20 DESEMBER 2021

Final Project Presentation

∂ erivative team

DigitalSkola

∂ erivative team

HackerEarth_how not to lose a customer in 10 days

WHAT HAVE WE DONE

Dataset Understanding

Identify the to do list

Analyzing the Data

Pre-processing

Model development

DigitalSkola

Preface

Churn rate is a marketing metric that describes the number of customers who leave a business over a specific time period.

Churn rate value may be predicted based on multiple factors such as the user's demographic, their browsing behavior, historical purchase data, etc.

On this Final Project, We try to predict the Churn Rate Value between 1 and 5.

∂ erivative team

DigitalSkola

∂ erivative team

Dataset Understanding

HACKEREARTH: HOW TO NOT LOSE A CUSTOMER IN 10 DAYS

DigitalSkola

THE GIVEN DATA MOSTLY CONSISTS OF USERS DATA RECORD AND ACTIVITY RECORD

THEIR RESPECTIVE LABELS CONSISTS OF THEIR CHURN RISK SCORE FROM 1 TO 5 CATEGORIC ORDINAL SCALE

25 Columns:
- 24 Features
- 1 Label

36992 Rows

Features (Pure Labels)

CUSTOMER_ID

the unique identification number of a customer

NAME

the name of a customer

SECURITY_NO

a unique security number that is used to identify a person

REFERRAL_ID

a referral ID

DigitalSkola

∂ erivative team

Features (Categoric Nominals)

GENDER

the gender of a customer

REGION_CATEGORY

the region that a customer belongs to

MEMBERSHIP_CATEGORY

the category of the membership that a customer is using

JOINED_THROUGH_REFERRAL

Represents whether a customer joined using any referral code or ID

∂ erivative team

DigitalSkola

Features (Categoric Nominals)

PREFERED_OFFER_TYPES

the type of offer that a customer prefers

MEDIUM_OF_OPERATION

the medium of operation that a customer uses for transactions

INTERNET_OPTION

the type of internet service a customer uses

USED_SPECIAL_DISCOUNT

Represents whether a customer uses special discounts offered

DigitalSkola

∂ erivative team

Features (Categoric Nominals)

OFFER_APPLICATION_PREFERENCE

Represents whether a customer prefers offers

PAST_COMPLAINT

Represents whether a customer has raised any complaints

COMPLAINT_STATUS

Represents whether the complaints raised by a customer was resolved

LAST_VISIT_TIME

Represents whether a customer has raised any complaints

JOINING_DATE

the date when a customer became a member

∂erivative team

DigitalSkola

Features (Categoric Ordinals)

FEEDBACK

the feedback provided by a customer

DigitalSkola

∂ erivative team

Features (Numeric)

AGE

the age of a customer

DAYS_SINCE_LAST_LOGIN

the no. of days since a customer last logged into the website

AVG_TIME_SPENT

the average time spent by a customer on the website

AVG_TRANSACTION_VALUE

the average transaction value of a customer

AVG_FREQUENCY_LOGIN_DAYS

the no. of times a customer has logged in to the website

POINTS_IN_WALLET

the points awarded to a customer on each transaction

DigitalSkola

∂erivative team

Label (Categoric Ordinal)

CHURN_RISK_SCORE

Represents the churn risk score that ranges from 1 to 5

DigitalSkola

∂ erivative team

EXPLORATORY DATA ANALYSIS

Now we are going to breakdown all of the data we have to get the better understanding of it

DigitalSkola

CHURN RISK SCORE

There is a negative value in churn risk score column, but in application there shouldn't be any negative prediction. The lowest should be 1

FEATURE DATA TYPE CLEANSING

We will make sure every data there were represented by its respective true data-type

VISUALIZATION FOR THE DATA

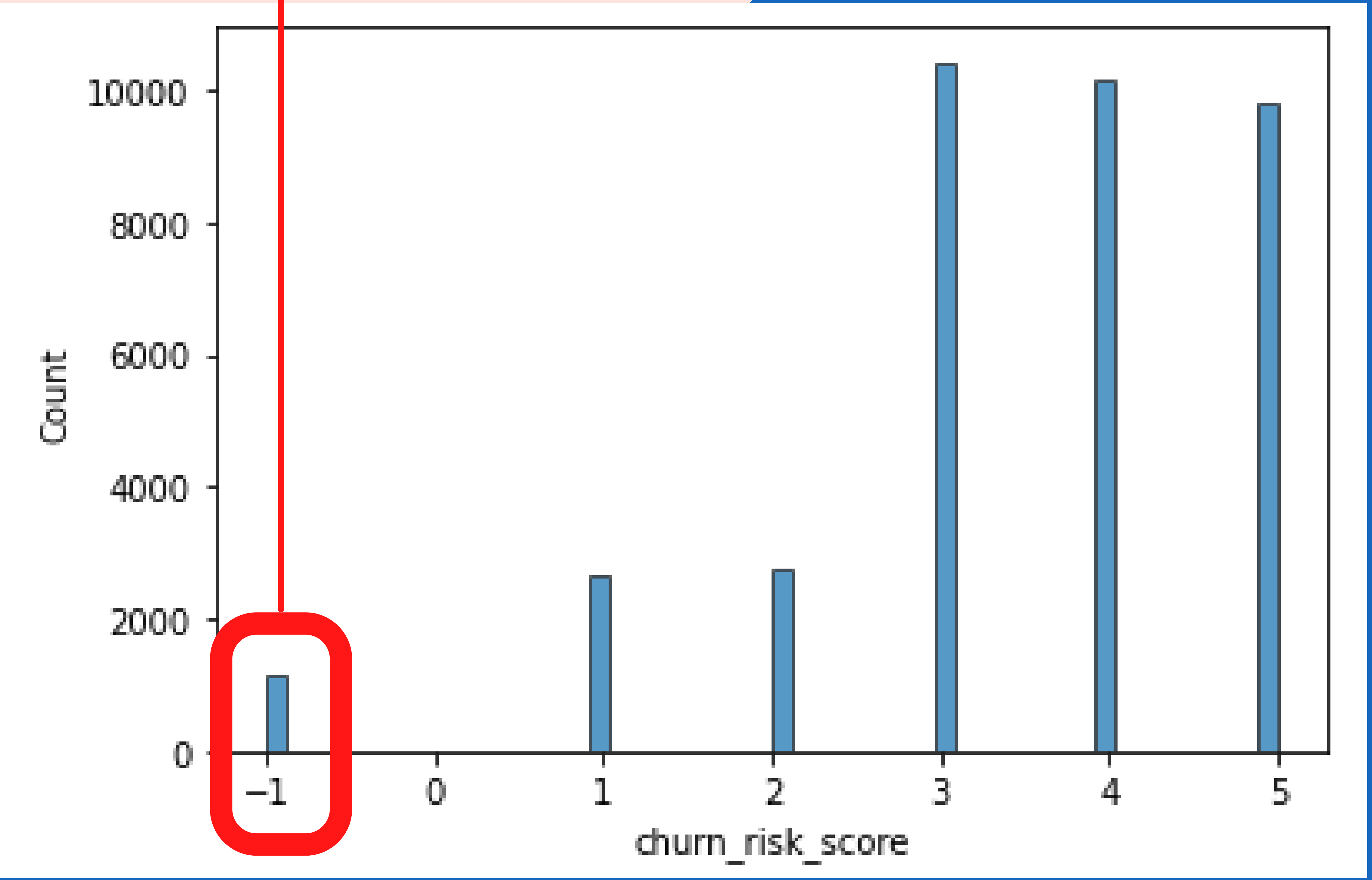
Visualize the data to further understand about the statistical condition of the data

CALCULATE THE CORRELATION BETWEEN FEATURES

Correlation counting to see how every numeric features correlate with one-another

EDA Processing

∂erivative team



There are invalid churn_risk_scores.

