

HOME CREDIT DEFAULT RISK



A Machine Learning approach
Made by: Heba Aladdin

CONTENT

Here's what you'll find in this presentation:

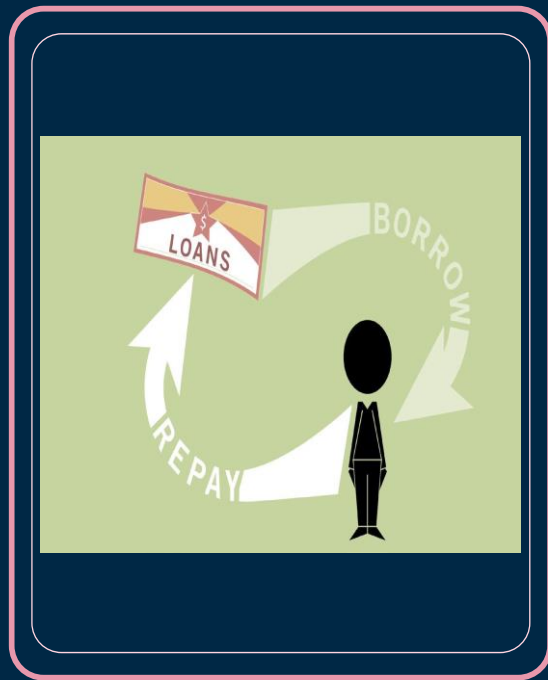
1. Business problem.
2. Value proposition.
3. Approach.
4. Key insights and findings.
5. Training ML models.
6. Results and conclusion

BUSINESS PROBLEM

HOME
CREDIT

Default risk is the chance that companies or individuals will be unable to make the required payments on their debt obligations. In other words, credit default risk is the probability that if you lend money, there is a chance that they won't be able to give the money back on time.

Lenders and investors are exposed to default risk in virtually all forms of credit extensions. To mitigate the impact of default risk, lenders often impose charges that correspond to the debtor's level of default risk. A higher level of risk leads to a higher required return



VALUE PROPOSITION

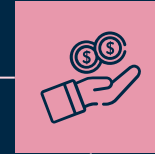
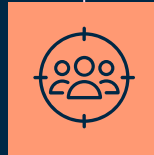
FASTER PROCESS

Shorter lending
process



INCREASE BORROWER BASE

Trust more applicants
and unbanked
population



DECREASE EXPOSURE

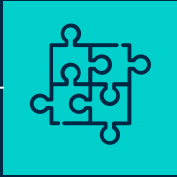
Decrease exposure
that lenders are
exposed to

INCREASE PROFIT

Lend more gain
more



APPROACH



01

DATA ANALYSIS & EDA

Exploaring the dataset
to find patterns and
business insights



02

FEATURE ENGINEERING & FEATURE SELECTION



03

MODELING

Training machine
learning models

UNDERSTANDING THE PROBLEM

REPAY

Applicants with no payment delays on their first few installments.

TARGET = 0



DEFAULT

Applicants with payment difficulties on their first few installments.

