Loan Default Study

By/ Mohamed Hassan

Bussines &Objective

- The world of finance is centered around lending money to individuals and businesses through loans.
- Loan applicants seek financial assistance for various purposes, and lenders evaluate their eligibility.
- **The key challenge** is predicting whether a loan applicant will repay or default based on their financial history and information.

Datasets &Strategy of work

- **Our dataset** is split into two key tables:
  - one for previous applications and another for current applications

- **Strategy of work** done in 3 phases:
  - Extracting aggregated insights from Previous Applications.
  - In-depth analysis of Current Applications.
  - Building a Predictive Model to Assess Default Risk

Phase 1

- We aggregated 3 main features for each current applicant who has a history in our databases.
- **Prev_status:** indicates combination between approval rate and rejection rate of previous loans:
  - 1 for one who has an approval rate greater than rejection **+ a threshold.**

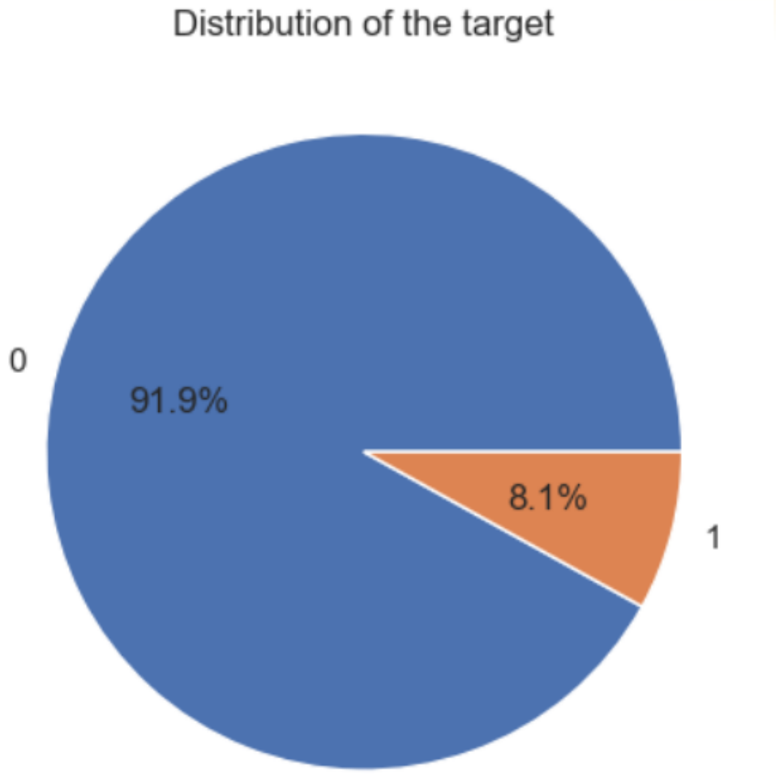
sk_id_curr	prev_avg_approved_credit	prev_avg_approved_annuity	prev_status
100002	179055.0000	9251.77500	1
100003	484191.0000	56553.99000	1
100035	203042.8125	16788.74625	-1
100072	133302.0000	12358.00500	0

Phase 2

Working with current applications
and getting insights from it.

