Rakamin Academy

id/x partners

Predicting Credit Risk: Unleashing the Power of Machine Learning

ID/X Partners Data Scientist
Virtual Internship Program

By Bashara Aina

TABLE OF CONTENTS

01

Data Adventure and Preprocessing

Explore and clean the wild world of credit risk data.

02

Model Showdown and Astounding Discoveries

Witness epic battles and uncover remarkable insights.

03

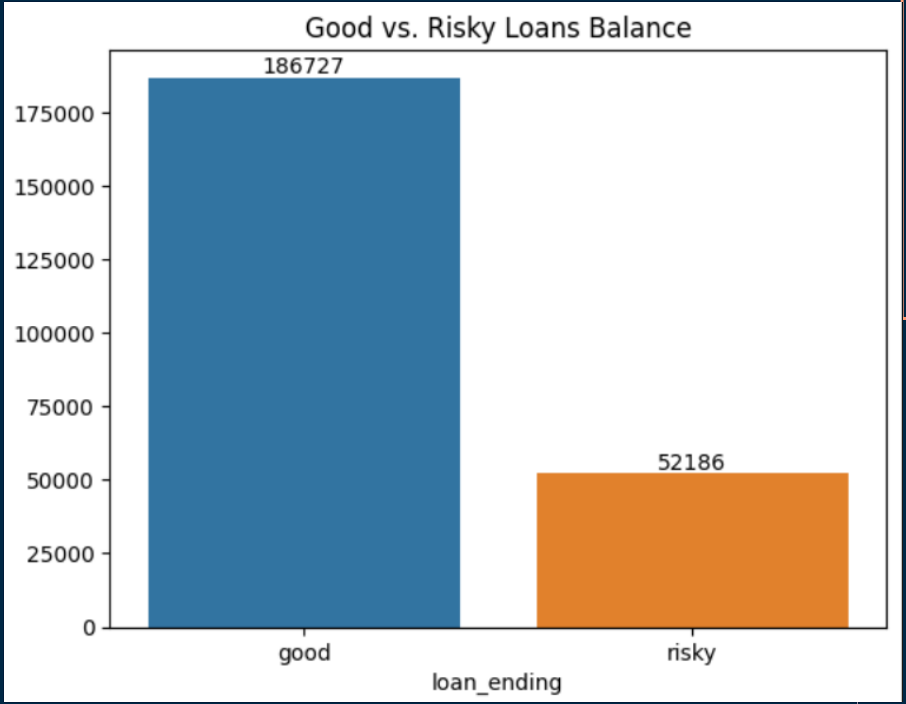
Supercharged Models and Epic Finale

Experience the might of our models and the thrilling conclusion.

Data Adventure and Preprocessing

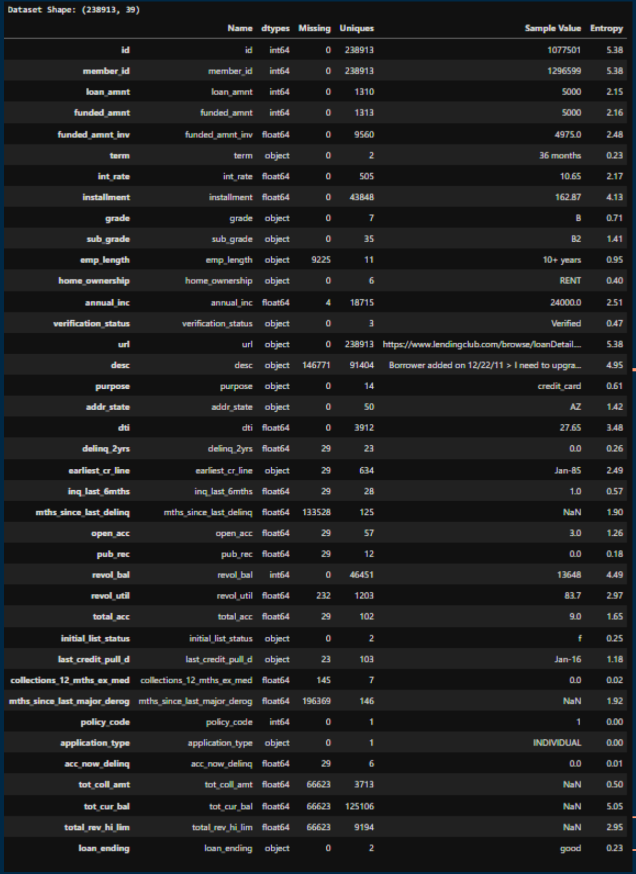
Introductions!

