Credit Card Customers - Predicting Customer Churn

UNIVERSITI MALAYA
KUALA LUMPUR

WQD 7005 : Data Mining - Group Assignment

Instructor: Prof. Dr. Loh Yuen Peng

Group Members	Student ID
Nicholas Tan Yu Zhe	S2180436
Yap Jia Xian	S2150857
Touhid Ahmed Choudhury	S2150778
Nasir Uddin Ahmed	S2015449

Intro & Recap

Problem Statement

The problem at hand is customer churn, which is a significant concern for businesses, particularly financial service providers such as banks. Customer attrition can have adverse effects on a bank's earnings and reputation. To address this issue, it is crucial for banks to forecast customer churn accurately. The objective of this project is to utilize data mining techniques on the Credit Card Customers dataset available on Kaggle to create a predictive model that can identify consumers likely to churn.

Analysis Goal

- Develop accurate customer churn prediction model for credit card firms to optimize marketing and retention efforts.
- Applying the SEMMA Methodology
- Evaluate the performance of the model

Data Overview

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Type	Format	Informat	Length
Attrition_Flag	Input	Nominal	No		No	.	.	Character	$17.	$17.	17
Avg_Open_To_E	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Avg_Utilization_I	Input	Nominal	No		No	.	.	Character	$20.	$20.	20
Card_Category	Input	Nominal	No		No	.	.	Character	$8.	$8.	8
CLIENTNUM	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Contacts_Count	Input	Nominal	No		No	.	.	Character	$1.	$1.	1
Credit_Limit	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Customer_Age	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Dependent_cour	Input	Nominal	No		No	.	.	Character	$1.	$1.	1
Education_Level	Input	Nominal	No		No	.	.	Character	$13.	$13.	13
Gender	Input	Nominal	No		No	.	.	Character	$6.	$6.	6
Income_Categor	Input	Nominal	No		No	.	.	Character	$14.	$14.	14
Marital_Status	Input	Nominal	No		No	.	.	Character	$8.	$8.	8
Months_Inactive	Input	Nominal	No		No	.	.	Character	$1.	$1.	1
Months_on_bool	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Naive_Bayes_Cl	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Total_Amt_Chng	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Total_Ct_Chng_	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Total_Relationsh	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Total_Revolving	Input	Nominal	No		No	.	.	Character	$4.	$4.	4
Total_Trans_Am	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Total_Trans_Ct	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
VAR23	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8

Property	Value
Data Source	https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers?resource=download
Data Name	Credit Card Customers
Data Size	1.44 MB
Year	2020
Dimension	10,000 Rows and 23 Columns

Analysis Data

Demographic data (age, gender, income, education) can reveal correlations between customer demographics and churn behavior.

Credit card usage data (type, limit, transactions, utilization) provides insights into churn likelihood based on customer behavior.

Customer tenure, product count, and account status offer insights into the relationship with the bank and churn potential.

Churn label identifies customers who have churned, enabling the development of accurate predictive models for churn.

SEMMA Methodology

Sample

Explore

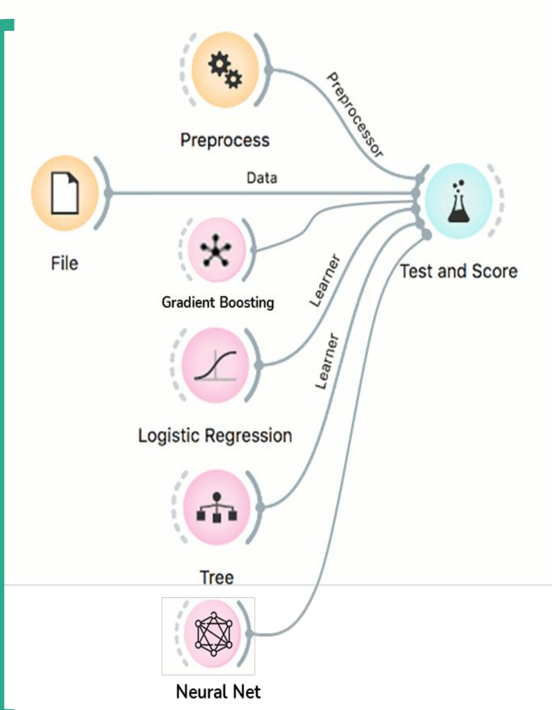
Done in Phase 1

Modify

Model

Assess

Phase 2



Platform: SAS Enterprise Miner Client 15.2

Modify - Importance

Spend Less Time & Money

As the likelihood of receiving inaccurate results after data cleansing is reduced.

Boost Data Quality

Improves the completeness, consistency, and reliability of data.

Enhance Decision Making

Quality of conclusions relies heavily on the quality of used data.

