

**Individual Assignment**

**TECHNOLOGY PARK MALAYSIA**

**MODULE CODE-DAP**

**Data Analytical Programming**

**APUMP2205-DSBA(PR)**

**Student’s TP: TP067696**

**Student’s Name: Mr.Muhammad Arif Bin Jamaluddin**

**Lecturer’s Name: Mr.DHASON PADMAKUMAR**

ACKNOWLEDGMENT

On this acknowledgement part, I want to take this opportunity to express my gratitude and very special thanks to my lecturer Mr. Dhason Padmakumar who guided me and my classmates for this subject (DAP). His years and years of experience of teaching this subject really showed when he was conducting the online class as the pace of the class was really good. Moreover, I want to thank my classmates for their participation in the classroom.

In addition, I would like to express my special thanks to my family, especially my mother who has been supporting me throughout my master program. It has been quite a roller coaster journey for me, and I am glad that I am able to pull it off.

Also, I want to express my gratitude towards the staff of APU for helping me with the students matters and without them student life would not be easy. In conclusion, I want to thank God for helping me getting through all the challenges that I faced throughout this study.

TABLE OF CONTENT

Contents

[ACKNOWLEDGMENT 2](#_Toc148303788)

[TABLE OF CONTENT 3](#_Toc148303789)

[1.0 Chapter 1 (Introduction) 12](#_Toc148303790)

[2.0 Chapter 2 (Problem Statement) 14](#_Toc148303791)

[3.0 Chapter 3 (Background Of Lasiandra Finance Inc. (LFI), New York, USA) 16](#_Toc148303792)

[4.0 Chapter 4 (ASSUMPTION, PROGRAM DEMONSTRATION, CODING AND JUSTIFICATION). 18](#_Toc148303793)

[5.0 Chapter 5 (Methodology) 20](#_Toc148303794)

[6.0 Chapter 6 ( Data Dictionary/ Metadata ) 22](#_Toc148303795)

[6.1 Introduction 22](#_Toc148303796)

[6.1 Location of the datasets on SAS 23](#_Toc148303797)

[6.2 Project datasets found inside the SAS permanent library/folder. 23](#_Toc148303798)

[6.2.1 Description 23](#_Toc148303799)

[6.2.2 Display the structure of the datasets-DAP67696.TRAINING\_DS 24](#_Toc148303800)

[6.2.3 Data dictionary of the dataset- DAP67696.TRAINING\_DS 24](#_Toc148303801)

[6.2.4 SAS PROC SQL Codes 24](#_Toc148303802)

[6.2.5 Screenshot(s) of the output 24](#_Toc148303803)

[6.2.6 Description 25](#_Toc148303804)

[6.2.7 SAS PROC SQL Codes 25](#_Toc148303805)

[6.2.8 Screenshot(s) of the output 26](#_Toc148303806)

[6.2.9 Description 27](#_Toc148303807)

[7.0 Chapter 7 ( Literature Review ) 28](#_Toc148303808)

[7.1 Modern banking system 28](#_Toc148303809)

[7.2 Determinant of profit and shareholder value creation in banking 30](#_Toc148303810)

[7.3 Factors that influence bank profitability 31](#_Toc148303811)

[7.4 Bank loan collection from its customers 32](#_Toc148303812)

[7.5 Machine learning in loan process application 34](#_Toc148303813)

[7.6 Outcome of the research 34](#_Toc148303814)

[8.0 Chapter 8 (Data Analysis and Data Cleansing) 35](#_Toc148303815)

[8.1 Analysis of the Categorical variables found in DAP67696.TRAINING\_DS 35](#_Toc148303816)

[8.1.1 Univariate Analysis of the Categorical variable 35](#_Toc148303817)

[8.1.2 Univariate Analysis of the Categorical variable - GENDER 35](#_Toc148303818)

[8.1.3 SAS Source Codes 36](#_Toc148303819)

[8.1.4 Screenshot(s) of the Output 36](#_Toc148303820)

[8.1.5 Description 36](#_Toc148303822)

[8.1.6 Univariate Analysis of the Categorical variable – MARITAL\_STATUS 37](#_Toc148303823)

[8.1.7 SAS Source Codes 37](#_Toc148303824)

[8.1.8 Screenshot(s) of the Output 37](#_Toc148303825)

