Rakamin Academy

id/x partners

ID/X Partners Data Scientist Project Based Internship Program

Credit Risk Loan Prediction

Muhammad Hadi Dermawan

https://www.linkedin.com/in/hadi dermawan/

Tujuan

Membuat model machine learning untuk memprediksi nasabah-nasabah yang mampu membayar tagihan credi

Membuat otomasi untuk menentukan keputusan dari model machine learning credit risk

Meningkatkan efisiensi perusahaan dalam menentukan nasabah yang layak menggunakan otomasi yang telah dibuat

Business Understanding

- **ID/X Partners mendapatkan projek dari sebuah lending company untuk menyediakan solusi teknologi bagi company tersebut. Kamu diminta untuk membangun model yang dapat memprediksi credit risk menggunakan dataset yang disediakan oleh company yang terdiri dari data pinjaman yang diterima dan yang ditolak.**

Data Understanding

Data yang digunakan untuk membuat model adalah data **loan record dataset,** yang terdiri dari 466285 baris, 74 kolom dengan 22 variabel berjenis kategorikal dan 52 variabel berjenis numerikal

Statistical Descriptive

Kolom Numerikal

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	466285.0	2.331420e+05	1.346050e+05	0.00	1.165710e+05	2.331420e+05	3.497130e+05	4.662840e+05
id	466285.0	1.307973e+07	1.089371e+07	54734.00	3.639987e+06	1.010790e+07	2.073121e+07	3.809811e+07
member_id	466285.0	1.459766e+07	1.168237e+07	70473.00	4.379705e+06	1.194108e+07	2.300154e+07	4.086083e+07
loan_amnt	466285.0	1.431728e+04	8.286509e+03	500.00	8.000000e+03	1.200000e+04	2.000000e+04	3.500000e+04
funded_amnt	466285.0	1.429180e+04	8.274371e+03	500.00	8.000000e+03	1.200000e+04	2.000000e+04	3.500000e+04
funded_amnt_inv	466285.0	1.422233e+04	8.297638e+03	0.00	8.000000e+03	1.200000e+04	1.995000e+04	3.500000e+04
int_rate	466285.0	1.382924e+01	4.357587e+00	5.42	1.099000e+01	1.366000e+01	1.649000e+01	2.606000e+01
installment	466285.0	4.320612e+02	2.434855e+02	15.67	2.566900e+02	3.798900e+02	5.665800e+02	1.409990e+03
annual_inc	466281.0	7.327738e+04	5.496357e+04	1896.00	4.500000e+04	6.300000e+04	8.896000e+04	7.500000e+06
dti	466285.0	1.721876e+01	7.851121e+00	0.00	1.136000e+01	1.687000e+01	2.278000e+01	3.999000e+01
delinq_2yrs	466256.0	2.846784e-01	7.973651e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	2.900000e+01
inq_last_6mths	466256.0	8.047446e-01	1.091598e+00	0.00	0.000000e+00	0.000000e+00	1.000000e+00	3.300000e+01
mths_since_last_delinq	215934.0	3.410443e+01	2.177849e+01	0.00	1.600000e+01	3.100000e+01	4.900000e+01	1.880000e+02
mths_since_last_record	62638.0	7.430601e+01	3.035765e+01	0.00	5.300000e+01	7.600000e+01	1.020000e+02	1.290000e+02
open_acc	466256.0	1.118707e+01	4.987526e+00	0.00	8.000000e+00	1.000000e+01	1.400000e+01	8.400000e+01
pub_rec	466256.0	1.605642e-01	5.108626e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	6.300000e+01
revol_bal	466285.0	1.623020e+04	2.067625e+04	0.00	6.413000e+03	1.176400e+04	2.033300e+04	2.568995e+06
revol_util	465945.0	5.617695e+01	2.373263e+01	0.00	3.920000e+01	5.760000e+01	7.470000e+01	8.923000e+02
total_acc	466256.0	2.506443e+01	1.160014e+01	1.00	1.700000e+01	2.300000e+01	3.200000e+01	1.560000e+02
out_prncp	466285.0	4.410062e+03	6.355079e+03	0.00	0.000000e+00	4.414700e+02	7.341650e+03	3.216038e+04
out_prncp_inv	466285.0	4.408452e+03	6.353198e+03	0.00	0.000000e+00	4.413800e+02	7.338390e+03	3.216038e+04
total_pymnt	466285.0	1.154069e+04	8.265627e+03	0.00	5.552125e+03	9.419251e+03	1.530816e+04	5.777758e+04
total_pymnt_inv	466285.0	1.146989e+04	8.254158e+03	0.00	5.499250e+03	9.355430e+03	1.523131e+04	5.777758e+04
total_rec_prncp	466285.0	8.866015e+03	7.031688e+03	0.00	3.708560e+03	6.817760e+03	1.200000e+04	3.500003e+04
total_rec_int	466285.0	2.588677e+03	2.483810e+03	0.00	9.572800e+02	1.818880e+03	3.304530e+03	2.420562e+04
total_rec_late_fee	466285.0	6.501292e-01	5.265730e+00	0.00	0.000000e+00	0.000000e+00	0.000000e+00	3.586800e+02
recoveries	466285.0	8.534421e+01	5.522161e+02	0.00	0.000000e+00	0.000000e+00	0.000000e+00	3.352027e+04
collection_recovery_fee	466285.0	8.961534e+00	8.549144e+01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	7.002190e+03

