Stage 3 DataRider

Cristanto Steven Benny Tri Setiawan Ulva Dewiyanti

Skenario Model ML

Setiap masing-masing anggota tim akan mengerjakan salah satu dari skenario yang ada, dimana penerapan algoritma, pemilihan features, dan tuning hyperparameter akan dilakukan pada setiap masing-masing skenario. Berikut skenario yang digunakan pada tahap modeling:

- 1. Data original
- 2. Data setelah outlier dihapus
- 3. Data yang distandarisasi
- 4. Data setelah outlier dihapus dan dinormalisasi/standarisasi

Summary Algoritma ML

Dari semua skenario yang ada, algoritma ML yang digunakan meliputi salah satu atau beberapa dari algoritma berikut:

- 1. Logistic Regression
- 2. KNN
- 3. Decision Tree
- 4. Random Forest
- 5. AdaBoost
- 6. XGBoost

Evaluation Metrics

Dari semua skenario yang ada, metrik yang digunakan adalah AUC (primary), Precision & Recall (secondary), dengan alasan berikut:

Technical Reason:

- 1. Dataset imbalance dengan proporsi hampir mencapai 8:2
- 2. Berfokus untuk mengurangi jumlah false negative dan false positive
- 3. Nilai accuracy menjadi kurang representatif karena banyak data sintetis hasil dari oversampling

Business Side Reasons:

- Tujuan utama dari model adalah untuk memprediksi jumlah default (yang benar-benar default) sebanyak-banyaknya sehingga dapat diberikan tindakan penanganan, sehingga diperlukan memperhatikan recall score
- 2. Namun, kita juga harus memperhatikan debitur yang seharusnya tidak default namun terprediksi default (false positive), karena jika terprediksi default namun tidak, maka debitur akan diberikan tindakan penanganan sebagai seorang default, hal tersebut tentunya dapat memicu ketidaknyamanan debitur tersebut yang dapat menyebabkan komplain atau churn, sehingga score precision juga perlu diperhatikan

Skenario 1: Data Original

Algoritma yang digunakan:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Feature yang digunakan:

Semua feature

Metrik Evaluasi Data Imbalance

					Orig	ginal								Tuni	ng Hype	rparame	eters			
Algoritma	Accu	iracy	Prec	ision	Re	call	Al	JC	F	1	Accu	uracy	Prec	ision	Re	call	A	JC	F	1
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.78	0.78	0.5	1	0	0	0.67	0.66	0	0	0.78	0.78	0.33	0	0	0	0.67	0.65	0	0
KNN	0.82	0.74	0.67	0.35	0.34	0.18	0.83	0.6	0.45	0.24	0.79	0.78	0.62	0.51	0.07	0.06	0.72	0.65	0.12	0.11
Decision Tree	0.99	0.69	0.99	0.33	0.97	0.35	1	0.58	0.98	0.34	0.81	0.76	0.64	0.43	0.29	0.19	0.82	0.65	0.4	0.26
Random Forest	0.99	0.78	0.99	0.52	0.98	0.2	1	0.7	0.98	0.29	0.82	0.78	0.9	0.56	0.23	0.12	0.86	0.72	0.37	0.19
Ada Boost	0.79	0.78	0.61	0.53	0.16	0.13	0.75	0.71	0.26	0.21	0.79	0.78	0.64	0.58	0.15	0.12	0.76	0.72	0.24	0.19
XGBoost	0.8	0.78	0.66	0.56	0.19	0.14	0.78	0.72	0.29	0.23	0.85	0.78	0.85	0.55	0.39	0.2	0.91	0.71	0.54	0.29

