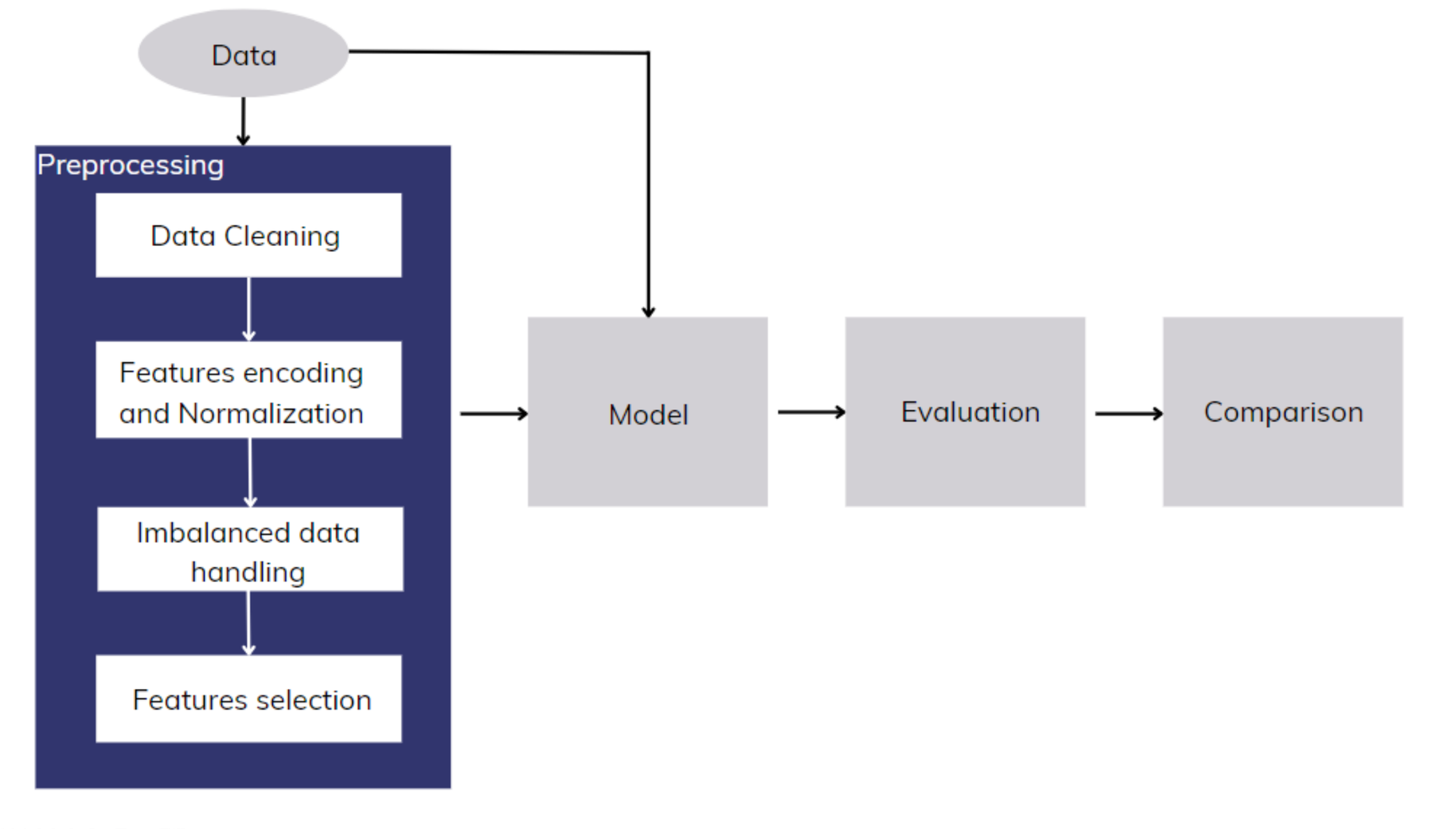
FEATURE SELECTION FOR LOAN REPAYMENT PREDICTION USING MACHINE LEARNING

System Degisn



DATA

- Home Credit Default Risk

- This dataset contains 308K anonymous clients’ with 122 unique features

DATA

SOME OF THE KEY FEATURES PRESENT IN THE DATASET:

- **SK_ID_CURR**: UNIQUE ID FOR EACH APPLICANT.
- **AMT_GOODS_PRICE**: PRICE OF GOODS FOR WHICH THE LOAN IS REQUESTED.
- **AMT_APPLICATION**: HOW MUCH CREDIT DID CLIENT ASK ON THE PREVIOUS APPLICATION
- **TARGET**: BINARY TARGET VARIABLE INDICATING IF THE APPLICANT EXPERIENCED DEFAULT (1) OR NOT (0).