[8.1.9 Description 37](#_Toc148303827)

[8.2 Univariate Analysis of the Categorical variable – LOAN\_LOCATION 38](#_Toc148303828)

[8.2.1 SAS Source Codes 38](#_Toc148303829)

[8.2.2 Screenshot(s) of the Output 38](#_Toc148303830)

[8.2.3 Description 38](#_Toc148303832)

[8.2.4 Univariate Analysis of the Continuous/Numeric variable 39](#_Toc148303833)

[8.2.5 Univariate Analysis of the Continuous/Numeric variable – GUARANTEE\_INCOME](#_Toc148303834)

[8.2.6 SAS Source Codes 39](#_Toc148303835)

[8.2.7 Screenshot(s) of the Output 40](#_Toc148303836)

[8.2.8 Description 41](#_Toc148303838)

[8.2.9 Univariate Analysis of the Continuous/Numeric variable – LOAN\_DURATION 42](#_Toc148303839)

[8.3 SAS Source Codes 42](#_Toc148303840)

[8.3.1 Screenshot(s) of the Output 43](#_Toc148303842)

[8.3.2 Description 43](#_Toc148303844)

[8.3.3 Univariate Analysis of the Continuous/Numeric variable – LOAN\_AMOUNT 44](#_Toc148303845)

[8.3.4 SAS Source Codes 44](#_Toc148303846)

[8.3.5 Screenshot(s) of the Output 45](#_Toc148303848)

[8.3.6 Description 45](#_Toc148303850)

[8.3.7 Bivariate analysis of the variables found in DAP67696.TRAINING\_DS 46](#_Toc148303851)

[8.3.8 Introduction 46](#_Toc148303852)

[8.3.9 Bivariate Analysis of the variables (categorical vs categorical) 46](#_Toc148303853)

[8.4. Bivariate Analysis of the variable – (GENDER vs MARITAL\_STATUS ); (Categorical vs categorical variable) 46](#_Toc148303854)

[8.4.1 SAS Source Codes 46](#_Toc148303855)

[8.4.2 Screenshot(s) of the Output 47](#_Toc148303856)

[8.4.3 Description 48](#_Toc148303859)

[8.4.5 SAS Source Codes 49](#_Toc148303860)

[8.4.6 Screenshot(s) of the Output 49](#_Toc148303861)

[8.4.7 Description 50](#_Toc148303863)

[8.4.8 Bivariate Analysis of the variable – (LOAN\_LOCATION vs LOAN\_HISTORY); (Categorical vs categorical variable) 51](#_Toc148303864)

[8.4.9 SAS Source Codes 51](#_Toc148303865)

[8.5 Screenshot(s) of the Output 51](#_Toc148303867)

[8.5.1 Description 52](#_Toc148303870)

[8.5.2 Bivariate Analysis of the variables (categorical vs continuous) 53](#_Toc148303871)

[8.5.3 Bivariate Analysis of the variable – (GENDER vs LOAN\_DURATION); (Categorical vs continuous variable) 53](#_Toc148303872)

[8.5.4 SAS Source Codes 53](#_Toc148303873)

[8.5.5 Screenshot(s) of the Output 53](#_Toc148303875)

[8.5.6 Description 54](#_Toc148303877)

[8.5.7 Bivariate Analysis of the variable – (LOAN\_HISTORY vs CANDIDATE\_INCOME); (Categorical vs continuous variable) 54](#_Toc148303878)

[8.5.8 SAS Source Codes 55](#_Toc148303879)

[8.5.9 Screenshot(s) of the Output 55](#_Toc148303881)

[8.6 Description 55](#_Toc148303882)

[8.6.1 Bivariate Analysis of the variable – (QUALIFICATION vs LOAN\_AMOUNT); (Categorical vs continuous variable) 56](#_Toc148303883)

[8.6.2 SAS Source Codes 56](#_Toc148303884)

[8.6.2 Screenshot(s) of the Output 56](#_Toc148303886)

[8.6.3 Description 57](#_Toc148303887)

[8.6.4 Analysis of the Categorical variables found in DAP67696.TESTING\_DS 57](#_Toc148303888)

[8.6.5 Univariate Analysis of the Categorical variable using SAS MACRO 57](#_Toc148303889)

