|  |
| --- |
| **D:\fcuc-logo.png** |
| **FACULTY OF ENGINEERING & COMPUTING** |

|  |  |  |
| --- | --- | --- |
| **Programme** | : | Bachelor of Information Systems (Hons) in Business Management |
| **Academic Year** | : | 2019 |
| **Module** | : | Data Science and Big Data Analytics |
| **Module Code** | : | CISY3016 |
| **Module Leader** | : | Kwan Lee |
| **Assignment Type** | : | Report |
| **Intake/Group** | : | January 2018 (Year 3 Semester 1) |
| **Distribution Date** | : | Thursday, 17 October 2019 |
| **Submission Date** | : | Thursday, 21 November 2019 4:55 PM |

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student ID** | **Class Code** |
| Liew Wei-Jenn | B0885 | CISY3016 |

**Table of Contents**

[Assignment Feedback Form 3](#_Toc25540022)

[Task 1 – Problem Analysis 5](#_Toc25540023)

[Task 2 – Using Data Science to Address the Problem 6](#_Toc25540024)

[References 9](#_Toc25540025)

[Appendix 10](#_Toc25540026)

|  |
| --- |
| Assignment Feedback Form Data Science and Big Data Analytics |

|  |  |  |
| --- | --- | --- |
| Criteria | Marks | Comments |
| Analysis the Problem faced by the Company | /5 |  |
| Discussion | /10 |  |
| Format Report | /5 |  |
|  | Total |  |

Marks:

|  |  |
| --- | --- |
| **General Comments:** | |
|  | |
| **Assessor’s Signature**: | **Date:** |
| **Name: Kwan Lee** | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Section: Documentation**  **/20** | | **Excellent** | **Good** | **Satisfactory** | **Marginal** | **Poor** | **Comments** |
| **(Out of 5)** | Analysis the Problem faced by the Company |  |  |  |  |  |  |
| **(Out of 10)** | Discussion |  |  |  |  |  |  |
| **(Out of 5)** | Format Report |  |  |  |  |  |  |

# Task 1 – Problem Analysis

In the case study, Domino’s had encountered a problem of unauthorized vouchers being used by customers in their purchases. Through their database, they had detected an unusual pattern of purchases which involved many irregularities in the value of orders that were received. There were high-volume orders with abnormally low billing amount. They discovered that it was the result of customers using free pizza voucher codes that were obtained from unauthorized third-party sources to make their purchases.

Rather than refusing service to the customers involved and refunding their money, Domino’s made matters worse by filing a police report which resulted in their customers being interrogated by the police. This decision received severe backlash from the customers as many of them felt that they had done nothing wrong. Indeed, the customers had every right to feel upset by Domino’s hasty actions. While the issue of unauthorized vouchers is a serious matter, the blame should not fall on the customers’ shoulders. They had technically committed no crime nor done anything illegal as they were not the ones who supplied the voucher codes and were just using them. Domino’s should have gone after the suppliers of the unauthorized vouchers instead as they were the actual source of trouble.

The customers were merely taking advantage of the fundamental logic flaws in the company’s online ordering system – flaws that existed long before Domino’s went on a high alert for unusual pattern of purchases. It was these flaws that allowed them to use a voucher code to get a hefty discount even though the code itself is illegitimate. A fully functional and fool-proof system should have been able to detect as well as reject the usage of such unauthorized voucher codes. Thus, the pizza company should turn their attention to the weaknesses in their online ordering system and put in the necessary effort to fixing the flaws.

# Task 2 – Using Data Science to Address the Problem

Data science can be used by Domino’s to resolve the flaws in their online ordering system. The first part of the process is to state and refine the questions that needs to be asked. Before choosing which questions to ask, it is important to take into consideration the different types of questions that exists. Altogether, there are six types: descriptive, exploratory, inferential, predictive, causal, and mechanistic.

