CSC423: DATA ANALYSIS AND REGRESSION / CSC 324: DATA ANALYSIS & STATISTICAL SOFTWARE II

Final Project | Total Points 100

The purpose of the final project is to demonstrate your ability to apply the knowledge and the techniques learned during this course. The final project for this class is more extensive analysis task, chosen by you from among the topics we discuss. Final projects will include a presentation to the rest of the class at the end of the quarter, in place of a final exam and a 2-part final paper.

Whenever it is possible, it is recommended that the online students attend the final presentations to participate in the live discussions of the final projects and to complete critiques of the other project presentations. Alternate arrangements will be provided for online students to do their presentations or submit it as a video recording.

DELIVERABLES FOR THE FINAL PROJECT:

I. Select Teammates (Week 6)

For both in-class and online students, the group size will be determined based on the class size. Online students can **ONLY** partner up with fellow online students. If students are looking for groupmates, post the following information under D2L discussion post "Final Project - Looking for Teammates" so that students without a groupmate can respond.

Information to include:

1. Full Name
2. Online or In-class student
3. Your research areas interest/types of dataset you like to analyze [optional]

Note: Online students will submit their presentations with a voiceover or video option. Voiceover is very straightforward. If you do not know how it works, please google the procedure and be familiar with the process ahead of time.

II. Select Dataset and Develop Project Proposal (Week 9) **[5 Points]**

Datasets / Data Sources:

There are several repositories for large datasets that are available online. The minimal requirement for the dataset is that it contains at least **6 variables** and **more than 300 records/observations**.

Below is a list of datasets and repositories you could get data from. A dataset could be used by a maximum of 2 groups. Once you select the dataset, post your name(s) and the dataset you selected on D2L under "Final Project – Topic Selection". Make sure to check this discussion thread before selecting the dataset to make sure two teams haven't already selected it.

Here is a list of datasets and repositories:

Simple Datasets:

This is a **last-resort datasets**, and if selected you will **lose 2 points**, as everything needed for the proposal is provided and you are really not learning anything in terms of putting together the proposal

- Datasets posted under "Group Project" section of D2L's CONTENT menu.

Data Repositories:

- UCI repository: http://archive.ics.uci.edu/ml/datasets.html. Select datasets that have "Default Task" listed as "Regression".
- Baseball players statistics: This site has data on the 2008 Major League Baseball season http://www.exploredata.net/Downloads/Baseball-Data-Set. Teams interested in analyzing the data can choose a response variable and a subset of 20 attributes to analyze.
- Statistical Data Sets - University of Massachusetts Amherst: Provides data sets for various types of analysis. Select a dataset from regression (multiple, or nonLinear) http://www.umass.edu/statdata/statdata/index.html
- University of Florida Statistics Professor's Miscellaneous Datasets: Prof Larry Winner, University of Florida Department of Statistics, provides links to a long list of data sets organized by statistical technique. Select a dataset from regression [LR 1a) Linear Regression or LR 1b) Nonlinear and Polynomial Regression] http://www.stat.ufl.edu/~winner/datasets.html

Complex Datasets:

(Will need data cleaning, recoding and/or summarization, etc. before you begin your analysis)

- KDnuggets is a great website that contains lots of information of interest to data scientists. It also includes a long list of data repositories: http://www.kdnuggets.com/datasets/index.html
- Datasets used for data analytics competitions at https://www.kaggle.com/datasets

Extra Credit: **Extra credit will be awarded to students based on 2 sets of criteria -**
Two extra credit points will be provided to students implementing items 1-3
Three extra credit points will be given for students implementing 1-3 and either item 4, 5 or 6

1. Data obtained from the complex dataset list (KDnuggets or Kaggle).
2. Dataset should contain at least 10+ predictors. To qualify for the extra credit, you have to use these predictors in your analysis.
3. Predictors should consist at least one qualitative predictor with more than 2-3 levels and quantitative predictors. To qualify for the extra credit, you have to use these predictors in your analysis.
4. Extensive data preprocessing. For example, combining multiple datasets, removing 1K+ outliers, recoding multiple variables.
5. Dataset consists of date/time and/or location-based predictors which are recoded and used in your analysis.
6. The number of observations should be 2,000 or more.

