CREDIT RISK PREDICTION

CLASSIFYING CUSTOMERS TO ASSESS THEIR PROBABILITY TO DEFAULT

DAVID | SERGI | TOMIL

IRON HACK Ironhack Private Bank Services

WHAT IS CREDIT RISK?

Credit risk is the possibility of
a loss resulting from a borrower's
failure to repay a loan.

In a global perspective, bad credit
risk assessments can lead
to a multilevel failure.

WHY IT MATTERS?

BANK LENDING	BUSINESS LENDING	MORTGAGE LENDING	CONSUMER CREDIT	CREDIT LOSSES
▲ 3.7%	▲ 2.3%	▲ 4.0%	▲ 2.6%	▲ 3.9%

overall, credit demand is expected to grow in the following years

AVERAGE GROWTH FORECAST
INDICATORS FOR 2023 TOTAL EUROZONE

TYPICAL SCORING TOOLS

FICO® Score Factors:

Payment history: 35%

Current amount of indebtedness: 30%

Length of credit history: 15%

Types of credit used: 10%

New credit accounts: 10%

VantageScore Factors:

Total credit usage: Extremely

Credit mix and experience: Highly

Payment history: Moderate

Age of credit history: Less

New accounts opened: Less

FACTORS IN CONSIDERATION…

factors USUALLY taken into consideration when calculating credit scores

CREDIT HISTORY	PAYMENT HISTORY	CAPACITY TO REPAY	AMOUNT OF INDEBTEDNESS	ASSOCIATED COLLATERAL

factors NOT taken into consideration when calculating credit scores

LIVING PLACE	EMPLOYMENT HISTORY	FAMILY BACKGROUND	EDUCATION HISTORY	LIVING FLAGS

GOALS

1. Agree on a definition of GOOD / BAD client based on credit repay record → **DEFINE TARGET VARIABLE**

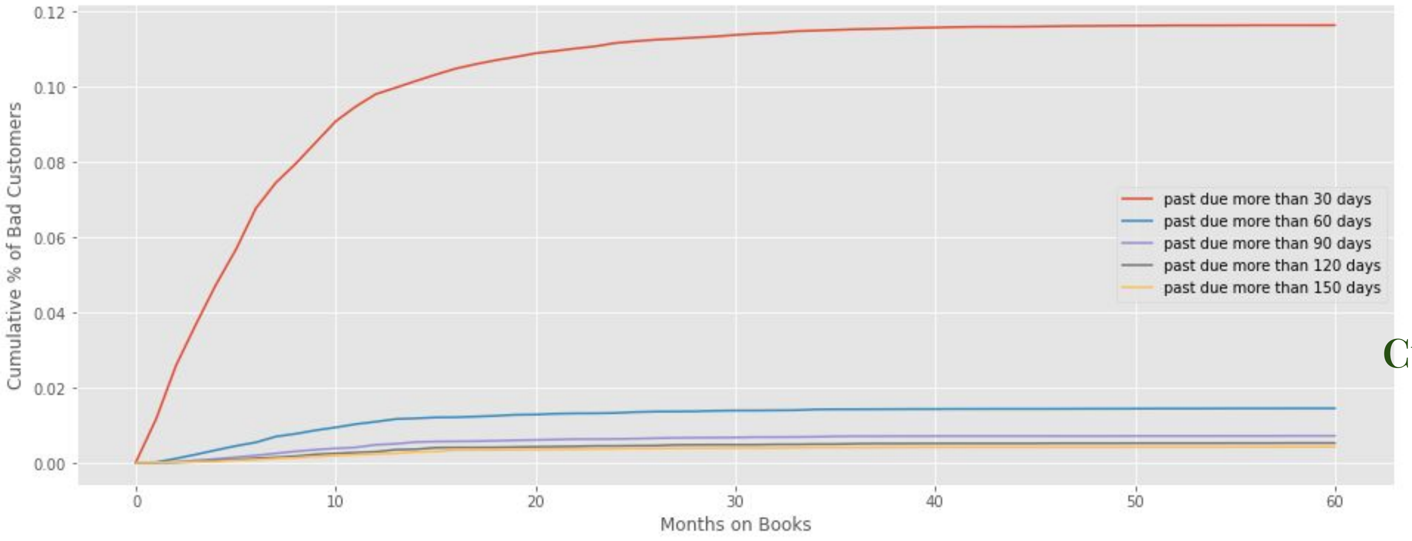
2. Analyze the factors that affect the determination of GOOD / BAD clients → **PROFILING AND FEATURE IMPORTANCE**

3. Build a model that classifies/predicts the customers as GOOD / BAD for a credit approval. → **MACHINE LEARNING PIPELINE**

