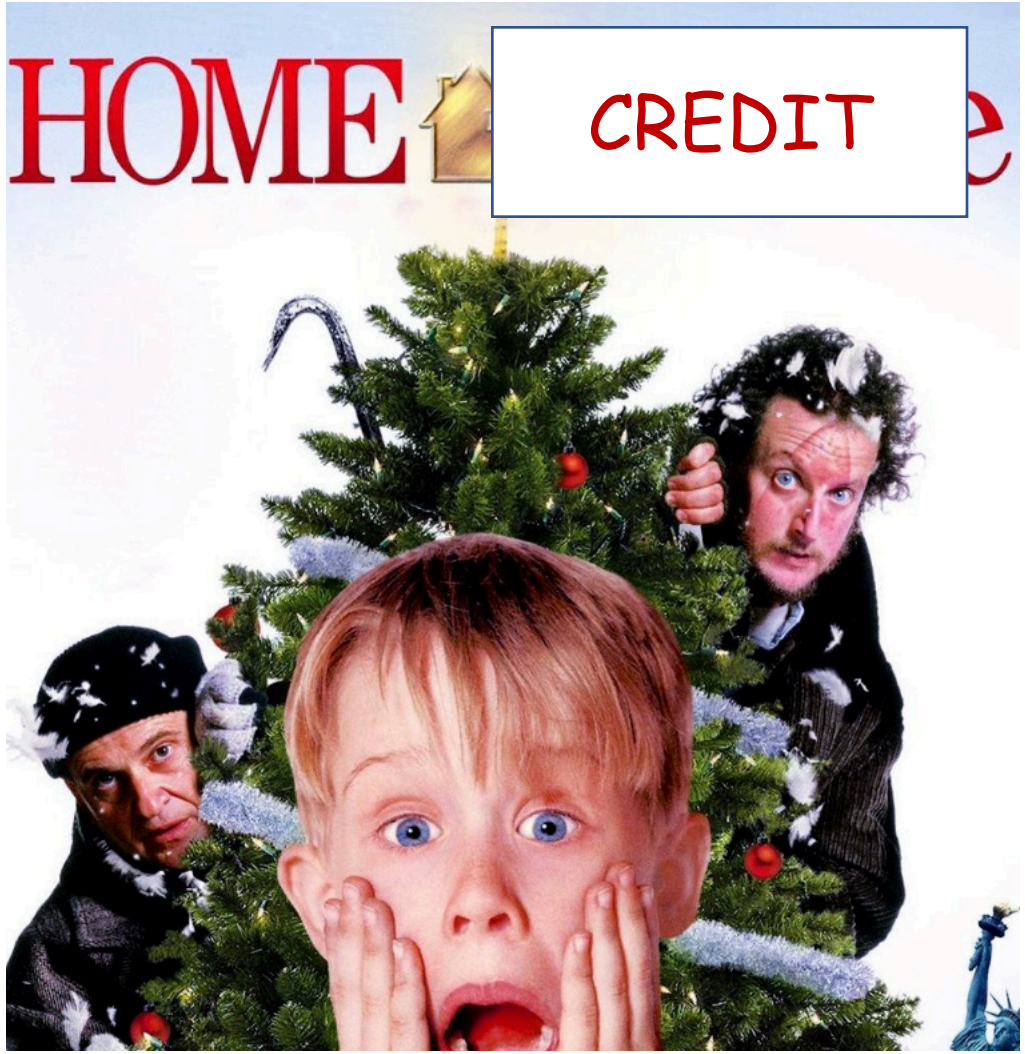
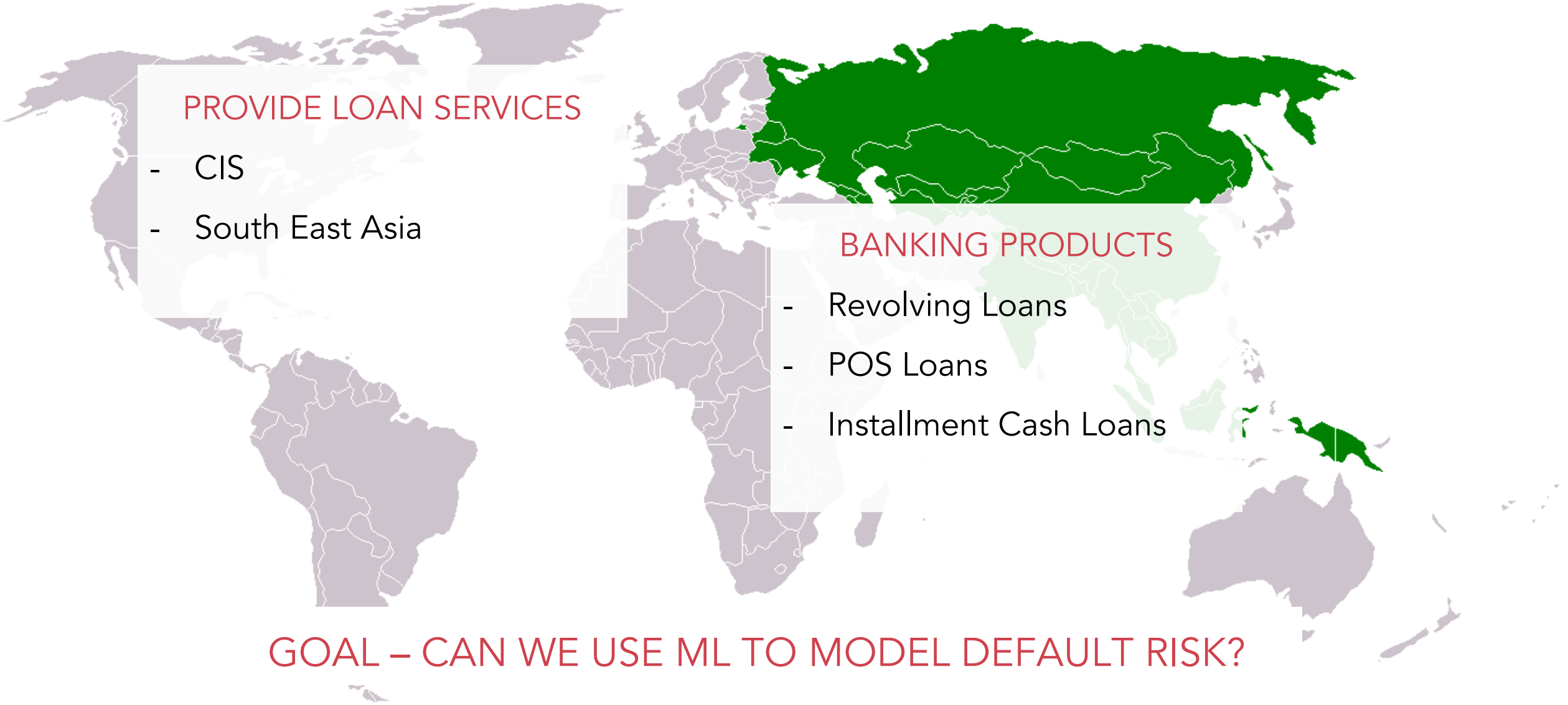

HOME
CREDIT

HOME CREDIT INTRODUCTION



PROVIDE LOAN SERVICES

- CIS
- South East Asia

BANKING PRODUCTS

- Revolving Loans
- POS Loans
- Installment Cash Loans

GOAL – CAN WE USE ML TO MODEL DEFAULT RISK?