We will process further without these data

DigitalSkola

FEATURE DATA TYPE CLEANSING

We will make sure every feature represented by their true data type (i.e.: 123 should be either int or float, ['female', 'male'] should be str or object, and '11-12-2008' into datetime64)

WHAT WE DO?

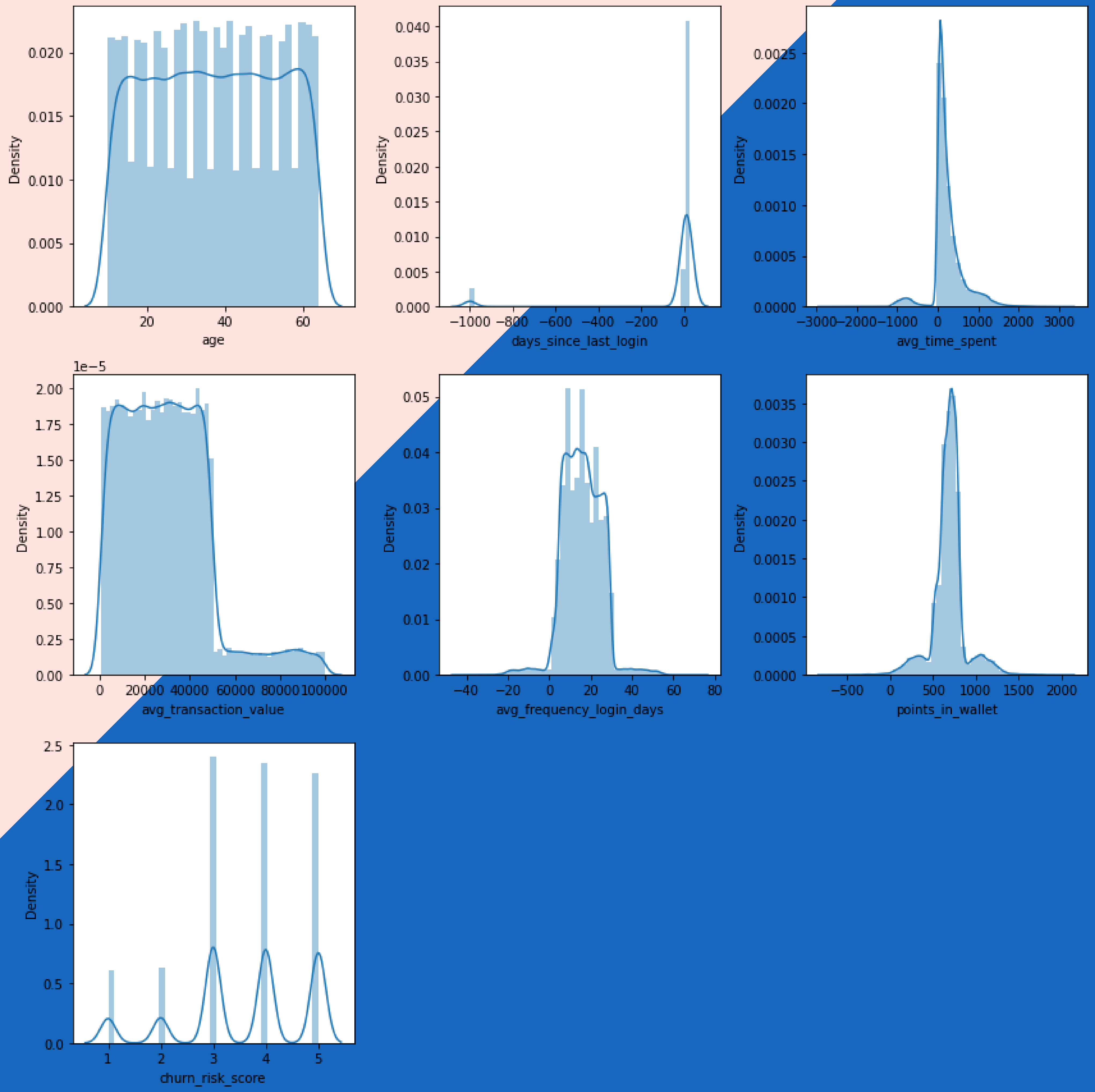
Replace the 'Error' value into NaN for avg_frequency_login_days feature

