MODELING DEFAULT RISK

ALIREZA GHASEMIEH

STUDENT ID: 500925479

TODAY PRESENTATION

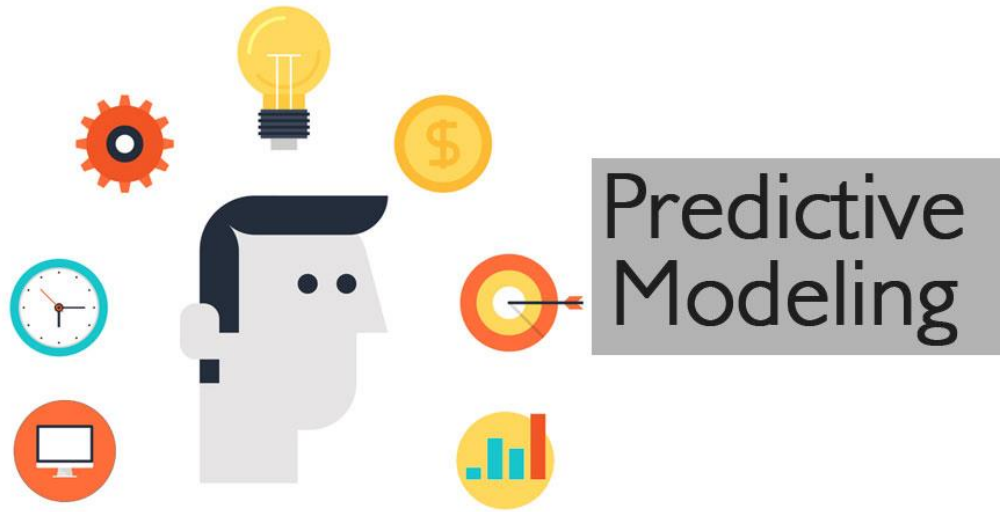
Problem

Dataset

Approach

Result

PROBLEM



Bank wants to calculate the default risk probability of loan applicants based on their financial history over all other financial institutes

DATASET

- Application Train / Test
- Bureau
- Bureau Balance
- Previous Application
- POS Cash Balance
- Credit Card Balance
- Installments Payment

DATASET

Total Attributes 219

application_{train|test}.csv
- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

124 * 350k

SK_ID_CURR

SK_ID_CURR

SK_ID_CURR

bureau.csv
- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

17 * 1.7m

previous_application.csv
- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

38*1.67m

SK_ID_BUREAU

SK_ID_PREV

SK_ID_PREV

bureau_balance.csv
- Monthly balance of credits in Credit Bureau
- Behavioral data

3 * 27.3m

POS_CASH_balance.csv
- Monthly balance of client's previous loans in Home Credit
- Behavioral data

