Rakamin Academy

HOME CREDIT

DEFAULT RISK PREDICTION

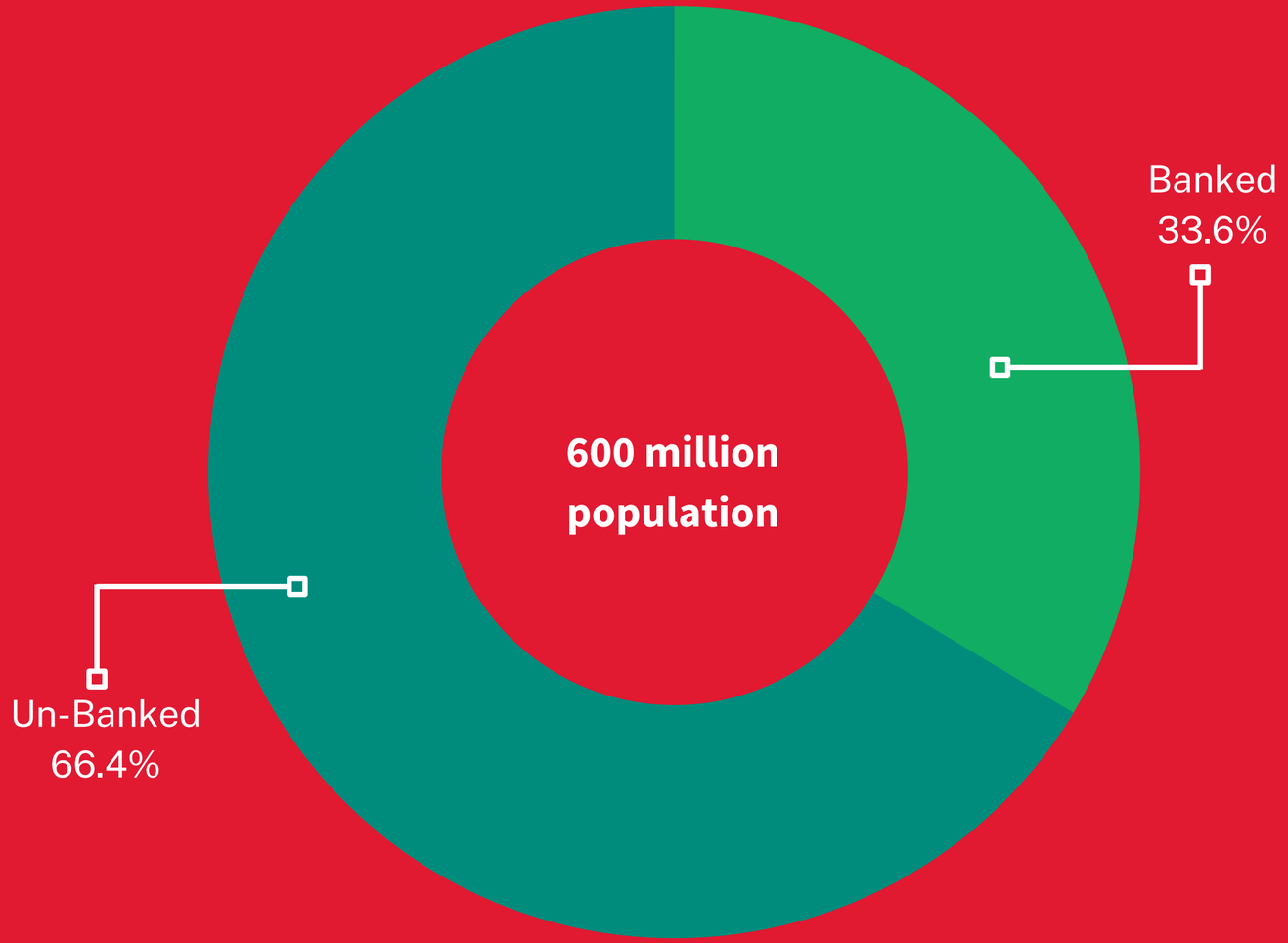
Home Credit Indonesia Data Scientist Virtual Internship Program

MANUEL SETYO SAPUTRO S.

GitHub Repository

https://github.com/msetyo15/Home-Credit-Scorecard-Model-Default-Risk

GO TO GITHUB



FINANCIAL INCLUSION FOR SOUTHEAST ASIA'S 'UNBANKED'

PROBLEM DESCRIPTION

PROBLEM ONE
Many people struggle to get loans due to insufficient credit histories

PROBLEM TWO
Ensure customer with capability to repay loans are NOT rejected

PROBLEM THREE
Aim to make use of alternative data (e.g. transactional info) to predict repayment ability

