Luddy School of Informatics, Computing and Engineering

# Home Credit Default Risk


Team Members:

1. Akash Gangadharan (akganga@iu.edu
2. Laya Harwin (lharwin@iu.edu)
3. Dhairya Shah (shahds@iu.edu)
4. Shobhit Sinha(shosinha@iu.edu)

# Agenda

01

## Project Objective

Understanding the business objective and data

02

## EDA and Data Preparation

Analyzing the historical data and generating insights out of it

03

## Modelling Pipelines

Exploring and preprocessing the data

04

## Comparing results and Deriving Insights

Comparing results from different model and evaluating

05

## Future Scope

Future work to be done

SECTION 1

# Project Objective

# What are we aiming to solve?

## Who?

Home Credit Group 

## What?

The goal of this Project is to predict whether a loan applicant is likely to default on a loan, given information about the applicant and their previous loan applications.

## Why?

To make informed decision that can lead to less default on the loan and thereby reducing the loss for the organization

## How?

Creating machine learning models using different demographic features and other usage data of the customer to predict whether the customer will default or not.

SECTION 2

# EDA and Data Preparation

# EDA and Data Preparation

### STEP 1

**Understand the Data and the requirement and align it with the Project Goals.**

### STEP 2

**Demographic analysis.**
Age, Gender, Occupations

### STEP 3

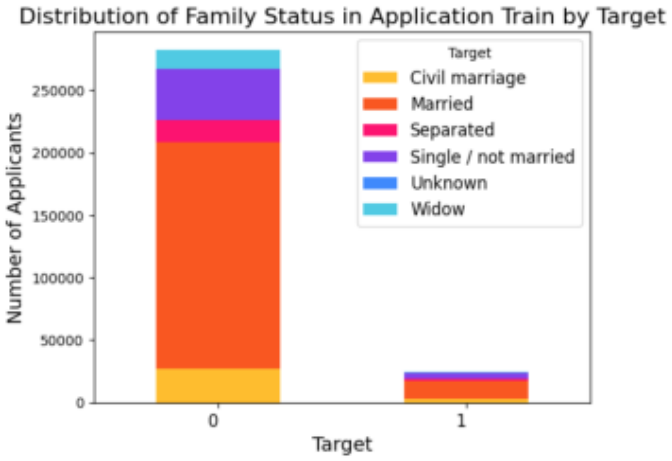
**Handling Missing values**
Drop columns with many null values and fill them based on the questions