Change the Data Type into Float for avg_frequency_login_days feature

Change the joining_date feature data type into datetime

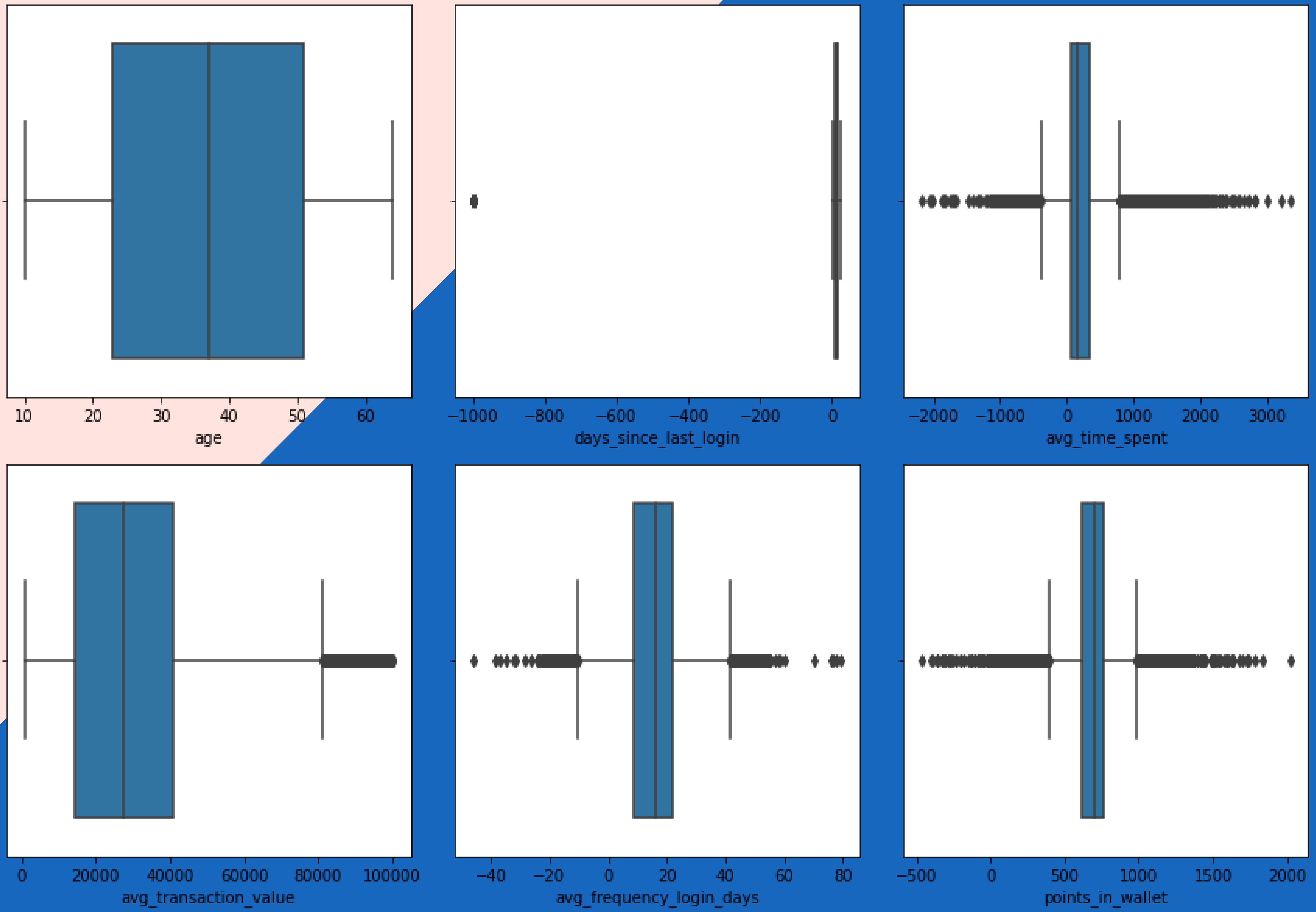
∂erivative team

DATA VISUALIZATION
-DISTRIBUTION PLOT-

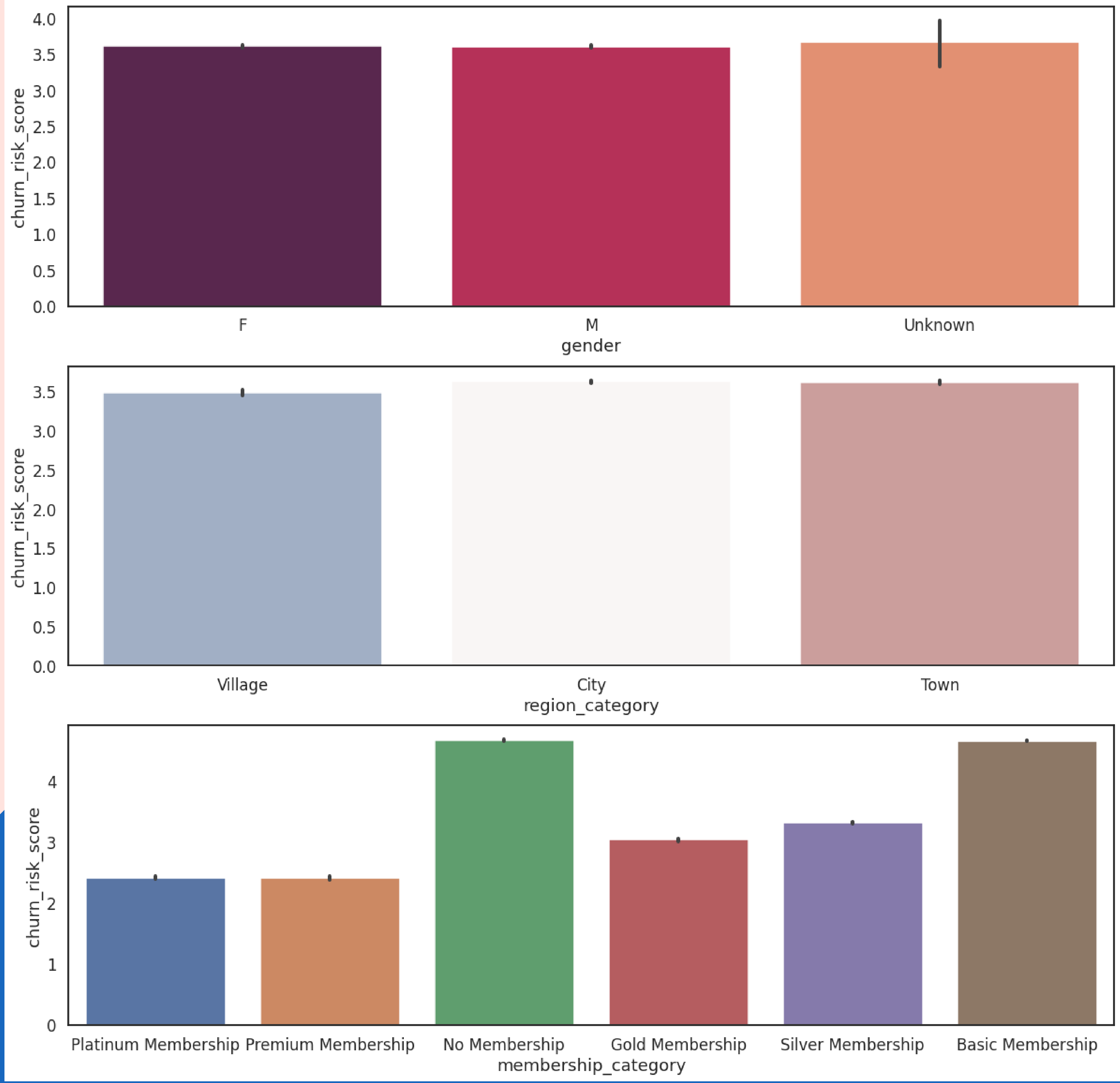


∂erivative team

