HOME CREDIT

Sijo VM
Pooja Shah
David Owen
Saurabh Bodas
Alisha Fernandes

Agenda

- The company
- Problem Statement
- Data Sources & Tools
- Hypothesis & Goals
- Exploratory Data Analysis
- Feature Engineering
- Model Building

1.
The company

$5.2 billion

Revenues!

132,000 employees

And a lot of customers!

1997

Founded in Czech Republic

Maps

10 Countries

Broaden financial inclusion to provide comfortable and safe borrowing experience

Focuses on the clients with little to no credit history

Transactional information, annual income, family status, housing type, etc. in order to predict their clients' repayment abilities

2.
Problem Statement

Hypothesis: Clients in careers with historically worse job security are most likely to default on their loan payments

Hypothesis: Clients with many previous credits are more likely to default on loans

Goal: Establish a trustworthy algorithm to validate/invalidate these claims and reveal other trends among the clientbase

Goal: Communicate results of said algorithm in a comprehensible manner

3.
Data Source and Tools

Data Source - **Kaggle**

Data Processing and modelling - **Pyspark and Python on the Databricks platform**

Data visualization - **Tableau , Draw.io**

Bureau

All clients previous credits provided by other financial institutions

Credit Card Balance

Monthly balance snapshots of previous credit cards owned by the applicant

Bureau Balance

Monthly balances of previous credits in the credit bureau

Previous Applications

All previous applications for loans by the client

Applications

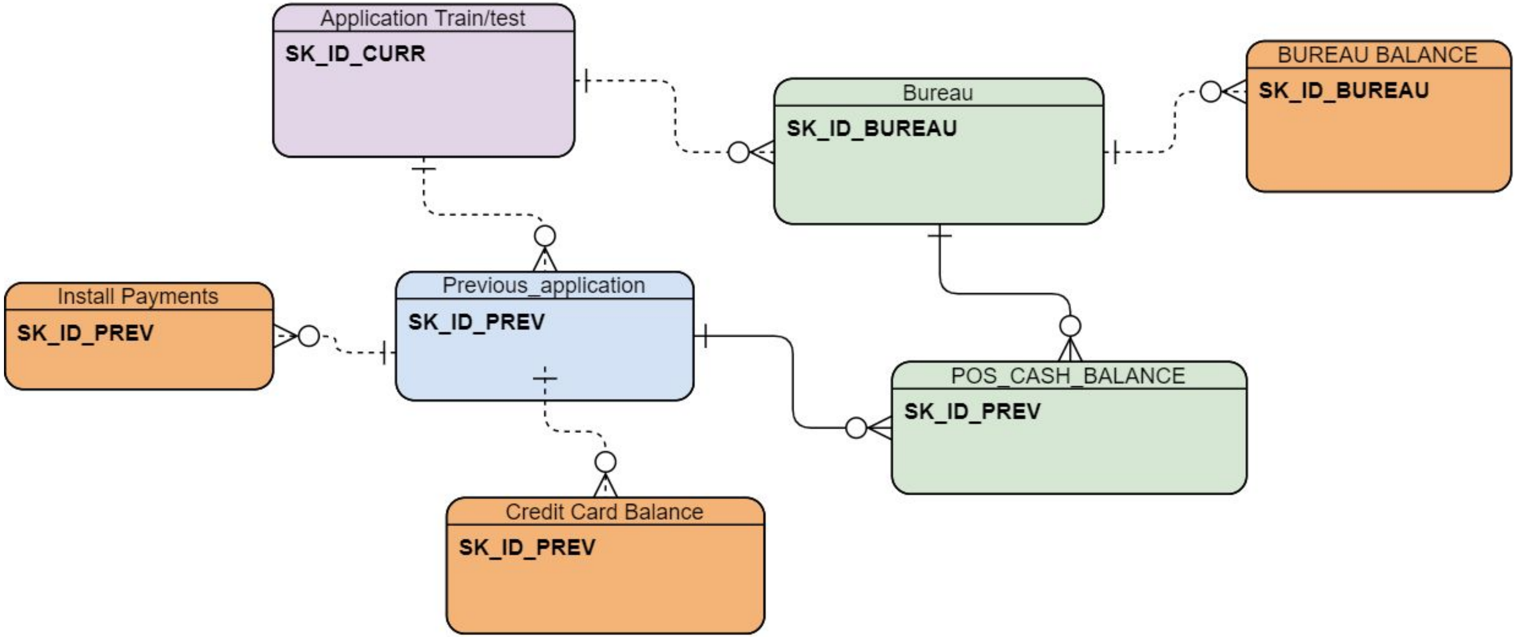
Main table depicting current loan applications for each applicant