DEFINE THE TARGET VARIABLE

WINDOW

60 MONTHS ON RECORD



Cumulative % of Bad Customers

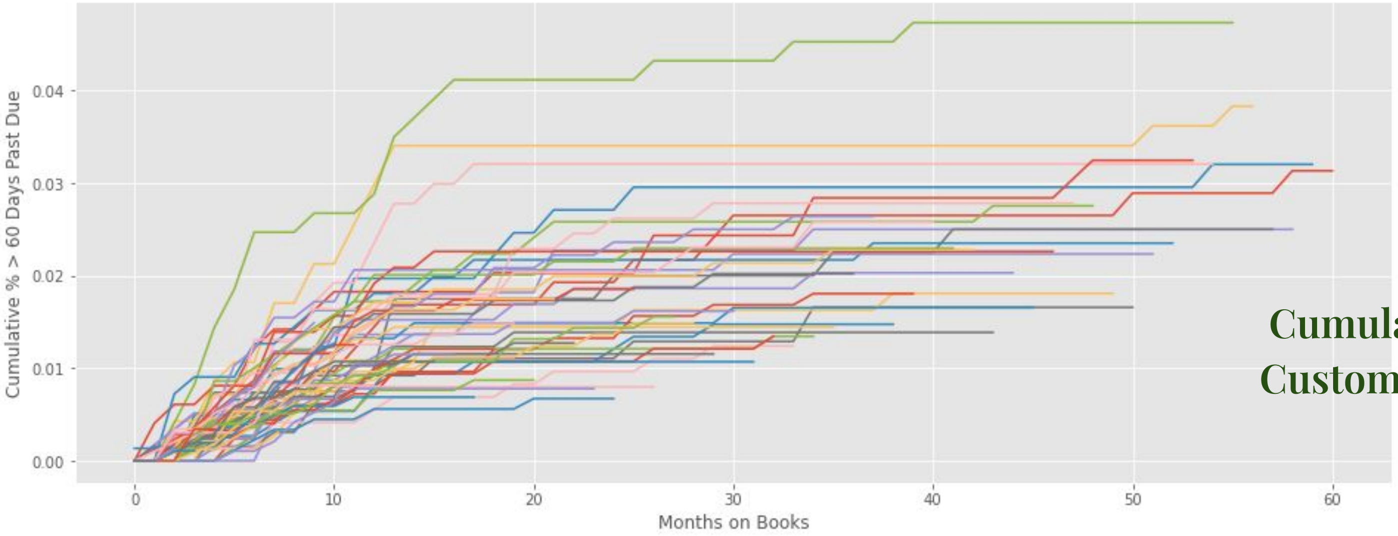
DEFINE THE TARGET VARIABLE

WINDOW

60 MONTHS ON RECORD

GOOD/BAD

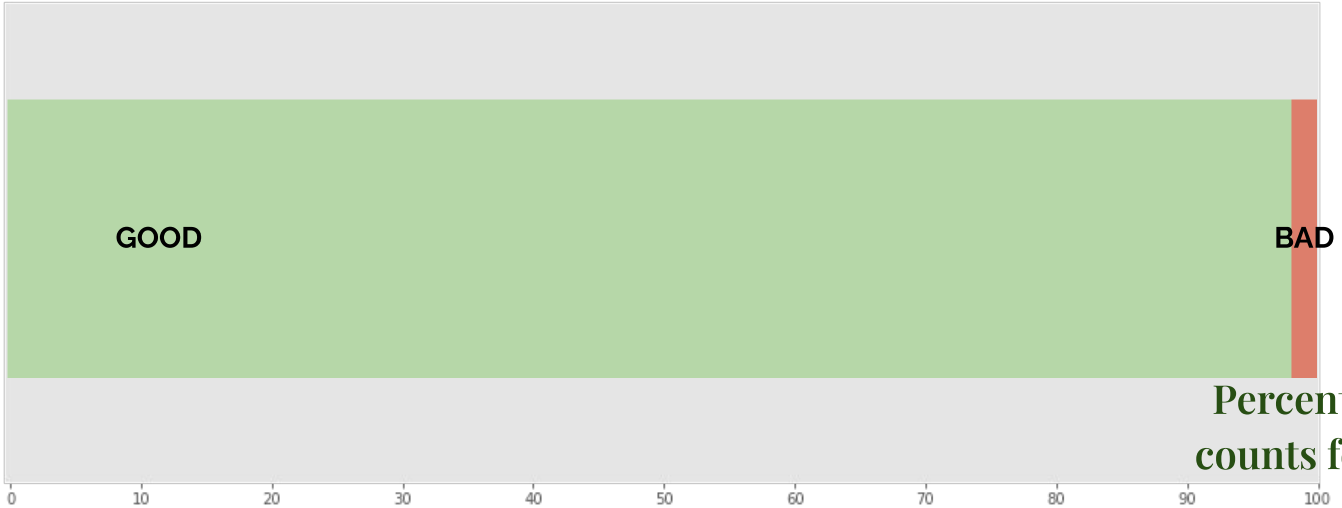
+60 DAYS PAST DUE



Cumulative % of Bad Customers (>60 days past due)

