Centre of Geographic Sciences
COGS | nscc

GDAA 2010 | Data Mining Modeling

Assignment #1: Predictive analytics with a continuous target

Academic Year 2021/2022
Value: 20%
Due date: March 25, 2022

Apply data mining models for the following two tasks:

1. Use IBM Modeler to build predictive models for a numeric (continuous) target variable with the following eight models:
    a. Neural network (node *Neural Net*)
    b. Linear regression (node *Regression*)
    c. Linear (node *Linear*)
    d. Generalized linear (*node GenLin*)
    e. General Linear Mixed Models (node *GLMM*)
    f. Support vector machine (node *SVM*)
    g. Linear support vector machine (node *LSVM*)
    h. K-nearest neighbor (node *KNN*)
2. Carry out PCA/Factor Analysis of your data using IBM Modeler

<u>Data preprocessing</u>

1. Use Business Analyst to extract a geospatial dataset (e.g., a series of dissemination areas)
2. Prepare your dataset with one numeric (continuous) target variable and at least 15 numeric predictors (note: you can have categorical predictors as well, though these should be in addition to the 15 numeric predictors).
3. All missing data should be replaced with mean values.
4. Analyze your data quality using the Data Audit node. All records should be valid.
5. Calculate the correlation between potential predictors and target. Eliminate potential predictors with weak correlation (node *Statistic*) or with low coefficients of variation (node *Feature Selection*).

<u>Task #1: Predicting a numeric (continuous) target variable</u>

1. Predict a numeric target variable with a neural network model using all numeric and categorical predictors.
2. Repeat the process using the neural network and the top most important predictors (you should decide how many to include).
3. Then, with the same set of the most important predictors, use seven more models:
    1. Linear regression
        i. Use stepwise method
        ii. Show the output from the coefficients table for the best model and comment which predictors is the most and least useful
        iii. Show the output from the collinearity diagnostics table for the best model

    2. Linear
        i. Use forward stepwise model selection method
    3. Generalized linear (use default settings)
    4. General Linear Mixed Models (use default settings)
        i. Use the panel *Main* to drop predictors
    5. Support vector machine (use default settings)
    6. Linear support vector machine (use default settings)
    7. K-nearest neighbor
        i. Do not normalize range input
        ii. Weight features by importance when computing distances
        iii. Perform feature selection
        iv. Use 3-4 top predictors (determined by neural network) as force entry

4. Merge output from eight models. Use the filter tab for editing names of duplicated output fields (*$L*) and eliminate all predictors and other output fields, except the target field and eight output fields (from eight models) representing predicted values
5. Compare results using the *Analysis* node. Pick the best model based on the minimum mean absolute error or the maximum correlation.
6. Calculate residuals for the best model. Create the histogram with normal curve.
7. Classify and label residuals. Create frequency table.
8. Map residuals using the appropriate sequence of colors.
9. Compare all available (11) models using the *Auto Numeric* modeling node. Do not partition data, do not build model for each split. Apply the stepwise method for the regression model and linear model.

Task #2: Using factor analysis

1. Use the *PCA/Factor* node. Use the maximum likelihood extraction method. Do not partition data. Extract three factors with the varimax rotation. Format the factor matrix by sorting values and hiding values below 0.2.
2. Show and comment on the total variance explained table.
3. Show and comment on the structure matrix, name factors based on loadings of variables (i.e., group them into meaningful categories of your own design, if possible).

Submit:

A report, poster, or narrated PowerPoint presentation with the following elements:

- The purpose of the assignment
- Your name and data sources
- Data preparation:
    - Data dictionary
    - Data quality report (using the Quality node)

- Output from Task #1:
    - Data quality output showing no missing data
    - Correlation table between target variable and predictors
    - Model accuracy and importance of top predictors from neural network
    - Part of the linear regression coefficients table for the best model with comments
    - Part of the linear regression collinearity diagnostics table for the best model with comments
    - Fragment of the output table coming from merging and filtering all predictions
    - Results from *Analysis* node comparing predictions from eight models with target variable with comments.
    - Histogram of residuals from the best model with normal curve
    - Frequency table
    - Map of reclassified residuals
    - Output from using the *Auto Numeric* node comparing used models
    - Importance of predictors for all compared models
    - Stream with edited modeling node names

- Output from task #2
    - Total variance explained table with comments
    - Structure matrix with comments and factor names
    - Stream

Any maps should have titles referring to a target variable and model used. Maps should be accompanied by legends, scale bars and north arrows. The same elements (maps, tables, graphs) should have appropriate and similar sizes, as well as descriptive captions.

Submit your work to the Assignment #1 folder on Brightspace.