POS_Cash

Monthly balance snapshots of previous POS and loans

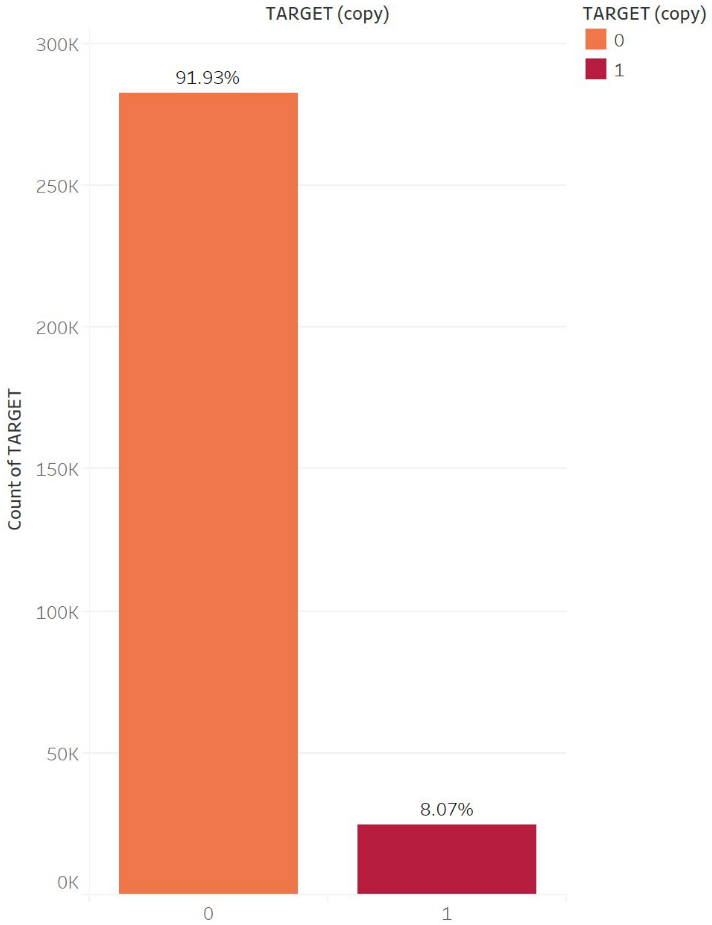
Installment Payments

Repayment history for previously disbursed credits


Application Train/test
SK_ID_CURR
Bureau
SK_ID_BUREAU
BUREAU BALANCE
SK_ID_BUREAU
Install Payments
SK_ID_PREV
Previous_application
SK_ID_PREV
POS_CASH_BALANCE
SK_ID_PREV
Credit Card Balance
SK_ID_PREV

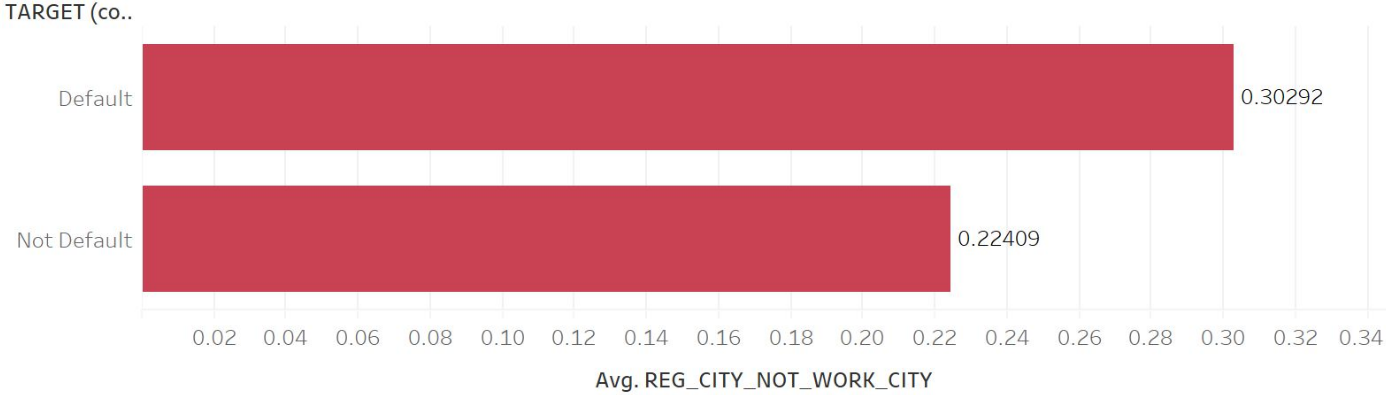
4.
Exploratory Data Analysis

Distribution of the Default Data



Count of TARGET for each TARGET (copy). Color shows details about TARGET (copy). The marks are labeled by % of Total Count of TARGET.

Proportion of discrepancy in residence and work location

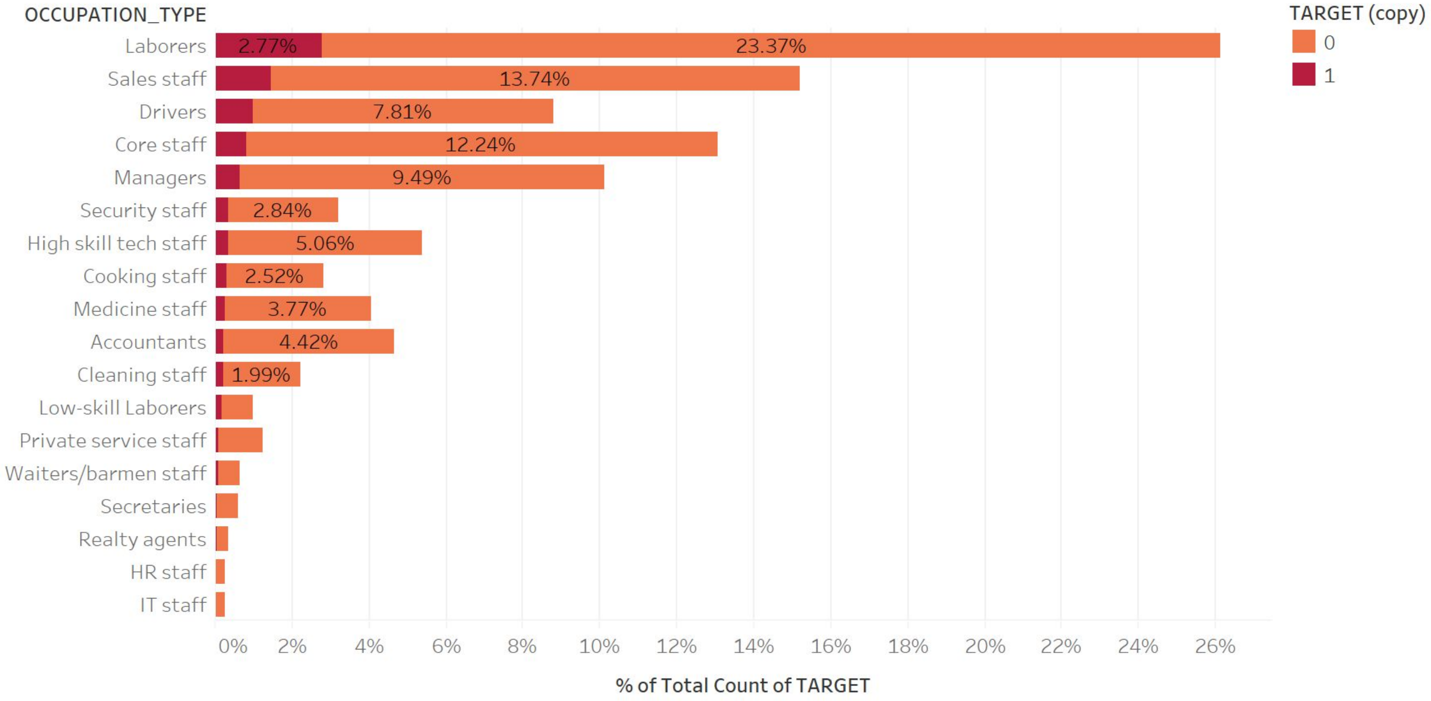


Average of REG_CITY_NOT_WORK_CITY for each TARGET (copy). The marks are labeled by average of REG_CITY_NOT_WORK_CITY. The view is filtered on TARGET (copy), which keeps Not Default and Default.

