**Final Project Report**

**Admin No: 19B599X**

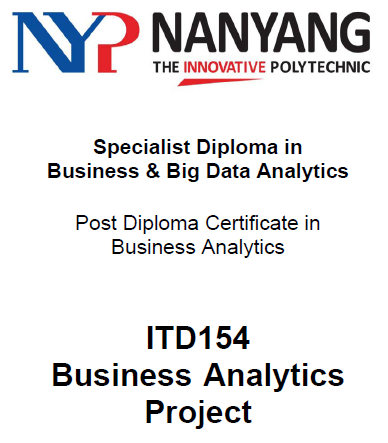


Table of **Contents**

[1 Introduction 3](#_Toc33480892)

[2 Work Accomplished 3](#_Toc33480893)

[2.1 Business Objective 4](#_Toc33480894)

[2.2 Data Mining Objective 4](#_Toc33480895)

[2.3 Data Understanding 4](#_Toc33480896)

[2.3.1 Describe and Explore Data 4](#_Toc33480897)

[2.3.2 Verify Data Quality 5](#_Toc33480898)

[2.4 Data Preparation 5](#_Toc33480899)

[2.4.1 Select and Clean Data 5](#_Toc33480900)

[2.4.2 Construct and Integrate Data 6](#_Toc33480901)

[2.4.3 Format Data 6](#_Toc33480902)

[2.5 Modeling 7](#_Toc33480903)

[2.5.1 Select Modeling Technique 7](#_Toc33480904)

[2.5.1.1 Weight by Correlation 7](#_Toc33480905)

[2.5.1.2 Build Model: Logistic Regression 8](#_Toc33480906)

[2.6 Evaluation 9](#_Toc33480907)

[2.6.1 Evaluate Result 9](#_Toc33480908)

[2.6.2 Review Process 9](#_Toc33480909)

[2.6.3 Improve the Prediction Accuracy 10](#_Toc33480910)

[2.6.4 Conclusion 11](#_Toc33480911)

[3 Reflection 12](#_Toc33480912)

[4 Proposed Usage of Skills Learned 13](#_Toc33480913)

[5 Appendix 14](#_Toc33480914)

[5.1 Figures 14](#_Toc33480915)

[5.2 Files 14](#_Toc33480916)

# Introduction

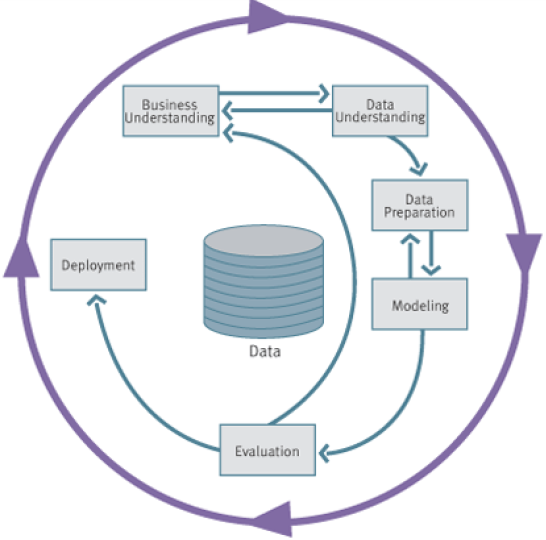
Identifying the customer's background and evaluating their needs in any type of loan application is a hectic process for the banking sector employees, especially when they have requested for high housing loans. The banking sectors will evaluate the customer's need with certain criteria to assess their repayment capability about the monthly repayment will pay in time without any due and late fee charges.

For the business project report, the NYB Bank dataset to analyze the Customer’s transaction history to understand the criteria behind their late payment. The project process follows the Cross Industry Standard Process for Data Mining Initiative (CRISP-DM) as general guidelines.

The project report describes the data preparation, cleansing and transformation of dataset to visualize the relevant attributes, data preparation is a most time-consuming process in providing solution to the business objective. The data mining identify the relationship among attributes in a dataset and model predict the customer who make late payment as our business objectives and, the analysis outcome give us recommendation to suggest to Bank employees. The report includes the learning process and knowledge attained to apply in a working place.

# Work Accomplished

This portion explains and guidelines the CRISP-DM process analytics to how the data has prepared from the data set, cleansed and transformed to obtain the business objective, its modeling technique and result to predict the data from the past transaction history and late payment. I have selected the NYB Bank data set to use for my Business project work.



CRISP-DM Process Analytic Model

## Business Objective

To find out the criteria for which type of customers will make late payment, if they have taken the housing loan.

## Data Mining Objective

To know what are the customer’s attributes (pattern and trends) show impact for late payment.

## Data Understanding

### Describe and Explore Data

The NYB Bank Data set has 28 individual table with many attributes correlation with other tables, its total rows, properties of each field, meaning of attributes in the table for business terms. The NYB bank database has relationship link between all the tables that is CSTCustomerMaster Table (Primary Key, CustomerID), which link with CSTCustomerEmployment table primary key, LNSAuto and LNSHousingLoanMaster Table, Credit card master table, Deposit and Fixed deposit tables.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Tables** | **No. of Records** | **Important Attributes** |
| 1 | CSTCustomerMaster | 6086 | CustomerID, Name, DOB, Nationality, No.of Dependent, Residential Address |
| 2 | CSTResidentialType | 12 | Type of residential property in Singapore |
| 3 | CSTCustomerEmployment | 6086 | Customer's Annual Income |
| 4 | LNSHousingLoanMaster | 1484 | Principle and Joint CustomerID, Housing Loan ID, Loan Start and Maturity period, Loan Principle Amount, Property Purchase Price |
| 5 | LNSHousingLoanPayment | 79771 | Monthly Payment, Balance amount |
| 6 | LNSHousingLoanLateFee | 1050 | LateFee charges |

Description of Tables

By exploring the data and drafting the statistical charts to see the correlation among the attributes to understand the customer’s background, refer the charts (Appendix 5.1 Figure A to C) from CSTCustomerMaster table and LNSHousingLoanMaster table.