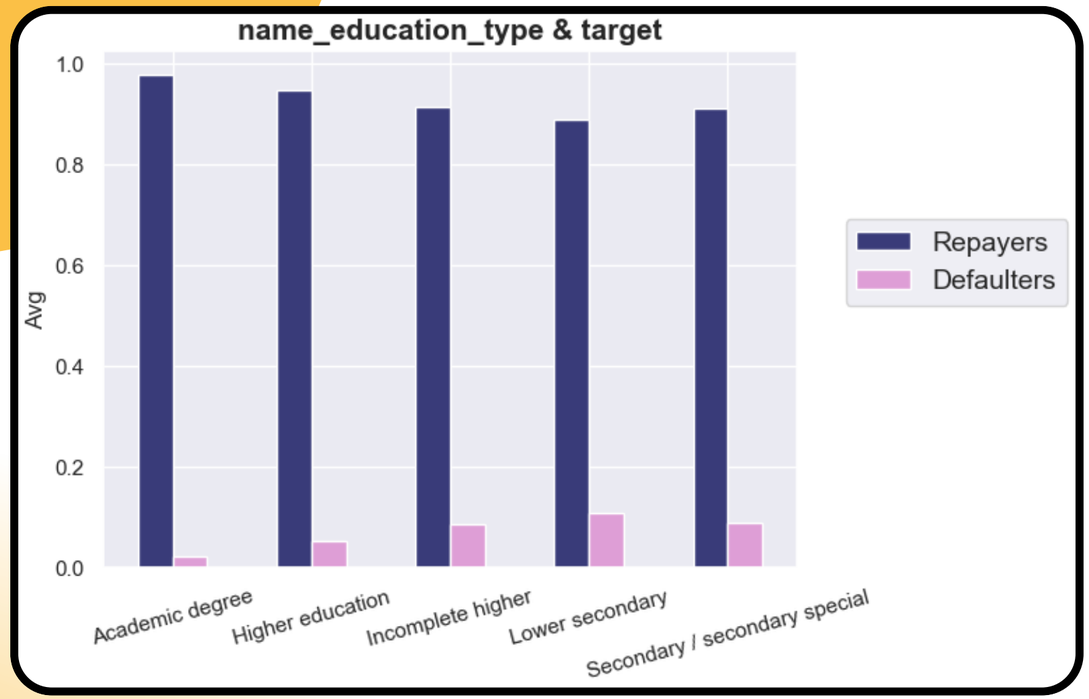
Identifying Problems & working with it

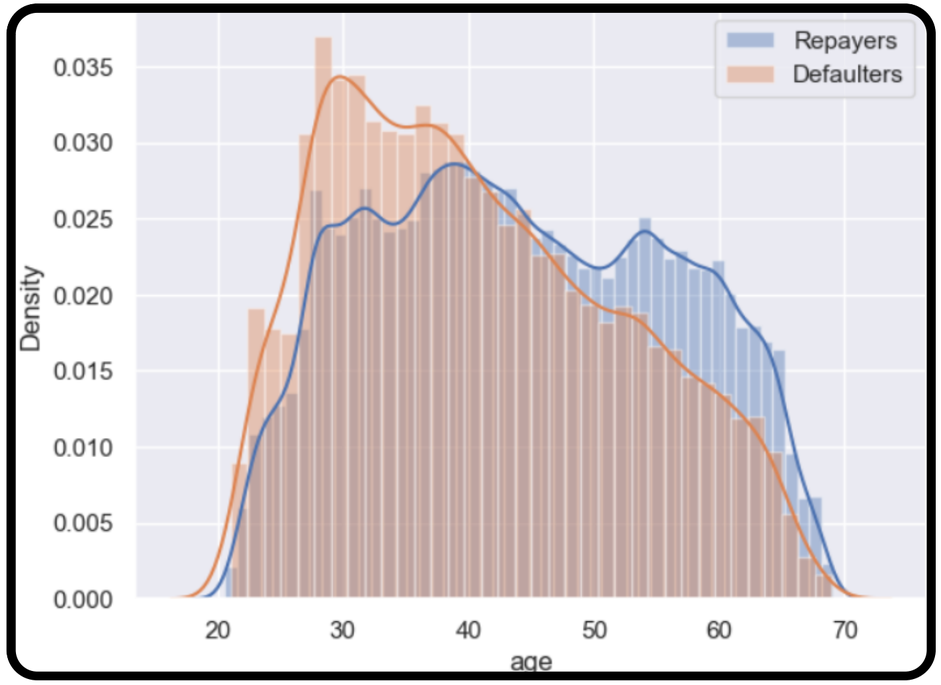
- **Huge number of missing values** and many features had more than 50% missing values. And handled these values by getting the best alternative values for missing values based on other columns.
- **Many categorical values** need to be converted to numeric values.
- **Imbalanced Data** as there are the repayers were ~92%