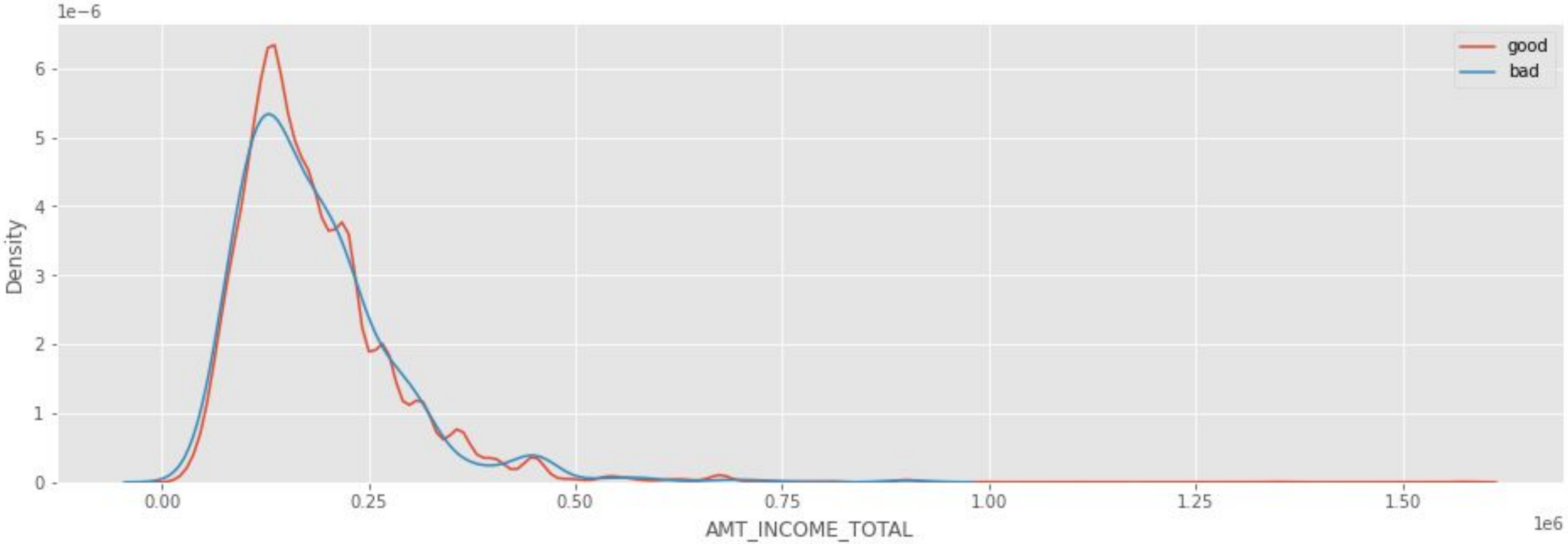
DEFINE THE TARGET VARIABLE

WINDOW	GOOD/BAD	IMBALANCE
60 MONTHS ON RECORD	+60 DAYS PAST DUE	98% vs. 2%

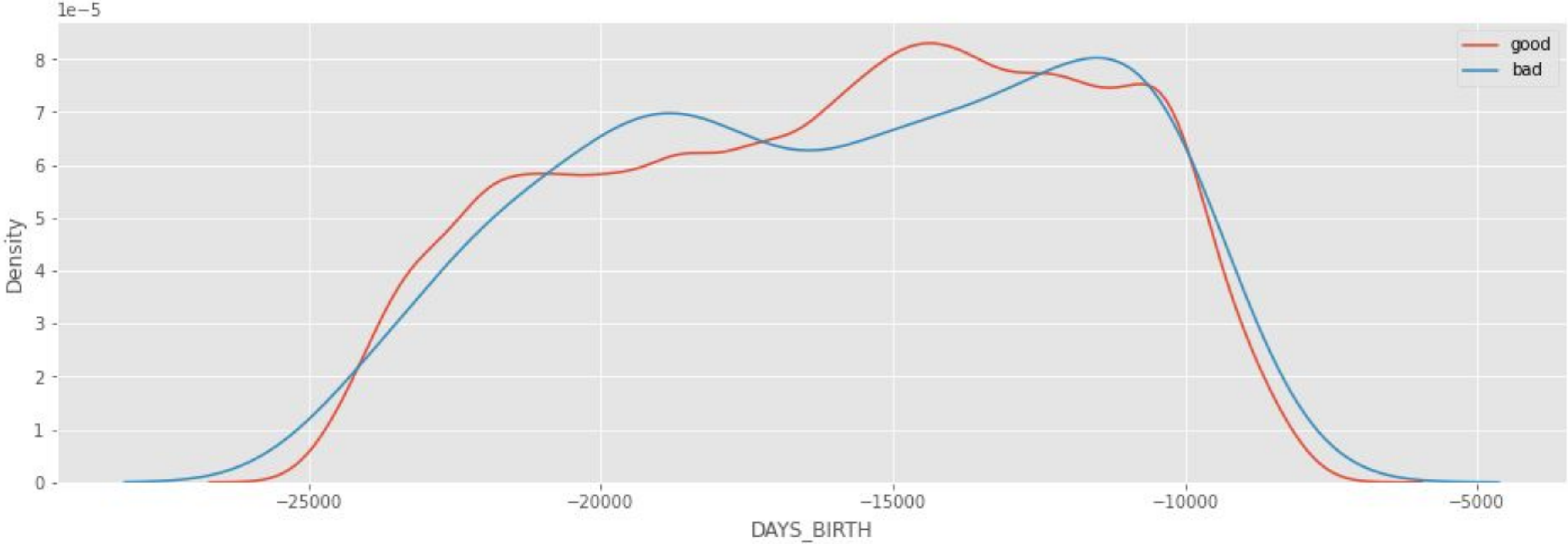


Percentage of value counts for the target variable.

