**Assessment Cover Sheet**

This Assessment Cover Sheet is only to be attached to

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ASSESSMENT DETAILS** | | | | | | |
| Unit title | | Introduction to Data Science | Tutorial /Lab Group | - | | Office use only |
| Unit code | | COS10022 | Due date | 07th Dec 2020 | |  |
| Name of lecturer/tutor | | Vong Wan Tze | | | |  |
| Assignment title | | Assignment 3 | | | | Faculty or school date stamp |
| **STUDENT(S) DETAILS** | | | | | | |
| Student Name(s) | | | | | Student ID Number(s) | |
| (1) | Masrur Rahman Zahin | | | | 101214608 | |

hard copy submission of assessments.

|  |  |  |  |
| --- | --- | --- | --- |
| **DECLARATION AND STATEMENT OF AUTHORSHIP** | | | |
| 1. I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment. 2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made. 3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/tutor concerned. 4. I/we have not previously submitted this work for this or any other course/unit. 5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.   I/we understand that:   1. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited. | | |
| **Student signature/s**  I/we declare that I/we have read and understood the declaration and statement of authorship. | | |
| (1) | Masrur |

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **https://www.swinburne.edu.my/current-students/manage-course/exams-results-assessment**

Copies of this form can be downloaded from the Student Forms web page at **https://www.swinburne.edu.my/current-students/manage-course/exams-results-assessment/how-to-submit-work.php**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **http://www.swinburne.edu.my/images/2013/logo.jpgSwinburne University of Technology Sarawak Campus**  **Faculty of Engineering, Computing and Science**  **COS10022 Introduction to Data Science**  Assignment 3 - Semester 2, 2020 | | | | |
|  | | | | |
| **Assessment Weighting** | | | : | 40% |
| **Due Date** | | | : | **Monday, 07th December 2020 at 11.59 pm (Malaysian local time)** |
|  | | | | |
| **Assignment Objectives** | | | : |  |
| 1. Appreciate the roles of data science and Big Data analytics in organisational contexts.  2. Compare and analyse the key concepts, techniques and tools for discovering, analysing, visualising and presenting data.  3. Describe the processes within the Data Analytics Lifecycle.  4. Analyse organisational problems and formulate them into data science tasks.  5. Evaluate suitable techniques and tools for specific data science tasks.  6. Develop and execute an analytics plan for a given case study. | | | | |
|  | | | | |
| **Instructions** | : |  | | |
| * This is an **INDIVIDUAL** assignment. | | | | |
| * You are expected to attempt all the questions. | | | | |
| * Write your answers in the grey box provided at the end of each question. | | | | |
| * The total available marks in this assignment is 100 marks. | | | | |
| * Adjust the size of the grey box to include all your answers. | | | | |
| * Submit your work as a PDF file via Canvas > COS10022 > Assignment 3. | | | | |
| * **Plagiarism shall immediately result in zero marks.** | | | | |
|  | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem Set 1: Data Analytics Lifecycle** | | | |
| [10 marks] The Data Analytics Life Cycle (DALC) comprises of six phases:    Fill in a phase of the DALC that is the most relevant to the data science concept below. | | | |
|  | **Example: Model Building** |  | **Overfitting** |
|  |  |  |  |
|  | Model Building |  | **Leave-one-out cross validation** |
|  |  |  |  |
|  | Data Preparation |  | **Analytic Sandbox** |
|  |  |  |  |
|  | Communicate Results |  | **Data-Ink Ratio** |
|  |  |  |  |
|  | Model Building |  | **Feature Selection** |
|  |  |  |  |
|  | Data Preparation |  | **Data Conditioning** |
|  |  |  |  |
|  | Discovery |  | **Initial Hypotheses** |
|  |  |  |  |
|  | Communicate Results |  | **Area Under the Curve** |
|  |  |  |  |
|  | Model Building |  | **Back-Testing** |
|  |  |  |  |
|  | Operationalize |  | **Pilot Project** |
|  |  |  |  |
|  | Data Preparation |  | **Dataset Inventory** |
|  |  |  |  |