### Verify Data Quality

Before proceed for data preparation, have verified the data quality of the tables chosen to construct the dataset for analysis. Analyze the tables for any missing value, blank fields and irrelevant information not related to the prediction and handling the missing value or removing in accessed in Data Preparation process.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tables** | **Which Field?** | **Qty** | **Missing or Irrelevant Value** |
| CSTCustomerMaster | Residential Address2 | 352 | Missing value |
| CSTResidentialType | - | - | - |
| CSTCustomerEmployment | - | - | - |
| LNSHousingLoanMaster | JointCustomerID | 96 | Missing Value |
| LengthofStayofJoint | 86 | Irrelevant Value |
| Resid.statusofJoint | 86 | Irrelevant Value |
| PropertyAddress 2 | 70 | Missing Value |
| LNSHousingLoanPayment | - | - | - |
| LNSHousingLoanLateFee | - | - | - |

Data quality (Missing Value)

## Data Preparation

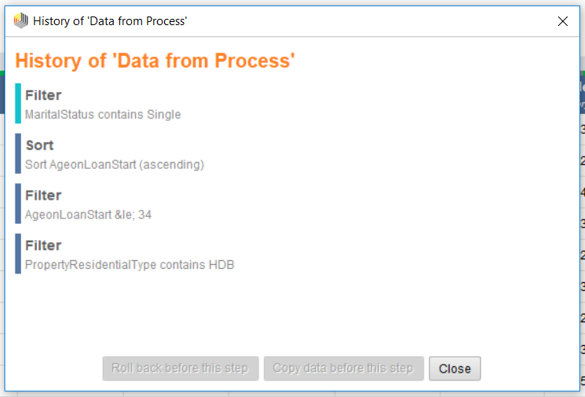
### Select and Clean Data

First step in a data preparation is Select the related tables and clean the tables without any issue to proceed for construct and integrate data. In a Rapidminer, read access the NYB database Access File, choose individual table name and execute the process, missing value and irrelevant field replace with meaningful value. For replacing missing value, use the TurboPrep in the Rapidminer (Appendix 5.1 Figure D TurboPrep for Data Clean).Similarly all other tables have read and replaced the missing value with meaningful value and removed the irrelevant information in the field, list of changes in the tables have listed below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tables** | **Which Field?** | **Qty** | **Missing or Irrelevant Value** | **Replace with** | **Reason** |
| CSTCustomerMaster | Residential Address2 | 352 | Missing value | "-" | Residential type of Shop House, Terrace House, Bungalows and Executive Condominium, do not have Address 2. |
| CSTResidentialType | - | - | - | - | - |
| CSTCustomerEmployment | - | - | - | - | - |
| LNSHousingLoanMaster | JointCustomerID | 96 | Missing Value | "0" (Zero) | No Joint Customer (PrincipleCustomer's Marital Status: Single) |
| LengthofStayofJoint | 86 | Irrelevant Value | "0" (Zero) | No Joint Customer, Some field have irrelevant value set to Zero |
| Resid.statusofJoint | 86 | Irrelevant Value | " " (Space) | No Joint Customer, Some field have irrelevant value set to Zero |
| PropertyAddress 2 | 70 | Missing Value | "-" | Property Residential type of Shop House, Terrace House, Bungalows and Executive Condominium may not have Address2. |
| LNSHousingLoanPayment | - | - | - | - | - |
| LNSHousingLoanLateFee | - | - | - | - | - |

Data Cleansing Details

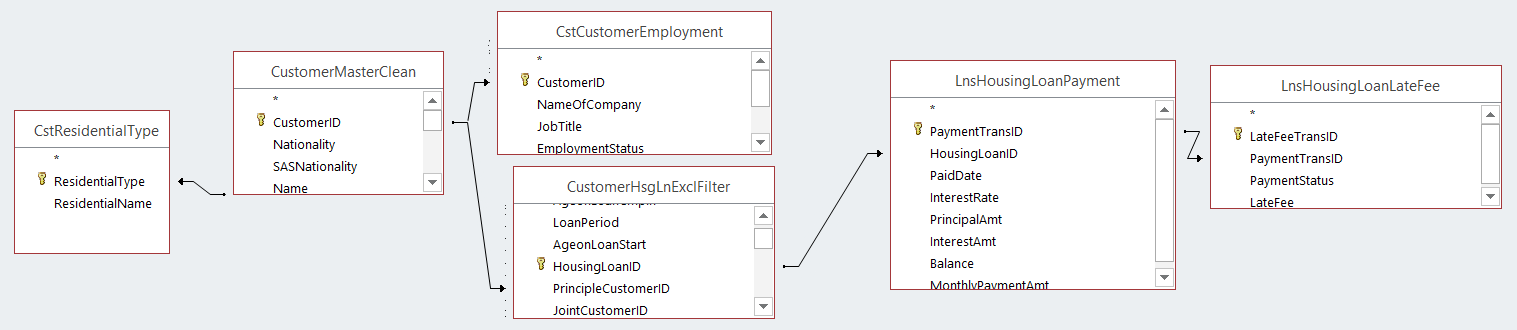
For those tables have replaced the missing values, saved in a new table named as CustomerMasterClean and combined table of Customer with Housing Loan Master after cleaned as CustomerHsgLnClean. The CustomerHsgLnClean file has further removed the data due to the Singapore National who are less than 35 years old Single cannot buy the HDB flats in Singapore from LNSHousingLoanMaster and New tables saved as CustomerHsgLnExclFilter (1456, original table 1484) as final data to use for prediction.