Hypothesis 1:

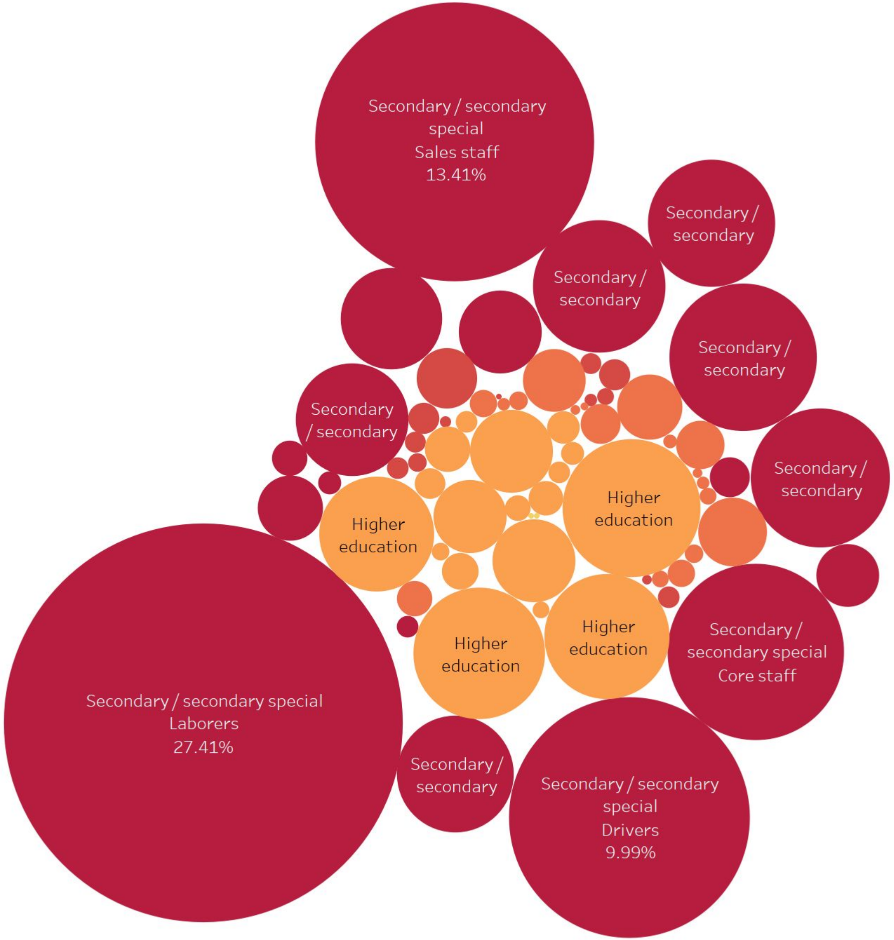
Are clients with historically worse job security more likely to default on their loan payments?

Number of people that default on a loan based on Occupation Types



% of Total Count of TARGET for each OCCUPATION_TYPE. Color shows details about TARGET (copy). The marks are labeled by % of Total Count of TARGET. The view is filtered on OCCUPATION_TYPE, which excludes Null.