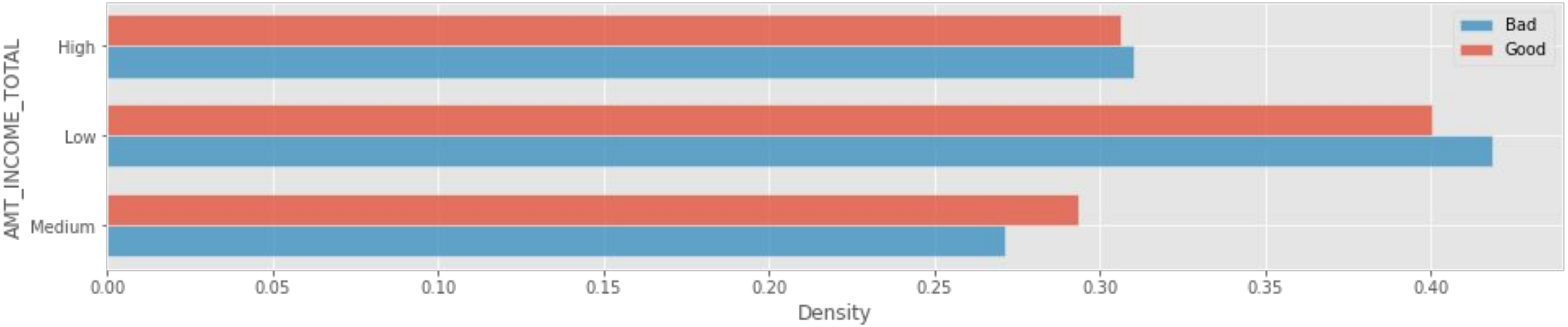
PROFILING AND FEATURE IMPORTANCE



PROFILING AND FEATURE IMPORTANCE

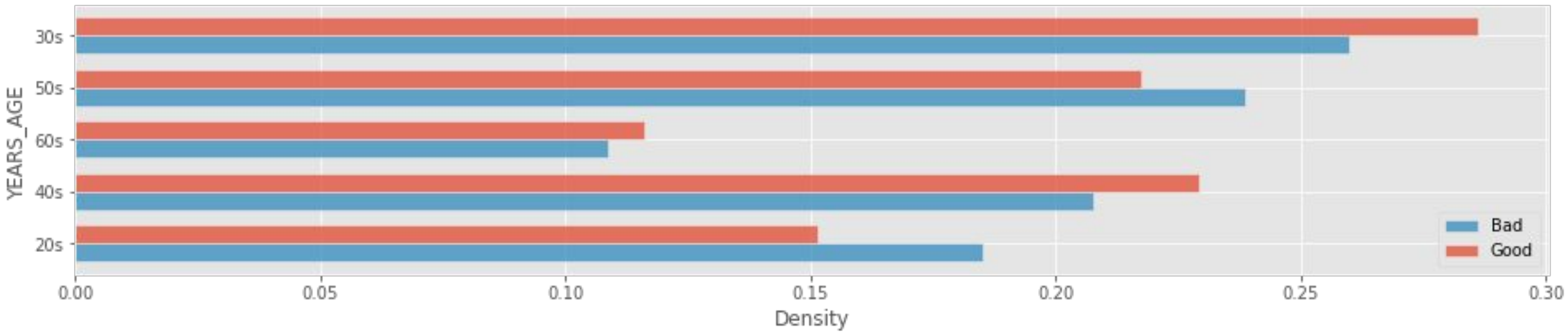


PROFILING AND FEATURE IMPORTANCE



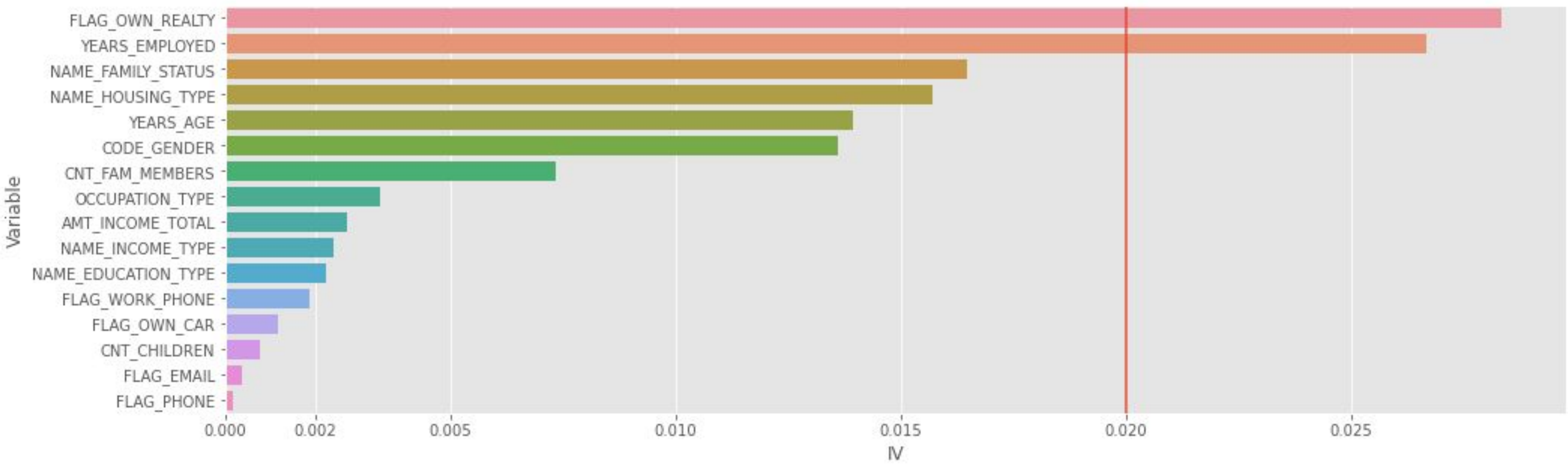
	Variable	Value	All	Good	Bad	Share	Bad Rate	Distribution Good	Distribution Bad	WoE	IV
0	AMT_INCOME_TOTAL	High	11161	10970	191	0.306141	0.017113	0.306074	0.310065	-0.012955	0.000052
1	AMT_INCOME_TOTAL	Low	14606	14348	258	0.400636	0.017664	0.400324	0.418831	-0.045195	0.000836
2	AMT_INCOME_TOTAL	Medium	10690	10523	167	0.293222	0.015622	0.293602	0.271104	0.079724	0.001794

