Foundation in Data Analytics II

Group Assignment

April 6, 2020

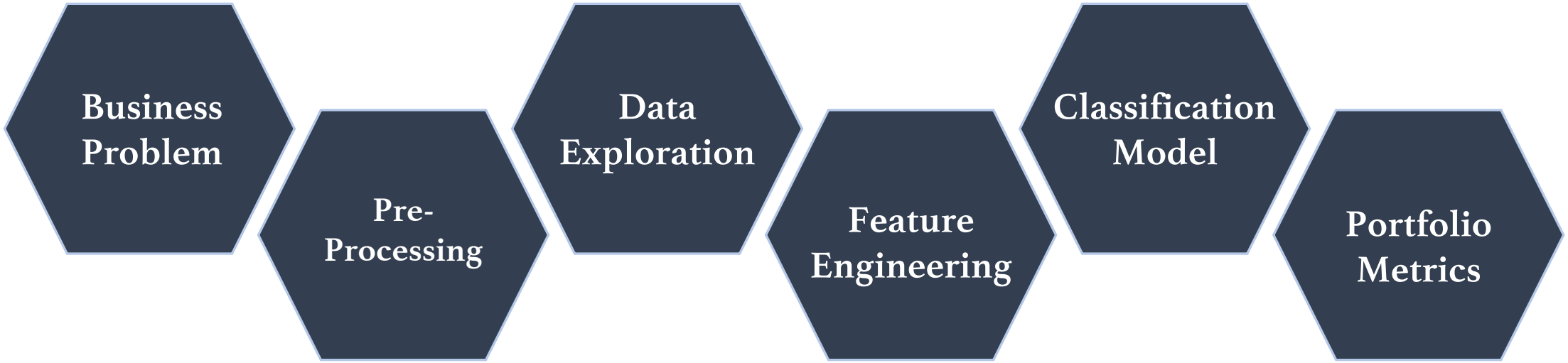
Nicholas David Gabriele - A0206490J

David Raj - A0056525E

Samir Swapna Shukla - A0206512U

Jonathan Simon Wagner - A0152784X

Agenda


Business Problem
Pre-Processing
Data Exploration
Feature Engineering
Classification Model
Portfolio Metrics

Business Problem

Predict the next month's default probability of Taiwanese customers on their credit cards. Customers are from different genders, educational backgrounds and live in different social environments. They were initially granted a different level of credit.

Expected Advantage

- Assist struggling customers with individual measures
- Provide risk management metrics to officers governing Value at Risk
- Refine initial credit card issuance process

Pre-Processing

Data Cleansing

Payment Made
Relabeled -2, 0 to -1
(pay duly)

Education
Relabeled 0, 5, 6 to 4
(others)

Marriage
Relabeled 0 to 3
(others)

