NORTHWESTERN
UNIVERSITY

SCHOOL OF
PROFESSIONAL
STUDIES

MSDS 498: Model Development Guide

Getting Started

Everyone in MSDS 498 will complete a capstone project in credit modeling. Everyone will work on the same data set – the credit card default data set from the UCI Machine Learning Repository. This document will outline the prescribed format for your modeling documentation. Additional documents will provide the prescribed outline for the production model document and the performance monitoring document.

Project Data

The data for the capstone project is from the UCI Machine Learning Repository. The raw data set and its documentation can be found on the UCI Machine Learning Repository website.

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

The paper that uses the data will also be provided in the course reserves.

For the project we will use the version of the data set provided in the course shell. The provided version will be cleaned to a format easily read into R, and it will already be split into a training data set, a testing data set, and a validation data set so that all students have exactly the same data sets for each part of the project. The ‘split’ will be in the form of flags – train, test, and validate which will take values 0/1. Students can subset the data set to create the three data sets needed for the project. For the part of the project outlined in this document we will use the training data set and the testing data set. We will save the validation data set for the final stage of the capstone project – performance validation.

General Paper Guidelines

The general guidelines are as follows.

(1) The paper should follow the prescribed final paper format outlined in this guide.

(2) The paper should be well written. The formats are designed to help you with the overall paper structure. The remainder of the presentation is left to you.

(3) The term data is the plural of datum. We should write 'the data are', not 'the data is'.

(4) The body of the paper should be between twenty and forty pages, double spaced. The title page, the bibliography, and any relevant appendices do not count against the page count. A good target number is the midpoint of thirty pages.

(5) Graphics and tables should be labeled so that they can be referenced. Typical labels are 'Figure 1' and 'Table 1'. Graphics and tables should also be large enough to be easily read and centered in the page.

(6) All papers should have page numbers on each page in the lower right hand corner.

General Grading Guidelines

Your final paper is worth a total of 500 points. Here is how those points will be broken down.

(1) Exploratory Data Analysis (100 points)

- This category will include the description of your data, your exploratory data analysis, and your feature engineering.
- You should only show relevant EDA and feature engineering results.
- If you do not feel that a result is worth discussing, then it should not be included in your report.

(2) Predictive Modeling (300 points)

- This category will focus on the section ‘Predictive Modeling: Methods and Results’.
- Be sure to present each method as its own subsection.
- Each subsection should be clear as to what model you are fitting and what results you obtained on both the training and test data sets.

(3) Comparison of Results (50 points)

- A summary of your results should be presented in a format that makes the effectiveness of the different models easily comparable.

(4) Overall Exposition (50 points)

- This category will include the ‘Introduction’ and ‘Conclusions’ sections and the overall exposition of the report. The report should be well written and use correct English grammar, punctuation, and spelling.
- The paper should be proofread and free of typos, poor grammar, or poor spelling.
- Paper should read well and be properly formatted with respect to the specific and general guidelines.
- The ‘Introduction’ should outline your problem and frame your paper.
- The ‘Conclusions’ section should wrap up your paper. It should not be two or three sentences long.
- The Conclusion can include discussion of what worked well and what did not, and suggest other avenues of investigation.

Final Paper Format

Here is an outline for the format of your model development documentation. Students are expected to follow this prescribed format. In addition all papers should contain a title page.

Title:

Model #101: Credit Card Default Model

Model Development Guide

1. Introduction

- Provide an overview and general statement of the problem.

- Provide a general discussion of how you have approached the problem and highlight some of the interesting results.

- Remember that the Introduction is an introduction to your paper.

2. The Data

- Provide a description of the data.

- Provide a table data dictionary of the default variables.

- Provide a data quality check through the use of data summaries.

- Your data is not clean! I have packaged and provided some structure to the data. However, I have not made it 100% clean for you. You should be finding some discrepancies between what you see in the data and what is supposed to be in the data based on the data dictionary. Please be aware of this fact. This is part of your data quality control exercise. You will want to note these discrepancies and your remedies for these data issues.

- You should not be deleting these ‘dirty’ observations. Instead you should be remapping them to valid values. Hint: Usually the observation will have a value that is not in the data dictionary. However, the data dictionary will have a coded value for ‘unkown’ or ‘other’. Any values that do not make sense should be mapped to ‘unknown’.

- Provide a table of the number of observations in the training data set, the testing data set, and the validation data set.

- Write the section with the intention that it will be read by someone who is not familiar with the problem. As members of the course we all are using the same data, and hence we are all familiar with the data, but we are not the primary audience.

3. Feature Engineering

- In this project we will need to engineer additional features from the variable provided. Admittedly there are some issues with the naivety of this data set. We will help alleviate those issues by giving you some guidance on feature engineering.

- Note that the billed amount is the monthly balance.

- Compute utilization = balance / credit line. How should you scale? [0,1] or [0,100]?