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE:

FEATURE NAME	FEATURE DESCRIPTION

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
TARGET	Target Variable (1 – default, 0 – all other)
NAME_CONTRACT_TYPE	ID if loan is cash or revolving
CODE_GENDER	Gender of client
CNT_CHILDREN	Number of children created by the client
AMT_INCOME_TOTAL	Income of the client
AMT_CREDIT	Credit amount of the loan

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: bureau.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item
CREDIT_ACTIVE	Status of credit bureau item
CREDIT_TYPE	Type of credit
MONTHS_BALANCE	How many months ago relative to application in application.csv
AMT_ANNUITY	Annuity of bureau credit
AMT_CREDIT_SUM	Current credit amount

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item
MONTHS BALANCE	How many months ago relative to application in application.csv
STATUS	Status of loan (active, closed, delayed payment)

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: pos_cash.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan
MONTHS_BALANCE	Month of balance relative to application in application.csv
CNT_INSTALMENT	Term of previous credit
CNT_INSTALMENT_FUTURE	Remaining installments on credit
NAME_CONTRACT_STATUS	Contract status during the month
SK_DPD	Days past due of credit

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: cc_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan
AMT_BALANCE	Balance for month
AMT_DRAWINGS_CURRENT	Total drawings during month
NAME_CONTRACT_STATUS	Type of contract
AMT_RECEIVABLE	Amount receivable on previous credit
AMT_PAYMENT_CURRENT	Current month payment on previous credit

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: previous_app.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan
AMT_CREDIT	Credit amount on previous application
AMT_APPLICATION	Total amount client applied for
NAME_SELLER_INDUSTRY	Industry of seller
PRODUCT_COMBINATION	Detailed product combination
DAYS_TERMINATION	Expected termination of previous application relative to application.csv

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

FILE: installments.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan
NUM_INSTALMENT_VERSION	Version of installment calendar
NUM_INSTALMENT_NUMBER	Installment number of payment
DAYS_INSTALLMENT	When the previous credit was to be paid
DAYS_ENTRY_PAYMENT	Actual payment day of installment
AMT_INSTALLMENT	Total installment amount

EDA – DATASET INTRODUCTION

FILE

application.csv

bureau.csv

bureau_balance.csv

pos_cash.csv

cc_balance.csv

previous_app.csv

installments.csv

EDA – DATASET INTRODUCTION

FILE	RECORDS
application.csv	356,251
bureau.csv	1,716,428
bureau_balance.csv	27,299,925
pos_cash.csv	10,001,358
cc_balance.csv	3,840,312
previous_app.csv	1,670,214
installments.csv	3,605,401

EDA – DATASET INTRODUCTION

FILE	RECORDS	FEATURES
application.csv	356,251	123
bureau.csv	1,716,428	17
bureau_balance.csv	27,299,925	3
pos_cash.csv	10,001,358	8
cc_balance.csv	3,840,312	23
previous_app.csv	1,670,214	37
installments.csv	3,605,401	8

EDA – DATASET INTRODUCTION

FILE	RECORDS	FEATURES	FEATURES AFTER GENERATION
application.csv	356,251	123	293
bureau.csv	1,716,428	17	133
bureau_balance.csv	27,299,925	3	0
pos_cash.csv	10,001,358	8	46
cc_balance.csv	3,840,312	23	118
previous_app.csv	1,670,214	37	363
installments.csv	3,605,401	8	31

FEATURE ENGINEERING

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
TARGET	Target Variable (1 – default, 0 – all other)

FILE: bureau.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item

FILE: previous_app.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item

FILE: installments.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FILE: pos_cash.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FILE: cc_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FEATURE ENGINEERING

CALCULATED FEATURES

- Groupby Aggregations
- Polynomial Features

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	
TARGET	

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY
1	2802425	108129	Cash loans	25188.615
335841	1536272	108129	Cash loans	21709.125
588441	2068863	108129	Consumer loans	4830.930
617224	2551979	108129	Consumer loans	6664.275
692217	2517198	108129	Revolving loans	11250.000
1380202	1760610	108129	Consumer loans	8593.965

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item

FILE: installments.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FILE: pos_cash.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FILE: cc_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

FEATURE ENGINEERING

CALCULATED FEATURES

- Groupby Aggregations
- Polynomial Features

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	
TARGET	

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY
1	2802425	108129	Cash loans	25188.615
335841	1536272	108129	Cash loans	21709.125
588441	2068863	108129	Consumer loans	4830.930
617224	2551979	108129	Consumer loans	6664.275
692217	2517198	108129	Revolving loans	11250.000
1380202	1760610	108129	Consumer loans	8593.965

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item

1. SPLIT DATAFRAME BY NUMERIC / CATEGORICAL VALUES
2. GROUPBY SK_ID_CURR
3. PERFORM AGGREGATION CALCULATIONS
4. APPEND TO APPLICATION DATAFRAME

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

SK_ID_CURR	previous_AMT_ANNUITY_count	previous_AMT_ANNUITY_mean	previous_AMT_ANNUITY_max	previous_AMT_ANNUITY_min	previous_AMT_ANNUITY_sum
108129	6	13039.485	25188.615	4830.93	78236.91

FEATURE ENGINEERING

DOMAIN FEATURES

- Time Series
- Handcrafted features

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
TARGET	Target Variable (1 – defa

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item

FILE: pos_cash.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_ID_PREV	Unique ID of previous home credit / loan

SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE
2562384	378907	-2	28.575
2562384	378907	-3	1897.875
2562384	378907	-4	4036.860
2562384	378907	-5	5228.280
2562384	378907	-6	56.970
2562384	378907	-7	1687.365
2562384	378907	-8	129.240
2562384	378907	-9	9169.380
2562384	378907	-10	11112.525
2562384	378907	-11	13002.075
2562384	378907	-12	0.000
2562384	378907	-13	0.000
2562384	378907	-14	0.000
2562384	378907	-15	0.000
2562384	378907	-16	0.000
2562384	378907	-17	0.000
2562384	378907	-18	0.000
2562384	378907	-19	0.000

FEATURE ENGINEERING

DOMAIN FEATURES

- Time Series
- Handcrafted features

FILE: application.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
TARGET	

SK_ID_CURR	AMT_ANNUITY	AMT_CREDIT	CNT_PAYMENT
271877	1730.430	17145.0	12.0
108129	25188.615	679671.0	36.0
122040	15060.735	136444.5	12.0
176158	47041.335	470790.0	12.0
202054	31924.395	404055.0	24.0
199383	23703.930	340573.5	18.0

