Rakamin Academy

In-Vehicle Coupon Recommendation

Dokumen Laporan Final Project

DEEP LEARNING 4.0

MODELING

DATASET COMBINATION

DATASET	FEATURE ENGINEERING		ENCODING METHOD			HANDLE CLASS IMBALANCE	Feature Selection Manual	USER
	Transform	Extraction	One Hot	Hash	Binary			
1	✔		✔				✔	Vias
2	✔	✔	✔				✔	Norisa
3	✔			✔			✔	Risa
4	✔	✔		✔			✔	Zul
5	✔	✔		✔		✔	✔	Dika
6	✔				✔		✔	Yovan
7	✔	✔			✔		✔	Humaidi

Di tahap modeling ini menggunakan 7 kombinasi dataset yang berbeda terhadap model yang akan diuji, pemetaan kombinasi dataset seperti yang tertera pada tabel di atas. Tujuannya ialah untuk mengetahui apakah perbedaan perlakuan ketika preprocessing mempengaruhi hasil modeling lalu memilih dataset mana yang paling bagus performanya terhadap kelima model tersebut.

MODELING

MODEL MACHINE LEARNING

Data splitting dilakukan dengan proporsi 70:30

Model yang digunakan dalam tahap ini ialah:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- CatBoost

Tahapan Modelling:

Main metrics yang digunakan pada tahap model evaluation ialah akurasi. Hal ini dikarenakan tujuan utama dari prediksi model ini adalah meningkatkan keefektifan dalam pemberian kupon ke pengemudi/pelanggan, maka dari itu dengan metrics akurasi hasil prediksi model menghitung secara keseluruhan yang menerima kupon dan tidak menerima kupon. Selain menggunakan evaluation metrics akurasi, ROC – AOC sebagai secondary metrics digunakan untuk memilih model yang termasuk ‘best-fit’.

MODELING

DATASET 1 (One Hot Encoding)

BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
1	LogisticRegression					
	DecisionTree	0.6	0.66	0.61	0.99	0.6
	RandomForest	0.64	0.68	0.71	0.99	0.67
	XGBoost	0.61	0.63	0.79	0.66	0.64
	CatBoost					

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
1	LogisticRegression					
	DecisionTree	0.63	0.64	0.82	0.68	0.65
	RandomForest	0.62	0.62	0.87	0.68	0.66
	XGBoost	0.62	0.61	0.89	0.7	0.66
	CatBoost					

Berdasarkan hasil modelling menggunakan dataset 1, performa model terbaik menggunakan model DecisionTree dengan nilai recall akurasi 63% dan gap ROC – AUC sebesar 0.03

MODELING

DATASET 2 (Feature Extraction + One Hot Encoding)

BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
2	LogisticRegression	0.67	0.69	0.77	0.73	0.73
	DecisionTree	0.67	0.71	0.71	0.67	0.99
	RandomForest	0.62	0.61	0.94	0.73	0.74
	XGBoost	0.73	0.73	0.81	0.79	0.83
	CatBoost	0.74	0.74	0.83	0.82	0.93

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
2	LogisticRegression	0.56	0.56	1.0	0.64	0.64
	DecisionTree	0.64	0.68	0.67	0.67	0.92
	RandomForest	0.72	0.71	0.85	0.79	0.91
	XGBoost	0.75	0.74	0.84	0.82	0.97
	CatBoost	0.74	0.74	0.83	0.82	0.93

Berdasarkan hasil modelling menggunakan dataset 2, performa model terbaik menggunakan model XGBoost dengan nilai akurasi sebesar 75% dan gap ROC – AUC sebesar 0.15

MODELING

DATASET 3 (Hash Encoding)

BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
3	LogisticRegression	0,66	0,67	0,76	0,72	0,71
	DecisionTree	0,66	0,7	0,71	0,66	0,99
	RandomForest	0,74	0,75	0,8	0,8	0,99
	XGBoost	0,74	0,75	0,8	0,82	0,96
	CatBoost	0,66	0,7	0,71	0,66	0,99

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
3	LogisticRegression	0.56	0.56	1.0	0.56	0.56
	DecisionTree	0.56	0.56	1.0	0.56	0.57
	RandomForest	0.73	0.73	0.83	0.96	0.81
	XGBoost	0.71	0.71	0.83	0.86	0.79
	CatBoost	0.73	0.73	0.84	0.84	0.79

Berdasarkan hasil modelling menggunakan dataset 3, performa model terbaik menggunakan model CatBoost dengan nilai recall sebesar 73% dan gap ROC – AUC sebesar 0.05

MODELING

DATASET 4 (Feature Extraction + Hash Encoding)

BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
4	LogisticRegression	0.67	0.69	0.77	0.72	0.72
	DecisionTree	0.67	0.71	0.71	0.66	1.00
	RandomForest	0.76	0.76	0.83	0.82	1.00
	XGBoost	0.75	0.76	0.82	0.82	0.82
	CatBoost	0.75	0.75	0.84	0.83	0.93

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
4	LogisticRegression	0.68	0.69	0.79	0.72	0.72
	DecisionTree	0.68	0.70	0.77	0.73	0.73
	RandomForest	0.74	0.73	0.86	0.81	0.94
	XGBoost	0.73	0.73	0.85	0.80	0.87
	CatBoost	0.76	0.77	0.83	0.83	0.97

Berdasarkan hasil modelling menggunakan dataset 4, performa model terbaik menggunakan model CatBoost dengan nilai recall sebesar 76% dan gap ROC – AUC sebesar 0.14

MODELING

DATASET 5 (Feature Extraction + Hash Encoding + Handle Class Imbalance)

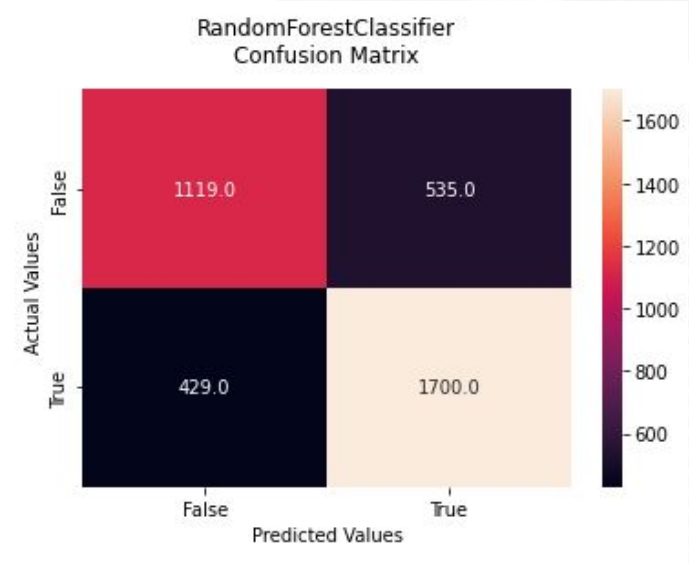
BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
5	LogisticRegression	0.67	0.69	0.75	0.78	0.72
	DecisionTree	0.66	0.70	0.70	1.00	0.65
	RandomForest	0.74	0.76	0.78	1.00	0.81
	XGBoost	0.74	0.76	0.79	0.97	0.81
	CatBoost	0.74	0.76	0.80	0.94	0.82

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
5	LogisticRegression	0.67	0.71	0.70	0.75	0.72
	DecisionTree	0.44	0	0	0.5	0.5
	RandomForest	0.74	0.76	0.79	1.00	0.81
	XGBoost	0.74	0.76	0.80	1.00	0.82
	CatBoost	0.73	0.75	0.78	0.90	0.81

Berdasarkan hasil modelling menggunakan dataset 5, performa model terbaik menggunakan model XGBoost dengan nilai recall sebesar 74% dan gap ROC – AUC sebesar 0.18

MODELING

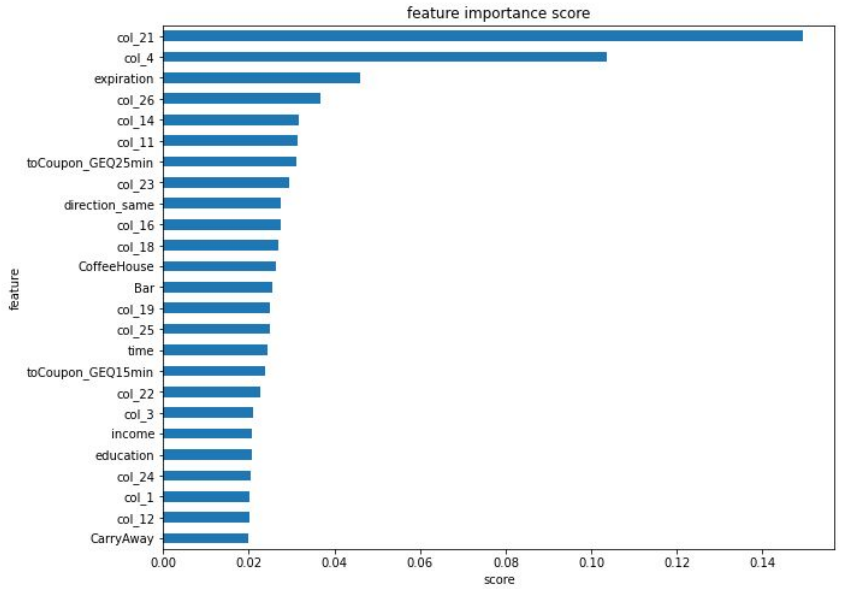
DATASET 5 (Feature Extraction + Hash Encoding + Handle Class Imbalance)