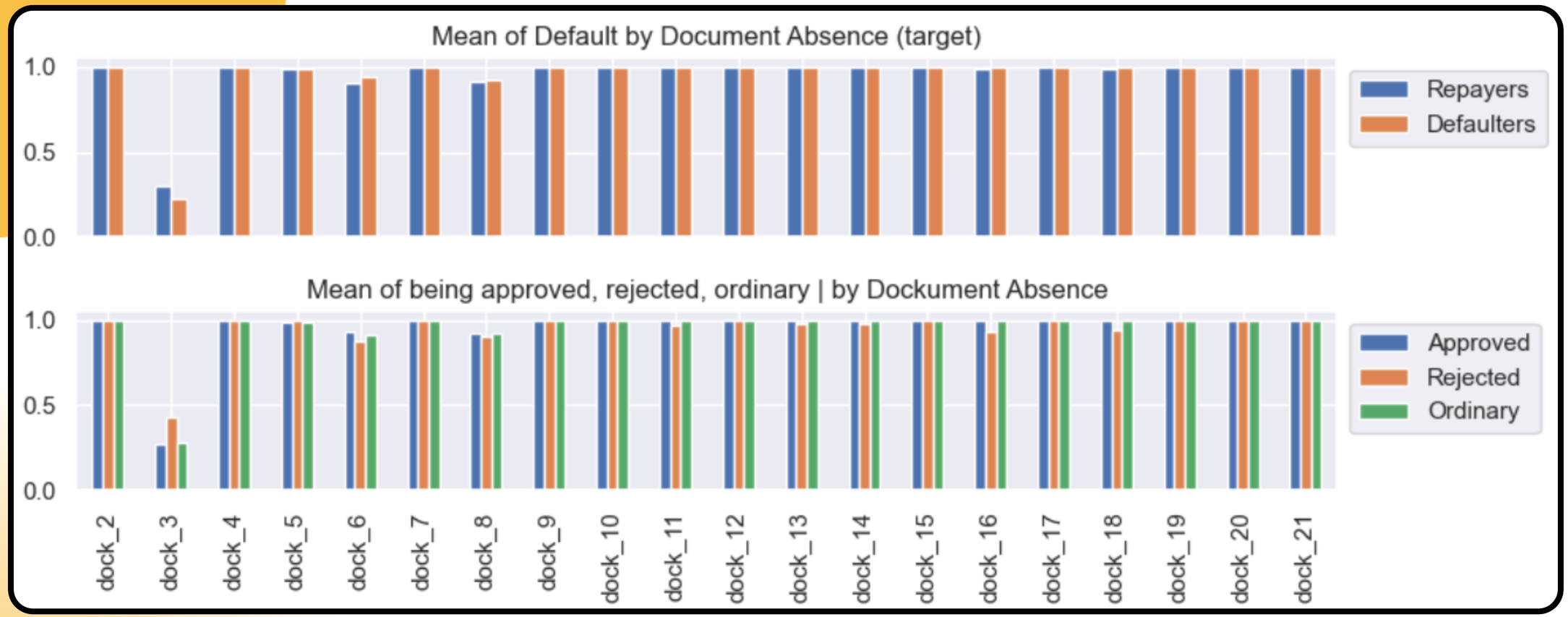
Q&A with Exploratory Data Analysis (EDA)

