KNN) model
This algorithm is used for classification. The principal of working is that storing data point and label. When making predictions, it selects the K nearest neighbor based on distance and assign the class label for classification.

KNN model in Training Set
Suite 1 Performance

- The model has 71% of accuracy that is medium performance.
- The precision is 50.9% that is low precision.
- The recall is 71% that is moderated performance.
- F1-score is 59% that is medium performance

Suite 2 Performance

- The model has 70% of accuracy.
- The precision is 50%(Macro) that is low performance.
- The recall is 70% of recall value.
- F1-score is 58 % that is low performance.

Suite 3 Performance

- The model has medium accuracy only 70% that is medium performance.
- The precision is 59% is low precision.
- The recall is 70% is medium recall.
- F1-score is 60%.

KNN model in Testing Set
Suite 1 Performance

- The model has 72% of accuracy that is medium performance.
- The precision is 52% is low precision.
- The recall is 72% is medium recall.
- F1-score is 60%.

Suite 2 Performance

- The model has 73% of accuracy that is medium performance.
- The precision is 59% is low precision.
- The recall is 73% is medium recall.
- F1-score is 63%.

Suite 3 Performance

- The model has 73% of accuracy that is medium performance.
- The precision is 59% is low precision.
- The recall is 73% is medium recall.
- F1-score is 63%.

Decision Tree model in Training Set
Suite 1 Performance

- The model has 100% of accuracy that is perfect performance.
- The precision is 100% is low precision.
- The recall is 55% is medium recall.
- F1-score is 56%.

Suite 2 Performance

- The model has 100% of accuracy that is perfect performance.
- The precision is 100% is perfect precision.
- The recall is 100% that is perfect recall.
- F1-score is 100% that is high performance.

Suite 3 Performance

- The model has 86% of accuracy that is perfect performance.
- The precision is 87% is perfect precision.
- The recall is 86% that is perfect recall.
- F1-score is 87% that is high performance.

Decision Tree model in Testing Set
Suite 1 Performance

- The model has 61% of accuracy that is low performance.
- The precision is 63% is low precision.
- The recall is 61% that is low recall.
- F1-score is 62% that is low performance.

Suite 2 Performance

- The model has 51% of accuracy that is low performance.
- The precision is 55% is low precision.
- The recall is 51% that is low recall.
- F1-score is 53% that is low performance.

Suite 3 Performance

- The model has 55% of accuracy that is low performance.
- The precision is 59% is low precision.
- The recall is 55% that is low recall.
- F1-score is 56% that is low performance.

Evaluation and Insight
K-Nearest Neighbors (KNN) model

- All three suites, K-Nearest Neighbors (KNN) model tend to have a good accuracy that led to predict correctly.
- Precision is consistency lower than recall both in train and test set.
- A balance assessment of the model is imbalanced because recall and precision have different trend.
- This model tend to generalisation or good fitting with data set because There are similar result of prediction in training and testing.

Decision tree model

- All three suites, Decision tree model tend to have a high accuracy that led to predict correctly in training
- Recall is consistency higher recall.

- A balance assessment of the model is imbalanced because recall and precision are not balance.
- This model indicate overfitting because model can perform efficiently during training while model predict incorrectly with testing.
- Model have dramatically high in training in all four value(accuracy, precision, recall and F1-score)

3.3. Comparing two models.

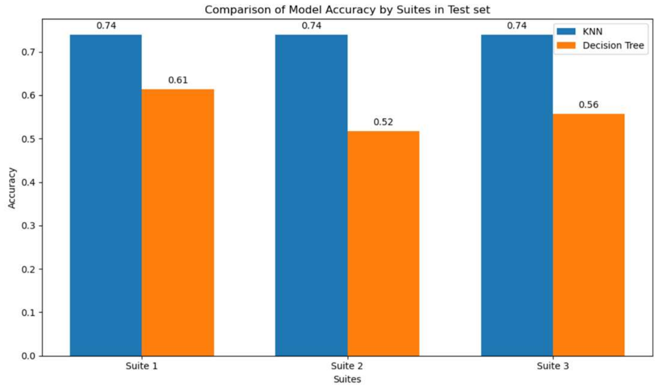


Figure 16- comparison of two Model accuracy by different suite in test set

Figure-16 show the accuracy of two model, Decision tree has higher accuracy than Decision tree model in all three suites. However, Decision tree model has lower accuracy than KNN model in three suites.

