點對點借貸信用風險預測

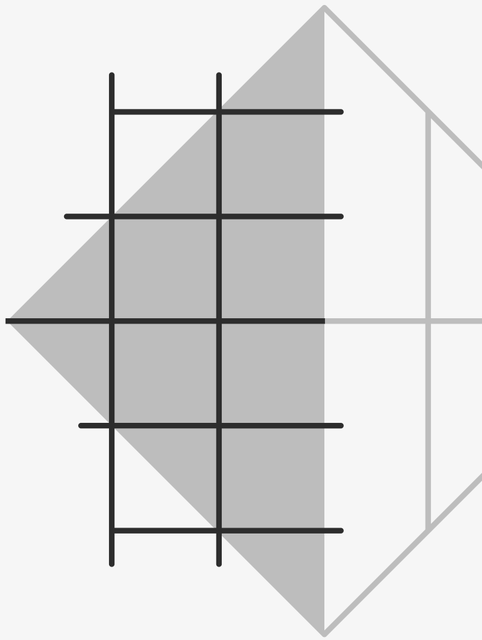
Members: 1091607陳卿雅、1091423郭羽蕎、1091420李婕綾、1090407謝宜庭

2024 June 21

Outline

1. Introduction
2. Method
3. Result
4. Conclusion

Introduction



Objective

Building a Reliable Model

Improving The Accuracy Of Credit Decisions

Increasing Operational Efficiency

Adapting to Market Changes

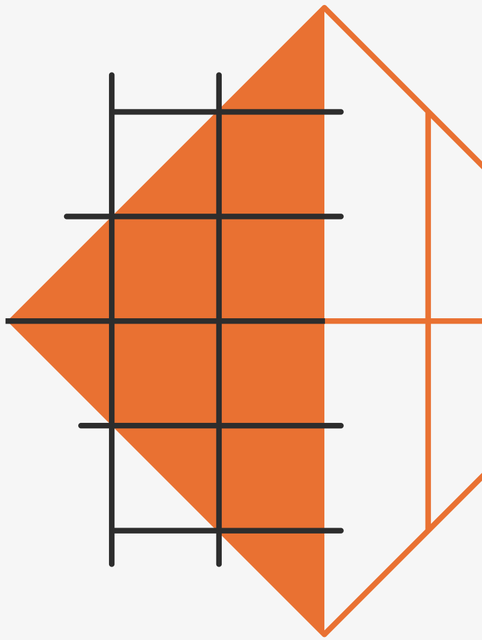
Dataset

- **Lending Club** is a peer-to-peer Lending company based in the US.
- The Lending Club dataset contains complete loan data for all loans issued through the 2007-2015, including the current **loan status** (Current, Late, Fully Paid, etc.) and latest payment information.
- **Features** (aka variables) include credit scores, number of finance inquiries, address including zip codes and state, and collections among others.
- Shape of dataset: **2260668 rows ✳145 columns**

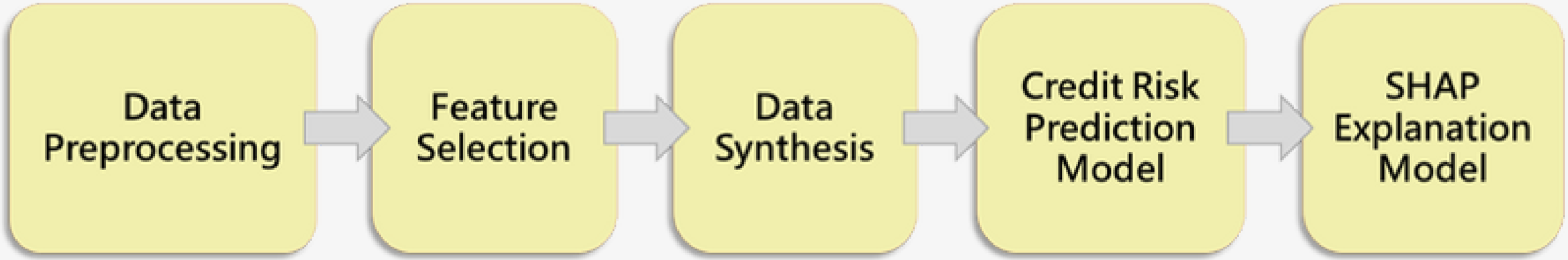
Dataset - Key Features

- **loan_status:** Indicates the current status of the loan (e.g., Current, Late, Fully Paid, Charged Off).
- **loan_amnt:** The amount of money requested by the borrower.
- **funded_amnt:** The total amount funded by investors.
- **term:** The number of months for loan repayment (36 or 60).
- **last_pymnt_amnt:** The amount of the last payment received.
- **debt_settlement_flag:** Indicates if the loan is flagged for debt settlement.
- **num_tl_30dpd:** Number of accounts currently 30 days past due.