DATA PREPROCESSING

DATA CLEANING

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

FEATURES ENCODING

1. COLUMN TYPES

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

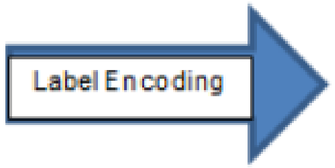
```
NAME_CONTRACT_TYPE              2
CODE_GENDER                     3
FLAG_OWN_CAR                    2
FLAG_OWN_REALTY                 2
NAME_TYPE_SUITE                 7
NAME_INCOME_TYPE                8
NAME_EDUCATION_TYPE             5
NAME_FAMILY_STATUS              6
NAME_HOUSING_TYPE               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
```

FEATURES ENCODING

2. ENCODING CATEGORICAL

- Label encoding: assign each unique category in a categorical variable with an integer. No new columns are created. An example is shown below

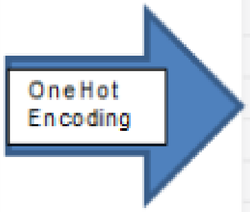
	occupation
0	programmer
1	data scientist
2	engineer
3	manager
4	ceo



	occupation
0	4
1	1
2	2
3	3
4	0

- One-hot encoding: create a new column for each unique category in a categorical variable. Each observation recieves a 1 in the column for its corresponding category and a 0 in all other new columns.

	occupation
0	programmer
1	data scientist
2	engineer
3	manager
4	ceo



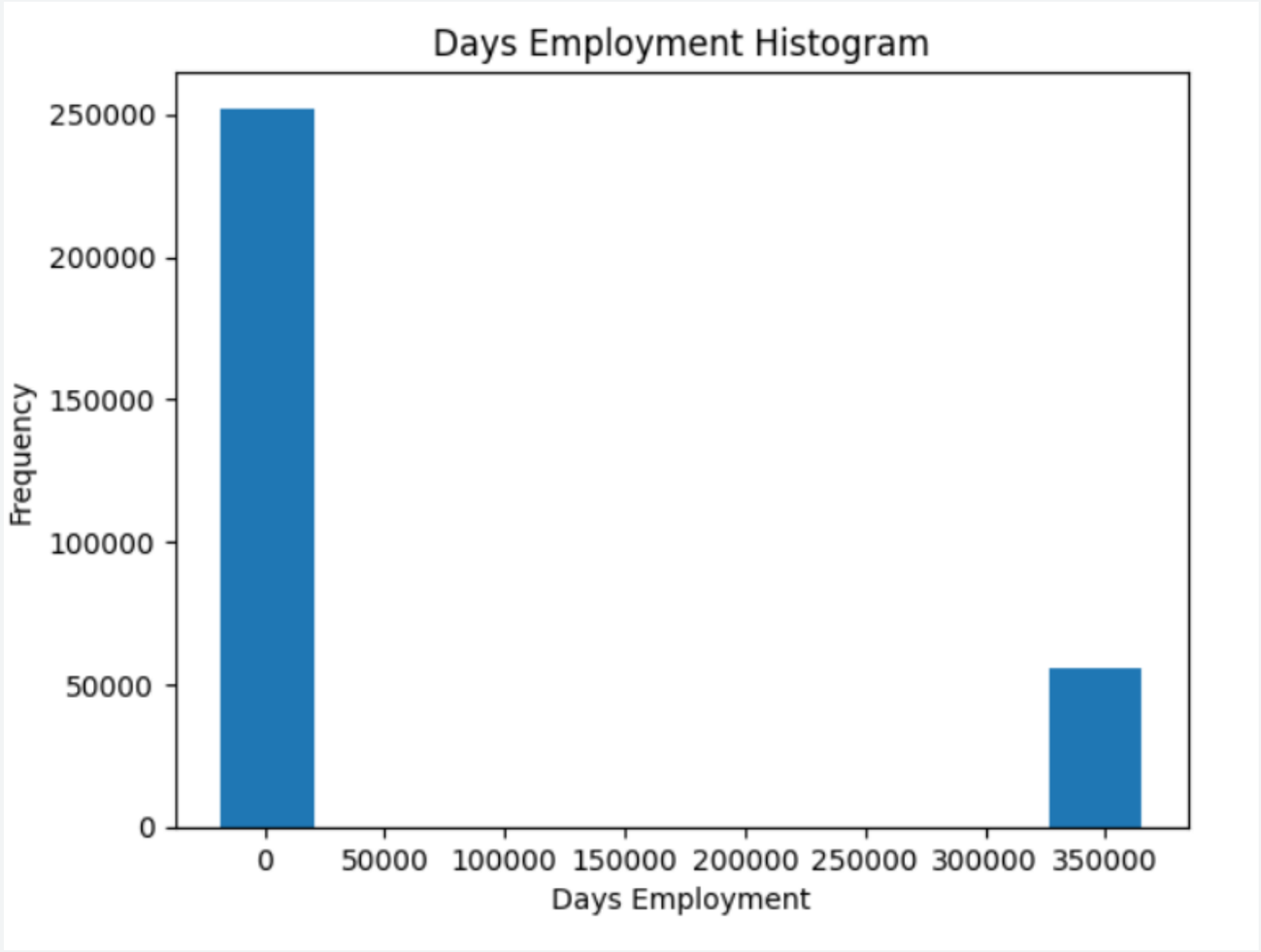
	occupation_ceo	occupation_data scientist	occupation_engineer	occupation_manager	occupation_programmer
0	0	0	0	0	1
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	1	0	0	0	0