|  |
| --- |
| **Problem Set 2: Bayes Theorem** |
| [10 marks] Tom consulted doctor for fever and cough. A nasopharyngeal swab was performed on him for RT-PCR testing, which is considered the gold standard for the diagnosis of Covid-19. The RT-PCR test has 95% sensitivity and 85% specificity.  Suppose that 10% of the population has the disease. Tom’s test came back positive. **What is the probability that Tom actually has Covid-19?** Show all calculation steps. |
| P(A) = Tom has COVID-19  P(B) = Tom is positive  Population = 0.1  Sensitivity = 0.95  Specificity = 0.85  P(A|B) = ( P(B|A) x P(A) ) / P(B)  = ( P(B|A) x P(A) ) / ( P(B|A) x P(A) + P(B|┐A) x P(┐A) )  = ( 0.95 x 0.1) / ( 0.95 x 0.1 + 0.15 x 0.9 )  = 0.095 / 0.23  = 0.413  = **41%** chance Tom has COVID-19 |
|  |

|  |  |  |
| --- | --- | --- |
| **Problem Set 3: Association Rules Mining** | | |
| Consider a retail store database, D, consisting of 9 transactions. As shown in the table below, there are five distinct household items and each transaction involves two to four household items. | | |
|  |  | |
| **Transaction ID** | | **List of Household Items** |
| 1 | | detergent, cleanser, towel, brush |
| 2 | | cleanser, towel |
| 3 | | cleanser, towel |
| 4 | | detergent, towel |
| 5 | | detergent, cleanser, soap |
| 6 | | detergent, towel |
| 7 | | cleanser, soap |
| 8 | | detergent, cleanser, towel |
| 9 | | detergent, cleanser, brush |
|  |  | |
| (a) | [7 marks] Suppose **minimum support** count required is **2**, apply the Apriori algorithm to generate all the frequent candidate itemsets and frequent itemsets . Show all your work. | |
| Itemset = {detergent, cleanser, towel, brush, soap}   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | trans\_id | detergent | cleanser | towel | brush | soap | | 1 | 1 | 1 | 1 | 1 | 0 | | 2 | 0 | 1 | 1 | 0 | 0 | | 3 | 0 | 1 | 1 | 0 | 0 | | 4 | 1 | 0 | 1 | 0 | 0 | | 5 | 1 | 1 | 0 | 0 | 1 | | 6 | 1 | 0 | 1 | 0 | 0 | | 7 | 0 | 1 | 0 | 0 | 1 | | 8 | 1 | 1 | 1 | 0 | 0 | | 9 | 1 | 1 | 0 | 1 | 0 |   Frequent candidate itemset C1   |  |  | | --- | --- | | Itemset | Support\_count | | {detergent} | 6 | | {cleanser} | 7 | | {towel} | 6 | | {bursh} | 2 | | {soap} | 2 |   applying minimum support count = 2  since, all the product in the Itemset were purchased atleast 2 times  Frequqent Itemset, **L1 = C1**  Frequent candidate itemset C2   |  |  | | --- | --- | | Itemset | Support\_Count | | {detergent,cleanser} | 4 | | {detergent,towel} | 4 | | {detergent,brush} | 2 | | {detergent,soap} | 1 | | {cleanser,towel} | 4 | | {cleanser,brush} | 2 | | {cleanser,soap} | 2 | | {towel,bursh} | 1 | | {towel,soap} | 0 | | {brush,soap} | 0 |   Frequency Itemset L2 with min supp = 2   |  |  | | --- | --- | | Itemset | Support\_Count | | {detergent,cleanser} | 4 | | {detergent,towel} | 4 | | {detergent,brush} | 2 | | {cleanser,towel} | 4 | | {cleanser,brush} | 2 | | {cleanser,soap} | 2 |   Frequent candidate itemset C3   |  |  | | --- | --- | | Itemset | Support\_Count | | {detergent,cleanser,towel} | 2 | | {detergent,cleanser,bursh} | 2 | | {detergent,cleanser,soap} | 1 | | {detergent,towel,bursh} | 1 | | {detergent,towel,soap} | 0 | | {detergent,brush,soap} | 0 | | {cleanser,towel,brush} | 1 | | {cleanser,towel,soap} | 0 | | {cleanser,brush,soap} | 0 |   Frequency Itemset L3 with min supp = 2   |  |  | | --- | --- | | Itemset | Support\_Count | | {detergent,cleanser,towel} | 2 | | {detergent,cleanser,bursh} | 2 | | | |
|  |  | |
| (b) | [3 marks] Let **minimum confidence** threshold is **80%**, find all strong association rules of the form {Item X, Item Y} 🡪 {item Z}. Show all your work. | |
| Confidence = (X -> Y) = supp(XUY)/supp(X)  here,  X = {item X, item Y}  Y = {item Z}   |  |  | | --- | --- | | Itemset | Frequency Count | | {detergent,cleanser} | 4 | | {detergent,towel} | 4 | | {detergent,brush} | 2 | | {detergent,soap} | 1 | | {cleanser,towel} | 4 | | {cleanser,bursh} | 2 | | {cleanser,soap} | 2 | | {towel,brush} | 1 |   minimum threshold = 80%  R1: detergent,cleanser -> towel  confidence = (detergent,cleanser,towel) / supp.count(detergent,cleanser)  = 2/4  = 50%  therefore, R1 is rejected  using these steps to find the confidence of items that have a strong association are.  (detergent,bursh) -> cleanser = **100%**  (detergent,soap) -> cleanser = **100%**  (cleanser,brush) -> detergent = **100%**  (towel,brush) -> detergent = **100%**  (towel,brush) -> cleanser = **100%** | | |