Skenario 2: Data setelah outlier dihapus

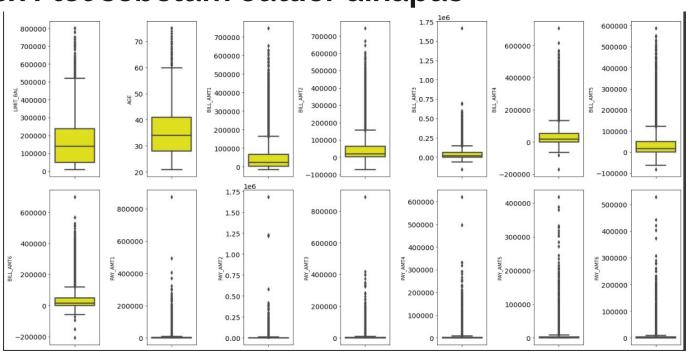
Algoritma yang digunakan:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest

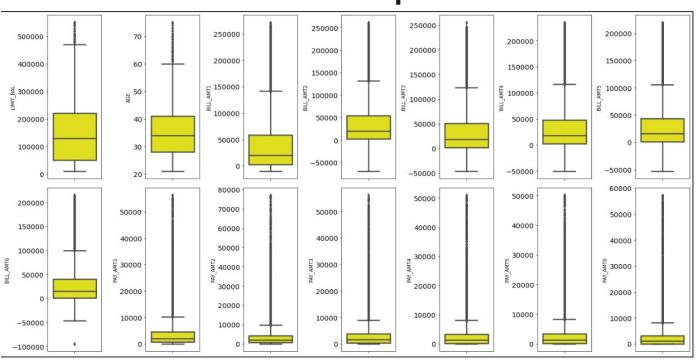
Feature yang digunakan:

LIMIT_BAL, PAY_X, BILL_AMTX, PAY_AMTX

Box Plot sebelum outlier dihapus



Box Plot setelah outlier dihapus



Metrik Evaluasi Data Imbalance

					Orig	inal								Tunii	ng Hype	rparam	eters			
Algoritma	Accu	iracy	Prec	ision	Re	call	Al	JC	F	1	Accı	ıracy	Prec	ision	Re	call	Al	JC	F	1
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.77	0.77	0.5	0.5	0	0	0.67	0.66	0	0										
KNN	0.81	0.74	0.67	0.36	0.35	0.2	0.83	0.6	0.46	0.26										
Decision Tree	1	0.7	1	0.37	1	0.43	1	0.61	1	0.39										
Random Forest	1	0.81	1	0.64	1	0.37	1	0.76	1	0.47	1	0.81	1	0.64	1	0.37	1	0.76	1	0.47

Metrik Evaluasi Data Setelah Oversampling

					Orig	inal								Tuni	ng Hype	rparam	eters			
Algoritma	Accu	iracy	Prec	ision	Re	call	Al	JC	F	1	Accu	iracy	Prec	ision	Re	call	Al	JC	F	1
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.62	0.61	0.62	0.32	0.63	0.61	0.66	0.65	0.63	0.42										
KNN	0.84	0.58	0.77	0.28	0.95	0.54	0.95	0.59	0.85	0.37										
Decision Tree	1	0.67	1	0.34	1	0.48	1	0.61	1	0.4										
Random Forest	1	0.79	1	0.53	1	0.5	1	0.76	1	0.51	1	0.79	1	0.53	1	0.5	1	0.76	1	0.51

Skenario 3: Data Distandarisasi

Algoritma yang digunakan:

- Logistic Regression
- KNN
- Decision Tree

Feature yang digunakan:

Numericals

Dinormalisasi / Distandarisasi

- Datanya banyak categoricals
- Metrics

```
=> main =F1,
```

second = Precission, Recall, Accuracy, AUC.

