Business Analytics Assignment

Kartheek Raju Mulasa

College of Engineering & Technology, University of Derby

k.mulasa1@unimail.derby.ac.uk.com

Contents

1 Introduction

The main idea of the research is to apply the machine learning algorithms on the Student Performance dataset in SAS environment. This report contains the methods used to extract insights from student performance dataset and Management information system for show casing the key steps of this process. Multiple algorithms are performed in order to find any relations between features and investigate the features causing the Student grade prediction. Algorithms like Linear Regression, Partial least square Regression and Ridge Regression performed on the mentioned set of data.

2 Dataset Description

The data set comprises of 32 features and 12640 data points. The “G3 - Final student grade” being the target variable and remaining features excluding being the independent variables. The target variable here is the continuous variable. The methodology here below data describes student success amongst two Portuguese schools in secondary education. The data characteristics include student ratings, features related to socioeconomic, financial, and education, and it was compiled using school records and questionnaires. Dataset of the results is given in two distinct subjects: Mathematics (mat) and Portuguese (por). In [Cortez and Silva, 2008], the dataset was modeled under five-level classification and regression tasks. Currently, it can be downloaded from the website https://archive.ics.uci.edu/ml/datasets/student+performance. The dataset has three files within it. The first is the ReadMe.txt file explaining the data information. The second file is a Features.txt file that explains all the features and their significance involved in the dataset. The third file is data.txt file, which is the real dataset to be used. Each of the files are in.txt format.

2.1 Feature description

1. School - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. Sex - student's sex (binary: 'F' - female or 'M' - male)
3. Age - student's age (numeric: from 15 to 22)
4. Address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. Famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's ((numeric: 0 - none, 1 - primary education (4th grade), 2 - (5th to 9th grade), 3 - (secondary education) or 4 - (higher education)).
8. Fedu - father's education ((numeric: 0 - none, 1 - primary education (4th grade), 2 - (5th to 9th grade), 3 - (secondary education) or 4 - (higher education)).

9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
11. Reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. Guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. Travel time - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. Study time - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. Failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. Schoolsup - extra educational support (binary: yes or no)
17. Famsup - family educational support (binary: yes or no)
18. Paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. Activities - extra-curricular activities (binary: yes or no)
20. Nursery - attended nursery school (binary: yes or no)
21. Higher - wants to take higher education (binary: yes or no)
22. Internet - Internet access at home (binary: yes or no)
23. Romantic - with a romantic relationship (binary: yes or no)
24. Famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. Free time - free time after school (numeric: from 1 - very low to 5 - very high)
26. Goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. Health - current health status (numeric: from 1 - very bad to 5 - very good)
30. Absences - number of school absences (numeric: from 0 to 93)
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

3 Problem Statement

The problem statement is to develop model to predict the student final grade that reflects best-adjusted R-Square and RMSE values. Secondly, Find the features influencing the target variable G3 (Final student grade).

4 Data Preprocessing

First of all, in order to implement the algorithm, the data must be correctly sorted, and any incomplete or null data assumptions should be handled. Different stages of data preprocessing will be involved in the process (Preprocessing, 2020).

4.1 Data Loading

For the study, details seem to have been taken from the local machine. So, set up the Working Directory initially. In addition, pull out all the data collection stored in the working directory at the designated site. The command Proc Import is used to load the data into SAS environment.

4.2 Data Summary

Whenever the data is available, the basic SAS commands are used to interpret data. This involves identifying data structure, data present characteristics, descriptive data figures, data forms correlated with data, and existence of incomplete or NA attributes, existence of outliers, data skewness.

4.3 Handling-missing values or Na values

If there is any in the data collection check for incomplete or NA values. Whether the data includes the missing values then these missing values will be imputed with the mean or median values or imputed on the basis of the domain knowledge with the individual values. Our dataset does not contain any missing values, so it does not require any modifications.

4.4 Handling the outliers

Often verify whether outliers are present in the data. This is achieved by measuring the plot of the box. The observations that fall above and beneath the whiskers are called outliers. For some of the features called "G1 and G2" (student grades), our data collection includes the outliers. Here the outliers are not substituted as outliers to certain data points, which are comparatively small.

4.5 Distribution of the target Variable

Test the distribution of target variable. It should follow through on normal distribution. By plotting the histogram, this can be calculated. If the target is biased left or right then apply the necessary transformations such as log, reverse transformations to render it distributed normally. The Proc univariate command is used to test goal variable normality. Our target variable “G3 Final student grade” is normally distributed.