Confussion matrix model Random Forest terhadap dataset 5

MODELING

DATASET 5 (Feature Extraction + Hash Encoding + Handle Class Imbalance)



AFTER TUNING HYPERPARAMETER – BEFORE FEATURE IMPORTANCES						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
5	XGBoost	0.74	0.76	0.80	1.00	0.82

AFTER TUNING HYPERPARAMETER + AFTER FEATURE IMPORTANCES						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
5	XGBoost	0.73	0.75	0.78	0.99	0.80

Setelah melakukan modeling dengan feature importances berdasarkan feature 25 teratas, nilai modeling mengalami penurunan. Nilai akurasi menurun dari 74% menjadi 73%

MODELING

DATASET 6 (Binary Encoding)

BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
6	LogisticRegression	0.68	0.69	0.78	0.74	0.73
	DecisionTree	0.70	0.72	0.77	0.78	0.75
	RandomForest	0.64	0.63	0.90	0.72	0.72
	XGBoost	0.66	0.65	0.88	0.74	0.74
	CatBoost	0.75	0.75	0.83	0.95	0.83

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
6	LogisticRegression	0.68	0.70	0.78	0.80	0.74
	DecisionTree	0.71	0.72	0.80	0.80	0.75
	RandomForest	0.70	0.75	0.92	0.78	0.76
	XGBoost	0.58	0.58	0.99	0.71	0.72
	CatBoost	0.76	0.75	0.82	0.87	0.80

Berdasarkan hasil modelling menggunakan dataset 6, performa model terbaik menggunakan model CatBoost dengan nilai recall sebesar 76% dan gap ROC – AUC sebesar 0.3

MODELING

DATASET 7 (Feature Extraction + Binary Encoding)

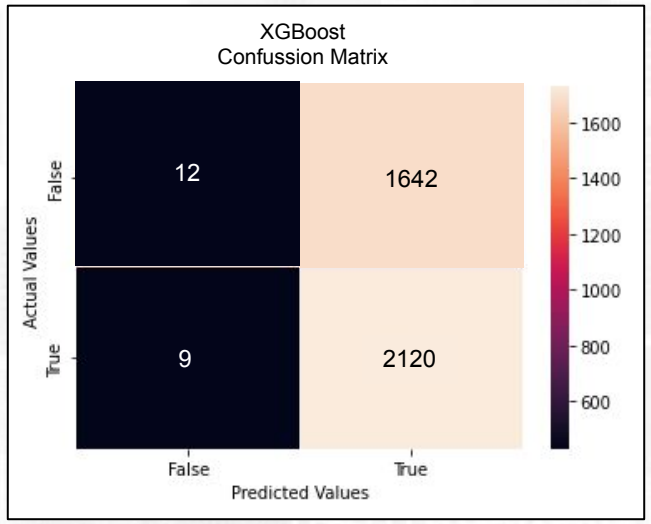
BEFORE TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
7	LogisticRegression	0.67	0.68	0.76	0.72	0.72
	DecisionTree	0.69	0.68	0.85	0.74	0.77
	RandomForest	0.65	0.63	0.90	0.74	0.73
	XGBoost	0.64	0.62	0.92	0.75	0.74
	CatBoost	0.75	0.75	0.82	0.82	0.93

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
7	LogisticRegression					
	DecisionTree					
	RandomForest					
	XGBoost	0.56	0.56	0.99	0.72	0.72
	CatBoost					

Berdasarkan hasil modelling menggunakan dataset 7, performa model terbaik menggunakan model XGBoost dengan nilai recall sebesar 56% dan tidak memiliki gap terhadap ROC – AUC