Imbalance data

Logistic Regression

```
Accuracy (Test Set): 0.78
Precision (Test Set): 0.00
Recall (Test Set): 0.00
F1-Score (Test Set): 0.00
AUC (Test Set): 0.66
Accuracy (Train Set): 0.78
Precision (Train Set): 1.00
Recall (Train Set): 0.00
F1-Score (Train Set): 0.00
AUC (Train Set): 0.67
Best penalty: 11
Best C: 0.6
Best solver: liblinear
```

KNN

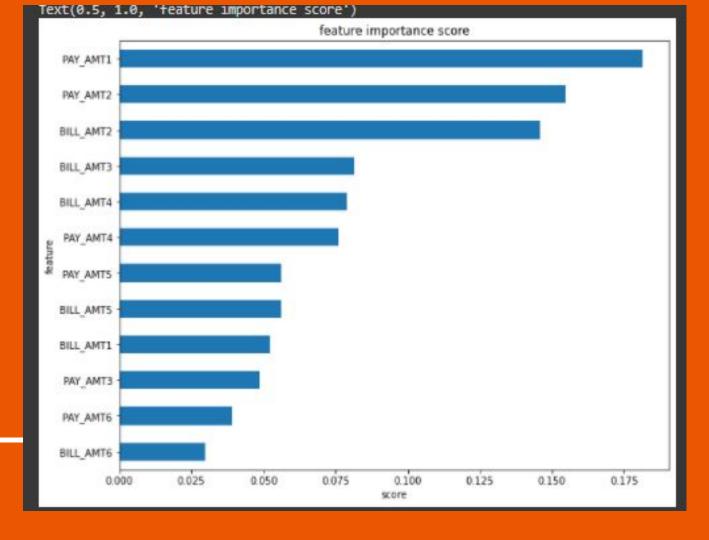
```
Accuracy (Test Set): 0.78
Precision (Test Set): 0.64
Recall (Test Set): 0.03
F1-Score (Test Set): 0.05
AUC (Test Set): 0.67
Accuracy (Train Set): 0.79
Precision (Train Set): 0.73
Recall (Train Set): 0.04
F1-Score (Train Set): 0.08
AUC (Train Set): 0.72
Best n neighbors: 50
Best weights: uniform
Best algorithm: auto
Best p: 1
```

Decision Tree

```
Accuracy (Test Set): 0.77
Precision (Test Set): 0.48
Recall (Test Set): 0.22
F1-Score (Test Set): 0.30
AUC (Test Set): 0.66
Accuracy (Train Set): 0.80
Precision (Train Set): 0.62
Recall (Train Set): 0.30
F1-Score (Train Set): 0.40
AUC (Train Set): 0.75
Best max depth: 8
Best min_samples_split: 5
Best min samples leaf: 1
Best max_features: sqrt
```

Evaluation Metrics

					Orig	ginal								Tuni	ng Hype	rparame	eters			
Algoritma	Accu	ıracy	Prec	ision	Re	call	Al	UC	F	1	Accı	ıracy	Prec	ision	Re	call	Al	UC	F	1
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0
KNN	0.81	0.75	0.65	0.42	0.41	0.25	0.84	0.64	0.5	0.31	0.77	0.78	0.68	0.74	0.06	0.04	0.72	0.67	0.1	0.08
Decision Tree	0.99	0.7	0.99	0.33	0.95	0.34	1	0.58	0.97	0.34	0.79	0.78	0.6	0.52	0.26	0.22	0.77	0.69	0.36	0.31



Skenario 4: Data setelah outlier dihapus dan distandarisasi

Algoritma yang digunakan:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Feature yang digunakan:

- Semua feature
- Berdasarkan nilai correlation heatmap
- PAY_AMTX dan BILL_AMTX only
- SelectKBest Features (sklearn library)