4.6 Data Exploration

This method is used to evaluate the data set's distributions and occurrence differences graphically. Some of the key strengths in this method is discovering the correlation between the variables input and target. There should be no multi-collinearity among the input variables according to the standards of machine learning algorithms. Data exploration is also used for analysis of the input variables distributions. This is achieved

through the plotting of different kinds of plots such as scatter plots, bar graphs, histogram, donut plots, pie charts etc. This identical method is observed in the study of the different attributes contained in our dataset. Key insights are drawn from this data exploration like G1 and G2 strong correlation to target variable thus for eliminate one feature when applying regression. Secondly more female students are failed in the final grade juxtapose with male students. Most scored mark across the students is 10 and followed by 12 for the second position. Percentage of male students are scored in the top marking 20 compared to female students Proc gchart/sgplot and ODS graphic designer also used for visualization. Proc corr is used to check collinearity within in input features.

4.7 Data Standardization

For improved readability and accurate forecasts, the data that is transmitted to the model will be on the same scale. The independent variables in the data collection should be weighted to be comparable. Implementation of min-max scaling and z-score scaling will do this. The input variables are taken to the same scale, by adopting one of these. No modification should be made to the target variable. The command Proc Stdize is used for standardizing input features.

4.8 Dummies

These would be the two ways the categorical or ordinal factors may be transformed to numerical values. First, search in the data collection for the categorical variables. If any, then add the dummies or single-hot encoding functions to translate them into numerical values. Since the algorithm is unable to interpret the data present in the form of strings, translating it into numerical values is necessary. Once all such procedures have been applied, the data should now be able to apply the appropriate algorithms. This report attached the SAS code in the appendix and highly encouraged to look each data preprocessing steps and insights drawn.

5 Application of the algorithms

If the appropriate data is available, then the necessary algorithm will be implemented according to the problem statement. In our list of issues, because the goal variable is a constant variable, we initially chose to use the linear regression method.

5.1 Linear Regression

This is the supervised algorithm for machine learning, where the target variable is a continuous variable. The contingent variables can be integer, ordinal or numerical (Linear Regression — Detailed View, 2020). This algorithm's key philosophy is to produce

a best fit line for predicting the target variable 'y' provided the input variable 'x'. For a single input and target variable, the hypothesis function will be as follows:

$$Y=m(x)+c \text{ where } m= \text{slope}, c= \text{intercept.}$$

The model gets the best fit line of regression by finding the best values for m and c. When the values of m and c are created, the model attempts to predict the value of 'y' to reduce the error discrepancy between the expected values and the real ones. The cost function used is sum of least squares/ordinary least squares.

Where Yi = actual value and other is the predicted value. This method is implemented based on the premises of the following pre model (Linear Regression — Detailed View, 2020).

- The aim variable should be distributed as normal.
- The input variable and the target variable should exhibit linear relationship.
- No multi collinearity (correlation between the input variables) will occur.
- There will be no outliers and the influence of leverage.

Once the assumptions above have been checked, the linear regression method is then implemented. This sets the predictions in motion. Once the predictions are generated, they need to verify post-model assumptions. The assumptions for the post model are as follows (Linear Regression — Detailed View, 2020).

- Errors must follow normal distribution.
- Homoscedasticity-Errors will continuously differ.
- No auto correlation- There should not be any correlation among the errors.

In our dataset, coefficients, residuals, R-square, Adjusted R-square values are created after implementation of the linear regression algorithm. RMSE, Cost is determined further. The least the RMSE, the better the model predictions appreciate. In our problem, **RMSE = 1.92** with **R-square value = 0.830**.

5.2 Performance Metrics

These are the metrics that are used to analyze model performance. The linear regression algorithm 's output metrics are given below.

- R-Square: - This metric is the statistical measure of how close to the fitted regression line the data points are to. This is also known as determination coefficient. There is always between 0 per cent -100 per cent. 0 percent [model] does not clarify the variation in the mean answer results. The variation of the response data around its mean is clarified by 100 percent allocation model.
- Negative r-square: - Model performs lower than the average. The higher the value of the r square, the better the model matches the results. However, the value r square

does not give the full picture. This value decreases as the number of variables input rises irrespective of the model's effect of the variable significance.

- Adjusted R-Square: - This is the updated variant of r square that was changed to represent the number of predictors in the model. This value only increases when the added new input variable improves the model's performance more than would have been expected by chance. That value decreases if the added input variable does not contribute to the performance of the model. The adjusted value r square is always less than the value r square value.