Excluded Age <35 years Old Single Singaporean

### Construct and Integrate Data

Another step in data preparation is to construct and integrate the multiple tables into a single database with all required attributes to use in the analysis process. A new query has designed with a combination of CustomerMasterClean Table, CSTCustomerEmployment, CustomerHsgLnExclFilter table, LNSHousingLoanPayment and LNSHousingLoanLatefee to construct a combined dataset table as single file and saved as “CustomerEmpHousingLnData”. The combined dataset has exported to Excel File in the same name as CustomerEmpHousingLnData.xlsx. Now, the exported data has a field count of 77764 rows with the multiple transaction of every customer over a period of few years.



Construct of Single Database

### Format Data

The objective to model the machine learning to detect the late payer from the late fee transaction history; noticed the total transaction from a combined table is 77764. It is not feasible to process the data with multiple transaction from Loan payment table, remove the duplicate field of transaction history to unique field with late fee information. I have considered the customer who had paid late any one time in few transactions as late payer and final data of CustomerHousingLnLatePay (1456 data) and rearranged the final data as below format.

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Description of Attributes** | **Data Type** |
| CustomerID | Employee ID (Primary key) | Integer |
| Full Name | Combination of Name and Family Name | Polynominal |
| Gender | Gender | Polynominal |
| Nationality | Nationality | Polynominal |
| SASNationality | SASNationality | Polynominal |
| Age | Age, derived from BirthDate | Integer |
| EducationLevel | Educational Qualification | Polynominal |
| MaritalStatus | Marital Status | Polynominal |
| NoOfDependents | No. of Dependents | Integer |
| JobTitle | Job Title | Polynominal |
| AnnualIncome | Annual Income | Integer |
| LoanStartDate | Loan Start Date | Date |
| LoanMaturityDate | Loan Maturity Date | Date |
| LoanPeriod | Loan Period | Integer |
| LoanPrincipalAmt | Loan Principle Amount | Real |
| PropertyResidentialType | Property Residential type Purchased | Polynominal |
| PropertyPurchasePrice | Property Purchased Price | Integer |
| MonthlyPaymentAmt | Monthly Payment Amount | Real |
| Late Payment | Late Payment | Polynominal |

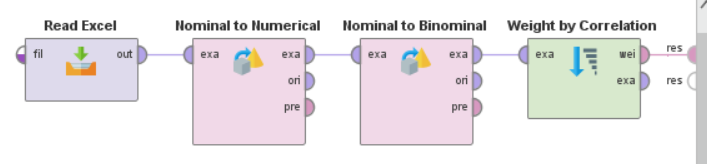
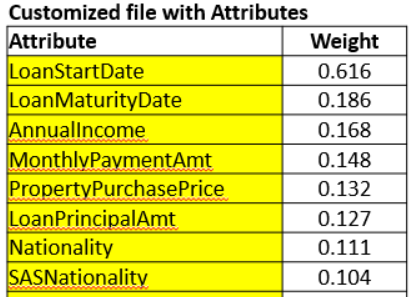
Format of Data

## Modeling

### Select Modeling Technique

The Final dataset CustomerHousingLnLatePay (1456) is having many polynominal data type attributes and the predicted field is a polynominal label. To predict the Polynominal label data, can use classification tree or logistics regression statistical modeling technique. For this project analysis, have selected the logistic regression to predict the label.

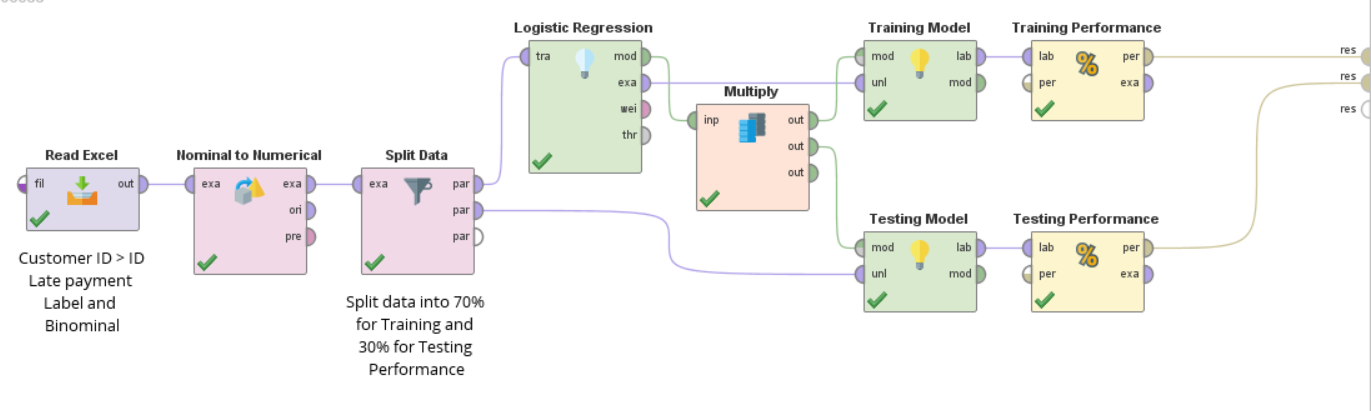
#### Weight by Correlation

Weight by Correlation Correlation Result

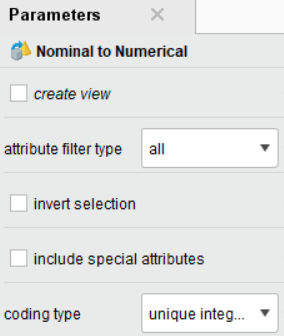
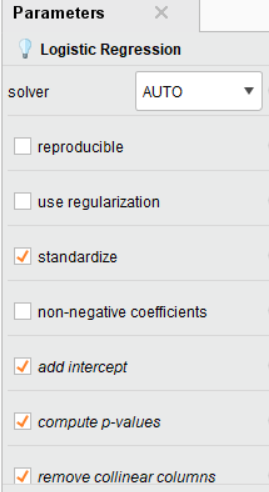
To determine the most relevant attributes for predicting label by weight the attributes by correlation operators. The selection of attributes to predict is normally from weight of correlation or Feature selection from all combined attribute. Here, have chosen the attributes based on personal decision. The customized attributes weight from the executed process shows the strong relation for predicting label is LoanStartDate (0.616), followed by LoanMaturityDate (0.186), AnnualIncome etc.