Number of Defaulters by Education Level and Occupation Type

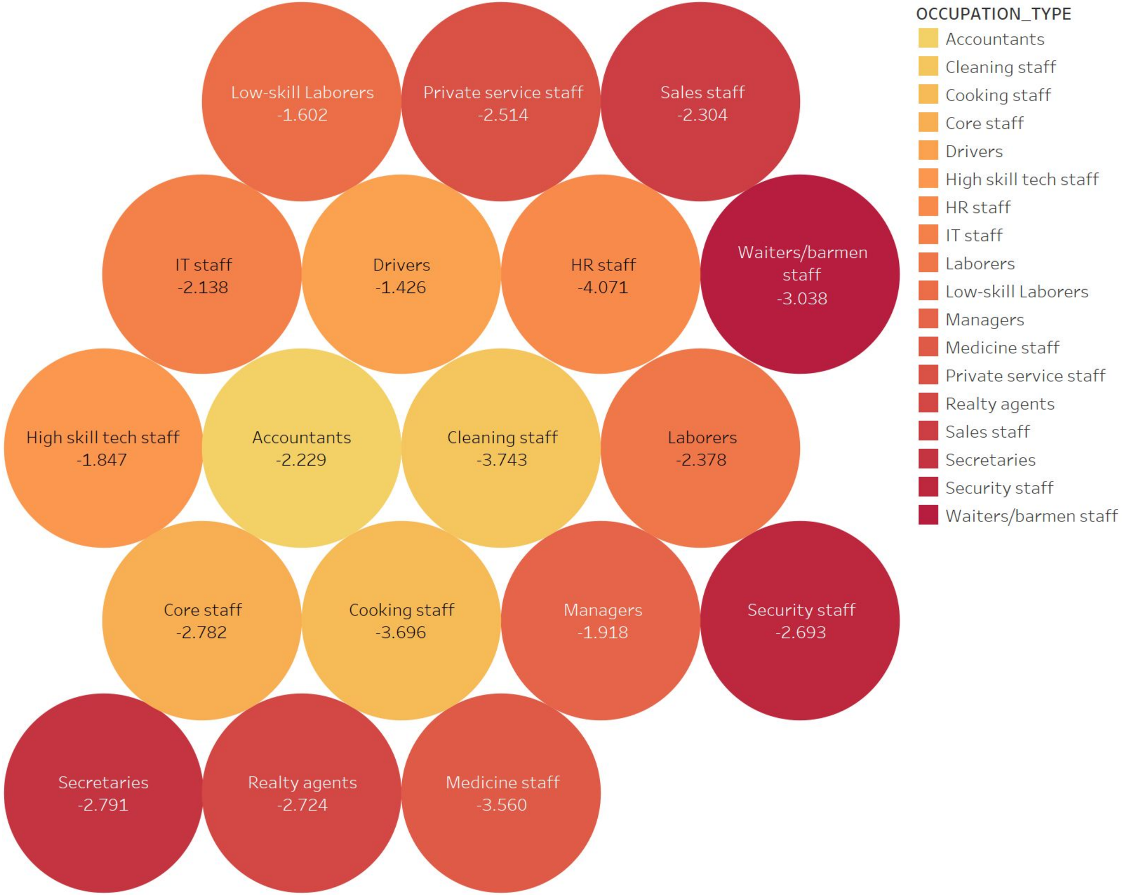


NAME_EDUCATION_TYPE

- 🟨 Academic degree
- 🟧 Higher education
- 🟧 Incomplete higher
- 🟥 Lower secondary
- 🟥 Secondary / secondary special

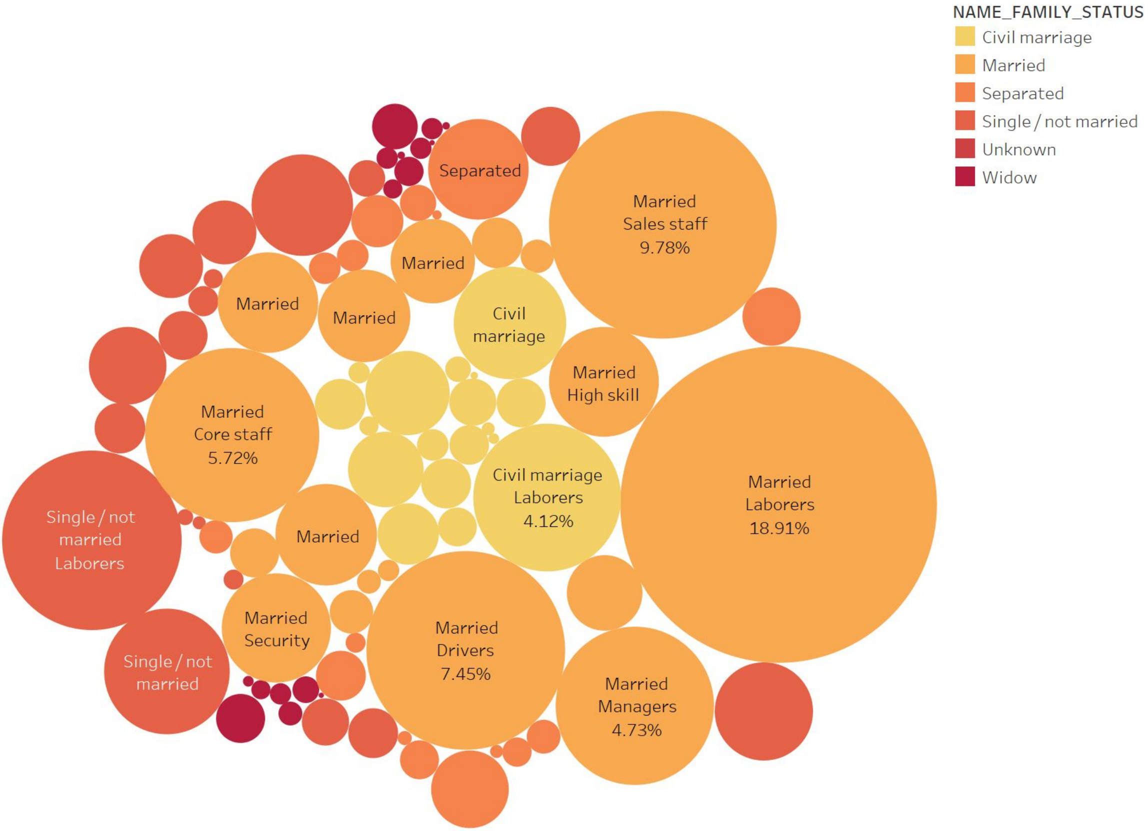
NAME_EDUCATION_TYPE, OCCUPATION_TYPE and % of Total TARGET. Color shows details about NAME_EDUCATION_TYPE. Size shows % of Total TARGET. The marks are labeled by NAME_EDUCATION_TYPE, OCCUPATION_TYPE and % of Total TARGET. The view is filtered on OCCUPATION_TYPE, which excludes Null.

Difference in Average Age between the Defaulters and Non Defaulters by Occupation Type



OCCUPATION_TYPE and Avg Age Diff. Color shows details about OCCUPATION_TYPE. Size shows Avg Age Diff. The marks are labeled by OCCUPATION_TYPE and Avg Age Diff. The view is filtered on OCCUPATION_TYPE, which excludes Null.