In refined model, the **RMSE value = 0.42** and Adjusted **R-square value =0.826.** This tells us that all the features have some significant impact on the model. In order to find the most significant features of the model, we opted for the partial least square regressor, Regularization concepts.

5.3 Regularization

If we pass to our model some 100 features along with the dataset, there's a lot of chances the model will overfit and become super complex. Regularization, therefore, means that only those functions are useful. It removes automatically features that aren't really helpful along with reducing the error. Regularization applies one additional parameter to the calculation (i.e., penalty). Below are the two types of regularization concepts, which are most frequently implemented (NintyZeros: A Simple Explanation of Regularization in Machine Learning, 2020).

Ridge Regularization

These are the regularization principles that are used in combination with several machine-learning algorithms to enhance the model efficiency. Ridge component reduces the functional impact but does not eliminate the features. We have implemented both Partial least regression with Principal Component Analysis Method and ridge regression to find out the most significant features and to reduce the error component from the model (Ridge Regression: Simple Definition - Statistics How To, 2020).

Partial Least Squares Regression

PLS Regression is a recent technique that generalizes features from Principal Component Analysis and Multiple Regression and combines all that. It is especially beneficial while we have to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). Because of the minimal demands on measurement scales, sample size and residual distributions, Partial Least Squares (PLS) can be a powerful method of analysis. Although PLS can be used to confirm theory, it can also be used to propose where relationships may or may not exist, and to suggest recommendations for subsequent testing (Partial Least Squares Regression and Principal Components Regression- MATLAB & Simulink Example, 2020).

After applying both of these regularizes, we found that Ridge model is better at eliminating features, and that MEDU, STUDYTIME, FAILURES, GOOUT, ABSENCES SEX and SCHOOLSUP are poised.
This report attached the SAS code in the appendix and highly encouraged to look Model development and hyperparameter turning steps and insights drawn.

6 MIS

The main idea is to develop the MIS system to present the whole data insight process from raw data to results including code optimization. Developed menu-based system including radio buttons for navigation like home and next. Radio buttons for the each the step like Preprocessing, Exploratory data analysis, Results and code optimization are developed. This can experience firsthand by selecting the Power point presentation which is attached to this report see attachments at last page of this report. Also attached the screenshots of the same in the appendix for redundancy. (MIS Report: Types, Examples and How to Effectively Prepare it | FineReport, 2020)

7 Code Optimization

Efficient programming is opted in this analysis to compile the code without compromising the performance and results to boost the run time and to decrease the load on the machine. Simple layout is opted like step by step by process including explanation followed by every code chunk contribute to flow of the code. Model 1 is developed by the raw data run time follows for real time is 4:29.51 seconds and the CPU time is 11.75 seconds. Final model decreased the run time follows for real time is 17.00 seconds and the CPU time is 2.67 seconds. This achieved by data wrangling like changing data types by label encoding and Feature engineering techniques like dropping the not significant features inferred from both statistical test and Machine learning models. This can experience firsthand by selecting the Power point presentation which attached to the report by clicking the radio button ‘code optimization’. (Code Optimization - an overview | ScienceDirect Topics, 2020)

8 Conclusion

In order to find out the most significant features and to reduce the error component from the model, we have implemented both Partial least square regression and ridge regression. After applying both these regularizes, we found that ridge model is better at elimination of features poised MEDU, STUDYTIME, FAILURES, GOOUT, ABSENCES SEX and SCHOOLSUP. In other words, Student grades prediction mostly tied with student centric features like how much student time spends on study, going

out, attendance, failures, gender, extra education support and interestingly student mother's education.

9 References

1. FineReport. 2020. MIS Report: Types, Examples And How To Effectively Prepare It | Finereport. [online] Available at: <https://www.finereport.com/en/reporting-tools/mis-report.html> [Accessed 25 May 2020].
2. Mathworks.com. 2020. Partial Least Squares Regression And Principal Components Regression- MATLAB & Simulink Example. [online] Available at: <https://www.mathworks.com/help/stats/examples/partial-least-squares-regression-and-principal-components-regression.html> [Accessed 25 May 2020].
3. Medium. 2020. Linear Regression — Detailed View. [online] Available at: <https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86> [Accessed 25 May 2020].
4. NintyZeros. 2020. Nintyzeros: A Simple Explanation Of Regularization In Machine Learning. [online] Available at: <https://www.nintyzeros.com/2020/03/regularization-machine-learning.html> [Accessed 25 May 2020].
5. Preprocessing, D., 2020. Data Preprocessing. [online] Communities.sas.com. Available at: <https://communities.sas.com/t5/Have-Your-Say/Data-Preprocessing/td-p/300414> [Accessed 25 May 2020].
6. Sciencedirect.com. 2020. Code Optimization - An Overview | Sciencedirect Topics. [online] Available at: <https://www.sciencedirect.com/topics/computer-science/code-optimization> [Accessed 25 May 2020].
7. Statistics How To. 2020. Ridge Regression: Simple Definition - Statistics How To. [online] Available at: <https://www.statisticshowto.com/ridge-regression/> [Accessed 25 May 2020].