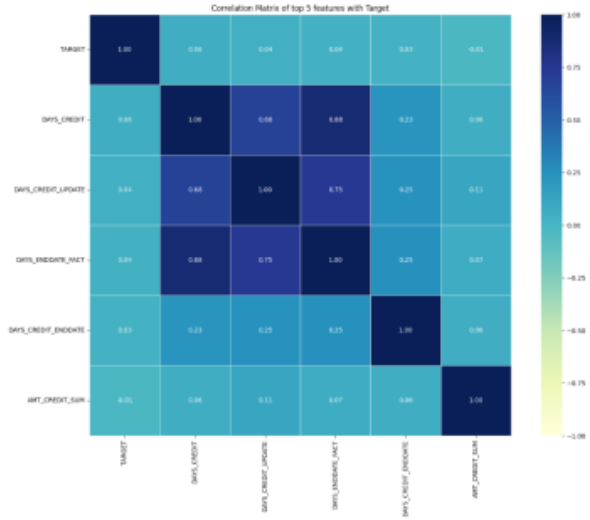
### STEP 5

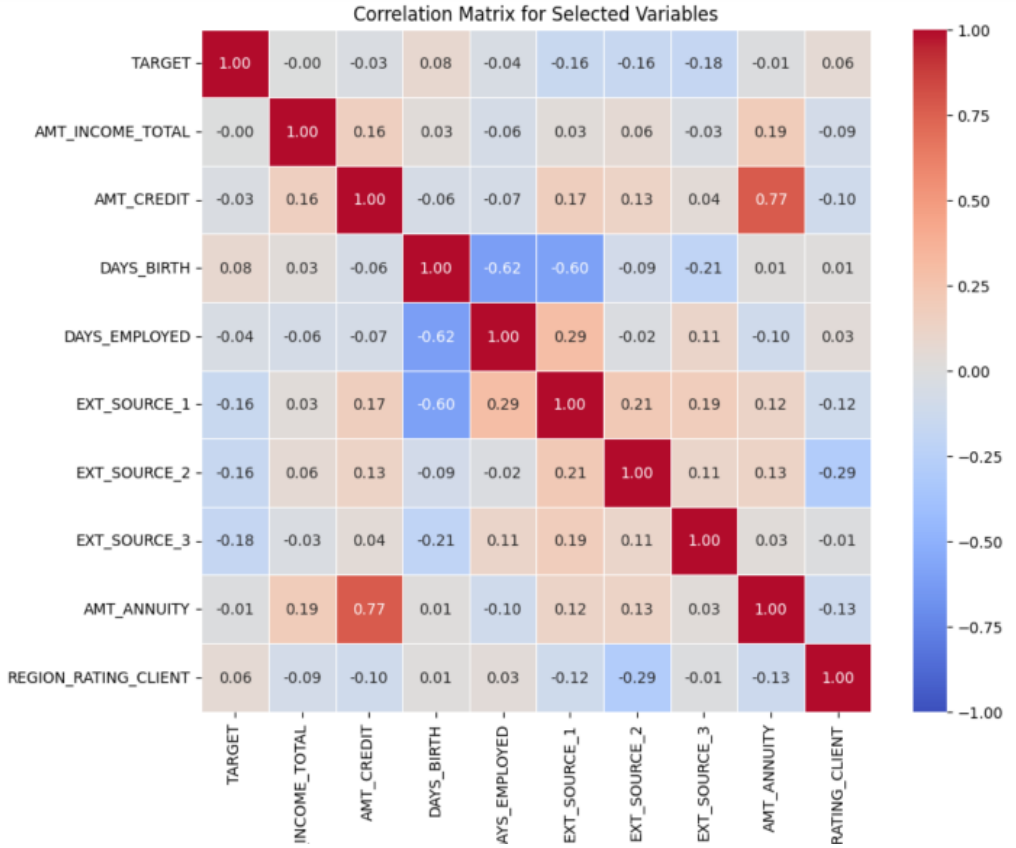
**Summary Statistics**
**Univariate analysis**
**Bivariate analysis**

### STEP 5

**Encoding categorical data**
Binary Encoding, Label Encoding, One Hot Encoding

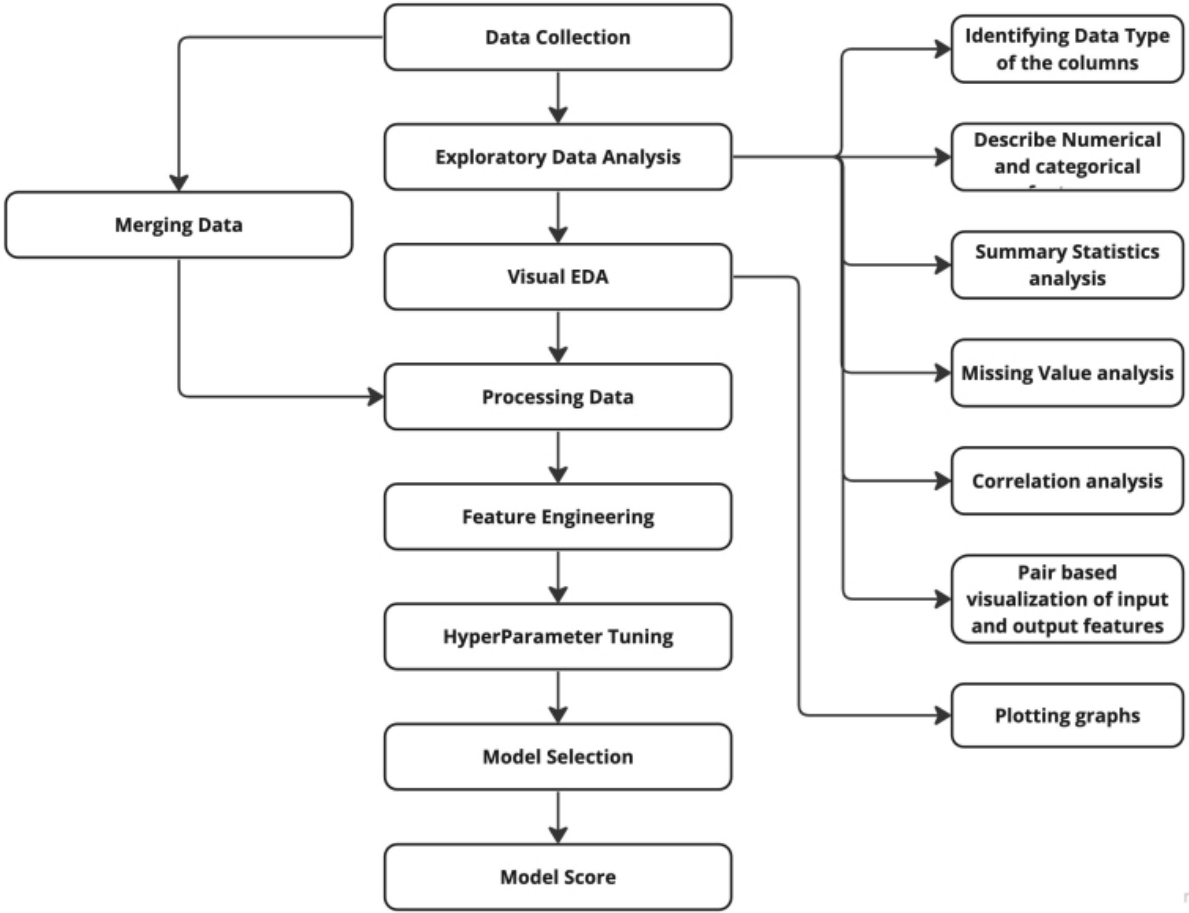

Distribution of Family Status in Application Train by Target
Target
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
Number of Applicants
250000
200000
150000
100000
50000
0
0
1
Target


