BUS 212a - Analyzing Big Data II

Project Assignment 2: Regression

Team Name \_\_\_\_Invincibles\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
|  | Max | Pts. |
| 1. .Rmd script runs without error (warnings okay). | 5 | 4 |
| 1. Descriptive statistics for relevant, important variables. For example, box-and-whisker plots, minimum, maximum; average (mean, median, mode) and standard deviation / variance. | 10 | 10 |
| 1. Scatterplot, matrix plot, or correlation matrix | 10 | 10 |
| 1. Justification of target variable. Why chosen? | 5 | 5 |
| 1. Diagnostics: reasonable-looking residual plots and/or Variance Inflation Factor (VIF) for each predictor < 10, most of them < 5. | 10 | 9 |
| 1. Attempt at including at least one interaction or polynomial term, i.e., higher-order term in model. | 10 | 10 |
| 1. Accuracy of final, best model. For Multiple Regression, show adjusted R-Squared or RMSE. | 10 | 10 |
| 1. How many of the observations were removed outliers? What percentage of the observations is that? Does that seem reasonable and acceptable? Any leverage points left in? If so, why? | 10 | 9 |
| 1. Interpretation of the final model. For Multiple Regression, interpret coefficients, e.g., for every unit increase/decrease of X, it’s an increase/ decrease of Y in the target, which means… | 10 | 7 |
| 1. Reflections about modeling process, at the start, end, or throughout the report. | 10 | 10 |
| 1. Attractive, professional-quality report overall. | 10 | 10 |
| TOTAL | 100 | 94 |

Comments:

In all\_variables.rmd – set the working directory to desktop. Better not to put as it will not run in another laptop. In interaction.rmd – to some other location is set as working directory.

Need to explain how arrived at the 87-outlier removal – criteria and evaluation – percentage of total data. Interpretation of residual plots is missing in many parts.