collection_recovery_fee	466285.0	8.961534e+00	8.549144e+01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	7.002190e+03
last_pymnt_amnt	466285.0	3.123914e+03	5.554737e+03	0.00	3.126200e+02	5.459600e+02	3.187510e+03	3.623444e+04
collections_12_mths_ex_med	466140.0	9.085253e-03	1.086484e-01	0.00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+01
mths_since_last_major_derog	98974.0	4.285255e+01	2.166259e+01	0.00	2.600000e+01	4.200000e+01	5.900000e+01	1.880000e+02
policy_code	466285.0	1.000000e+00	0.000000e+00	1.00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
annual_inc_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dti_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
verification_status_joint	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
acc_now_delinq	466256.0	4.002093e-03	6.863680e-02	0.00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
tot_coll_amt	396009.0	1.919135e+02	1.463021e+04	0.00	0.000000e+00	0.000000e+00	0.000000e+00	9.152545e+06
tot_cur_bal	396009.0	1.388017e+05	1.521147e+05	0.00	2.861800e+04	8.153900e+04	2.089530e+05	8.000078e+06
open_acc_6m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_6m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_il_24m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mths_since_rcnt_il	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_bal_il	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
il_util	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_rv_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
open_rv_24m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max_bal_bc	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
all_util	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_rev_hi_lim	396009.0	3.037909e+04	3.724713e+04	0.00	1.350000e+04	2.280000e+04	3.790000e+04	9.999999e+06
inq_fi	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total_cu_tl	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
inq_last_12m	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Statistical Descriptive

Kolom Kategorikal

	count	unique	top	freq
term	466285	2	36 months	337953
grade	466285	7	B	136929
sub_grade	466285	35	B3	31686
emp_title	438697	205475	Teacher	5399
emp_length	445277	11	10+ years	150049
home_ownership	466285	6	MORTGAGE	235875
verification_status	466285	3	Verified	168055
issue_d	466285	91	Oct-14	38782
loan_status	466285	9	Current	224226
pymnt_plan	466285	2	n	466276
url	466285	466285	https://www.lendingclub.com/browse/loanDetail....	1
desc	125983	124436		234
purpose	466285	14	debt_consolidation	274195
title	466265	63099	Debt consolidation	164075
zip_code	466285	888	945xx	5304
addr_state	466285	50	CA	71450
earliest_cr_line	466256	664	Oct-00	3674
initial_list_status	466285	2	f	303005
last_pymnt_d	465909	98	Jan-16	179620
next_pymnt_d	239071	100	Feb-16	208393
last_credit_pull_d	466243	103	Jan-16	327699
application_type	466285	1	INDIVIDUAL	466285