No Missing or NAN values found in dataset

Data Description

	Limit balance	default
count	30,000	30,000
mean	167,484	0.221
std	129,747	0.415

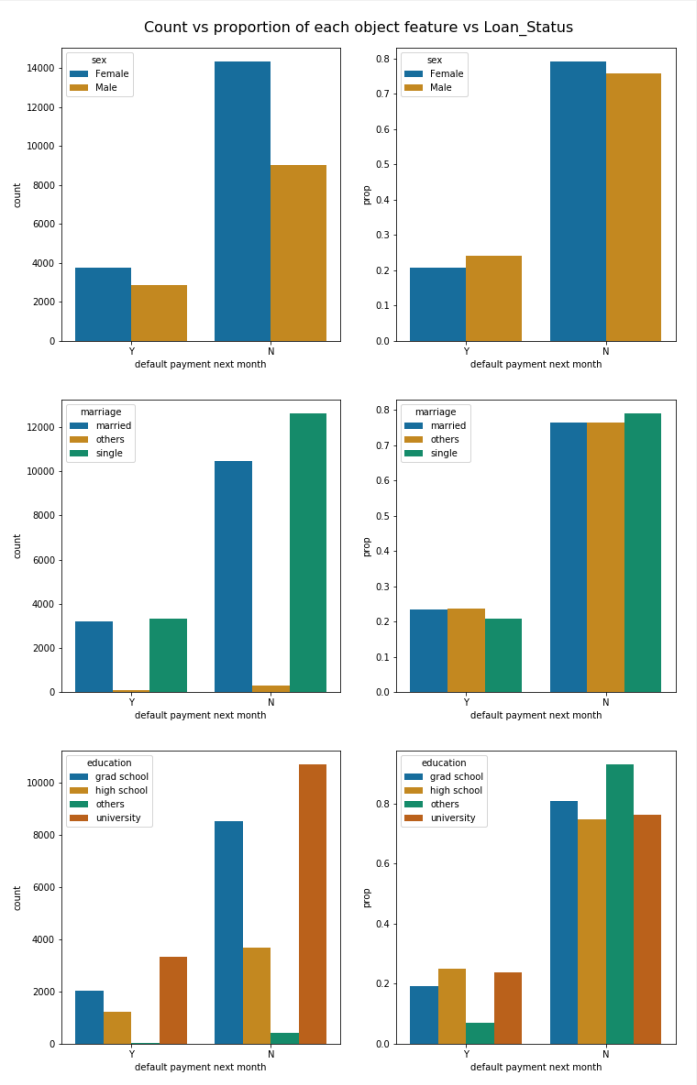
30,000 Credit Card clients

Large standard deviation of limit balance

22.1% of clients are defaulting on their payments next month

Data Exploration

Data Visualization

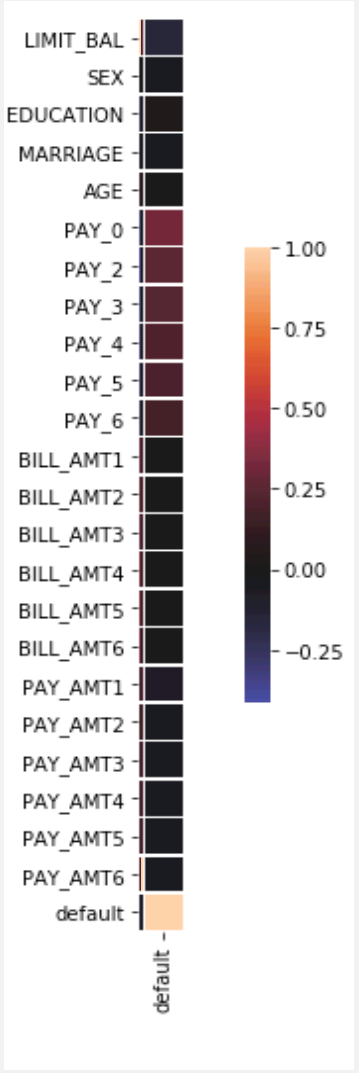


Distribution of gender is even between male and female, indicating that gender may not be a favorable factor

Distribution of marriage status is even all categories, indicating that marriage status may not be a favorable factor

Distribution of education status is quite varied, indicating that education status could be a deciding factor.

Pearson Correlation



The variables indicating how many months payment has not been made seems to show the most correlation with the default status. Hence, we shall keep them during the model training.

Remaining original features do not provide much correlation at first sight.

There is the need to explore new engineered features to provide the model with significant relationships to work with.

Feature Engineering

1) Age Binning

- According to Taiwan's education and employment system, population can be binned into 4 groups (Studying, Junior Executive, Senior Executives, Retired)
- Assumption is that certain groups with income would be less likely to default

2) Average Credit Utilization

- Assumption is that higher the ratio, it would be harder to clear the loan

$$\frac{1}{6}\left(\sum_{k=1}^{6}\frac{Amount\ Bill_k}{Limit\_Bal}\right)$$

3) Delinquent Payment Indicator (DPI)