How do the clients' Demographics such as the age stages related to de

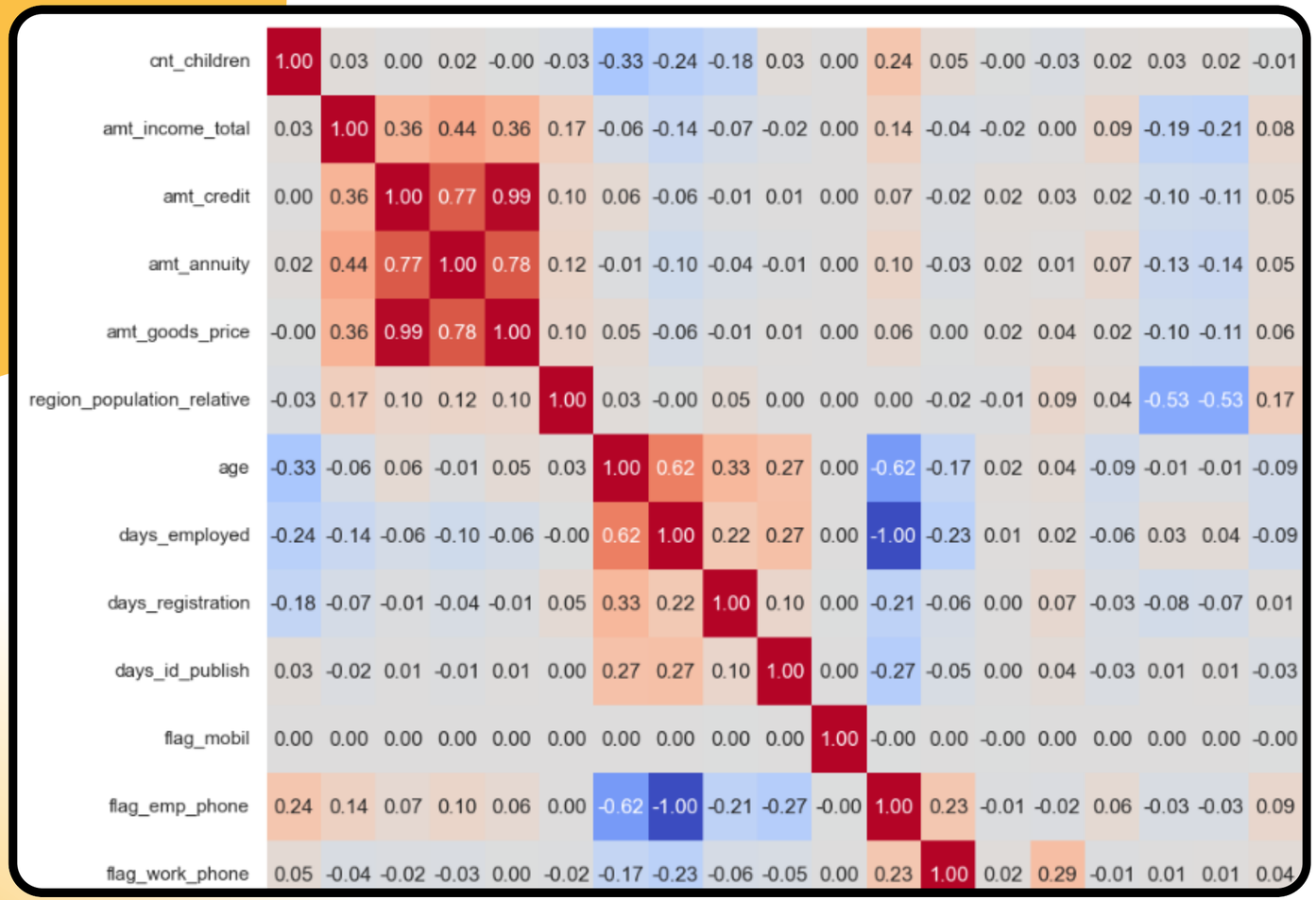




What impact does the absence of docks have on loan rejection or applicant default?