DATA VISUALIZATION
-BOX PLOT-

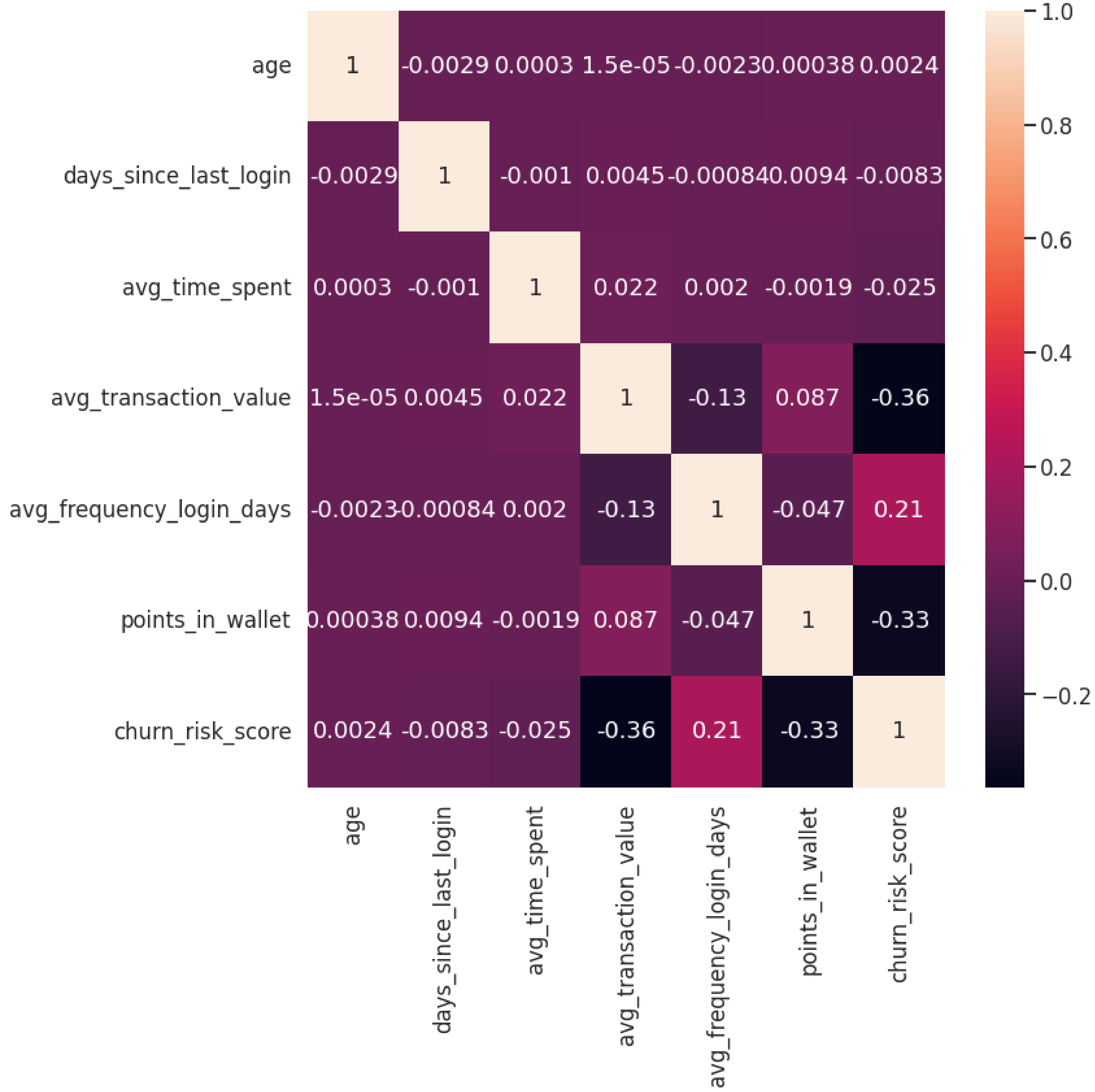


∂ erivative team



DATA VISUALIZATION
-BAR CHART-

∂erivative team



Calculating Correlation Between Features

∂erivative team

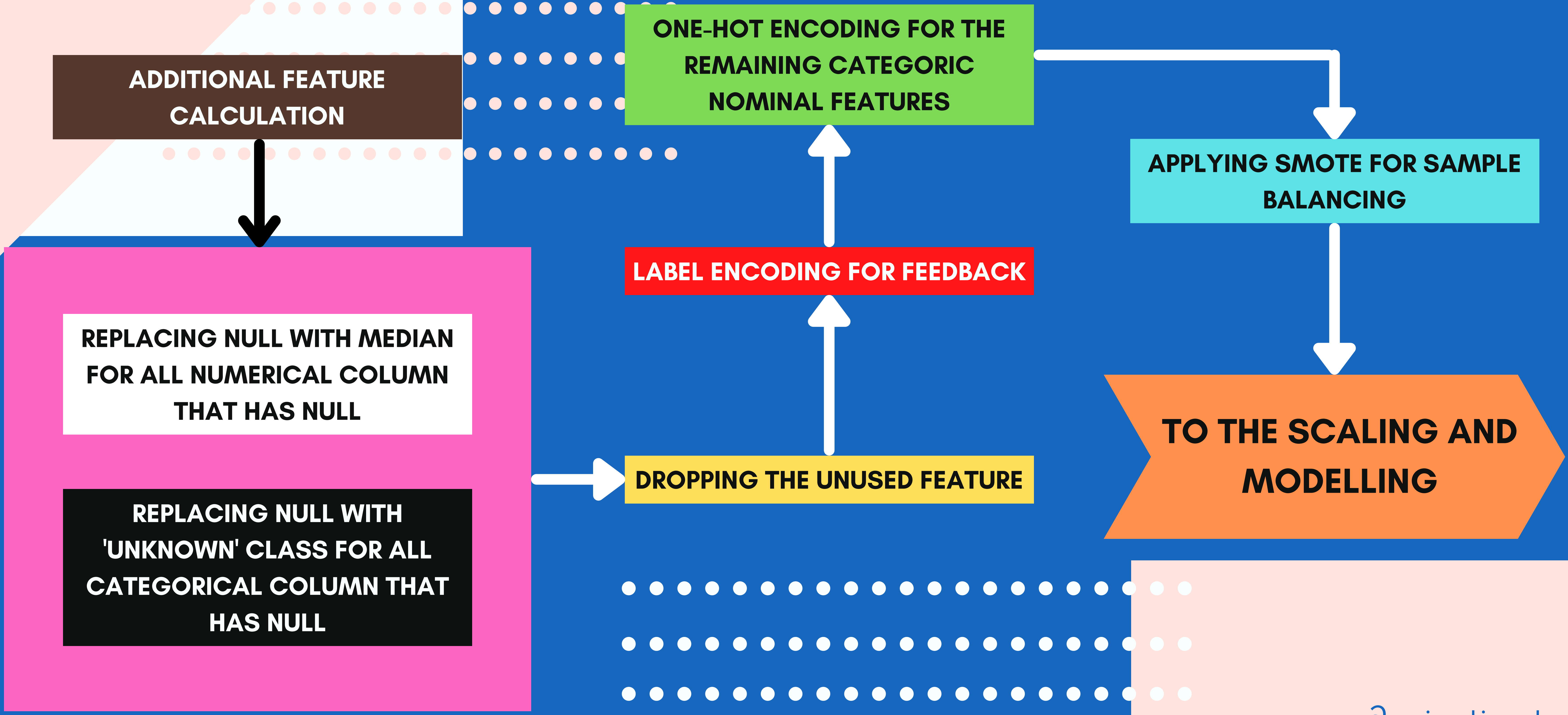
∂ erivative team

DATA PRE-PROCESSING

SO HOW WAS IT?