- Flag if a person has ever delayed their payment

$$1\ if\ Repayment\ Status_k > 0\ for\ k \in [1,6]$$
$$otherwise\ 0$$

4) Rehabilitation Indicator

- Flag if a person has CONTINUOUSLY delayed their payment

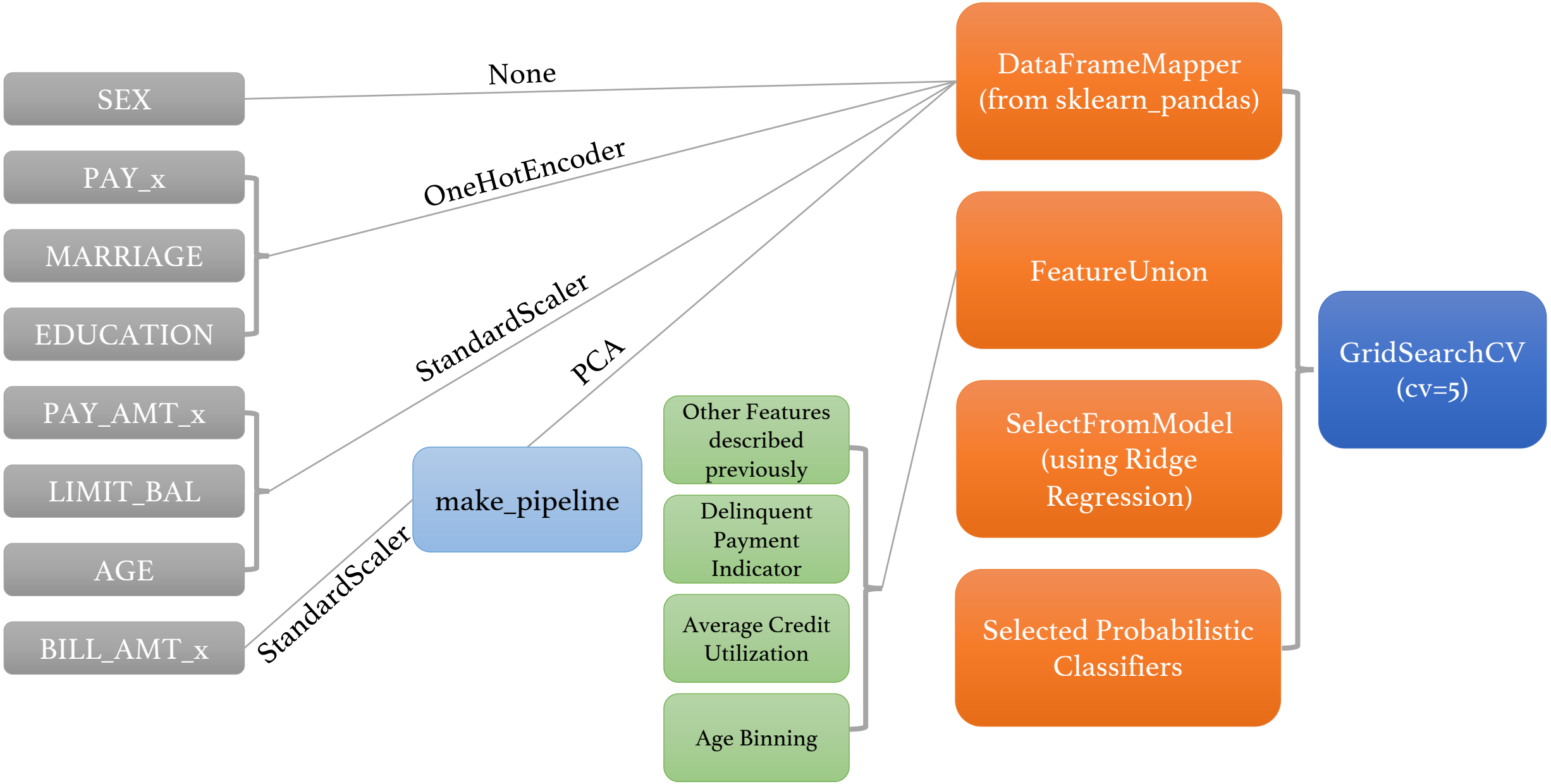
$$1\ if\ Repayment_{\ t+1} - Repayment_{\ t} > 0\ for\ t \in [1,5],$$
$$otherwise\ 0$$

5) Proportion Time Delinquent

- A ratio of how often a person has delayed their payment

$$\frac{1}{6}\left(\sum_{k=1}^{6} DPI_k\right)$$

ML Pipeline


SEX
PAY_x
MARRIAGE
EDUCATION
PAY_AMT_x
LIMIT_BAL
AGE
BILL_AMT_x
None
OneHotEncoder
StandardScaler
StandardScaler
PCA
make_pipeline
Other Features described previously
Delinquent Payment Indicator
Average Credit Utilization
Age Binning
DataFrameMapper (from sklearn_pandas)
FeatureUnion
SelectFromModel (using Ridge Regression)
Selected Probabilistic Classifiers
GridSearchCV (cv=5)

Model Selection

Considered models (only probabilistic models due to business application):
MLPClassifier, LightGBM, Logistic Regression, XGBoost, CATBoost