PROFILING AND FEATURE IMPORTANCE

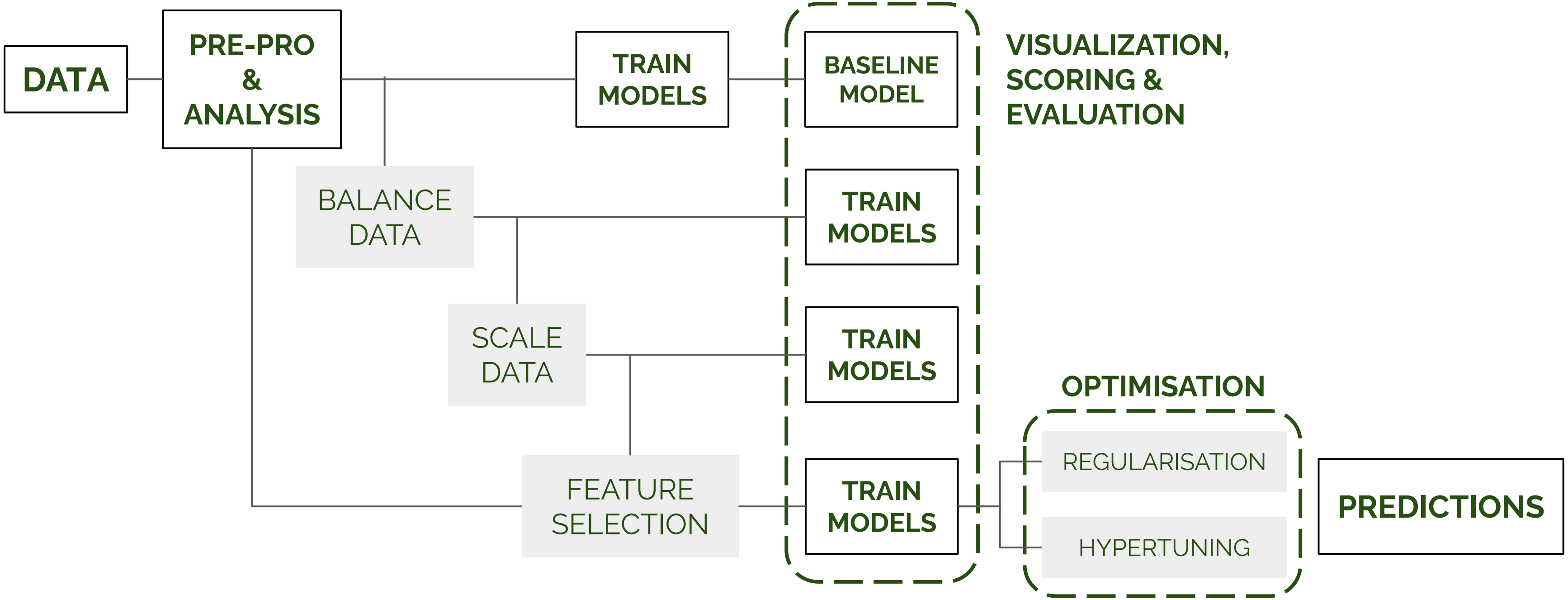


	Variable	Value	All	Good	Bad	Share	Bad Rate	Distribution Good	Distribution Bad	WoE	IV
0	YEARS_AGE	20s	5539	5425	114	0.151932	0.020581	0.151363	0.185065	-0.201026	0.006775
1	YEARS_AGE	30s	10419	10259	160	0.285789	0.015357	0.286236	0.259740	0.097136	0.002574
2	YEARS_AGE	40s	8340	8212	128	0.228763	0.015348	0.229123	0.207792	0.097721	0.002084
3	YEARS_AGE	50s	7933	7786	147	0.217599	0.018530	0.217237	0.238636	-0.093951	0.002010
4	YEARS_AGE	60s	4226	4159	67	0.115917	0.015854	0.116040	0.108766	0.064737	0.000471

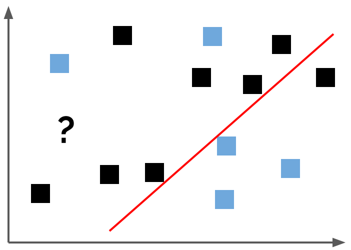
PROFILING AND FEATURE IMPORTANCE



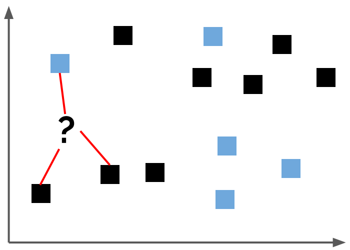
MACHINE LEARNING WORKFLOW (pipeline)



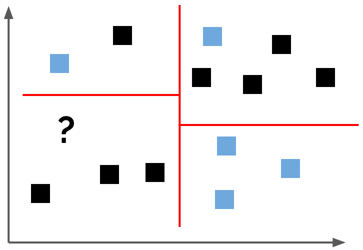
THE MODELS TESTED



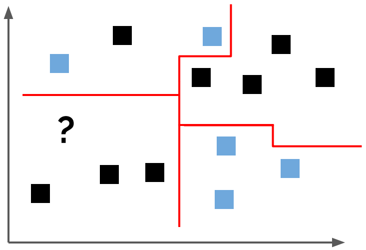
LOGISTIC REGRESSION



K-NEAREST NEIGHBORS



DECISIONS TREES



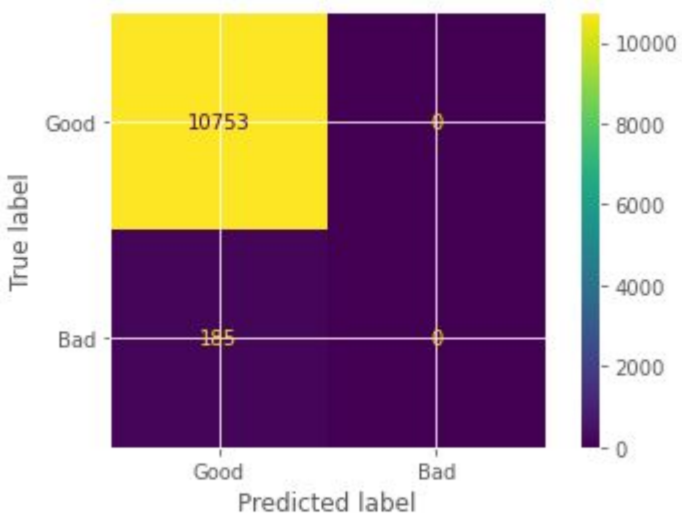
RANDOM FOREST

RAW DATA (UNBALANCED)

nice scores...

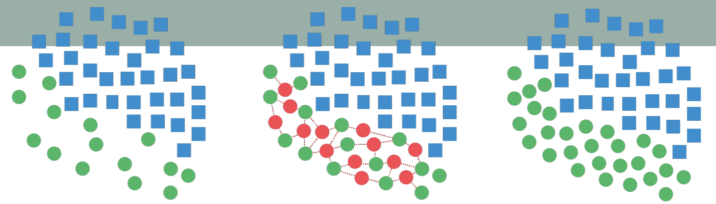
classifiers	accuracy	presicion	recall	f1_score
LogisticRegression	0.983086	0.966459	0.983086	0.974702
KNeighbors	0.981715	0.971128	0.981715	0.975116
DecisionTree	0.979978	0.972917	0.979978	0.975848
RandomForest	0.983086	0.966459	0.983086	0.974702

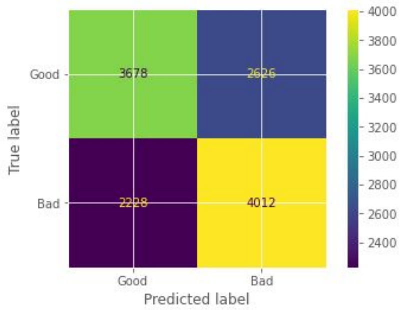
bad performance!



