Business Flow Understanding

ID/X PARTNER X RAKAMIN
VIRTUAL INTERNSHIP EXPERIENCE

MUHAMMAD KHALIFA UMANA

Credit risk assesment project

LOAN APPLICATION

User / borrower

Admin / Assesor

Step-by-step loan application:

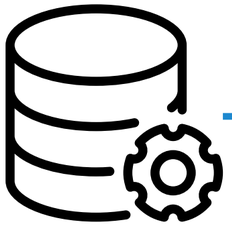
1. user apply a loan and filling required data
2. complimentary outsourced data will be acquired to help the assessment process
3. the data is recorded in the database
4. the application will be assessed by admin
5. the approval result will be based on admin evaluation

User apply for a loan

User-filled data

Credit complimentary data

Data Recorded



Data send into admin

Admin Assesing the data

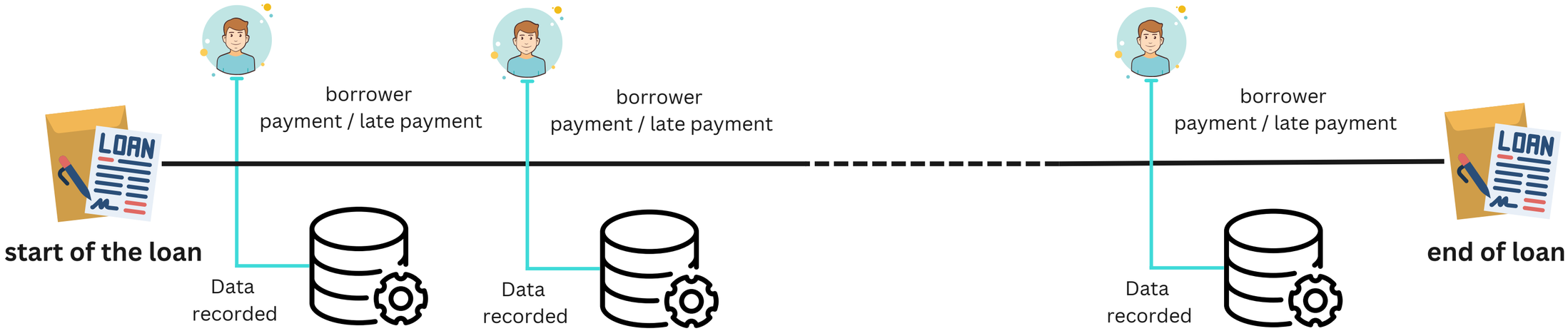
Approval decision

LOAN ACTIVITY

User / borrower

what **happen** during the loan is active:

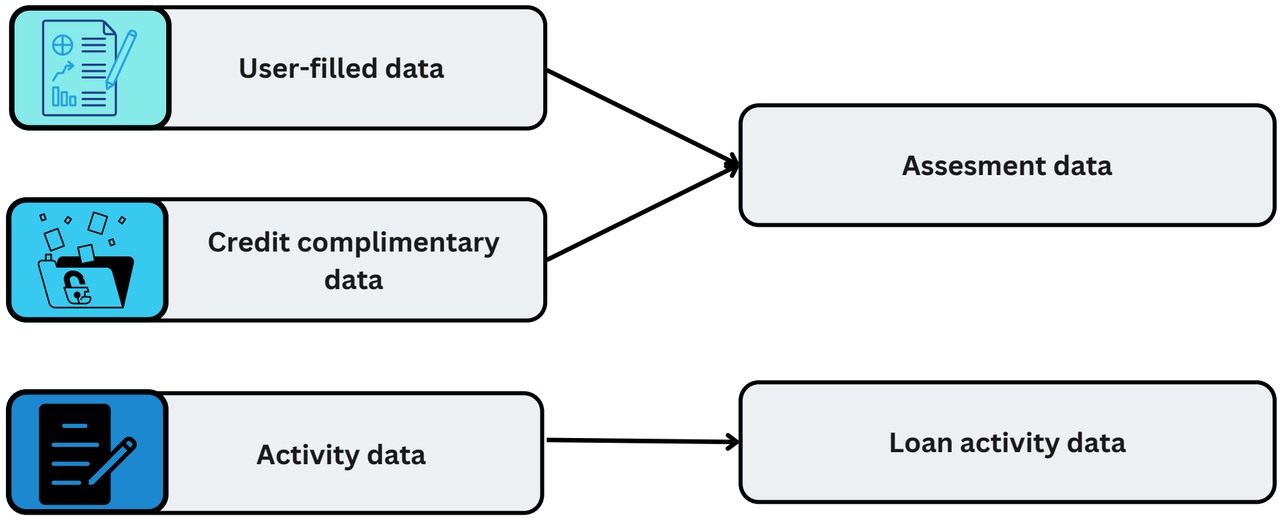
- every transaction will be recorded based on what action is performed
- if the record is already exist, current transaction data will **update** the record in database
- **loan status** will be updated in the end of loan



DATA TYPE BASED ON ITS COLLECTION

 **User-filled data**	Data that acquired from input of borrower e.g employment, duration of employment, annual income
 **Credit complimentary data**	Data that acquired from outside soruce that available to be accessed e.g public record, account number
 **Activity data**	This data is sourced from the activity toward the loans which recorded inside the company e.g upcoming payment day, last payment amount

