CDS503 MACHINE LEARNING ACADEMIC SESSION: SEMESTER 1, 2019/2020

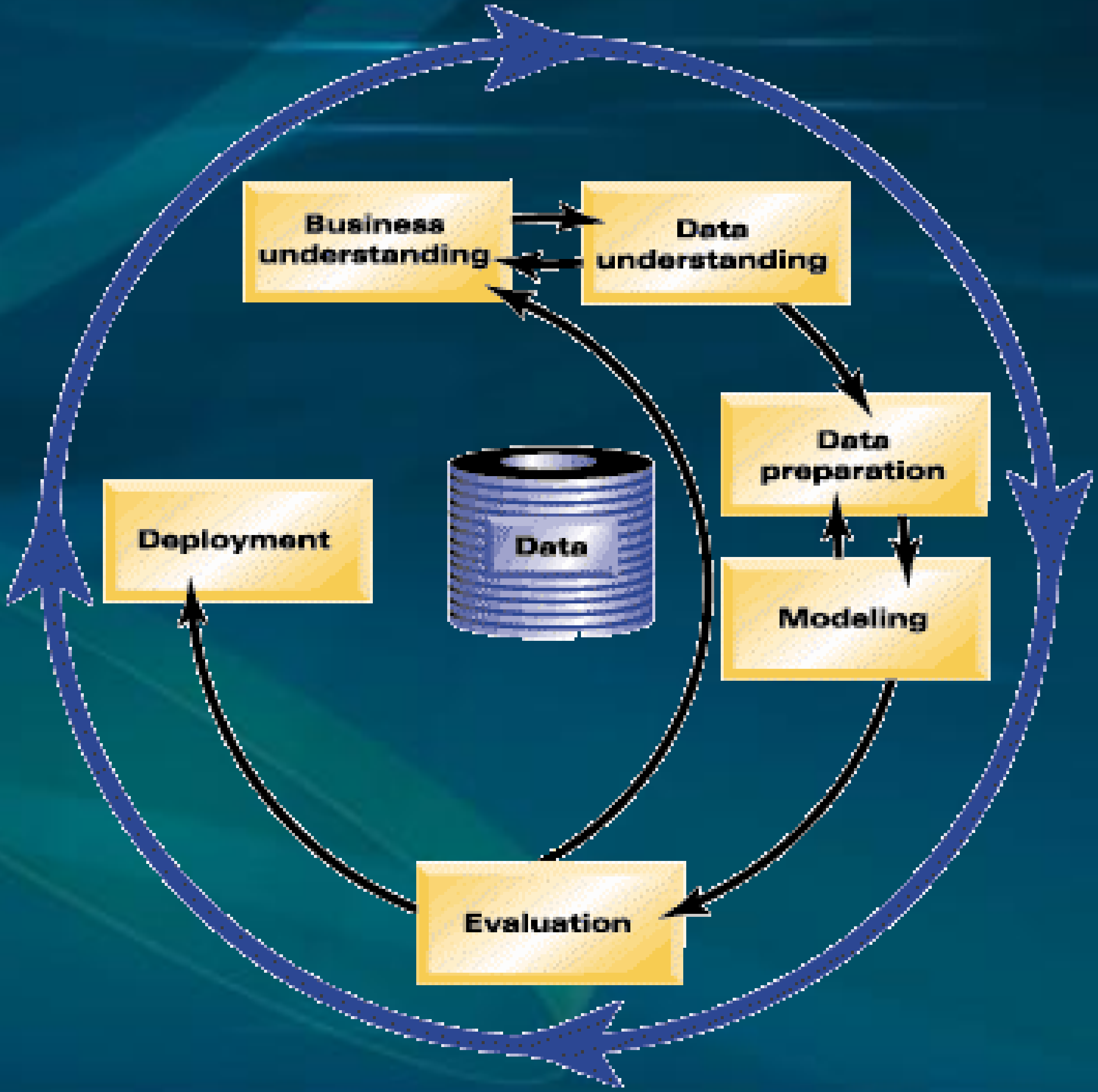
SCHOOL OF COMPUTER SCIENCES, USM, PENANG

GROUP PROJECT
VEHICLE LOAN DEFAULT PREDICTION

VEERAYEN MOHANADAS	P-COM0136/19
CHAN HUAN YANG	P-COM0068/19
LOGANATHAN MUNIANDY	P-COM0115/19

Choice of approach

- CRISP-DM methodology
  - Business understanding
  - Data understanding
  - Data preparation
  - Modeling
  - Evaluation
  - Deployment



Problem statement

- Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates. The need for a better credit risk scoring model is also raised by these institutions.

Business objectives

- Business Goal - To maximize bank profit by ensuring those clients capable of repayment are not rejected and minimize the bank loses by identifying those clients with a high risk of defaulting loan.