SK_ID_PREV	Unique ID of previous home credit / loan

FILE: bureau_balance.csv

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_BUREAU_ID	ID of previous credit bureau item

EXAMPLE FEATURE	FEATURE DESCRIPTION
SK_ID_CURR	Unique ID of client / loan
SK_BUREAU_ID	ID of previous credit bureau item

$$P = A \cdot \frac{1 - (1 + r)^{-n}}{r}$$

FILE: installments.csv

SK_ID_CURR	AMT_ANNUITY	AMT_CREDIT	CNT_PAYMENT	INTEREST_RATE
271877	1730.430	17145.0	12.0	3.08
108129	25188.615	679671.0	36.0	1.65
122040	15060.735	136444.5	12.0	4.61
176158	47041.335	470790.0	12.0	2.91
202054	31924.395	404055.0	24.0	5.91
199383	23703.930	340573.5	18.0	2.49

SK_ID_PREV	Unique ID of previous home credit / loan

FILE: pos_cash.csv

EXAMPLE FEATURE	FEATURE DES
SK_ID_CURR	Unique ID of client
SK_ID_PREV	Unique ID of previous home credit / loan

MODEL SELECTION

BASELINE MODELS

- Logistic Regression
- Random Forest Classifier

LIGHTGBM

MODEL SELECTION

BASELINE MODELS

- Logistic Regression
- Random Forest Classifier

LIGHTGBM

- Fast

```
df_merged.shape
```

```
(307511, 986)
```

MODEL SELECTION

BASELINE MODELS

- Logistic Regression
- Random Forest Classifier

LIGHTGBM

- Fast

```
df_merged.shape
```

```
(307511, 986)
```

TABLE III
TIME AND AUC USING LIGHTGBM

#Rows	AUC	Time
307507	0.789996	786
250000	0.788589	638
200000	0.786344	512
150000	0.786215	393
100000	0.782477	263
50000	0.777649	121

TABLE II
TIME AND AUC USING XGBOOST

#Rows	AUC	Time
307507	0.788320	4306
250000	0.784516	3550
200000	0.781219	2892
150000	0.773347	2098
100000	0.772771	1219
50000	0.768899	9487

Source: "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset", Essam Al Daoud

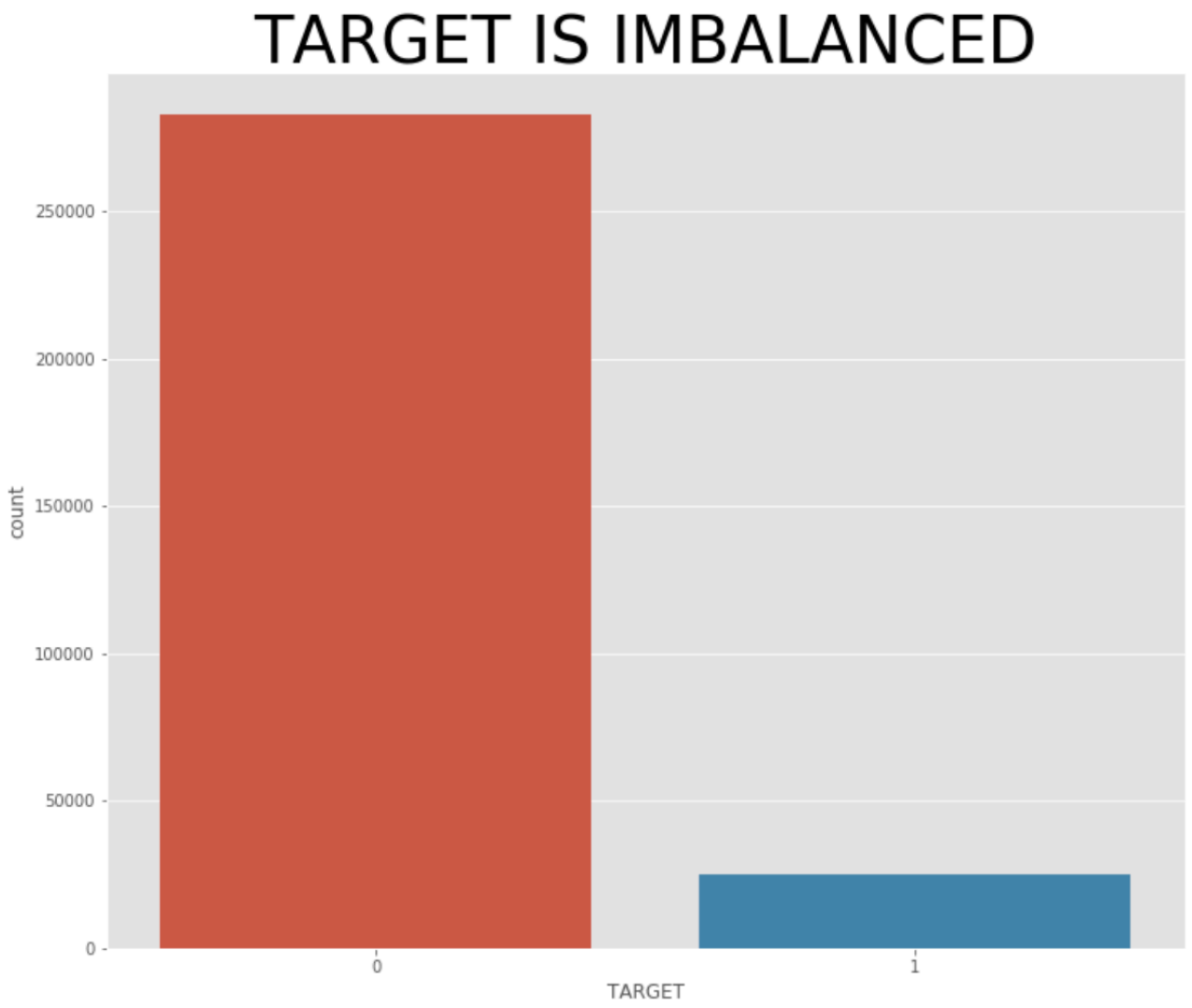
MODEL SELECTION

BASELINE MODELS

- Logistic Regression
- Random Forest Classifier

LIGHTGBM

- Fast
- Versatile
  - Regularization
  - Scale class weights
  - Handles null values



LIGHTGBM

LIGHTGBM – PROCESS

- Manual parameter tuning with subset of train data
- Stratified 5-Fold Cross Validation
- Assess AUC