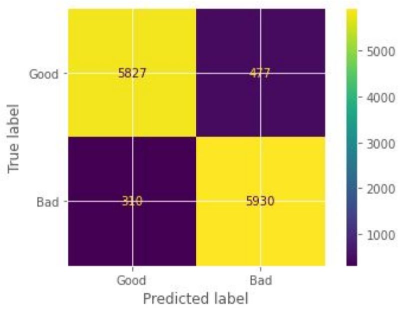
BALANCED DATA

SMOTE OVERSAMPLING

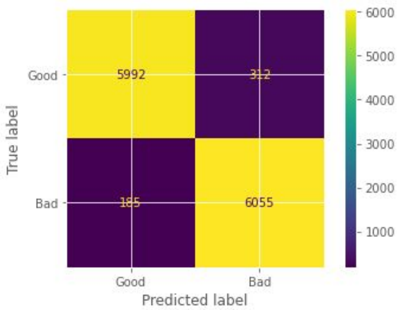




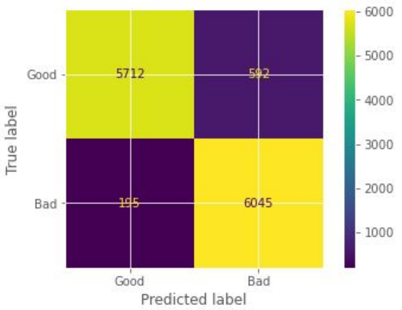
LogisticRegression()
score on train set: 0.6169
score on test set 0.6120



KNeighborsClassifier()
score on train set: 0.9464
score on test set 0.9312



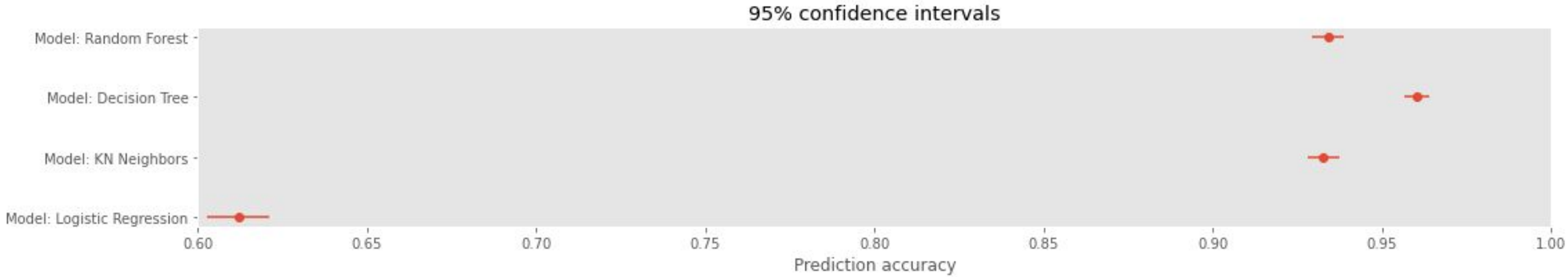
DecisionTreeClassifier()
score on train set: 0.9686
score on test set: 0.9601



RandomForestClassifier()
score on train set: 0.9391
score on test set: 0.9362

BALANCED DATA

CONFIDENCE INTERVALS



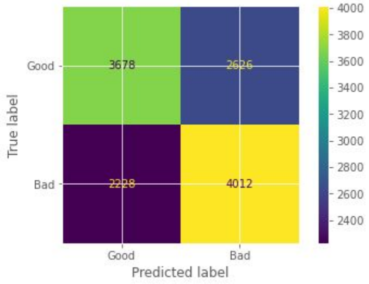
LOGISTIC REGRESSION OPTIMIZATION

SCALER TESTING

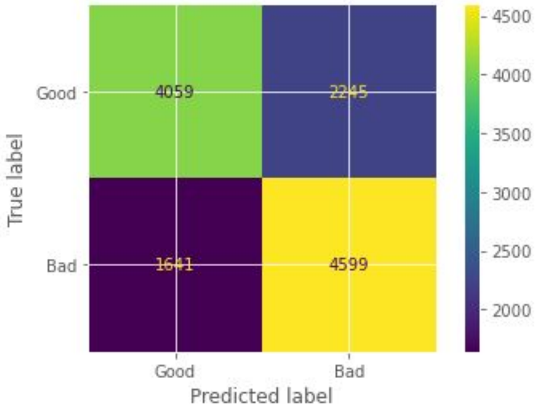
Logistic Regression Model improvement	
Scaler	Test set f1-Score
StandardScaler	0.61
MinMaxScaler	0.60
PolynomialScaler	0.70
RobustScaler	0.60