Method



Workflow


Data Preprocessing
Feature Selection
Data Synthesis
Credit Risk Prediction Model
SHAP Explanation Model

Data PreProcessing - Raw Data

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	s
0	NaN	NaN	2500	2500	2500.0	36 months	13.56	84.92	C	
1	NaN	NaN	30000	30000	30000.0	60 months	18.94	777.23	D	
2	NaN	NaN	5000	5000	5000.0	36 months	17.97	180.69	D	
3	NaN	NaN	4000	4000	4000.0	36 months	18.94	146.51	D	
4	NaN	NaN	30000	30000	30000.0	60 months	16.14	731.78	C	
...	...	...	...	...	...	...	...	...	...	
2260663	NaN	NaN	12000	12000	12000.0	60 months	14.08	279.72	C	
2260664	NaN	NaN	12000	12000	12000.0	60 months	25.82	358.01	E	
2260665	NaN	NaN	10000	10000	10000.0	36 months	11.99	332.10	B	
2260666	NaN	NaN	12000	12000	12000.0	60 months	21.45	327.69	D	
2260667	NaN	NaN	16550	16550	16550.0	60 months	21.45	451.94	D	

2260668 rows × 145 columns

Data PreProcessing - Delete Columns

將欄位缺失值大於資料總筆數80%的欄位刪除

```
Index(['id', 'member_id', 'url', 'desc', 'mths_since_last_delinq',
       'mths_since_last_record', 'next_pymnt_d', 'mths_since_last_major_derog',
       'annual_inc_joint', 'dti_joint', 'verification_status_joint',
       'open_acc_6m', 'open_act_il', 'open_il_12m', 'open_il_24m',
       'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m',
       'open_rv_24m', 'max_bal_bc', 'all_util', 'inq_fi', 'total_cu_tl',
       'inq_last_12m', 'mths_since_recent_bc_dlq',
       'mths_since_recent_revol_delinq', 'revol_bal_joint',
       'sec_app_earliest_cr_line', 'sec_app_inq_last_6mths',
       'sec_app_mort_acc', 'sec_app_open_acc', 'sec_app_revol_util',
       'sec_app_open_act_il', 'sec_app_num_rev_accts',
       'sec_app_chargeoff_within_12_mths',
       'sec_app_collections_12_mths_ex_med',
       'sec_app_mths_since_last_major_derog', 'hardship_type',
       'hardship_reason', 'hardship_status', 'deferral_term',
       'hardship_amount', 'hardship_start_date', 'hardship_end_date',
       'payment_plan_start_date', 'hardship_length', 'hardship_dpd',
       'hardship_loan_status', 'orig_projected_additional_accrued_interest',
       'hardship_payoff_balance_amount', 'hardship_last_payment_amount',
       'debt_settlement_flag_date', 'settlement_status', 'settlement_date',
       'settlement_amount', 'settlement_percentage', 'settlement_term'],
      dtype='object')
```