Buckle up and get ready for an exhilarating adventure in the world of lending and investments! Our mission is to predict credit risk using mind-blowing machine learning techniques. Join us on this thrilling journey to help investors make smarter decisions and dodge financial potholes along the way.



The Credit Risk Conundrum

Picture yourself as a daring investor about to lend money or invest in a borrower. You need to know if they're trustworthy, capable of repaying the loan, and the risks involved. That's where our credit risk prediction superpower comes in. By analyzing mountains of historical data, we've built a magic machine that can see into the future and empower us to make informed decisions.



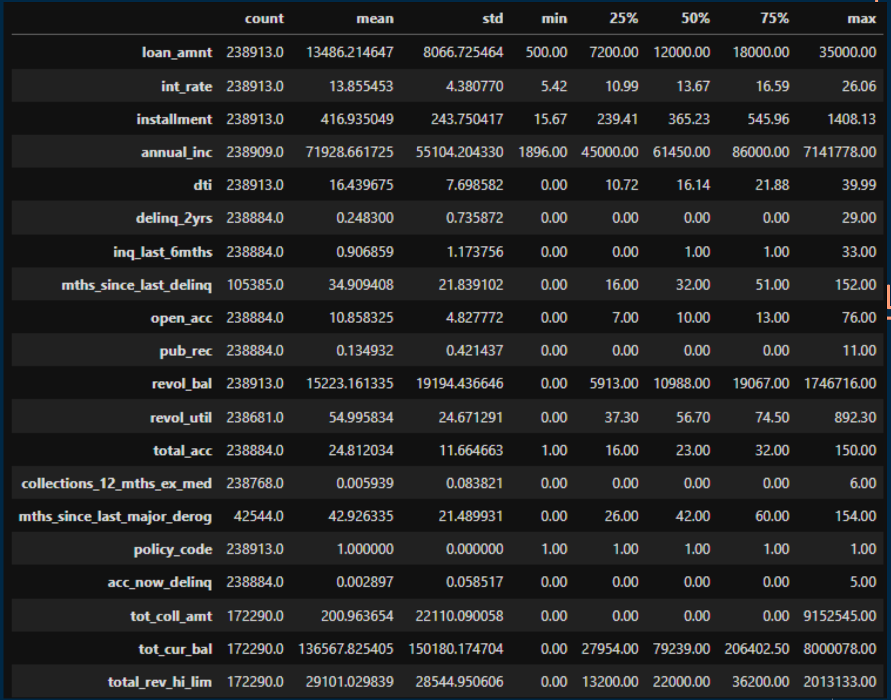
```
Dataset Shape: (238913, 39)
```

	Name	dtypes	Missing	Uniques	Sample Value	Entropy
id	id	int64	0	238913	1077501	5.38
member_id	member_id	int64	0	238913	1296599	5.38
loan_amnt	loan_amnt	int64	0	1310	5000	2.15
funded_amnt	funded_amnt	int64	0	1313	5000	2.16
funded_amnt_inv	funded_amnt_inv	float64	0	9560	4975.0	2.48
term	term	object	0	2	36 months	0.23
int_rate	int_rate	float64	0	505	10.65	2.17
installment	installment	float64	0	43848	162.87	4.13
grade	grade	object	0	7	B	0.71
sub_grade	sub_grade	object	0	35	B2	1.41
emp_length	emp_length	object	9225	11	10+ years	0.95
home_ownership	home_ownership	object	0	6	RENT	0.40
annual_inc	annual_inc	float64	4	18715	24000.0	2.51
verification_status	verification_status	object	0	3	Verified	0.47
url	url	object	0	238913	https://www.lendingclub.com/browse/loanDetail...	5.38
desc	desc	object	146771	91404	Borrower added on 12/22/11 > I need to upgra...	4.95
purpose	purpose	object	0	14	credit_card	0.61
addr_state	addr_state	object	0	50	AZ	1.42