Proposal:

Submit 1-page proposal that includes:

1) Project title: be creative, come up with a catchy title (if possible)
2) Team mates: Full names of all team mates as it appears on campusconnect
3) Dataset: This should include
    a. dataset name
    b. brief description of the dataset
    c. # of DV(s) and description of the dependent variables
    d. # of IV and description of the independent variables, # of numeric variables, # of text variables, # of date/time variables, # of location related variables
    e. number of rows/observations and
    f. the URL to the site where you got the data from
4) Problem description: What you plan to predict, analyze, etc. and why
5) Proposed methodology: Proposed approach as to what steps you will follow to address when you mentioned in (4) above.
6) References: Use at least three references other than text book or class notes. References are journal articles, or research papers that will be helpful in understanding what scholars and industry experts suggest in terms of methodology, variables, or future direction for similar datasets. See document on References_How to cite them.PDF to see how to cite and use references.

Extra Credit: **Two extra credit points will be provided to those who implement approach or methodology obtained from the references that was not covered during the course.**

Presentation (Week 11) **[20 Points]**

Presentations should be 8-10 minutes long (subject to change depending on the number of groups). Each member should present a portion of the analysis and answer any question that the instructor or the class mates have.

Each member will have to come up with their own final model that is distinct from their teammates, and do their own coding and analysis of the model as well.

What to present: Should be geared towards technical audience

1. Project title and team member names as it appears on campus connect
2. Objective and goal
3. Data – variables and definition, dummy variables, interaction terms, transformations, polynomials, etc. – high level
4. Methodology used – high level
5. Compare the models (all aspects discussed in class) and its performance on train/test – for each team member
6. Selection of final model(s) and reason(s) why?
7. Discussion on limitations of study, and future direction

8. Compare and contrast your findings with your references and if methodology from references **that wasn't covered in class** were implemented. Explain how.
9. **Presentation slides should be uploaded to D2L before class**

How to present it:

1. Text should be readable from a few feet away. So consider using a minimum font size of 18 points or larger
2. Use graphs with adequate labeling and titles to ensure that graphs will be easy to interpret by everyone
3. It is important that you don't overcrowd your slides. Simplicity is the key to make effective presentations.
4. Use text effectively
    a. To introduce the project (what problem was investigated? what data was used?);
    b. To explain graphs and highlights important aspects of your study;
    c. To present the techniques you used and
    d. More importantly to interpret the results of your analysis. Conclusions can be summarized in a bullet-point list
5. Colors and graphics should be used to enhance the presentation and not just for decoration
6. **Do not**
    a. Copy and paste software output with no editing
    b. Use variable names in your output. You should use meaningful variable labels
    c. Include basic diagnostics results (QQ plots, residual analysis, etc...). These are details that you should have in the report. For the presentation focus on the main results

Reports (Week 11) **[75 Points]**

Each group member should write his/her SAS code to solve the problem, but the group will write a single analysis of the results using a word processor. The analysis should be submitted as 2 separate reports:

1. **A Non-technical Summary Report (15 Points):** No longer than one typewritten page, describing the objective or goal and conclusions of your statistical analysis to a non-technical audience. It should be understandable to a person who does not know regression analysis or statistics. Often, in your workplace, you will have to present your findings to non-technical folks to convince them without using any technical jargons. Make sure to include the model statement.
2. **A Technical Summary Report (60 Points):** A 5-8 page technical report should include the following sections. The appendix, code and references don't count towards the 5-8 page limit. It should also include all the important outputs and the codes in the appendix section. This report is intended for a statistically literate audience and must be written in a clear organized fashion using the correct terminology. It should consist of the following sections:

Abstract	Give a short summary of the goal, approach/methodology and important findings and recommendations	Group Effort	2 Points
Introduction	Describe the goal or objective and any hypothesis, any literature review or background research you did using the references, why it is important, context, motivation etc.	Individual Effort if Goals are different, otherwise,	2 Points