Number of Defaulters by Marital Status and Occupation Type

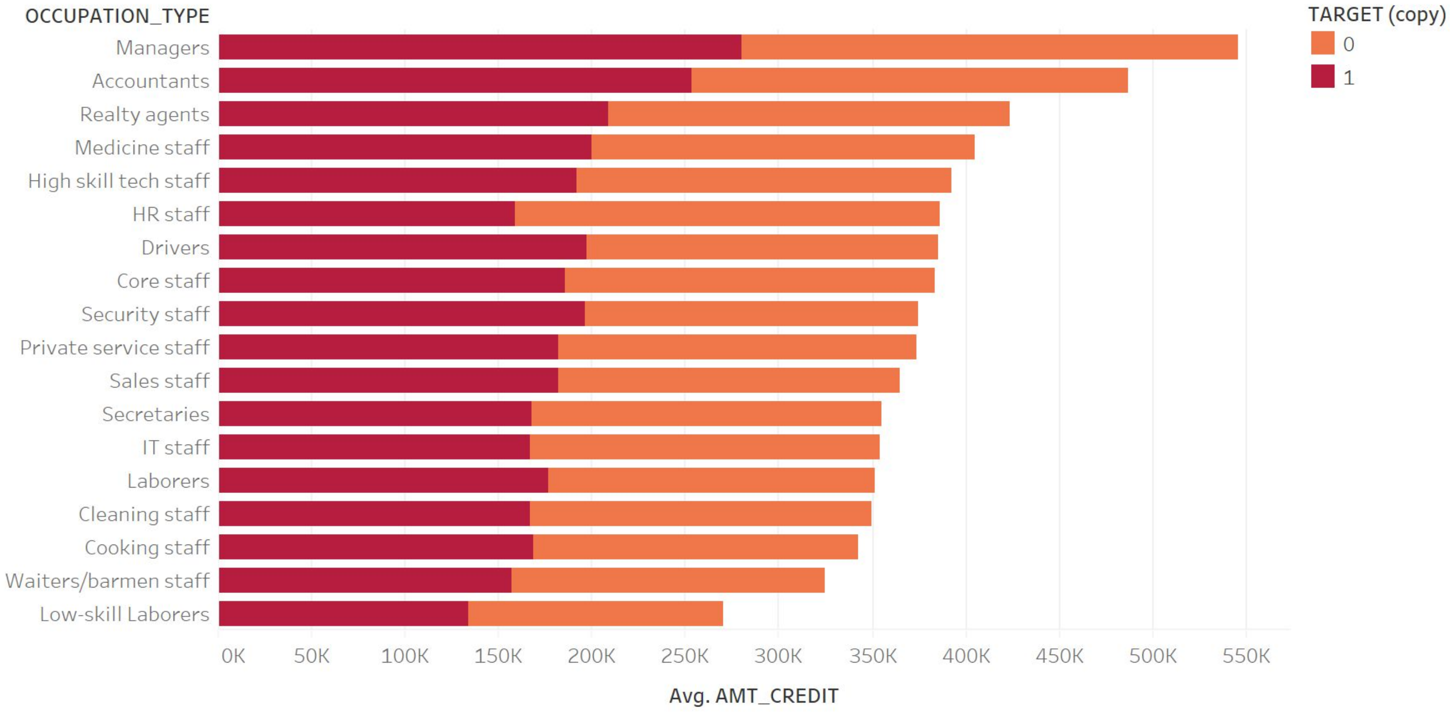


NAME_FAMILY_STATUS, OCCUPATION_TYPE and % of Total TARGET. Color shows details about NAME_FAMILY_STATUS. Size shows % of Total TARGET. The marks are labeled by NAME_FAMILY_STATUS, OCCUPATION_TYPE and % of Total TARGET. The view is filtered on OCCUPATION_TYPE, which excludes Null.

Hypothesis 2:

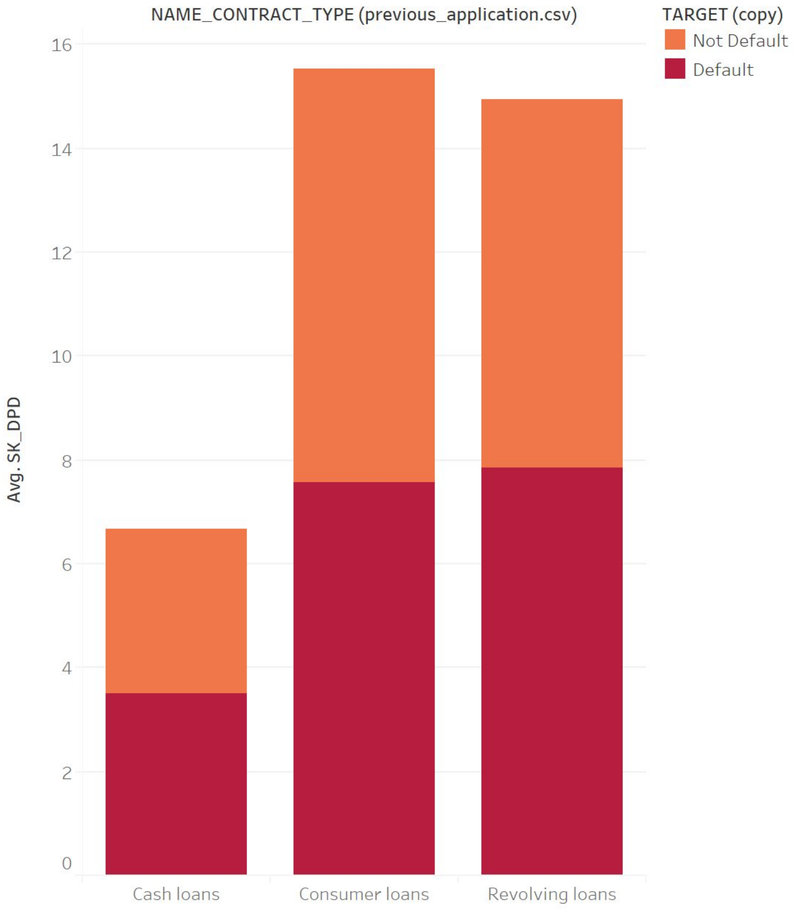
Are clients with many previous credits more likely to default on loans?

