BUSS6002 Assignment 2

Due Date: Friday 31 May 2019

Value: 25% of the total mark

Rationale

This group assignment has been designed to help students develop valuable collaboration and communication skills when working in a team, as well as to allow students to contextualise their data science skills on a real-world problem from business.

Instructions

1. **Required submission items via Canvas:**
    1. **ONE** written report (PDF format).
        - Assignments > Report Submission (Assignment 2)
    2. **ONE** Jupyter Notebook .ipynb
        - Assignments > Upload Your Code File (Assignment 2)
    3. **ONE** csv file of test results
        - Assignments > Test Results Submission (Assignment 2)
2. The assignment is due at **12:00pm (noon) on Friday, 31 May 2019**. The late penalty for the assignment is 5% of the assigned mark per day, starting after 12:00pm on the due date. The closing date Friday, 07 June 2019, 12:00pm (noon) is the last date on which an assessment will be accepted for marking.
3. As per anonymous marking policy, please include the Group ID and Student IDs of all group members. Do **NOT include names**. The name of the report and code file must follow: **GroupID_BUSS6002_Assignment2_S12019,** and the name of test results must follow: **GroupID_BUSS6002_Assignment2_Test_Results.csv**.
4. Your answers shall be provided as a word-processed report giving full explanation and interpretation of any results you obtain. Output without explanation will receive **zero** marks. You are required to also submit your code that can reproduce your reported results, as reproducibility is a key component to data science. Not submitting your code will lead to a loss of 50% of the assignment mark.
5. Be warned that plagiarism between individuals is always obvious to the markers of the assignment and can be easily detected by Turnitin.
6. Presentation of the assignment is part of the assignment. There will be **10%** marks for the presentation of your report and code submission.
7. Numbers with decimals should be reported to the **third-decimal point**.

Meeting Minutes

1. Each group is required to submit at least 3 meeting minutes as the appendix attached to the final report. Templates will be provided for preparing meeting agendas and meetings minutes. You may use the templates provided or a template you choose.
2. In case of a problem within a group we will request minutes of the previous meetings. We can make an individual adjustment to the group mark if there is sufficient

evidence that a student has done very little. If the student has done nothing, we will award a mark of zero.

Peer Review

1. If you encounter any issues with your group members, please report and discuss with your unit coordinator as early as possible.
2. We may ask for peer review from each student within a group. The instructions about how to do this will be released later on.
3. Each group will be awarded a group mark per the marking criteria. In special cases, individual marks may be applied if there is dispute in a group and the quality/quantity of contributions made by individuals are significantly different, in which cases the unit coordinator will seek peer review reports from individuals within a group to decide on individual marks.

Group Competition

We will allocate **10%** marks for competition among groups. The winning group with the highest test score will secure a full 10% mark, while other groups will be awarded a mark according to their test scores relative to the best test score.

Project Description and Dataset

Today, e-commerce has revolutionized the way companies do business and how consumers make purchasing decisions. It has become common practice for consumers to read online reviews to inform their decision making. The increasing popularity of online reviews also stimulates the business of "illegal" activities (e.g., fake review writing) that try to mislead readers by producing deceptive reviews to influence readers' opinions and decision making. Typical examples include posting of undeserving positive reviews in order to promote a product and/or by posting of fake negative reviews for competitors in order to damage their reputation.

In this assignment, you are tasked with developing a data science technique that can automatically detect fake reviews from a collection of reviews submitted to a leading online shopping portal. You must summarise your findings in a report according to the task instructions given below.

You are provided with two data files: `review_train.csv` and `review_test.csv`. The files contain a collection of reviews submitted for a range of products and customers' purchase information. Only `review_train.csv` contains `Label` values (target), where 1 indicates "fake" and 0 indicates "normal". You should train and validate your model building using `review_train.csv` and report the test score on `review_test.csv`. It is your goal to identify fake reviews based on the text of reviews and other features available in the data sets. The meanings of features are obvious according to their variable names in the data sets. However, it may not be feasible to directly use these features (especially reviews represented as raw text) to build a good classification model. One of your tasks is to carefully extract or construct meaningful features as input for the modelling process.