10 Appendix

CODE

```
  ./*Importing Data*/
  PROC IMPORT OUT= DERBY.kartheek
              DATAFILE= "C:\Users\mkart\OneDrive - Univer-
sity of Derby\SAS
  \Data\Data\student-mat.csv"
              DBMS=CSV REPLACE;
       GETNAMES=YES;
       DATAROW=2;
  RUN;
  run;
  /*Data Snapshot*/
  proc print data=derby.kartheek;
  run;
  /*Checking Data types*/
  proc contents data=derby.kartheek;
```

```
  run;
  /**/
  PROC FREQ DATA = dataset;
  TABLES variables
  /CHISQ TESTP = (percentage values);

  /*Correlation test */
  /*INSIGHT 1: G1,G2,G3 and failures are correlated*/
  proc corr data=derby.kartheek;
  var AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES FAMREL
FREETIME GOOUT DALC WALC HEALTH ABSENCES G1 G2 G3;
  run;

  /*Scatter Matrix all Numerical variables*/
  proc template;
  define statgraph sgdesign;
  dynamic _AGE _MEDU _FEDU _TRAVELTIME _STUDYTIME _FAILURES
_FAMREL _FREETIME _GOOUT _DALC _WALC _HEALTH _ABSENCES _G1A
_G2A _G3A;
  begingraph / designwidth=811 designheight=804;
     layout lattice;
        scatterplotmatrix _AGE _MEDU _FEDU _TRAVELTIME
_STUDYTIME _FAILURES _FAMREL _FREETIME _GOOUT _DALC _WALC
_HEALTH _ABSENCES _G1A _G2A _G3A / name='scatterplotmatrix'
diagonal=(histogram normal kernel ) markerat-
trs=(color=CXFF0000 );
     endlayout;
  endgraph;
  end;
  run;

  proc sgrender data=DERBY.KARTHEEK template=sgdesign;
  dynamic _AGE="AGE" _MEDU="MEDU" _FEDU="FEDU"
_TRAVELTIME="TRAVELTIME" _STUDYTIME="STUDYTIME"
_FAILURES="FAILURES" _FAMREL="FAMREL" _FREETIME="FREETIME"
_GOOUT="GOOUT" _DALC="DALC" _WALC="WALC" _HEALTH="HEALTH"
_ABSENCES="ABSENCES" _G1A="G1" _G2A="G2" _G3A="G3";
  run;

  /*Checking the  missing values */
  /*INSIGHT 2:No missing values*/
  proc means data=derby.kartheek NMISS N;
  run;
  /*Checking the normality for target variable*/
  proc univariate data=derby.kartheek;
  var G3;
  run;
  /* Checking the Outliers target variable */
```