|  |  |
| --- | --- |
| **Problem Set 4: k-Means Clustering** | |
| Observe the following data points plotted on a Cartesian plane:    The horizontal axis is the -axis and the vertical axis is the -axis. Now assume that the-means clustering algorithm is to be executed on these data points, where the circled data points are the randomly chosen *initial* centroids. | |
|  | |
| (a) | [0.5 mark] What is the value of of the current clustering configuration? |
| 2 | |
|  | |
| (b) | [0.5 mark] What centroid initialization method was used? |
| Forgy | |
|  | |
| (c) | [6 marks] Let **x1** be the initial centroid for Cluster A and **x3** be the initial centroid for Cluster B. In the 1st iteration of the -means algorithm, to which cluster will **x5** and **x6** be assigned to? Show all the necessary calculations to support your answer. |
| |  |  |  |  | | --- | --- | --- | --- | |  | Cluster A **(1,4)** | Cluster B **(2,6)** | Centroid | | X1 (1,4) | |1-1| + |4-4| = 0 | |1-2| + |4-6| = 3 | A | | x2 (1,6) | |1-1| + |6-4| = 2 | |1-2| + |6-6| = 1 | B | | X3 (2,6) | |2-1| + |6-4| = 3 | |2-2| + |6-6| = 0 | B | | x4 (3,8) | |3-1| + |8-4| = 6 | |3-2| + |8-6| = 3 | B | | x5 (4,3) | |4-1| + |3-4| = 4 | |4-2| + |3-6| = 5 | A | | x6 (5,2) | |5-1| + |2-4| = 6 | |5-2| + |2-6| = 7 | A |   as we can see by calculating the distance between the point and the centroid using Manhattan distance, it is apparent, x5 and x6 belongs to **Cluster A** | |
|  | |
| (d) | [3 marks] Calculate the positions of the new centroids after the 1st iteration of the -means algorithm. Show your calculation steps |
| Initial centroid:  Cluster A (1,4)  Cluster B (2,6)  X1 (1,4) = A x3 (2,6) = B  x5 (4,3) = A x2 (1,6) = B  x6 (5,2) = A x4 (3,8) = B  New centroid:  Coordinates = ((sum of x coordinates)/(count of x) , (sum of y coordinates)/(count of y))  Cluster A = ((1 + 4 + 5)/(3) , (4 + 3 + 2)/(3)) = **(3.3,3)**  Cluster B = ((2 + 1 + 3)/(3) , (6 + 6 + 8)/(3)) = **(2,6.6)**  Therefore, the position of the new centroids after the 1st iteration are **(3.3,3)** for Cluster A and **(2,6.6)** for Cluster B | |