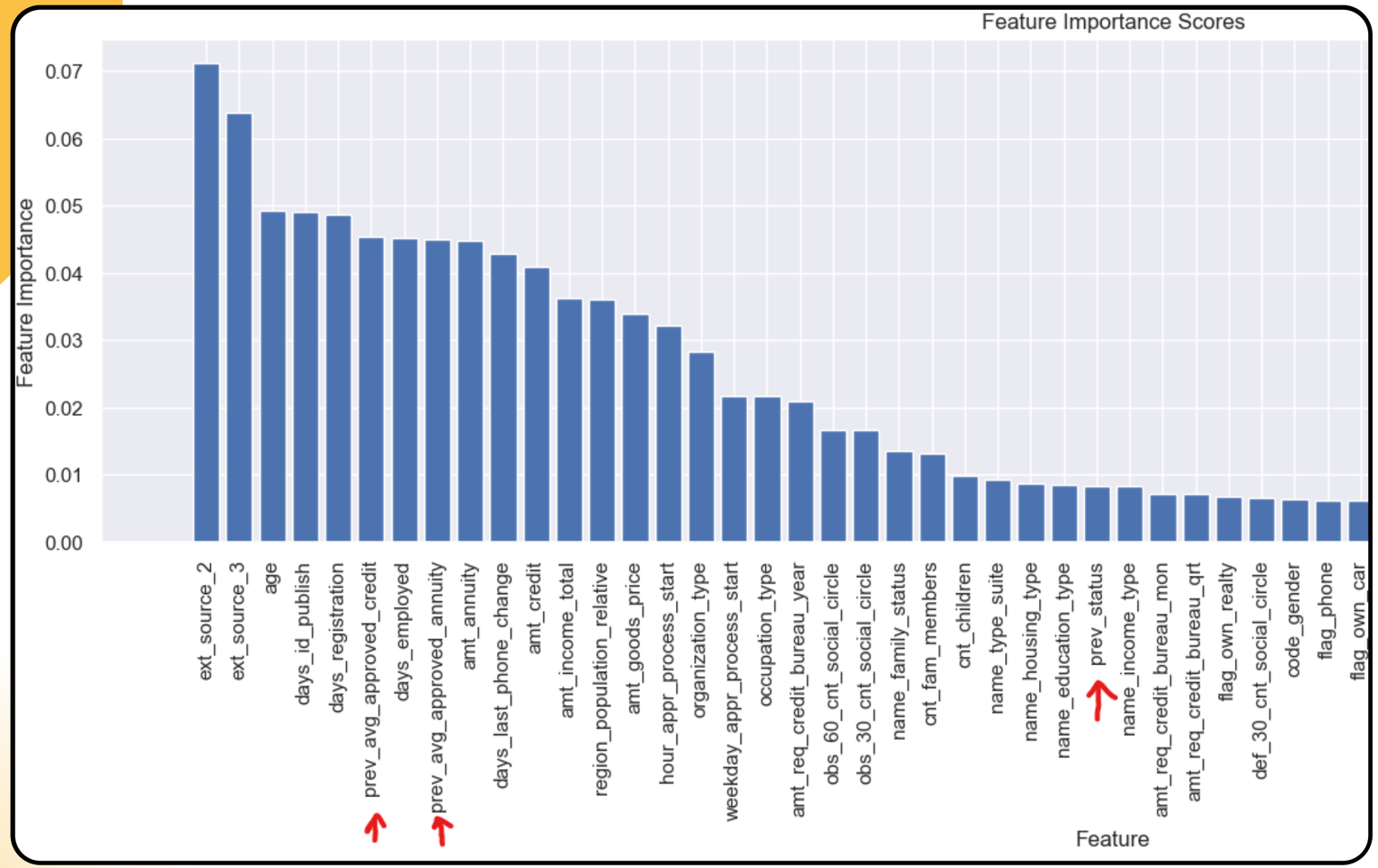
Indicating Multicolleniarity between features



	cnt_children	amt_income_total	amt_credit	amt_annuity	amt_goods_price	region_population_relative	age	days_employed	days_registration	days_id_publish	flag_mobil	flag_emp_phone	flag_work_phone
cnt_children	1.00	0.03	0.00	0.02	-0.00	-0.03	-0.33	-0.24	-0.18	0.03	0.00	0.24	0.05
amt_income_total	0.03	1.00	0.36	0.44	0.36	0.17	-0.06	-0.14	-0.07	-0.02	0.00	0.14	-0.04
amt_credit	0.00	0.36	1.00	0.77	0.99	0.10	0.06	-0.06	-0.01	0.01	0.00	0.07	-0.02
amt_annuity	0.02	0.44	0.77	1.00	0.78	0.12	-0.01	-0.10	-0.04	-0.01	0.00	0.10	-0.03
amt_goods_price	-0.00	0.36	0.99	0.78	1.00	0.10	0.05	-0.06	-0.01	0.01	0.00	0.06	0.00
region_population_relative	-0.03	0.17	0.10	0.12	0.10	1.00	0.03	-0.00	0.05	0.00	0.00	0.00	-0.02
age	-0.33	-0.06	0.06	-0.01	0.05	0.03	1.00	0.62	0.33	0.27	0.00	-0.62	-0.17
days_employed	-0.24	-0.14	-0.06	-0.10	-0.06	-0.00	0.62	1.00	0.22	0.27	0.00	-1.00	-0.23
days_registration	-0.18	-0.07	-0.01	-0.04	-0.01	0.05	0.33	0.22	1.00	0.10	0.00	-0.21	-0.06
days_id_publish	0.03	-0.02	0.01	-0.01	0.01	0.00	0.27	0.27	0.10	1.00	0.00	-0.27	-0.05
flag_mobil	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	-0.00	0.00
flag_emp_phone	0.24	0.14	0.07	0.10	0.06	0.00	-0.62	-1.00	-0.21	-0.27	-0.00	1.00	0.23
flag_work_phone	0.05	-0.04	-0.02	-0.03	0.00	-0.02	-0.17	-0.23	-0.06	-0.05	0.00	0.23	1.00

