WQD7005 DATA MINING ASSIGNMENT PART 2

CREDIT SCORE CLASSIFICATION

GROUP 1

YOUTUBE LINK: https://youtu.be/YrV0HPwoJUg

22056764	Devayani A/P Balkrishnan
S2195613	Li Xin Qi
S2155659	Lim Yu Xuan
S2192763	Navaneeta A/P P Shanmugam

Presented by: Li Xin Qi

CONTENT

Presented by: Li Xin Qi

PROJECT RECAP

Credit Score Classification Model

- Reflect individual's creditworthiness
- Allow financial institution to:
  - optimize capability in determining individual's risk level
  - minimize credit risk and default cost
  - increase sales and overall expected profits

SEMMA

01 Sample

02 Explore

03 Modify

04 Model

05 Assess

Presented by: Li Xin Qi

OBJECTIVES

To explore the key factors that have significant impact on an individual's credit score.

To build a credit score classification model.

To assess various models and determine the best model in classifying customers' credit score.

Presented by: Li Xin Qi

MODIFY

Pre-Modifying via Base SAS → **Replacement Value** → **Dropped Old Variables** → **Imputation via Median/ Count** → **Log10 Transformation**

Presented by: Lim Yu Xuan

MODIFY

New Dataset Variable Type and Role Summary

Role	Variable	Count
Input	Nominal	4
	Interval	17
	Ordinal	1
	Binary	9
Target	Binary	1

Presented by: Lim Yu Xuan

DATA EXPLORATION

Characteristics

Low Spending with Small Value Payments

Scientist

Correlation Matrix

Annual Income & Monthly In Hand Salary

Loan & Total Estimated Monthly Instalment per month

Interest Rate and & Credit History Age

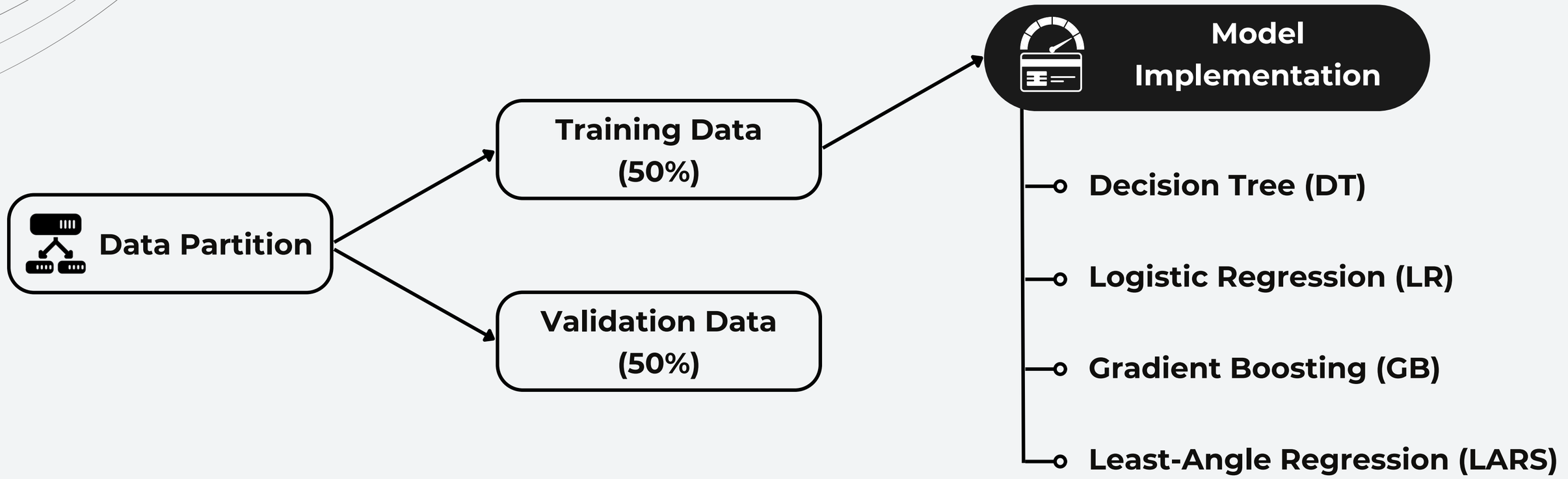
Chi-Square Plot

Credit Mix

Payment of Minimum Amount

Presented by: Lim Yu Xuan

MODEL



Presented by: Navaneeta A/P P Shanmugam

DECISION TREE (DT)