DigitalSkola

PRE-PROCESSING STEPS



∂erivative team

JOINING DAYS

From **joining_date** feature, will be extracted ***the number of days*** a user had been registered and the calculation will be written in a new column of **joining_days** which has an integer data type

JOINING_DAYS = LAST_JOIN_DATE - JOINING_DATE

ADDITIONAL FEATURE CALCULATION

∂erivative team

DigitalSkola

REPLACE NULL WITH MEDIAN

Untuk sisa kolom fitur numerik yang memiliki nilai Null:
"avg_frequency_login_days",
"points_in_wallet"

REPLACE NULL WITH 'UNKNOWN' CLASS

Untuk sisa kolom fitur kategorik yang memiliki nilai Null:
"region_category",
"preferred_offer_types

∂ erivative team

DROPPING THE UNUSED FEATURE

CUSTOMER_ID

NAME

SECURITY_NO

JOINING_DATE

REFERRAL_ID

LAST_VISIT_TIME

LAST_JOIN_DATE

∂ erivative team

LABEL ENCODING FOR 'FEEDBACK'

0

'POOR WEBSITE'

'NO REASON SPECIFIED'

'POOR PRODUCT QUALITY'

'POOR CUSTOMER SERVICE'

'TOO MANY ADS'

1

'PRODUCT ALWAYS IN STOCK'

"QUALITY CUSTOMER CARE'

'USER FRIENDLY WEBSITE'

'REASONABLE PRICE'

DigitalSkola

ONE HOT ENCODING FOR CATEGORIC NOMINAL COLUMN

GENDER

REGION_CATEGORY

MEMBERSHIP_CATEGORY

JOINED_THROUGH_REFERRAL

PREFERRED_OFFER_TYPES

MEDIUM_OF_OPERATION

INTERNET_OPTION

USED_SPECIAL_DISCOUNT

OFFER_APPLICATION_PREFERENCE

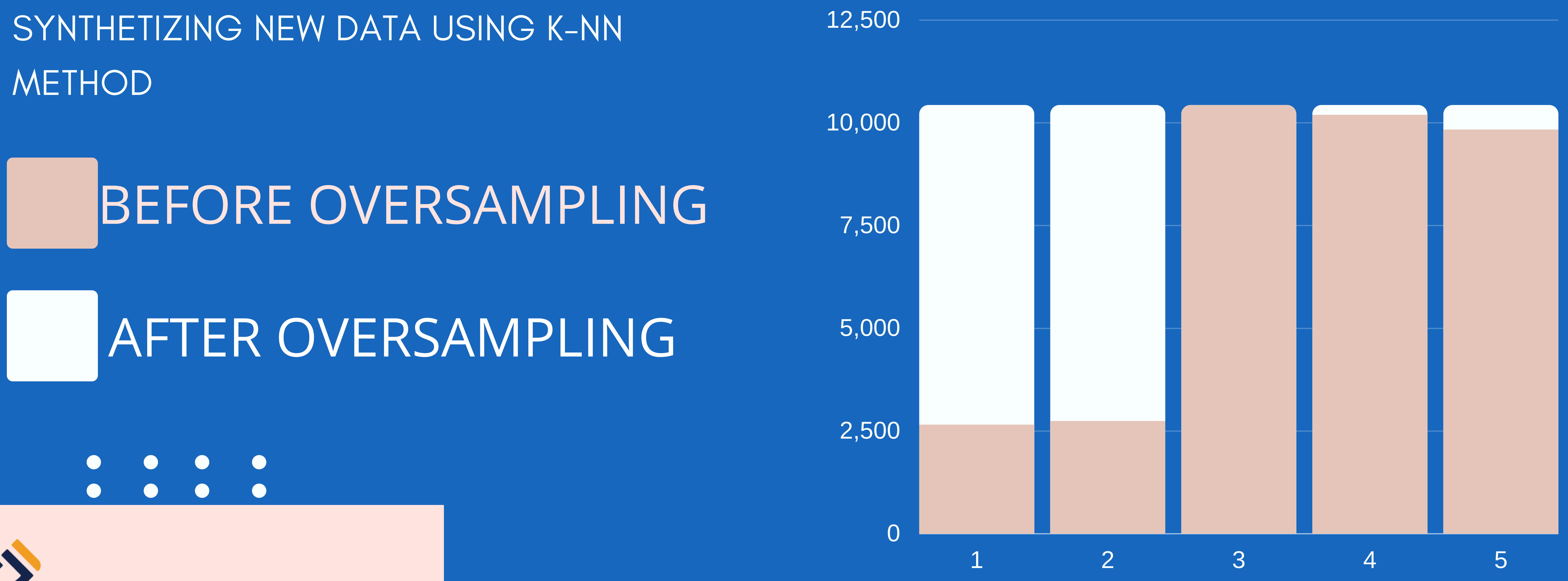
PAST_COMPLAINT

COMPLAINT_STATUS

∂erivative team

SMOTE USAGE

TO BALANCE THE SAMPLE BY SYNTHETIZING NEW DATA USING K-NN METHOD



∂ erivative team

DigitalSkola

USING ROBUST SCALING

```
array([[-0.77042858,  0.625     ,  0.44389089, ...,  0.        ,
         0.        ,  0.        ],
       [-0.20274436,  0.5       ,  0.46372345, ...,  1.        ,
         0.        ,  0.        ],
       [ 0.28384211,  0.25      ,  1.19249174, ...,  0.        ,
         1.        ,  0.        ],
       ...,
       [-0.33719064,  1.37205026, -0.47785962, ...,  0.        ,
         0.        ,  0.        ],
       [ 0.63767368, -0.24357934,  0.92370912, ...,  0.98287825,
         0.01712175,  0.        ],
       [ 0.07573499,  0.07765599, -0.37156102, ...,  0.        ,
         0.        ,  0.        ]])
```