刪除的欄位

145 columns

87 columns

Data PreProcessing - Non-numeric Columns

Filtering

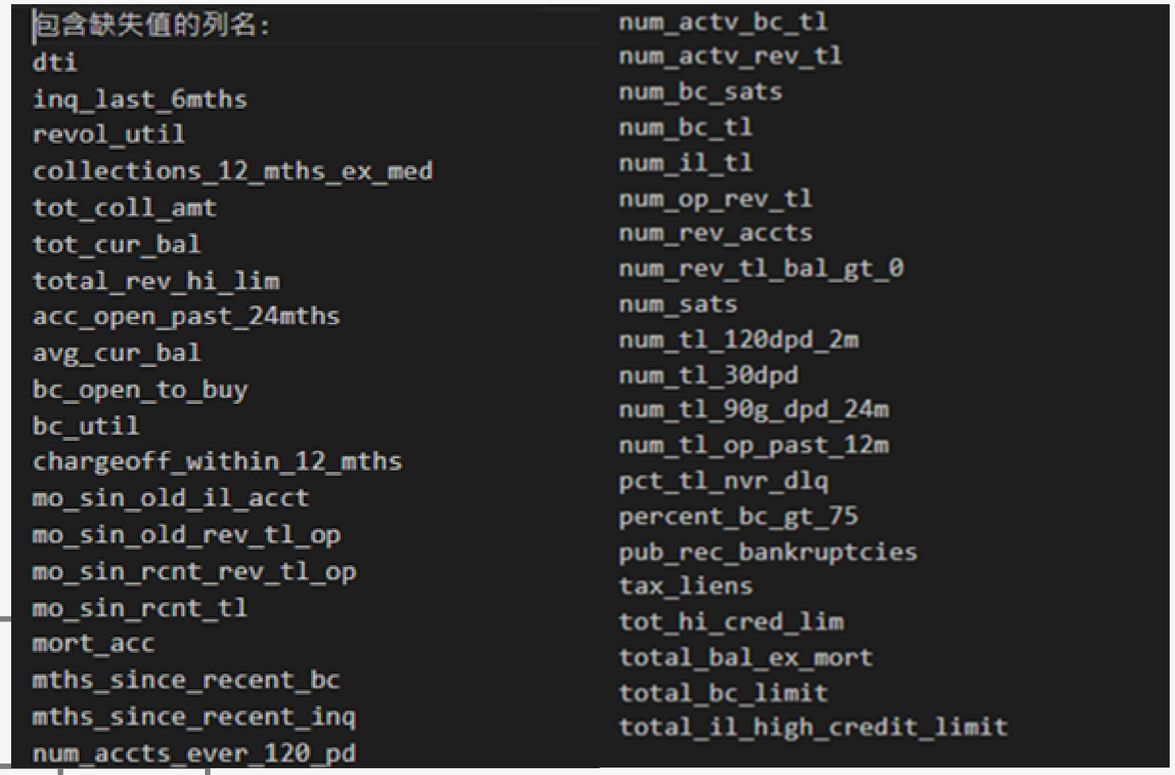
loan_status=[Current, Late, Fully Paid, Charged Off]

Encoding

觀察數值選擇對欄位做Label encoding或是One hot encoding

Data PreProcessing - Impute Missing Value

針對欄位有缺失值的地方進行補值



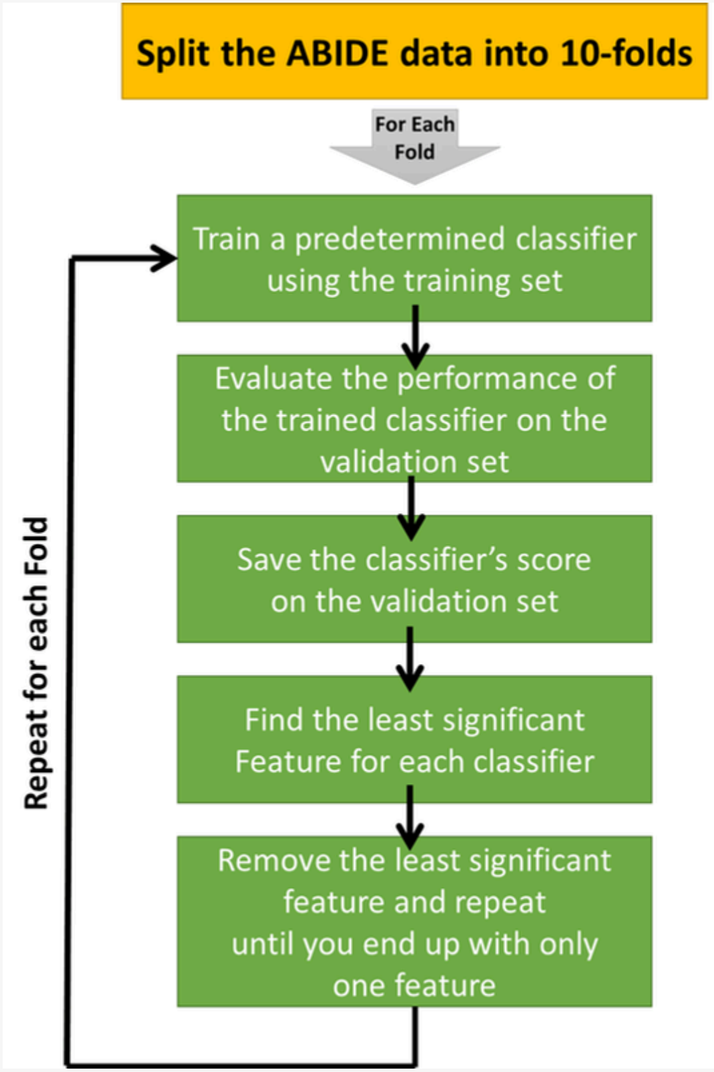
包含缺失值的欄位

fillna(df[].mean)

Feature Selection

RFECV(Recursive Feature Elimination with Cross-Validation)