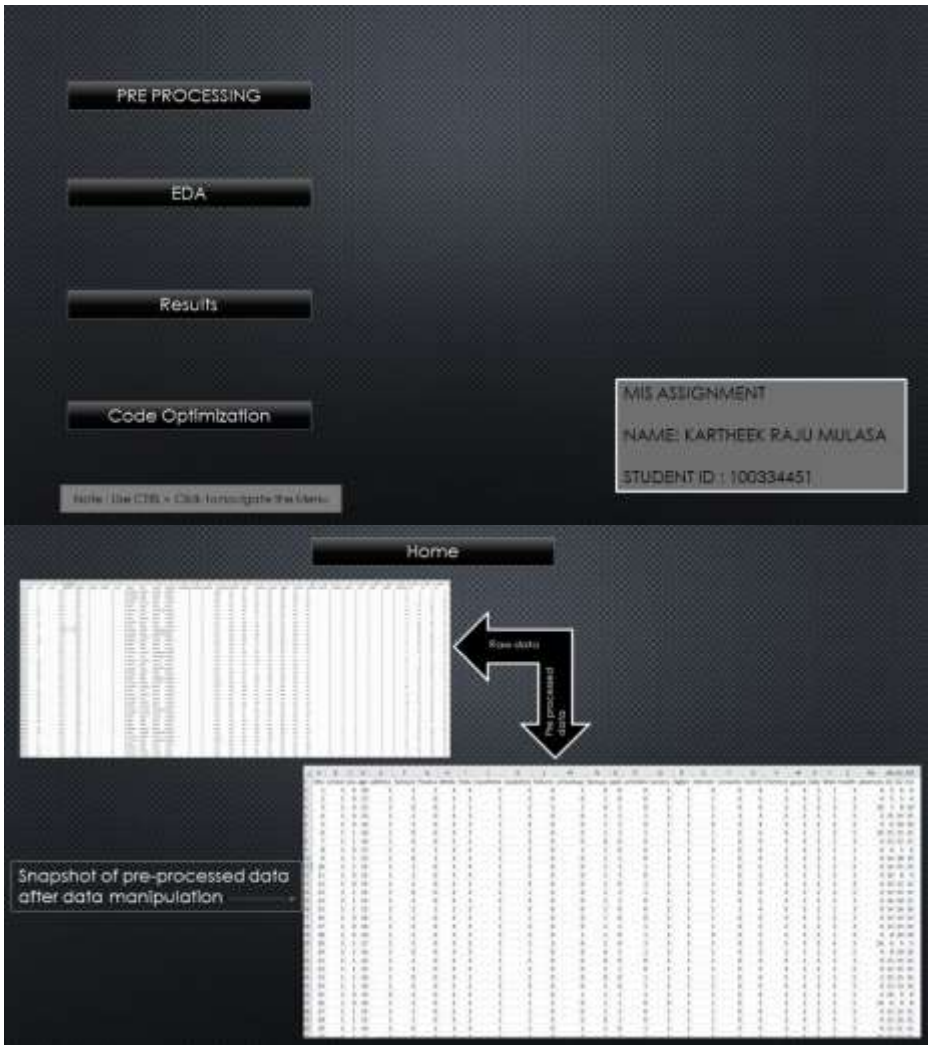
```
  /* INSIGHT 3: No outliers in the target variable. Across
the numerical features very few outliers no need to han-
dling */
  proc template;
  define statgraph sgdesign;
  dynamic _G3A _G3A2;
  begingraph;
     entrytitle halign=center 'G3';
     entryfootnote halign=left 'Type in your footnote...';
     layout lattice / rowdatarange=data columndata-
range=union rows=2 rowgutter=10 columngutter=10 row-
weights=(0.8 preferred);
        layout overlay;
           histogram _G3A / name='histogram'
binaxis=false;
           densityplot _G3A / name='Normal' include-
missinggroup=true normal();
           densityplot _G3A / name='Kernel' include-
missinggroup=true kernel() lineattrs=GraphData2(thick-
ness=2 );
           discretelegend 'Normal' 'Kernel' / opaque=false
border=true halign=right valign=top displayclipped=true
across=1 order=rowmajor location=inside;
        endlayout;
        layout overlay / yaxisopts=( discreteopts=( tick-
valuefitpolicy=none));
           boxplot y=_G3A2 / name='box(h)' groupdis-
play=Cluster orient=horizontal;
        endlayout;
        columnaxes;
           columnaxis / label=('G3 Boxplot');
        endcolumnaxes;
     endlayout;
  endgraph;
  end;
  run;

  proc sgrender data=DERBY.KARTHEEK template=sgdesign;
  dynamic _G3A="G3" _G3A2="G3";
  run;

  /*Final grade visualisation in percentage  */
  proc gchart data=derby.kartheek;
  pie G3/type = percent;
  run;
  /*Final grade visualisation distrubation with gender  */
  proc gchart data=derby.kartheek;
  vbar G3/SUBGROUP =sex
```