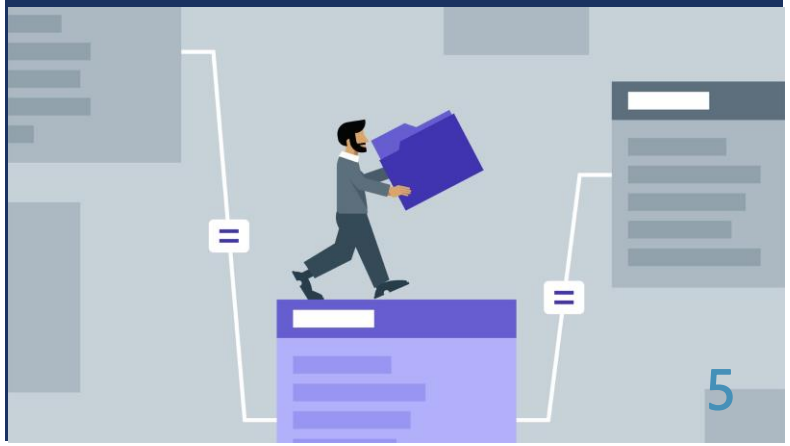
8 * 10m

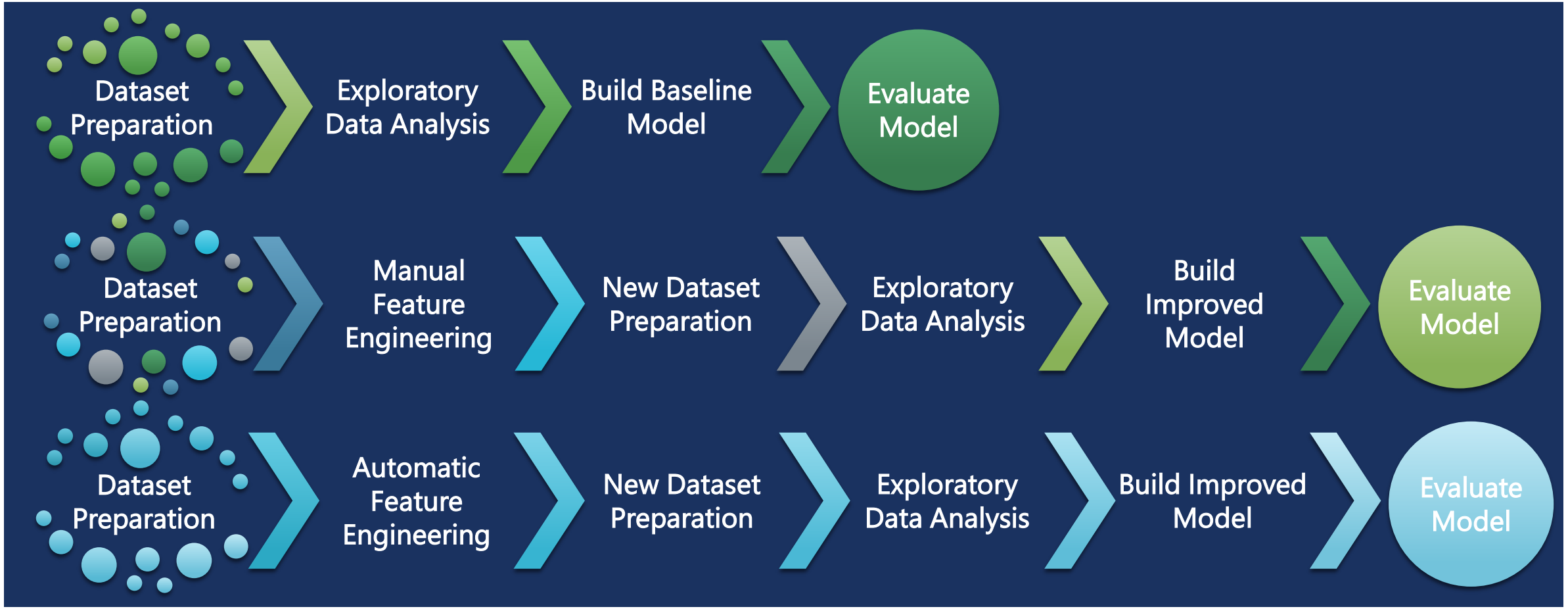
instalments_payments.csv
- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

8*13.6m

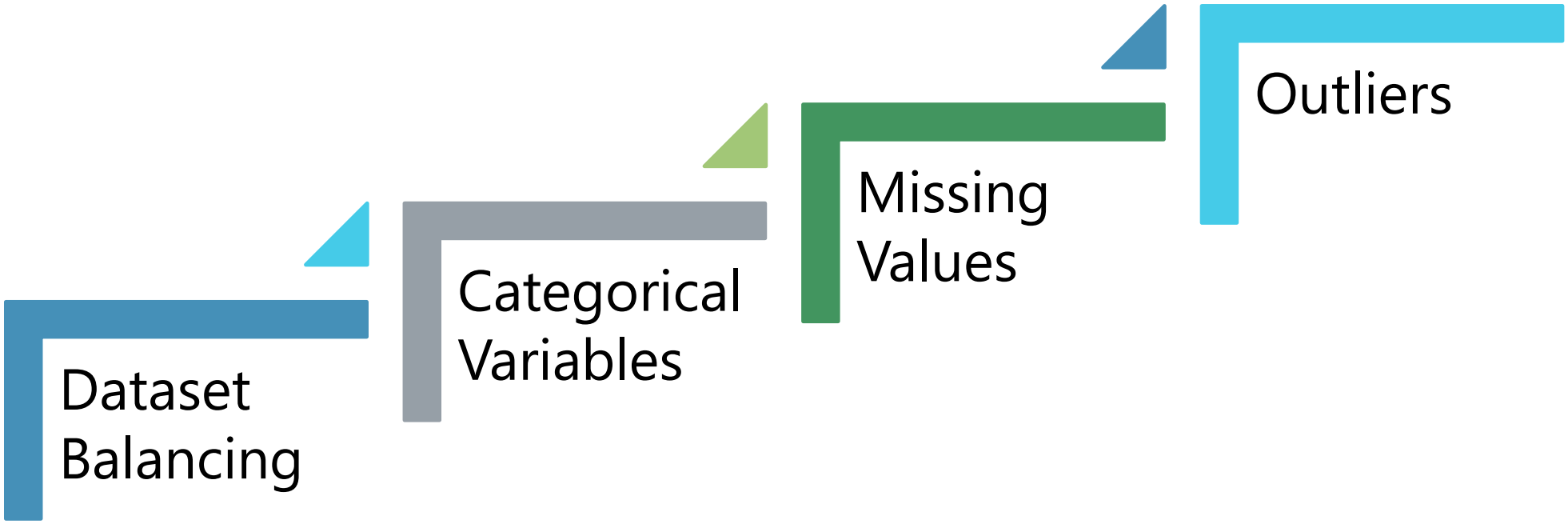
credit_card_balance.csv
- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