[8.6.6 Introduction 57](#_Toc148303890)

[8.6.7 SAS Source Codes 58](#_Toc148303891)

[8.6.8 Screenshot(s) of the Output 59](#_Toc148303895)

[8.6.9 Description 59](#_Toc148303899)

[8.7 Univariate Analysis of the Continuous/Numeric variable – using the SAS MACRO 60](#_Toc148303900)

[8.7.1 SAS Source Codes 60](#_Toc148303901)

[8.7.2 Screenshot(s) of the Output 62](#_Toc148303905)

[8.7.3 Description 65](#_Toc148303913)

[8.7.4 Bivariate Analysis of the variables – (Categorical vs categorical variable) using SAS Macro 66](#_Toc148303914)

[8.7.5 SAS Source Codes 66](#_Toc148303915)

[8.7.6 Screenshot(s) of the Output 67](#_Toc148303919)

[8.7.7 Description 70](#_Toc148303926)

[8.7.8 Bivariate Analysis of the variables – (Categorical vs continuous variable) using SAS Macro 71](#_Toc148303927)

[8.7.9 SAS Source Codes 71](#_Toc148303928)

[8.8 Screenshot(s) of the Output 72](#_Toc148303932)

[8.8.1 Description 73](#_Toc148303936)

[8.8.2 Data Cleaning 74](#_Toc148303937)

[8.8.3 Imputing the missing values found in the categorical variables in the datasets DAP67696.TRAINING\_DS 74](#_Toc148303938)

[8.8.4 Imputing the missing values found in the categorical variables - GENDER. 74](#_Toc148303939)

[8.8.5 SAS Source Codes 74](#_Toc148303940)

[8.8.6 Screenshot(s) of the Output 75](#_Toc148303942)

[8.8.7 Description 75](#_Toc148303944)

[8.8.8 SAS Source Codes 75](#_Toc148303945)

[8.8.9 Screenshot(s) of the Output 76](#_Toc148303947)

[8.9 Description 76](#_Toc148303949)

[8.9.1 SAS Source Codes 76](#_Toc148303950)

[8.9.2 Screenshot(s) of the Output 77](#_Toc148303952)

[8.9.3 Description 77](#_Toc148303954)

[8.9.4 SAS Source Codes 77](#_Toc148303955)

[8.9.5 Screenshot of the Output 78](#_Toc148303957)

[8.9.6 Description 78](#_Toc148303959)

[8.9.7 SAS Source Codes 78](#_Toc148303960)

[8.9.8 Screenshot(s) of the Output 79](#_Toc148303962)

[8.9.9 Description 79](#_Toc148303964)

[8.0.1.1 SAS Source Codes 80](#_Toc148303965)

[8.0.1.2 Screenshot(s) of the Output 80](#_Toc148303967)

[8.0.1.3 Description 80](#_Toc148303969)

[8.0.1.4 SAS Source Codes 80](#_Toc148303970)

[8.0.1.5 Screenshot(s) of the Output 81](#_Toc148303972)

[8.0.1.6 Description 81](#_Toc148303974)

[8.0.1.7 Imputing the missing values found in the categorical variables – MARITAL\_STATUS. 81](#_Toc148303975)

[8.0.1.8 SAS Source Codes 81](#_Toc148303976)

[8.0.1.9 Screenshot(s) of the Output 82](#_Toc148303978)

[8.0.2 Description 82](#_Toc148303980)

[8.0.2.1 SAS Source Codes 82](#_Toc148303981)

[8.0.2.2 Screenshot(s) of the Output 83](#_Toc148303983)

[8.0.2.3 Description 83](#_Toc148303985)

[8.0.2.4 SAS Source Codes 83](#_Toc148303986)

[8.0.2.5 Screenshot(s) of the Output 84](#_Toc148303988)

[8.0.2.6 Description 84](#_Toc148303990)

[8.0.2.7 SAS Source Codes 84](#_Toc148303991)

[8.0.2.8 Screenshot(s) of the Output 84](#_Toc148303993)

[8.0.2.9 Description 84](#_Toc148303995)

[8.0.3 Imputing the missing values found in the categorical variables – FAMILY\_MEMBERS. 85](#_Toc148303996)

[8.0.3.1 SAS Source Codes 85](#_Toc148303997)