Metrik Evaluasi Data Imbalance

201		Imbalan	ice			21						Tuning	Imbalanc	e					120		
Algoritma	Features	Accu	iracy	Prec	ision	Re	call	Al	UC	F	1	Acci	uracy	Preci	ision	Rei	call	A	UC	FI	Į.
700		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	All	0.81	0.82	0.68	0.67	0.37	0.37	0.78	0.77	0.48	0.47	0.81	0.82	0.68	0.67	0.36	0.36	0.78	0.77	0.47	0.47
Logistic Regression	Corr Heatmap	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.47	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.48
rog but neglession	PAY_AMT & BILL_AMT only	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0
	SelectKBest	0.81	0.82	0.68	0.67	0.37	0.37	0.77	0.76	0.48	0.48	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.47
	All	0.83	0.78	0.74	0.53	0.43	0.32	0.87	0.7	0.54	0.4	0.8	0.8	0.69	0.67	0.22	0.22	0.78	0.74	0.33	0.33
KNN	Corr Heatmap	0.83	0.79	0.7	0.54	0.46	0.36	0.79	0.69	0.55	0.43	0.81	0.81	0.68	0.66	0.35	0.34	0.78	0.75	0.46	0.45
KININ	PAY_AMT & BILL_AMT only	0.81	0.75	0.65	0.42	0.41	0.25	0.84	0.64	0.5	0.31	0.77	0.78	0.75	0.89	0.02	0.01	0.69	0.68	0.04	0.03
	SelectKBest	0.84	0.79	0.74	0.54	0.45	0.33	0.88	0.7	0.56	0.41	0.8	0.8	0.69	0.65	0.26	0.25	0.78	0.75	0.38	0.36
	All	1	0.72	1	0.39	1	0.42	1	0.61	1	0.4	0.81	0.79	0.66	0.57	0.38	0.35	0.8	0.73	0.49	0.43
Decision Tree	Corr Heatmap	0.87	0.79	0.87	0.54	0.54	0.35	0.9	0.66	0.67	0.42	0.81	0.81	0.69	0.67	0.35	0.34	0.77	0.75	0.47	0.45
Decision free	PAY_AMT & BILL_AMT only	0.99	0.7	0.99	0.33	0.95	0.34	1	0.58	0.97	0.34	0.79	0.77	0.61	0.5	0.22	0.16	0.76	0.66	0.32	0.24
	SelectKBest	0.98	0.72	1	0.39	0.91	0.41	1	0.61	0.95	0.4	0.81	0.8	0.67	0.62	0.38	0.35	0.78	0.73	0.49	0.45
	All	1	0.81	1	0.64	1	0.37	1	0.76	1	0.47	0.84	0.82	0.79	0.68	0.41	0.35	0.9	0.78	0.54	0.47
Random Forest	Corr Heatmap	0.87	0.79	0.83	0.54	0.58	0.37	0.89	0.71	0.68	0.44	0.82	0.82	0.7	0.68	0.37	0.36	0.8	0.76	0.48	0.47
National Forest	PAY_AMT & BILL_AMT only	0.99	0.78	0.99	0.52	0.95	0.2	1	0.7	0.97	0.29	0.82	0.78	0.89	0.61	0.24	0.13	0.85	0.72	0.38	0.21
	SelectKBest	0.98	0.79	0.99	0.54	0.92	0.4	1	0.73	0.95	0.46	0.82	0.82	0.71	0.68	0.37	0.35	0.81	0.77	0.49	0.46
	All	0.81	0.81	0.69	0.66	0.35	0.35	0.79	0.77	0.47	0.46	0.81	0.81	0.7	0.66	0.35	0.34	0.8	0.78	0.46	0.45
Ad- B	Corr Heatmap	0.81	0.81	0.68	0.66	0.37	0.37	0.77	0.76	0.48	0.47	0.81	0.81	0.69	0.66	0.34	0.34	0.77	0.76	0.45	0.45
Ada Boost	PAY_AMT & BILL_AMT only	0.78	0.78	0.61	0.63	0.13	0.12	0.73	0.71	0.21	0.2	0.78	0.78	0.62	0.62	0.11	0.1	0.74	0.72	0.19	0.17
	SelectKBest	0.81	0.82	0.69	0.67	0.36	0.37	0.77	0.76	0.47	0.48	0.81	0.81	0.69	0.67	0.34	0.34	0.77	0.77	0.45	0.45
	All	0.82	0.82	0.72	0.67	0.38	0.36	0.81	0.78	0.49	0.47	0.86	0.81	0.82	0.64	0.49	0.38	0.91	0.78	0.62	0.47
VCBood	Corr Heatmap	0.82	0.82	0.69	0.66	0.38	0.37	0.78	0.77	0.49	0.48	0.82	0.82	0.7	0.67	0.38	0.37	0.79	0.77	0.49	0.48
XGBoost	PAY_AMT & BILL_AMT only	0.79	0.79	0.65	0.6	0.18	0.15	0.77	0.72	0.28	0.24	0.85	0.78	0.87	0.53	0.42	0.22	0.91	0.71	0.57	0.31
	SelectKBest	0.82	0.82	0.71	0.67	0.38	0.37	0.79	0.77	0.49	0.48	0.81	0.81	0.7	0.67	0.33	0.32	0.79	0.77	0.45	0.43

