DSC 423: Data Analysis And Regression / DSC 323: 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 2-part final paper.

FINAL PROJECT:

I. Find Teammates (Lecture 1-5) *[Individual Effort]*

The group size will be 4 or 5 member teams. You can find teams in class or 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. Your research areas interest/types of dataset you like to analyze [optional]

II. Select Dataset and Develop Project Proposal (May. 10) *[Group Effort]* **[5 Points]**

Datasets / Data Sources:

There are several public repositories available online which includes datasets from various industries. **The dataset(s) you pick has to be publically available, and cannot be from private source or sensitive in nature**. The minimal requirement for the dataset is that it contains at least **6 predictors (before creating dummy variables)** and **more than 600 records/observations. If you have taken this class before, you cannot use the dataset you used previously. Do not go beyond 3,000 observations.**

Suggested Dataset Sources:

(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

Proposal:

Submit 2-3 page proposal (can exceed the page limit) 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 D2L
3. Dataset: This should include
   a. dataset name
   b. brief description of the dataset
   c. # of DV(s) and description of the dependent variables including data type (number, text, etc.)
   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. (*Hint: Make a list of everything we have learnt in class, and put them in order)*
6. Time line: Provide a timeline for completing the different phases of the project. Here is a rough timeline, **but you have to modify it** to include the various sections you included in your methodology as this doesn't breakdown the last part.

   Suggested Timeline

   Lectures 1-5 : Form teams

   Lectures 6-7 : Pick dataset and write proposal

   Lecture 7-10 : Data cleaning, exploration, analysis, writing report, submission of all deliverables, team evaluations

Report (Jun. 7) **[95 Points]**

Each group should write their SAS code to solve the problem and write a single analysis of the results using a word processor.

Note: Page length does not matter, if you need more pages to explain what you need to it is OK.

A Technical Summary Report: ***[Group Effort]*** A 30-50 page technical report should include the following sections. The appendix and SAS code do not count towards the limit. SAS cod should be submitted in a SAS file. It should also include all the important outputs 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. The format should be like a report or journal article, as **NOT** like your assignment. It should consist of the following sections:

Abstract	Give a short summary of the goal, approach/methodology and important findings and recommendations	5 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.	5 Points
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	10 Points
Analysis, Results and Findings	Your analysis should address the following points: **Note:** This is a list of techniques is provided only for guidance, and is not given in the order of execution. Students should review all the course materials to come up with a list of all techniques learnt in class, and implement the relevant techniques for your dataset in the right sequence. 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. Use either liner, logistic, polynomial regression techniques and/or transformations 4. Check for collinearity among the independent variables. 5. 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.	55 Points

	6. You may want to fit a few models that seem adequate for your data and then select the model among them that provides the ``best'' prediction of Y. 7. Analyze the residual plots to look for patterns that might suggest a failure in the assumptions and some inadequacies in the selected model. 8. The existence of outliers and influential points may have dramatic effects on your analysis. Check also if there are outliers. 9. Can your model be improved? Are you satisfied with the model you have chosen? 10. 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. 11. Compute two predictions including the prediction intervals using the regression model. 12. Apply validation techniques to evaluate the predictive power of your model. Split the original dataset at random into a training and test set. Test set should have at least 15 observations in order to compute meaningful validation statistics. Discuss the model performance using training, and testing sets. ***Hints for the Statistical Analysis:*** It is possible that you may not 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	For **Graduate Students ONLY** (if there are grad and undergrad students in a team, the undergrads do not have to do this part) Explain in detail 1. Any additional avenues worth exploring based on what you have discovered so far? 2. Does the current results obtained suggest new directions worth exploring by you? 3. Did any of the research and references you read help improve your research further?	10 Points
Research References	For **Graduate Students ONLY** (if there are grad and undergrad students in a team, the undergrads do not have to do this part) 1. Read 5 Research papers, articles, journals relevant to your project 2. Cite them correctly 3. Cross reference them in your narrative under Introduction, methodology, analysis and results and/or future work. ***Note:*** Just adding only reference as a bibliography will result in zero for this section.	10 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.	- 5 Points if not submitted

IMPORTANT

- If the code is not included you will receive a 0% as the code determines the analysis, and the outputs necessary for the report
- If the code doesn't execute, you will lose 50% of your grade. However, if the errors are at the beginning or at the time of importing, you will receive a 0% as code determines everything else
- If the data file is not provided or incorrect file is provided, you will lose a significant portion of the grade as I won't be able to run your code to check for accuracy
- If validation is not included, 20% of the grade will be lost as the whole purpose of building the model is to see if it can successfully predict unknown values, and model validation tests for that
- If team evaluation is not submitted you will lose 10% of your grade
- If you didn't contribute to the project, you will receive 0% for the project

Team Contribution (Jun. 7) ***[Individual Effort]*** *(-10% points if not submitted)*

Group members should also submit the team evaluation document via D2L (see TeamEvaluation.docx under "Group Project" section under Content). Ten percent will be deducted if evaluations are not submitted by the due date/time.

What to Submit

1. Proposal (See Schedule for due dates)
2. Final Project Deliverables

Group Effort

1) SAS Code – should be in .SAS file format. Double check you are submitting the code and not the log file
2) Data file - Double check if it is the correct version of the data file that you ran your code on
3) Report – Should be in Word file format. Make sure all sections mentioned in the instructions are included in the report. Also, the appendix should be correctly annotated and cross-referenced in your narrative. All outputs should go in the appendix section at the end of the report.

Individual Effort

4) Team evaluation – template is provided under group project section on D2L