DATA TYPE BASED ON ITS COLLECTION

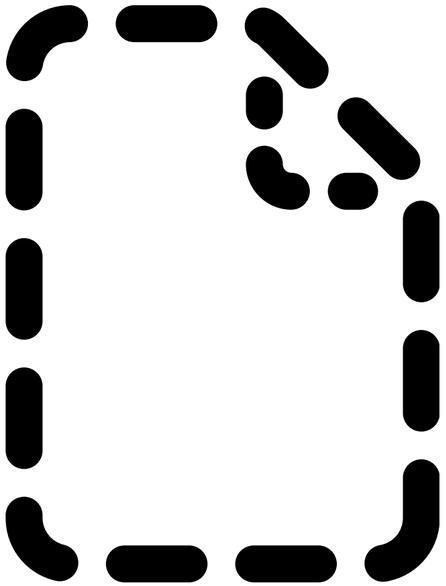


Based on the data acquiring process and how it will be used, the data divided into 2 category in which are:

- Assesment data
- Loan Activity data

This category will divide the data timeline into **before** and **after** the approval process

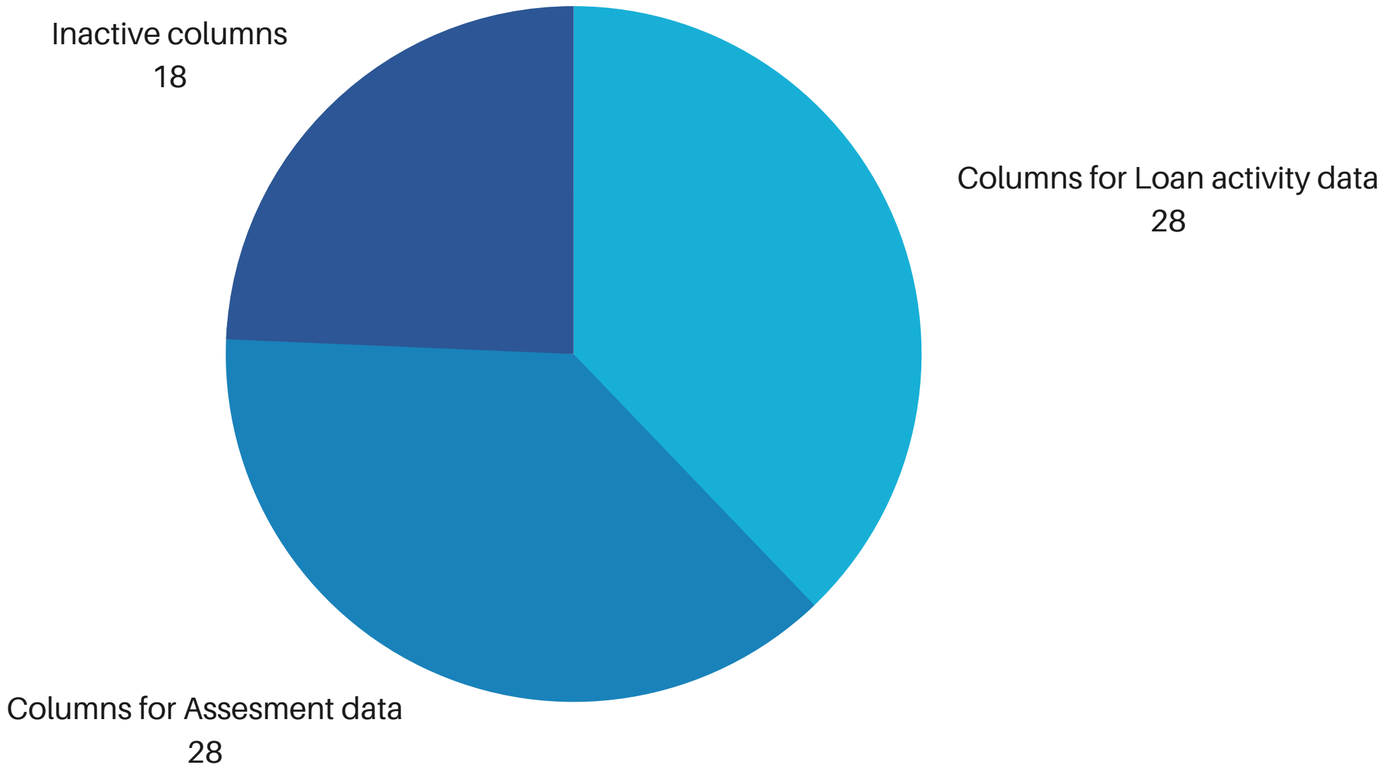
COLUMNS WITH NO DATA



Looking further there are **17 columns** with no input at all (100% empty)

These feature will be considered as **Inactive** since it is not being used in the loan process

SUMMARY OF DATA CATEGORIZATION



The data category will determine on how the data will be used

As Assessment data will be available to be accessed pre-approval of the loan, this category will be used on the machine learning modeling and analysis

With the Loan activity data in which only exist after the loan is active, it can be used for data analaysis

Inactive Columns will not be used at all

There is a column named "policy_code" in which only have 1 as the input value therefore it will be considered as inactive since it does not contain any information to be used