Average Amount of Previous Credit based on Occupation Type



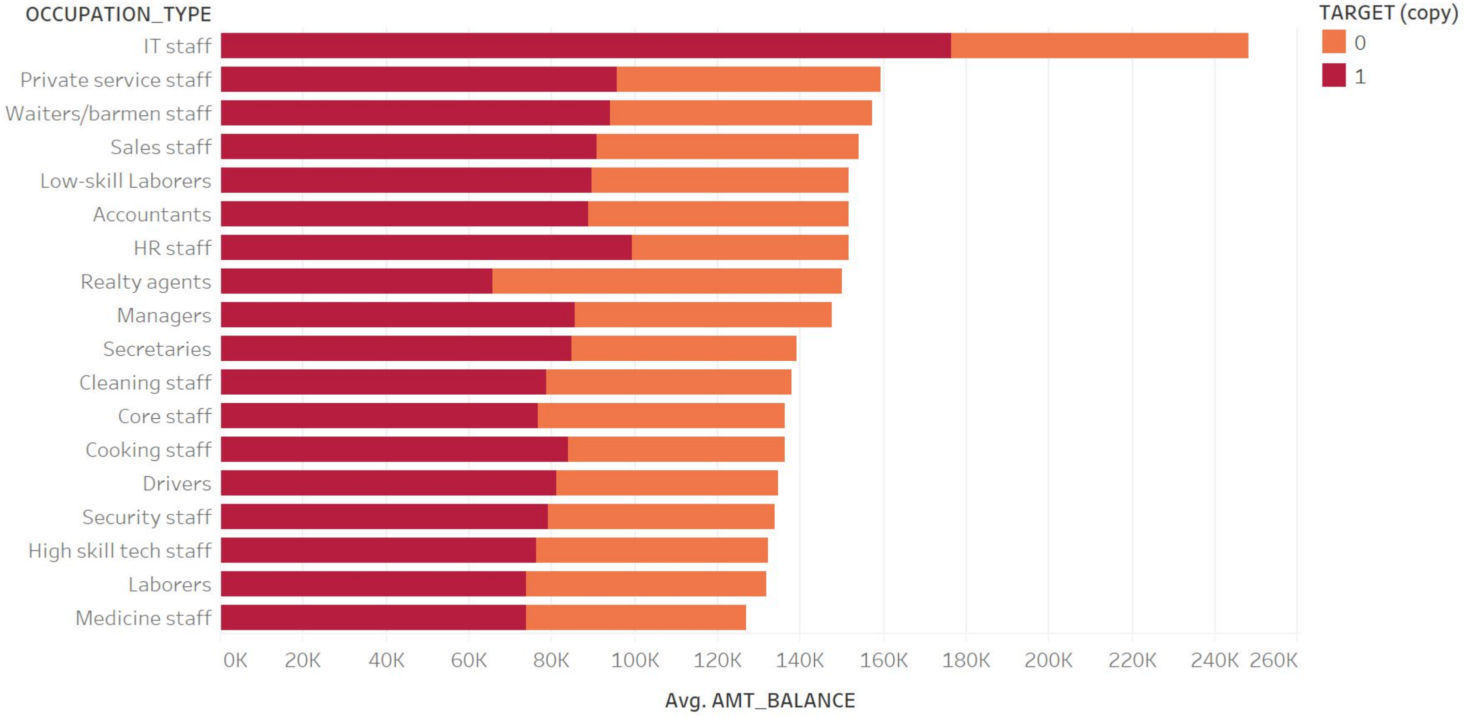
Average of AMT_CREDIT for each OCCUPATION_TYPE. Color shows details about TARGET (copy). The view is filtered on OCCUPATION_TYPE and TARGET (copy). The OCCUPATION_TYPE filter excludes Null. The TARGET (copy) filter keeps 0 and 1.

Days past due for different loan types



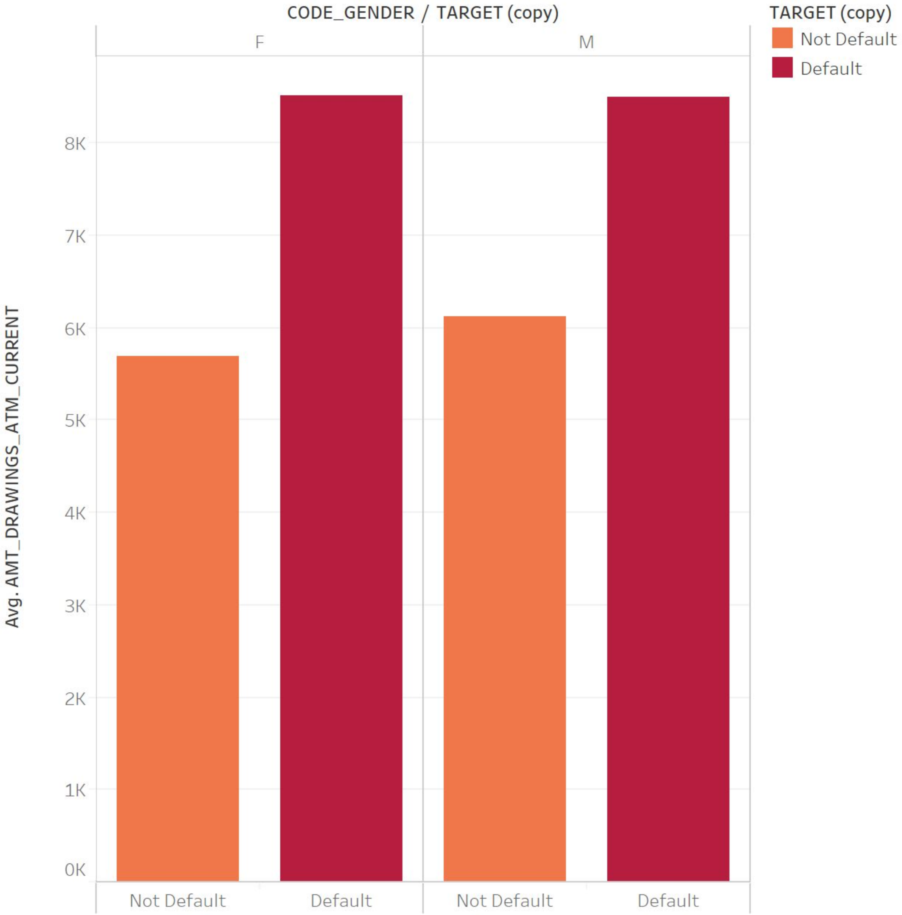
Average of SK_DPD for each NAME_CONTRACT_TYPE (previous_application.csv). Color shows details about TARGET (copy). The view is filtered on NAME_CONTRACT_TYPE (previous_application.csv), which keeps Cash loans, Consumer loans and Revolving loans.

Average Credit Card Balance by Occupation



Average of AMT_BALANCE for each OCCUPATION_TYPE. Color shows details about TARGET (copy). The view is filtered on OCCUPATION_TYPE, which excludes Null.

Average Amount Withdrawn from ATM by Gender and Default Status



Average of AMT_DRAWINGS_ATM_CURRENT for each TARGET (copy) broken down by CODE_GENDER. Color shows details about TARGET (copy).