#### Build Model: Logistic Regression



Logistic Regression Model

The CustomerHousingLnLatePay file is imported and set the CustomerID role as ID, exclude name as it don’t serve any purpose in decision making, set the polynominal label to Binominal and Label. We make use of same dataset to train the logistic regression model by splitting the dataset into 70% for training and 30% for testing the effectiveness of trained model. Set the Nominal to Numerical operator parameter coding type to “Unique Integers” and Logistic regression parameter solver to “AUTO”. “Performance Classification” operator predicts the training model performance with “Apply model”. The focus of building model is to predict the Sensitivity (True Positivity of the Late payment) of the Dataset.

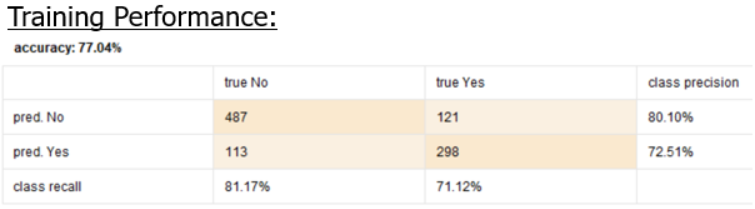
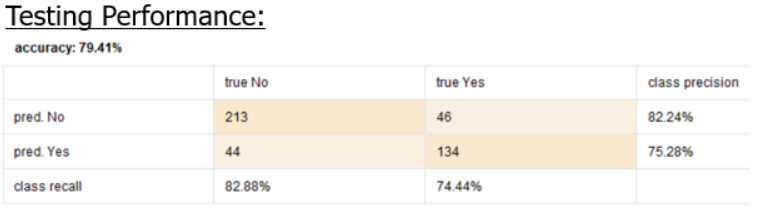
 

Nominal to Numerical Logistic Regression

## Evaluation

### Evaluate Result

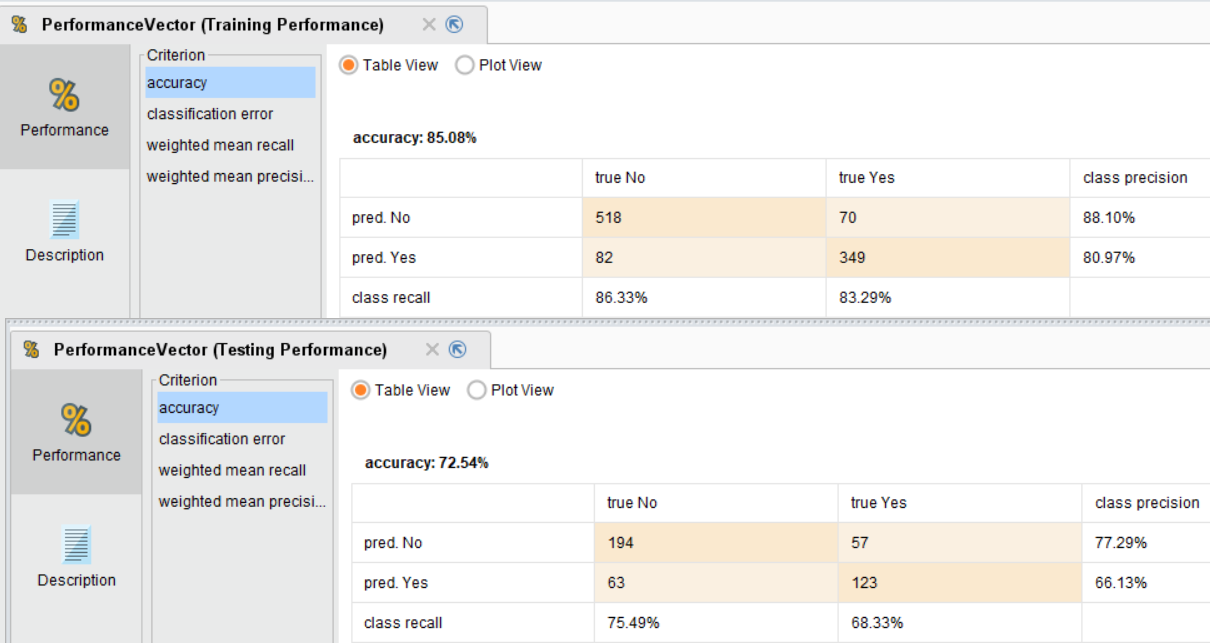
By evaluating the predicted result, training and test performance accuracy of build model is 77.04%, whereas the True Negativity, Specificity is 80.1%) and 79.41% whereas the True Positivity, Sensitivity is 72.5%) respectively. Our focus is to predict the late payment, true positivity, Sensitivity of the dataset. The reason for test predicting accuracy is better than training due to NO Late Paid (857 data) ~60 % and Late Paid (599 data) 40%.; it is slightly skewed to NO Late Paid and still acceptable to build the data without any balancing required to do in the dataset.

. 

Training and Testing Performance

With the prediction, help the Bank employee to analyze the Loan Payment capacity in time with main attributes of LoanStartDate, LoanMaturityDate, AnnualIncome and MonthlyPaymentAmt, PropertyPurchasePrice, PrincipleLoanAmt and their Nationality. Upon knowing their Loan Payment capacity, the bank employee can suggest the customer to go longer period of loan payment by reducing the MonthlyPaymentAmt.