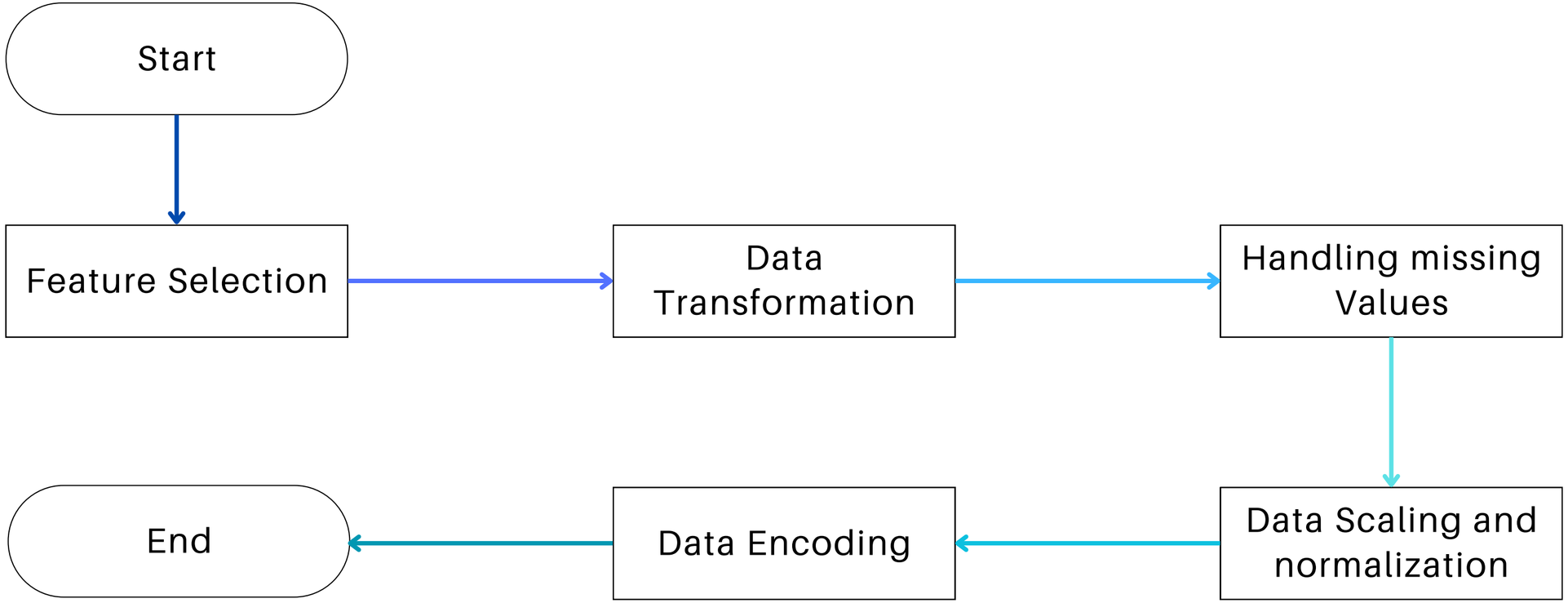
Data Preprocessing

ID/X PARTNER X RAKAMIN
VIRTUAL INTERNSHIP EXPERIENCE

MUHAMMAD KHALIFA UMANA

Credit risk assesment project

DATA PREPROCESSING STEPS



DATA PREPROCESSING STEPS

Feature Selection

What to consider to be selected:

- Active columns (not 100% missing values)
- Not a category column with high cardinality (e.g state_addr, zip_code), we limit to max 14 classes
- Disregarding pymnt_plan because of the unbalance class only 7 input for 'yes' vs 466.284 'no'

Target feature selection:

- considering 'loan_status' to define the creditor bad or good.

DATA PREPROCESSING STEPS

Data Transformation

Target feature transformation('loan_status'):

- 'Charged Off', 'Default', 'Late (31-120 days)', 'Late (16-30 days)', and 'Does not meet the credit policy. Status:Charged Off' are cosidered as **BAD**
- 'Current', 'Fully Paid', 'In Grace Period', and 'Does not meet the credit policy. Status:Fully Paid' are cosidered as **GOOD**

DATA PREPROCESSING STEPS

Data Transformation

After examining further there are 2 numerical columns that can be transformed into a categorical columns. Since they are consist more than

The columns are **mths_since_last_record** and **mths_since_last_delinq** then it transformed to **last_record** and **last_delinq** to indicates when is the last of respective information occured. then the null can be identified as no record/delinquencies

both columns then binned into:

- 'under 3 Months'
- 'under 6 months'
- 'under a year'
- 'under 5 year'
- 'under 10 years'
- 'more than 10 years'
- 'no records/delinq'

DATA PREPROCESSING STEPS

Data Transformation

Another columnd transformed is that the categorization of **emp_length** reduced the class number so from how many years the length of employment into range of :

- 1 year and under
- 3 years and under
- 5 years and under
- under 10 years
- 10 years and more

This also helps inputing the missing data into a new class '**No Employment**'

DATA PREPROCESSING STEPS

Handling missing Values

Since some of the missing values are handled on the data transformation step, the rest of missing data is dropped as the columns with missing data are no more than 0.03% of the data.

DATA PREPROCESSING STEPS

Data Scaling and normalization

- There are counts table that do not need any scaling/normalization such as pub_rec, delinq_2yrs, collections_12_mths_ex_med, inq_last_6mths
- on open_acc, scaling could be implemented but as it is a count table and the data range is 0-84, no scaling is necessary
- for dti feature as it is a ratio and ranging from 0-39.9, it suggests that the values are already on a relatively similar scale and cover a limited range. In this case, the need for scaling becomes less critical.
- loan_amnt and anual_inc are large and vary significantly, scaling the feature will be necessary

DATA PREPROCESSING STEPS

Data Encoding

All the encoding applied to categorical columns using one hot encoding which are applied to:

- 'purpose'
- 'emp_length'
- 'home_ownership'
- 'last_record'
- 'verification_status'
- 'last_delinq'
- 'term'

this is the ends of data preprocessing steps resulting in total of 53 columns for training feature.