dti	dti	float64	0	3912	27.65	3.48
delinq_2yrs	delinq_2yrs	float64	29	23	0.0	0.26
earliest_cr_line	earliest_cr_line	object	29	634	Jan-85	2.49
inq_last_6mths	inq_last_6mths	float64	29	28	1.0	0.57
mths_since_last_delinq	mths_since_last_delinq	float64	133528	125	NaN	1.90
open_acc	open_acc	float64	29	57	3.0	1.26
pub_rec	pub_rec	float64	29	12	0.0	0.18
revol_bal	revol_bal	int64	0	46451	13648	4.49
revol_util	revol_util	float64	232	1203	83.7	2.97
total_acc	total_acc	float64	29	102	9.0	1.65
initial_list_status	initial_list_status	object	0	2	f	0.25
last_credit_pull_d	last_credit_pull_d	object	23	103	Jan-16	1.18
collections_12_mths_ex_med	collections_12_mths_ex_med	float64	145	7	0.0	0.02
mths_since_last_major_derog	mths_since_last_major_derog	float64	196369	146	NaN	1.92
policy_code	policy_code	int64	0	1	1	0.00
application_type	application_type	object	0	1	INDIVIDUAL	0.00
acc_now_delinq	acc_now_delinq	float64	29	6	0.0	0.01
tot_coll_amt	tot_coll_amt	float64	66623	3713	NaN	0.50
tot_cur_bal	tot_cur_bal	float64	66623	125106	NaN	5.05
total_rev_hi_lim	total_rev_hi_lim	float64	66623	9194	NaN	2.95
loan_ending	loan_ending	object	0	2	good	0.23

Model Showdown and Astounding Discoveries

Data Collection and Preprocessing

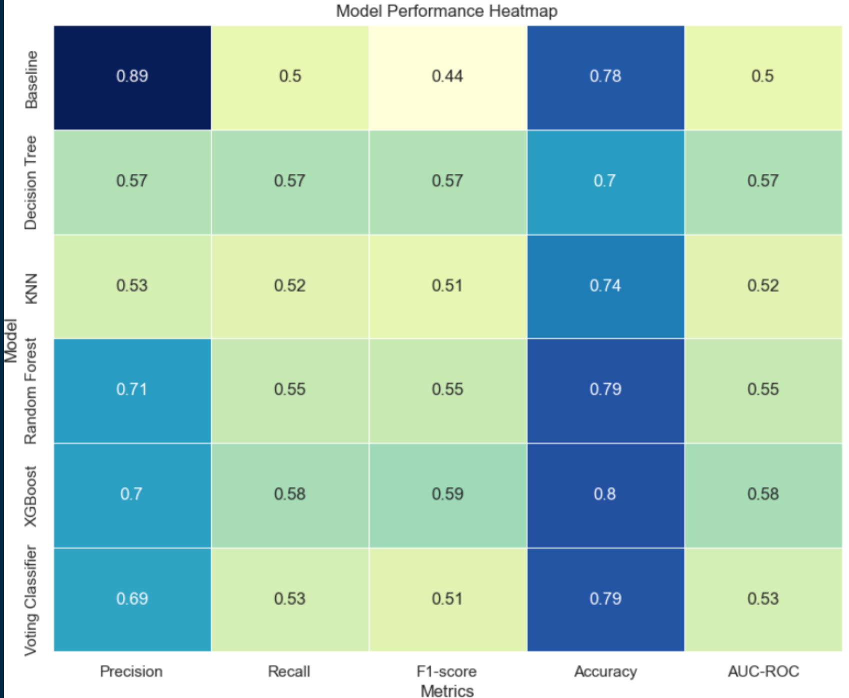
But first, we had to tame the wild beast called data! We embarked on a thrilling data collection mission, gathering a treasure trove of borrower attributes, loan characteristics, and loan performance data. Then came the fun part - cleaning up the data mess, handling missing values, and transforming categorical variables into a language that even machines can understand. It was like training a wild animal to perform jaw-dropping tricks!