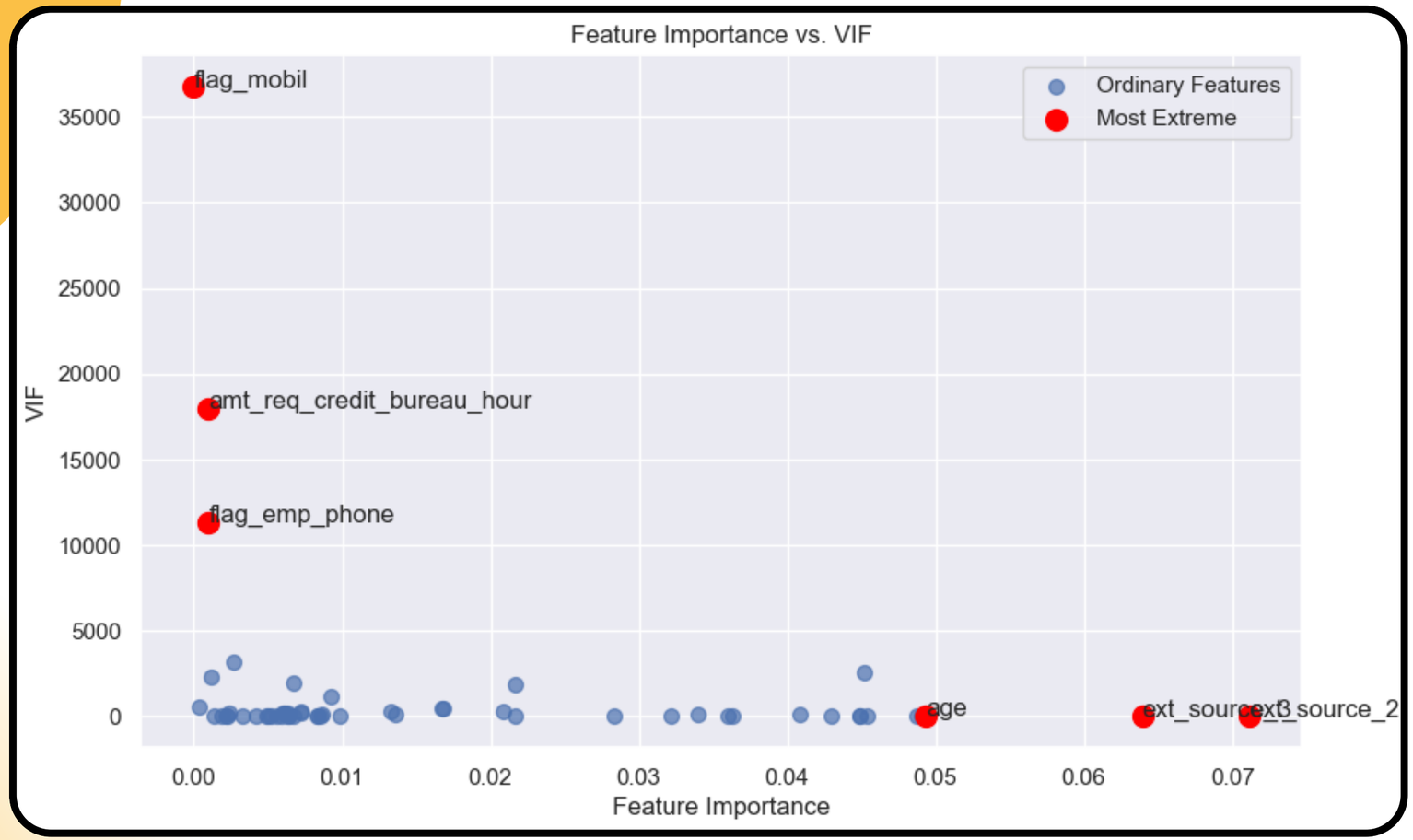
	Features	VIF
19	flag_mobil	36769.290717
47	amt_req_credit_bureau_hour	17975.672110
20	flag_emp_phone	11326.287111
49	amt_req_credit_bureau_week	3162.710802
16	days_employed	2557.738436