OBJECTIVE

Predict if each application is

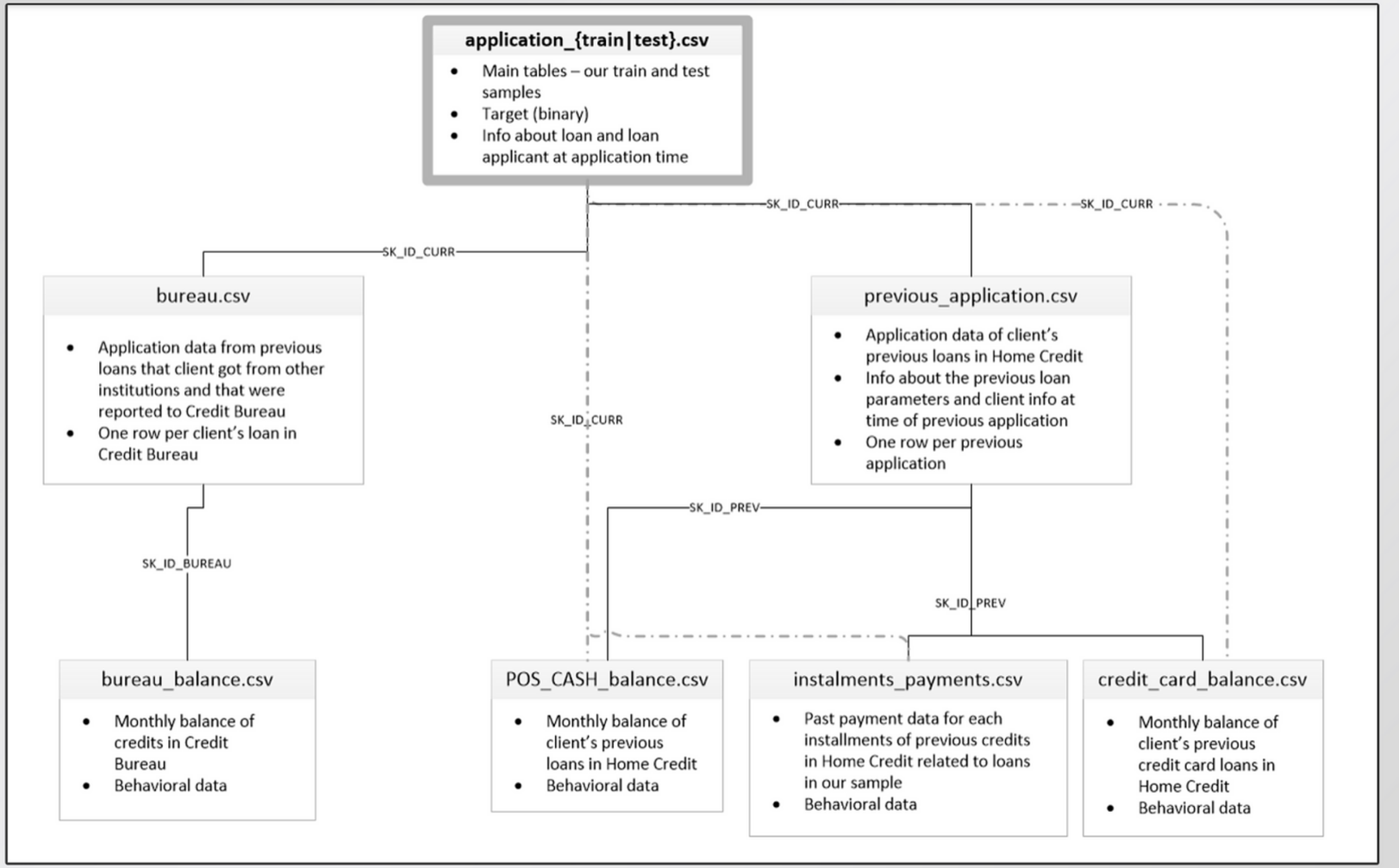
NOT capable of repaying a loan

CLASSIFY INTO
2 CLASSES :