### Review Process



Nominal to Numerical (Dummy coding)

The predicted outcome is not satisfactory to expectation more than 80% accuracy and for predicting new customer capability. From the accuracy 1 in 4 prediction is not correctly predicted (about 20 ~ 23% as classification error %). By changing the Nominal to Numerical operator parameter coding type as “Dummy coding”, the training accuracy increased to 85.08% (from 77.04%) and testing accuracy dropped to 72.54% (from 79.14%), the difference between train and test is more than 5% (12.54%) which is under sampling. There is no change in predicting accuracy by changing the default setting in Solver for Logistic regression.

### Improve the Prediction Accuracy

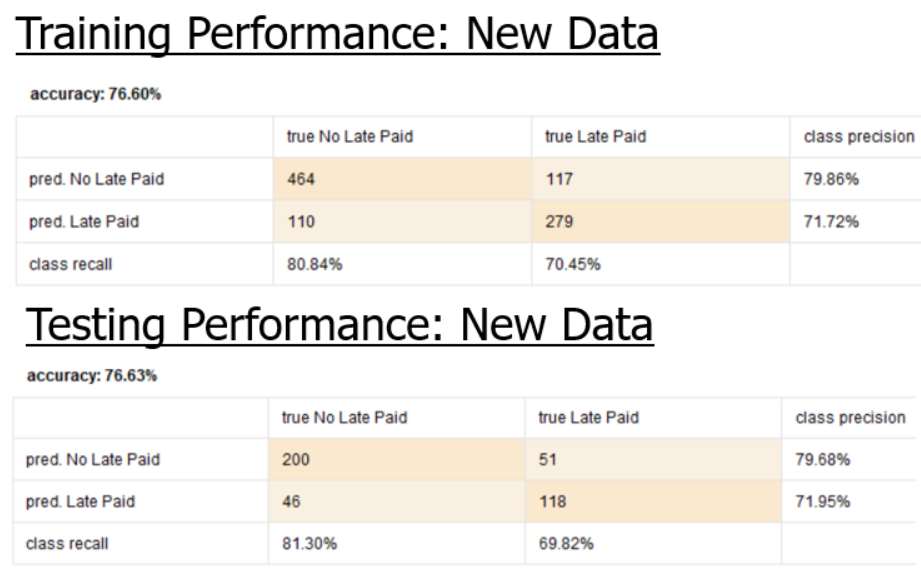
There are few ways to improve the accuracy of prediction of customer capability

* balancing the data to have equal “No Late Paid and Late Paid”
* Use other data modeling technique
* Improve the data quality (Cleansing for New file, Refer Appendix 5.1 Figure E to G )

|  |  |  |
| --- | --- | --- |
| **CustomerHousingLnData** | 1484 | **Rows Removed Reason** |
| Final | 28 | Singaporean single aged <35years old |
| 1456 |  |
| **CustomerHousingLnData (New)** | 1456 | **Rows Removed Reason** |
| Final | 17 | Total income is very much lower than MonthlyPayAmt, suspect mostly payout from CPF account |
| 1 | Very low monthly Payment which is $70 per month (Outlier) |
| 43 | Principle & JointCustomer age is < 21 years at the time of Loan Start |
| 10 | Marital Status of PrincipleCustomer is Single and Joint Customer is married, Wrong Information |
| 1385 |  |

Data Cleaning (New)

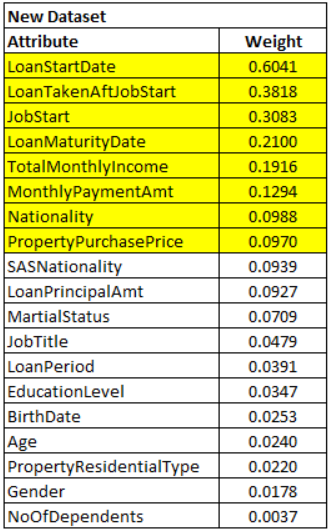
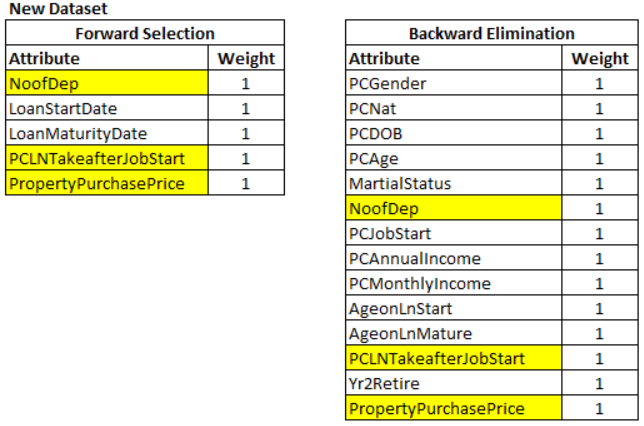
* Logistic regression with new Data set



With improved dataset the Training and Testing accuracy not much improved than the original dataset whereas the accuracy difference between Training and Testing has improved in New Dataset (0.03%) compared to Original dataset (2.37%), confirmed the New Dataset with further cleansing enhance the modeling to predict the Late payment.

### Conclusion

The new dataset with feature selection operator find out the most relevant attribute in predicting the customer’s late payment. The weight by correlation operator and feature selection operator identified attributes and common attributes at feature selection has some commonalities of PropertyPurchasePrice and how many years after the housing loan has taken from Job start (LoanTakenAftJobStart) attribute in the new table.

Weight by Correlation (New Data) Feature Selection (New Data)

With the cleaned data, the accuracy of predicting defines the No. of dependent, Property Purchase Price (housing) and How many years of work has the customer has while applying for the loan determine the loan paying capacity.It is clearly shows the correlation that the customer who had started their career after 1980 and taken the housing loan after that will go for late payment once in a while.

