DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

EXAM ASSIGNMENT

Study Programme and level	**MSc Business Intelligence**						
Term	S16o						
Course name and exam code(s)	Data Mining for Business Decisions						460152E041 / 460142E002
Exam form and duration	Take-home, written (WHAI)						48 hours
Date and time	24 – 26 May						9:00-9:00
Supplementary material/aids	All	X	Specified		No		
Other relevant information	Assignment posted: Tuesday, May 24 at 09:00 on WISEflow Deadline for reports: Thursday, May 26 at 09:00 on WISEflow						
Hand-in of hand-written material allowed	Yes		No	X	Comments:		
Number of pages (incl. front page)	4 (A zip-file name *pumpitup* is also uploaded as appendix)						

Your exam paper must comply with the following format requirements:

- Your **student ID number** must appear on every page.
- Write **page numbers and total number of pages** on all pages of your paper (e.g. 1 of 15, 2 of 15 and so on).
- Your exam paper MUST be handed in as one **PDF** file, but additional material/appendices may be uploaded in other file formats.
- **The file name** must be your student ID number **AND** the name of the exam.

DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

Pump it up:
Predicting the operating status of water pumps in Tanzania

This is a 48-hour take-home exam. You receive a business problem and a data set. Besides, all other materials are allowed. You have 48 hours to formulate a data mining project, process the data, build a prediction model, document your modelling decisions, report the results, and discuss how the final model contributes to the solution of the business problem.

Background

Can you predict which water pumps are faulty? Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, or which ones do not work at all? Predict their operating status based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed. A smart understanding of which water points are likely to fail can help the government improve inspection and maintenance operations and thereby ensure that clean, potable water is available to communities across Tanzania. Additional background information about the Water Point Mapping system in Tanzania is available on the Water Point Mapping Tanzania website.

The case

The goal of this exam is to build a model that predicts, whether a water pump in Tanzania is likely to fail. The model should be optimised in such a way that it enables the Tanzanian government and its partners to prioritise their water pump inspection and maintenance operations and thereby ensure that as many users as possible will have stable access to clean, potable water. The exam case is adapted from an ongoing data mining competition on DrivenData sponsored by Taarifa.

The data

The data set pumpitup.csv (zipped, archive file: pumpitup.zip) contains data from a cross-section of 59400 water pumps in Tanzania. There are 37 raw input variables:

- **amount_tsh** - Total static head (amount water available to water point)
- **date_recorded** - The date the row was entered
- **funder** - Who funded the well
- **gps_height** - Altitude of the well
- **installer** - Organisation that installed the well
- **longitude** - GPS coordinate
- **latitude** - GPS coordinate
- **wpt_name** - Name of the water point, if there is one
- **basin** - Geographic water basin
- **subvillage** - Geographic location
- **region** - Geographic location
- **region_code** - Geographic location (coded)

- **district_code** - Geographic location (coded)
- **lga** - Geographic location
- **ward** - Geographic location
- **population** - Population around the well
- **recorded_by** - Group entering this row of data
- **scheme_management** - Who operates the water point
- **scheme_name** - Who operates the water point
- **permit** - If the water point is permitted
- **construction_year** - Year the water point was constructed
- **extraction_type** - The kind of extraction the water point uses
- **extraction_type_group** - The kind of extraction the water point uses
- **extraction_type_class** - The kind of extraction the water point uses
- **management** - How the water point is managed
- **management_group** - How the water point is managed
- **payment** - How the users of the water point pay
- **payment_type** - How the users of the water point pay
- **water_quality** - The quality of the water
- **quality_group** - The quality of the water
- **quantity** - The quantity of water
- **quantity_group** - The quantity of water
- **source** - The source of the water
- **source_type** - The source of the water
- **source_class** - The source of the water
- **waterpoint_type** - The kind of water point
- **waterpoint_type_group** - The kind of water point

And one raw target variable:

- **status_group** – operating status of the water point

Note that the raw target variable (status_group) is nominal and has three possible values:

- **functional** - the water point is operational and there are no repairs needed
- **functional needs repair** - the water point is operational, but needs repairs
- **non-functional** - the water point is not operational

It is your decision, if you want to model the target variable in its raw form, if you want to transform it into a binary target variable (and if so, how), or if you want to treat it as ordinal. Depending on the specific objectives you define for your project and the way you would like your final model to be deployed, one of these transformations might be preferable.

Your task

- ***Please develop a model that predicts, whether a water pump is likely to fail.***
- ***The model should be optimised in such a way that it enables the Tanzanian government and its partners to prioritise their water pump inspection and maintenance operations and thereby ensure that as many users as possible will have stable access to clean, potable water.***

- ***Please document your data mining project in one single PDF file and upload the file before the end of the 48-hour examination period.***
- ***Please structure your report as follows:***

Title page

Executive summary

1. Introduction
    - 1.1. Background
    - 1.2. Objectives
2. Method
    - 2.1. Data
    - 2.2. Target variable
    - 2.3. Input variables
    - 2.4. Data pre-processing
        - 2.4.1. Transformations and derived features
        - 2.4.2. Treatment of outliers and missing values
        - 2.4.3. Sampling and data partitioning
    - 2.5. Prior probabilities and profit/loss matrix *(if relevant)*
3. Results
    - 3.1. Candidate models
    - 3.2. Model selection approach
    - 3.3. Final model
        - 3.3.1. Overall predictive accuracy
        - 3.3.2. Observed versus predicted target values
        - 3.3.3. Improvement over baseline
        - 3.3.4. Profit and loss *(if relevant)*
4. Discussion
    - 4.1. Assessment of model performance
    - 4.2. Contribution to the solution of the business problem
    - 4.3. Deployment recommendations
    - 4.4. Recommended follow-up activities

Appendices *(if relevant)*