LIGHTGBM – ROC CURVES

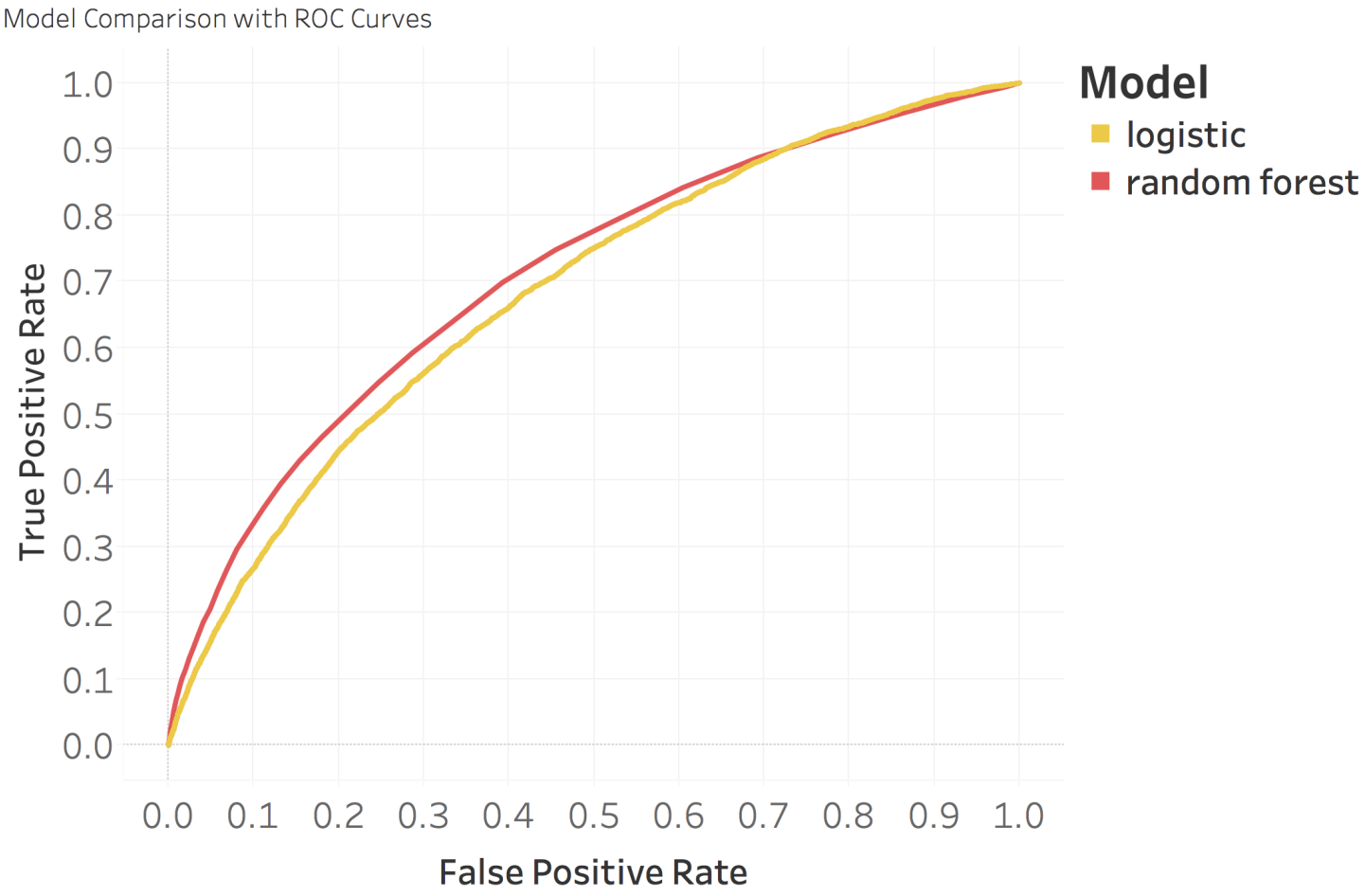
LIGHTGBM – PROCESS

- Manual parameter tuning with subset of train data
- Stratified 5-Fold Cross Validation
- Assess AUC

AUC SCORES BY MODEL

Logistic Regression: 0.6833

Random Forest: 0.7089



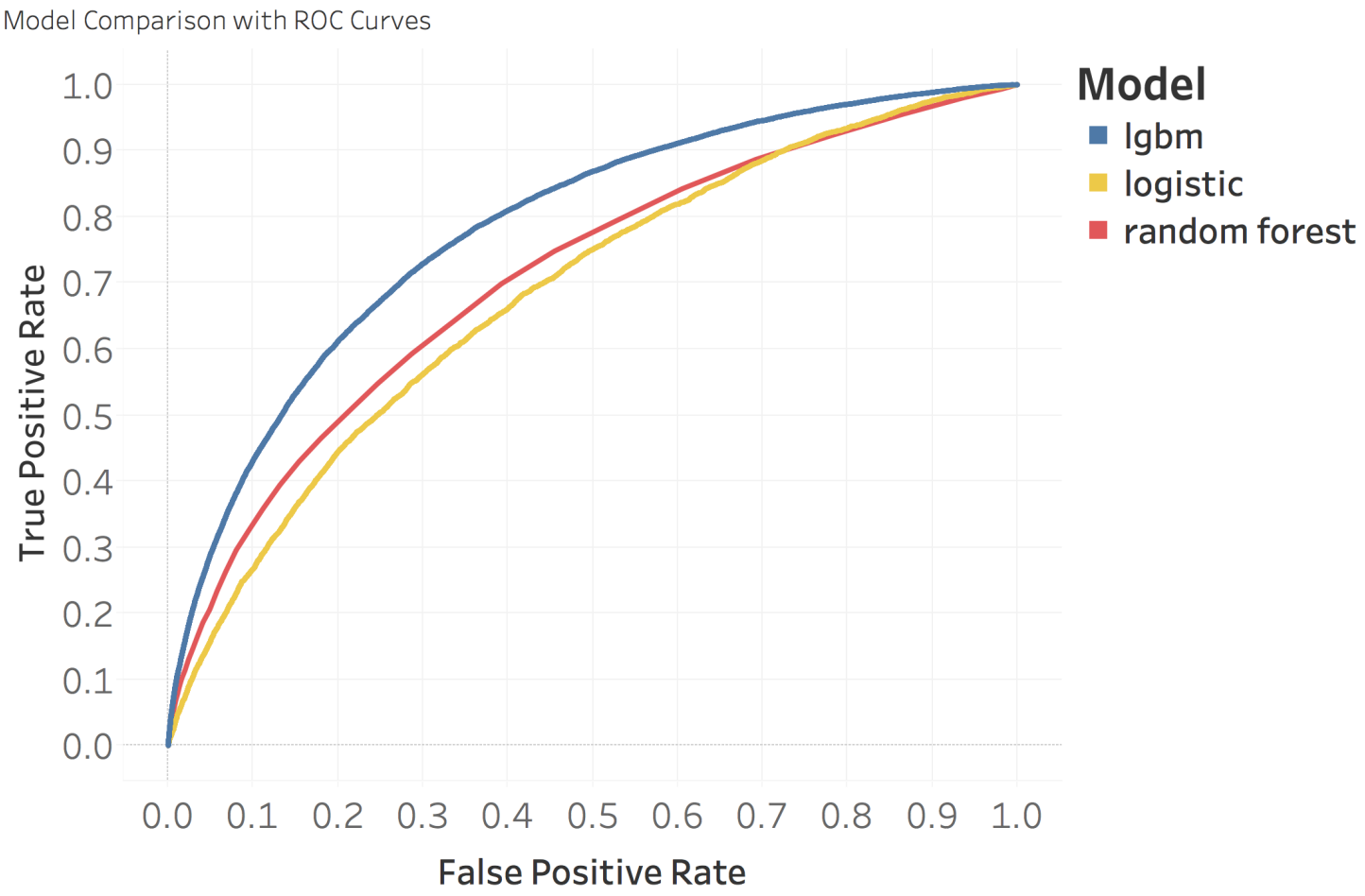
LIGHTGBM – ROC CURVES

LIGHTGBM – PROCESS

- Manual parameter tuning with subset of train data
- Stratified 5-Fold Cross Validation
- Assess AUC

