PART 3

10.0 Chapter 10 (Data Visualization and Report Generation)

10.1.1 Data visualization

10.1.2 Introduction

Data visualization is one of the important parts of any statistical data science project where it is important to tell the stories about the dataset and visualize them in charts, bars, and graphs. It is one of the most important skills to have as a data scientist to be able to do data storytelling and SAS offers many kinds of techniques and tools for data visualization. Some of them are using SAS Graphic Procedures where it provides various functions such as PROC SGPLOT, PROC SGPANEL, PROC GCHART that allow the data scientists to create scatter plots, charts, bar graphs and histograms. Another one is ODS graphics where it can create graphics in formats such PDF, HTML and SAS Visual Analytics where it provides web-based tools to visualize data and also provide platform for advanced analytics and predictive modelling.

10.1.3 SAS Source Codes

A screen shot of a computer

Description automatically generated

Figure 145

10.1.4 Screenshot(s) of the Output

A graph showing a few different types of loan applications

Description automatically generated with medium confidence

Figure 146

10.1.5 Description

Figure above shows the simple bar chart to visualize the relationship between the loan application and the loan location. It is found that from the figure above the variables town has the highest number of loan applications followed by city and the least number of applicants is village.

10.1.6 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 147

10.1.7 Screenshot(s) of the Output

A graph showing a couple of squares

Description automatically generated

Figure 148

10.1.8 Description

Figure above shows the simple bar chart that describes the relationship between the variables of marital status against the loan applicants. For the marital status it consisted of married and not married where from the information above the married loan applicant is higher than not married applicants.

10.1.9 SAS Source Codes

A computer code with text

Description automatically generated with medium confidence

Figure 149

10.2 Screenshot(s) of the Output

A graph of a family member

Description automatically generated

Figure 150

10.2.1 Description

Figure above shows the stacked bar chart of the relationship between the number of family members by loan location where for this data visualization, the groups were stacked side by side. Hence as shown above, the loan location has 3 categories which are city, town, and village which indicated by the color blue, red and green. The number of family members are categorized into 0 family members, 1,2 and 3 whereby loan applicants 0 family members have the highest number of applicants compared to 1,2 and 3. Moreover, by looking at figure above in the applicant with 0 family members the loan location town is the highest compared to city and village. For applicant with number of family members 1 has loan location city the highest and for applicant with number of family members 2 has loan location town the highest and for applicant with number of family members 3 has loan location town the highest as well.

10.2.2 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 151

10.2.3 Screenshot(s) of the Output

A graph of a family member

Description automatically generated

Figure 152

10.2.4 Description

Figure above shows the stacked bar chart that shows the relationship between number of family members and gender, and it is known that the gender variable has 2 categories which are the female and male. From the information above the gender female is represented by red color meanwhile for male it is green. The number of family members are categorized into 0 number of family members, 1,2, and 3 where loan applicant with 0 family members is the highest compared to number of family members 1,2, and 3. From the information above, in general male applicant are the highest to apply for loan compared to female.

10.2.5 SAS Source Codes

A computer screen shot of a code

Description automatically generated

Figure 153

10.2.6 Screenshot(s) of the Output

A pie chart with a blue and red center

Description automatically generated

Figure 154

10.2.7 Description

Figure above shows the data visualization using pie chart approach where a pie chart is a representation of values as slices of circle with different colors. The output for this SAS code shows the pie chart above where the title is about loan approval status by location. It is known that Y indicates the number of loan applicants with loan approval status accepted meanwhile for N indicates number of loan applicants with loan approval status not accepted. From pie chart above, loan approval status accepted is 523 applicants which is higher than loan approval status not accepted at 91.