Negative – 0 : Able to repay a loan

Positive – 1 : Unable to repay a loan

DATASET DESCRIPTION



MacBook Air

DATASET DESCRIPTION

TOTAL NO. OF FEATURES : 121

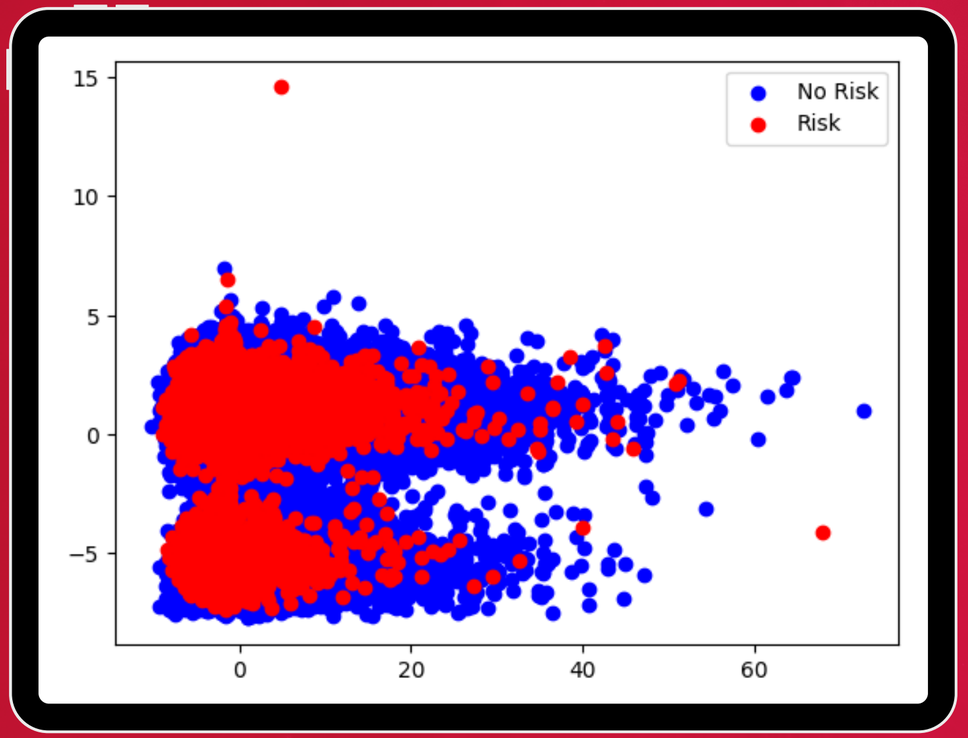
TOTAL NO. OF OBSERVATIONS : 304531

ACCURACY PARADOX IMBALANCED CLASS:

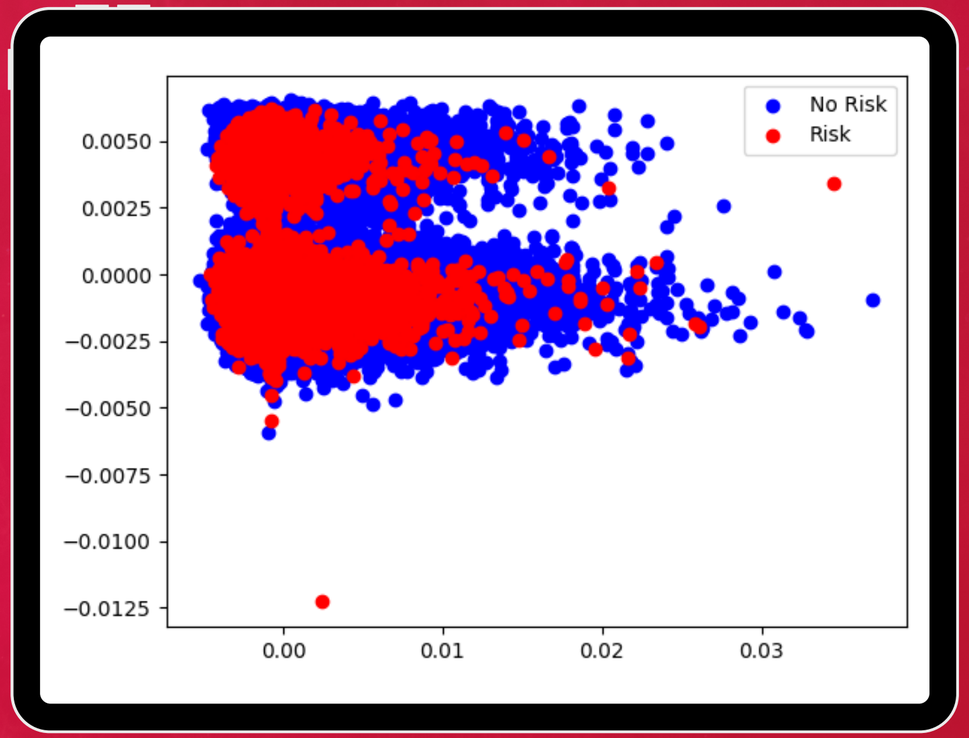
- More loans were repaid than loans that were not repaid