- RFE combine with KFOLD
- Split data into K-folds and separately execute a RFE until remain one feature
- Get the best feature subset



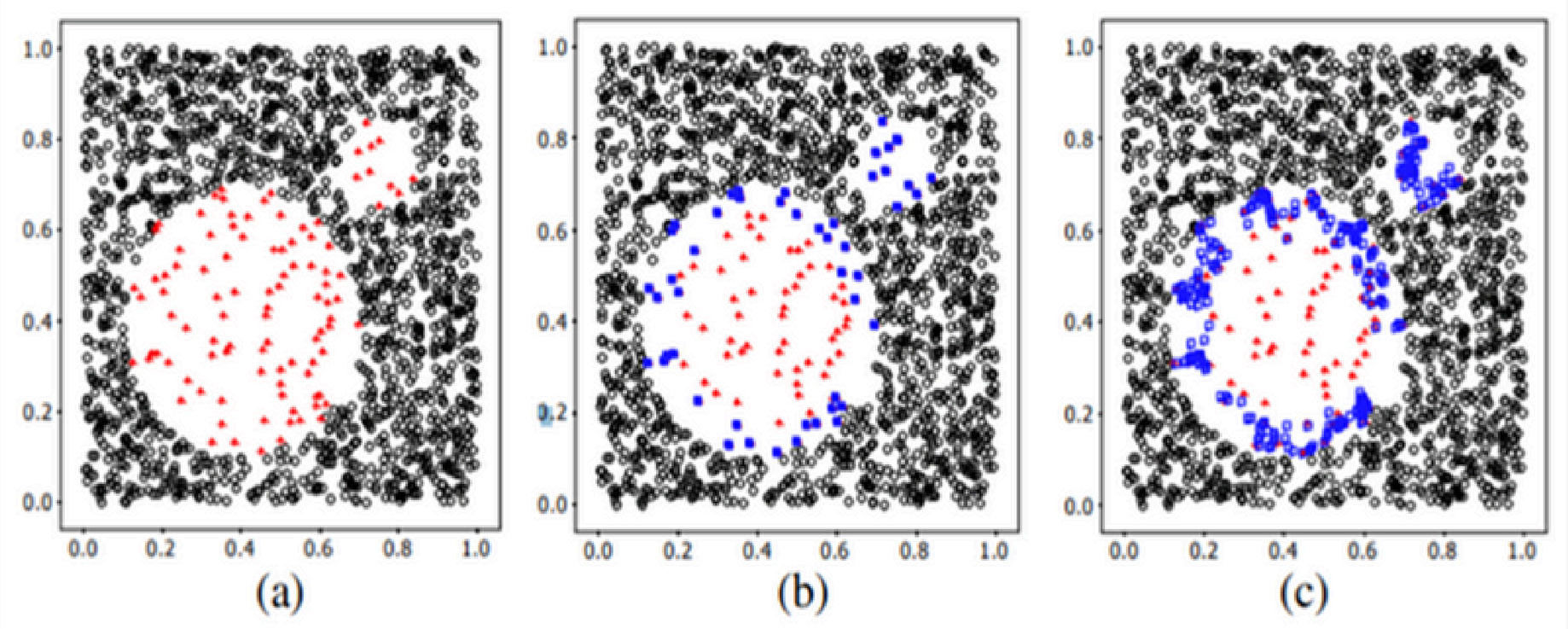
https://www.researchgate.net/publication/357756192/figure/fig1/AS:1114511230672897@1642731592990/The-flowchart-of-The-RFECV-algorithm.ppm

Feature Selection - Top 10 Features

Feature Name	Description	Feature Name	Description
recoveries	Post charge-off gross recovery.	**last_pymnt_amnt**	Last payment amount received.
funded_amnt	The total amount committed to the loan.	**loan_amnt**	The amount of money the borrower requested.
term	The number of months for the loan repayment (36 or 60).	**num_tl_30dpd**	Number of accounts currently 30 days past due.
total_rec_prncp	Total principal received to date.	**funded_amnt_inv**	The total amount funded by investors.
debt_settlement_flag	Indicates if the loan is flagged for debt settlement.	**total_rec_late_fee**	Late fees received to date.