[8.0.3.2 Screenshot(s) of the Output 85](#_Toc148303999)

[8.0.3.3 Description 85](#_Toc148304001)

[8.0.3.4 SAS Source Codes 86](#_Toc148304002)

[8.0.3.5 Screenshot(s) of the Output 86](#_Toc148304004)

[8.0.3.6 Description 86](#_Toc148304006)

[8.0.3.7 SAS Source Codes 87](#_Toc148304007)

[8.0.3.8 Screenshot(s) of the Output 87](#_Toc148304009)

[8.0.3.9 Description 88](#_Toc148304011)

[8.0.4 SAS Source Codes 88](#_Toc148304012)

[8.0.4.1 Screenshot(s) of the Output 88](#_Toc148304014)

[8.0.4.2 Description 88](#_Toc148304016)

[8.0.4.3 SAS Source Codes 89](#_Toc148304017)

[8.0.4.4 Screenshot(s) of the Output 89](#_Toc148304019)

[8.0.4.5 Description 89](#_Toc148304021)

[8.0.4.6 SAS Source Codes 90](#_Toc148304022)

[8.0.4.7 Screenshot(s) of the Output 90](#_Toc148304024)

[8.0.4.8 Description 90](#_Toc148304026)

[8.0.4.9 SAS Source Codes 90](#_Toc148304027)

[8.0.5 Screenshot(s) of the Output 91](#_Toc148304029)

[8.0.5.1 Description 91](#_Toc148304031)

[8.0.5.2 Imputing the missing values found in the continuous variables – LOAN\_AMOUNT.](#_Toc148304032)

[8.0.5.3 SAS Source Codes 92](#_Toc148304033)

[8.0.5.4 Screenshot(s) of the Output 92](#_Toc148304035)

[8.0.5.5 Description 92](#_Toc148304037)

[8.0.5.6 SAS Source Codes 93](#_Toc148304038)

[8.0.5.7 Screenshot(s) of the Output 93](#_Toc148304040)

[8.0.5.8 Description 93](#_Toc148304042)

[8.0.5.9 SAS Source Codes 94](#_Toc148304043)

[8.0.6 Screenshot(s) of the Output 94](#_Toc148304045)

[8.0.7 Description 94](#_Toc148304047)

[8.0.7.1 SAS Source Codes 95](#_Toc148304048)

[8.0.7.2 Screenshot(s) of the Output 95](#_Toc148304050)

[8.0.7.3 Description 95](#_Toc148304052)

[8.0.7.4 Imputing the missing values found in the continuous variables – LOAN\_DURATION. 95](#_Toc148304053)

[8.0.7.5 SAS Source Codes 96](#_Toc148304054)

[8.0.7.6 Screenshot(s) of the Output 96](#_Toc148304056)

[8.0.7.7 Description 96](#_Toc148304058)

[8.0.7.8 SAS Source Codes 97](#_Toc148304059)

[8.0.7.9 Screenshot(s) of the Output 97](#_Toc148304061)

[8.0.8 Description 97](#_Toc148304063)

[8.0.8.1 SAS Source Codes 98](#_Toc148304064)

[8.0.8.2 Screenshot(s) of the Output 98](#_Toc148304066)

[8.0.8.3 Description 98](#_Toc148304068)

[8.0.8.4 SAS Source Codes 99](#_Toc148304069)

[8.0.8.5 Screenshot(s) of the Output 99](#_Toc148304071)

[8.0.8.6 Description 99](#_Toc148304073)

[8.0.8.7 Imputing the missing values found in the continuous variables – GUARANTEE\_INCOME. 99](#_Toc148304074)

[8.0.8.8 SAS Source Codes 100](#_Toc148304075)

[8.0.8.9 Screenshot(s) of the Output 100](#_Toc148304077)

[8.0.9.1 Description 100](#_Toc148304079)

[8.0.9.2 SAS Source Codes 101](#_Toc148304080)

[8.0.9.3 Screenshot(s) of the Output 101](#_Toc148304082)

[8.0.9.4 Description 101](#_Toc148304084)

[8.0.9.5 SAS Source Codes 102](#_Toc148304085)

[8.0.9.6 Screenshot(s) of the Output 102](#_Toc148304087)