DATASET VISUALIZATION

NON TRIVIAL PROBLEM – OVERLAPPING CLASSES

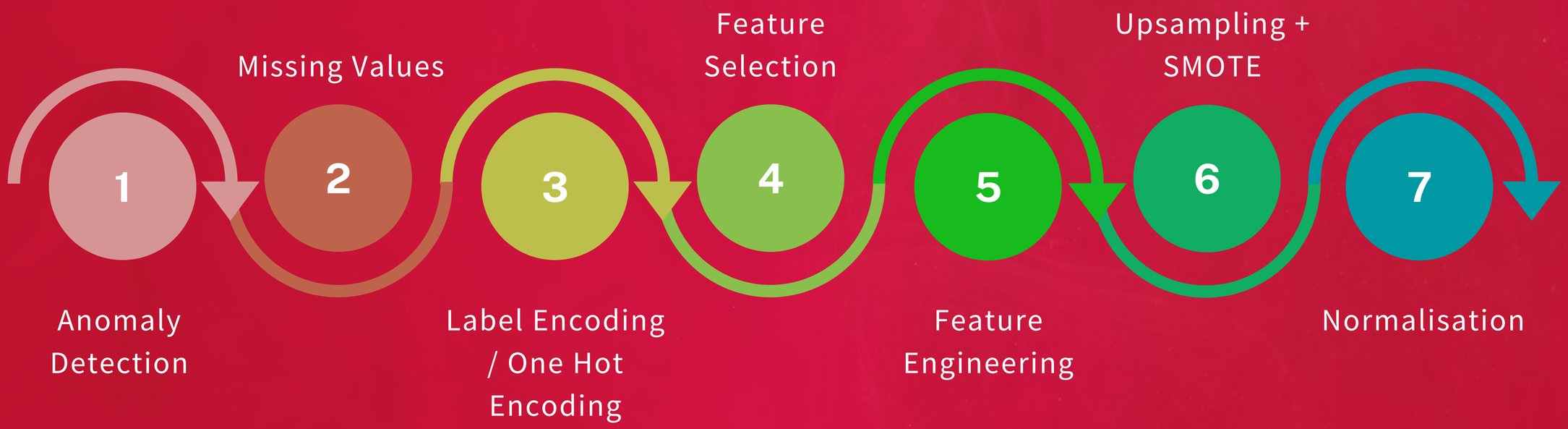


PCA



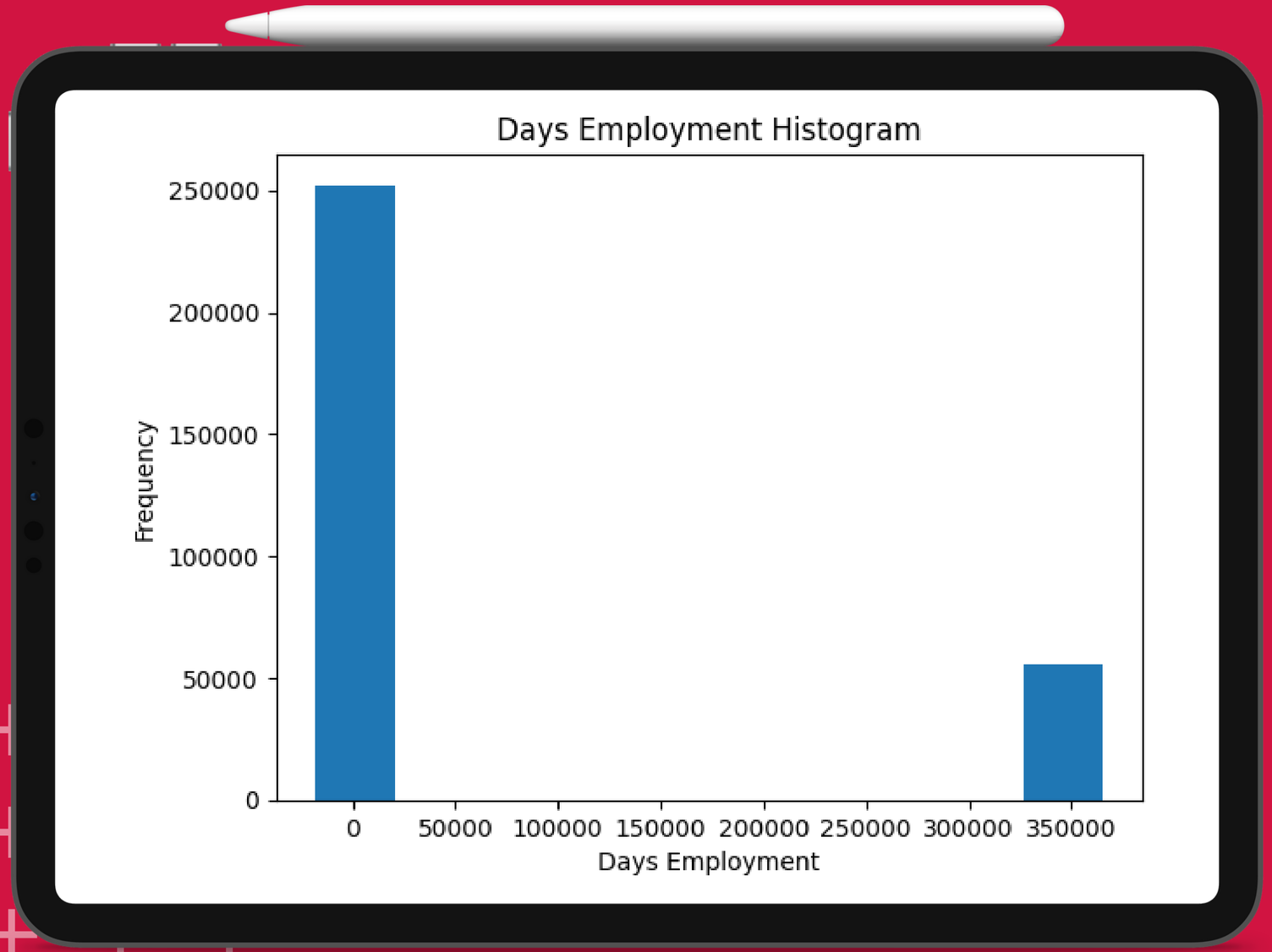
ICA

DATASET DESCRIPTION



ANOMALI DETECTION

MAXIMUM NUMBER OF DAYS EMPLOYED IS 1000 YEARS (ANOMALY)



ANOMALI DETECTION

OBSERVATION :

All the anomalies have the same value and have lower rate of default

SOLUTION :

Set the anomalies to median

IMPUTATE - MISSING VALUES

OBSERVATION :

Total of 67 columns with missing values

SOLUTION :

Median imputing

Your selected dataframe has 122 columns.
There are 67 columns that have missing values.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

LABEL ENCODING

- For columns with 2 or fewer categories (binary values)
- Not done for columns >=3 categories because it will imply that there is order
- A total of 4 columns were labelled encoded

ONE HOT ENCODING

- For columns with more than 3 categories
- Except the 4 columns, the rest of the columns were One Hot Encoded