10.2.8 SAS Source Codes

A computer code with text

Description automatically generated

Figure 155

10.2.9 Screenshot(s) of the Output

A red pie chart with a green one

Description automatically generated

Figure 156

A red pie chart with a green one

Description automatically generated

Figure 157

A pie chart with a green and red center

Description automatically generated

Figure 158

10.3 Description

The figure above shows an advanced version of data visualization of pie chart to show the relationship between loan approval status by loan location where the loan locations are divided into city, town, and village. For the loan location city, the number of loan applicants with loan approval status accepted is 169 compared to not accepted at 33. Meanwhile for loan location town, the number of loan applicants with loan approval status accepted is 203 compared to not accepted at 30 and for loan location village, the number of loan applicants with loan approval status accepted is 151 compared to not accepted at 28. It can be deduced that loan approval status accepted is higher than not accepted and most of the accepted loan approval status lives in town.

10.3.1 SAS Source Codes

A screen shot of a computer code

Description automatically generated

Figure 159

10.3.2 Screenshot(s) of the Output

A pie chart with different colored circles

Description automatically generated

Figure 160

10.3.2 Description

The figure above shows an advanced pie chart that shows the relationship between family members vs loan location where it shows much detail explanation of each compositions in percentage of the loan location for town, city, and village. At the same time, it visualizes the number of family members of 0,1,2,and 3 with the color of blue, red, green, and brown representing each of the number of family members. From the pie chart above most of the loan applications have 0 family members which comprised at 58.68% of the applicants, meanwhile for 1 family members at 16.58%, 2 family members at 16.45% and 3 family members at 8.31%.

10.3.3 SAS Source Codes

A screenshot of a computer code

Description automatically generated

Figure 161

10.3.4 Screenshot(s) of the Output

A diagram of a person's population

Description automatically generated

Figure 162

10.3.5 Description

Figure above shows the scatterplot to find out relationship between candidate income and loan amount by grouping both of it using gender where for this data visualization, it is found that most of the candidates income or salary is lower than 20000 for both male and female. Also, most of the loan amount approved by the bank is lower than 200 but another trend can be seen is that there are few outliers in the data points where male has the highest candidate income and loan amount.

10.3.6 SAS Source Codes

A close-up of a computer code

Description automatically generated

Figure 163

10.3.7 Screenshot(s) of the Output

A graph of a person and person

Description automatically generated

Figure 164

10.3.8 Description

Figure above shows another data visualization that shows the box plot of candidate income versus gender whereby looking at the box plot for female all of the candidate income is under 20000. Meanwhile for male candidate income majority of it is under 20000 but there are few outliers where the income is at 40000, 60000, and 80000.

10.3.9 Report Generation

10.4 Physical location of SAS library

10.4.1 SAS Source Codes

A white background with black and blue text

Description automatically generated

Figure 165

Location of the file

10.4.2 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 166

A screenshot of a data

Description automatically generated

Figure 167

10.4.2 Description

The SAS code used is PROC function on the TESTING\_DS whereby it is used to locate the file location in the SAS library. The physical file location is indicated by /home/u61522473/DAP\_PT\_AUG\_2023\_TP067696 which is a cloud based and Libref shows the name of the library which indicated by DAP67696. The other figure shows the other intermediate and temporary dataset created during the assignment where to know it is data is by looking at the Member Type and the file size can be also known and last modified shown when the last time the specific dataset is modified in the SAS studio.

10.4.3 Introduction to ODS

This part here is to discuss the report generation using SAS ODS – Output Delivery System where ODS where in the field of data processing, it refers to software component used in SAS software. ODS or Output Delivery System provide advanced control and the flexibility to generate and manage output from SAS procedures and program.

Using ODS the data scientist is allowed to customize the type, format, and appearance of the SAS output, for example create graphs, chart, tables and generate reports and deliver the output in multiple and various formats. These formats include the PDF,HTML and RTF( Rich Text Format) and main features of SAS ODS are output customization, various output formats, selective output, data manipulation and destination control.

For output customization the features allow the data scientist to modified the layout, style, and the formatting of the output to meet the data scientist preferences. On the other hand, for the various output format shows that ODS allow same analysis done on the report to be delivered in various format at same time. Thus, this allows easy access to information and makes it very easy to share information with other users or data scientists.

For selective output, ODS allow the data scientist or users to choose which parts of the output to be included and excluded meanwhile data manipulation allow the data scientist to manipulate the output results of the SAS programming. Destination control allows the data scientist to specify the output directory whether it’s in a file, email, or printer format.