AUC SCORES BY MODEL

Logistic Regression: 0.6833
Random Forest: 0.7089
LightGBM: 0.7839



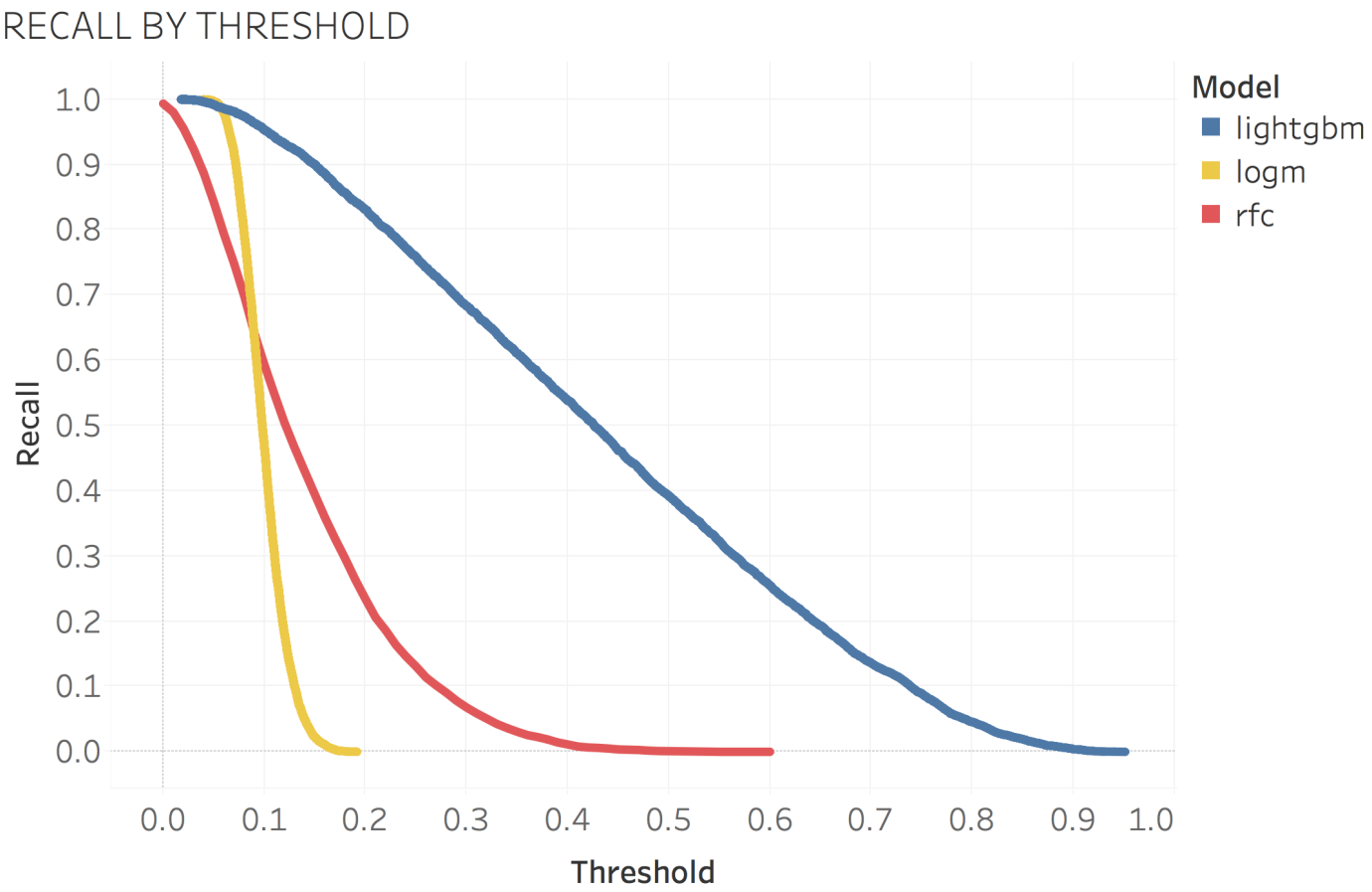
LIGHTGBM - RECALL

LIGHTGBM – PROCESS

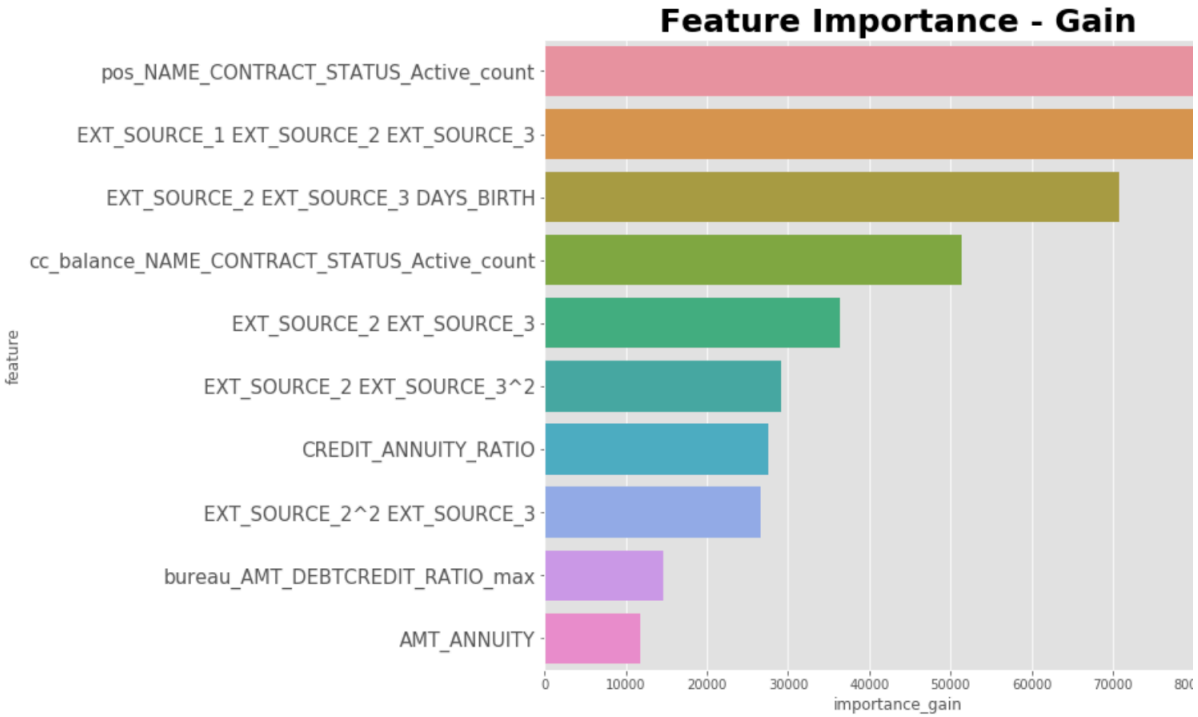
- Manual parameter tuning with subset of train data
- Stratified 5-Fold Cross Validation
- Assess AUC