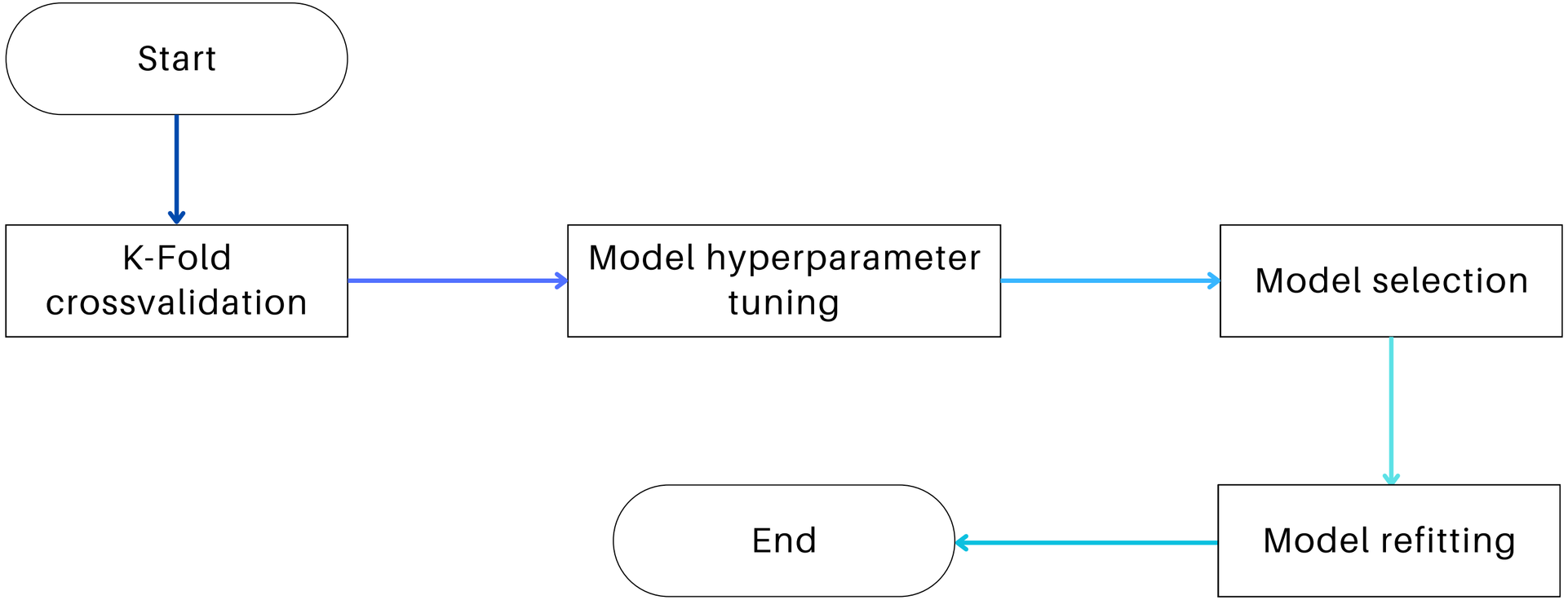
Model Training

ID/X PARTNER X RAKAMIN
VIRTUAL INTERNSHIP EXPERIENCE

MUHAMMAD KHALIFA UMANA

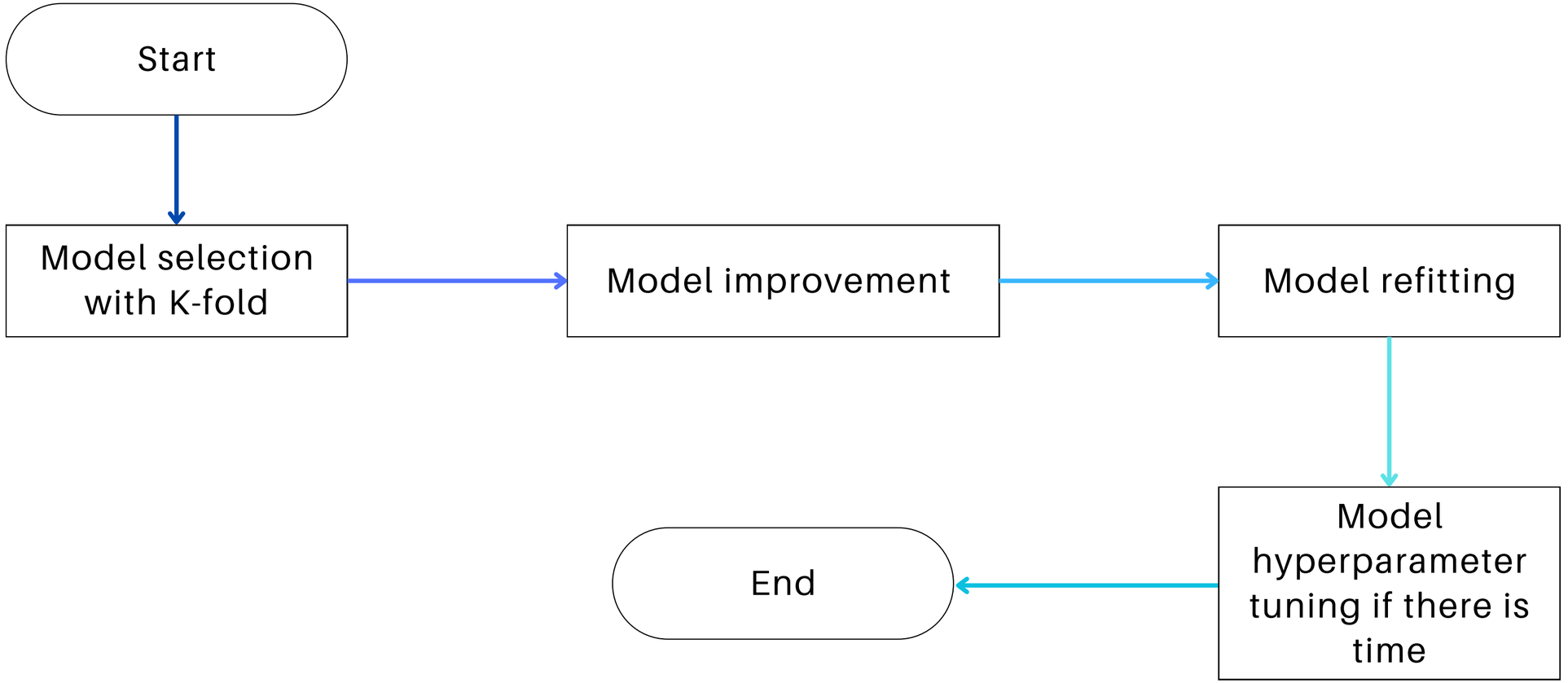
Credit risk assesment project

MODEL TRAINING STEPS (IDEALLY)



MODEL TRAINING STEPS (ADJUSTED)

Due to the limited amount of time, we will readjust the steps to:



MODEL TRAINING STEPS

Model selection

Firstly all classification model are trained using the preprocessed data and default setting for hyperparameter using 5 fold of K-fold validation.