10.4.4 SAS Source Codes

A screenshot of a computer program

Description automatically generated

Figure 168

10.4.5 Screenshot



Figure 169

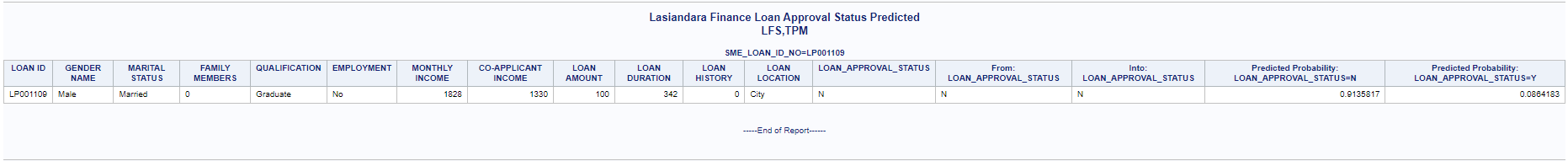


Figure 170

A screenshot of a computer

Description automatically generated

Location of the pdf, LFI\_LAS

Figure 171

A screenshot of a computer

Description automatically generated

Figure 172

10.4.6 Description

Figure above shows the SAS code and the output to generate report using SAS ODS – Output Delivery System where it omits the output Lasiandara Finance Loan Approval Status Predicted. Figure above shows that result of the output Lasiandara Finance have both train and test dataset cleaned which can be indicated no table generated is empty. In the SAS code above both categorical and continuous/numeric variables are used and for the SAS pdf report function of ODS HTML CLOSE; and ODS PDF CLOSE is used. The HTML CLOSE; and ODS PDF CLOSE used is to generate the output of the PDF where the location of the pdf can be seen in figure above and saved as LFI\_LAS.pdf.

10.4.7 SAS Source Codes

A computer screen shot of a message

Description automatically generated

Figure 173

10.4.8 Screenshot(s) of the Output

A screenshot of a computer

Description automatically generated

Figure 174

10.4.9 Description

Figure above shows the SAS code to generate report carrying the loan approval status without using SAS ODS where the data used is DAP67696.TESTING\_PREDICTED\_DS where it is different from the previous one that used SAS ODS.

10.5 SAS Source Codes

A screenshot of a computer

Description automatically generated

Figure 175

10.5.1 Screenshot(s) of the Output

A screenshot of a document

Description automatically generated

Figure 176

A screenshot of a computer

Description automatically generated

Figure 177

A screenshot of a document

Description automatically generated

Figure 178

A screenshot of a data sheet

Description automatically generated

Figure 179

A screenshot of a document

Description automatically generated

Figure 180

A screenshot of a computer

Description automatically generated

Figure 181

10.5.2 Description

Figure above shows the output of SAS code to shows and visualize the dataset used is DAP67696.TESTING\_PREDICTED\_DS and based on the SAS code above the loan location and loan location id are used to generate the report of the variables. The variables shown in the figure above are candidate income, loan amount, loan duration, and loan approval status where i\_loan\_approval\_status is what the machine learning algorithm read, and loan approval status is the variable name defined by the data scientist. From the table above, it can be deduced that the sum for candidate income and loan amount for loan location city is 1090446 and 28770.53, for loan location town its 1233097 and 33907.06 and for loan location village its 994181 and 279219.47. It can be deduced that the sum of candidate income for all 3-loan locations is 3317724 and for loan amount is 89897.06. For the loan approval status can be seen the outcome is either Yes or No indicates by Y and N for the loan applicants.

10.5.3 Discussion

Hence, it is known by the data scientists beforehand that the train and test dataset which is TRAINING.DS and TESTING.DS. It is known that from the SAS library for TRAINING DS has 614 total rows and 13 total number of columns. Meanwhile for TESTING DS it has 367 total rows and 13 total columns. Both of the dataset has the same variables which are the SME Loan ID, Gender, Marital Status, Family Member, Qualification, Employment, Candidate Income, Guarantee Income, Loan Amount, Loan History, LOAN Duration, Loan Location and Loan Approval Status.