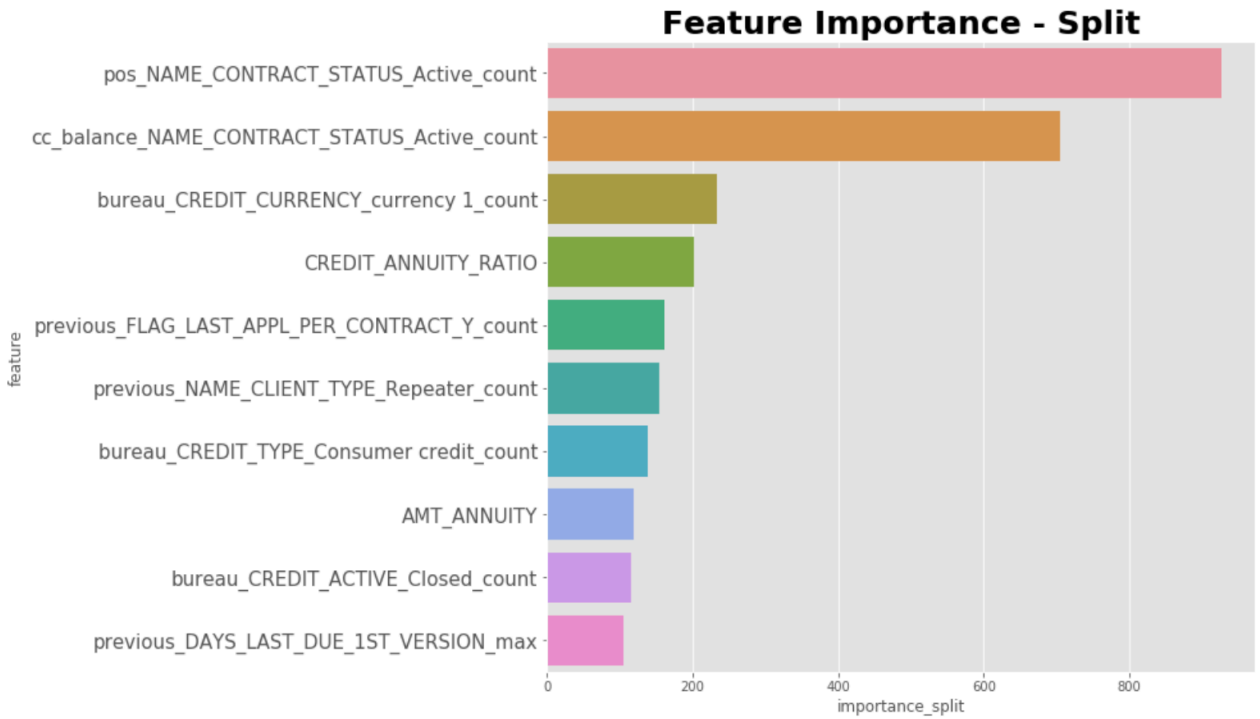
AUC SCORES BY MODEL

Logistic Regression: 0.6833
Random Forest: 0.7089
LightGBM: 0.7839

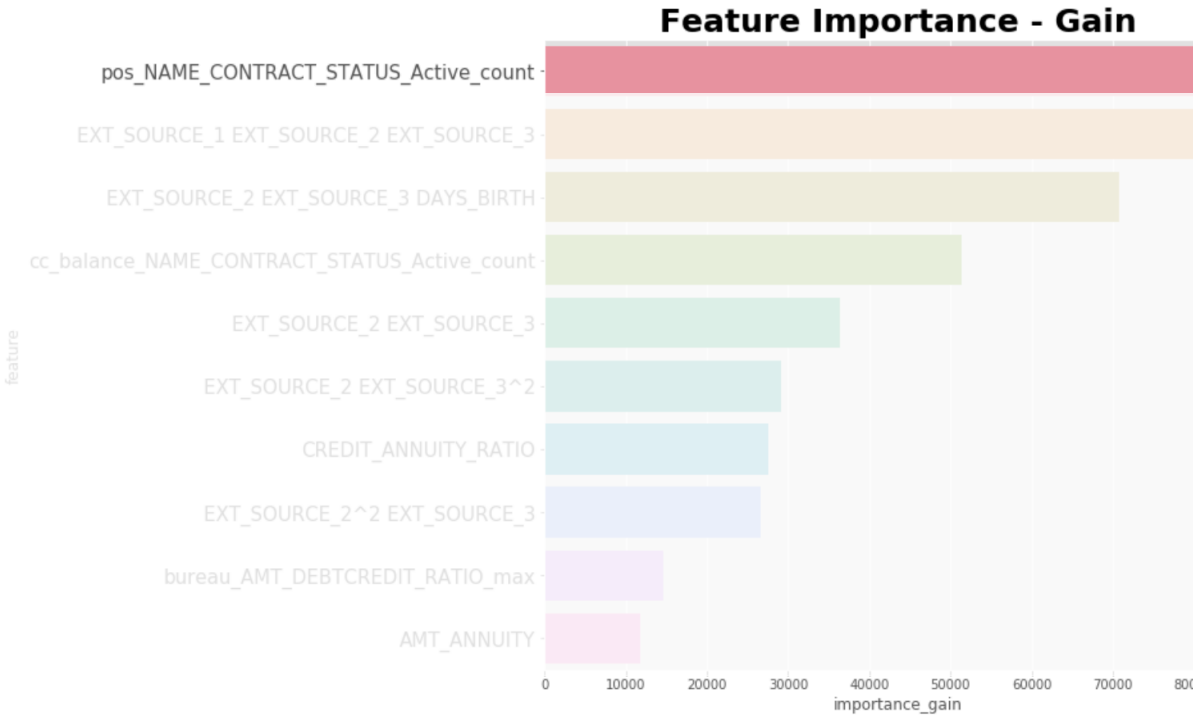


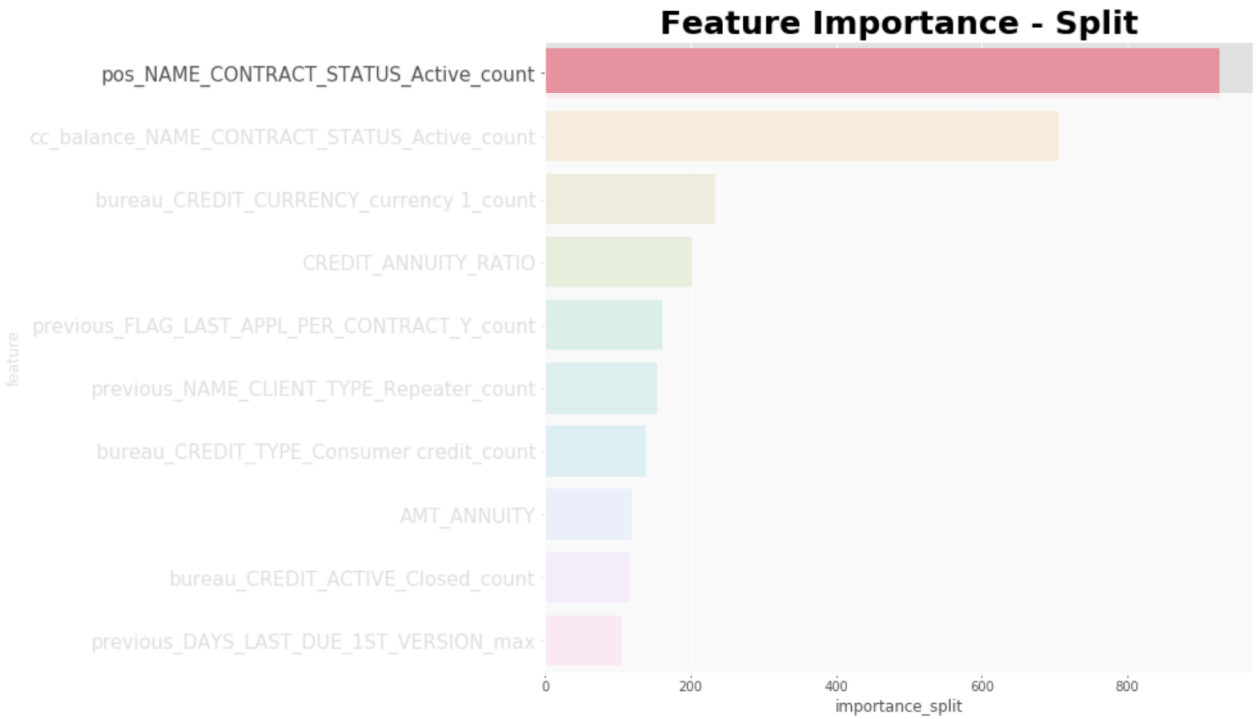
FEATURE IMPORTANCE



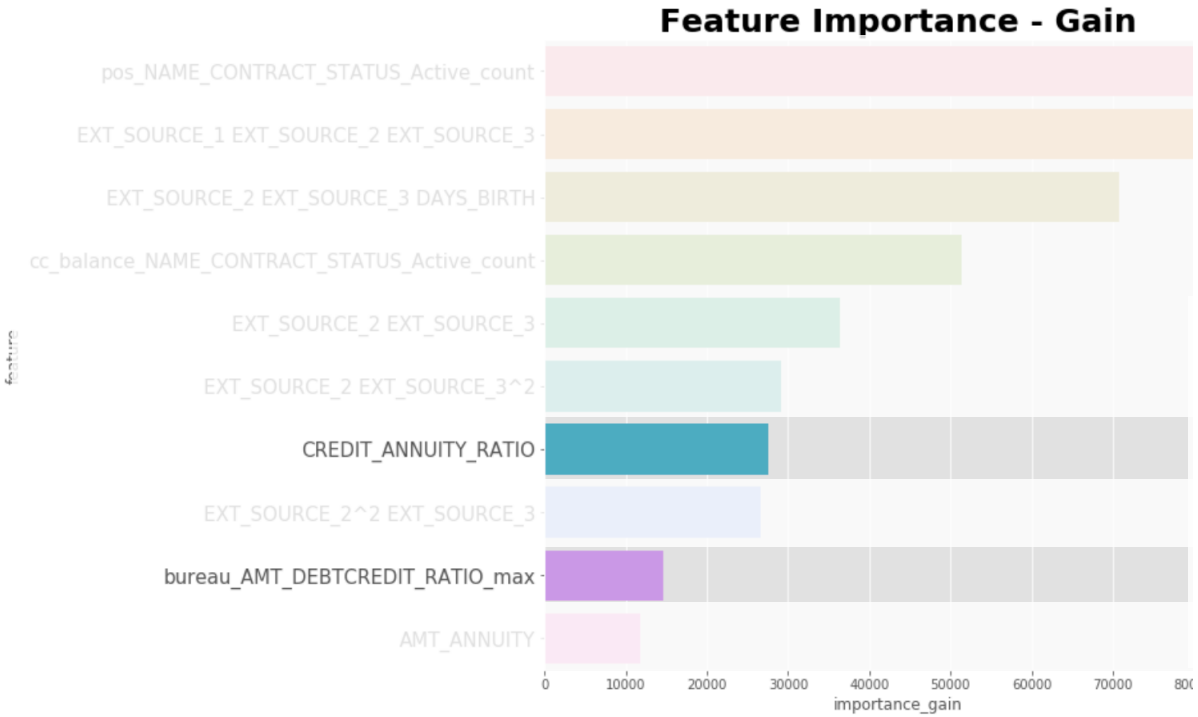


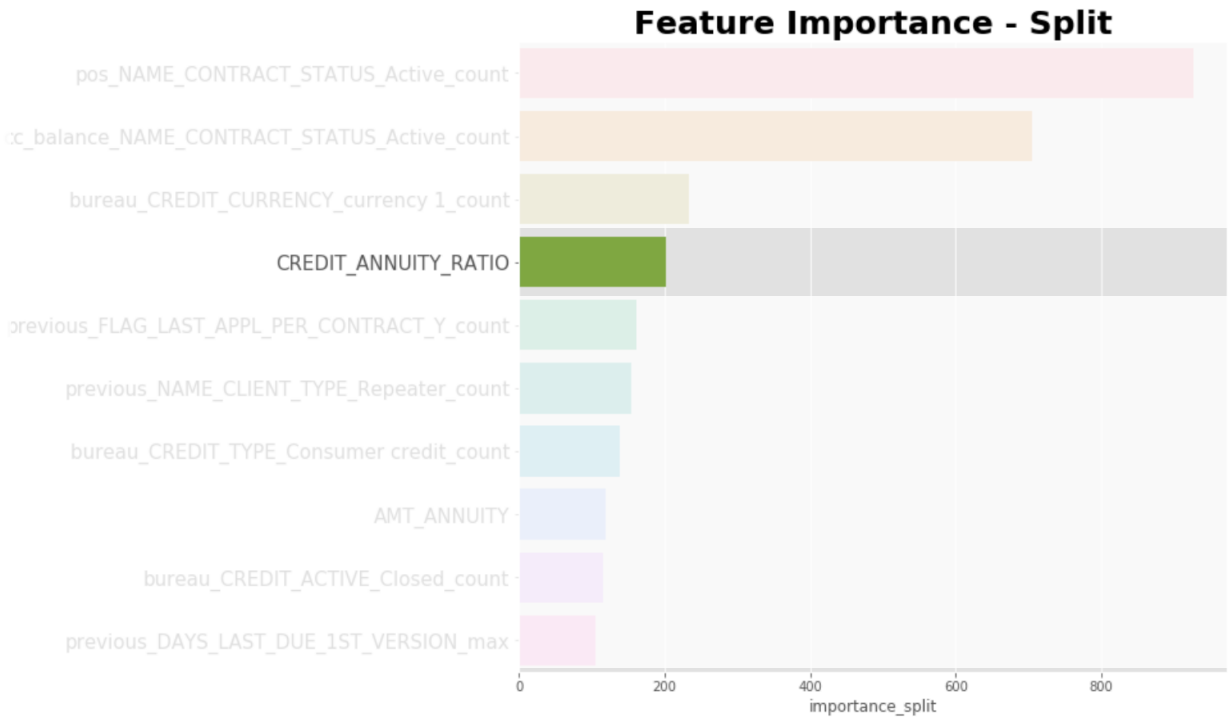
FEATURE IMPORTANCE



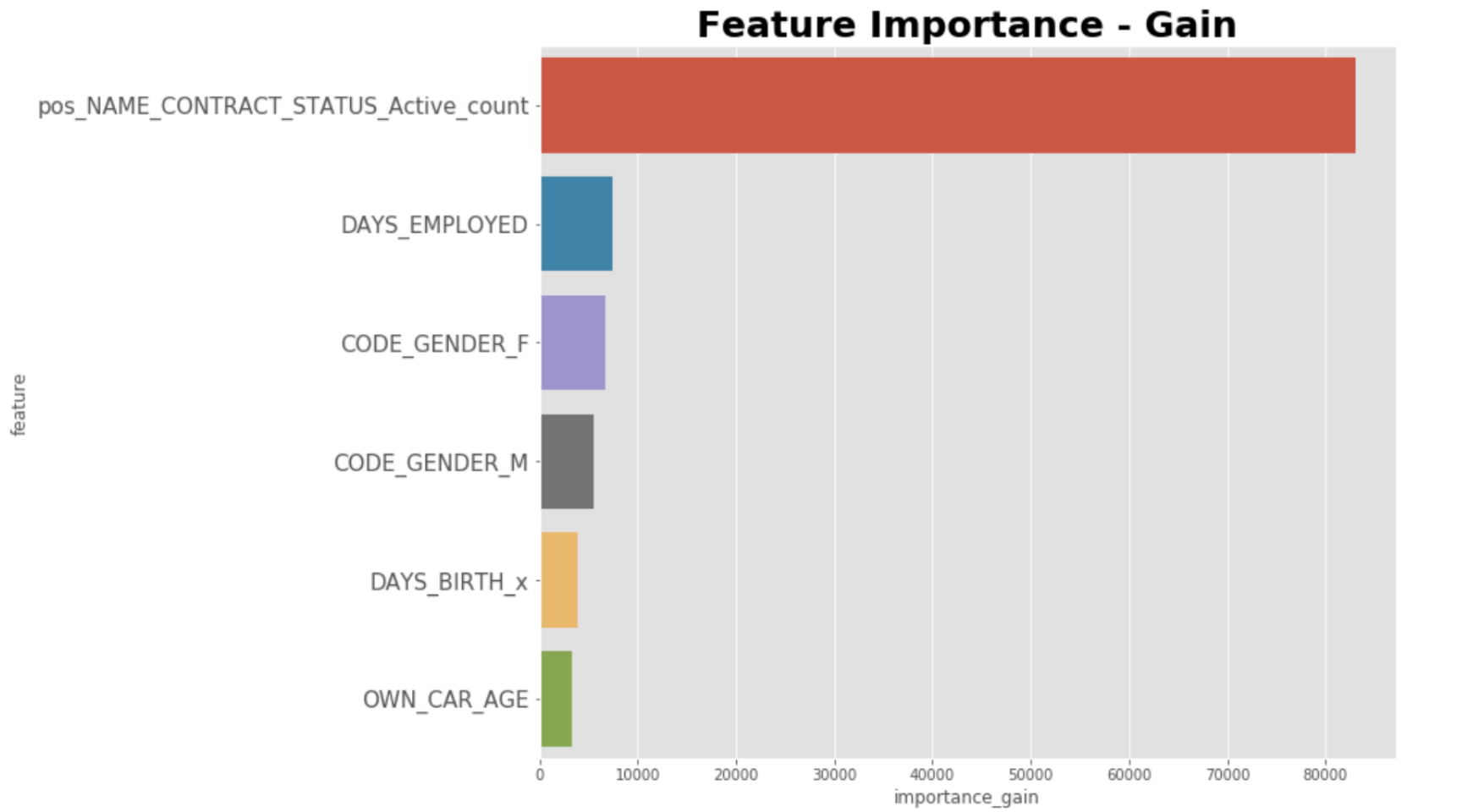


HANDCRAFTED FEATURES PERFORMED WELL