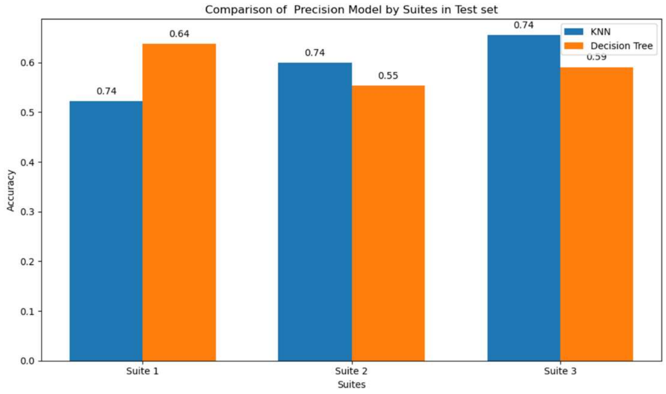


Figure 17-comparison of Precision in test set by different suites

Figure 17 show the precision in different suites of KNN and Decision tree model. The precision of KNN is higher than Decision tree in suite2 and suite3 while Decision tree model has higher precision than KNN in suite 1.

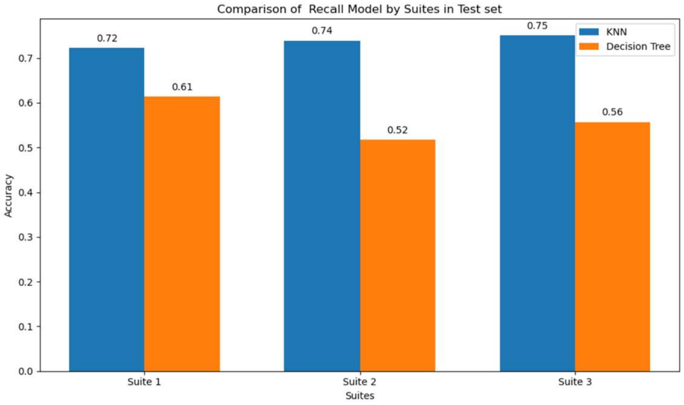


Figure 18-comparison of Recall in test set by different suites

Figure 18 shows comparing the recall in in different suites of KNN and Decision tree model on Micro average. KNN has higher precision in all three suites. KNN has higher than Decision tree. KNN has higher precision in three suites. However, KNN has much less high precision than Decision tree in suite3.Thus, KNN model tends to have higher recall than decision tree model.

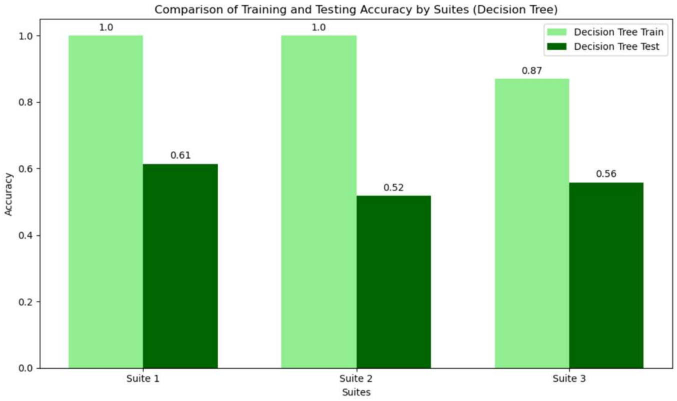


Figure 19- comparison of Accuracy in Decision tree modell by different suites.

Figure 19 shows comparison of training and testing in three suite of Decision tree model. This graph can indicate underfitting because the model can predict dramatically correctly during training. In the other hand, model can predict much less correctly in testing.

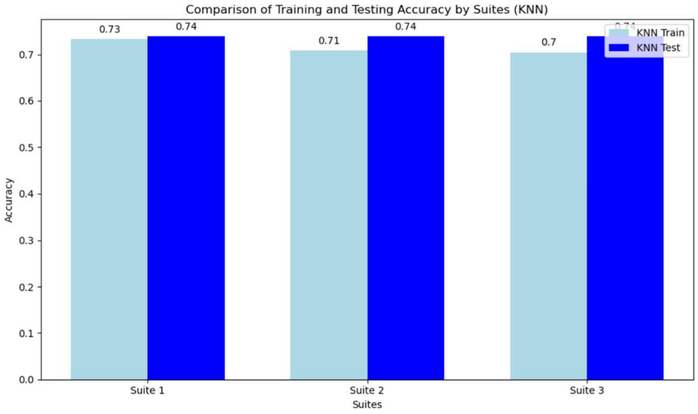


Figure 20-comparison of accuracy between training and testing by different suites

Figure 20 shows comparison of training and testing in three suit of K-Nearest Neighbors model. This graph can indicate generalization or good fitting because the predicting result is much similar both in traing and testing. This model tend to be used with other dataset. The model can perform with new dataset efficiently.