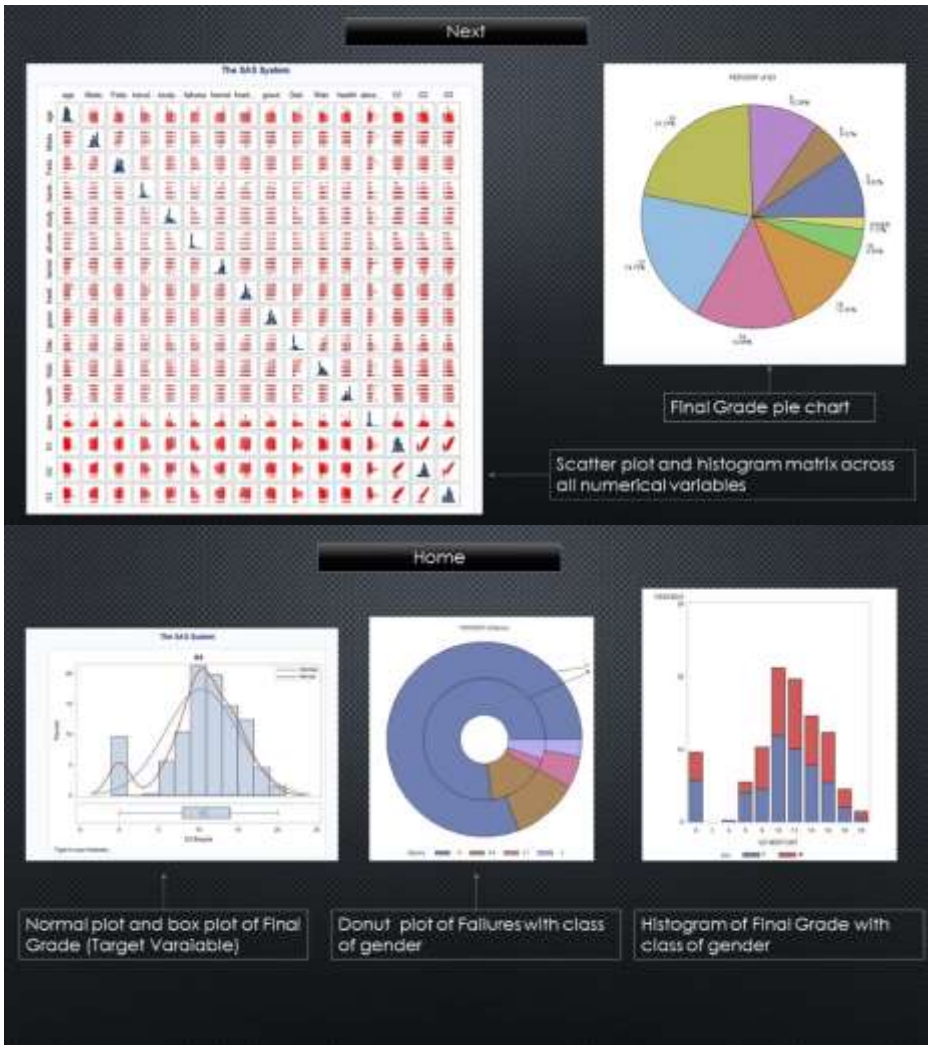
```
  type = percent;
  run;
  /*Final failure visualisation distrubation with gender
*/

  proc gchart data=derby.kartheek;
  donut failures/SUBGROUP =sex
  type = percent;
  run;
  /* Data Manipilation 1 - label  encoding creating the
custom formats for data set */
  PROC FORMAT;
      VALUE $GENDERLABEL
      "M"   = 1
      "F"   = 0;
    VALUE $YESANDNO
      "yes"   = 1
      "no"   = 0;
    VALUE $Pstatus
      "A"   = 1
      "T"   = 0;
    VALUE $address
      "U"   = 1
      "R"   = 0;
    VALUE $school
      "GP"   = 1
      "MS"   = 0;
    VALUE $famsize
      "GT3"   = 1
      "LE3"   = 0;

  RUN;
  /*Checking the new format */
  Proc print data= derby.kartheek;
  format sex $GENDERLABEL. schoolsup $YESANDNO. famsup
$YESANDNO. paid $YESANDNO. activities $YESANDNO. nursery
$YESANDNO.
  higher   $YESANDNO.   internet   $YESANDNO.   romantic
$YESANDNO. address $address. Pstatus $Pstatus. school
$school. famsize $famsize.;
  run;
  /*Created the SAS data set  with label encoding */
  data derby.kartheekraj;
  set derby.kartheekraj;
  run;
  /*Data Standardization */
  proc stdize data=derby.kartheekraj method=Std pstat;
```

