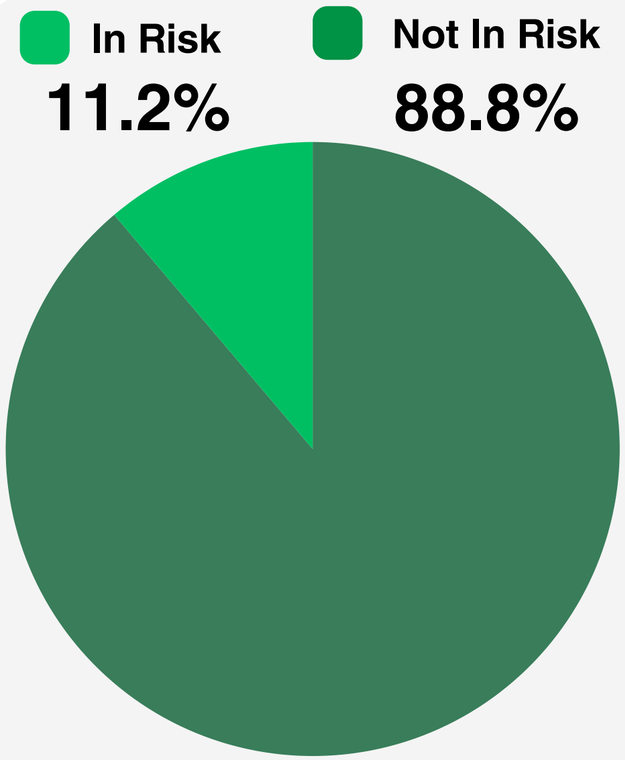
Project-Based Virtual Internship:

DATA SCIENTIST - ID/X PARTNERS x RAKAMIN ACADEMY

by Bhima Fairul Rifqi

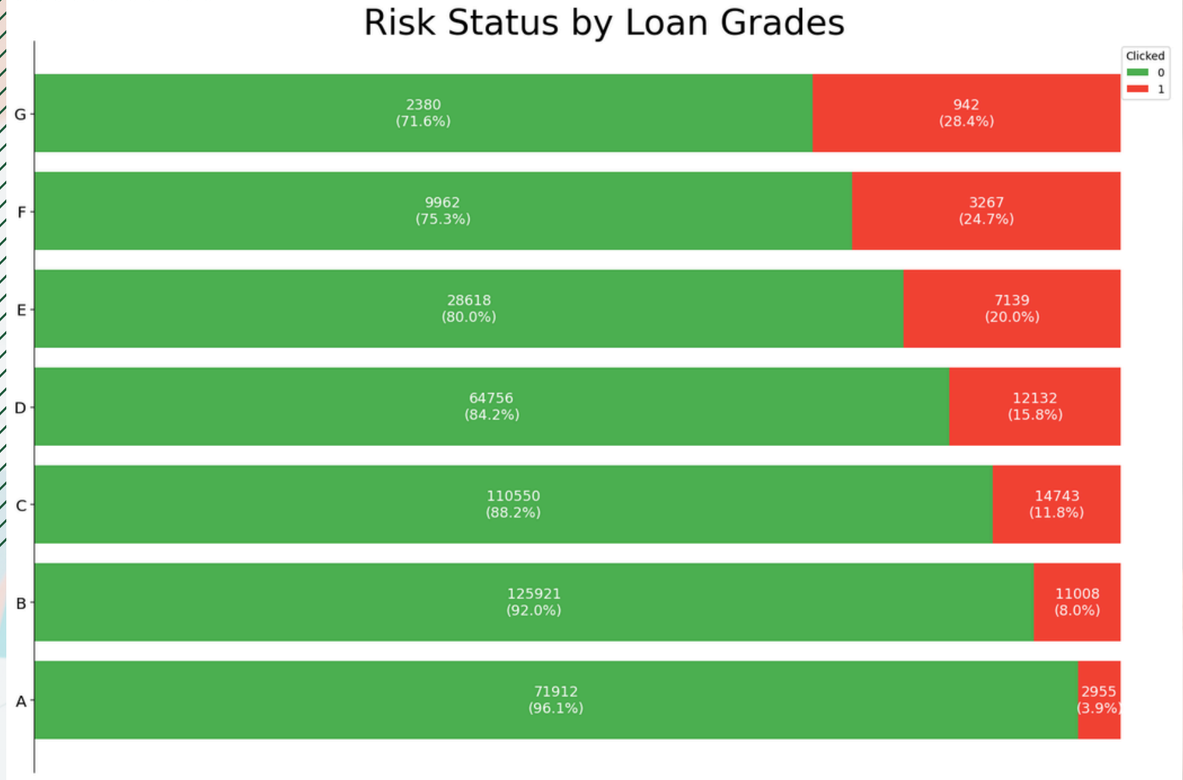
Business Overview



LendMeSome, a company with over 10% of its clients struggling to repay their loans, faced financial challenges as a funding company. To address this, they leveraged machine learning to identify clients who were potentially at risk

Business Metric

Struggling with its loan portfolio, LendMeSome aims to issue loans only to low-risk clients while maximizing returns from the interest they pay.

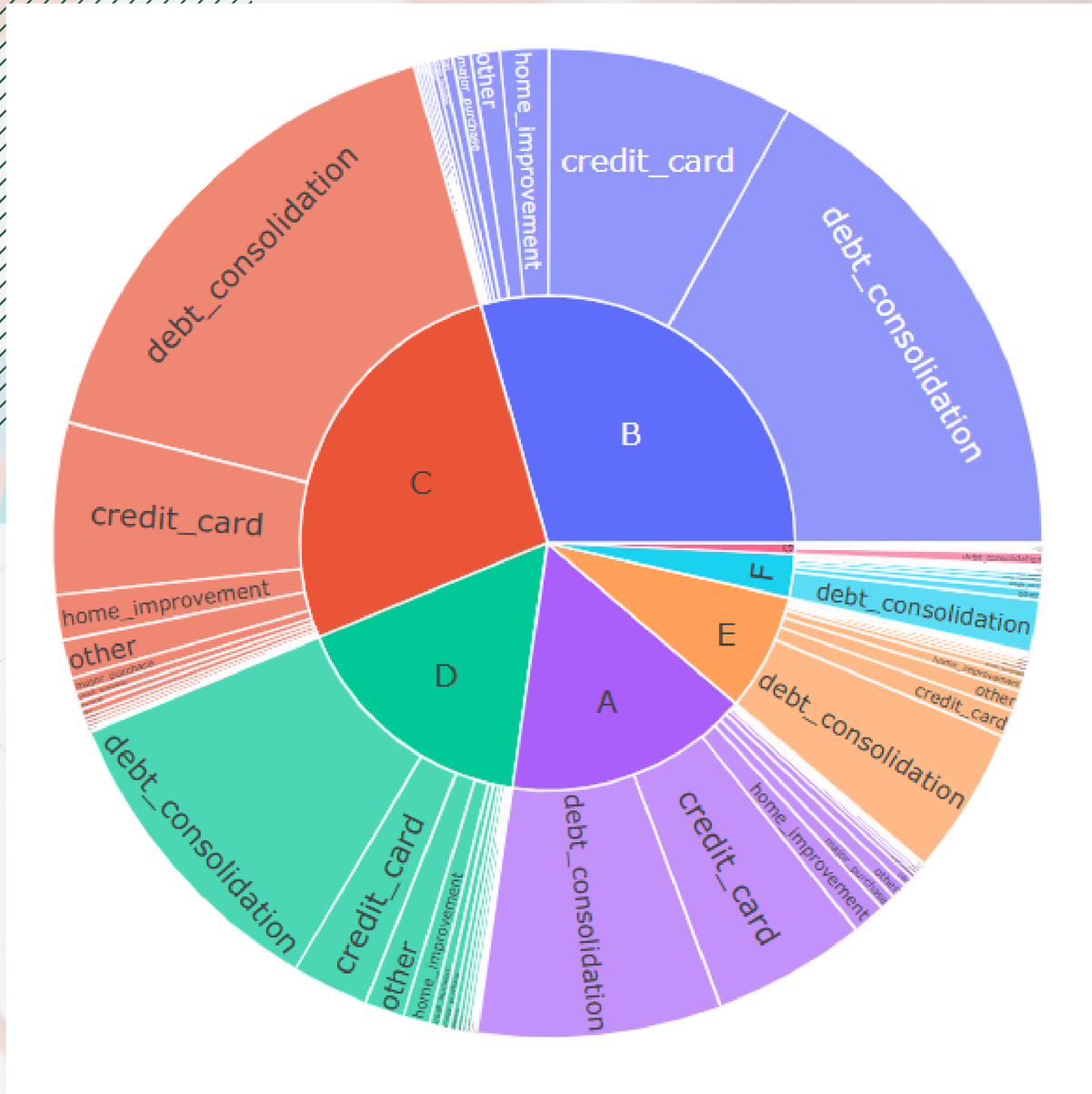


Insights:

Risk Status by Loan Grades

It was found that higher loan grades are associated with higher risk of default. This is likely because the interest rate significantly affects the risk (p-value ≈ 0), where a 1% increase in interest rate increases the odds of default by approximately 13%. Descriptive analysis shows that loan grade G has the highest median interest rate, while lower grades have progressively lower rates.

In summary, higher loan grades carry higher default risk, primarily due to their higher interest rates.



Insights:

Reasons of Taking Loans

Most clients take loans for debt consolidation, combining multiple debts into one. While it can help manage debt, it also increases risk if repayment fails. The second most common reason is paying off credit cards. Most take loan is loan grade B and loan grade C.

Recommendation

It is recommended to strengthen risk assessment for borrowers, especially those taking loans for debt consolidation and credit card repayment, as they tend to carry higher existing debt burdens. Additionally, interest rates and credit limits should be tailored according to each loan grade's risk profile to effectively minimize the risk of default.

Features Engineering

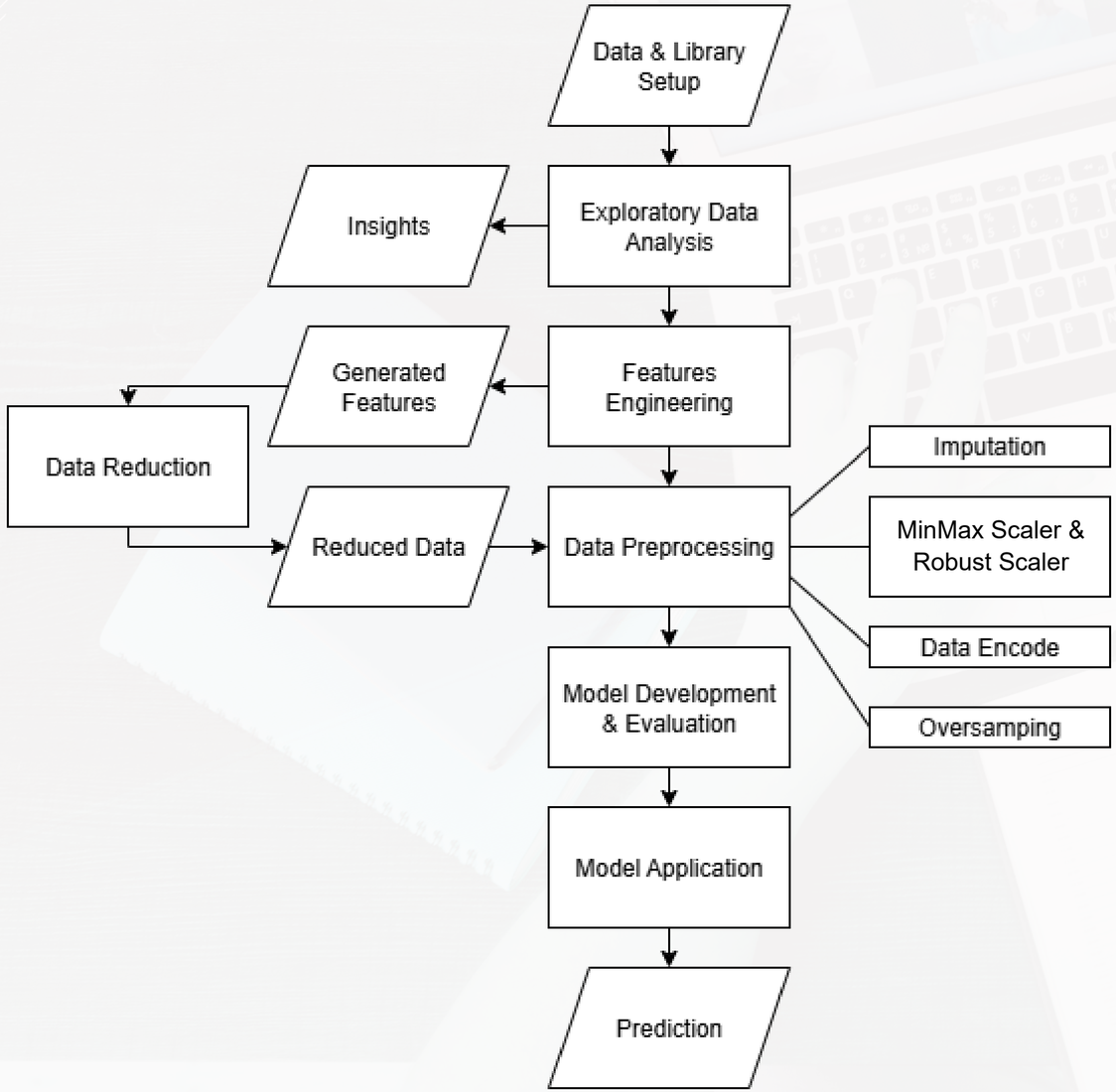
Generated features:

- in_risk: Simplified loan status indicating potential risk.
- first_credit_to_issue_gap: Time gap between the creation of the first credit account and the loan issuance.
- income_source_verification: Simplified status of income source verification.
- long_term: Simplified representation of loan term duration.
- initial_listing_whole: Simplified version of the initial loan listing.

Final Features

Features	
loan_amnt	revol_bal
int_rate	revol_util
sub_grade	collections_12_mths_ex_med
emp_length	acc_now_delinq
home_ownership	tot_cur_bal
annual_inc	first_credit_to_issue_gap
dti	income_source_verification
delinq_2yrs	long_term
inq_last_6mths	initial_listing_whole
open_acc	in_risk
pub_rec	

Modelling Workflow



Data Splitting

Since there is no label on testing dataset, the training dataset are splitted into two sets, data train and data validator within 80:20 ratio respectively, to evaluate the model performance

Applied Models and Its Evaluation

Model	Accuracy	Recall (class 0)	Precision (class 0)
Logistic Regression	0,628	0,626	0,932
Decision Tree	0,791	0,866	0,865
Random Forest	0,881	0,987	0,891
XGBoost	0,887	0,998	0,998
MLP	0,748	0,796	0,909

Model Evaluation Metrics:

The main metric could be recall for class 0, since the company's primary aim is to approve loans to the right borrowers (those likely to repay) in order to reliably earn returns from the interest they pay.

Model Impact Analysis

	Pred 0	Pred 1
Actual 0	82.686	144
Actual 1	10.357	70

Safe Borrowers
Approved: 82.686
Incorrectly rejected: 144 rejected

Risky Borrowers
Approved: 10.357
Rejected (correctly detected): 70

→ The model captures nearly all profit-generating borrowers, but this focus on maximizing safe approvals also allows more risky borrowers through as a trade-off.