Then we examined which model is the best using current data (without class balancing or hyperparameter tuning).

we want to highlight the performance metric of Recall because we also want to consider not to reject all potentially defaulted borrower, late payments and other penalties in which add some contribution to the company.

MODEL TRAINING STEPS

Model selection

Model	Accuracy	Precision	Recall	F1-Score
Linear Regression	88.74%	44.83%	0.12%	0.25%
RidgeClassifier	88.76%	33.33%	0.01%	0.02%
BaggingClassifier	88.20%	19.73%	1.66%	3.06%
GradientBoostingClassifier	88.77%	62.5%	0.05%	0.10%
RandomForestClassifier	88.70%	20.68%	0.23%	0.45%
DecisionTreeClassifier	79.08%	13.15%	15.40%	14.20%

MODEL TRAINING STEPS

Model selection

We observed that the best model is DecisionTreeClassifier but the recall is low but with high accuracy, this might indicates imbalance class.

next we will apply SMOTE for the clas balancing.

MODEL TRAINING STEPS

Model improvement

We applied SMOTE for the DecisionTreeClassifier resulting in

Model	Accuracy	Precision	Recall	F1-Score
DecisionTreeClassifier with SMOTE	99.99%	100%	99.99%	99.99%

It seems that it is now overfitting, now let's proceed to use other model but applying SMOTE

MODEL TRAINING STEPS

Model improvement

We applied SMOTE for the RandomForestClassifier and XGBoostClassifier resulting in

Model	Accuracy	Precision	Recall	F1-Score
RandomForestClassifier with SMOTE	99.99%	100%	99.99%	99.99%
XGBoostClassifier with SMOTE	91.93%	99.68%	84.12%	91.24%

It seems that using RandomForestClassifier still resulting on overfit problem. But using XGBoost now is at the acceptable rate of model performance metric

MODEL TRAINING STEPS

Model improvement

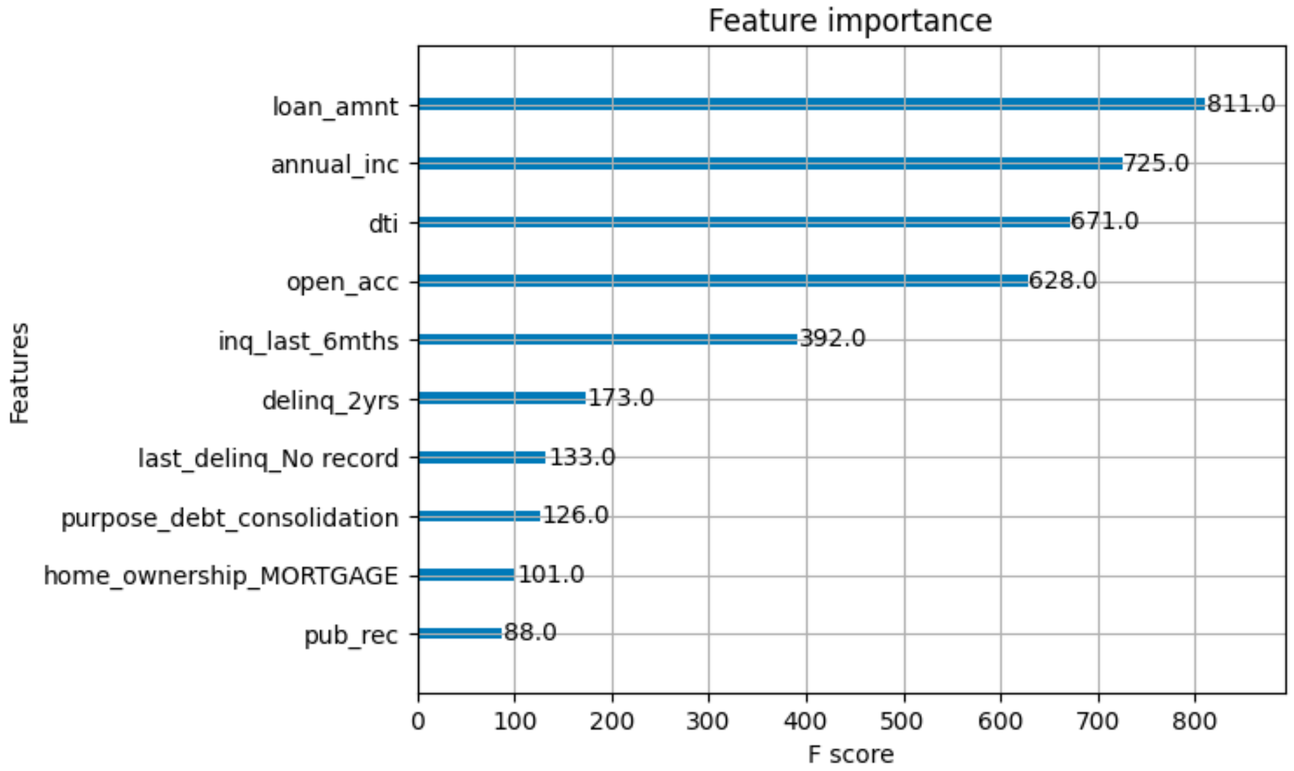
There are some references that imbalance class is not really a problem so we tried XGBoost without SMOTE because it is not tried in the previous model selection

Model	Accuracy	Precision	Recall	F1-Score
XGBoostClassifier without SMOTE	91.60%	99.62%	83.59%	90.90%

Okay now the result are not much different so it seems SMOTE is improving only **some** model but in XGBoost it does not showing much effect.