Label Encoding

```
app_train.dtypes.value_counts()
```

```
float64    65
int64      41
object     16
dtype: int64
```

One Hot Encoding

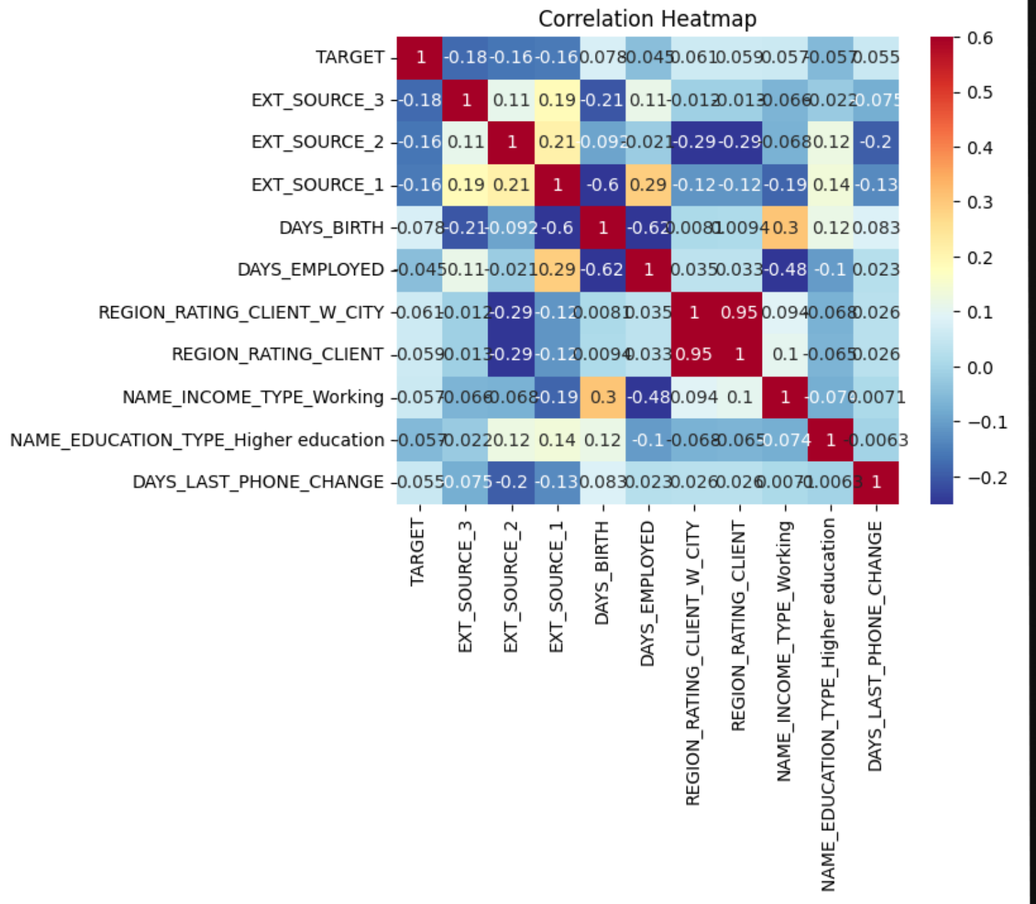
```
app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)
```

```
NAME_CONTRACT_TYPE              2
CODE_GENDER                     3
FLAG_OWN_CAR                    2
FLAG_OWN_REALTY                 2
NAME_TYPE_SUITE                 7
NAME_INCOME_TYPE                8
NAME_EDUCATION_TYPE             5
NAME_FAMILY_STATUS              6
NAME_HOUSING_TYPE               6
OCCUPATION_TYPE                18
WEEKDAY_APPR_PROCESS_START      7
ORGANIZATION_TYPE              58
FONDKAPREMONT_MODE              4
HOUSETYPE_MODE                  3
WALLSMATERIAL_MODE              7
EMERGENCYSTATE_MODE             2
dtype: int64
```

FEATURE SELECTION

```
Most Positive Correlations:
 NAME_EDUCATION_TYPE_Secondary / secondary special    0.049824
REG_CITY_NOT_WORK_CITY                                0.050994
DAYS_ID_PUBLISH                                       0.051457
CODE_GENDER_M                                         0.054713
DAYS_LAST_PHONE_CHANGE                                0.055218
NAME_INCOME_TYPE_Working                              0.057481
REGION_RATING_CLIENT                                  0.058899
REGION_RATING_CLIENT_W_CITY                           0.060893
DAYS_BIRTH                                            0.078239
TARGET                                                1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                          -0.178919
EXT_SOURCE_2                          -0.160472
EXT_SOURCE_1                          -0.155317
NAME_EDUCATION_TYPE_Higher education  -0.056593
CODE_GENDER_F                         -0.054704
NAME_INCOME_TYPE_Pensioner            -0.046209
ORGANIZATION_TYPE_XNA                 -0.045987
DAYS_EMPLOYED                         -0.044932
FLOORSMAX_AVG                         -0.044003
FLOORSMAX_MEDI                        -0.043768
Name: TARGET, dtype: float64
```



FEATURE ENGINEERING

POLYNOMIAL FEATURE ENGINEERING

Top 5 features with highest correlation selected

Do polynomial feature engineering with degree-3

- If an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2].
- Resultant 56 features