TO MAKE SURE EVERY FEATURE HAS THE FAIR RANGE OF VALUE

∂ erivative team

∂ erivative team

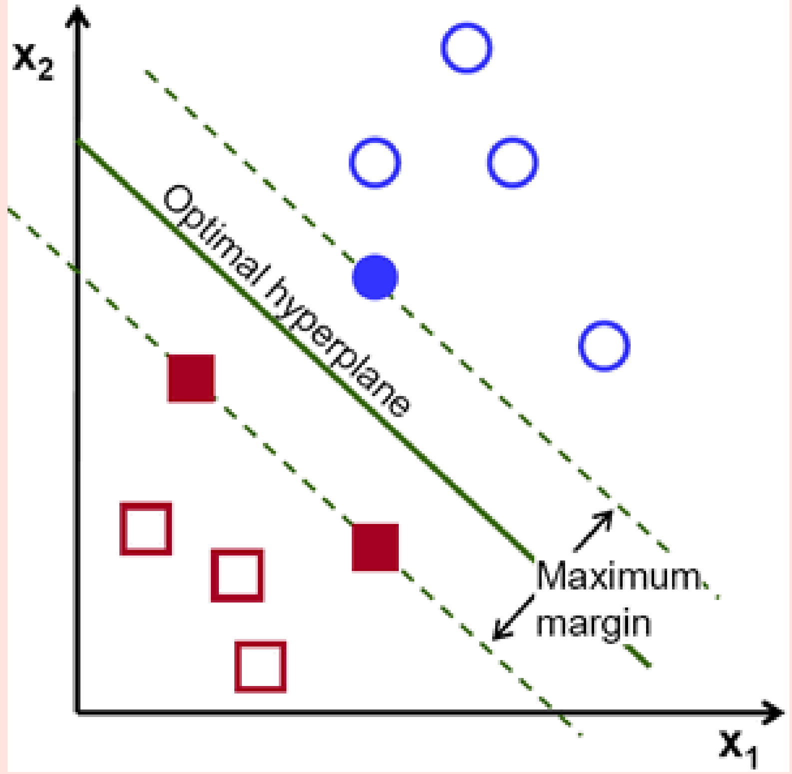
MODELLING

OF COURSE NOT A CATWALK MODEL

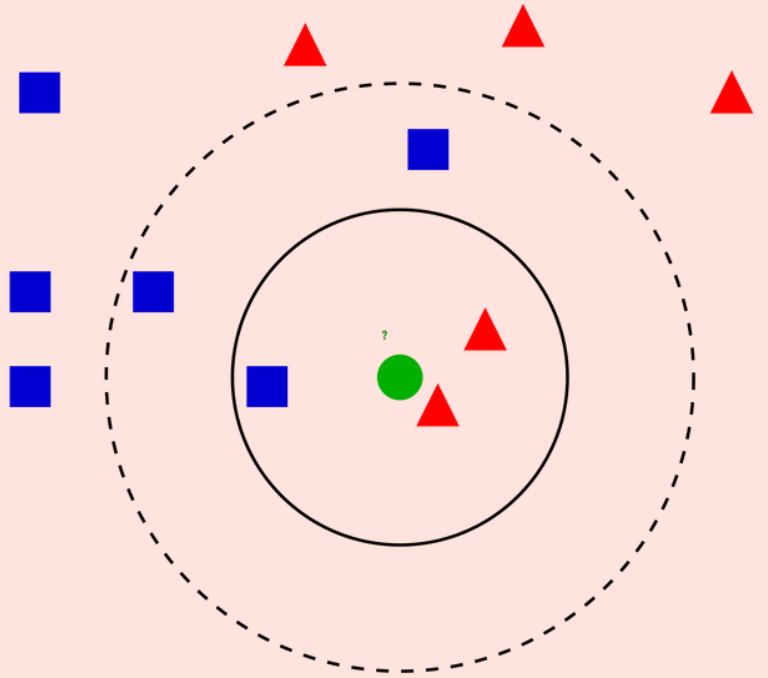
DigitalSkola

METHOD USED

RANDOM FOREST



SVM



K-NEAREST NEIGHBOR

dmlc
XGBoost

∂erivative team

DigitalSkola

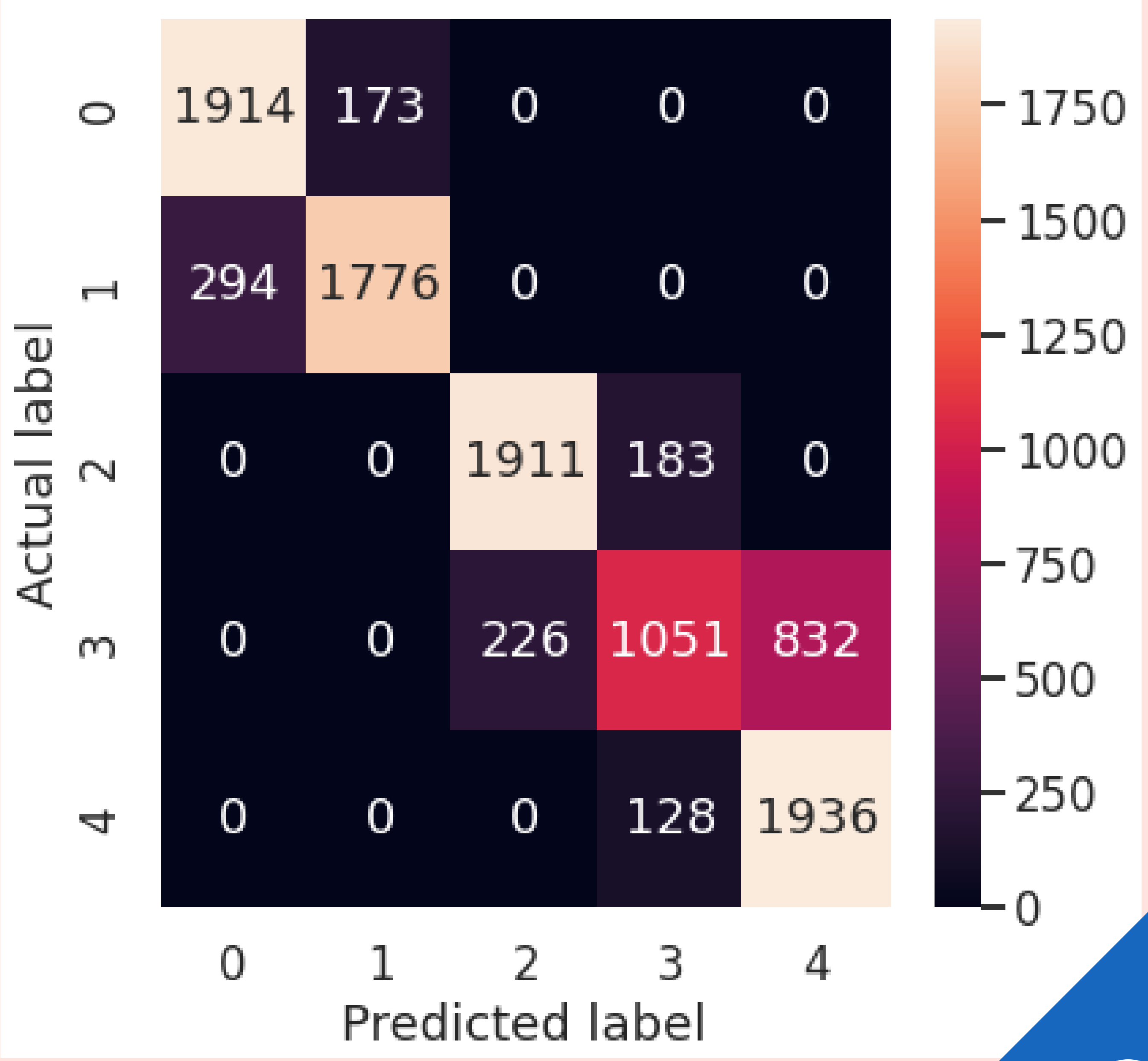
RANDOM FOREST

f1-score :0.8170634148645475

accuracy : 0.8238679969301612

	precision	recall	f1-score	support
1	0.87	0.92	0.89	2087
2	0.91	0.86	0.88	2070
3	0.89	0.91	0.90	2094
4	0.77	0.50	0.61	2109
5	0.70	0.94	0.80	2064
accuracy			0.82	10424
macro avg	0.83	0.82	0.82	10424
weighted avg	0.83	0.82	0.82	10424