FEATURE IMPORTANCE



EXAMPLE

	171125
NAME_CONTRACT_TYPE	1.0
FLAG_OWN_CAR	0.0
FLAG_OWN_REALTY	1.0
CNT_CHILDREN	12.0
AMT_INCOME_TOTAL	225000.0
AMT_CREDIT	202500.0
AMT_ANNUITY	10125.0
AMT_GOODS_PRICE	202500.0
REGION_POPULATION_RELATIVE	0.04622
DAYS_BIRTH_x	13894.0
DAYS_EMPLOYED	2542.0
DAYS_REGISTRATION_x	1867.0
DAYS_ID_PUBLISH_x	3709.0
OWN_CAR_AGE	-999999.0
FLAG_MOBIL	1.0

Probability of Default: 8.1%

	171125
NAME_CONTRACT_TYPE	1.0
FLAG_OWN_CAR	0.0
FLAG_OWN_REALTY	1.0
CNT_CHILDREN	12.0
AMT_INCOME_TOTAL	225000.0
AMT_CREDIT	202500.0
AMT_ANNUITY	20[illegible]5.0
AMT_GOODS_PRICE	202500.0
REGION_POPULATION_RELATIVE	0.04622
DAYS_BIRTH_x	13894.0
DAYS_EMPLOYED	2542.0
DAYS_REGISTRATION_x	1867.0
DAYS_ID_PUBLISH_x	3709.0
OWN_CAR_AGE	-999999.0
FLAG_MOBIL	1.0

Probability of Default: 8.1%

GOING FORWARD

QUESTIONS