Goal :

To increase correlation to target

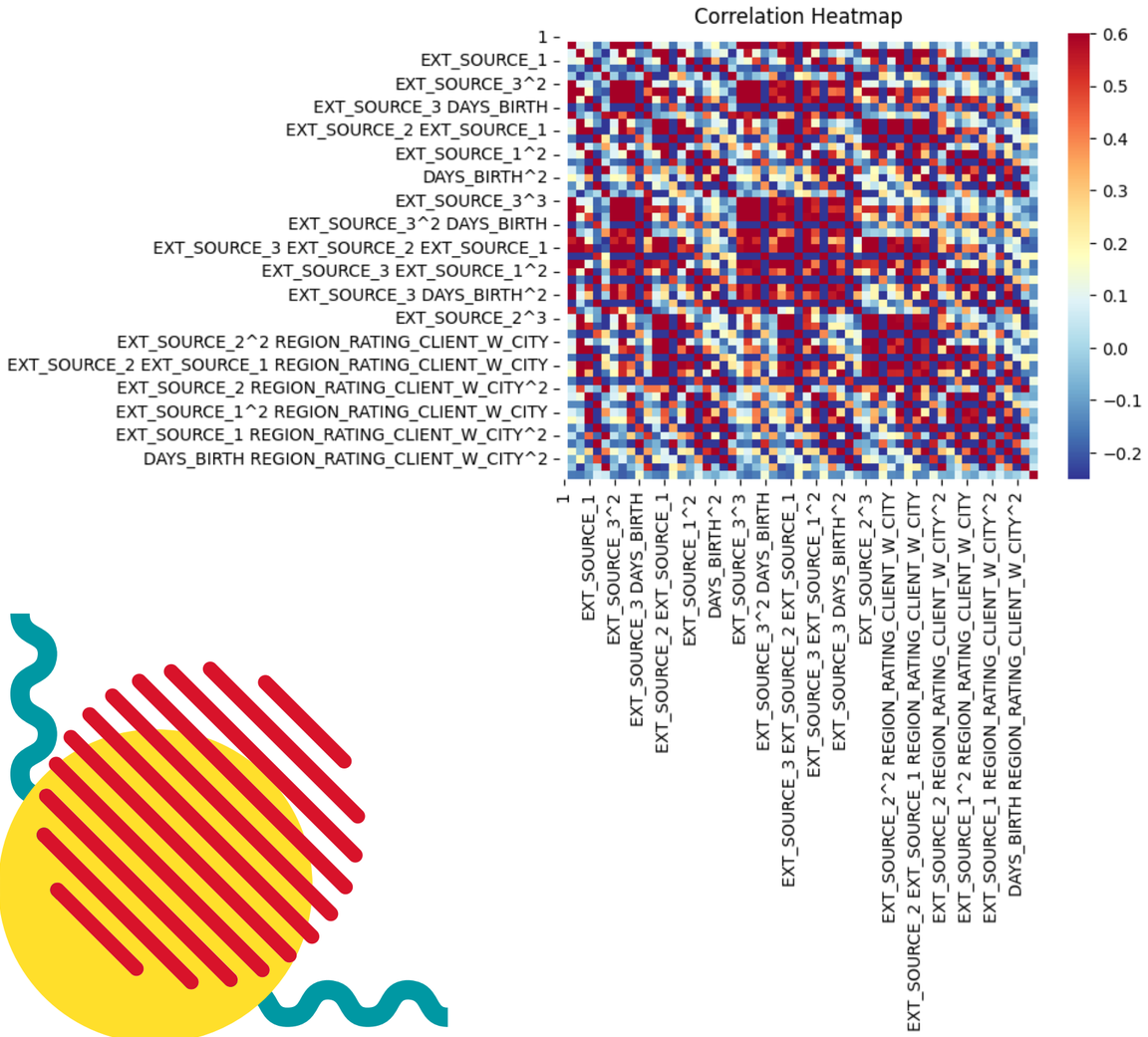
FEATURE ENGINEERING

Correlations after feature engineering

New features correlations to target increased

```
Most Negative Correlations:
EXT_SOURCE_3 EXT_SOURCE_2                                  -0.193939
EXT_SOURCE_3 EXT_SOURCE_2 EXT_SOURCE_1                     -0.189605
EXT_SOURCE_3 EXT_SOURCE_2^2                                -0.176428
EXT_SOURCE_3^2 EXT_SOURCE_2                                -0.172282
EXT_SOURCE_2 EXT_SOURCE_1                                  -0.166625
EXT_SOURCE_3 EXT_SOURCE_2 REGION_RATING_CLIENT_W_CITY      -0.165126
EXT_SOURCE_3 EXT_SOURCE_1                                  -0.164065
EXT_SOURCE_2                                               -0.160295
EXT_SOURCE_2^2 EXT_SOURCE_1                                -0.156867
EXT_SOURCE_3                                               -0.155892
Name: TARGET, dtype: float64

Most Positive Correlations:
EXT_SOURCE_2 DAYS_BIRTH REGION_RATING_CLIENT_W_CITY     0.126300
EXT_SOURCE_3^2 DAYS_BIRTH                               0.141777
EXT_SOURCE_2^2 DAYS_BIRTH                               0.149313
EXT_SOURCE_3 DAYS_BIRTH                                 0.150109
EXT_SOURCE_3 EXT_SOURCE_1 DAYS_BIRTH                    0.151816
EXT_SOURCE_2 EXT_SOURCE_1 DAYS_BIRTH                    0.155891
EXT_SOURCE_2 DAYS_BIRTH                                 0.156873
EXT_SOURCE_3 EXT_SOURCE_2 DAYS_BIRTH                    0.181283
TARGET                                                  1.000000
1                                                            NaN
Name: TARGET, dtype: float64
```