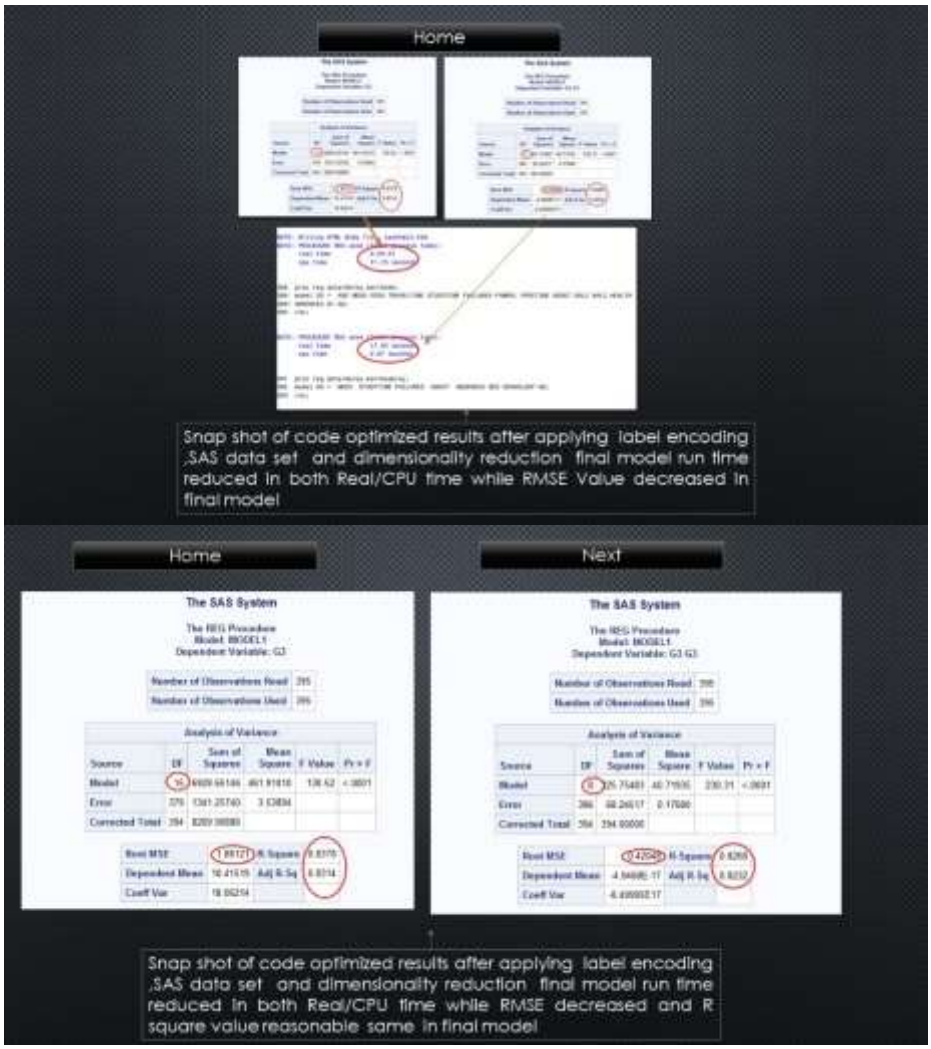
```
    var   AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES
FAMREL FREETIME GOOUT DALC WALC HEALTH ABSENCES G1 G2;
  run;
  proc print data =pstat
  run;
  /*MODEL 1*/
  /*Applying regression model ignoring the insight 1 on
original data set*/
  /*Insight 4: G1 and G2 are highly influence the model
reaching Rsqaure value to 0.8 but we are looking other
features can
  contribute to grade prediction*/
  proc reg data=derby.kartheek;
  model G3 =  AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES
FAMREL FREETIME GOOUT DALC WALC HEALTH ABSENCES G1 G2;
  run;
  /*MODEL 2*/
  /*Applying regression model considering  the insight 1
on orginal data set */
  /*Insight 5 : RMSE values significant increase */
  proc reg data=derby.kartheek;
  model G3 =  AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES
FAMREL FREETIME GOOUT DALC WALC HEALTH ABSENCES ;
  run;
  /*MODEL 3*/
  /*Applying regression model considering the insight 1 on
label encoding data set*/
  /*Insight 6 : RMSE slight changed but RSquare value in-
creased */
  proc reg data=derby.kartheekraj;
  model G3 =  AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES
FAMREL  FREETIME  GOOUT  DALC  WALC  HEALTH  ABSENCES  SEX
SCHOOLSUP FAMSUP PAID NURSERY
  HIGHER INTERNET ROMANTIC ADDRESS PSTATUS SCHOOL FAMSIZE;
  run;
  /*MODEL 4*/
  /*Applying partial least regression with Principal com-
ponent analysis model considering the insight 1 on label
encoding data set*/
  /*This model is to find Dimensionality reduction not for
actual analysis  */
  /*Insight 7: Found few variables can explain the target
variable */
  proc pls data=derby.kartheekraj method=PCR ;
    model  G3  =   AGE  MEDU  FEDU  TRAVELTIME  STUDYTIME
FAILURES FAMREL FREETIME GOOUT DALC WALC HEALTH ABSENCES
SEX SCHOOLSUP FAMSUP PAID NURSERY
```