[8.0.9.7 Description 102](#_Toc148304089)

[8.0.9.8 SAS Source Codes 103](#_Toc148304090)

[8.0.9.9 Screenshot(s) of the Output 103](#_Toc148304092)

[8.0.0.1 Description 103](#_Toc148304094)

[9.0 Chapter 9 ( Model Creation and Prediction) 104](#_Toc148304095)

[9.1.1 SAS Source Codes 104](#_Toc148304096)

[9.1.2 Screenshot(s) of the Output 105](#_Toc148304098)

[9.1.3 Description 105](#_Toc148304101)

[9.1.4 Screenshot(s) of the Output 106](#_Toc148304102)

[9.1.5 Description 106](#_Toc148304105)

[9.1.6 Screenshot(s) of the Output 106](#_Toc148304106)

[9.1.7 Description 106](#_Toc148304108)

[9.1.8 Screenshot(s) of the Output 107](#_Toc148304109)

[9.1.9 Description 107](#_Toc148304111)

[9.2 Screenshot(s) of the Output 108](#_Toc148304112)

[9.2.1 Description 108](#_Toc148304114)

[9.2.2 SAS Source Codes\ 108](#_Toc148304115)

[9.2.3 Screenshot(s) of the Output 109](#_Toc148304117)

[9.2.4 Description 109](#_Toc148304119)

[9.2.5 List the details of the dataset carrying the loan approval status predicted- DAP67696.TESTING\_PREDICTED\_DS 109](#_Toc148304120)

[9.2.6 SAS Source Codes 109](#_Toc148304121)

[9.2.7 Screenshot(s) of the Output 110](#_Toc148304123)

[9.2.8 Description 110](#_Toc148304125)

[10.0 Chapter 10 (Data Visualization and Report Generation) 111](#_Toc148304126)

[10.1.1 Data visualization 111](#_Toc148304127)

[10.1.2 Introduction 111](#_Toc148304128)

[10.1.3 SAS Source Codes 111](#_Toc148304129)

[10.1.4 Screenshot(s) of the Output 112](#_Toc148304131)

[10.1.5 Description 112](#_Toc148304133)

[10.1.6 SAS Source Codes 112](#_Toc148304134)

[10.1.7 Screenshot(s) of the Output 113](#_Toc148304136)

[10.1.8 Description 113](#_Toc148304138)

[10.1.9 SAS Source Codes 114](#_Toc148304139)

[10.2 Screenshot(s) of the Output 114](#_Toc148304141)

[10.2.1 Description 114](#_Toc148304143)

[10.2.2 SAS Source Codes 115](#_Toc148304144)

[10.2.3 Screenshot(s) of the Output 115](#_Toc148304146)

[10.2.4 Description 116](#_Toc148304148)

[10.2.5 SAS Source Codes 116](#_Toc148304149)

[10.2.6 Screenshot(s) of the Output 117](#_Toc148304151)

[10.2.7 Description 117](#_Toc148304153)

[10.2.8 SAS Source Codes 118](#_Toc148304154)

[10.2.9 Screenshot(s) of the Output 118](#_Toc148304156)

[10.3 Description 119](#_Toc148304160)

[10.3.1 SAS Source Codes 120](#_Toc148304161)

[10.3.2 Screenshot(s) of the Output 120](#_Toc148304163)

[10.3.2 Description 121](#_Toc148304165)

[10.3.3 SAS Source Codes 121](#_Toc148304166)

[10.3.4 Screenshot(s) of the Output 121](#_Toc148304168)

[10.3.5 Description 122](#_Toc148304170)

[10.3.6 SAS Source Codes 122](#_Toc148304171)

[10.3.7 Screenshot(s) of the Output 122](#_Toc148304173)

[10.3.8 Description 123](#_Toc148304175)

[10.3.9 Report Generation 123](#_Toc148304176)

[10.4 Physical location of SAS library 123](#_Toc148304177)

[10.4.1 SAS Source Codes 123](#_Toc148304178)

[10.4.2 Screenshot(s) of the Output 123](#_Toc148304181)

[10.4.2 Description 124](#_Toc148304183)

[10.4.3 Introduction to ODS 125](#_Toc148304184)

[10.4.4 SAS Source Codes 126](#_Toc148304185)