TARGET = 1

92%



8%

300,000+

KEY INSIGHTS & FINDINGS

BORROWER GENDER



66%

Female

34%

Male

DEFAULT

Repay on time



93%

90%

Default



7%

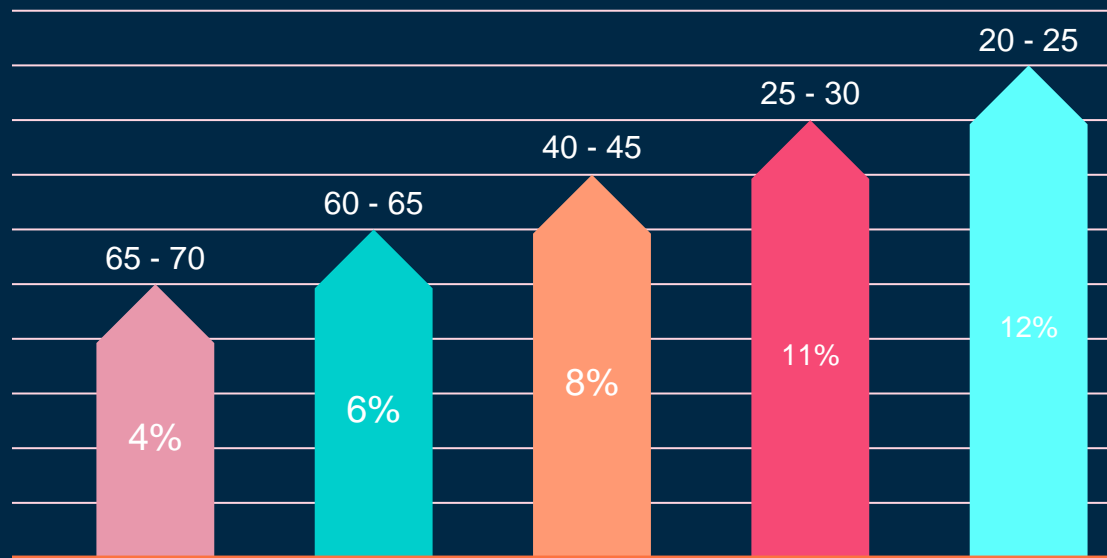
10%



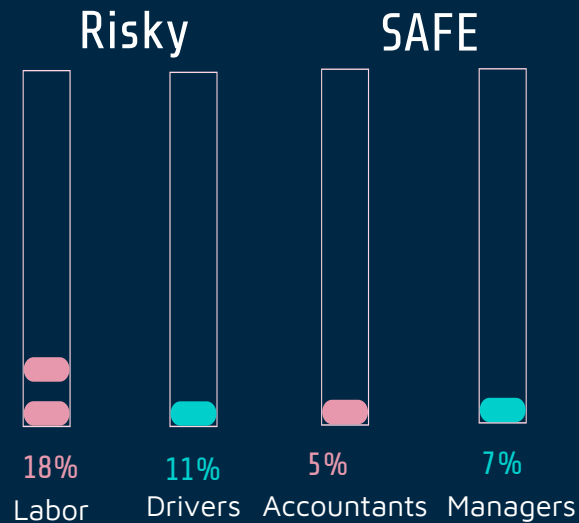
300,000+

Applicants

KEY INSIGHTS & FINDINGS



BORROWER AGE GROUP VS FAILURE TO REPAY



BORROWER JOB VS DEFAULT

For more insights check the [EDA notebook](#)

OUR PIPELINE

Handling missing feature,
anomalies, encoding

DATA CLEANING

FEATURE ENGINEERING

Adding two sets of features

- Domain Knowledge
- Aggregated features

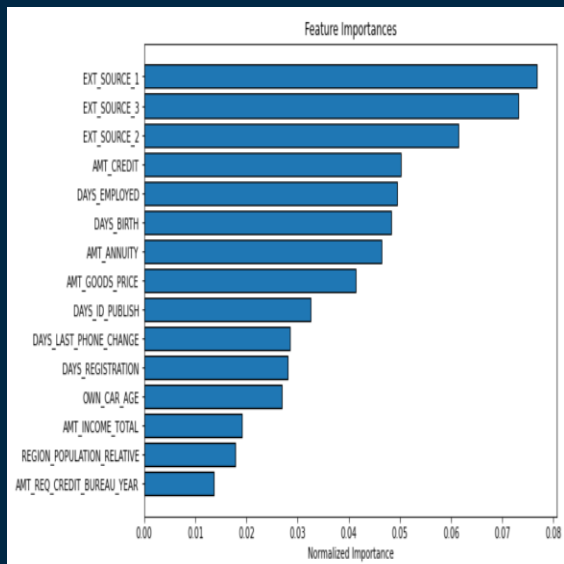
- Remove missing columns
- Correlation chart
- Collinear features removal