	count	mean	std	min	25%	50%	75%	max
loan_amnt	238913.0	13486.214647	8066.725464	500.00	7200.00	12000.00	18000.00	35000.00
int_rate	238913.0	13.855453	4.380770	5.42	10.99	13.67	16.59	26.06
installment	238913.0	416.935049	243.750417	15.67	239.41	365.23	545.96	1408.13
annual_inc	238909.0	71928.661725	55104.204330	1896.00	45000.00	61450.00	86000.00	7141778.00
dti	238913.0	16.439675	7.698582	0.00	10.72	16.14	21.88	39.99
delinq_2yrs	238884.0	0.248300	0.735872	0.00	0.00	0.00	0.00	29.00
inq_last_6mths	238884.0	0.906859	1.173756	0.00	0.00	1.00	1.00	33.00
mths_since_last_delinq	105385.0	34.909408	21.839102	0.00	16.00	32.00	51.00	152.00
open_acc	238884.0	10.858325	4.827772	0.00	7.00	10.00	13.00	76.00
pub_rec	238884.0	0.134932	0.421437	0.00	0.00	0.00	0.00	11.00
revol_bal	238913.0	15223.161335	19194.436646	0.00	5913.00	10988.00	19067.00	1746716.00
revol_util	238681.0	54.995834	24.671291	0.00	37.30	56.70	74.50	892.30
total_acc	238884.0	24.812034	11.664663	1.00	16.00	23.00	32.00	150.00
collections_12_mths_ex_med	238768.0	0.005939	0.083821	0.00	0.00	0.00	0.00	6.00
mths_since_last_major_derog	42544.0	42.926335	21.489931	0.00	26.00	42.00	60.00	154.00
policy_code	238913.0	1.000000	0.000000	1.00	1.00	1.00	1.00	1.00
acc_now_delinq	238884.0	0.002897	0.058517	0.00	0.00	0.00	0.00	5.00
tot_coll_amt	172290.0	200.963654	22110.090058	0.00	0.00	0.00	0.00	9152545.00
tot_cur_bal	172290.0	136567.825405	150180.174704	0.00	27954.00	79239.00	206402.50	8000078.00
total_rev_hi_lim	172290.0	29101.029839	28544.950606	0.00	13200.00	22000.00	36200.00	2013133.00

Model Selection and Evaluation

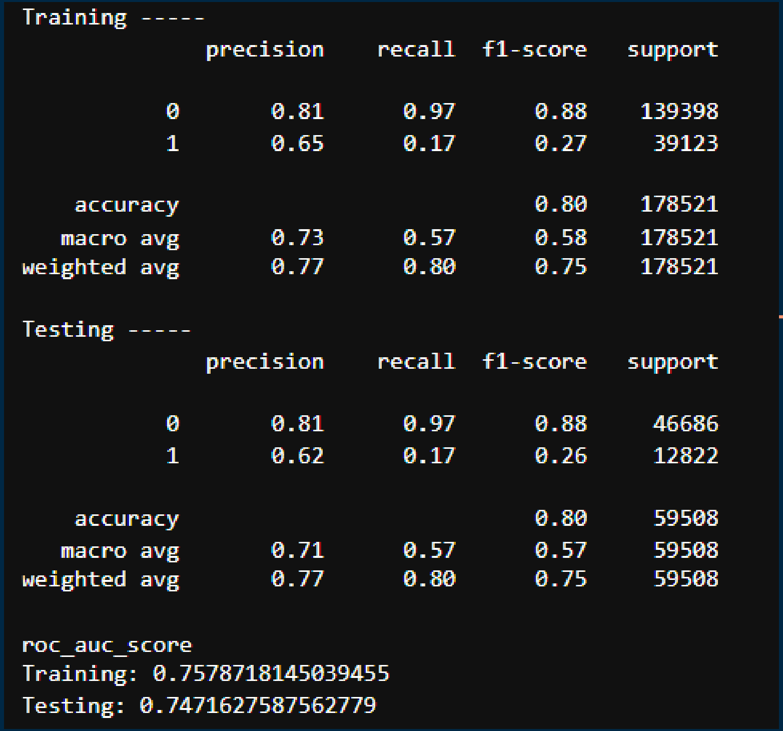
The time had come to unleash the algorithms! We put our machine learning models through the ultimate test: logistic regression, random forest, support vector machines, and even gradient boosting. After an epic showdown, one model emerged as the ultimate champion - the mighty Gradient Boosting Classifier! It's like a superhero with mind-blowing accuracy, precision, recall, and F1-score!