ANOMALIES

One way to support anomalies quantitatively is by looking at the statistics of a column using the describe method

```
app_train['DAYS_EMPLOYED'].describe()
```

```
count    307511.000000
mean      63815.045904
std      141275.766519
min      -17912.000000
25%       -2760.000000
50%       -1213.000000
75%        -289.000000
max      365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```



Anomalous clients and see if they tend to have higher or low rates of default than the rest of the clients.

```
The non-anomalies default on 8.66% of loans
The anomalies default on 5.40% of loans
There are 55374 anomalous days of employment
```

WE WILL FILL IN THE ANOMALOUS VALUES WITH NOT A NUMBER (NAN) AND THEN CREATE A NEW BOOLEAN COLUMN INDICATING WHETHER OR NOT THE VALUE WAS ANOMALOUS.

TOTAL COLUMNS COME TO 241

STRATEGY N°3

EXPERT KNOWLEDGE

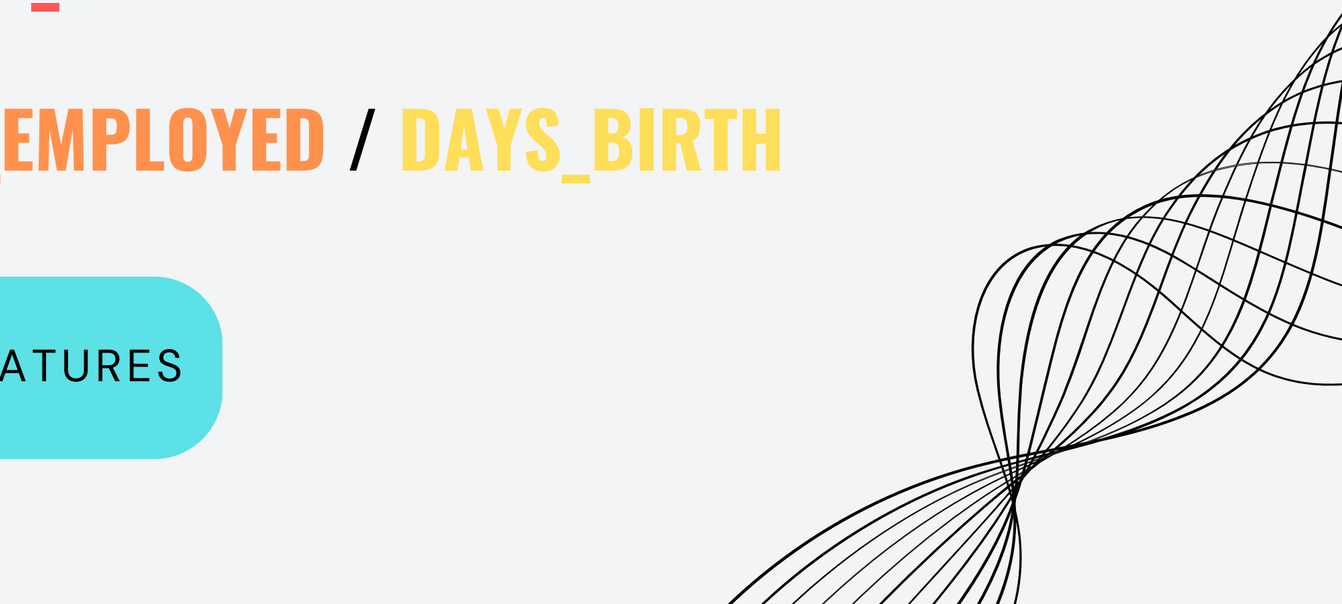
CREDIT_INCOME_PERCENT = AMT_CREDIT / AMT_INCOME_TOTAL

ANNUITY_INCOME_PERCENT = AMT_ANNUITY / AMT_INCOME_TOTAL

CREDIT_TERM = AMT_ANNUITY / AMT_CREDIT

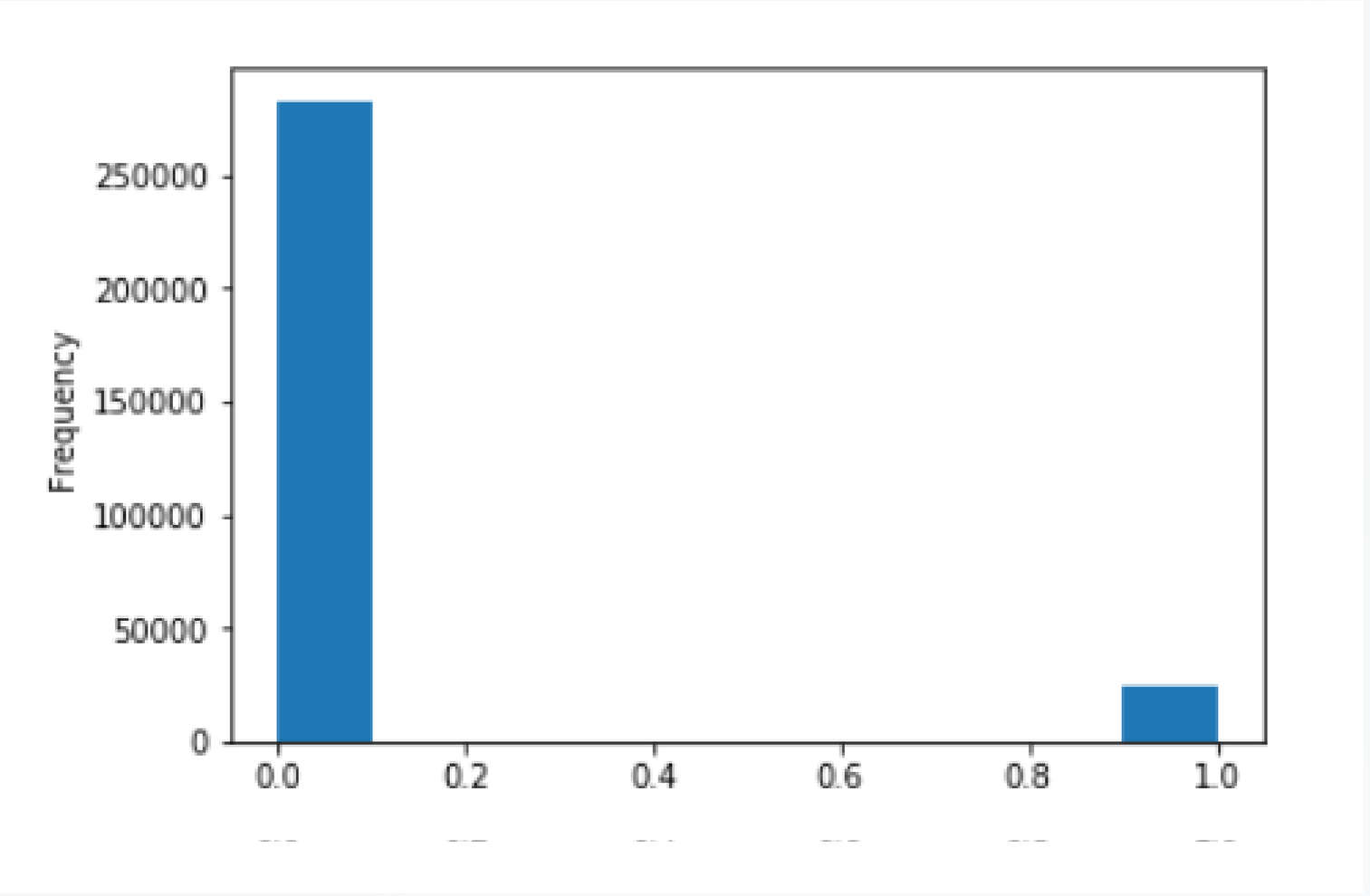