[10.4.5 Screenshot 126](#_Toc148304187)

[10.4.6 Description 127](#_Toc148304192)

[10.4.7 SAS Source Codes 128](#_Toc148304193)

[10.4.8 Screenshot(s) of the Output 128](#_Toc148304195)

[10.4.9 Description 128](#_Toc148304197)

[10.5 SAS Source Codes 129](#_Toc148304198)

[10.5.1 Screenshot(s) of the Output 130](#_Toc148304200)

[10.5.2 Description 134](#_Toc148304207)

[10.5.3 Discussion 135](#_Toc148304208)

[11.0 Chapter 11 (Conclusion) 137](#_Toc148304209)

[12.0 REFERENCES 138](#_Toc148304210)

Part 1

1.0 Chapter 1 (Introduction)

Banks have been the bedrock of modern economies in the modern world right now where it plays crucial role in financial infrastructure of many countries around the globe. Banks act as an intermediary between the depositors (institution who loan money to the bank) and borrowers (to whom the bank lends money). A banking system is a group or network of institutions that provide financial services and typically banking institutions operate a payment system, provide loans, take deposits, and help people, individuals, or companies with investments. There is a certain amount of money banks pay for deposits and interest is often the income the bank receives when giving out loans. So, there are two key people here, the depositors and the borrowers where depositors can be the common man such as individuals, financial and non-financial firms, and local governments.

For example, payments and loans from typical commercial banks allow people to use deposit funds and use checking and debit cards to pay bills or make any purchases. So, what is a bank? A bank like many others is a business but unlike any other business, bank don’t manufacture a physical product like factory or industry but instead bank provide services. These services such as business loans which is related to this assignment, Lasiandra Finance Inc. (LFI), that provide loan to small startups. Other services bank provided are car loans, home mortgages loan, credit card services and retirement accounts.

On the other hand, central banks such the federal reserve bank in USA and in Malaysia (Bank Negara Malaysia) main role is to distribute currency and establish money related policies. Another type and form of bank is investment bank where it conducts trades or deals with capital markets and most banks are profit-seeking entities where the major goal is to protect the interest of the shareholders. Bank make profit by charging more interest in loans and paying less interest on deposits. For example, a typical interest rate for a house mortgage with a 30-year loan would be around 3.29%. There are several key takeaways such as both banking systems operate differently and have different target clientele. Private banks focus on providing the best services and caters to wealthy individuals, companies, and corporations where often the goal is to generate profit for shareholders.

The public banking system’s main goal is to adhere to societal role, provide apprehensive and accessible financial services to the masses, contribute to economic growth and development as instructed by the government. So, in conclusion, the way banks and the banking industry work is that it manages and operates the flow or exchange of money between business and people where it offers deposit account where people can store their money in a secured place. This money is used by the banks to provide loans to businesses. In return the bank makes a profit from the interest payments they receive on those loans from the borrowers.

2.0 Chapter 2 (Problem Statement)

As a data scientist at this company in the Headquarters of LFI, Washington, D.C., United, it is required by the company to analyse the datasets obtained from the past customers and build an accurate model or in other word model building to predict the approval process of loan applications either it is being approved or being rejected.

SAMPLE

EXPLORE

MODIFY

Workflow of model proposed.

MODEL

Above shows how the workflow of the model proposed where further detail explanation will be done to explain the model output of the analytical study of the Lasiandra Finance Inc.

A list of loan information

Description automatically generated with medium confidence

Figure 1 : Datasets of the bank loan

Figure above shows the datasets used for the assignment and further details about the datasets is that it contained training and testing datasets in the csv format where it is imported into the SAS software.

3.0 Chapter 3 (Background Of Lasiandra Finance Inc. (LFI), New York, USA)

Its leading private financing company called Lasiandra Finance Inc. (LFI) New York, USA which caters to small startup or small companies especially (SME) Small and Medium Enterprises that needed funding. This company clearly understands that some businesses need a dream of extra push of fund so that it can accelerate the business growth faster. This company requires to upgrade their loaning process or system to be tailor made and customer centric. Also, in the past few years this company massively upgraded its wings to speed up and increase its process as it needs to automate loan eligibility process based on customer portfolio entered online.