Descriptive questions are used to summarize the characteristic of a dataset [1]. Examples include determining the proportions of male and female students in a college and determining the mean amount of sales of a shop per day. The point is to just obtain the results and not to interpret those results. Exploratory questions are used for the purpose of analyzing the data to identify the patterns, trends, or relationships between the variables [1]. For instance, whether playing video games have an impact on a person’s health. This type of questions helps to generate hypotheses rather than test them.

An inferential question is when a proposed hypothesis is restated as a question which would be answered through analyzing a set of data that is different from the one in the exploratory analysis [1]. For an example, analyzing a representative sample of teenagers in Malaysia. The purpose is to compare it to the results from the exploratory analysis in order to determine whether the association that was previously observed holds true in both analyses. Thus, an inference can be made on the entire population based on the analysis that was performed on the representative sample. Predictive questions, as its name implies, are used for predicting future behaviors, actions, or occurrences [1]. Examples include what types of students will take an IT course and which age group will be the most interested in purchasing the latest iPhone. These questions are concerned with determining the potential outcomes in the future rather than the causes of the outcomes themselves.

Causal questions are used for determining the effect of one variable on another variables [1]. For instance, what is the effect of exercise on a person’s cholesterol level. This type of questions will help determine the effects of one factor on another by comparing of the outcomes of the subjects in two different states – i.e. the cholesterol levels of the group of people who exercise often and the group who don’t. A mechanistic question asks how exactly a factor causes an effect [1]. For an example, how does frequent exercise lead to low cholesterol level. These questions help to explain how one variable can affect another, rather than just identifying their relationship.

Descriptive, exploratory, inferential, and predictive questions are suitable for this case study. For instances, descriptive questions such as ‘what is the average monthly sales of Domino’s outlet in Kepong?’; exploratory questions like ‘whether the usage of pizza vouchers have an impact on sales?’; inferential questions such as ‘whether the usage of pizza vouchers have a negative impact on sales?’; and predictive questions like ‘which pizza voucher is likely to be legitimate?’. These questions will eventually lead to the cause of unusual pattern of purchases – the use of unauthorized voucher codes. The ultimate goal is to determine which voucher codes are genuine and which are not.

Once the questions are properly defined, the next step is to decide which type of learning method should be used: supervised or unsupervised. Supervised learning method refers to when the learner is provided with target information along with a few examples [2]. For instance, which group of customers is likely to discontinue the service after the price increase. In this case, the target is specified – will a customer discontinue the service after the price is increased? A training set of labelled data is then created such as the type of customer, their income, and the price increase. The learner will determine the relationship between all those input data and the output data which is the continuation/discontinuation of service after price increase.

Unsupervised learning method refers to when a set of examples is provided but there is no target information [2]. For an example, what groups can the customers be divided into. In this case, no purpose of the learning is specified which leaves the learner to draw its own conclusions about the common characteristics between the customers and sort them into groups. Perhaps the customers will be grouped based on their age, gender, income, or their loyalty status. The point is to let the learner determine the structure of the data without the use of labelled data.

Supervised learning method is appropriate for the purpose of this case study as it requires the use of labelled data such as average monthly sales and the usage of pizza vouchers. The target is to determine the impact of the vouchers on the sales. There are a couple of techniques that are mainly used in supervised learning – classification and regression [2].

Classification is when the target output is categorical [3]. This model attempts to form some conclusion based on the values that are observed. It will try to predict the value of the outcomes of the input data. Examples include emails being labelled as ‘spam’ or ‘non-spam’ and credit card details being labelled as ‘genuine’ or ‘fraudulent’. In other words, classification predicts the value of a new input data based on the training set. Regression is when the target output is continuous [3]. This model predicts the real value data rather than class labels. Examples include predicting stock prices and predicting expected sales. In other words, the values predicted by regression cannot be classified into discreet groups but are actual real values instead.