By knowing the important attribute like Job started, No. of dependent, Purcahse price help the banking sector Employee to predict the customer capability of Loan payment. Further, the banking employee can suggest the customer to go for longer payment duration or minimize the monthly payment.

# Reflection

I have gained knolwedge in familiarization and understanding the dataset using CRISP-DM Statistical process flow. In this project, the selected dataset is MSAccess which is backbone of any SAP system, through this have gained the knowledge to use Query Wizard, Simple Query, Create new table from exisitng mulitple tables, create calculated field from fields and export the Access table into Excel format.

Understanding the data and Preparing the data set for modeling is a toughest job, it requires constant revisit to the original table and make additional columns to give more meaningful in predict the required Label, enjoyed and kept constant thinking of how to improve and what kind of data to include and exclude according to the business objective.

Additionally, I have make use of TurboPrep which is help in cleansing and predicting the model which is not thought in our PDC1. A new tool has learned from the project.

Have applied all the learning skills from PDC1 (Business Essential, Statistics and Analysis technique, Data Visualization from Rapidminer) to make the raw data to useful insight data to predict the Late payment Label.

Here, I have used other modelling technique to compare and analysis the predicting accuracy like Cross Validation with SVM (Appendix 5.1 Figure H, result of SVM model in Figure I and Comparison result of all models in Figure J) Decision Tree and Feature selection operator to select relevant attributes and then compared the result from different model to understand better in learning point of view.

Gained data handling knowledge to improve the predicting accuracy from the dataset like balancing and cleansing.

# Proposed Usage of Skills Learned

From the business project learning, can make use of gained knowledge in many sectors, such as

* Banking Sector

To analyze the customers capabilities in terms of age, income, education, saving from CPF account, deposit account and suggest customer and bank to offer or not.

Help to incorporate the learnt skill in identifing fraud transaction.

* Employment Sector

To understand the employee’s behaviour and their attrition reason that impact the business output.

* Trading investment prediction

Review and monitor the stock transaction details and divident payout prediction.

* Estimation of power usage in Industries and service sector.

To Predict the household power consumption over the climate change and for servicing sector power consumption over the market fluctuation (Semiconductor industry).

* Customer satisfaction

To predict the customer turn around again to dine from the customer satisfaction and suggestion to improve their turn around.

* Graduation students career or further study prediction

To predict the student pursue the post graduation or employment from the family background, GPA scored, Current market trend.

* Medical sector

To predict the infection condition based on analysis.

* Transportation Sector

To predict the fastest and shortest route for travelling based on current condition.

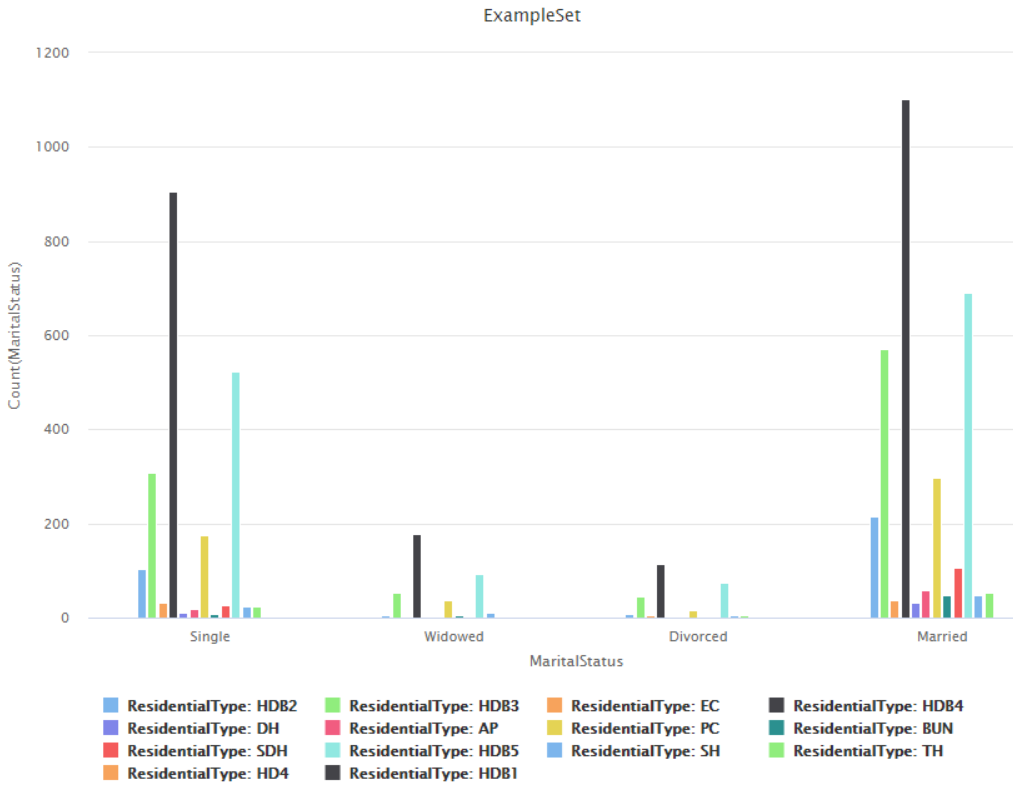
* Food Delivery App

To predict the customer’s need in the menu and suggest recommended Top up with the original purchase based on ethnic groups.