Submission ID: 67376341	Result	Score
21 seconds ago	Accepted	75.28612



Our Model

∂ erivative team

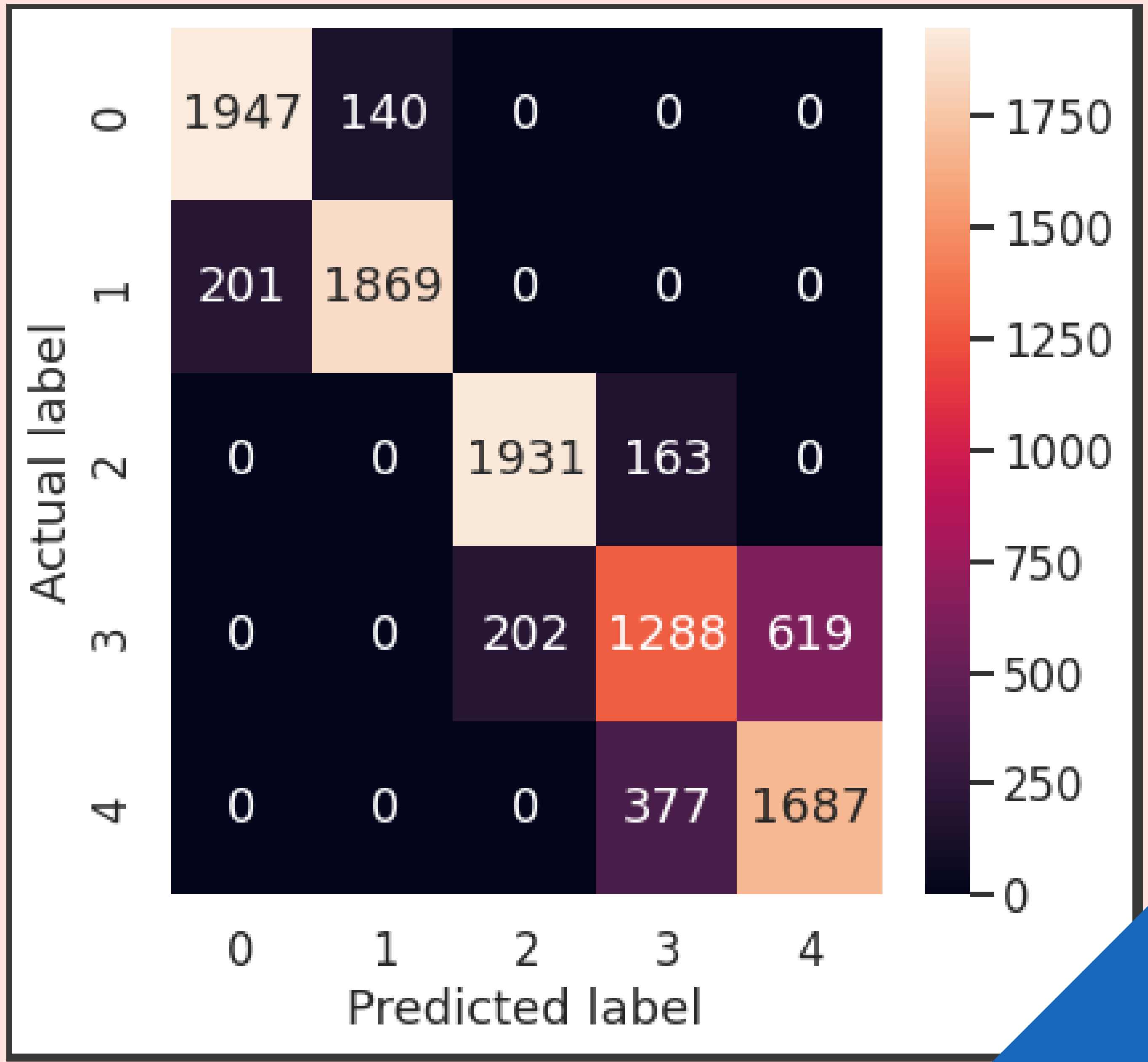
XGBOOST

f1-score : 0.8351839259364959

accuracy : 0.8367229470452802

	precision	recall	f1-score	support
1	0.91	0.93	0.92	2087
2	0.93	0.90	0.92	2070
3	0.91	0.92	0.91	2094
4	0.70	0.61	0.65	2109
5	0.73	0.82	0.77	2064
accuracy			0.84	10424
macro avg	0.84	0.84	0.84	10424
weighted avg	0.84	0.84	0.83	10424