Data Synthesis

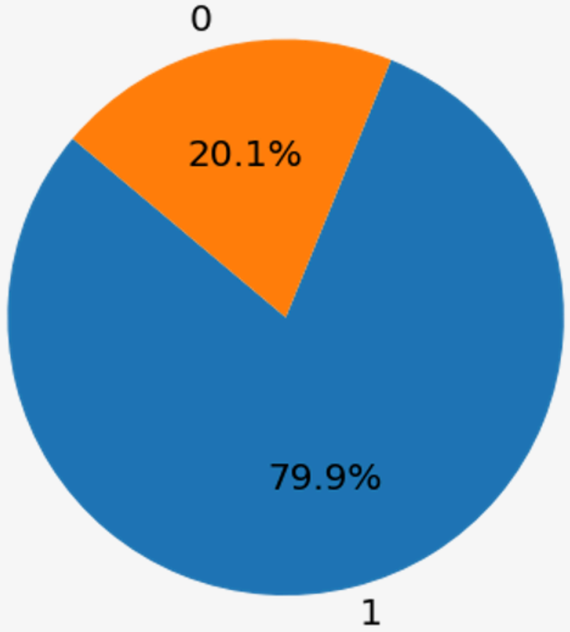
Borderline SMOTE



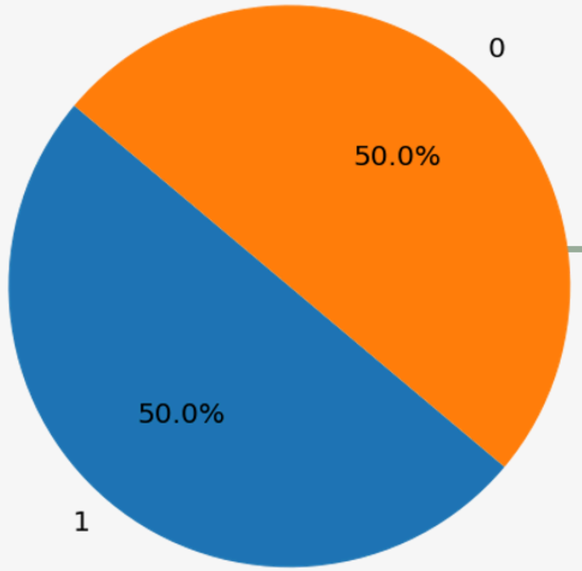
https://lufor129.medium.com/%E8%B3%87%E6%96%99%E7%A7%91%E5%AD%B8-%E4%B8%80-%E8%99%95%E7%90%86%E4%B8%8D%E5%B9%B3%E8%A1%A1%E8%B3%87%E6%96%99%E5%B9%BE%E7%A8%AE%E6%96%B9%E6%B3%95-39c25d06a6cb

Data Synthesis

- 將目前的資料切成訓練跟測試資料，比例8:2(1,042,885:260,722)，再將訓練資料做Borderline SMOTE
- 訓練資料總筆數: 1,042,885 -> 1,667,122