MODEL TRAINING STEPS

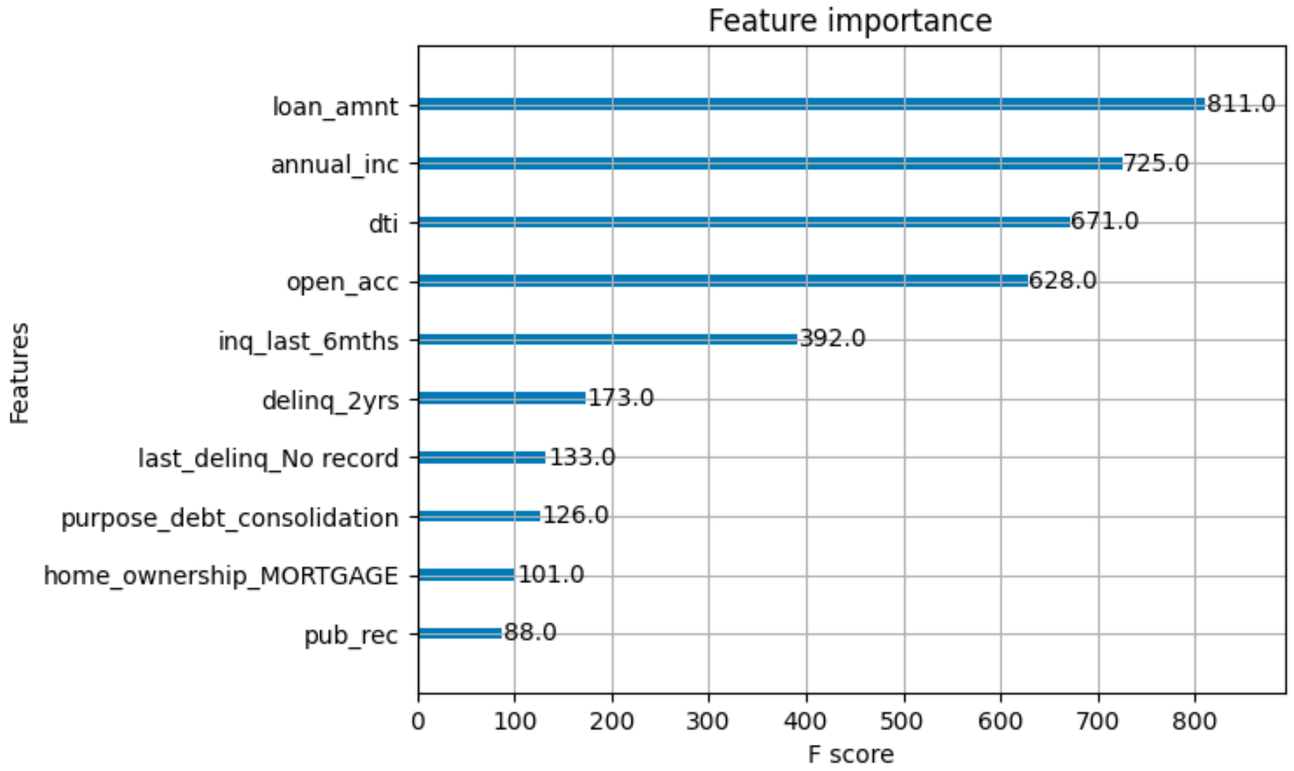
Model improvement



observing the top 10 of feature importance, the top 5 feature are only consist of numerical feature and the categorical only shows after the top 6 (last_delinq).

MODEL TRAINING STEPS

Model improvement



To reduce the computation power, we might want to see the model performance by only calculating 5 features.

MODEL TRAINING STEPS

Model refitting

XGBoost only with 5 feature training:

Model	Accuracy	Precision	Recall	F1-Score
XGBoostClassifier (5 important features only)	89.12%	93.30%	84.40%	88.63%

Now the model will need less input which will lead to faster computing but not much differences in performances as before.

MODEL TRAINING STEPS

Model hyperparameter tuning if there is time

unfortunately with the time limit of this project, the hyperparameter tuning is not be able to be done