Hence the main problem faced by the company is the process of approving the loans as the process of loan approval is complicated. The procedure is very complex as it needed constant verification and validation such loan criteria and eligibility as there is no guarantee that the chosen or eligible applicant will be accepted. There are several key takeaways here such as companies like Lasiandra Finance Inc typically fall under private sector banking where it is a type of banking system that generally held by private companies or very wealthy individuals. The table below shows the benefits of private banks and its counterpart public bank although both of them served the same purpose to make money or profit.

Table 1

|  |  |
| --- | --- |
| **Private Bank** | **Public bank** |
| Offers fast and quick services to its customers | Low interest charges on loans |
| Provide customized services according to customers preferences | High interest rate on deposits |
| Fast and quick financial decision making | Full security jobs for employees |
| Streamlined management system | Offers its services to large customer base |

The way manual loan applications work is that it consists of 7 steps before the loan application is approved by the bank. The process consists of pre-qualification process, loan application, application processing, underwriting process, credit decision, quality check and loan funding.

The table below shows the detailed steps of each manual loan application process.

Table 2

|  |  |
| --- | --- |
| **Step** | **Details** |
| pre-qualification process | In the original step of loan application process, this method works whereby the borrower needs to submit several lists of items to the lender to get loan. |
| Loan Application | At this stage the borrower completed the loan application and mostly nowadays it is done through web and mobile app. |
| Application Processing | The credit department then received the application where it reviewed for it completeness, accuracy and how genuine it is |
| Underwriting Process | At this stage there is several criteria need to be check by the lenders such as credit scores and risk scores |
| Credit decision | From the results of the underwriting process, the applicants will be notified by the banker whether the loan approval is approved or denied. |
| Quality Check | The lenders must ensure the quality of the manual loan application process since the rate of the lending is high and must be regulated most of the time. |
| Loan Funding | After the loan application process is completed when the documents are signed, it is hence loans fund shortly afterwards. |

4.0 Chapter 4 (ASSUMPTION, PROGRAM DEMONSTRATION, CODING AND JUSTIFICATION).

The deliverables and fulfillment of the assignment is that it required to conduct the data analysis on the datasets using SAS program and the report requires to introduce the data, method/technique, and coding problems. One of ultimate goals is to discuss the objective of the analysis and from that the output of the code results are interpreted in datasets exploration.

Regards to the program demonstration and coding, there are a few steps that need to be taken such as ensuring that break down the complex procedure of the coding into the important steps. These essential steps include showing the ability to demonstrate subsets of PROC, SQL code, macro code or supplementary code. Hence, each step the code explanation must done the way it anticipated and identify the section where improvement needed to be done and repeat again.

After that, discuss the obstacle when doing the SQL code programming and show the mistakes done. It is reasonable to show the programming problem and later show some modifications to make the project/assignment better. Often the times the modifications done when there are some changes occur on the goal and can be done by tweak the code and the output.

The end goal of the program is to have a complete code that demonstrate the datasets used which in this assignment the banking loan datasets, along with the product of code, tables and ODS output (Output Delivery System). Before that let dives deep into understand the PROC SQL, the relational database management system where relational database is an organized and structured collection of information where the data are arranged into a two-dimensional table. Each of the table contain usually one or more row and columns. Often the time the data are related, based upon their values and not according to other data structures.

For this data analytical assignment (DAP), SQL statements is used to read and update table. PROC SQL used structured query language that talks to relational database management system as SQL provide familiar action or task such as CREATE, SELECT, UPDATE, INSERT, DROP and DELETE, Below shows the task of SQL.

* Create and delete dataset.
* Retrieve and manipulate SAS dataset.
* Add or modify data values in a dataset.
* Create and delete indexes on columns in a dataset.
* Add, modify, or drop columns in a dataset.

It is known that PROC SQL can be used on SAS files, databases tables and combinations of these to execute query operations. The reason why SAS are chosen is that SAS as a programming language is very user friendly and it is easy to use compared to other programming language. This is due to the SAS used very simple syntax that uses abbreviated and direct commands and hence this makes it is an excellent language for people that has very little or no knowledge for programming. SAS programming language provides a simple user interface whereby this software includes charts, graph and plots and hence makes it easier to plot bars, graphs, and charts.