DAYS_EMPLOYED_PERCENT = DAYS_EMPLOYED / DAYS_BIRTH

→ 245 FEATURES

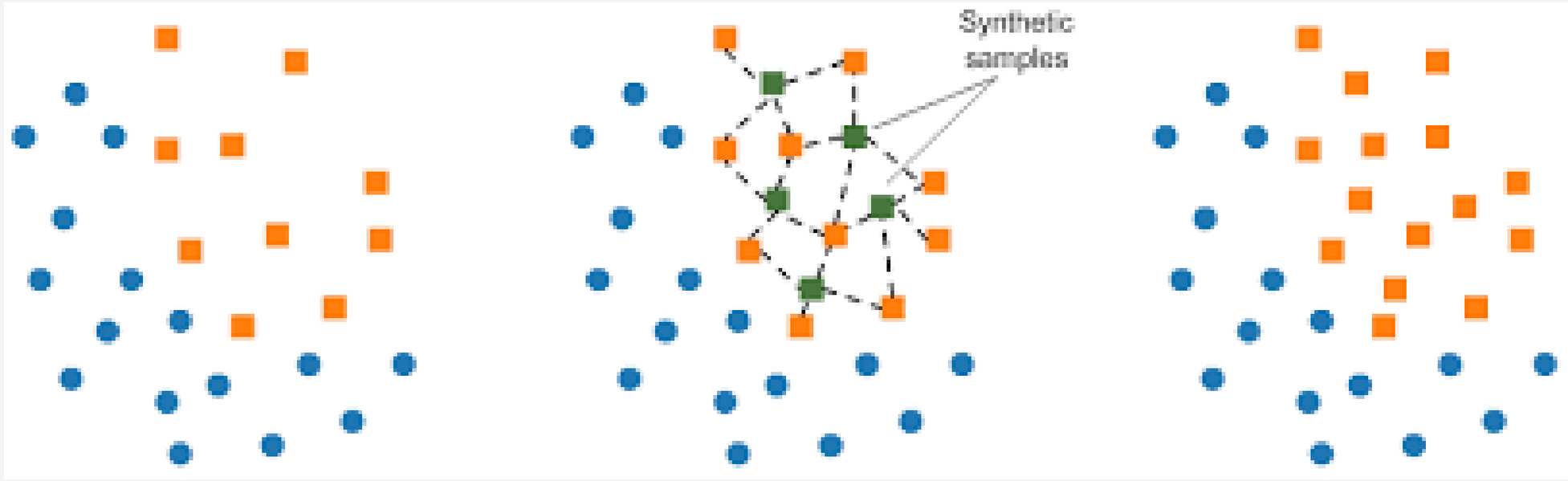


Imbalanced data handling

Imbalanced data



SMOTE (Synthetic Minority Over-sampling)



Feature Selection

Permuation feature importance

The permutation feature importance algorithm based on Fisher, Rudin, and Dominici (2018)

Input: Trained model **f**, feature matrix **X**, target vector **y**, error measure **L(y, f)**

1. **Estiamate** the original model error **e(orig)** = **L(y, f(X))**
2. For each feature **j (**column of **X):**
    - Randomly shuffle columns **j** generate feature matrix **X(perm)**
    - Estimate error **e(perm)** = **L(y, f(Xperm))**
    - Calculate permutation feature importance

        FI(j) = **e(perm)** / **e(orig)**

        FI(j) = **e(perm)** - **e(orig)**
3. Sort featurs by **descending** FI

Feature importance for LightGBM model

index	feature	weight
0	CODE_GENDER_M	0.0009473843952566118
1	CODE_GENDER_F	0.0008606672881393163
2	EXT_SOURCE_3	0.0007045764953281797
3	EXT_SOURCE_2	0.0003837231989941257