|  |  |
| --- | --- |
| **Problem Set 5: Linear Regression** | |
| (a) | [2 marks] The histogram below shows the frequency distribution of the outcome variable *Income*. The values of *Income* tend to be highly skewed to the left (distribution of value has a range tail to the left). Does such non-normallly distributed outcome variable violate the general assumption of a linear regression model? Provide supporting arguments. |
| Linear regression assumes a linear relationship between the dependent and independent variable, therefore, the histogram here violates the principle of linear relationship. | |
|  |  |
| (b) | [2 marks] Suggest a graphical method to test for violation of the normality assumption. |
| Histogram and normality plot | |
|  |  |
| (c) | [2 marks] Recommend a data transformation method to normalise the distribution of the *Income* variable. |
| Rescaling | |
|  |  |
| (d) | [4 marks] Using the ROC curve below, discuss how it can be used to determine an appropriate threshold value for the classification of *Income*.  ROC curve analysis with MedCalc |
| From the ROC curve, it is evident that, if our specificity is 100% then our sensitivity is also 100% but having a false positive rate of 100% will result in incorrectly classifying all the false positives. Therefore, a good balance of true positive rate and false positive rate would be around (15, 90) meaning, 90% correct classification and 15% incorrect classification. | |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Problem Set 6: Naïve Bayes Classifier** | | | | | |
| [10 marks] Suppose you have the following training set of positive (+) and negative (-) online products. There are only 5 training instances and 3 features. | | | | | |
| **Great** | | **Fine** | **Terrible** | **Class** |  |
| 1 | | 0 | 0 | + |  |
| 0 | | 1 | 1 | - |  |
| 0 | | 1 | 1 | - |  |
| 0 | | 0 | 0 | + |  |
| 1 | | 0 | 1 | - |  |
| Assume that the Laplace smoothing constant = 1 is used when building a **Naïve Bayes** model from the training set above. What would be the ***class*** label (+ or -) assigned to the product reviews below:  Review 1: “The product is ***great*** and works ***fine***.”  Review 2: “***Great*** product. Packaging was ***terrible***.”  Show your complete work, including the probability score calculations. Round the probability scores to two decimal places. | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | Great | Fine | Terrible | | Positive | 1 | 0 | 0 | | Negative | 1 | 2 | 3 |   After laplace smoothing (+1)   |  |  |  |  | | --- | --- | --- | --- | |  | Great | Fine | Terrible | | Positive | 2 | 1 | 1 | | Negative | 2 | 3 | 4 |   **Positive Reviews:**  P(Great|+) = 2/4 = 0.5  P(Fine|+) = 1/4 = 0.25  P(Terrible|+) = 1/4 = 0.25  **Negative Reviews:**  P(Great|-) = 2/9 = 0.22  P(Fine|-) = 3/9 = 0.33  P(Terrible|-) = 4/9 = 0.44  P(+) = 2/5 = 0.4  P(-) = 3/5 = 0.6  **Review1:**  Postive = P(+) x P(Great|+) x P(Fine|+)  = 0.4 x 0.5 x 0.25  = 5%  Negative = P(-) x P(Great|-) x P(Fine|-)  = 0.6 x 0.22 x 0.33  = 4.3%  The % of review 1 being postive is greather. Thus, it is a **Positive Review**  **Review2:**  Postive = P(+) x P(Great|+) x P(Terrible|+)  = 0.4 x 0.5 x 0.25  = 5%  Negative = P(-) x P(Great|-) x P(Terrible|-)  = 0.6 x 0.22 x 0.44  = **5.8%**  The % of review 2 being negative is greather. Thus, it is a **Negative Review** | | | | | |
|  |  | | | | |