Before Borderline SMOTE



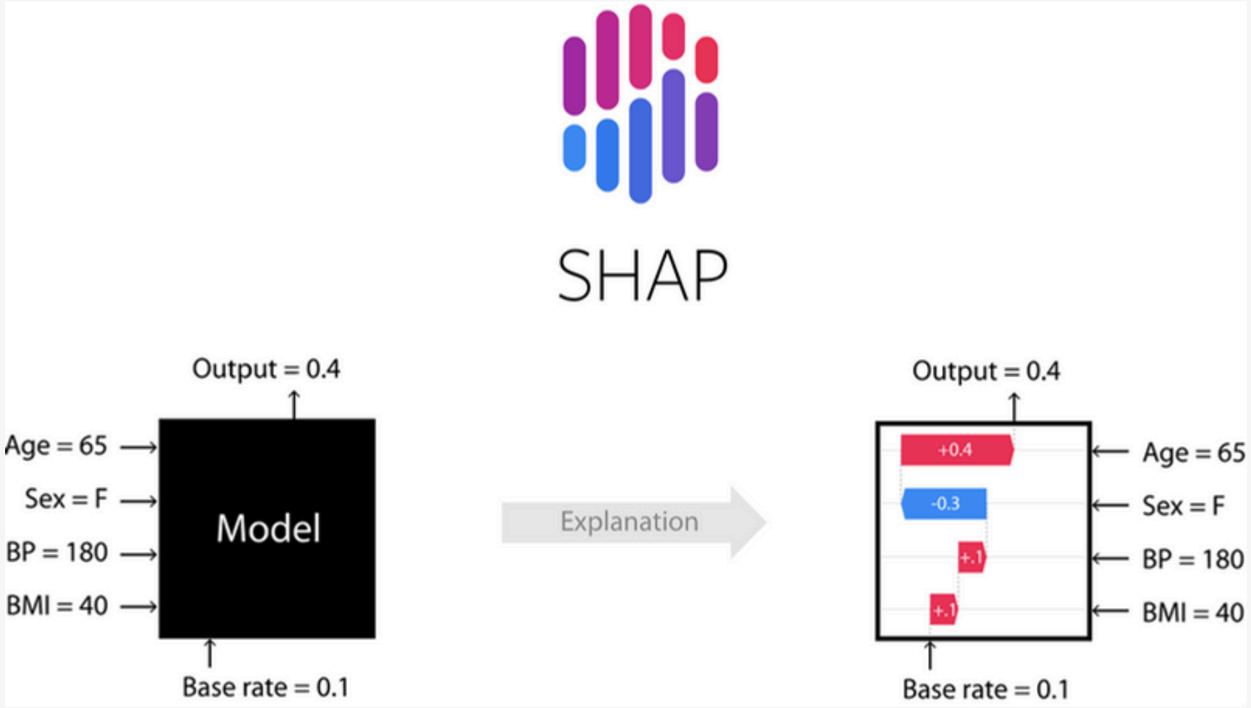
After Borderline SMOTE

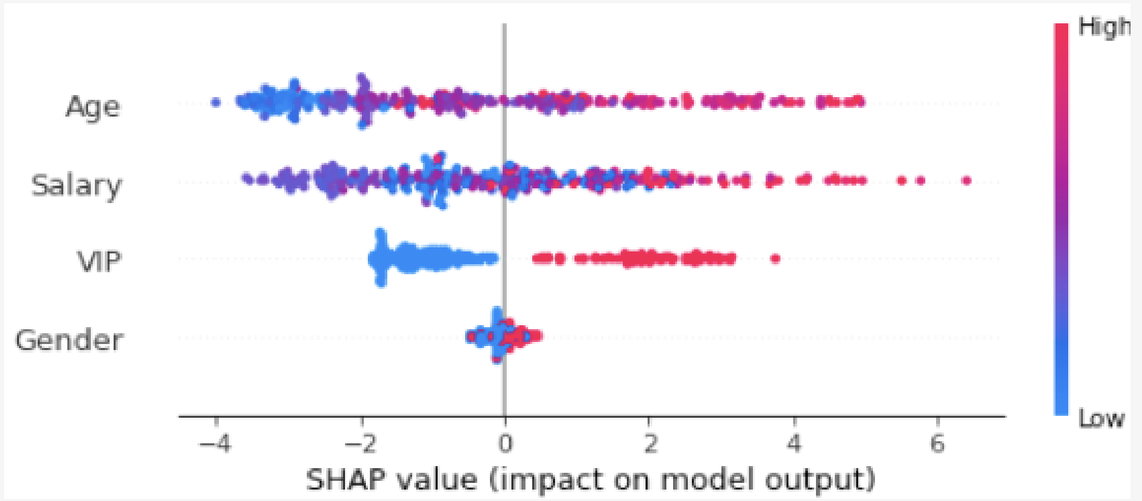
Credit Risk Prediction Model

- DNN
- Random Forest
- Logistic Regression
- CatBoost
- XGBoost

SHAP Explanation Model

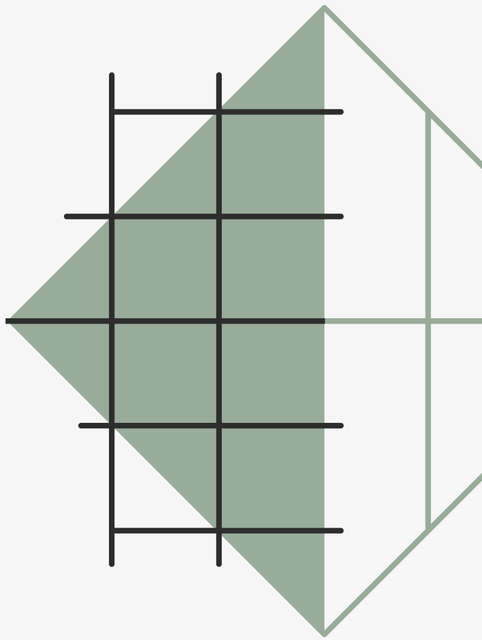
Calculate shapley value of each feature to evaluate the impact of features on the final predicted value.