Data Quality Issues Faced

Incomplete

Marital Status

Noisy

Income Category

Inconsistent

Total_Revolving_Bal,
Total_Amt_Chng_Q4_Q1
Dependent_count,
Months_Inactive_12_mon, Gender,
Contacts_Count_12_mon,
Avg_Utilization_Rati

Data Cleaning & Solving Quality Issues

Tools Used talend sas

Solution of Incompleteness

Mode values with simulated missing values
Marital_Status

Solution of Noisiness

"Unknown" category removed
Income Category

Solution of Inconsistencies

Changed erroneous values for inconsistent values.
eg: '-' to 0

Transformation Dataset

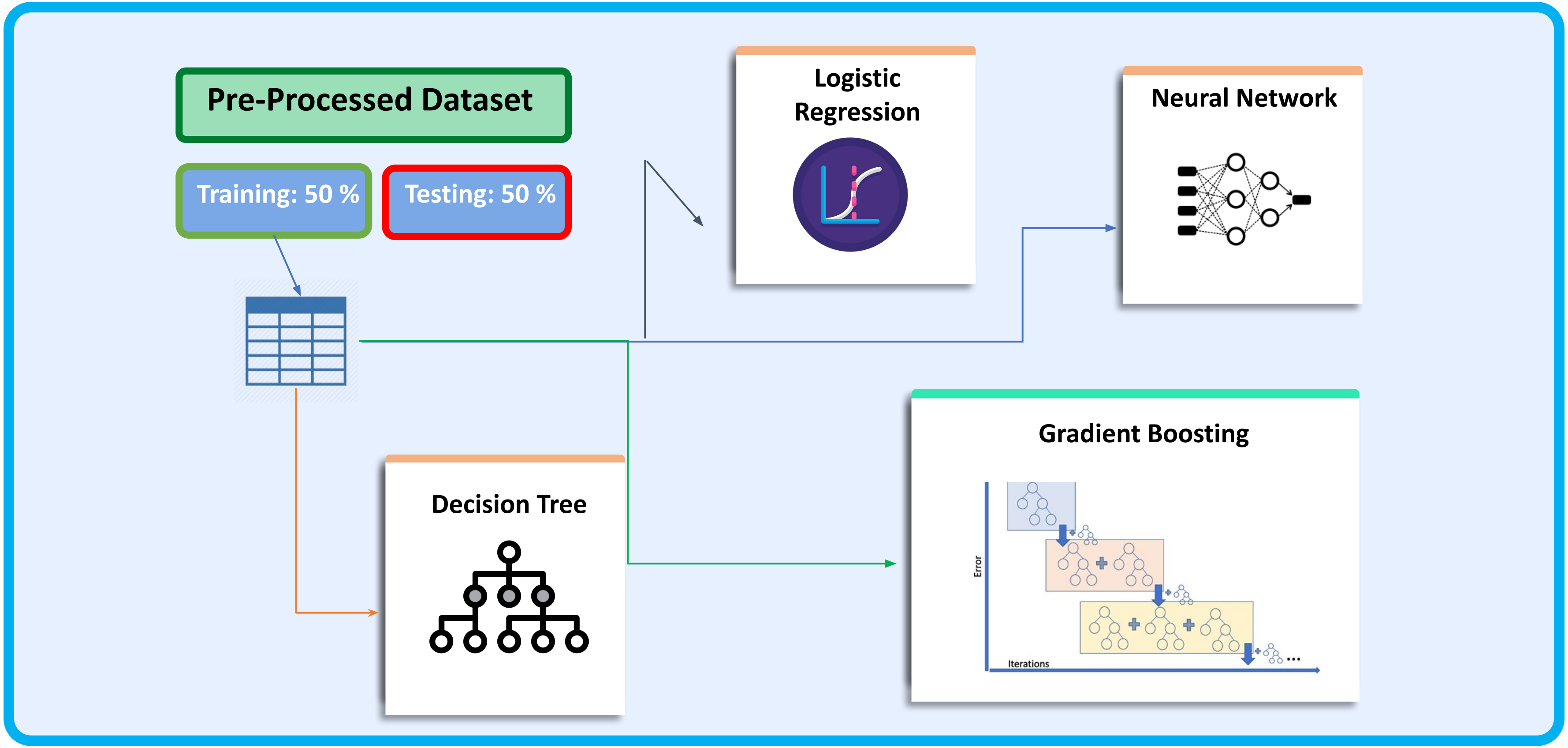
Cleaning Results for Data

After Data Cleaning & Modification

Pre-Processed Dataset

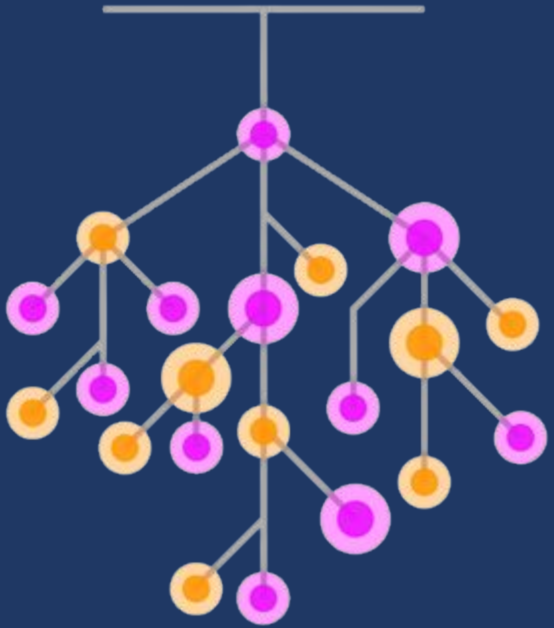
No Duplicates **No Missing Values** **No Inconsistency** **No Noisiness**