5.
Feature Engineering

- ‘Days’ variable made positive and in terms of years
- ‘Family size’ converted to binned categorical variable
- Technical information added to improve model performance
    - Credit term
    - % of days employed

- Missing values imputed with the median
- Data split into 80% train, 20% test
- SMOTE package used for resampling to deal with imbalanced classes

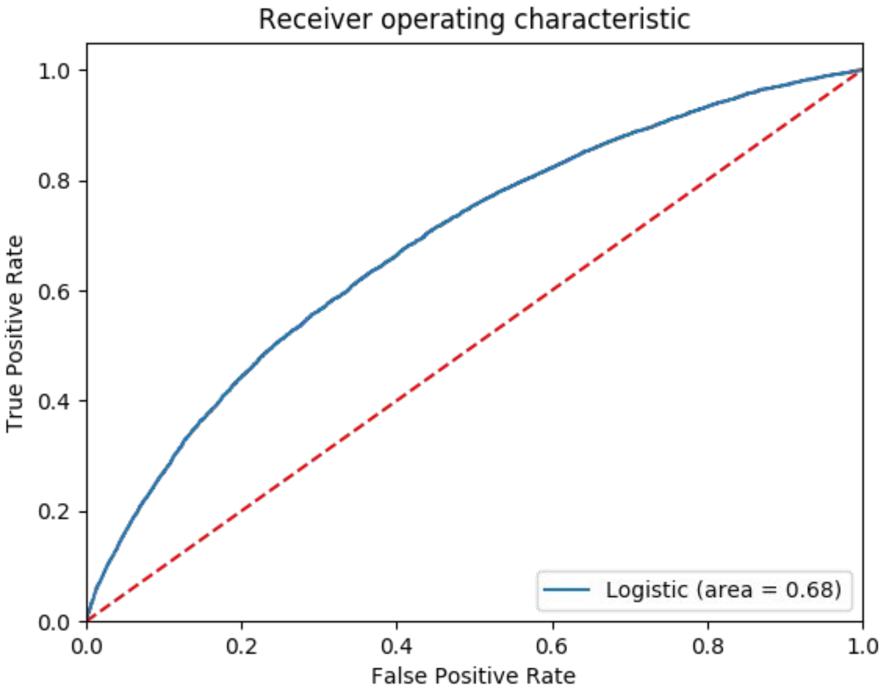
6.
Model Building

Insights and Recommendations

Models Built: A Comparison

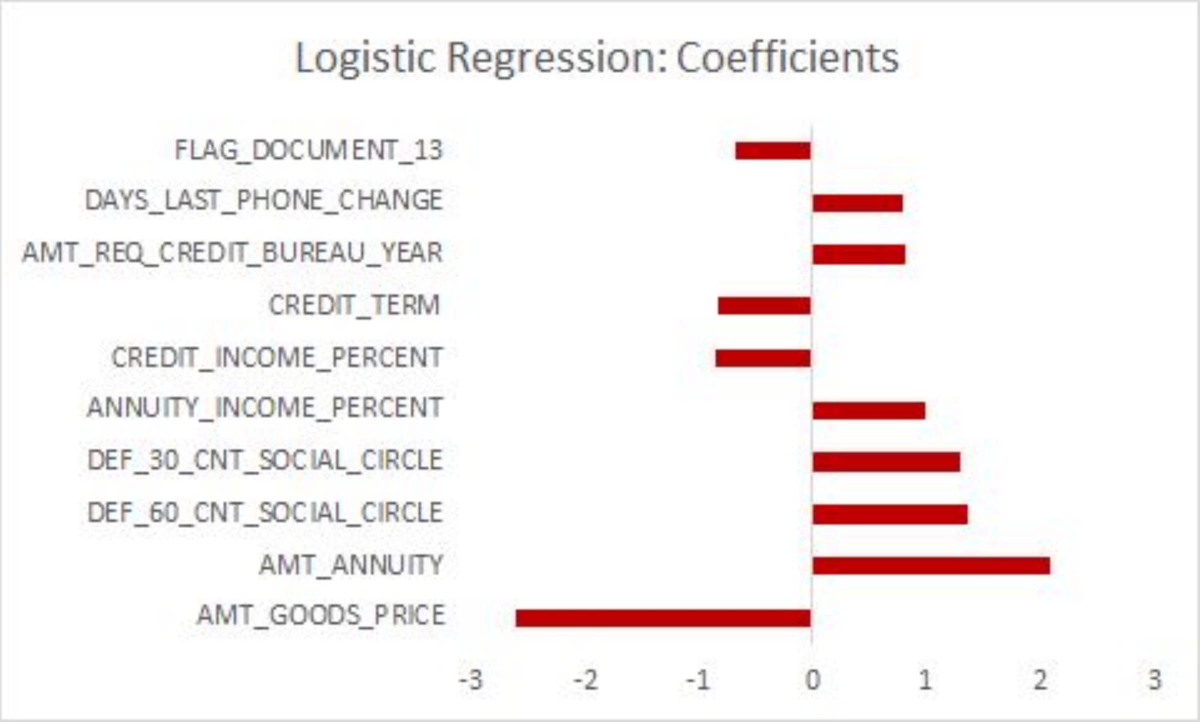
01	Random Forests	● AUROC Score: 0.65
02	Random Forests: Resampling	● AUROC Score: 0.63
03	Logistic Regression	● AUROC Score: 0.68
04	Cat Boosting	● AUROC Score: 0.69