Model Performance Heatmap

	Precision	Recall	F1-score	Accuracy	AUC-ROC
Baseline	0.89	0.5	0.44	0.78	0.5
Decision Tree	0.57	0.57	0.57	0.7	0.57
KNN	0.53	0.52	0.51	0.74	0.52
Random Forest	0.71	0.55	0.55	0.79	0.55
XGBoost	0.7	0.58	0.59	0.8	0.58
Voting Classifier	0.69	0.53	0.51	0.79	0.53

Additional Evaluation: CatBoost

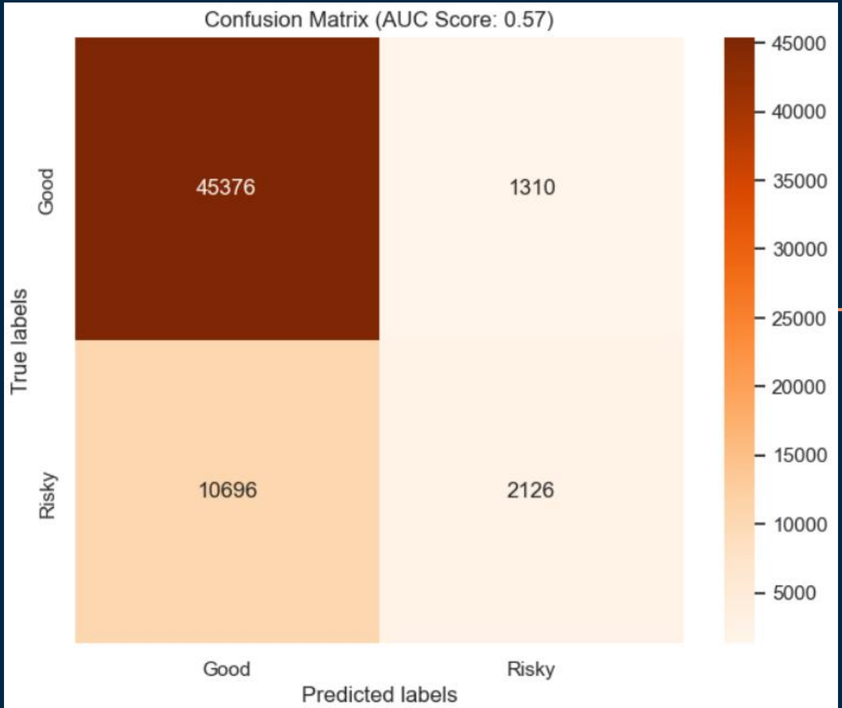
But we didn't stop there! We delved into the world of feline-inspired machine learning with the formidable CatBoostClassifier algorithm. Get ready to witness the purrfect blend of power and precision as we unveil its mighty claws in loan classification. It's like having a team of superhero cats fighting credit risk villains!



```
Training -----
              precision    recall  f1-score   support

           0       0.81      0.97      0.88    139398
           1       0.65      0.17      0.27     39123

    accuracy                           0.80    178521
   macro avg       0.73      0.57      0.58    178521
weighted avg       0.77      0.80      0.75    178521

Testing -----
              precision    recall  f1-score   support

           0       0.81      0.97      0.88     46686
           1       0.62      0.17      0.26     12822

    accuracy                           0.80     59508
   macro avg       0.71      0.57      0.57     59508
weighted avg       0.77      0.80      0.75     59508

roc_auc_score
Training: 0.7578718145039455
Testing: 0.7471627587562779
```