UPSAMPLING + SMOTE

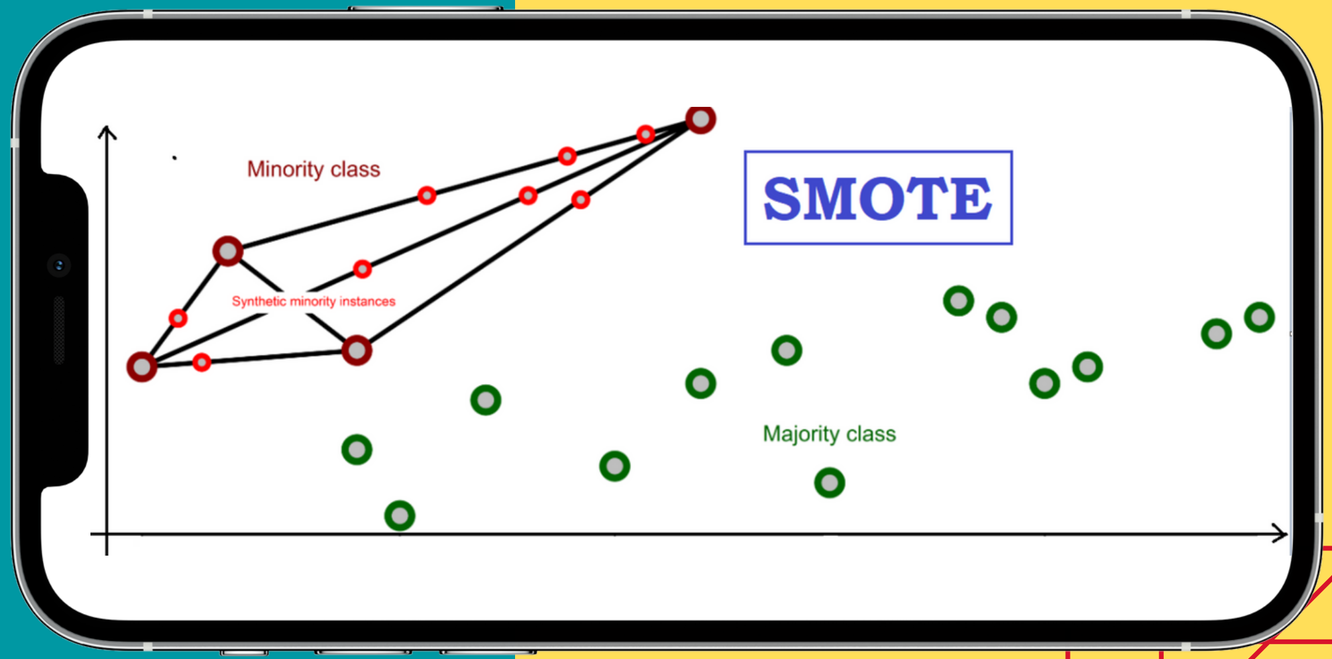
IMBALANCED DATASET :

10x number of negative class compared to positive

PROBLEMS :

Overall high accuracy but low recall for minority class (positive)

- High prediction for person who are able to repay
- Low prediction for person who are unable to repay



SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

- Uses KNN to generate new synthetic instances
- Instead of cloning the same data

UPSAMPLING RESOLVES THIS ISSUE BY ADDING MORE MINORITY CLASS

Increases sensitivity to the minority class

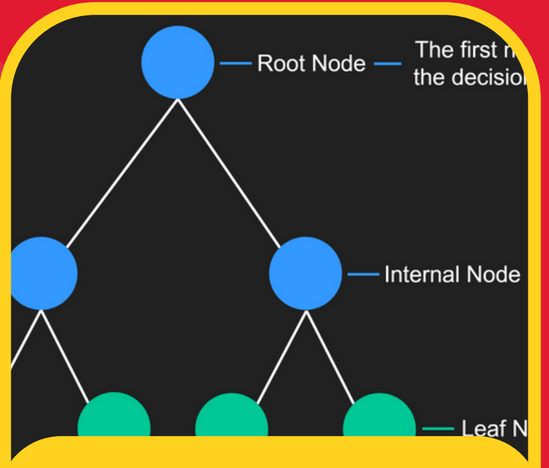
NORMALISATION

- Different features have non-standardised weightage
- Prevent dominance in dimensions, weighs all features equally in their representation
- Various scaling methods were used
- MinMax scaling works BEST due to difference in test and training dataset