As a data scientist it is important to study all of these variables as these variables influencing the decision making of the data scientist to approve the loans to the applicant. For this analysis, the dependent variable is the loan approval status meanwhile the rest of the variables are the independent variables. The loan approval status of the loan applicants will have outcome of either Yes or No depending on the independent variables.

For the model creation, logistic regression is used to predict the outcome of the loan approval status which is the response variable. It is known that the dependent variable is loan approval status is a categorical variable and the independent variable is the mixture of categorical and continuous/numeric variables. The data is cleansed when the number of observations read is equal to number of observations used, hence the model convergence status is accepted. Model convergence status is defined as convergence of the machine learning model during the training of the data, and which indicates it reach satisfactory state. The output for model convergence status used precision of as it is the most common value convergence threshold in iterative algorithm optimization.

As a data scientist, it is important to know what the definition of Akaike Information Criterion (AIC) and Schwarz Criterion (SC) is where both of it are the statistical techniques used in model selection especially in linear regression machine learning model. AIC is used in model selection particularly in maximum likelihood estimation where , where k is the number of parameters in the model and ln(L) is the natural algorithm of the maximum likelihood of model with an input of data. On the other hand, Schwarz Criterion (SC) is the same as AIC but put more penalty on model with a greater number of parameters. It is defined as , where there is addition to previous criterion which n, the sample size. In general AIC and SC value must be lower which indicating the model is better fit which in this case the value of SC is 769.311 which is higher than AIC at 764.891

There is another criterion which is the -2logL where it is under the model fit statistics the term “-2logL” refers to one of the components used in model selection where L representing the likelihood of the data given in the model and -2logL indicates the measure of model goodness of fit. On the other hand, in the model fit statistics, there is intercept and covariates values of AIC and SC respectively at 589.101 and 659.821 where both of intercept and covariates play important role in calculation of the criterions to help determine the best fitting models.

Intercept usually often included in linear regression model where it shows the value of dependent variable which in this case loan approval status meanwhile all of the values of the independent variables (covariates) are equal to zero. On the other hand, the covariates are the independent variables in which it causes and effect on the dependent variables. In general, the relationship between covariates and dependent variables are quantified using regression coefficients.

For this analysis, there are 3 variables which act as the most contributing factors which are the loan location, loan history, and marital status. This is due to the value of Pr>Chisq for these variables is less than 0.05 compared to other independent variables. The term “Pr>Chisq” represents for the probability greater than chi-square, where it’s a p-value that is related to chi-square statistic in statistical analysis. A chi-square test indicates that whether there is a significant relationship between two categorical variables and in general Pr>Chisq is used to test and see is there any statistical significance between the variables.

If the P value is less than 0.05 in which in this case for loan location, loan history, and marital status where each of these p values are lesser than 0.05. Hence the data scientist should conclude that these variables are the contributing factors to approve the loan status whether Yes or No.

11.0 Chapter 11 (Conclusion)

It can be deduced that for this assignment, there are 3 parts which are divided into part 1, part 2, and part 3. For part 1, it comprises of chapter 1: introduction, chapter 2: problem statement, chapter 3: background of Lasiandra Finance, chapter 4: assumption, program demonstration, coding and justification, chapter 5: methodology and chapter 6: data dictionary/metadata. On the other hand, for part 2, it comprises of chapter 7: literature review, chapter 8: data analysis/data cleansing and chapter 9: model creation and prediction. The last section is part 3 where it consists of chapter 10: data visualization and report generation.

In general, this assignment has lot of knowledge that can be gained especially, understanding data management system using SQL programming in SAS studio. As a data scientist, a lot of knowledge and skills can be gained during this assignment where some of the outcomes gained are assess and study variety type and forms of datasets by reading, combined and categorizing the datasets used which is the training and testing datasets using data analytical programming method. Other outcomes from this assignment are that it taught the data scientist to produce analytical data models by creating reports and enhanced listings. It also gives new skills for the data scientist to learn about data visualization using SAS software.