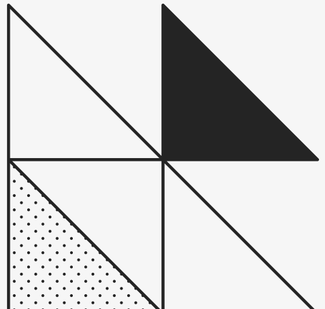




https://ithelp.ithome.com.tw/articles/10276922

Result





Evaluation

		Predicted Condiction	
		Yes	No
Actual Condiction	Yes	TP	FN
	No	FP	TN

Confusion Matrix

無風險借貸: 208391 / 高風險借貸: 52331

- **TP**: 無風險借貸被正確預測
- **FN**: 無風險借貸被錯誤預測
- **TN**: 高風險借貸被正確預測
- **FP**: 高風險借貸被錯誤預測

DNN

		Predicted Condiction	
		Yes	**No**
Actual Condiction	**Yes**	208039	352
	No	20	52311

Confusion Matrix

Random Forest

		Predicted Condiction	
		Yes	**No**
Actual Condiction	**Yes**	208387	4
	No	78	52253

Confusion Matrix

Logistic Regression

		Predicted Condiction	
		Yes	**No**
Actual Condiction	**Yes**	208391	0
	No	7	52324

Confusion Matrix

CatBoost

		Predicted Condiction	
		Yes	**No**
Actual Condiction	**Yes**	208389	2
	No	121	52210

Confusion Matrix

XGBoost

		Predicted Condiction	
		Yes	**No**
Actual Condiction	**Yes**	208390	1
	No	55	52276

Confusion Matrix

Model Result Conclusion

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

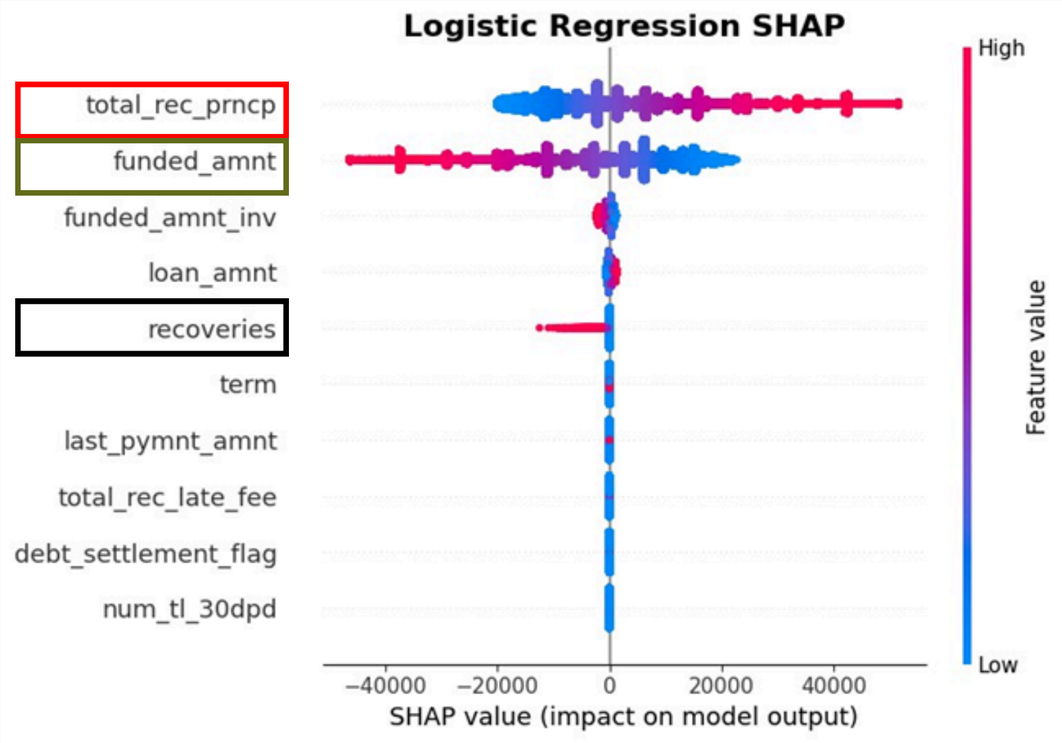
$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Acc: Logistic Regression
- Recall: Logistic Regression
- Precision: Logistic Regression
- F1-score: Logistic Regression