- Compute payment_ratio = payment / balance. Define a 0 payment / 0 balance as 100. Do we understand why we should define 0 payment / 0 balance as a 100 for our modeling problem? What is the payment_ratio measuring? Hint: you will want to make sure that you line up the payments with the correct bills.

- Consider binning age. How should we do this? Maybe use a decision tree as an exploratory tool for age. You can also consider binning other variables if you are interested.

- Compute some measure of balance velocity or increase. You can look at increases in utilization instead of balance to have the increase be normalized. You can define several types of measures here. One would be the increase in the utilization over the history of the series. Another would be the difference between the minimum utilization and the current utilization. Try a couple of different measures. Be creative.

- If you use statistical graphics, algorithms, or any other exploratory data analysis methods to define your features, then show the output.

- Final Note on Feature Engineering – Any variable that you define should have an exact definition in this section so that any reader can reproduce the feature.

- **Note that I have provided a separate document with feature engineering instructions. See the Weekly Video Library.**

4. Exploratory Data Analysis

4a. Traditional EDA

- Provide highlights of important features in your data using statistical graphics and data summaries.

- Provide a table of data summaries to act as a data quality check for your engineered features.

- What statistical graphics are appropriate for a binary classification problem? (Hint: Scatterplots are not appropriate.)

- Remember that the purpose of EDA is to explore the relationship between the predictor variables and the response variable, not the relationship between the predictor variables. Also note that we are interested in the engineered features, not the raw data elements.

4b. Model Based EDA

- Fit a decision tree using rpart. Plot the tree dendogram using this example.

http://blog.revolutionanalytics.com/2013/06/plotting-classification-and-regression-trees-with-plotrpart.html

- Does the decision tree inspire you to make a few additional data summaries that are related to its output? Maybe the decision tree keys on one or two variables as we should go back and make some box plots with this variable to see a separation of the two classes. Maybe we should make some other type of plot.

- As data sets get larger (i.e. wider) we tend to use computational methods to explore the data. Some common approaches are to use decision trees, random forest, and automated variable selection with generalize linear models. We will use the decision tree here, and random forest and logistic regression with automated variable selection in the next section of the paper.

- Use OneR on the final set of features. Which features does OneR suggest are important? Does OneR order them?

- You could also use simple logistic regression models to provide some model based insights.

- Tree = mandatory, OneR = mandatory, Simple Logistic Regression Models = optional

5. Predictive Modeling: Methods and Results

- In this section we will provide the results from four modeling approaches. Three of these modeling approaches are defined for you, and you get to choose the fourth approach from the list of choices.

- For each model provide any relevant or useful model output and a table of the model performance in-sample (i.e. on the training data set) and out-of-sample (i.e. on the test data set).

- The metrics to be measured are: (1) true positive rate or sensitivity, (2) false positive rate, and (3) the accuracy.

5.a Random Forest

- Include the variable importance plot.

5.b Gradient Boosting

- Use GBM or XGBoost packages. Include the variable importance plot.

5.c Logistic Regression with Variable Selection

- Random Forest and Gradient Boosting will identify a pool of interesting predictor variables. Use that information to help you choose an initial pool of predictor variables. List your initial pool of predictor variables in a table.

- Choose a variable selection algorithm. Use that variable selection algorithm to arrive at an 'optimal' logistic regression model.

- Since this is a linear model, you should provide a table of the model coefficients and their p-values.

5.d Your Choice – CHAID, Neural Network, SVM, or some other method appropriate for binary classification.

- Provide the relevant output for the model of choice. For example SVM has margin plots that are useful, and a neural network allows you to plot out the network topology. If your chosen method has a 'standard' plot that is typically shown with it, then we all expect to see that plot, and you should be providing that plot with the model.

6. Comparison of Results

- Aggregate your results from Section 5 and discuss. All of your model metrics should be presented in a single table for all models for both the training and test data sets. You should be able to compare and contrast the model performance with discussion and easily determine which model performed best.

7. Conclusions

- Conclude your paper. Reiterate your problem and highlight your results.

- How would you characterize the overall quality of your results?

- Do you have any recommendations for approaching the problem in a different manner or with different techniques? Would you recommend any particular avenues for future research?

- A lot of times a good conclusion reads like a good abstract.

8. Bibliography (if needed)

- References can be constructed using any valid style format. However, references can be cited in your paper using a name-year citing or by using the number scheme, e.g. Bhatti [1].

- One type of citing used in the program is the APA style format. However, it can be difficult to use with some types of sources, hence we will allow the number format for convenience.

- When citing references consider using: Bhatti [1] for a single author, Bhatti and Lucas [2] for two authors, and Bhatti et. al. [3] for three or more authors.

X. Appendices (if needed)

- Appendices are where we put useful, but not primary information. Useful but not primary information could be summary statistics, auxiliary plots, or any additional or auxiliary details and discussion.