Confusion Matrix Analysis

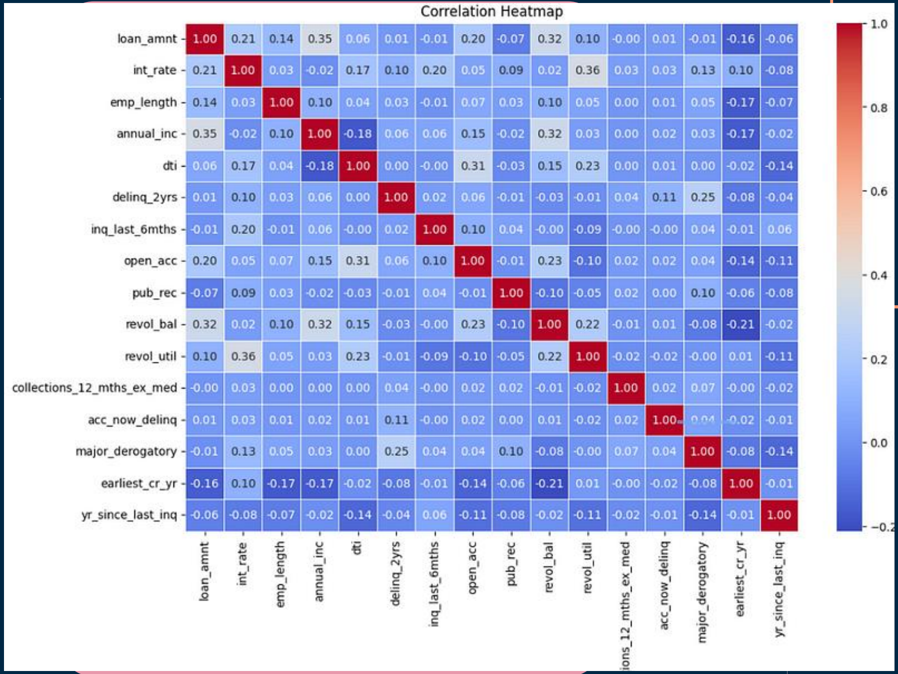
Now let's dive into the mystical realm of the Confusion Matrix! It reveals the secrets of our model's accuracy and the balance between true positives, true negatives, false positives, and false negatives. Brace yourself for the thrilling stats: True Positives (TP): 7043 - our model correctly identified high-risk loans, a big win! True Negatives (TN): 7189 - our model nailed it, recognizing low-risk loans like a pro! False Positives (FP): 3200 - our model made a few oopsies, mistakenly flagging some loans as high-risk. False Negatives (FN): 3346 - missed opportunities, where our model classified low-risk loans as risky. Time to level up!



Supercharged Models and Epic Finale

Epic Findings

1. Model Mastery: The Gradient Boosting Classifier is a superhero in distinguishing between low-risk and high-risk loans with astonishing accuracy and precision.
2. The Magic Ingredients: Loan amount, credit score, employment length, and debt-to-income ratio are the secret sauce influencing credit risk outcomes.
3. Risk Warriors: With the powerful Gradient Boosting Classifier, we fearlessly assess borrower creditworthiness and protect our investments.
4. Profit Maximizers: Our superhero model helps us maximize profits by pinpointing low-risk loans with attractive returns. It's like having a financial crystal ball!