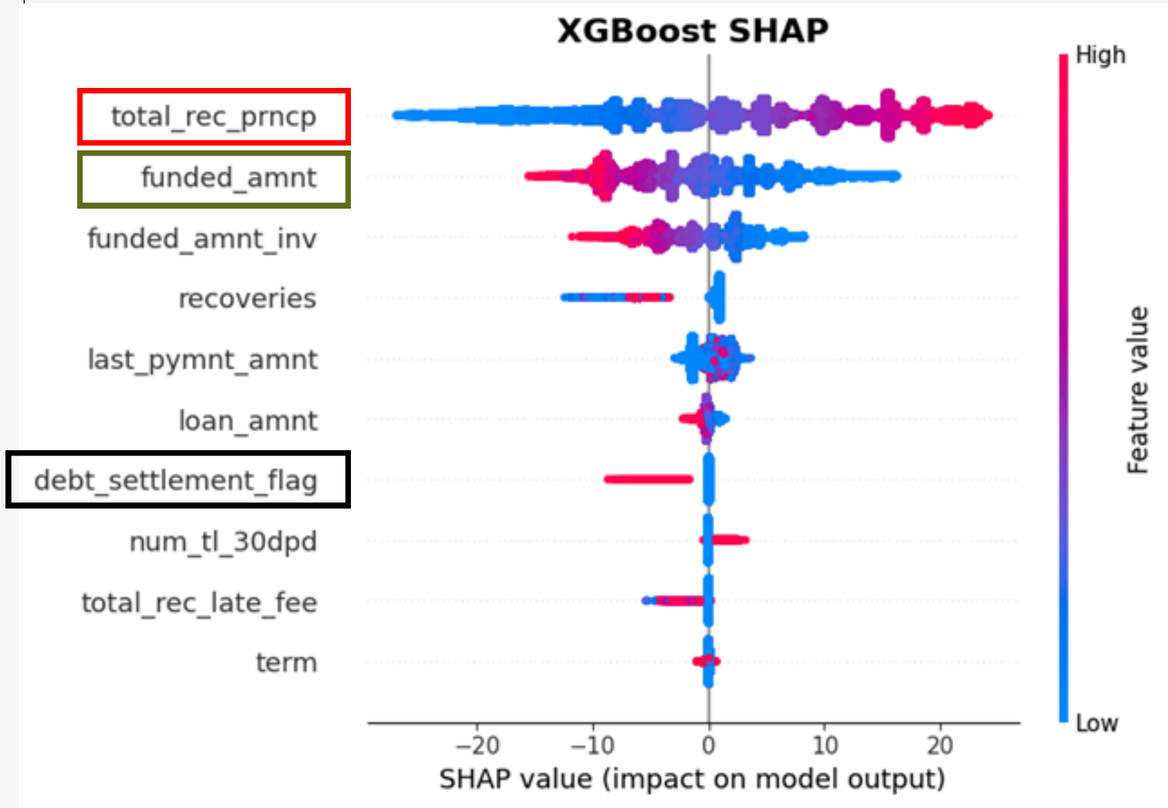
SHAP Explanation - Logistic Regression



Logistic Regression

- **total_rec_prncp:** 至今收到的本金，代表數值越大越可以借錢，越小代表會欠錢。
- **funded_amnt:** 承諾的貸款總額，數值越大代表會欠錢，數值越小越可以借錢。
- **recoveries:** 扣除回收後的費用，數值越大代表越可能會欠錢。

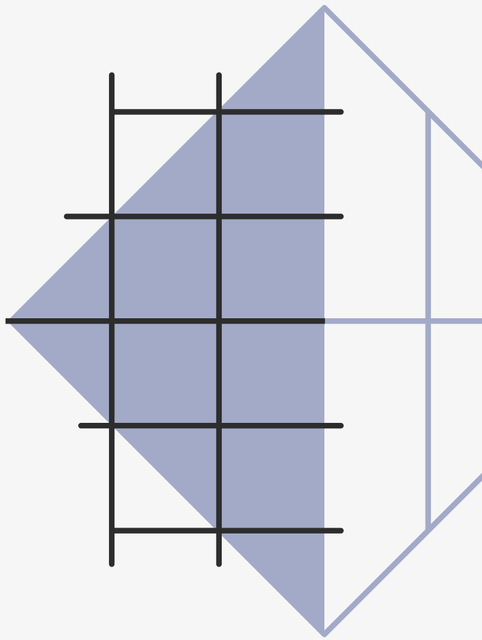
SHAP Explanation - XGBoost



XGBoost

- **total_rec_prncp:** 為至今收到的本金，數值越大越可以借錢，越小代表會欠錢。
- **funded_amnt:** 承諾的貸款總額，數值越大代表會欠錢，數值越小越可以借錢。
- **debt_settlement_flag:** 標記已註銷的借款人是否正在與債務結算公司合作，數值越大代表會欠錢。

Conclusion



Conclusion

Improved model performance with RFECV and SMOTE

- Increased correct predictions
- Reduced computation

SHAP explanation model

- Enhanced model transparency
- Improved interpretability, supporting decision-making

Improved credit risk management

- More accurately identify high-risk borrowers
- Lower bad debt rate

Thanks for listening