Submission ID: 67370677	Result	Score
15 seconds ago	Accepted	75.63348



Our Model

∂ erivative team

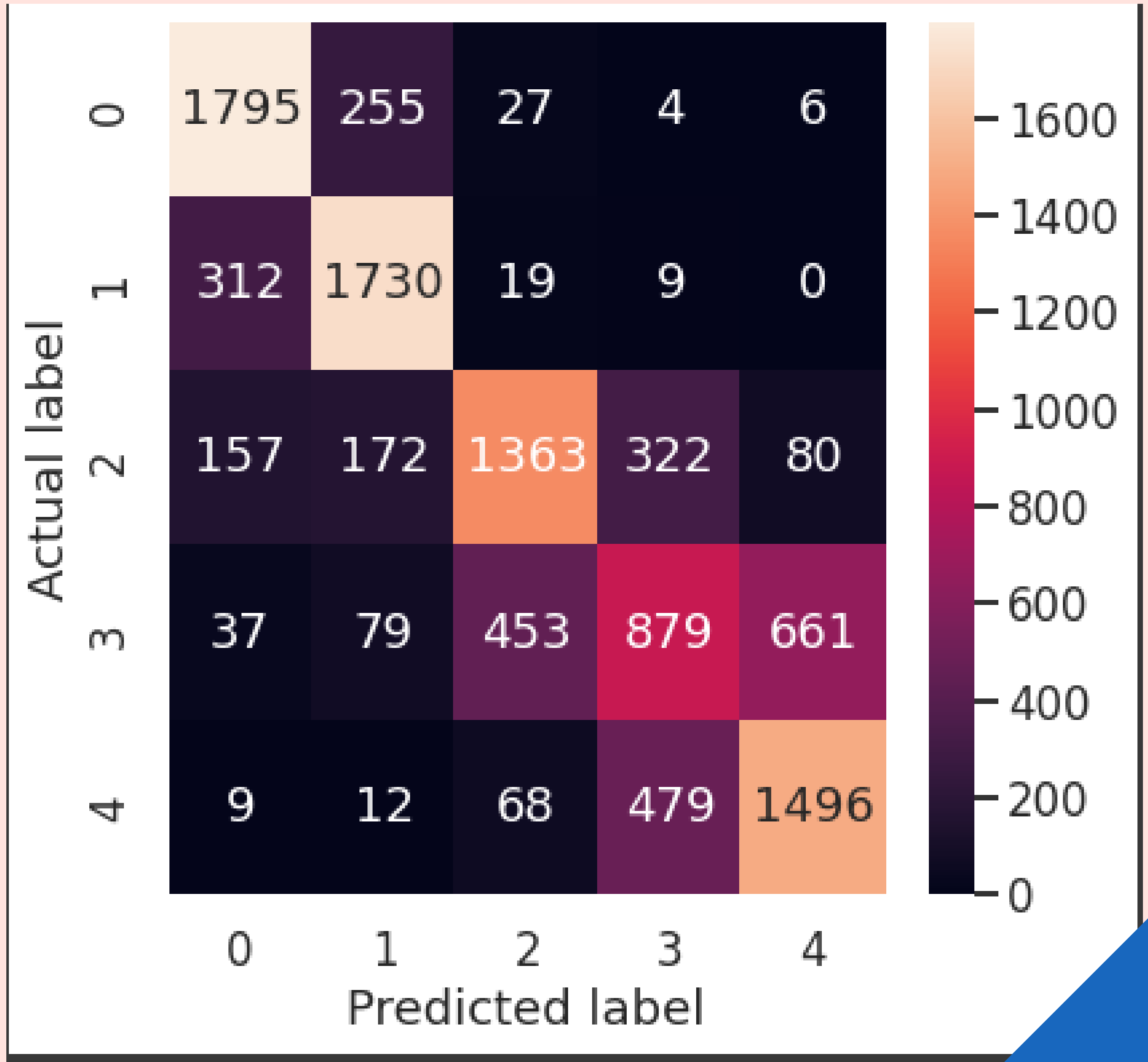
KNN

f1-score : 0.6904538694592525

accuracy : 0.6967574827321565

	precision	recall	f1-score	support
1	0.78	0.86	0.82	2087
2	0.77	0.84	0.80	2070
3	0.71	0.65	0.68	2094
4	0.52	0.42	0.46	2109
5	0.67	0.72	0.69	2064
accuracy			0.70	10424
macro avg	0.69	0.70	0.69	10424
weighted avg	0.69	0.70	0.69	10424

Submission ID: 67376491	Result	Score
19 seconds ago	Accepted	53.34862



Our Model

∂erivative team

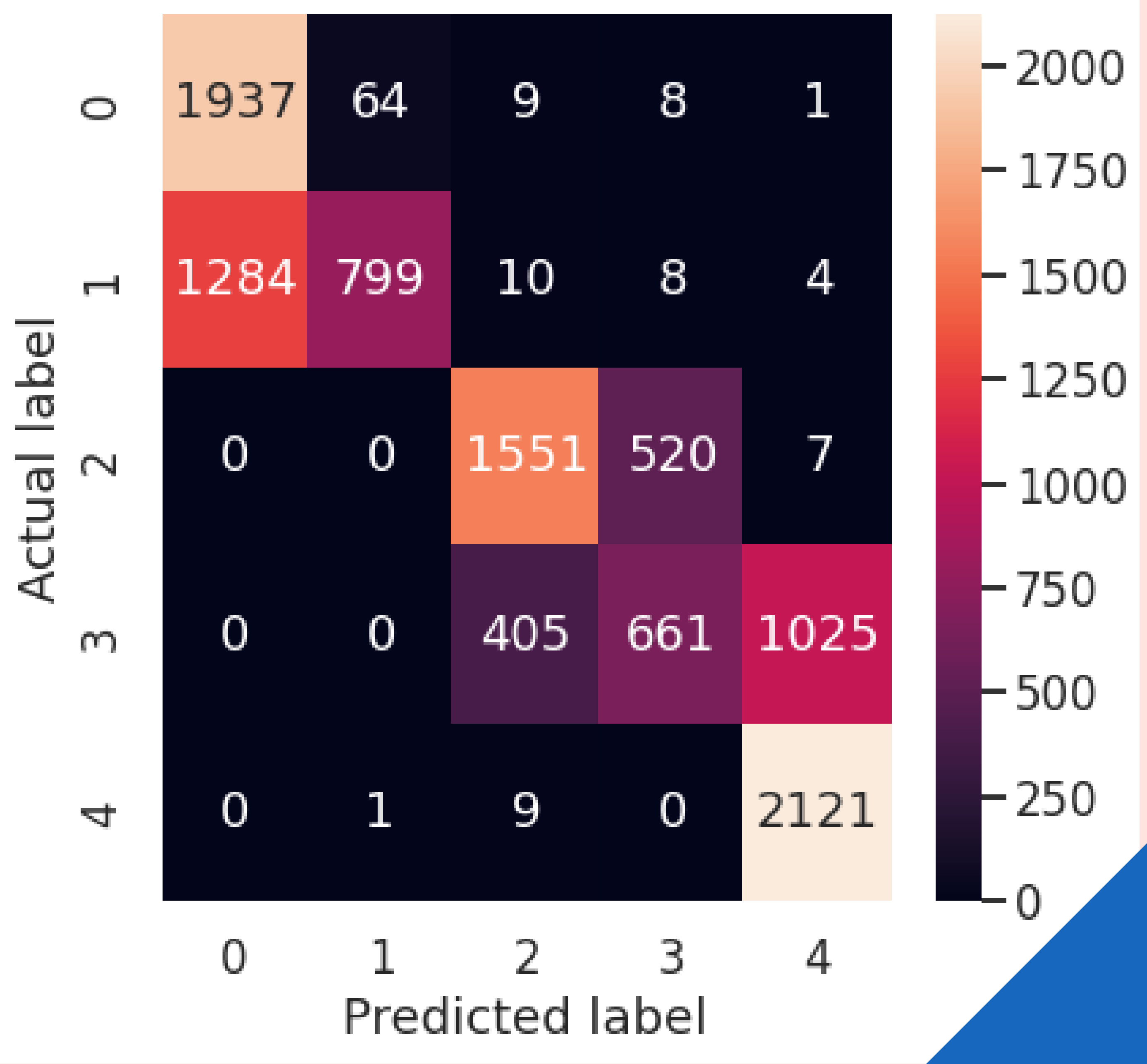
SVM

f1-score : 0.6490629314667464

accuracy : 0.6781465848042978

	precision	recall	f1-score	support
1	0.60	0.96	0.74	2019
2	0.92	0.38	0.54	2105
3	0.78	0.75	0.76	2078
4	0.55	0.32	0.40	2091
5	0.67	1.00	0.80	2131
accuracy			0.68	10424
macro avg	0.71	0.68	0.65	10424
weighted avg	0.71	0.68	0.65	10424

Submission ID: 67377054	Result	Score
6 seconds ago	Accepted	61.44761



Our Model

∂ erivative team

Summary

XGBoost and Random Forest have highest accuracy both on test and validation dataset (about 83 % on test data and 75% on validation data)

While KNN and SVM don't performed well on this dataset
(KNN has 53.35 % accuracy and
SVM has 61.45 % accuracy
on validation dataset)

DigitalSkola

∂ erivative team

∂erivative team

Temui Tim Kami

BENNITO
SIANTURI

YUDHI
NUGRAHA
RIYANSYAH

MUHAMMAD
GALANG
GARDAMUKTI

ALDRICHO A.
POLLARDO