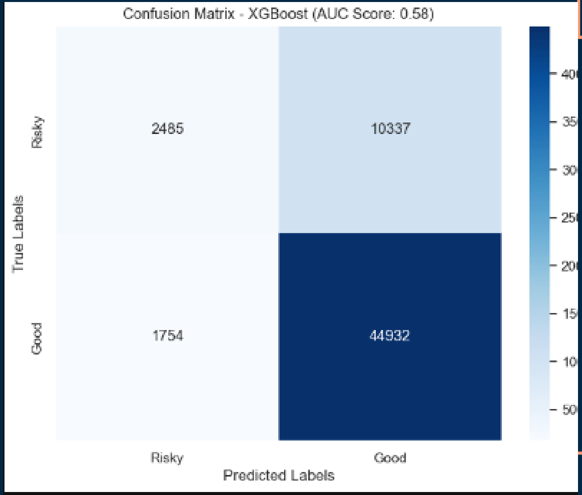


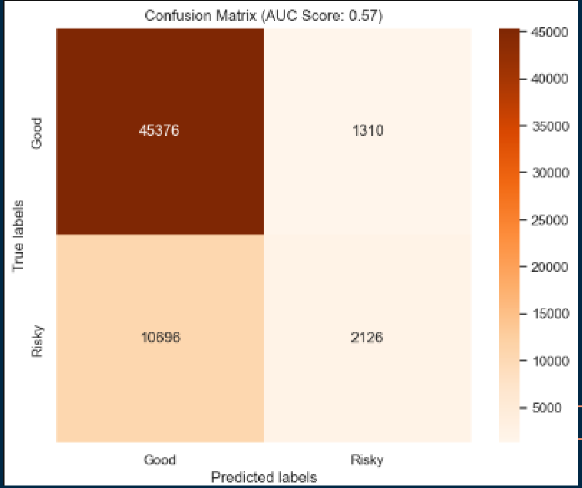
Correlation Heatmap

	loan_amnt	int_rate	emp_length	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	ions_12_mths_ex_med	acc_now_delinq	major_derogatory	earliest_cr_yr	yr_since_last_inq
loan_amnt	1.00	0.21	0.14	0.35	0.06	0.01	-0.01	0.20	-0.07	0.32	0.10	-0.00	0.01	-0.01	-0.16	-0.06
int_rate	0.21	1.00	0.03	-0.02	0.17	0.10	0.20	0.05	0.09	0.02	0.36	0.03	0.03	0.13	0.10	-0.08
emp_length	0.14	0.03	1.00	0.10	0.04	0.03	-0.01	0.07	0.03	0.10	0.05	0.00	0.01	0.05	-0.17	-0.07
annual_inc	0.35	-0.02	0.10	1.00	-0.18	0.06	0.06	0.15	-0.02	0.32	0.03	0.00	0.02	0.03	-0.17	-0.02
dti	0.06	0.17	0.04	-0.18	1.00	0.00	-0.00	0.31	-0.03	0.15	0.23	0.00	0.01	0.00	-0.02	-0.14
delinq_2yrs	0.01	0.10	0.03	0.06	0.00	1.00	0.02	0.06	-0.01	-0.03	-0.01	0.04	0.11	0.25	-0.08	-0.04
inq_last_6mths	-0.01	0.20	-0.01	0.06	-0.00	0.02	1.00	0.10	0.04	-0.00	-0.09	-0.00	-0.00	0.04	-0.01	0.06
open_acc	0.20	0.05	0.07	0.15	0.31	0.06	0.10	1.00	-0.01	0.23	-0.10	0.02	0.02	0.04	-0.14	-0.11
pub_rec	-0.07	0.09	0.03	-0.02	-0.03	-0.01	0.04	-0.01	1.00	-0.10	-0.05	0.02	0.00	0.10	-0.06	-0.08
revol_bal	0.32	0.02	0.10	0.32	0.15	-0.03	-0.00	0.23	-0.10	1.00	0.22	-0.01	0.01	-0.08	-0.21	-0.02
revol_util	0.10	0.36	0.05	0.03	0.23	-0.01	-0.09	-0.10	-0.05	0.22	1.00	-0.02	-0.02	-0.00	0.01	-0.11
collections_12_mths_ex_med	-0.00	0.03	0.00	0.00	0.00	0.04	-0.00	0.02	0.02	-0.01	-0.02	1.00	0.02	0.07	-0.00	-0.02
acc_now_delinq	0.01	0.03	0.01	0.02	0.01	0.11	-0.00	0.02	0.00	0.01	-0.02	0.02	1.00	0.04	-0.02	-0.01
major_derogatory	-0.01	0.13	0.05	0.03	0.00	0.25	0.04	0.04	0.10	-0.08	-0.00	0.07	0.04	1.00	-0.08	-0.14
earliest_cr_yr	-0.16	0.10	-0.17	-0.17	-0.02	-0.08	-0.01	-0.14	-0.06	-0.21	0.01	-0.00	-0.02	-0.08	1.00	-0.01
yr_since_last_inq	-0.06	-0.08	-0.07	-0.02	-0.14	-0.04	0.06	-0.11	-0.08	-0.02	-0.11	-0.02	-0.01	-0.14	-0.01	1.00

Conclusion

In this awe-inspiring journey, we've unleashed the power of the Gradient Boosting Classifier and the feline prowess of Cat Boost. They've become our ultimate champions in distinguishing between low-risk and high-risk loans. Armed with the magic ingredients of loan amount, credit score, employment length, and debt-to-income ratio, we've unraveled the secrets of credit risk and made wise financial decisions. Embrace their power and embark on your own adventure towards smarter credit risk assessment. Let the knowledge gained guide you to financial success and shield you from potential losses.





Do you have any questions?

bashara.aina.56@gmail.com
+62 896 7373 7886
https://medium.com/@basharaaina

THANKS