BACKGROUND & PROPERTIES

Decision Tree is

- a supervised learning technique that commonly used to solve classification problem.
- tree-like structure with root nodes, interval nodes, leaf nodes and branches.

Properties:

- Significance level set to 1.0.
- Maximum depth and leaf size set to 10.

DECISION RULES & VARIABLE IMPORTANCE

- 41 decision rules and 81 nodes.
- Root Node : Interest rate (highest information gain).
- Total EMI per month has the highest number of splitting.
- If the interest rate is lesser, then the credit score is 'standard to poor'.
- Top 3 important variables from high to low: Interest_Rate, Credit_Mix, Month

CLASSIFICATION RESULTS

Training Set Outcome:
Misclassification Rate - 0.1521

Validation Set Outcome:
Misclassification Rate - 0.1599

Presented by: Navaneeta A/P P Shanmugam

LOGISTIC REGRESSION (LR)

BACKGROUND & PROPERTIES

Logistic Regression is

- a statistical model used in classification problem and predictive analytics.
- where dependent variables bounded between 0 and 1 to produce a probabilistic outcome.

Properties:

- Polynomial terms set as Yes.
- Polynomial degree of 2.

CLASSIFICATION RESULTS

Training Set Outcome:
Misclassification Rate - 0.1783

Validation Set Outcome:
Misclassification Rate - 0.1858

Presented by: Navaneeta A/P P Shanmugam

GRADIENT BOOSTING (GB)

BACKGROUND & PROPERTIES

Gradient Boosting is

- a technique where each decision tree reduces the error of the previous decision tree progressively by predicting its error.
- where sequential method is used in prediction.

Properties:

- N Iterations set to 700.
- Maximum depth and reuse variable set to 3.

CLASSIFICATION RESULTS

Training Set Outcome:
Misclassification Rate - 0.0717

Validation Set Outcome:
Misclassification Rate - 0.1538

Presented by: Devayani Balkrishnan

LEAST-ANGLE REGRESSION (LARS)

BACKGROUND & PROPERTIES

Least-Angle Regression is

- an algorithm used in regression for high-dimensional data.
- similar to forward selection but sometimes may be more accurate.

Properties:

- Variable Selection Method set to LASSO.
- Model Selection Criteria set to Cross Validation.
- CV Fold set to 200.

CLASSIFICATION RESULTS

Training Set Outcome:
Misclassification Rate - 0.1856

Validation Set Outcome:
Misclassification Rate - 0.1833

Presented by: Devayani Balkrishnan

ASSESS

Models	DT	LR	GB	LARS
Misclassification Rate	0.1599	0.1858	**0.1538**	0.1833
Accuracy	84.01%	81.42%	**84.62%**	81.67%
Precision	0.8569	0.8327	**0.8778**	0.8381
Recall	**0.8166**	0.7864	0.8044	0.7850
F1-Score	0.8362	0.8089	**0.8395**	0.8107
AUC	2nd	3rd	**1st**	4th

Presented by: Devayani Balkrishnan

CONCLUSION

Sample

Stratified equal-sized sampling using sample node on the connected data source

Explore

Univariate, Bivariate and Multivariate Analysis to investigate data type error and explore relationships between variables and target

Modify

Pre-modifying dataset using Base SAS &
Replace, drop, impute and Log 10 transformation to prepare data for modelling

Model

Partition data into training and validation
&
Build DT, LR, GB and LARS credit score classification model

Assess

Assess models based on metrices like precision, recall, accuracy, misclassification rate, F1-score and AUC

Presented by: Li Xin Qi

THANKS FOR LISTENING