LOGISTIC REGRESSION OPTIMIZATION

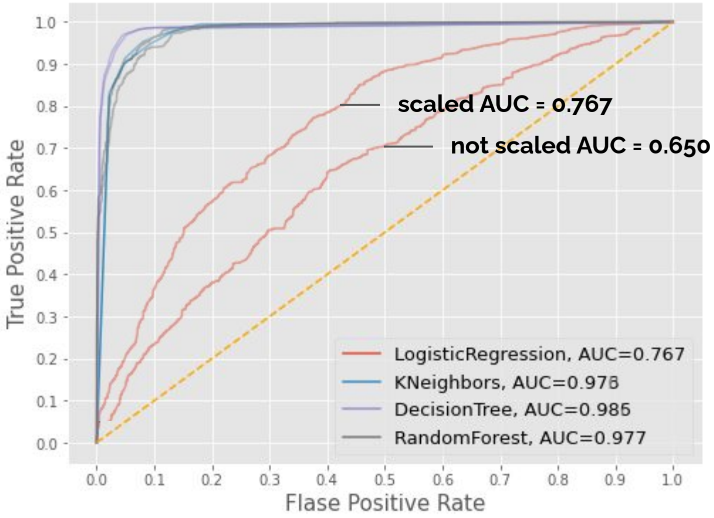
POLYNOMYAL SCALER



LogisticRegression()
score on train set: 0.6169
score on test set 0.6120



LogisticRegression()
score on train set: 0.6994
score on test set 0.6979



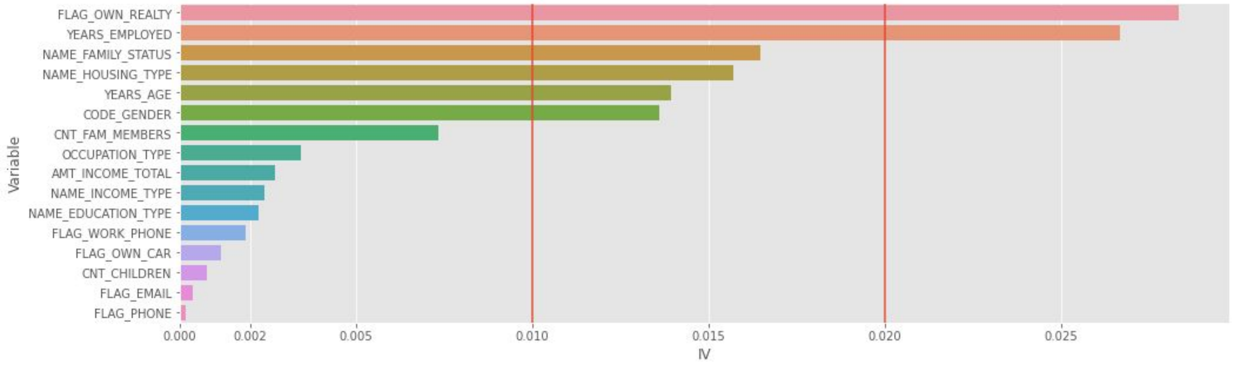
	accuracy	presicion	recall	f1_score
classifiers				
LogisticRegression	0.697943	0.700399	0.697943	0.69716

LOGISTIC REGRESSION OPTIMIZATION

FEATURE SELECTION (WoE & IV)

	Variable	IV
3	FLAG_OWN_REALTY	0.028349
14	YEARS_EMPLOYED	0.026667
7	NAME_FAMILY_STATUS	0.01645
15	NAME_HOUSING_TYPE	0.015711
1	YEARS_AGE	0.013915
0	CODE_GENDER	0.013581
9	CNT_FAM_MEMBERS	0.007332
12	OCCUPATION_TYPE	0.003413
13	AMT_INCOME_TOTAL	0.002682
11	NAME_INCOME_TYPE	0.00239
10	NAME_EDUCATION_TYPE	0.00223
4	FLAG_WORK_PHONE	0.001865
2	FLAG_OWN_CAR	0.001144
8	CNT_CHILDREN	0.000748
6	FLAG_EMAIL	0.000366
5	FLAG_PHONE	0.00015
16	CREDIT_SCORE	None

- less than 0.02: useless for prediction
- 0.02 to 0.10: weak predictor
- 0.10 to 0.3: medium predictor
- 0.30 to 0.5: strong predictor
- more than 0.5: too good to be true (suspicious)

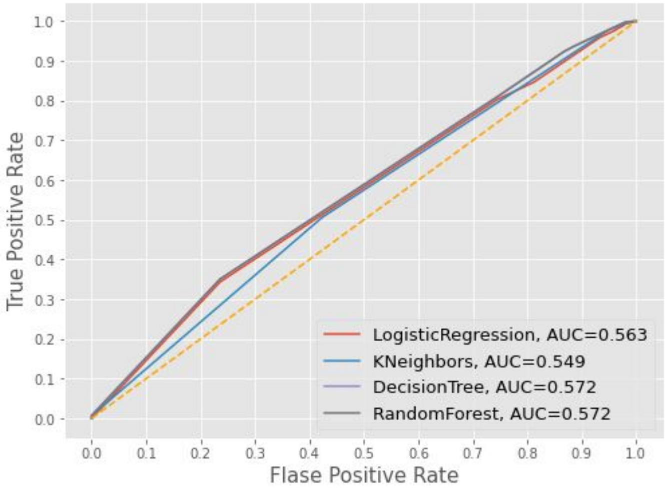


LOGISTIC REGRESSION OPTIMIZATION

FEATURE SELECTION (WoE & IV)

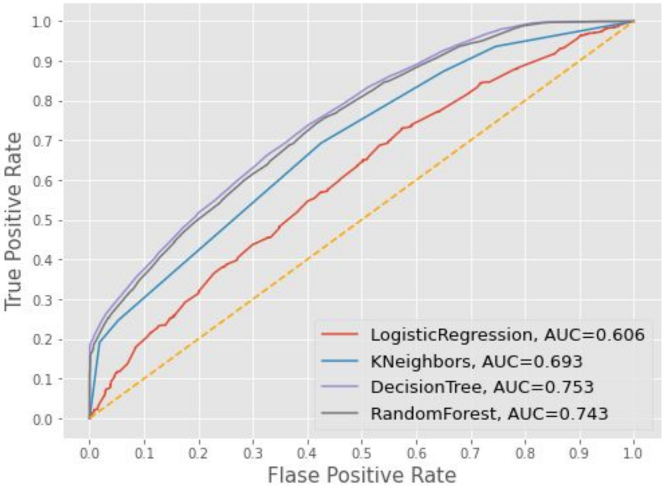
FEATURES USED (IV > 0.02):

