**Data Mining Final Project**

**Report, Excel file, PowerPoint & Presentation due on or before 10pm March 13, 2021  
No Extensions and No Resubmits; 10% penalty for each day late**

Submit only ONE assignment per team; each team member will receive their own final project grade.

**Datasets**:

* Suggested datasets for the final project are in the XLMiner Module Dataset section on Canvas.
* Please add your group members names to the following google sheet by the end of second week: [ISM6136TeamsGoogleSheet](https://docs.google.com/spreadsheets/d/1uN-SiDHJgEKTIyrzFJyfDHAoc3JZkbLo9eiaaa8E5y8/edit?usp=sharing) and add the name of the dataset selected for the final project before February 20, 2021.
* Each team must use a unique dataset.

**Actions:**  **Data Mining Final Project Checklist of things to do and to report**

* Prepare/clean and understand the data (Very important to know your data)
  + identify the outcome variable and how it is represented in the data set
  + explain any recoding processes completed in XLMiner or Excel
  + explain the handling of missing data and/or outliers
  + if a variable was removed state why.
  + at a minimum provide three visuals (graphs) of your data relative to your key (outcome) variable
  + provide descriptive statistics for variables of interest
  + state the original total number of records and number of variables and final number of records and variables in the clean data
* Provide a complete descriptive data dictionary for the clean data
  + describe each variable and its data type (numeric, binary, dummy, categorical)
  + identify any coded values (i.e. Yes=1 and No=0)
* State the percent of records used as training and validation records (should be the same ratio for each model).
* Apply your dataset to three of the data mining techniques used in the class projects.
  + Linear Regression or Logistic Regression
  + Classification Trees or Regression Trees
  + Neural Network Classification or Neural Network Prediction
* Provide a sample of about 10 records of the prepared data in the report
* State your analysis process such as the alternative runs (parameters) you tried with each model. The following are suggestions not limitations to the discussion and your previous assignments can be helpful in knowing what is expected for each model.
  + For a regression model, did you remove the non-significant variables and rerun the regression?
  + For the Classification trees, what were the decision tree options (layers, minimum # in the terminal node)? Identify at least 2 rule paths to the outcome.
  + For a neural network, did you run a classification or prediction model (# of epochs, etc.)?
  + Provide key/important indicators and evaluation information about each technique as they apply such as:
  + How good was the model at predicting or classifying what you were looking for (True positive)?
  + Accuracy rates, RMSE, F1-score, etc.
  + Regression - r2, significant variables, is the model significant, state the regression equation.
  + Tree rules - do not use the words right or left, state what the node values mean, what variable(s) did the model find to be the most important.
* Provide anything else that you feel is pertinent or interesting about your project and the data story.
* Analyze each technique and provide a comparison chart of the techniques and evaluations.
* State your conclusions and recommendations.
* Submit a presentation recording, based on your PowerPoint slides, that is between 15 and 20 minutes, all team members must present. **Presentation is due by March 13th at 10pm**
  + Create using the method you prefer; Screen-cast-o-matic is one suggestion.
  + Submit the file to your team’s discussion board.
  + Monitor your discussion board for any questions from other students.
* Watch at least two other team’s presentations and post a question, comment, or suggestion for each of those presentations. Constructive criticism is helpful and should be appreciated.  
  **Comments and responses are due by March 18th 10pm**.

Submission to Canvas:

* PowerPoint, Word Document, Excel file and Presentation recording.
* The PowerPoint must follow the Presentation-Case Study template provided on Canvas.
* The Report document should closely follow the case study template as a guideline.
  +  Executive Summary  Objective  Dataset  Graphical Analysis  Statistical Analysis  Implementation  Performance Evaluation  Conclusion  References
  + The report must include a cover page and page numbers. Use type size 11.
  + The XLMiner results that are discussed should appear as tables or figures, either in an Appendix, or in the body of the text.
  + Do not simply tack on a pile of output which the reader will have to wade through to find the relevant analyses.
  + All output should be clearly labeled and edited to remove non-essential portions; remove extraneous blank lines. Numbers in the report and PowerPoint should have no more than 4 decimal places.
  + As a rule, do not include any output which you do not mention, describe, or discuss in the text of your paper.
* Evaluation of team. The last submission is a Canvas message to me by each individual team member expressing how well the team worked together. While each team member is responsible for the documents turned in and knowing how to complete the tasks, your team may have separated the tasks. Please list which team member worked on the elements of the projects.

As Shakespeare said, *brevity is the soul of wit*!  
 be brief and to the point, stimulating your reader's interest.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_