Business understanding

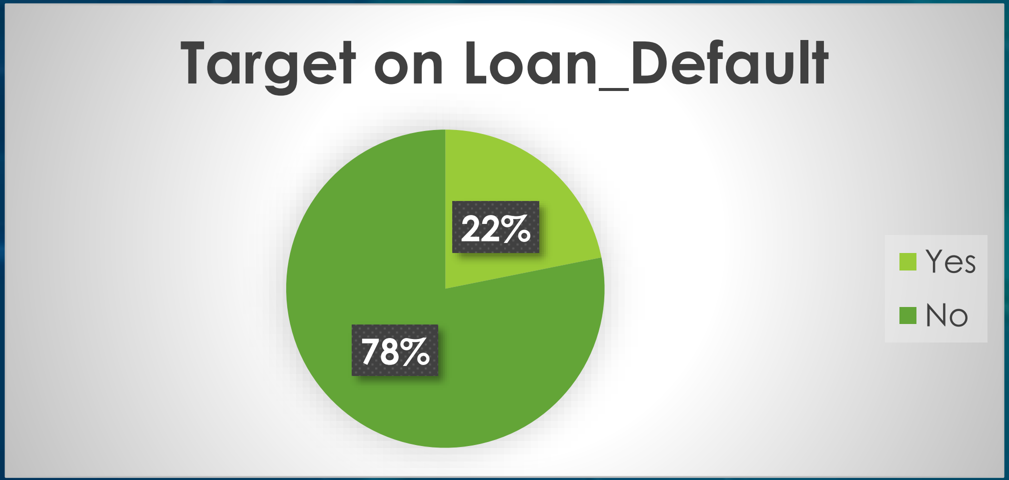
Describing data

- The dataset used in this project is an extract from a real-world dataset. It can be downloaded directly from Kaggle via https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-prediction.
- Dataset consists of 233154 records of loaner/borrower's data with the target (loan_default) describing the default status using 40 features
- The target is binary. It is a binary classification problem

Data understanding

Describing data …

- Our training dataset is imbalanced as the ratio of loan_default-0 to loan_default-1 instances is 177520: 49645 or more concisely 7:2
- “Employment. Type" showing a total of 7661 records of missing value
- All the numerical features fail the Shapiro-Wilk Test. They do not look Gaussian



Data understanding

Handling Data

- Cleaned Employment.Type feature (7661 missing value)
- Construct a new feature "Age" which is easier for interpretation to replace the feature "Date.of.Birth".
- Data consolidation,

  e.g. "PRI.NO.OF.ACCTS" and feature "SEC.NO.OF.ACCTS“ consolidated to new feature called "NO.OF.ACCTS“ and then dropped the earlier two.
- Handle imbalanced data by down-sampling majority class (class 0)
- Principle Component Analysis for Feature selection

Data preparation

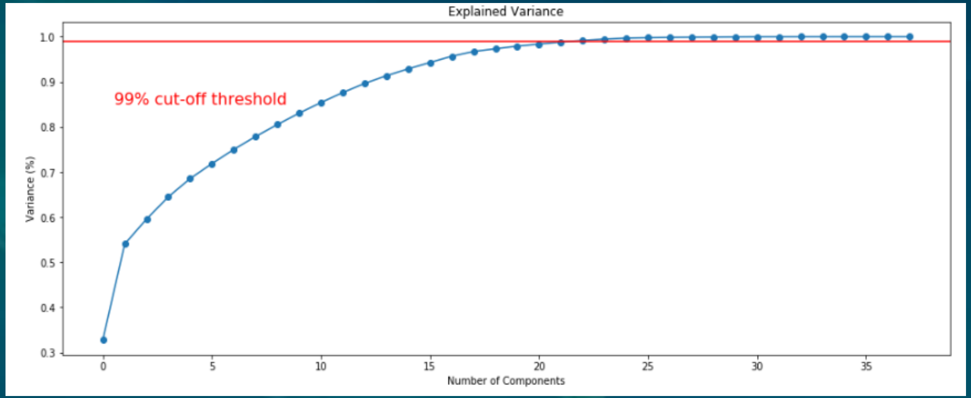
Algorithms to model

- Linear Discriminant Analysis
- Naive Bayes
- Logistic Regression
- Decision Tree
- Support Vector Machine
- Random Forest