4	ELEVATORS_MODE	0.0003186853686561352 7
5	DAYS_BIRTH	0.0002319682615388174 4
6	CREDIT_TERM	0.0001669304312008623 7
7	FLAG_PHONE	0.0001539228651332713 6
8	AMT_REQ_CREDIT_BUREAU_QRT	0.0001344115160318848 4
9	EXT_SOURCE_1	0.0001235718776422256 6
10	AMT_GOODS_PRICE	0.0001105643115746124 4
11	AMT_REQ_CREDIT_BUREAU_YEAR	0.0001083963838966806 1
12	AMT_INCOME_TOTAL	9.322089015115776e-05
13	NAME_INCOME_TYPE_Working	9.105296247324812e-05
14	AMT_REQ_CREDIT_BUREAU_MON	8.888503479529409e-05
15	CNT_CHILDREN	8.238125176149858e-05
16	FLAG_OWN_CAR	6.286990266011206e-05
17	DAYS_EMPLOYED	6.286990266008985e-05
18	OBS_30_CNT_SOCIAL_CIRCLE	5.203026427043067e-05
19	FLOORSMAX_MODE	4.76944089145892e-05
20	REGION_RATING_CLIENT_W_CITY	4.552648123663516e-05
21	FLOORSMAX_MEDI	3.6854770524930024e-05

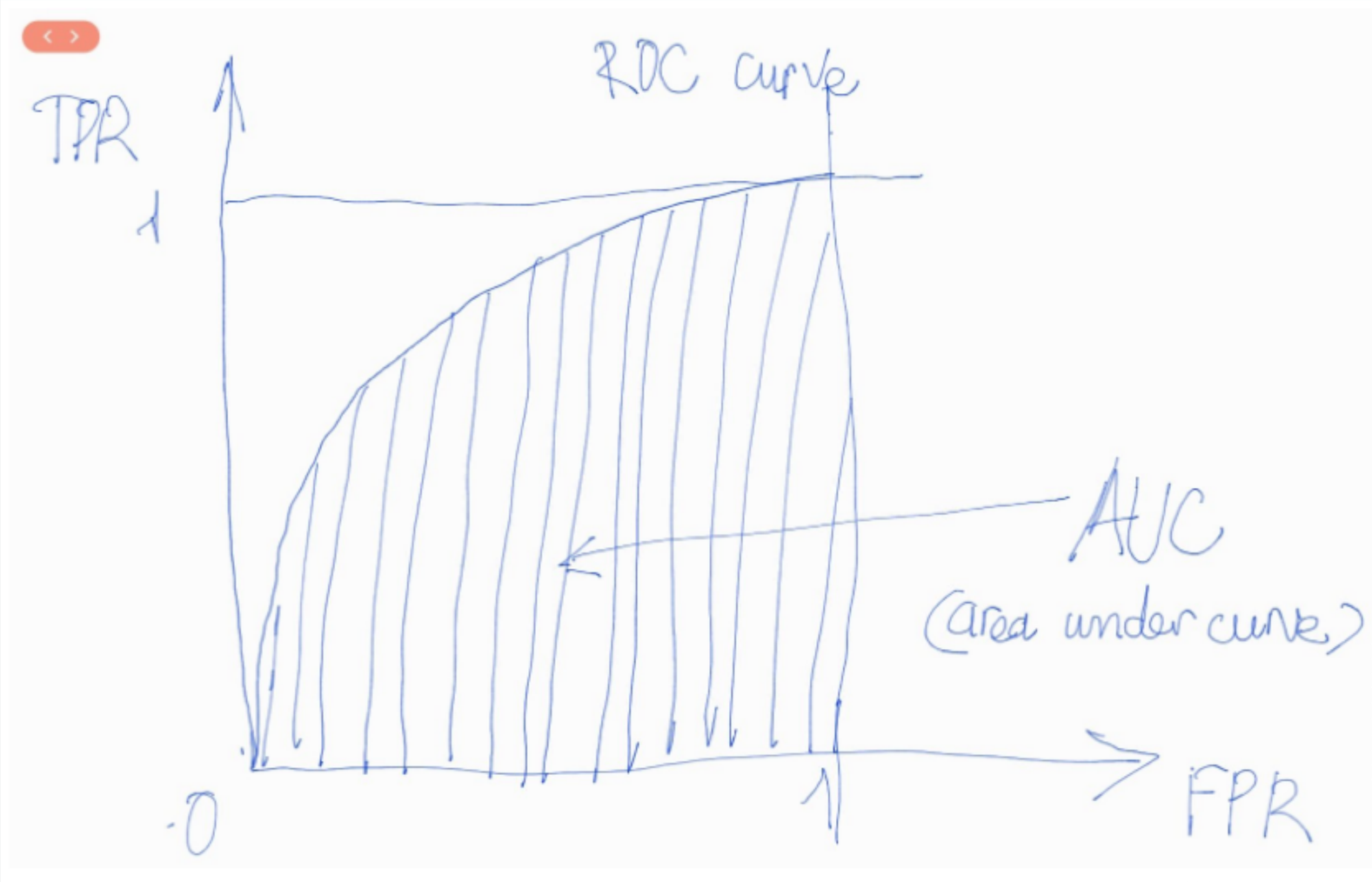
index	feature	weight
50	REGION_RATING_CLIENT	4.3358553558858756e-06
51	NAME_EDUCATION_TYPE_Lower secondary	4.3358553558636714e-06
52	NAME_HOUSING_TYPE_Municipal apartment	4.3358553558636714e-06
53	YEARS_BEGINEXPLUATATION_AVG	2.1679276779540403e-06
54	DAYS_LAST_PHONE_CHANGE	2.1679276779540403e-06