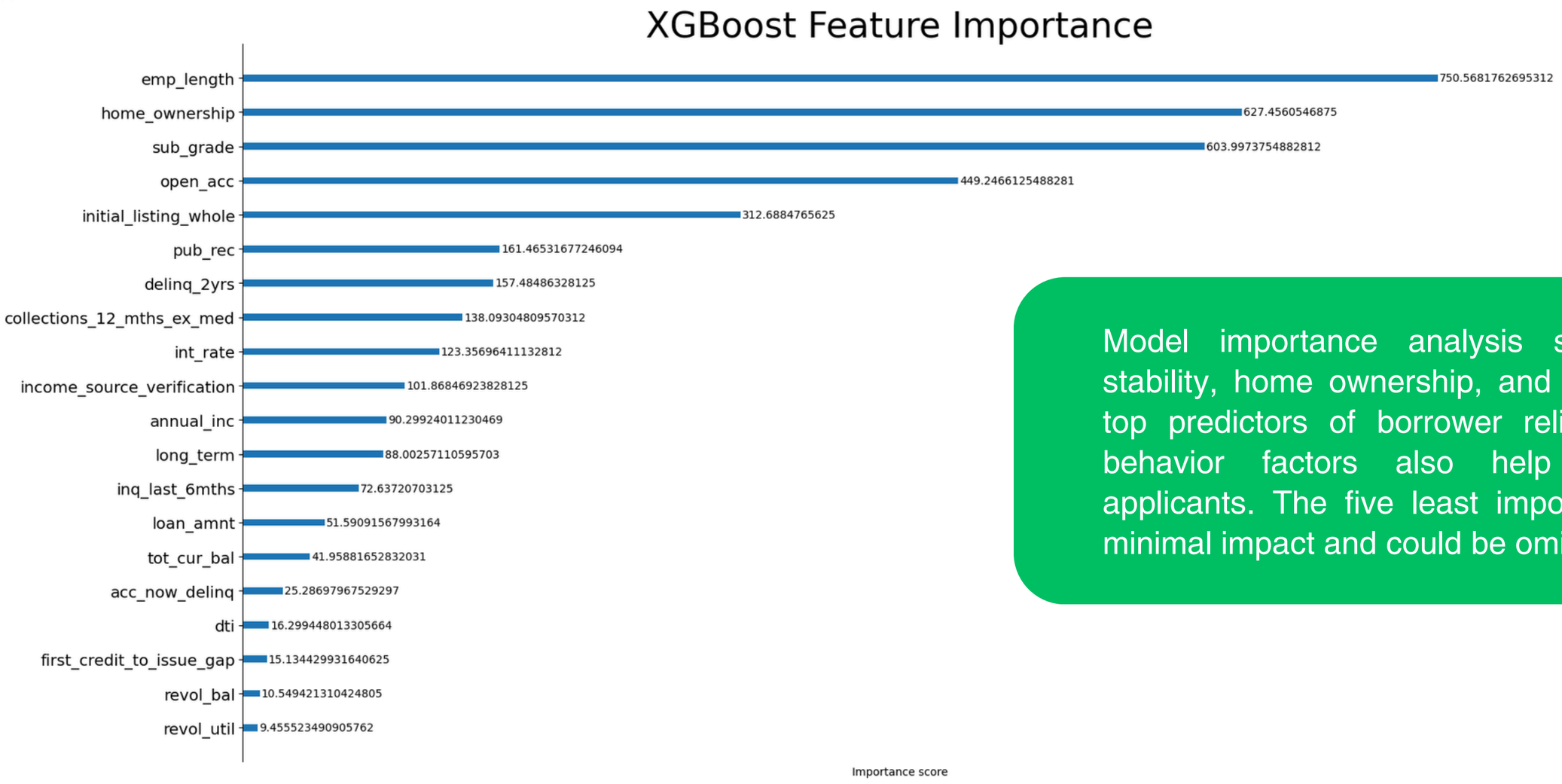
Model Impact Analysis: Financial Simulation

NB:
Median of interest rate: 13.66%
Median of loan amount: 12,000

Impact Type	Count	Unit Value	Total Value	Interpretation
Revenue from Safe Approvals	82,686	$1,639	$135,537,354	from 99,8% safe borrowers approved
Loss from Risky Approvals	10,357	−$12,000	−$124,284,000	from 10,3k risky borrowers approved
Net Profit Impact	—	—	**$11,253,354**	net impact from safe vs risky approvals

The model successfully generates more than $11 million in net return.

Model Interpretability



Model importance analysis shows employment stability, home ownership, and loan sub-grade are top predictors of borrower reliability, while credit behavior factors also help identify high-risk applicants. The five least important features have minimal impact and could be omitted.

Recommendation

Business:

Focus lending strategies on borrowers with stable employment, home ownership, and higher loan sub-grades, as these are the strongest indicators of repayment reliability. Monitor key credit behavior variables, such as open accounts, past collections, and public records, to manage risk. Less important factors (bottom features) can be deprioritized in decision-making, simplifying the approval process without affecting predictive accuracy.

Technical:

Focus on top features like employment length, home ownership, and loan grade, and consider removing low-importance features to simplify the model. Fine-tune the classification threshold if reducing risky approvals is needed. Regularly retrain the model and monitor its performance to maintain accuracy over time.

Things to Adjust

Interest Rate and Loan Regulations by Grade.

Impact of Model Implementation

Developed model could help in generating net return.

Main Factors of Default Risk

Employment length, home ownership, and loan grades.

Conclusions

THANK YOU

Access the notebook here

GitHub / Repo

Bhima Fairul Rifqi | Data, Machine Learning, & Statistics

fairulrifqi962@gmail.com

+62 81347492018

linkedin.com/in/fairulrifqi

github.com/fairulrifqi