A lot of knowledge had been taught by the course instructor Dr Dhason Padmakumar where for the past 2 months it has been quite a journey learning about data analytical programming. The course is very crucial to data scientists as it teaches one of the pillars of programming language that need to be mastered by data scientists in their profession. The domain of the datasets used is for banking and finance industry but after learning the skills of understanding the sql coding it can be applied to other domain of datasets in the field of engineering, healthcare, insurance and more. Dr Dhason lay out the structures of teaching the fundamentals of the sql programming techniques first before going through the assignment together which has been great way to understand about this assignment.

12.0 REFERENCES

Stiglitz, J.E. and A. Weiss (1988), ‘‘Banks as Social Accountants and Screening Devices for the Allocation of Credit’’, Working Paper No. 2710, Washington: National Bureau of Economic Research.

Coase, R.H. (1937), ‘‘The Nature of the Firm’’, *Economica*, **4**, 386–405.

Alchian, A. and H. Demsetz (1972), ‘‘Production, Information Costs and Economic Organisation’’, *American Economic Review*, **62**, 777–795.

Williamson, O. (1981), ‘‘The Modern Corporation: Origins, Evolution, Attributes’’, Journal of Economic Literature, 19, 1537–1568.

Briault, C. (2000), Draft Paper, ‘‘FSA Revisited and Some Issues for European Securities Markets Regulation’’, London: Financial Services Authority, December.

Beccalli, E., Casu, B., Girardone, C., 2005. Efficiency and Stock Performance in European Banking. Journal of Business, Accounting and Finance (forthcoming)

Fernández, P., 2002. EVA, Economic Profit and Cash Value Added do not measure shareholder value creation. University of Navarra, IESE, Research paper no. 453

Eisenbeis, R.A., Ferrier, G.D., Kwan, S.H., 1999. The Informativeness of Stochastic Frontier and Programming Frontier Efficiency Scores: Cost Efficiency and Other Measures of Bank Holding Company Performance. Federal Reserve Bank of Atlanta, Working Paper no. 99-23.

Chu, S.F., Lim, G.H., 1998. Share Performance and Profit Efficiency of Banks in an Oligopolistic Market: Evidence from Singapore. Journal of Multinational Financial Management 8, 155--168.

Jacobson, T., Lindé, J., Roszbach, K., 2006. Credit risk versus capital requirements under Basel II: are SME loans and retail credit really different?. Journal of Financial Services Research, forthcoming

Duffie D., 2005. Credit risk modeling with affine processes. Journal of Banking & Finance, 29, 11, pp. 2751-2802

Lucas, A., Klaassen, P., 2006. Discrete versus continuous state switching models for portfolio credit risk. Journal of Banking & Finance, 30, 1, pp. 23-35

Galluccio, S., Roncoroni, A., (2006). A new measure of cross-sectional risk and its empirical implications for portfolio risk management, Journal of Banking & Finance, (forthcoming)

Zheng H., 2006. Interaction of credit and liquidity risks: Modelling and valuation. Journal of Banking & Finance, 30, 2, pp. 391-407

Jobst, N.J., Mitra G., Zenios S.A., 2006. Integrating market and credit risk: A simulation and optimisation perspective. Journal of Banking & Finance, 30, 2, pp. 717-742

Scandizzo, S., 2005. Risk Mapping and Key Risk Indicators in Operational Risk Management. Economic-Notes, 34, 2, pp. 231-56

De Fontnouvelle, P., Jordan, J., Rosengren, E., 2005. Implications of Alternative Operational Risk Modeling Techniques. National Bureau of Economic Research, Inc, NBER, Working Papers: no. 11103

Barth, M.E., Beaver, W.H., Landsman, W., 1998. Relative valuation roles of equity book value and net income as a function of financial health. Journal of Accounting and Economics 25, 1--34.

Athanasoglou, P.P., Brissimis, S. N., Delis, M. (2008). Bank specific, industry specific and macroeconomic determinants of bank profitability. Journal of international financial markets, institutions and money, 18(2), 121-136.

Kumar, Rajiv, et al. (2019). Prediction of loan approval using machine learning. *International Journal of Advanced Science and Technology, 28(7), 455*-460.