55	FLAG_DOCUMENT_16	2.1679276779318357e-06
56	NAME_HOUSING_TYPE_House / apartment	2.1679276779318357e-06
57	ORGANIZATION_TYPE_Industry: type 9	2.1679276779318357e-06
58	LIVINGAREA_AVG	2.1679276779318357e-06
59	NAME_HOUSING_TYPE_Rented apartment	2.1679276779318357e-06
60	FLAG_DOCUMENT_3	2.2204460492503132e-17
61	FLAG_DOCUMENT_15	0.0
62	AMT_REQ_CREDIT_BUREAU_HOUR	0.0
63	FLAG_DOCUMENT_14	0.0
64	FLAG_DOCUMENT_17	0.0
65	FLAG_DOCUMENT_19	0.0
66	FLAG_DOCUMENT_13	0.0
67	FLAG_DOCUMENT_20	0.0
68	FLAG_DOCUMENT_21	0.0
69	FLAG_DOCUMENT_12	0.0
70	FLAG_DOCUMENT_11	0.0
71	LIVINGAPARTMENTS_AVG	0.0

Result and Comparation

Area under curve (AUC)

$$\text{TPR} = \frac{TP}{\text{total positive}}$$

$$\text{FPR} = \frac{FP}{\text{total negative}}$$



Paper to compare

"Credit Risk Scoring Analysis Based on Machine Learning Models"

from Research Lab for Knowledge and Wisdom at Xi'an Jiaotong-Liverpool University in China and the Department of Computer Science at the University of Liverpool in the UK.

Published in: 2019 6th International Conference on Information Science and Control Engineering (ICISCE)

Date Added to IEEE Xplore: 04 June 2020

Comparation

FEATURE NUMBER AFTER FEATURE GENERATION

Original dataset	Polynomial generated dataset	Expert knowledge generated dataset	FeatureTools Toolkit generated dataset
240	274	249	268

Feature Number After Feature Selection

Original	Permitation for Random forest model	Permitation for Logictis Regression model	Permitation for LightGBM model
245	90	112	62

Comparation

TABLE II
EXPERIMENT RESULT

Feature Method	Random Forest	Logit Regression	LightGBM
Original	0.684	0.706	0.721
Polynomial	0.601	0.720	0.733/0.738*
Expert Knowledge	0.677	0.703	0.755/0.778*
FeatureTools	0.681	0.711	0.724
Result with * is recorded after dropping 20 least correlated features.			

Feature Method	Random Forest	Logictis Regression	LightGBM
Original	0.700	0.736	0.736
Permitation Importance Feature	0.668	0.732	0.740

Future works

Concatenate all features from multiple CSV files into a single CSV file with a larger number of features.

Finding another method about features selection to apply this model.

THANK'S FOR WATCHING