	MLPClassifier			**LightGBM**			**Logistic Regression**		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0	**0.84**	0.96	0.89	0.83	0.96	0.89	0.83	0.97	0.89
1 (Default)	**0.77**	0.36	0.48	0.71	0.35	0.47	0.74	0.32	0.45
Macro Avg	**0.77**	**0.66**	**0.69**	0.77	0.65	0.68	0.78	0.64	0.67
Weighted Average	0.81	0.82	0.8	0.81	0.82	0.8	0.81	0.82	0.79
Test Accuracy	**0.82**			**0.82**			**0.82**		
AUC	**Train: 0.7605, Test: 0.7738**			**Train: 0.7771, Test: 0.7738**			**Train: 0.7583, Test: 0.7738**		

The accuracy metrics suggest that despite the similar test accuracy, the MLP Classifier performed better in terms of F1- score and has a higher Recall score.

Higher Recall Score suggest that the model is less likely to label a false positive (less likely to label a “Default” when it is not)

Variables

Significant Variables

Main Variable	Variation	Option
Repayment Status	September	-1
		1
		2
		3
		4
		6
		7
	August	1
		4
		5
		6
		7
	July	4
		5
		6
	June	1
		4
		5
		6
	May	4
		7
	April	-1
		3
		6
		8
Payment Amount	August	numerical
Education Status	Education Status	Others
DPI	DPI	T/F

8 Variables from dataset
1 Variable from Feature Engineering

Sanity Check

With education, only Others is significant as Grad School, University and High School have exact same proportions. (verified in slide 5)

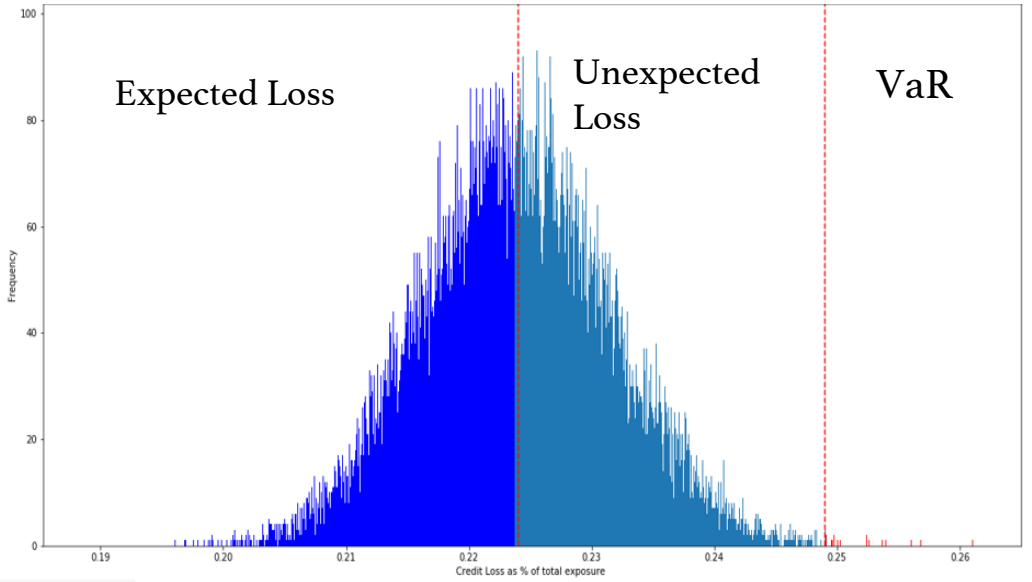
Options for repayment status gets larger towards September. Since this looking at the default the following month (October), more options in months closer to that does impact the outcome.

Portfolio Metrics

Using Probabilistic Default Rates to derive Portfolio Metrics

Case Study of an Institution that buys Credit Card debt to prepare an Asset backed Security
How would it like to make provisions for future losses ?

- **Expected Loss (EL)**
  - PD (output of Probabilistic classifiers)
  - Exposure at Default (Total Bill Amount - Total Payment)
  - Loss given Default = 1 ,no collateral for credit cards
- **Unexpected Loss**
  - Difference between VaR and EL
- **Value at Risk (VaR)**
  - 99.9% confidence level



The credit losses are expressed as a percentage of overall exposure at default.

Using the probability default rates derived from logistic regression for 24000 accounts (train data), and running a Monte Carlo simulation, we obtained:

EL= 22.4% and VaR=24.9%