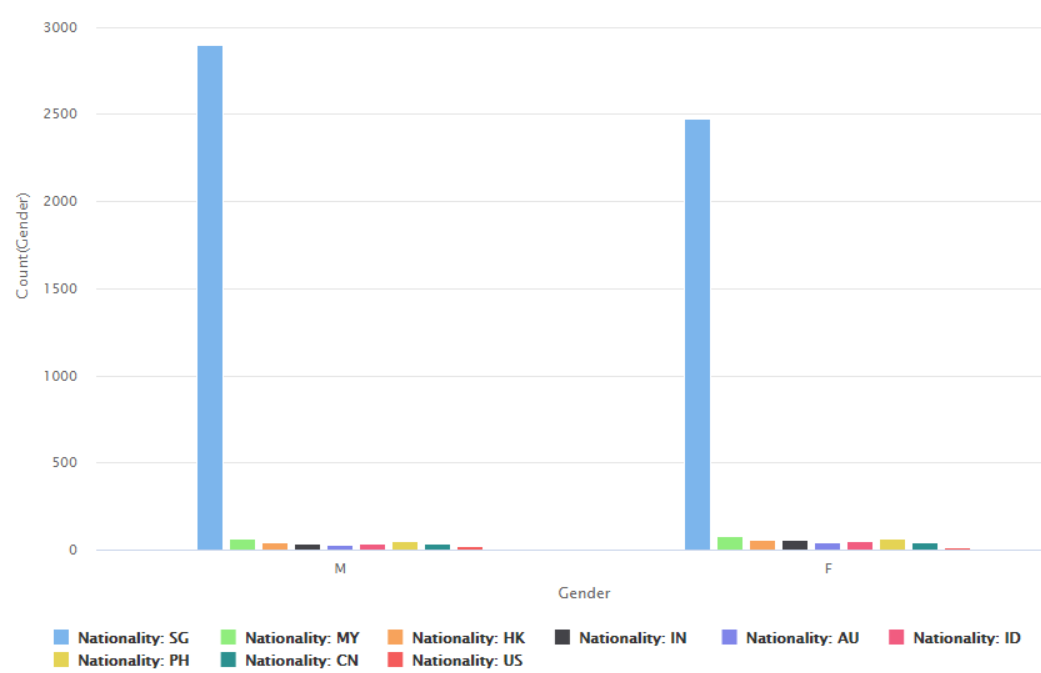
# Appendix

## Figures

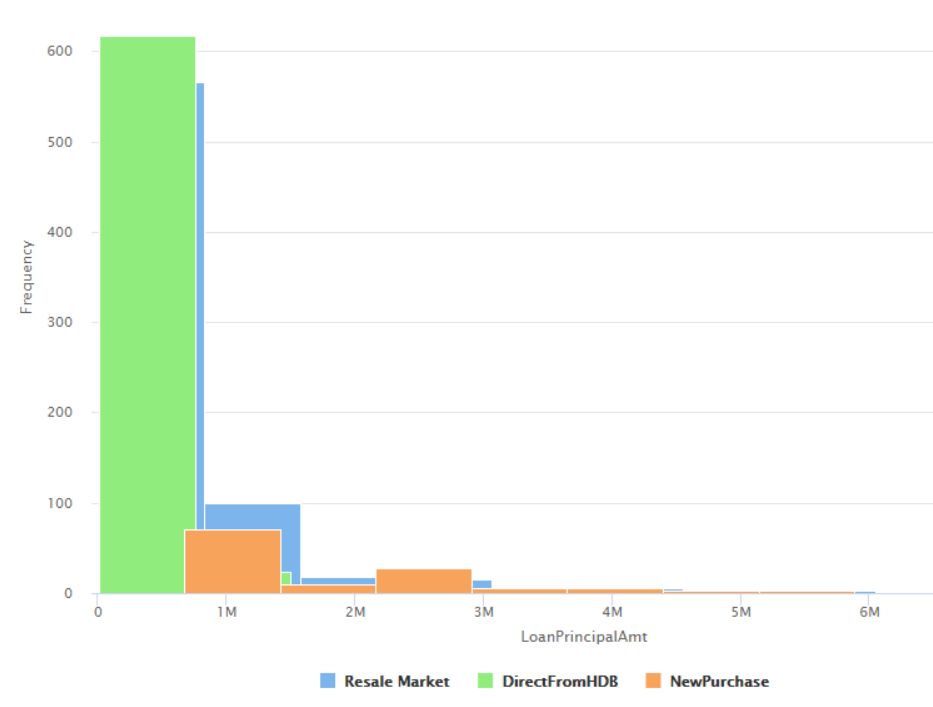
1. Marital Status vs Residential Type (CSTCustomerMaster)



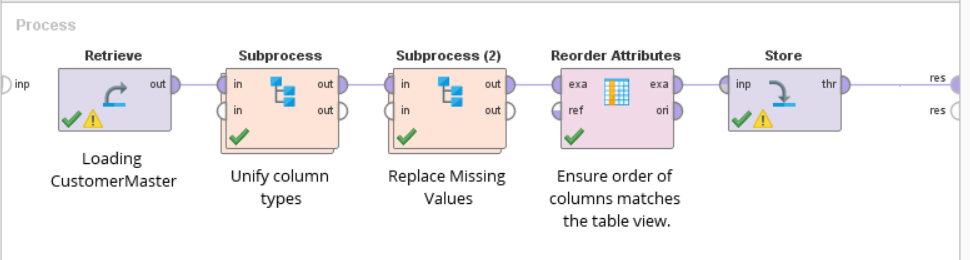
1. Gender vs Nationality (CSTCustomerMaster)