Evaluation of Logistic Regression

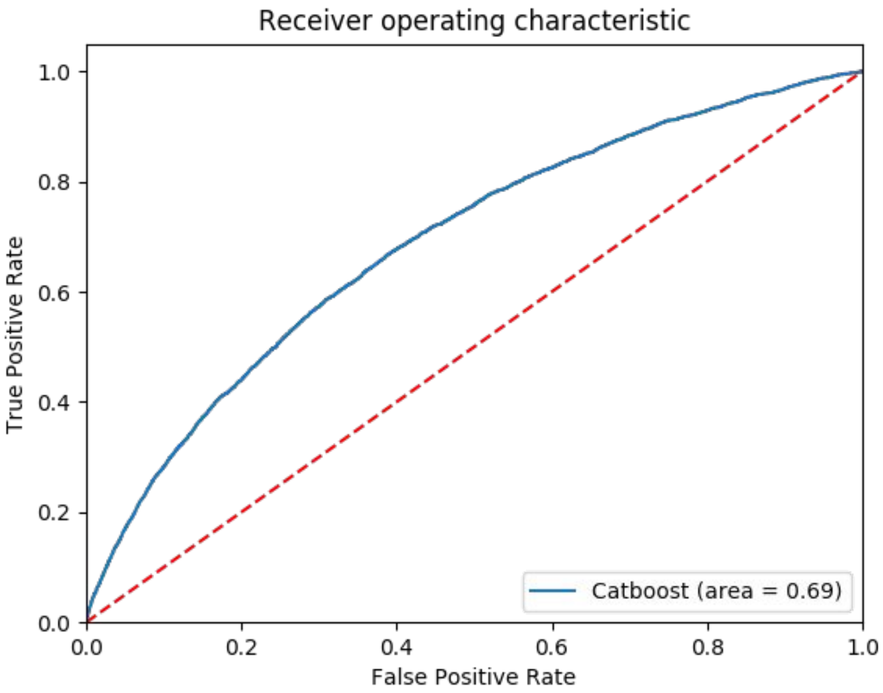


	0-Predict	1-Predict
0-Actual	55880	768
1-Actual	4574	281

Reduce False Negatives!
People who default but the model predicts the client won't!

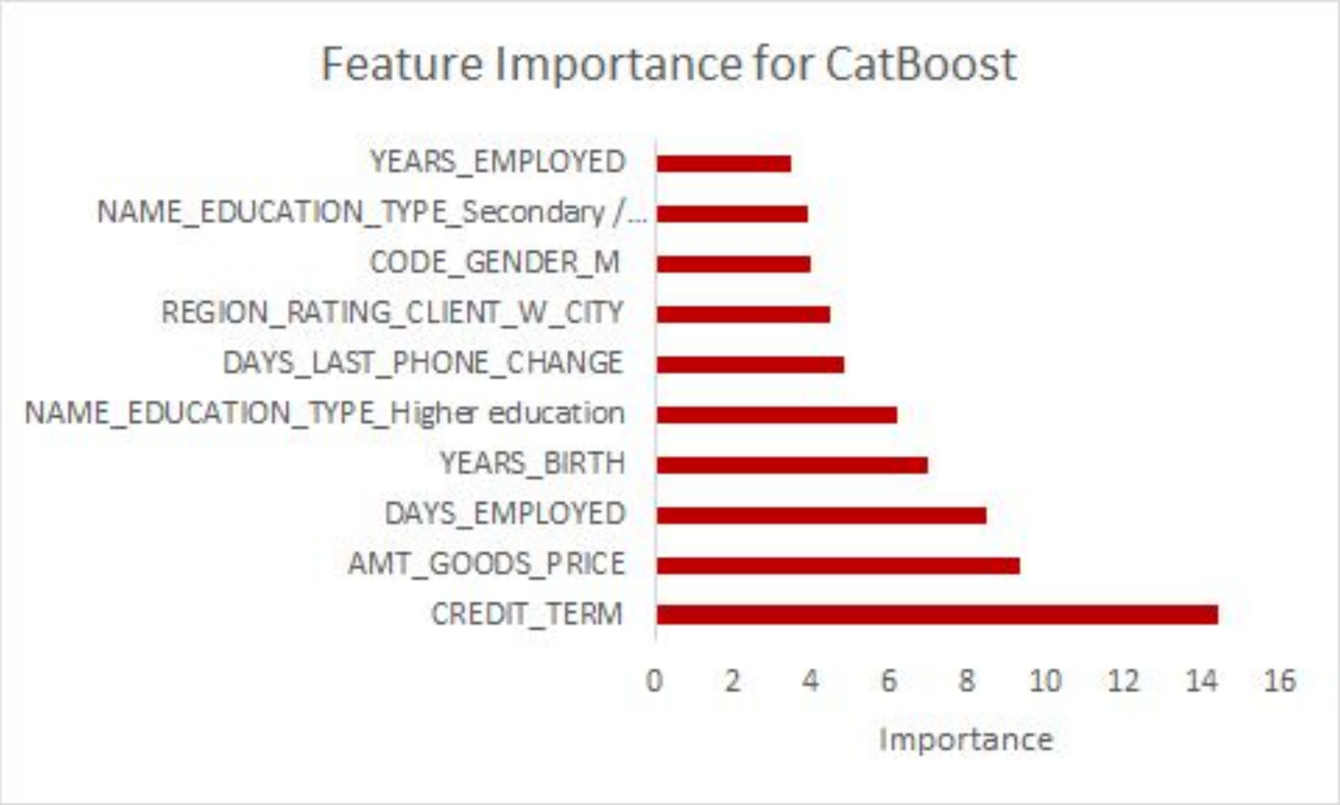

Logistic Regression: Coefficients
FLAG_DOCUMENT_13
DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_YEAR
CREDIT_TERM
CREDIT_INCOME_PERCENT
ANNUITY_INCOME_PERCENT
DEF_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
AMT_ANNUITY
AMT_GOODS_PRICE
-3
-2
-1
0
1
2
3

Evaluation of CatBoost



	0-Predict	1-Predict
0-Actual	54233	2415
1-Actual	4136	719

Reduce False Negatives!
People who default but the model predicts the client won't!


Feature Importance for CatBoost
YEARS_EMPLOYED
NAME_EDUCATION_TYPE_Secondary /...
CODE_GENDER_M
REGION_RATING_CLIENT_W_CITY
DAYS_LAST_PHONE_CHANGE
NAME_EDUCATION_TYPE_Higher education
YEARS_BIRTH
DAYS_EMPLOYED
AMT_GOODS_PRICE
CREDIT_TERM
0
2
4
6
8
10
12
14
16
Importance

Insights and recommendations

- Be more cautious when you are lending to labourers and not highly educated clients
- The recent withdrawals from the ATM has an impact on the default risk
- Defaulting is not instant - If the credit balance increases over time, then the client is highly likely to default
- Region rating from the model as well as the discrepancy in the work and residence location
- Amts_goods_price the proposed loan purpose higher - more likely to default
- If a person has recently changed the phone number, then the propensity to default increases

Questions?