FEATURE SELECTION

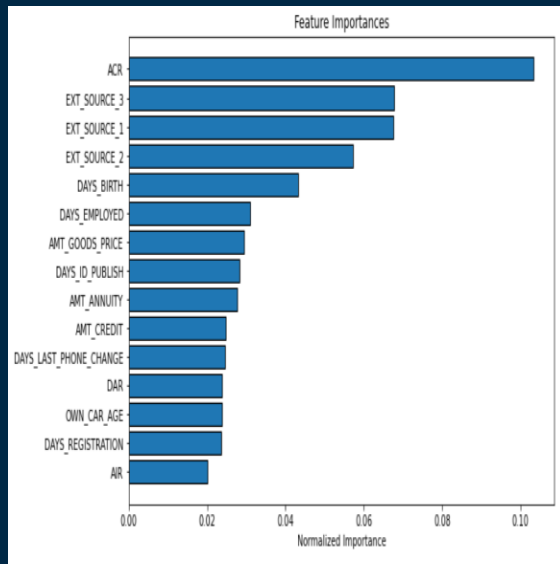
MODELING

- Baseline model: logistic regression
- Random forest
- LightGBM 10 Kfold

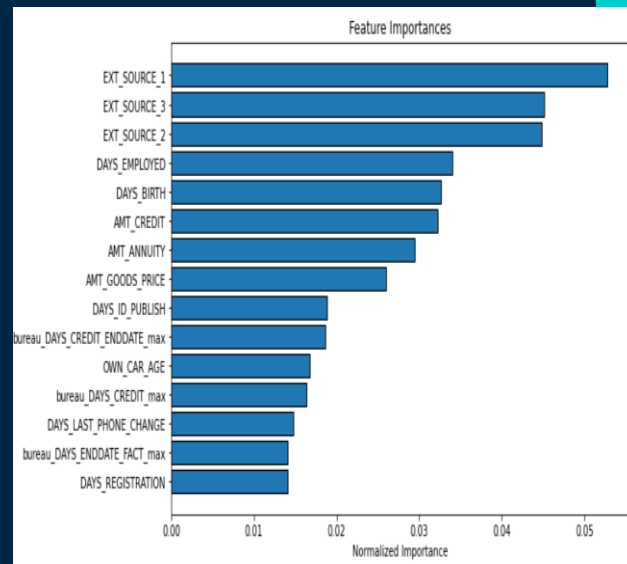
MODELING



LightGBM on raw Data



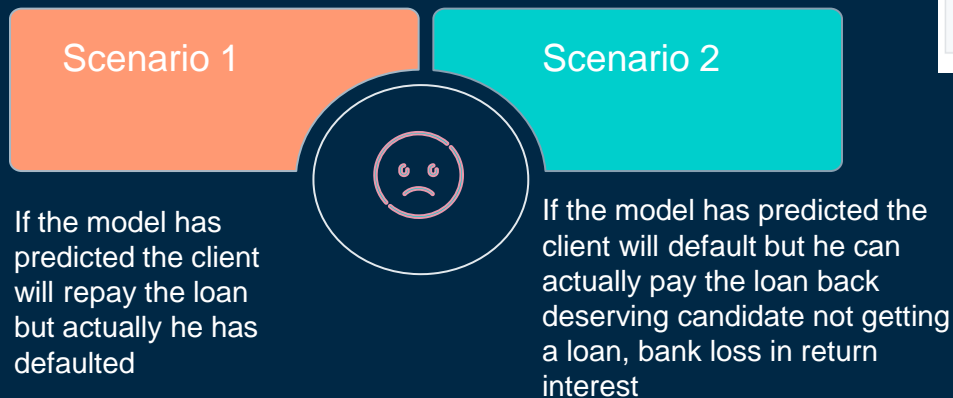
LightGBM with Domain
knowledge features



LightGBM with Aggregated
features

RESULTS

- The metric we choose to evaluate our models is Receiver Operating Characteristic Area Under the Curve (ROC AUC, also sometimes called AUROC) due to the high unbalanced labels.
- Home credit will face losses if the model prediction is wrong in two scenarios:



Experiment	Train AUC	Validation AUC
Raw dataset	0.806430	0.758923
Domain dataset	0.815190	0.766038
Aggregated dataset v1	0.825520	0.766415
Aggregated dataset v2	0.815504	0.763560

LightGBM 10 k-fold

THANKS.