Metrik Evaluasi Data Setelah Oversampling

111		Balano	2									Tuning	Balance								
Algoritma	Features	Acci	iracy	Preci	ision	Re	call	Al	UC	F	1	Accu	racy	Prec	ision	Red	call	Al	JC	F:	1
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	All	0.76	0.76	0.74	0.72	0.44	0.44	0.78	0.77	0.55	0.54	0.76	0.76	0.74	0.72	0.43	0.43	0.78	0.77	0.55	0.54
Logistic Regression	Corr Heatmap	0.76	0.76	0.74	0.72	0.45	0.42	0.77	0.76	0.56	0.53	0.76	0.75	0.74	0.72	0.44	0.41	0.77	0.76	0.56	0.53
rog prir vegi ezatori	PAY_AMT & BILL_AMT only	0.66	0.67	0.45	0.4	0	0	0.67	0.67	0.01	0	0.66	0.67	0.45	0.4	0	0	0.67	0.67	0.01	0
	SelectKBest	0.76	0.76	0.74	0.73	0.45	0.44	0.77	0.76	0.56	0.54	0.76	0.76	0.75	0.73	0.44	0.43	0.77	0.77	0.55	0.54
	All	0.84	0.75	0.75	0.61	0.76	0.63	0.91	0.79	0.75	0.62	1	0.79	1	0.71	1	0.64	1	0.84	1	0.67
KNN	Corr Heatmap	0.78	0.73	0.68	0.59	0.62	0.53	0.82	0.73	0.65	0.56	0.76	0.75	0.73	0.71	0.46	0.42	0.78	0.76	0.56	0.53
KININ	PAY_AMT & BILL_AMT only	0.81	0.71	0.74	0.56	0.68	0.53	0.87	0.72	0.71	0.54	0.98	0.71	1	0.74	0.95	0.2	1	0.73	0.97	0.32
	SelectKBest	0.83	0.74	0.75	0.61	0.71	0.58	0.9	0.78	0.73	0.59	0.98	0.77	1	0.67	0.95	0.58	1	0.79	0.97	0.62
	All	1	0.73	1	0.58	1	0.61	1	0.7	1	0.59	0.79	0.75	0.75	0.65	0.57	0.51	0.85	0.77	0.65	0.58
Decision Tree	Corr Heatmap	0.85	0.76	0.89	0.69	0.63	0.49	0.91	0.72	0.74	0.57	0.77	0.75	0.71	0.66	0.52	0.49	0.78	0.75	0.6	0.56
Decision free	PAY_AMT & BILL_AMT only	0.98	0.67	1	0.5	0.95	0.51	1	0.64	0.97	0.51	0.75	0.69	0.68	0.55	0.47	0.38	0.8	0.69	0.55	0.45
	SelectKBest	0.98	0.75	1	0.61	0.95	0.61	1	0.72	0.97	0.61	0.78	0.77	0.75	0.72	0.51	0.48	0.79	0.76	0.6	0.58
	All	1	0.81	1	0.79	1	0.58	1	0.86	1	0.67	1	0.81	1	0.79	0.99	0.59	1	0.86	0.99	0.67
Random Forest	Corr Heatmap	0.85	0.76	0.87	0.68	0.66	0.51	0.9	0.76	0.75	0.58	0.79	0.76	0.78	0.71	0.51	0.46	0.83	0.78	0.62	0.56
Nandom roresc	PAY_AMT & BILL_AMT only	0.98	0.78	1	0.73	0.95	0.5	1	0.8	0.97	0.59	0.98	0.77	0.99	0.73	0.95	0.49	1	0.8	0.97	0.59
	SelectKBest	0.98	0.79	0.99	0.72	0.96	0.61	1	0.82	0.97	0.66	0.95	0.8	0.97	0.75	0.87	0.6	0.99	0.83	0.92	0.67
	All	0.8	0.79	0.8	0.79	0.52	0.5	0.83	0.82	0.63	0.61	0.8	0.79	0.81	0.79	0.52	0.51	0.84	0.83	0.64	0.62
Ada Boost	Corr Heatmap	0.77	0.76	0.75	0.73	0.45	0.43	0.78	0.77	0.56	0.54	0.77	0.76	0.75	0.73	0.45	0.42	0.78	0.77	0.56	0.54
Aud DOOSE	PAY_AMT & BILL_AMT only	0.72	0.71	0.67	0.65	0.31	0.29	0.73	0.71	0.42	0.4	0.71	0.71	0.68	0.65	0.28	0.26	0.74	0.72	0.39	0.37
	SelectKBest	0.79	0.79	0.79	0.78	0.52	0.5	0.81	0.81	0.62	0.61	0.79	0.79	0.8	0.78	0.5	0.49	0.82	0.81	0.62	0.6
	All	0.82	0.81	0.84	0.81	0.57	0.54	0.87	0.85	0.68	0.65	0.99	0.82	0.99	0.79	0.97	0.63	1	0.86	0.98	0.7
X GBoost	Corr Heatmap	0.77	0.76	0.76	0.73	0.47	0.44	0.79	0.78	0.58	0.55	0.8	0.77	0.79	0.71	0.55	0.5	0.83	0.79	0.65	0.59
AGDOOS	PAY_AMT & BILL_AMT only	0.73	0.72	0.71	0.66	0.34	0.31	0.78	0.73	0.46	0.42	0.98	0.77	0.99	0.69	0.95	0.53	1	0.8	0.97	0.6
	SelectKBest	0.81	0.8	0.83	0.8	0.55	0.53	0.85	0.83	0.66	0.64	0.83	0.82	0.85	0.83	0.59	0.56	0.87	0.85	0.7	0.67