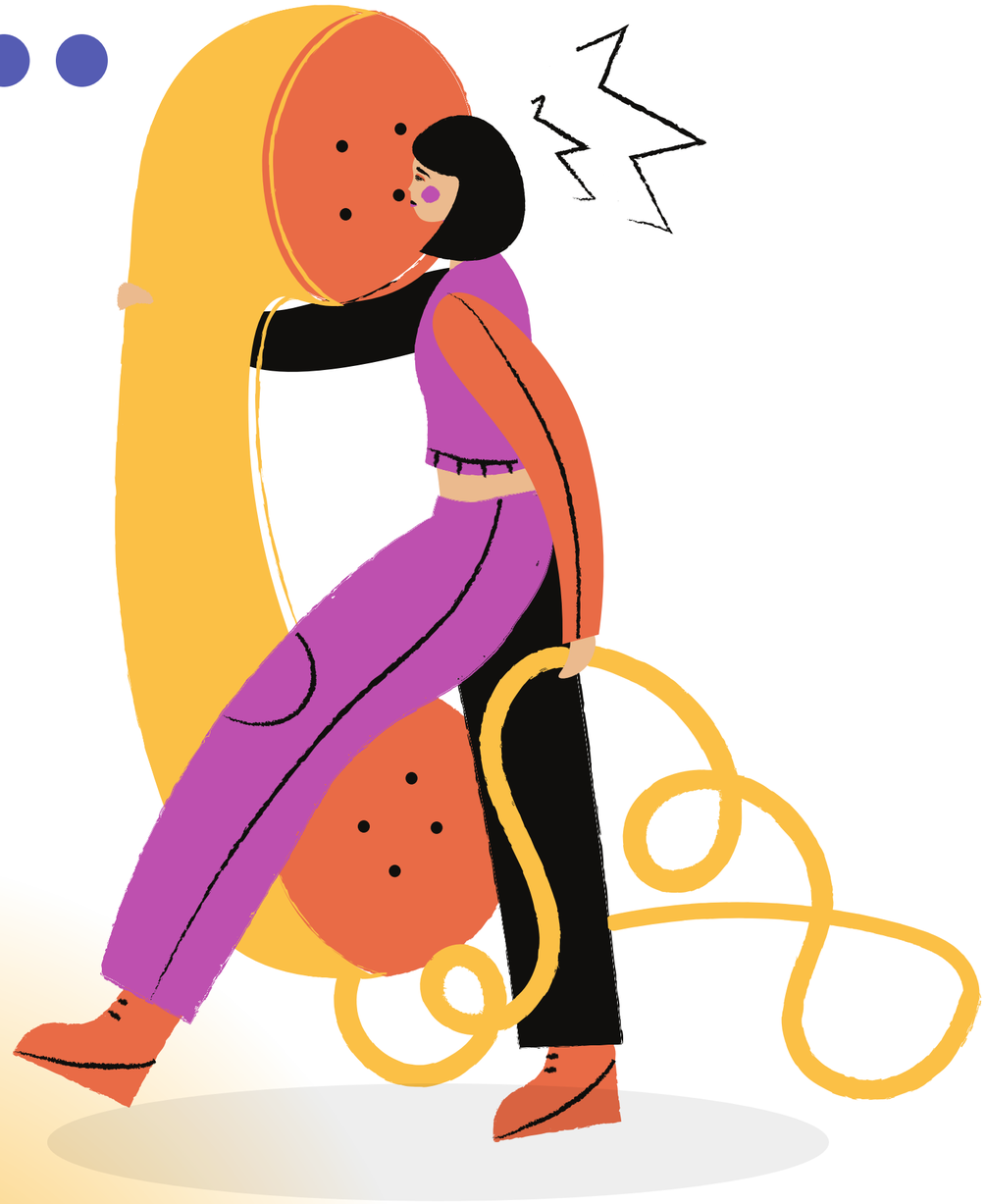
Getting most important features to the target