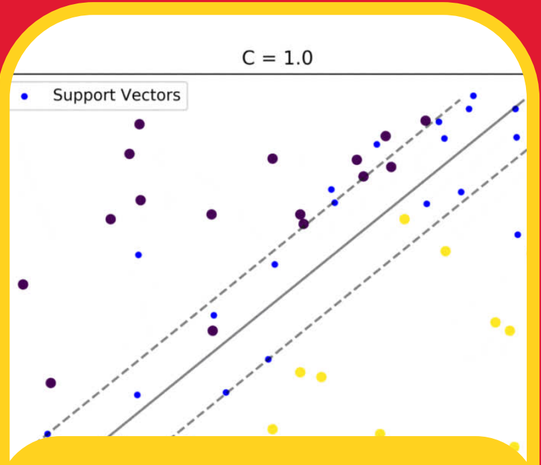
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	202500.0	406597.5	24700.5
0	270000.0	1293502.5	35698.5
0	67500.0	135000.0	6750.0
0	135000.0	312682.5	29686.5
0	121500.0	513000.0	21865.5

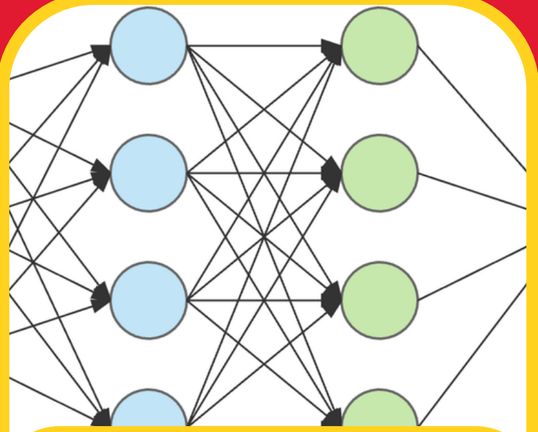
ALGORITHMS USED



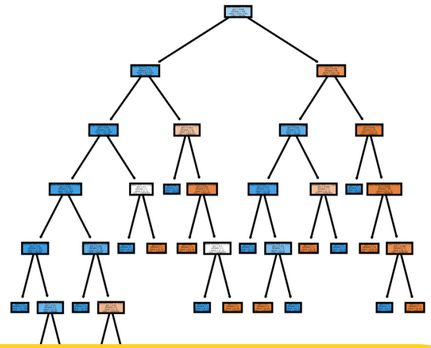
Decision Tree



Linear SVM



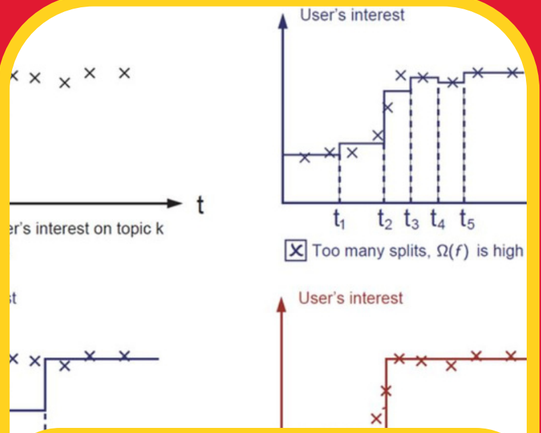
Neural Network (MLP)



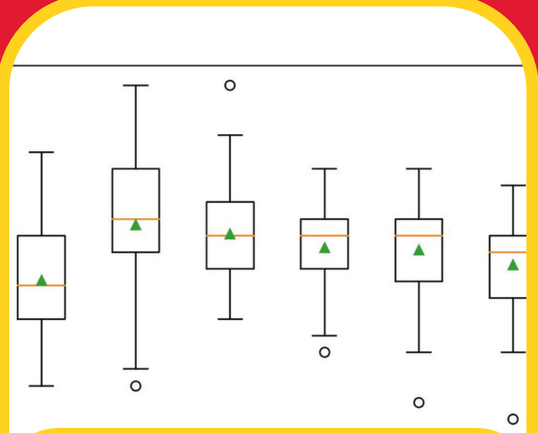
Random Forest

$$|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(}{P(B)}$$

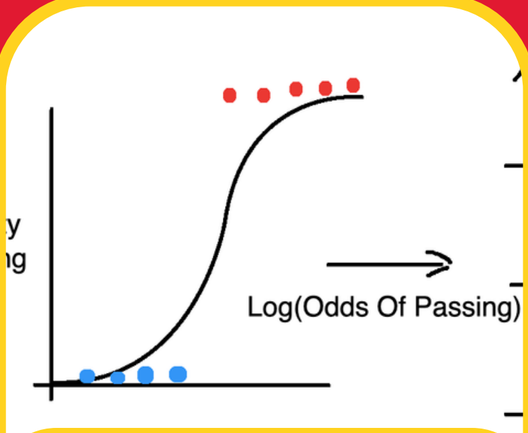
Naive Bayes



XGBoost



Adaboost



Logistic Regression

BASELINE MODELS PERFORMANCE COMPARASON

Models	Accuracy	F1-Score	ROC AUC
Adaboost	66.52 %	0.24	0.72
Random Forest	81.81 %	0.20	0.64
Neural Network (MLP)	67.98 %	0.25	0.72
Naive Bayes	59.11 %	0.22	0.71
Linear SVM	67.20 %	0.25	0.67
Decision Tree	76.79 %	0.18	0.56
Logistic Regression	64.40 %	0.24	0.72