Summary Tuning Hyperparameter

Terdapat lebih dari 100 eksperimen yang dilakukan dari skenario 1-4 termasuk eksperimen dengan melakukan tuning hyperparameter, setelah dianalisis secara menyeluruh pengaruh tuning hyperparameter pada model dapat disimpulkan bahwa penggunaan tuning dapat membuat model menjadi tidak terlalu overfit maupun underfit, namun pengaruhnya tidak terlalu signifikan antara score evaluasi hasil tuning dengan tidak.

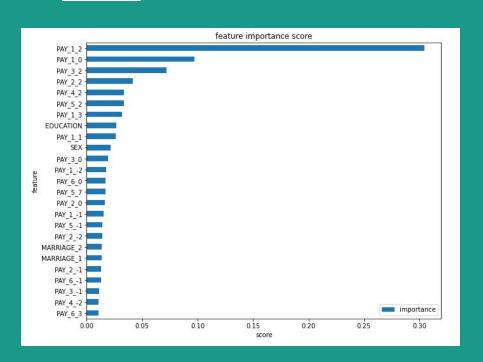
Summary Modeling

4 Nilai metrik evaluasi dengan AUC terbesar

Algoritma	Features	Condition	Accuracy	Precision	Recall	AUC	F1
	All features	Overcompling	0.81	0.81	0.54	0.85	0.65
XGBoost		Oversampling	0.8	8.0	0.53	0.83	0.64
AG BOOST	SelectKBest	Oversampling & Tuning Hyperparameters	0.82	0.83	0.56	0.85	0.67
AdaBoost	All features	Oversampling & Tuning Hyperparameters	0.79	0.79	0.51	0.83	0.62