|  |  |
| --- | --- |
| **Problem Set 7: Tree-Based Classifier** | |
| [10 marks] You are given a dataset with 4 attributes in the table below. The first row contains attribute names. Each row after the first represents the values for one data instance. The output attribute is ***Credit***. Construct a decision tree with root node ***Age*** using the information gain measure for attribute selection. Show all calculation steps.   |  |  |  |  | | --- | --- | --- | --- | | **Income** | **Age** | **Marriage** | **Credit** | | High | ≥ 30 | Married | Class 1 | | Low | ≥ 30 | Married | Class 1 | | Low | < 30 | Married | Class 2 | | Low | < 30 | Single | Class 2 | | Low | ≥ 30 | Single | Class 3 | | High | ≥ 30 | Single | Class 3 | | High | < 30 | Married | Class 3 | | |
|  | |
| Diagram  Description automatically generated | |
|  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Problem Set 8: Model Evaluation** | | | | | | | | | |
| Suppose for 500 emails, a classifier predicts 100 to be **Spam** and 400 to be **Genuine**. Of the 400 genuine predictions, 330 are truly **Genuine**. Of the 100 spam predictions, 70 are truly **Spam**. Let “spam” be the positive class and “others” be the negative class. | | | | | | | | | |
| (a) | [3 marks] Fill in the confusion matrix below indicating the number of true positive, false positive, false negative, and true negative classifications. | | | | | | | | |
|  |  | **Predicted Class** | | | | | **Total** |
| **Spam** | | | | **Genuine** |
| **Actual Class** | **Spam** | 70 TP | | | | 70 FN | 140 |
| **Genuine** | 30 FP | | | | 330 TN | 360 |
|  | **Total** | 100 | | | | 400 | **500** |
| (b) | [4 marks] Calculate the following evaluation metric score based on the confusion matrix above. Show the formula of each metric and round your answers to two decimal places. | | | | | | | | |
|  | True Positive Rate (Recall) = | | | | | False Positive Rate (Type I Error Rate) = | | | |
|  | 70 / ( 70 + 70 ) = **0.5** | | | |  | | 30 / ( 30 + 330 ) = **0.08** | | |
|  |  | | | | |  | | | |
|  | False Negative Rate (Type II Error Rate) = | | | | | Precision = | | | |
|  | 70 / ( 70 + 70 ) = **0.5** | | | |  | | 70 / ( 70 + 30 ) = **0.7** | | |
|  |  | | | | | | | | |
| (c) | [3 marks] Suppose we have two users (Tom and Jerry) with the following preferences.  *Tom hates seeing spam mails in his Inbox. However, he doesn’t mind periodically checking the ‘Junk’ directory for genuine emails incorrectly marked as spam.*  *Jerry doesn’t even know where the ‘Junk’ directory is. He would much prefer to see spam emails in his inbox than to miss genuine emails without knowing.*  Which user is more likely to satisfied with this classifier? Explain you answer using the evaluation metric scores from Question 8 (b). | | | | | | | | |
|  | | Tom hates seeing any spam message, according to our classifier out of 100 emails 30 were labelled incorrectly and was sent to inbox, therefore this classifier is not optimum for Tom. On the other hand jerry does not mind having spam email infeltrating his primary mail, therefore the classifier is more more suitable for Jerry. | | | | | | | |