Confusion Matrix

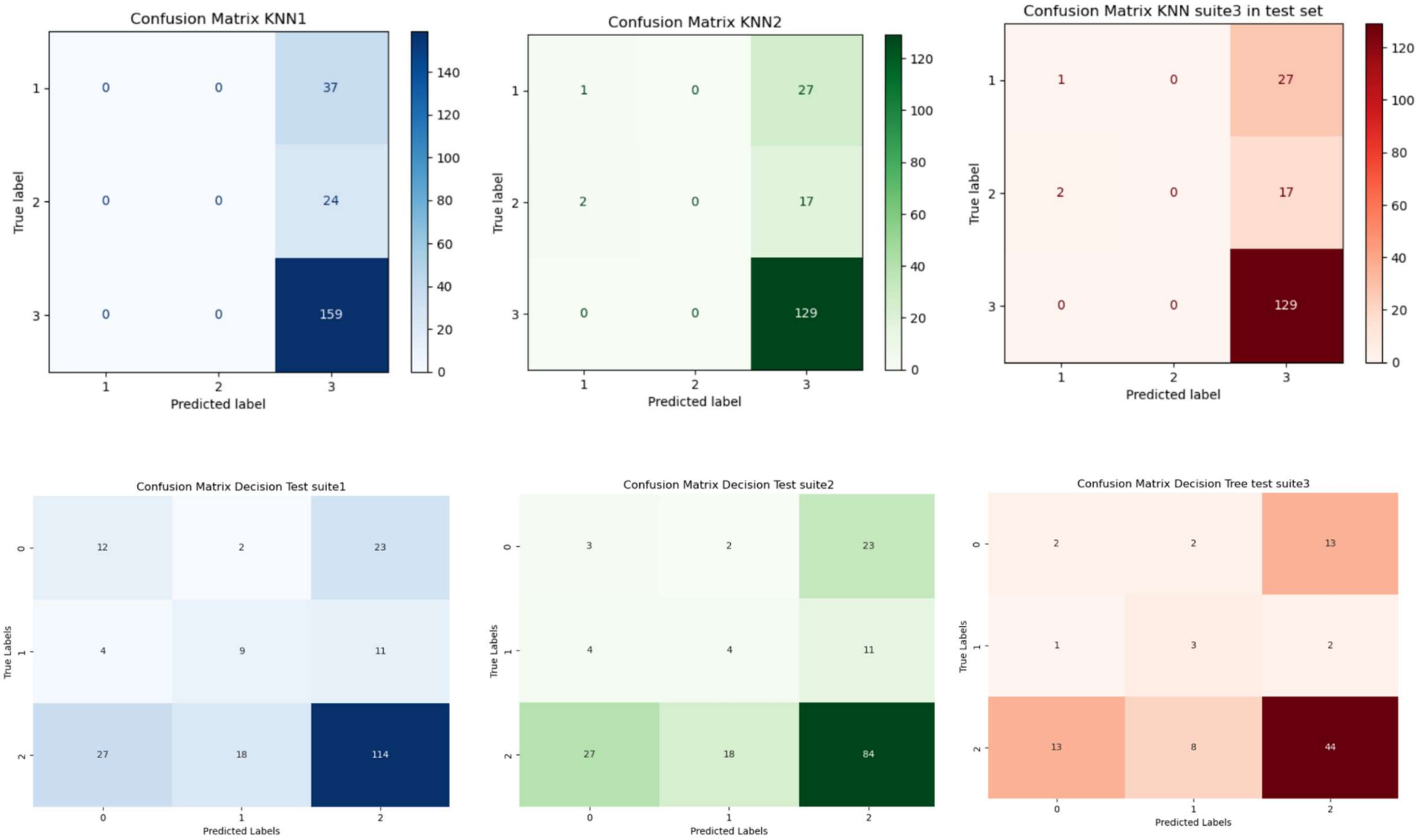


Figure 21- Confusion Matrix of two model in different suites

K-Nearest Neighbors and Decision Tree is compared.

- The K-Nearest Neighbors model has a higher true negative when compared to Decision tree models.
- Decision tree has higher true positive than K-Nearest Neighbors.
- There are low false negative in K-Nearest Neighbors and Decision Tree.
- Decision tree has higher false positive than K-Nearest Neighbors.

Conclusions

The analysis of wholesaler customer dataset reveals several insights. The exploratory data analysis illustrates a positive correlation between annual spending of Grocery – Detergents Paper, Grocery-Milk and Milk – Detergent and Paper that indicate the crucial insight of customer behaviour in different region and different channel. On the other hand, the negative correlation is found in detergent paper and frozen product with a few negative numbers. In terms of model, K-Nearest Neighbors (KNN) model demonstrate imbalance between precision and recall, while Decision Tree model perform better than K-Nearest Neighbors model with higher accuracy but show sign overfitting because it can predict much correctly in training whereas KNN can predict correctly in similar same in training and testing. Overall, the finding provides valuable insights for wholesaler to make market plan to enhance annual revenue for their customers.