Correlation Matrix of top 5 features with Target


Correlation Matrix for Selected Variables
TARGET
AMT_INCOME_TOTAL
AMT_CREDIT
DAYS_BIRTH
DAYS_EMPLOYED
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
AMT_ANNUITY
REGION_RATING_CLIENT
1.00 -0.00 -0.03 0.08 -0.04 -0.16 -0.16 -0.18 -0.01 0.06
-0.00 1.00 0.16 0.03 -0.06 0.03 0.06 -0.03 0.19 -0.09
-0.03 0.16 1.00 -0.06 -0.07 0.17 0.13 0.04 0.77 -0.10
0.08 0.03 -0.06 1.00 -0.62 -0.60 -0.09 -0.21 0.01 0.01
-0.04 -0.06 -0.07 -0.62 1.00 0.29 -0.02 0.11 -0.10 0.03
-0.16 0.03 0.17 -0.60 0.29 1.00 0.21 0.19 0.12 -0.12
-0.16 0.06 0.13 -0.09 -0.02 0.21 1.00 0.11 0.13 -0.29
-0.18 -0.03 0.04 -0.21 0.11 0.19 0.11 1.00 0.03 -0.01
-0.01 0.19 0.77 0.01 -0.10 0.12 0.13 0.03 1.00 -0.13
0.06 -0.09 -0.10 0.01 0.03 -0.12 -0.29 -0.01 -0.13 1.00
TARGET
INCOME_TOTAL
AMT_CREDIT
DAYS_BIRTH
AYS_EMPLOYED
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
AMT_ANNUITY
RATING_CLIENT
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

SECTION 3

# Modelling Pipeline


Data Collection
Exploratory Data Analysis
Merging Data
Visual EDA
Processing Data
Feature Engineering
HyperParameter Tuning
Model Selection
Model Score
Identifying Data Type of the columns
Describe Numerical and categorical
Summary Statistics analysis
Missing Value analysis
Correlation analysis
Pair based visualization of input and output features
Plotting graphs
miro

SECTION 4

# Comparing results

| | Pipeline | Parameters | TrainAcc | ValidAcc | TestAcc | Train Time(s) | Test Time(s) |
|---|---|---|---|---|---|---|---|
| 1 | Baseline Pipeline(steps=[('dt', DecisionTreeClassifier())]) with 320 inputs | {'dt__max_depth': [5, 10]} | 93.83% | 93.75% | 93.46% | 0.196276 | 0.934601 |
| 2 | Baseline Pipeline(steps=[('lr', LogisticRegression())]) with 320 inputs | {'lr__C': [0.01], 'lr__penalty': ['l1', 'l2']} | 91.90% | 92.25% | 91.93% | 0.078855 | 0.919290 |
| 3 | Baseline Pipeline(steps=[('ada',\n AdaBoostClassifier(base_estimator=DecisionTreeClassifier()))]) with 320 inputs | {'ada__n_estimators': [50], 'ada__learning_rate': [0.01, 0.1], 'ada__base_estimator__max_depth': [1, 5]} | 92.60% | 92.69% | 92.46% | 10.076050 | 0.924596 |

```
ROC AUC score
Logistic Regression : 0.7470648772667208
Decision Tree : 0.8074479725505205
Random Forest : 0.9999999934281516
ADA Boost : 0.8129134572402118
```


Overview Data Code Discussion Leaderboard Rules Team Submissions **Late Submission** ···

## Submissions

You selected 0 of 2 submissions to be evaluated for your final leaderboard score. Since you selected less than 2 submission, Kaggle auto-selected up to 2 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

0/2

■ Submissions evaluated for final score

All Successful Selected Errors Recent

| Submission and Description | Private Score | Public Score | Selected |
|---|---|---|---|
| **submission.csv** Complete (after deadline) · now · Group15 AML Logistic Regression | 0.73315 | 0.73762 | ☐ |
| **submission.csv** Complete (after deadline) · 7h ago · group 15 AML | 0.66457 | 0.65961 | ☐ |
| **submission.csv** Complete (after deadline) · 7h ago · Submission from Group 15 AML | 0.65139 | 0.66622 | ☐ |

SECTION 5

# Future Scope

To conclude, in this phase we cleaned the data and found the most relevant features to the Target variable and prediction.

Further we are planning to improvise the feature engineering, perform hyperparameter tuning for our models alongside using K-Fold cross validation and GridSearchCV, we might also use some advanced gradient boosting models so that we could get as close to the best accuracy as we can.

INDIANA UNIVERSITY BLOOMINGTON