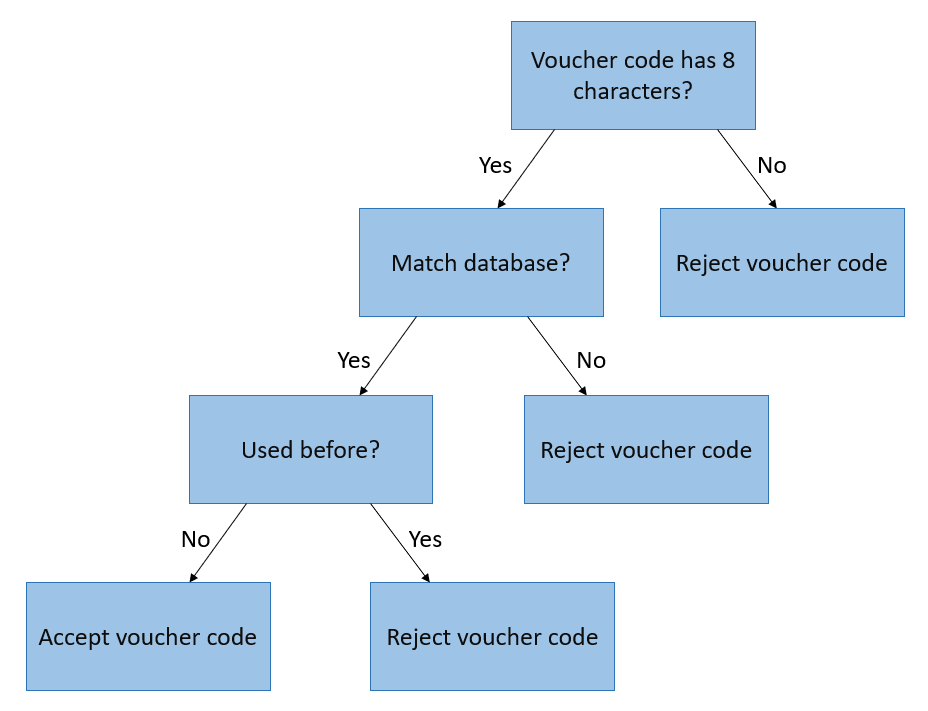
In this case study, the classification model is suitable to be used to predict the legitimacy of the pizza vouchers as the values can be classified into ‘legitimate/authorized’ and ‘illegitimate/unauthorized’. To be more specific, decision trees classification will be used. Decision trees are recursive partitioning algorithms that presents the patterns of an underlying dataset in a tree-like structure [4]. It helps users to make step-by-step decisions by considering the different factors affecting them and the outcomes that could result from them. The use of a decision tree can help the system to accept or reject a voucher code by determining its legitimacy.

Assuming the voucher code has been fixed to 8 characters, the decision tree would first start with a question such as ‘Does the voucher code have 8 characters?’ which would be answered with ‘Yes’ or ‘No’; answering ‘Yes’ would lead to the next question ‘Does the voucher code match the ones available in the database?’, which would also be answered with ‘Yes’ or ‘No’; ‘Yes’ would lead to another question like ‘Has the voucher code been used before?’, which would be answered with ‘Yes’ or ‘No’; ‘Yes’ would prompt the system to reject the code as it had already been used by another customer while ‘No’ would prompt the system to reject it. Answering ‘No’ to the first and second question would straight away prompt the system to reject the code as it is not a legitimate code according to the database. An illustration of this example (Figure 1) is attached in the appendix.

# References

1. R. Peng and E. Matsui, "The Art of Data Science", Bookdown.org, 2017. [Online]. Available: https://bookdown.org/rdpeng/artofdatascience/types-of-questions.html. [Accessed: 24- Nov- 2019]
2. F. Provost and T. Fawcett, Data Science for Business, 1st ed. Sebastopol, California: O'Reilly Media, Inc., 2013, pp. 24-25.
3. B. Baesens, Analytics in a big data world. Hoboken, New Jersey: John Wiley & Sons, Inc., 2014, p. 35.
4. B. Baesens, Analytics in a big data world. Hoboken, New Jersey: John Wiley & Sons, Inc., 2014, p. 42.

# Appendix



**Figure 1: Decision Tree – Accept/Reject voucher codes**