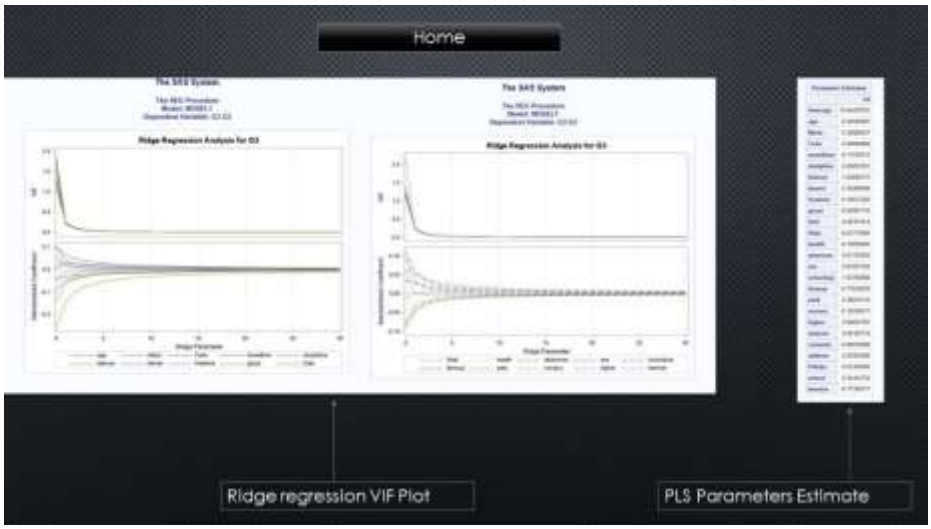
```
  HIGHER INTERNET ROMANTIC ADDRESS PSTATUS SCHOOL FAMSIZE
/ solution;
  run;
  /*MODEL 5*/
  /*Applying ridge regression model considering the insight
1 on label encoding data set*/
  /*This model is to find Dimensionality reduction not for
actual analysis */
  /*Insight 8: Going out is the best feature came out sig-
nificant in VIF    */
  proc reg data=derby.kartheekraj outvif
  outest=PRD_vif ridge=0 to 30 by 1 edf TABLEOUT;
    model G3 =  AGE MEDU FEDU TRAVELTIME STUDYTIME FAILURES
FAMREL  FREETIME  GOOUT  DALC  WALC  HEALTH  ABSENCES  SEX
SCHOOLSUP FAMSUP PAID NURSERY
  HIGHER INTERNET ROMANTIC ADDRESS PSTATUS SCHOOL FAMSIZE;
  run;

  /*MODEL 6*/
  /*Applying regression model considering the insight6 and
7  on label encoding data set*/
  /*Insight 9: RMSE remains around the same value but r
sqaure value slight changed compared to model 3 */
  proc reg data=derby.kartheekraj;
  model G3 =  MEDU  STUDYTIME FAILURES  GOOUT  ABSENCES SEX
SCHOOLSUP;
  run;
  /*MODEL 7 */
  /*Applying regression model considering the insight6 and
7  on label encoding data set*/
  /*Insight 10: RMSE increases and but r square value in-
creased  compared to model 1 */

  proc reg data=derby.kartheekraj;
  model G3 =  MEDU  STUDYTIME FAILURES  GOOUT  ABSENCES SEX
SCHOOLSUP G2;
  run;
```

MIS Screenshots.


PRE PROCESSING
EDA
Results
Code Optimization
MIS ASSIGNMENT
NAME: KARTHEEK RAJU MULASA
STUDENT ID : 100334451
Home
Raw data
Pre processed data
Snapshot of pre-processed data after data manipulation


Next
Final Grade pie chart
Scatter plot and histogram matrix across all numerical variables
Home
Normal plot and box plot of Final Grade (Target Varaiable)
Donut plot of Failures with class of gender
Histogram of Final Grade with class of gender


Home
Snap shot of code optimized results after applying label encoding ,SAS data set and dimensionality reduction final model run time reduced in both Real/CPU time while RMSE Value decreased in final model
Home
Next
The SAS System
The SAS System
Snap shot of code optimized results after applying label encoding ,SAS data set and dimensionality reduction final model run time reduced in both Real/CPU time while RMSE decreased and R square value reasonable same in final model



Attachments.

Raw data.csv

Raw data-

kartheekraj.sas7bdat

SAS data set-

mis.pptx

MIS PPT-