|  |  |  |
| --- | --- | --- |
| **Problem Set 9: Model Selection** | | |
| [10 marks] Determine which analytic model (classification, *regression*, or *clustering*) should be the first choice for answering the following business questions. Briefly justify your answers. | | |
| Business Question (a): | | “Will this customer purchase product P1 if given incentive X?” |
| Using clustering we can group customers based on their behavior or purchase patterns. This will  take out the guess work and provide us a clearer direction about if certain customer are more likely  to buy a thing given certain incentive. | | |
|  | | |
| Business Question (b): | | “Which service package (**S1**, **S2**, or none) will a customer be likely to purchase if given incentive **X**?” |
| Classification (more precisely Decision tree classification) can be used to answer such questions. | | |
|  | | |
| Business Question (c): | | “How much will this customer utilise the service?” |
| This question can be answered with regression. Because regression can provide us an actual  value whereas, classification classifies only true/false value and clustering creates similar  groups. | | |
|  | | |
| Business Question (d): | | “What products or services should we develop or customise?” |
| We can use classification to determine which product or services should be developed or customised. | | |
|  | | |
| Business Question (e): | | “Who are the target buyers of the product?” |
| Clustering can create cluster of people who might be interested in simialr products. Therefore, clustering can be used to solve the business problem. | | |
|  |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Problem Set 10: Data Analytic Plan** | | | | |
| [10 marks] **Joey Bags O’Donuts Wholesale Wine Emporium** is an import-export business focused on bringing bulk wine to the United States (New Jersey) and selling it to select wine and liquor stores across the country. The way the business works is that Joey Bags travels the globe finding incredible offers on large quantities of wine. Joey then ships the wine back to New Jersey. Like most businesses, his business objective is to sell this stuff to various stores at a profit.  The company obtains a lot of business via its email newsletters. Usually, there are two or three wine deals in each email, for example one for Champagne, and another on Malbec. Some deals are amazing, 80% or more off of retail price. In total, the company has offered 32 campaign offersor deals last year in 2017, all of which have gone quite well.  Not satisfied with the current sales performance, the company wants to do even better.This requires understanding its customers’ interests a little bit more. To achieve that goal, Joey was advised to employ *K*-Means clustering to find the best customer segment and to understand why it is the best segment to focus on.  Figure 1 shows data about Joey’s wine campaign offers this year. Figure 2 provides a partial snapshot of 324 customer transactions in response to these offers (assume every customer name is unique).  Figure 3 to Figure 6 depict the clustering outcomes. It was noted that the average Silhouette score of the resulting clusters was 0.1696. | | | | |
| Given the information above, complete the analytic plan below. | | | | |
|  |  | | | |
| **Analytic Plan** | | | | |
| **Components of Analytic Plan** | | **Joey Bag O’Donuts Wholesale Wine Emporium** | | |
| **Framing a business problem into an analytic problem**  [2 marks] | | Business problem: | | |
| Joey Bag O’Donuts have seen decline in their sale in current years, they want to improve their sales and want to advertise their product more effectively. | | |
| Analytic problem: | | |
| *K*-Means clustering can be used to find the best performing campaign among certain groups of people. It can also help us understand the consumer segment better. | | |
| **Initial Hypothesis**  [1 mark] | | Initial Hypothesis (only 1): | | |
| Even though champagne had more consistent sale (at least 72kg) than Cabernet Sauvignon, Cabernet Sauvignon had more past peeks (4 times out of 6) and the discount been relatively lower than champagne (50.8% vs 69.8%). Therefore, based on past peek performance and discount rate, Cabernet Sauvignon should be advertised more. | | |
| **Data and Scope**  [4 marks] | | Campaign offers data | | |
| Dimension | : | 7 |
| Size | : | 32 |
| Time scope | : | January to December 2017 |
| Name of identifier var. | : | Offer # |
| Transaction data | | |
| Dimension | : | 2 |
| Size | : | 324 |
| Time scope | : | January to December 2017 |
| Name of identifier var. | : | Offer # |
| Possible data transformation: | | |
| Data Aggregation | | |
| **Model Planning** | | K-Means Clustering (***k*** = 4) | | |
| **Result and Key Findings**  [3 marks] | | Cluster profiles | | |
| Cluster 1 has a greater number of sales than any other cluster. | | |
| **Business Impact** | | Develop future email campaigns tailored to specific groups of customers with noticeable preferences for certain types of wine product offerings. These highly targeted marketing campaigns are expected to boost the company’s overall sales performance. | | |
|  |  | | | |
|  |  | | | |

**--- End of Assignment ---**