# **GOLDSMITHS, UNIVERSITY OF LONDON**

**Department of Computing**

# **B. Sc. Examination 2017-2018**

**IS53023B Data Mining – Resit**

**Duration: 2 hours 15 minutes**

**Date and time:**

This paper is in two parts: part A and part B.  You should answer ALL questions from part A and TWO questions from part B.  Part A carries 40 marks, and each question from part B carries 30 marks.  The marks for each part of a question are indicated at the end of the part in [.] brackets.

Electronic calculators must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

**THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM**

**PART A**

**Question 1**

In a customer attrition / churning application which involves a dataset with details about customers, the output attribute called **churn** has two values, **yes** and **no**. One builds a predictive model that, before being used in practice, is evaluated on a test dataset. Below one provides two columns from the scored dataset, namely the column **churn**, and the column **predicted\_churn** which provides the predictions by applying the model on this test dataset.

You are required to:

1. Explain what the confusion matrix is and compute it in this case. [8]
2. Calculate the accuracy, error, precision, sensitivity, specificity, and lift with respect to the class **yes** by showing your work. [12]
3. Briefly explain what the accuracy, error, precision, sensitivity, specificity and lift represent in the context of this mentioned application of predicting churners. [6]
4. What is the most important performance measure among the ones mentioned in (b) above, as a criterion for the selection of a predictive model? Justify your answer. [5]
5. List the names of 6 algorithms that can be used to predict churn, and the names of 3 algorithms that cannot predict churn. [9]

****

 **page 1**

**PART B**

**Question 2**

a) Explain why k-Means cannot detect outliers. [5]

b) Explain why k-Means cannot work with missing data. [5]

c) Apply k-Means algorithm using the following dataset for k=2. The cluster centres to start with are the rows (40, 50) and (60, 50). Only the first two iterations are required to be illustrated. Would more iterations be needed for the completion of the algorithm? Justify your answer. [20]

 **page 2**

**Question 3**

a) Briefly explain the conditional independence assumption that the Naive Bayes algorithms makes regarding the data. [4]

b) Apply the Naive Bayes algorithm on the dataset below in order to predict the value for LifeInsurance attribute for an instance defined by HomeOwner=No, Retired=Yes, Car=No, and CreditCardInsurance=No. You are required to illustrate how the algorithm works in this case. [26]



 **page 3**

**Question 4**

The following dataset having the attributes **lifestyle**, **family status,** **car** and **sports** is to be used for the extraction of association rules. You are required to apply the Apriori algorithm with the minimum support of 0.15 and the minimum confidence 0.6. In particular:

a) Calculate the minimum count. [2]

b) Generate all the frequent itemsets. [12]

c) Generate all the association rules formed of two items and decide which ones are strong rules. Finally you are required to compute also the support for the strong rules only. [16]



  **END OF EXAMINATION**