Moreover, benefits of using SAS are that it enhances data security as it provides high data security to many businesses and enterprises as SAS is one of the major key players of analytics tool used in most companies, business, and corporations. Hence due to this high security measures guaranteed by SAS, data manipulation is nearly impossible due data security provided by SAS.

5.0 Chapter 5 (Methodology)

The methodology used in this assignment is SEMMA method where it stands for "Sample, Explore, Modify, Model, and Assess.", whereby this method is a data mining and predictive analytics method, or technique developed and created by SAS institute. SAS is a software company widely known for data management tools and analytics solutions and below shows the detailed step of each of the SEMMA methods. Figure below shows the SEMMA method.

A diagram of a process

Description automatically generated

Figure 2

The first one is Sample whereby representative sample of the datasets is selected which in this case the TRAINING AND TESTING datasets. The sample datasets are very crucial to analyze a subset of data instead of the whole datasets. Next is Explore where data exploration is done to gain insights on the datasets chosen as for this part it is done to study is there any observations, and anomalies between the variables. Modify is a method where the data undergoes pre-processing and it undergoes imputation, data transformation, cleaning data and handling missing values.

Next is the Model where after the data preparation from Modify, this step participating in building a predictive model where the data is prepared to train the machine learning model selected. Hence the proposed machine learning model used is for predictions or classification based on new data. The last one is Assess where after model building was done the performance of the model is assessed using evaluation matrices such as performance matrices to determined their accuracy at making predictions. This method is also crucial to study the effectiveness of the proposed model and for these datasets some of the categorical variables are education level, gender, loan amount, loan approval, and loan type.

6.0 Chapter 6 ( Data Dictionary/ Metadata )

6.1 Introduction

Exploration on the datasets is done where the datasets used contain train and test datasets which are imported into the SAS studio software. EDA stands for "Exploratory Data Analysis" where it is an approach for data analysis which solely focuses on studying the characteristics of the datasets using statistical measures. Some of key features of EDA include the data summarization, data visualization, data cleaning, pattern recognition, hypothesis study, feature selection and data transformation. Figure below shows the loan datasets used in this assignment.

A list of loan information

Description automatically generated with medium confidence

Figure 3

Data exploration is very crucial to find insights on the datasets by understanding the data structure or in this case metadata which means by data that describes other data.

6.1 Location of the datasets on SAS

A screenshot of a computer

Description automatically generated

Figure 4

6.2 Project datasets found inside the SAS permanent library/folder.

A close up of numbers

Description automatically generated

Figure 5



Figure 6

6.2.1 Description

The purpose of SAS library is simply a collection of SAS files that are stored in the same folder or directory of the computer. Figure above shows the datasets stored inside SAS permanent library whereby a permanent SAS library is stored inside external storage and will be not deleted when SAS session terminated. The maximum length of a library is typically 32 characters, and the length is maximum 8 for numeric letters. The permanent library name for this assignment is DAP67696 where it stored the TESTING\_DS and TRAINING\_DS files.

6.2.2 Display the structure of the datasets-DAP67696.TRAINING\_DS

6.2.3 Data dictionary of the dataset- DAP67696.TRAINING\_DS

A white text with green text

Description automatically generated with medium confidence

Figure 7

6.2.4 SAS PROC SQL Codes

A close up of a number

Description automatically generated

Figure 8

6.2.5 Screenshot(s) of the output

A screenshot of a computer program

Description automatically generated

Figure 9

6.2.6 Description

The diagram above shows the structure of the TRAINING\_DS dataset where CREATE TABLE command or syntax is used to create a table shown in diagram above. The table above shows the variable name along with data type and length in the new table created and for this project the library created is DAP67696 and from the output of the SAS log window the table is created successfully.

6.2.7 SAS PROC SQL Codes

A screenshot of a computer

Description automatically generated

Figure 10

6.2.8 Screenshot(s) of the output

A screenshot of a computer

Description automatically generated

Figure 11

A screenshot of a computer

Description automatically generated

Figure 12

A screenshot of a table

Description automatically generated

Figure 13

6.2.9 Description

Based on the figure above the it shows the data structure of the DAP6796.TRAINING\_DS whereby the number of variables is 13, the number of observations is 614 and the observation length is 96. The purpose of this method is to store data in the SAS system memory and the data can easily be monitored and retrieved by organizations to understand the characteristics of the datasets better.