Tasks

Please note that most tasks are deliberately designed to be open ended as this is a real-world problem.

Exploratory Data Analysis (EDA): You should conduct a thorough EDA for the given data sets. For example, check/deal with missing data, visualise the distributions of features, identify which features can better distinguish different target values, and perform feature correlation analysis, etc. Carefully present your analysis and findings in your report.

Feature Engineering: You may consider feature engineering strategies to extract useful features for your modeling task. Given that reviews are represented as raw text, you may consider pre-processing raw text and transforming them into text features, e.g., unigrams, bigrams, or TF-IDF. Carefully think about what features can be 辨别 discriminative for fake/normal reviews, for example, length of review, percentage of 大写词 capitalized words, etc. Justify your choice of feature engineering strategies and report your findings.

Benchmark Model: Build a logistic regression model to assess the feasibility of the project and establish a baseline model. You may choose to split the given training data into a training/validation set or use cross-validation to validate your model building. You need to validate your model building against whatever hyperparameters apply to the model. For this task, you can build your baseline model using simplistic text-based features. Document your analysis and findings.

Improving your Model: You are requested to improve the performance of your benchmark model as much as you can. You may consider the feature selection strategies, i.e., choosing the top K most important features to re-build your model, or when reasonable, adding new constructed features to re-build your model. Report your setting and comparisons with the benchmark model.

Note: For this task, if you want to use a classification model that is not taught in this unit, you must clearly explain the principle of the model, justify why you choose that model, and present your analysis. For any model you choose, you need to validate your model building against whatever hyperparameters apply.

Final Test Results: Finally, according to your analysis, decide your best model and apply the model on the test data. You are asked to report the classification results on the test data. Save your results into a csv file containing two columns, one for the Review indices (REVIEW_ID from `review_test.csv`) and the other column for the predicted labels (0's or 1's). Name your file as **GroupID_BUSS6002_Assignment2_Test_Results.csv**. The results will be assessed by our markers in order to decide your group performance among the entire class (group competition!).

Note that, we will use **F1-score** as the test score for group competition.

Presentation

- The assignment material to be submitted will consist of a final report that:

    1) Takes a research article form in which you shall have a number of sections such as introduction, methodology, experiment results, findings/interpretation, and conclusion. All references should be properly cited and take a full bibliographical format. Here are a few examples
    http://cs229.stanford.edu/proj2015/007_report.pdf
    http://cs229.stanford.edu/proj2015/188_report.pdf
    http://cs229.stanford.edu/proj2015/031_report.pdf

    2) Details ALL steps and decisions taken by the group regarding requirements above.

    3) Demonstrates an understanding of the problem being addressed and the relevant principles of data science techniques used.

    4) Includes an executive summary to conclude your best performing model and the recommended features for modeling.

    5) Clearly and appropriately presents any relevant graphs and tables.

- The report should be **NOT more than 20 pages** with font size no smaller than 11pt, including everything like text, figures, tables, small sections of inserted code, etc., but excluding the cover page and the appendix containing the meeting minutes. Think about the best and most structured way to present your work, summarise the procedures implemented, support your results/findings and prove the originality of your work. You will provide your code as a separate submission to the report; however, you may insert small sections of your code into the report when necessary.
- Your code submission has no length limit, however marks are assigned for code presentation, so make your code as concise as possible and add comments when necessary to explain the functionality of your code segments. Make sure to remove any unnecessary code and ensure that your code can be run without error.
- Your group is required to submit at least 3 meetings minutes. Your group may use the attached templates for preparing meeting agendas and minutes. Documentation should include attendance, discussion points, actions decided, etc. You may use your own form or find something online.
- You, as a member of a group, may be also required to submit your peer review. Please use the provided criteria sheet and assessment form for this purpose. You will be advised how to use an online form when it becomes available.