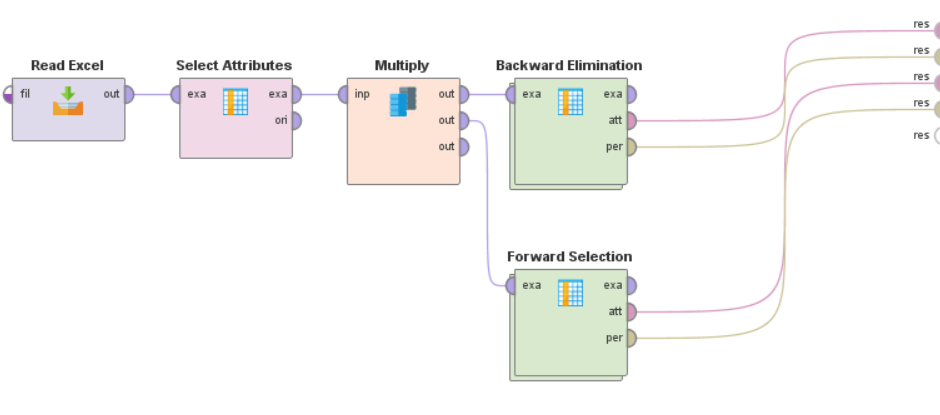
1. LoanPrincipleAmt vs PropertyPurchaseType(LNSHousingLoanMaster)



1. TurboPrep for Data Clean

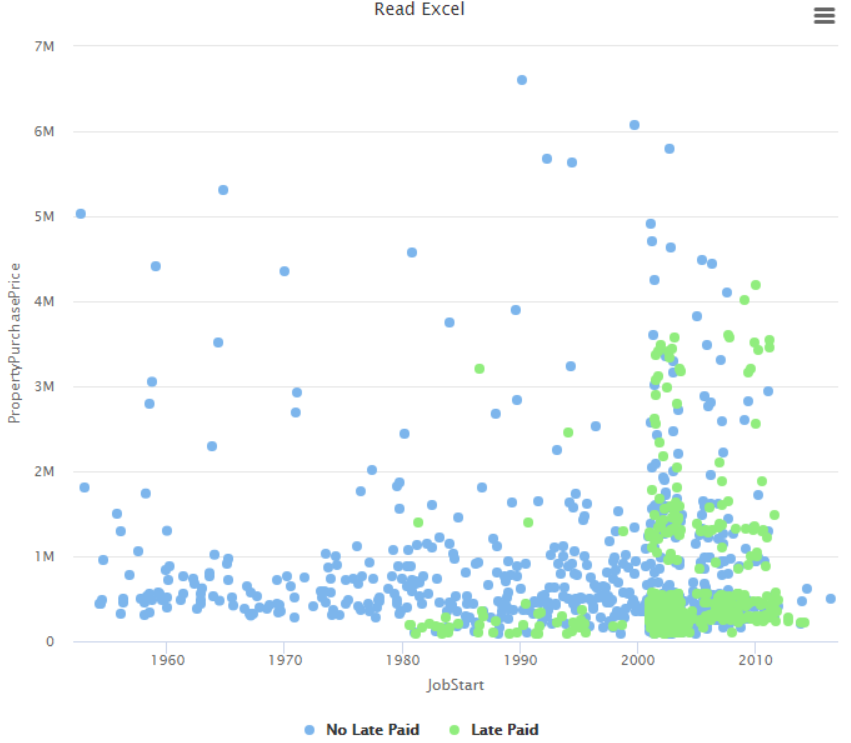


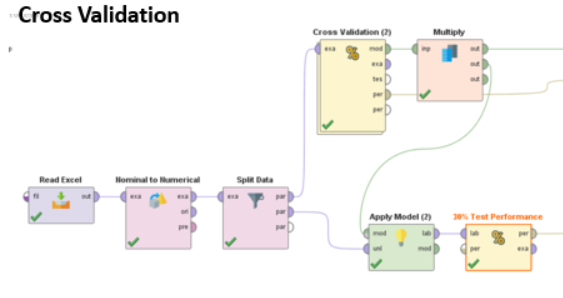
1. Logistic Regression Model – New Data (for Result Improvement)



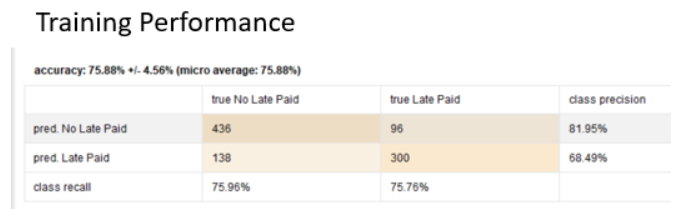
1. Correlation chart between Loan taken after Job Start vs Job Start, Group by Late Payment (New Data)

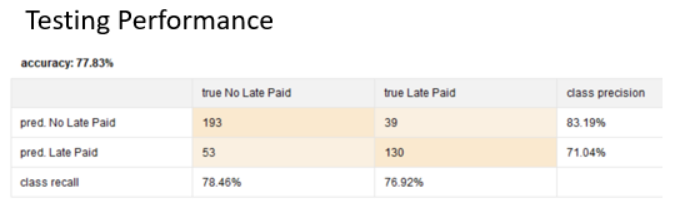


1. Correlation chart between PropertyPurchasePrice vs Job Start Group by Late Payment (New Data)
2. Cross Validation with SVM

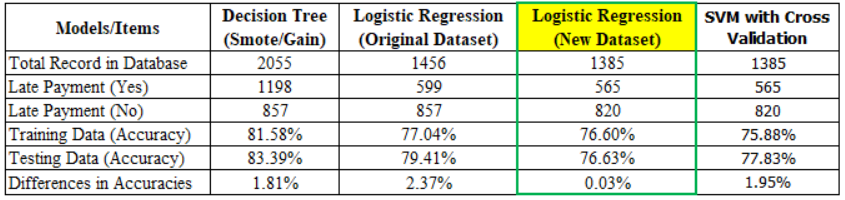


1. Cross Validation with SVM Results





1. Comparison of all models results



## Files

1. CustomerHousingLnData (Old File)



1. CustomerHousingLNData (New File)



1. Rapid Miner Files

