While big data is the “talk of the town”, the reality is that most of us deal with small to medium data sets, or data sets that could be crunched on a single computer rather than a “big data” set up. The more important skill is therefore the ability to apply analytics to data sets, which is what we shall focus on in this assessment.

**Background**

The Vodafone PR nightmare in 2010 continue to remain a vivid telecommunication event in Australia. One of the outcomes was one person’s creation of a Website call “Vodafail”, where many Vodafone customers vented their frustration about the service provider. Years later, this Website is now a great data source for Vodafone and its competitors to learn about issues that concerns customers most.

Earlier, a competitor has scraped the Vodafail.com Website for comments left by past angry Vodafone customers. Your company has been hired to perform analytics for the competitor – Optus. Their analytical brief are as follows.

 An intern helped scraped the data off the Vodafail.com website and is saved as a Microsoft Access database with two tables.

 Optus reckons that they may one day face similar issues and would like to build an “early detection” system where they would monitor comments about Optus in social media. The early detection system would analyse each comment (i.e., a string of text) and then decide what the issue was. If the detected issue starts to appear very often, then Optus will be alerted so that they can look at addressing the problem before a similar PR nightmare occurs.

 After listening to the brief, you realise that the core engine of the said early detection system is a classifier that could predict the topic of a given string of text. These topics would then be counted so that if there are more than X of them are detected within a given timeframe, an alert could be raised to inform Optus to investigate further.

 To explain this, you sketched the following system’s view of the “early detection” system to show to your customer. He is happy and asked that you proceed.



**Task 1**

Download “vodafail.mdb” (this will be provided to you) and start exploring the data. As mentioned before, it is important that you get acquainted with the data so that you know what the facts mean (i.e., giving it appropriate context).

 What did you observe about the data? Write down at least two and no more than five observations. If you have more than five, give the five that you think are the most important for the analytics to be performed later.

**Task 2**

Before your proposed analytics could be applied, you realise that the ‘target class’ (i.e., the “Issue Category” field) is in the table call “Category” and the “Comments” from customers are in another table call “Vodafail”. Therefore, your first task is to merge the two tables into a single flat table (i.e., data transformation) so that you can load them into your analytical tool.

 Write down the SQL statement that you used to merge the two tables. If you do not know how to run a SQL query in Microsoft Access, check out the following tutorial: http://goo.gl/GqieUx

 Export the output of your SQL query as an Excel spreadsheet and upload this as part of your submission for this assignment.

o You can export the SQL query output by copying from Access and pasting the data into Excel; or you can click on the “External Data” ribbon and use the “Export to Excel” functionality.

**Task 3**

After completing Task 2, you also realise that this file looks like the “language\_id” problem in one of the workshops you completed. Therefore, you decided that you will need to split the comments into one word per column.

 Produce this newly formatted file by reviewing the workshops you completed. You may want to Google how to split a column of text into a word term per column from within Excel.

 Make sure that the file can be loaded using the “File” widget in Orange. If not, describe the steps you undertook to make the file ‘loadable’ into Orange.

 When you have a file that can be loaded by the “File” widget in Orange, upload this file as part of your submission for this assignment.

**Task 4**

Now that the file can be loaded, your final task is to produce a classifier that could predict comments to as high an accuracy as possible when tested using the “test.tab” file.

 You are allow to use any classifier, and/or any combination of the classifiers in Orange.

 You can also tune the parameters of each widget to achieve the best prediction accuracy possible.

o For the purpose of this assessment, accuracy is determined by the ratio of the number of test cases that were correctly predicted against all test cases applied.

 In no more than half a page, present the key aspects of your classification system. Details such as your choice of classifier, the parameters setting, your tuning process, your evaluation of the classifiers you produced, etc., can be discussed where appropriate. The aim is to provide a brief to your marker to help him or her understand your Orange Scheme file so that the most marks could be awarded whenever possible.

**Hints**

There are many ways to improve accuracy. If you recall, the key principle with analytics is “garbage in, garbage out”. Hence, the biggest improvement can often be made by pre-processing the data file effectively before using Orange.

 English as a language contains wide variation due to vocabulary, spelling and grammar. A word like ‘purchase’ is likely to have the same meaning as ‘buy’, ‘purchasing’ or ‘purchased’ but are different to a classifier. There are English language processing tools to help ‘reduce’ word terms with the same meaning down to a common form, e.g., stemming.

 Some comments in the file are very long while some are very short. Consider whether you need to include all word terms (i.e., all the columns) for each training example (i.e., a row in the spreadsheet you exported in Task 3).

 Generally, spending more time preparing the data is more important than tuning parameters of the widgets. Therefore, look at the spreadsheet and see what you can do with it. Recall that the analytical process is governed by the KDD framework. This framework, if you recall, requires the analyst to reiterate through each phase and between phases to achieve (i) understanding of the data; (ii) the objective of the analytics; and (iii) the best tools to apply to achieve the outcomes desired.

**Marking rubrics**

There are four tasks in this assignment. A grade will be awarded to each task and then an overall mark determined for the entire assessment task. The rubric below gives you an idea of what you must achieve to earn a certain ‘grade’.

As a general rule, to meet a ‘credit’, you must first satisfy the requirements of a ‘pass’. And for a High distinction, you must first satisfy the requirements of a ‘distinction’, which must of course first meet the requirements of a ‘credit’, and so on.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Your final grade will be decided based on the grades for each component in the assessment. It is not a simple average but rather, a final consideration of how you performed as a whole informed by the grades of each component. **Pass** | | **Credit** | | **Distinction** | | **High Distinction** | |
| **Task 1:** Observation about data | Two acceptable observation of data is presented. | | Three acceptable observation of data is presented. | | Four acceptable observation of data is presented. | | All five acceptable observation of data is presented. |
| **Task 2:** SQL and data transformation | SQL statement and spreadsheet produced. | | SQL statement shows correct join conditions and spreadsheet reflects the SQL statement. | | N.A. | | In addition to meeting the requirements of a ‘credit’, spreadsheet accurately reflects the data input expected. |
| **Task 3: Input file preprocessing** | File could be loaded into Orange. | | File shows word terms processed correctly into their respective columns. | | File shows adequate and essential preprocessing to load data for analytics. | | File shows relevant and adequate preprocessing to improve analytical outcomes. |
| **Task 4: Analytics** | Orange Scheme file loadable by marker and meets basic client’s brief. | | Orange Scheme file loadable by marker and achieves a majority accuracy (i.e., > 50% correct) under the marker’s test file. | | In addition to the ‘credit’ requirements, Orange Scheme file achieves an accuracy of > 75% under marker’s test file. | | In addition to the ‘distinction’ requirements, Orange Scheme file achieve an accuracy of above 80% under the marker’s test file and that, the Orange Scheme set up is judged to be more advanced than |