Modeling

Experimental setup

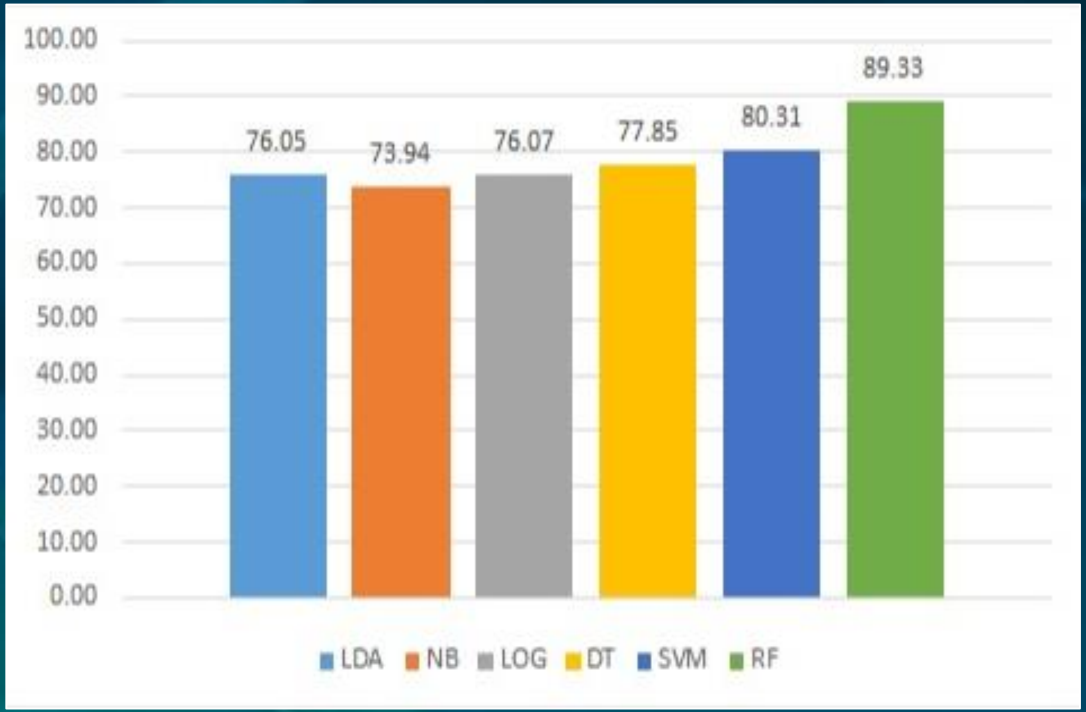
- Split the dataset into the train set and test set.
- Down sample majority class to make it balance
- Feature Scaling on both training and testing sets
- Apply Principal Component Analysis (PCA)
- Build 6 different models
- Evaluate the model



Modeling

Modeling metric and result

- AUCROC used as performance metric as measure the separability.
- It tells how much model is capable of distinguishing between classes. Higher the AUC, better at distinguishing between loan defaulter or non-defaulter.

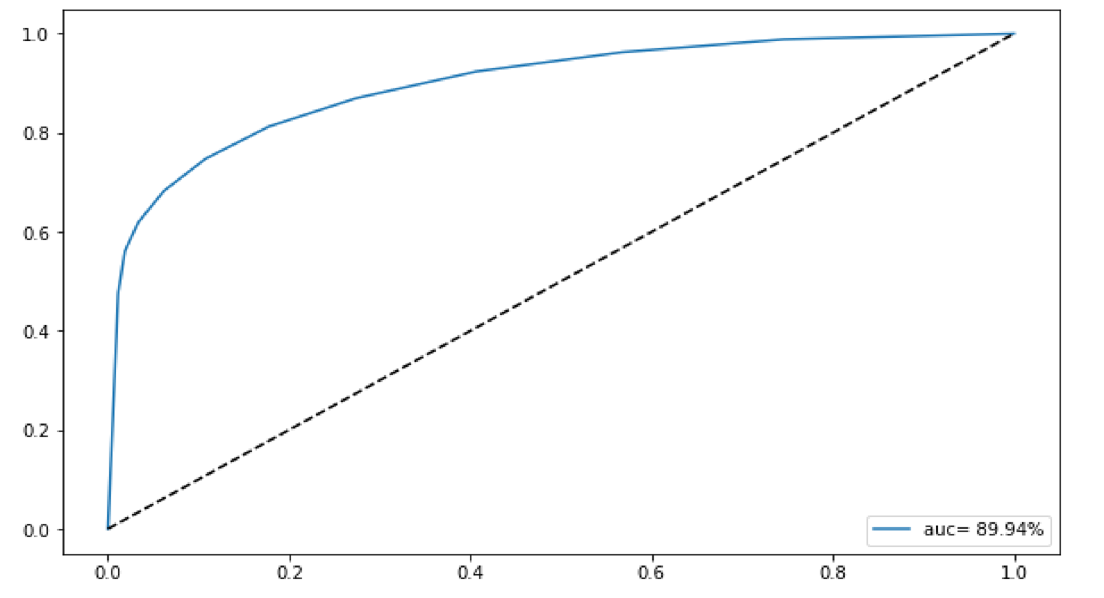


Random Forest

Modeling

Test evaluation

- AUC scored at 89.94%
- Accuracy at 84.51%



AUC at 89.94%

	Predicted 0	Predicted 1
Actual 0	18744 (TN)	2261 (FP)
Actual 1	2563 (FN)	7583 (TP)

Confusion matrix

Evaluation

Conclusion and Future Work

- Overall, the Random Forest algorithm gave the highest performance metric result.
- Decision based algorithm like Random Forest could provide clear structure on what basis on is given loan or rejected loan. This is very useful for financial institutions especially when there are subject to audit.
- The future work will focus more on using more advanced machine learning algorithms such as Deep Learning and Neural Network.

Thank you !!!