Model Deployment

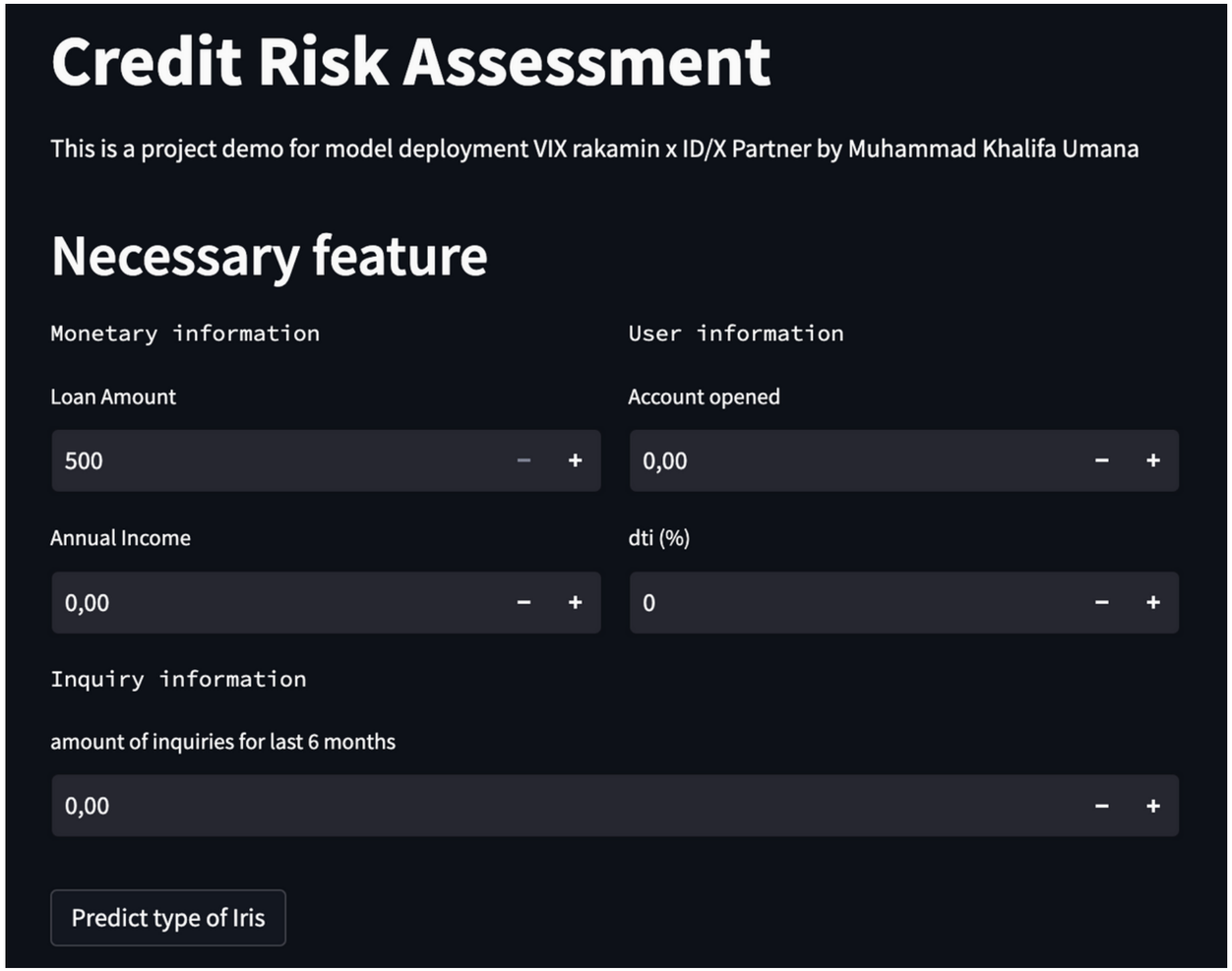
ID/X PARTNER X RAKAMIN
VIRTUAL INTERNSHIP EXPERIENCE

MUHAMMAD KHALIFA UMANA

Credit risk assesment project

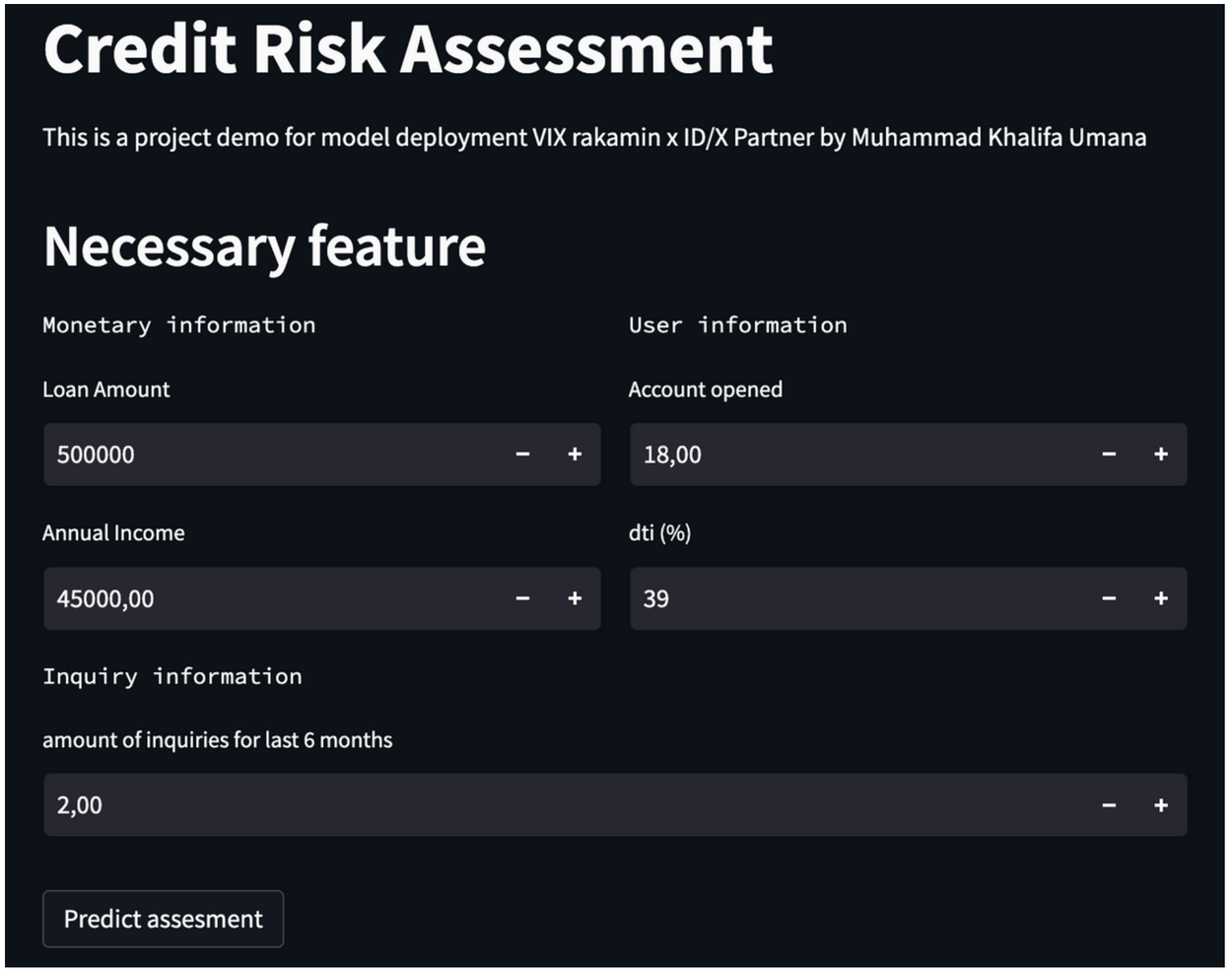
MODEL DEPLOYMENT

Default interface



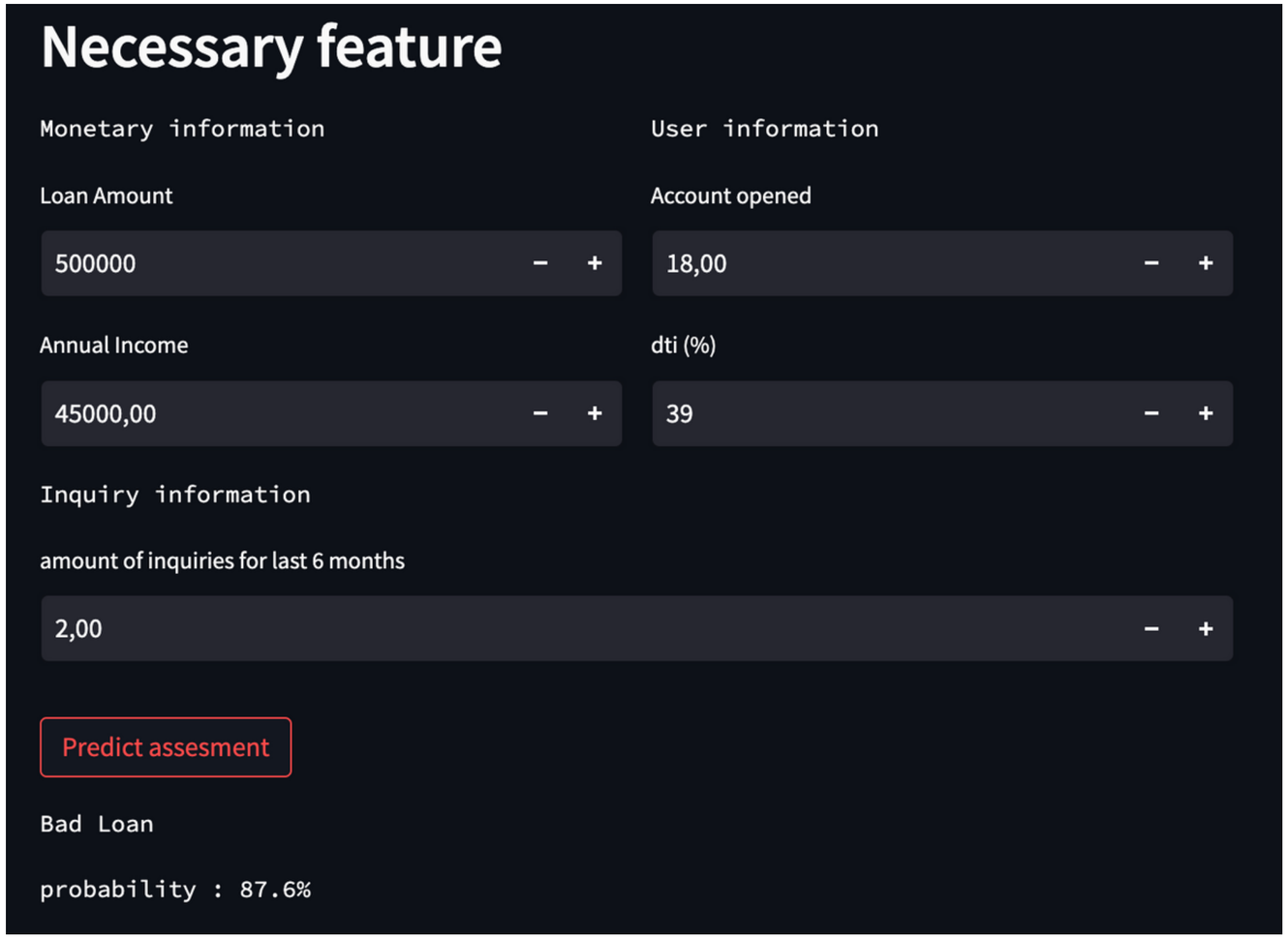
MODEL DEPLOYMENT

Filled form



MODEL DEPLOYMENT

Result of prediction



Evaluation

ID/X PARTNER X RAKAMIN
VIRTUAL INTERNSHIP EXPERIENCE

MUHAMMAD KHALIFA UMANA

Credit risk assesment project

KEYPOINTS TO BE HIGHLIGHTED

- More time can provide more exploration and more experimentation
- Clear business processes can help ease the understanding of data acquisition
- The model still have room for improvement by using hyperparameter tuning
- Data analysis can be help if there is more guidance on the business flow understanding