MODELING

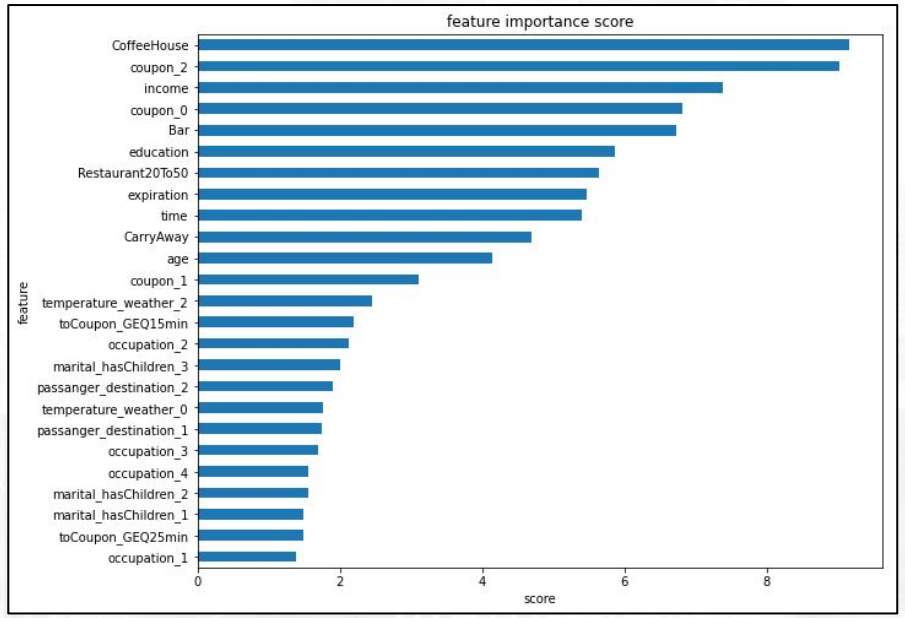
DATASET 7 (Feature Extraction + Binary Encoding)



Confussion matrix model XGBoost terhadap dataset 7

MODELING

DATASET 7 (Feature Extraction + Binary Encoding)



Berdasarkan hasil Feature Importance di atas, dapat dikatakan bahwa 25 feature di atas merupakan feature terpenting yang dapat secara signifikan mempengaruhi penerimaan atas kupon yang kita tawarkan. Dan berdasarkan diagram di atas penerima kupon dengan frekuensi ke coffee house, jenis kupon dan income memiliki pengaruh besar diterimanya kupon atau tidak.

MODELING

MODEL SELECTION

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
1	LogisticRegression					
	DecisionTree	0,63	0,64	0,82	0,68	0,65
	RandomForest	0,62	0,62	0,87	0,68	0,66
	XGBoost	0,62	0,61	0,89	0,7	0,66
	CatBoost					
2	LogisticRegression	0,56	0,56	1,0	0,64	0,64
	DecisionTree	0,64	0,68	0,67	0,67	0,92
	RandomForest	0,72	0,71	0,85	0,79	0,91
	XGBoost	0,75	0,74	0,84	0,82	0,97
	CatBoost					
3	LogisticRegression	0.56	0.56	1.0	0.56	0.56
	DecisionTree	0.56	0.56	1.0	0.56	0.57
	RandomForest	0.73	0.73	0.83	0.96	0.81
	XGBoost	0.71	0.71	0.83	0.86	0.79
	CatBoost	0.73	0.73	0.84	0.84	0.79
4	LogisticRegression	0.68	0.69	0.79	0.72	0.72
	DecisionTree	0.68	0.70	0.77	0.73	0.73
	RandomForest	0.74	0.73	0.86	0.81	0.94
	XGBoost	0.73	0.73	0.85	0.80	0.87
	CatBoost	0.76	0.77	0.83	0.83	0.97

AFTER TUNING HYPERPARAMETER						
DATASET	MODEL	EVALUATION METRICS				
		Accuracy	Precision	Recall	ROC-AUC Train	ROC-AUC Test
5	LogisticRegression	0.67	0.71	0.70	0.75	0.72
	DecisionTree	0.44	0	0	0.5	0.5
	RandomForest	0.74	0.76	0.79	1.00	0.81
	XGBoost	0.74	0.76	0.80	1.00	0.82
	CatBoost	0.73	0.75	0.78	0.90	0.81
6	LogisticRegression	0.68	0.67	0.69	0.72	0.70
	DecisionTree	0.71	0.72	0.70	0.80	0.75
	RandomForest	0.70	0.75	0.76	0.85	0.76
	XGBoost	0.68	0.77	0.76	0.82	0.72
	CatBoost	0.62	0.63	0.67	0.80	0.74
7	LogisticRegression					
	DecisionTree					
	RandomForest					
	XGBoost	0.56	0.56	0.99	0.72	0.72
	CatBoost					

Berdasarkan hasil modelling dari keseluruhan dataset dengan pertimbangan evaluation metrics akurasi, dataset 4 dengan model CatBoost memiliki perorma terbaik yakni nilai akurasi sebesar 76%.

GITHUB

https://github.com/dikaaka/In-Vehicle-Coupon-Recommendation-Project.git