Best model yaitu XGBoost dengan fitur hasil dari SelectKBest dan telah di oversampling serta tuning hyperparameters, dimana model tersebut memiliki nilai AUC, Precision, dan Recall tertinggi.

Best Model & Feature Importances



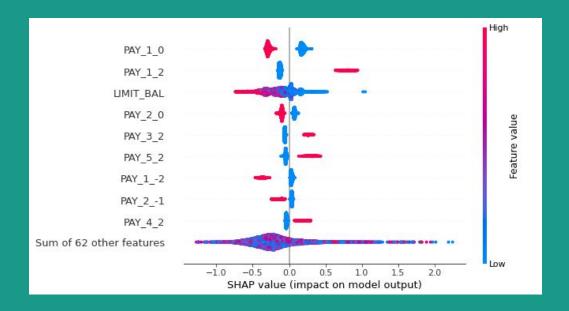
eksperimen terhadap best model dengan 10 features terpenting

```
Accuracy (Test Set): 0.76
Precision (Test Set): 0.71
Recall (Test Set): 0.43
F1-Score (Test Set): 0.54
AUC (Test Set): 0.76

Accuracy (Train Set): 0.76
Precision (Train Set): 0.73
Recall (Train Set): 0.45
F1-Score (Train Set): 0.55
AUC (Train Set): 0.76
```

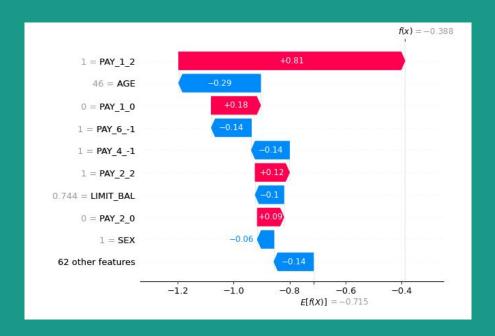
metric evaluation yang dihasilkan oleh model tidak memberikan score yang lebih bagus dibandingkan dengan model sebelumnya

Best Model & Feature Importances



Dapat disimpulkan bahwa yang paling berpengaruh terhadap status default adalah status pembayaran (PAY) terutama pembayaran yang telat 2 bulan dan juga limit kredit, sehingga dapat dilakukan penindakan kepada debitur ketika terdeteksi default 2 bulan berturut-turut, dengan menawarkan restructuring limit kredit, payment atau jumlah tagihan

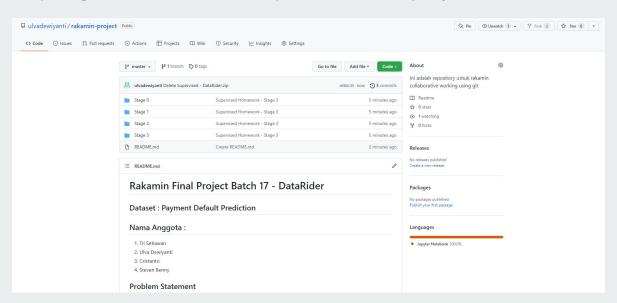
Model Interpretation



berdasarkan SHAP value disamping, terlihat bahwa PAY_1 dengan status 2 (telat 2 bulan) memberikan pengaruh yang sangat besar dalam penentuan default debitur tersebut, sehingga tindakan yang dapat diberikan seperti opsi restructuring Bill amount di bulan berikutnya sesuai kemampuan bayar debitur untuk mencegah default

Github Link

https://github.com/ulvadewiyanti/rakamin-project



Terima kasih

DataRider