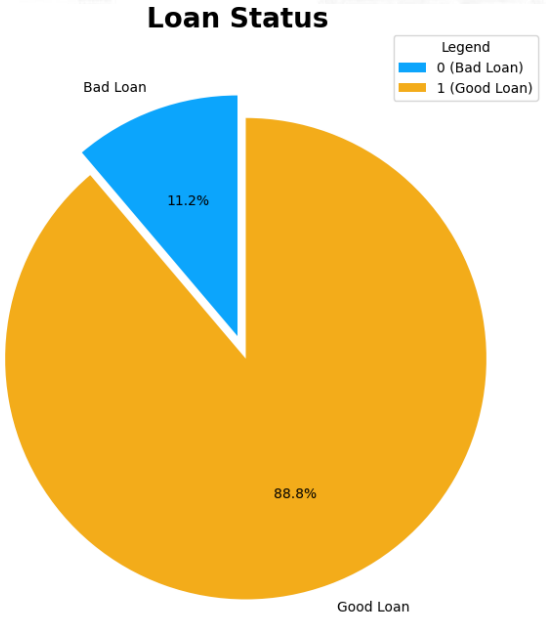
Statistical Descriptive

Kolom Kategorikal

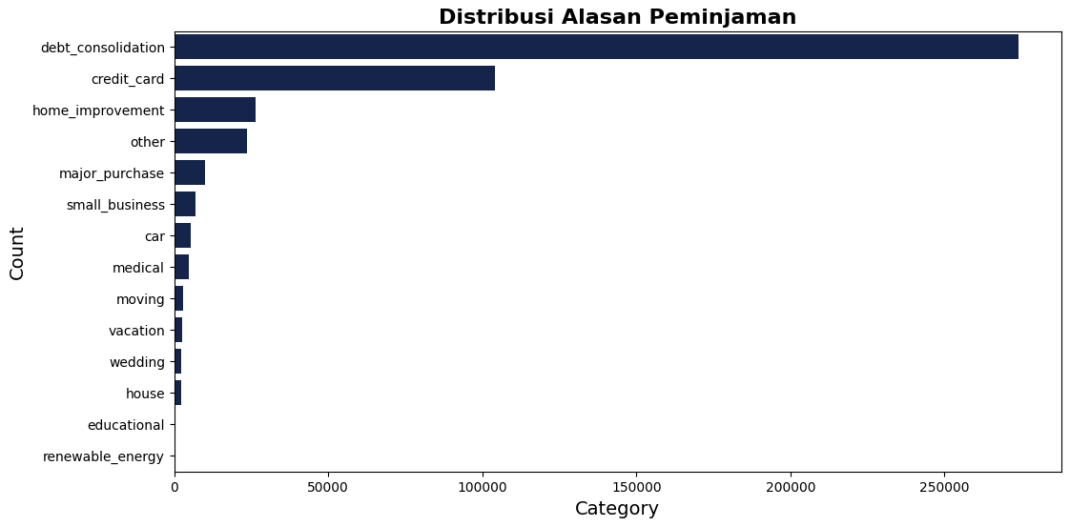
	count	unique	top	freq
term	466285	2	36 months	337953
grade	466285	7	B	136929
sub_grade	466285	35	B3	31686
emp_title	438697	205475	Teacher	5399
emp_length	445277	11	10+ years	150049
home_ownership	466285	6	MORTGAGE	235875
verification_status	466285	3	Verified	168055
issue_d	466285	91	Oct-14	38782
loan_status	466285	9	Current	224226
pymnt_plan	466285	2	n	466276
url	466285	466285	https://www.lendingclub.com/browse/loanDetail....	1
desc	125983	124436		234
purpose	466285	14	debt_consolidation	274195
title	466265	63099	Debt consolidation	164075
zip_code	466285	888	945xx	5304
addr_state	466285	50	CA	71450
earliest_cr_line	466256	664	Oct-00	3674
initial_list_status	466285	2	f	303005
last_pymnt_d	465909	98	Jan-16	179620
next_pymnt_d	239071	100	Feb-16	208393
last_credit_pull_d	466243	103	Jan-16	327699
application_type	466285	1	INDIVIDUAL	466285

Exploratory Data Analysis

Diketahui bahwa client company memiliki **performa bad loan** di angkat 11.2%, hal ini menandakan company ini berada diatas rata-rata dari rate Asia yaitu 3.84% (TheGolbalEconomy.com) yang mengakibatkan kerigian bagi client company.



Exploratory Data Analysis



Tujuan dari peminjaman credit oleh nasabah kebanyakan bertujuan untuk tujuan pembayaran debt consolidation atau pembayaran credit card

Data Prepocessing

Data Cleaning

Data cleaning dilakukan untuk menghapus variabel-variabel yang memiliki nilai null dan menghapus variabel yang tidak dapat memberikan informasi penting untuk pemodelan pada variabel bertipe kategorikal

Data Prepocessing

Missing Value

Untuk melakukan handling missing value dilakukan penghapusan variabel-variabel yang memiliki nilai null > 50% dan melakukan imputasi median untuk variabel bertipe numerikal dan modus untuk yang bertipe kategorikal

Data Prepocessing

Feature Engineering

Hal yang dilakukan pada feature engineering adalah melakukan feature encoding untuk data yang bertipe ordinal dan kategorikal

Data Prepocessing

Handling Outliers

Proses handling outlier dilakukan untuk variabel yang memiliki value diatas nilai IQRnya, variabel yang perlu dilakukan handling outliers, yaitu:

- 'installment’
- 'annual_inc’
- 'open_acc’
- 'total_rec_late_fee’
- 'last_pymnt_amnt’
- 'total_rev_hi_lim',
- 'tot_coll_amt’
- 'collection_recovery_fee’
- 'tot_cur_bal'

Modeling

Pada project ini proses eksperimen pemodelan dilakukan menggunakan 6 jenis model yaitu Random Forest, Decision tree, Logistic Regression, Adaboost, XG boost, dan Gradien boosting

Modeling

Berdasarkan hasil dari table disamping didapatkan bahwa model XGBoost memberikan hasil terbaik dari metric ROC AUC pada data test train maupun test dan tidak adanya indikasi overfitting

	Model	Accuracy	Precision	Recall	F1 Score	AUC (Test)	AUC (Train)
1	Random Forest	0.9784	0.9773	0.9998	0.9885	0.9377	1.0000
2	Decision Tree	0.9548	0.9795	0.9715	0.9755	0.8590	1.0000
3	Logistic Regression	0.9468	0.9815	0.9606	0.9709	0.9393	0.9579
4	AdaBoost	0.9626	0.9697	0.9905	0.9800	0.9087	0.9889
5	XGBoost	0.9806	0.9798	0.9996	0.9896	0.9519	0.9986
6	Gradien Boosting	0.9746	0.9733	0.9999	0.9864	0.9238	0.9936

Rakamin Academy

Terima Kasih