THE UNIVERSITY OF SYDNEY

Data Analytics Project

2ND SEM 2017

INFO3406

LECTURER: CLAUDIO DIAZ. TUTORS: AUGUSTO DIAS AND HARRISON TRI

Contents

Learning Outcomes

The objective of the project is to put in practice all the theory learned in class with a tool and methodology that help students in their future as professionals.

Specifically the learning outcomes are,

- Process large data sets using appropriate technologies
- Select statistical techniques appropriate for summarization and analysis of a data set, and can justify their choice
- Select statistical techniques appropriate for evaluation of a predictive model that is based on data analysis, and can justify their choice
- Find out details of how to use a method or tool in the data analytic process.
- Carry out (in guided stages) the whole design and implementation cycle for creating a pipeline to analyse a large heterogeneous dataset.
- Apply concepts and terms from social science to describe and analyse the role of a data analysis task in its organizational context
- Communicate the results produced by an analysis pipeline, in oral and written form, including meaningful diagrams
- Communicate the process used to analyse a large data set, and justify the methods used.

The Project

In this project the students have to follow the CRISP-DM methodology to achieve a specific goal, using data analytics contents and tools learned in this course.

The project is divided in four stages.

- Project Stage 1: Obtain data, clean it and load it.
- Project Stage 2: Summarize and analyse the data.
- Project Stage 3: Develop and test a predictive model.
- Project Stage 4: presentation of results.

Each stage in detail,

Project Stage 1: Obtain data, clean it and load it.

- Business Understanding
    - Determine Business Objectives
        - Background
        - Business Objectives
        - Business Success Criteria
    - Assess Situation
        - Inventory of Resources
        - Requirements, Assumptions, and Constraints
        - Risks and Contingencies
        - Terminology
        - Costs and Benefits
    - Determine Data Mining Goals
        - Data Mining Goals
        - Data Mining Success Criteria
    - Produce Project Plan
        - Project Plan
        - Initial Assessment of Tools and Technique
- Data Understanding
    - Collect Initial Data
        - Initial Data Collection Report
    - Describe Data
        - Data Description Report

Project Stage 2: Summarize and analyse the data.

- Data Understanding
    - Explore Data
        - Data Exploration Report
    - Verify Data Quality
        - Data Quality Report
- Data Preparation
    - Select Data
        - Rationale for Inclusion/ Exclusion

    - Clean Data
        - Data Cleaning Report
    - Construct Data
        - Derived Attributes
        - Generated Records
    - Integrate Data
        - Merged Data
    - Format Data
        - Reformatted Data
    - Dataset
        - Dataset Description

Project Stage 3: Develop and test a predictive model.

- Modeling
    - Select Modeling Techniques
        - Modeling Technique
        - Modeling Assumptions
    - Generate Test Design
        - Test Design
    - Build Model
        - Parameter Settings Models
        - Model Descriptions
    - Assess Model
        - Model Assessment
        - Revised Parameter Setting
- Evaluation
    - Evaluate Results
        - Assessment of Data Mining Results w.r.t. Business Success Criteria
        - Approved Models
    - Review Process
        - Review of Process
    - Determine Next Steps
        - List of Possible Actions Decision

Project Stage 4: Presentation of results.

- Deployment
    - Plan Deployment
        - Deployment Plan
    - Plan Monitoring and Maintenance
        - Monitoring and Maintenance Plan
    - Produce Final Report
        - Final Report Final Presentation
    - Review Project
        - Experience Documentation

Assessments

- 10% Project Stage 1: Obtain data, clean it and load it [individual work]
- 10% Project Stage 2: Summarize and analyze the data [individual work]
- 10% Project Stage 3: Develop and test a predictive model [group work]
- 10% Project Stage 4: presentation of results [group work]

For the stages 3 and 4, the groups are pairs of students (two students). Students make the groups. If for some reason a student don't have a group by the end of week 8, please send a mail to the lecturer or tutors.

Submissions

At the week of the assessment, at the end of the class, you have to

- Deliver to the lecturer the report (document) printed
- Upload on BlackBoard the report, raw data, processed data, the model and the output, accordingly.

In specific,

Assignment	**%**	**Due**	**Individual/Group**	**Notes**
Stage 1	10	Week 4 (Thu 24 Aug, 1:00 pm)	Individual	Submission: - The report and raw data, electronically via eLearning (blackboard) - The report physically at the end of the lecture
Stage 2	10	Week 8 (Thu 21 Sept, 1:00 pm)	Individual	Submission: - The report, raw data and processed data, electronically via eLearning (blackboard) - The report physically at the end of the lecture
Stage 3	10	Week 11 (Thu 19 Oct, 1:00 pm)	Group (pairs)	Submission: - The report, raw data, processed data, the model and the output, electronically via eLearning (blackboard) - The report physically at the end of the lecture - Only one of the group make the submission
Stage 4	10 (5 report, 5 video)	Week 13 (Thu 2 Nov, 1:00 pm)	Group (pairs)	Submission: - The report, raw data, processed data, the model and the output, electronically via eLearning (blackboard) - A video explaining your project - The report physically at the end of the lecture - Only one of the group make the submission

Expectations

In this course is expected, in each stage, a submission of a report (document) including each content of each phase, as explained detailed in the chapter “The Project”. Is recommended to use pictures and screenshots to help understanding.

Also, in every stage, must be delivered the corresponding output work of every phase (e.g., in the phase “Modelling” the output is a model). This is detailed in the chapter “Submissions”, column “Notes”.

Specifically for stage 4, additional to the report, you must make a video explaining all you project. This video must be 2 minutes long.

Remember that the methodology make the stages dependant, so every stage of the project is summative. So, in phase 2 you have to deliver phase 1 and phase 2, in phase 3 you have to deliver phase 1, 2 and 3, and in phase 4 you have to deliver phase 1, 2, 3 and 4 plus the video. Be careful and check out that all the ideas in the document are aligned.

Finally, don’t forget to write down in the document the source of every information you use (remember plagiarism)

If you need further explanation of each phase and sub stage, please ask the lecturer and/or tutors. Also there is a link to a book, in BlackBoard, created by IBM, with examples.

Late Submission Policy

- A penalty of minus 20% (one) mark per each day after the due date.

-Maximum delay is 5 (five) days, after that assignments will not be accepted