		Group Effort	
Methodology	Steps of your approach, specifically where you obtained the data (site the exact data source/link), how you pre-processed or cleaned the data (recoding, transformations, interaction variables, etc.), model approach, validation method, and any type(s) of analysis did you performed	Individual Effort	6 Points
Analysis, Results and Findings	Your analysis should address the following points: 1. The exploratory analysis of the data including descriptives that may suggest a possible model that is adequate for fitting the data. Do the data show a non-linear relationship? Should a transformation of the response variable and/or the predictors be useful? 2. Try interaction variables. 3. Check for collinearity among the independent variables. 4. A variable selection method will enable you to select suitable models and find the set of predictor variables, which are more informative for predicting the response variable. 5. You may want to fit a few models (at least 1 final model per teammate) that seem adequate for your data and then select the model among them that provides the ``best'' prediction of Y. 6. Analyze the residual plots to look for patterns that might suggest a failure in the assumptions and some inadequacies in the selected model. 7. The existence of outliers and influential points may have dramatic effects on your analysis. Check also if there are outliers. 8. Can your model be improved? Are you satisfied with the model you have chosen? 9. Use the selected regression model to examine the relationship and associations among the variables in your study and to identify, among the observed independent variables, the strongest predictors for the response variable. 10. Compute two predictions including the prediction intervals using the regression model. 11. Apply validation techniques to evaluate the predictive power of your model. Split the original dataset at random into a testing set and a training set. Trainings set should have at least 15 observations in order to compute meaningful validation statistics. Use the testing set to estimate the parameters for the selected regression model, and apply the fitted model to predict the values in the training set. Compare predictions and observed values in the training set using mean square error and cross-validating statistics. Discuss the model predictive performance. **IMPORTANT:** Each team member should come up with a final model that is distinct from the rest of the team and evaluate the performance metrics using test sets. There will be a 20% reduction in both presentation and final report grade if test set performance is not included. There will be a 50% reduction in both presentation and final report grade if SAS Code is not included or doesn't run. ***Hints for the Statistical Analysis:*** It is possible that you may not	Individual Effort	40 Points

	find a satisfactory model that fits adequately your data. Sometimes a data set may admit more than one satisfactory answer; sometimes there may be none. If the statistical analysis shows that no regression models are suitable for your data set, mention what approaches you have tried and what was unsatisfactory about them. If there is more than one suitable model, mention the pros and cons and compare their performance in predicting the response variable. The final aim of any statistical analysis is the understanding of a phenomenon or the investigation of a scientific problem, which your data arise from. Remember that the regression function is a mathematical representation of such a problem and the interpretation of the parameters values will give you insights about the relationships of the variables in the problem.		
Future Work	Any additional avenues worth exploring based on what you have discovered so far? Does the current results obtained suggest new directions worth exploring by you? Explain how?	Individual Effort if Goals are different, otherwise, Group Effort	2 Points
Appendix	All relevant outputs should be included here and cross referenced in your Analysis, Results & Findings section. Appendix should be the last section of your report.	Individual Effort	2 Points
Code	Attach the SAS code from each member to the zip file along with the dataset used. **If different datasets were used by each member, provide the code and the dataset labeled with your name as a prefix (e.g. Jake_SAScode_loans.sas and Jake_loans.csv), so that when the code is run, I see the same output as what is provided in the report. If the code is not included or doesn't execute, you will lose 50% of your grade for the report and presentations.**	Individual Effort	2 Points
References	Papers that you read and cited in your paper/final report, data sources. See document on References_How to cite them.PDF to see how to cite and use references. There should be at least one citation per team member.	Individual Effort	2 Points
Zip File	Zip the folder with all the data files (raw and cleaned excel/PDF files, all corresponding files used), each team member's SAS code and the 2 reports.	Group Effort	2 Points

Team Contribution (Week 11) *(-2 points if not submitted)*

Group members should also submit the team evaluation document via D2L (see TeamEvaluation.docx under "Group Project" section under Content). Two points will be deducted if evaluations are not submitted.