FLAG_OWN_REALTY
YEARS_EMPLOYED



FEATURES USED (IV > 0.01):

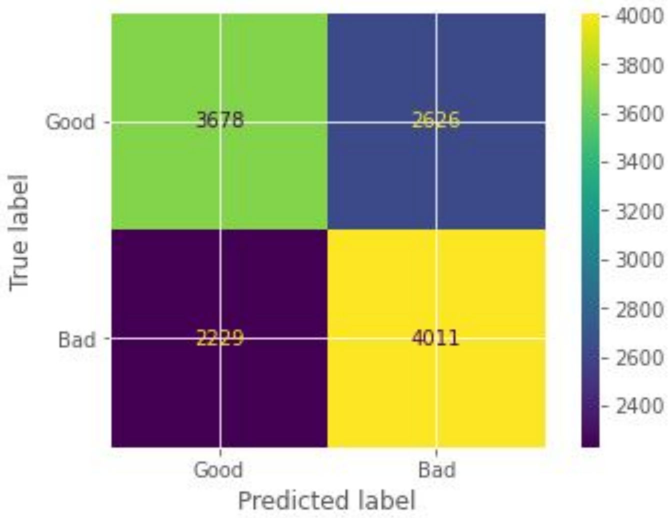
FLAG_OWN_REALTY
YEARS_EMPLOYED
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
YEARS_AGE
CODE_GENDER



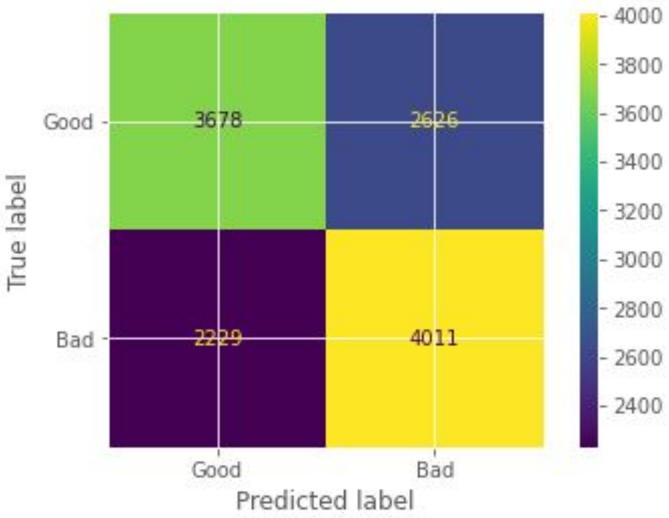
LOGISTIC REGRESSION OPTIMIZATION

HYPERTUNING

SAGA



MULTI_CLASS = ‘ovr’



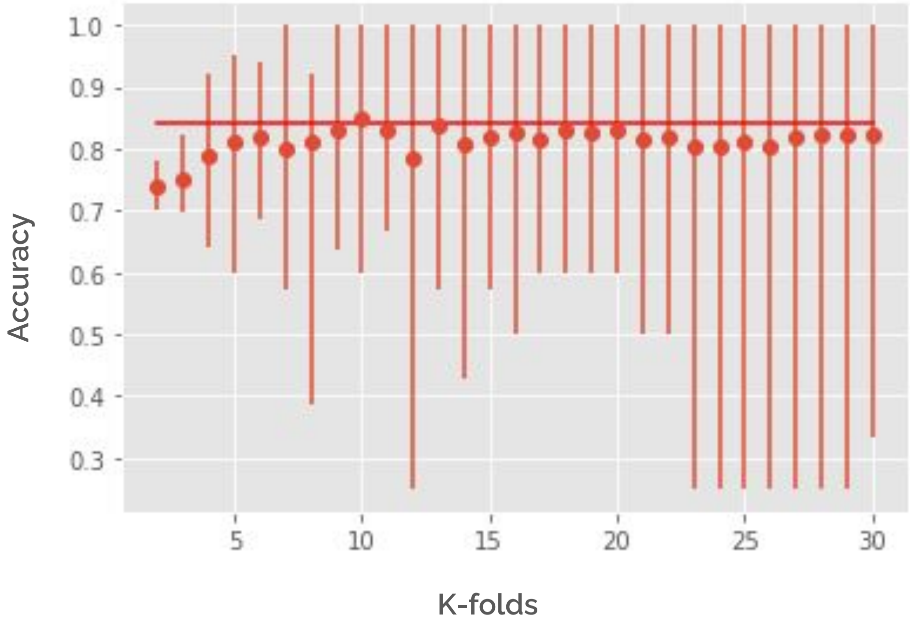
OVERFITTING

LOGISTIC REGRESSION OPTIMIZATION

K FOLD CROSS VALIDATION

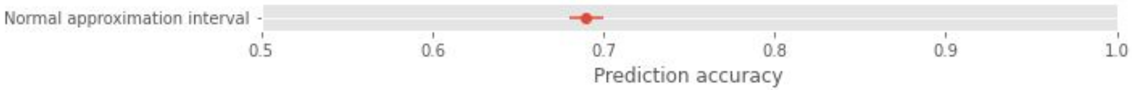
K-folds = 13

Ideal Accuracy = 0.84



LOGISTIC REGRESSION PREDICTIONS

CONCLUSIONS:



Jack is a young man who just started at the University. He lives in a rented apartment, is single and available on Tinder and enjoys travelling.
Is he gonna get credit?

NO

Susan has been retired for 3 years now. She joined salsa group after his husband passed away last year. She lives in a municipal apartment at the center of the city.
Is she gonna get credit?

NO

Lily married 10 years ago a musician. She inherited a house from her parents at the outskirts of the city. She has one kid and they are planning to have another one.
Is she gonna get credit?

YES

THANK YOU

IRON HACK

Ironhack Private Bank Services