Featuer Selection - most importand & little VIF

Analysis finished with 38 features / 125



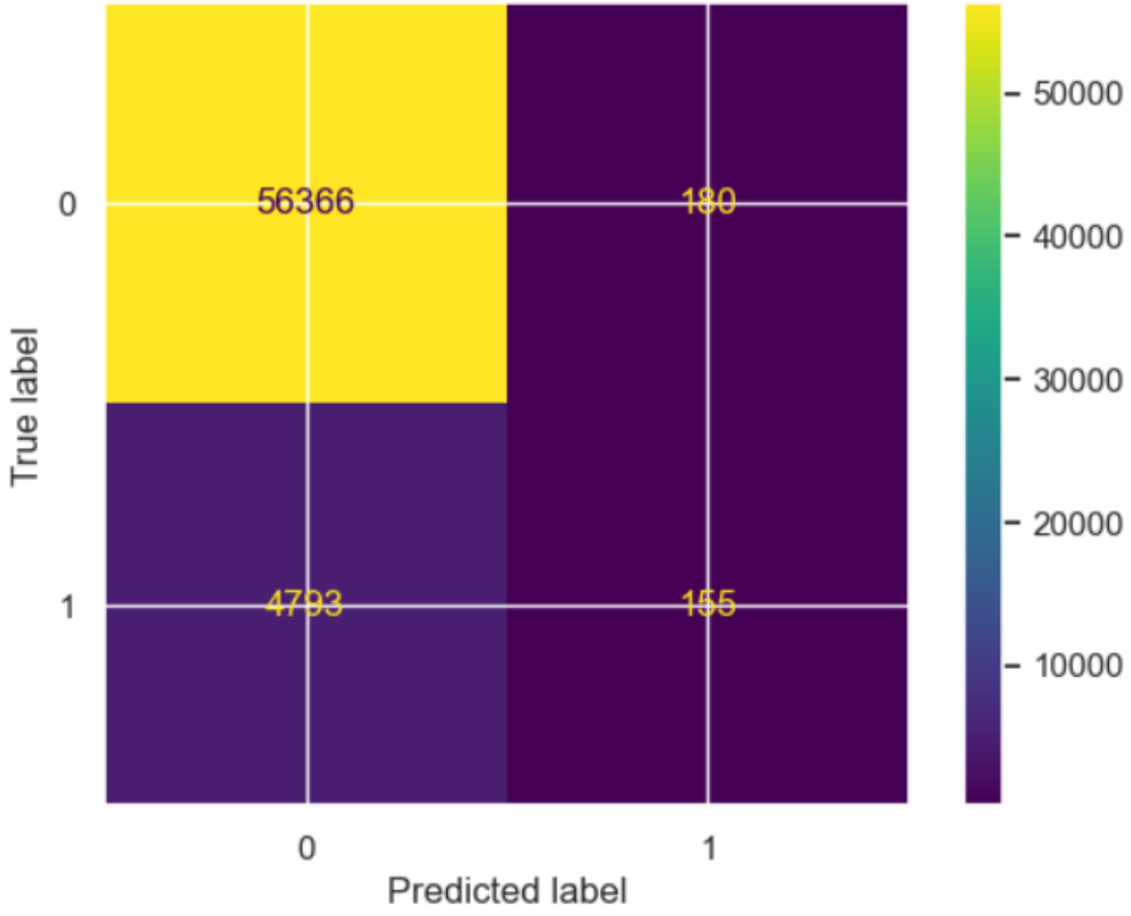


Phase 2

Modeling

Trying more than one model

Confiusion matrix of XGBoost - without tuning



Results of 4 models without tuning

```
Model: Logistic Regression
Accuracy: 0.9195368653852408
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
Model: Random Forest
Accuracy: 0.9196344358799232
Precision: 0.5833333333333334
Recall: 0.004244139046079224
F1 Score: 0.008426966292134831
Model: XGBoost
Accuracy: 0.9191303216573975
Precision: 0.4626865671641791
Recall: 0.03132578819725142
F1 Score: 0.05867878099564642
Model: Decision Tree
Accuracy: 0.851367613100465
Precision: 0.1410958904109589
Recall: 0.16653193209377526
F1 Score: 0.15276232851316277
```

Thanks for your attention