Modelling



Pre-Processed Dataset
Training: 50 %
Testing: 50 %
Logistic Regression
Neural Network
Decision Tree
Gradient Boosting
Error
Iterations

Decision Tree



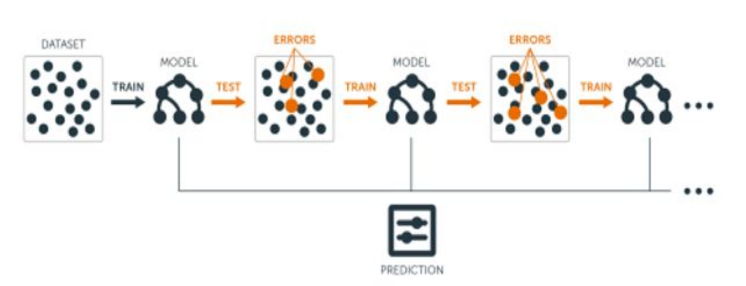
Model Accuracy:
93.68%

Findings

The top 5 variables that have the highest importance to customer attrition are Total Transaction Count, Total Revolving Balance, Total Relationship Count, Total Transaction Amount, and Total Change of Count in Quarter 4 to Quarter 1

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LG10 Total Trans Ct	Transformed Total Trans Ct	4	1.0000	1.0000	1.0000
LG10 Total Revolving Bal	Transformed Total Revolving Bal	2	0.8309	0.8382	1.0089
Total Relationship Count		4	0.5533	0.5495	0.9931
LG10 Total Trans Amt	Transformed Total Trans Amt	5	0.4939	0.5003	1.0130
LG10 Total Ct Chng Q4 Q1	Transformed Total Ct Chng Q4 Q1	3	0.3895	0.2739	0.7032
LG10 Total Amt Chng Q4 Q1	Transformed Total Amt Chng Q4 Q1	1	0.2767	0.2389	0.8635
Gender		1	0.2053	0.1106	0.5386
LG10 Months Inactive 12 mon	Transformed Months Inactive 12 mon	1	0.1951	0.2190	1.1227
LG10 Customer Age	Transformed Customer Age	0	0.0000	0.0000	.
LG10 Credit Limit	Transformed Credit Limit	0	0.0000	0.0000	.
LG10 Months on book	Transformed Months on book	0	0.0000	0.0000	.
Dependent count		0	0.0000	0.0000	.
Card Category		0	0.0000	0.0000	.
LG10 Avg Utilization Ratio	Transformed Avg Utilization Ratio	0	0.0000	0.0000	.
LG10 Contacts Count 12 mon	Transformed Contacts Count 12 mon	0	0.0000	0.0000	.
LG10 Avg Open To Buy	Transformed Avg Open To Buy	0	0.0000	0.0000	.
Education Level		0	0.0000	0.0000	.
Marital Status		0	0.0000	0.0000	.
Income Category		0	0.0000	0.0000	.

Gradient Boosting



Model Accuracy:
93.39%

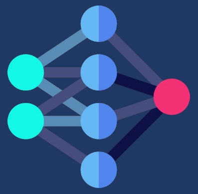
Properties

a. Ensembled Method

b. Combination of Multiple Weak Learners

c. Uses Decision tree as base model

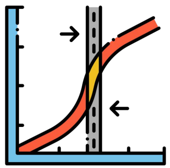
Neural Network



a. Non-linear decision boundaries
b. High model capacity
c. Can automatically learn features from the data

```
Model Accuracy:
92.43%
```

Logistic Regression



a. Linear decision boundaries
b. Simple and interpretable
c. Fast and efficient training and prediction times

```
Model Accuracy:
90.35%
```

Hidden Patterns Found

Sl.	Hidden Patterns	Phase
1	Most Customer Age range is between 35-54	Explore
2	The number of months the customer has been a credit card holder was 30.2 to 38.8.	Explore
3	The client transaction amount was 1000 to 2500 range.	Explore
4	Most of the customer majority is in the blue category	Explore
5	Most of the customers have at least "Graduate" education level.	Explore
6	Total Change of Amount in Quarter 4 to Quarter 1 is having 28 times the odds of having customer attrition than existing customers.	Model
7	Average Open to Buy is having 24 times the odds of having customer attrition than existing customers.	Model

Conclusion

We applied SEMMA methodology to mine impactful insights from Credit Card Customer Dataset.

We sampled the data into SAS Enterprise Miner to begin the mining process, and Talend was used to clean it.

We have developed four distinct models to forecast customer churn by analyzing customer demographics and predicting whether the customer churn or not.