23 * 3.8m



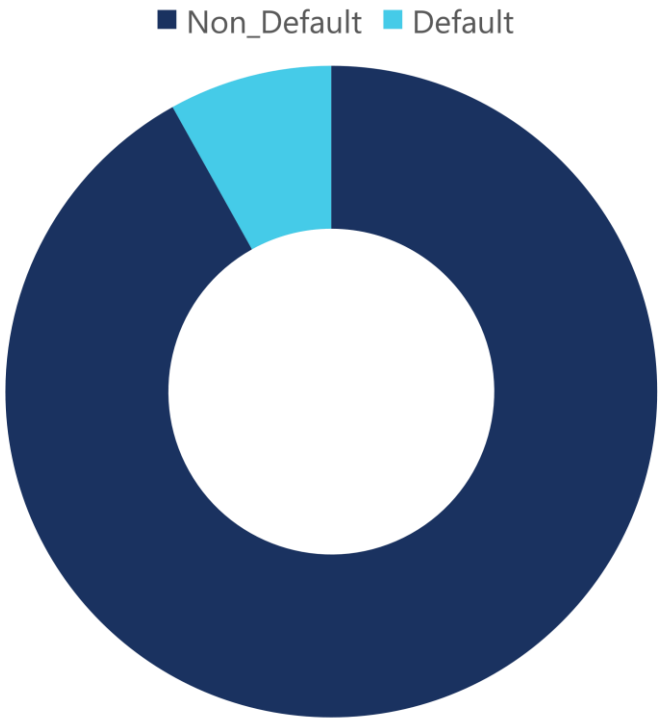

Dataset Preparation
Exploratory Data Analysis
Build Baseline Model
Evaluate Model
Dataset Preparation
Manual Feature Engineering
New Dataset Preparation
Exploratory Data Analysis
Build Improved Model
Evaluate Model
Dataset Preparation
Automatic Feature Engineering
New Dataset Preparation
Exploratory Data Analysis
Build Improved Model
Evaluate Model

APPROACH

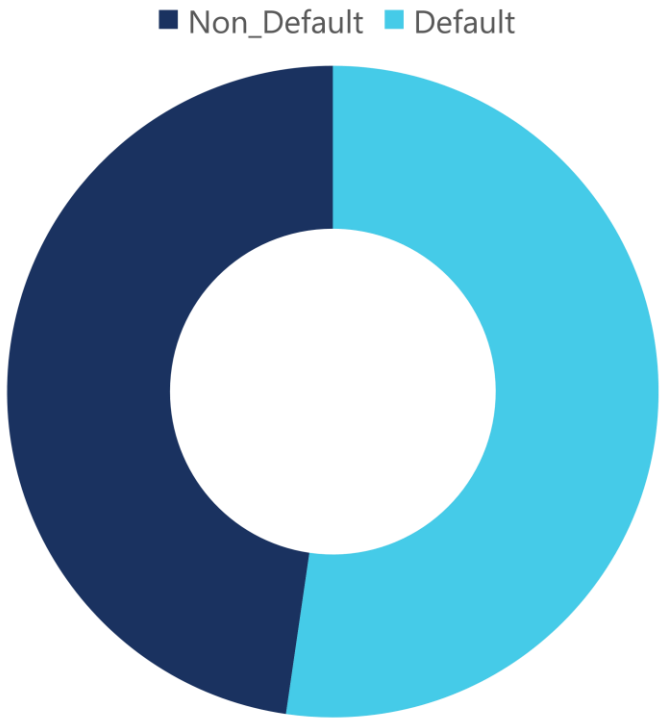


DATASET PREPARATION

DATASET BALANCING

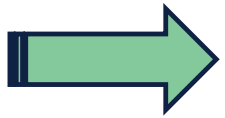


- Over Sampling
- Under Sampling
- SMOT Method



DATASET PREPARATION (CATEGORICAL VARIABLES)

Own Car
Y
N
N
Y



Own Car
1
0
0
1

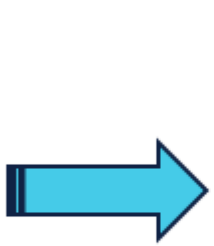
- Categorical Variables
  - Convert to Binomial Variables
  - Convert to Dummy Variables

Car Model
Ford
BMW
Hyundai
Honda

Car Model_Ford	Car Model _Hyundai	Car Model _Honda	Car Model _BMW
1	0	0	0
0	0	0	1
0	1	0	0
0	0	1	0

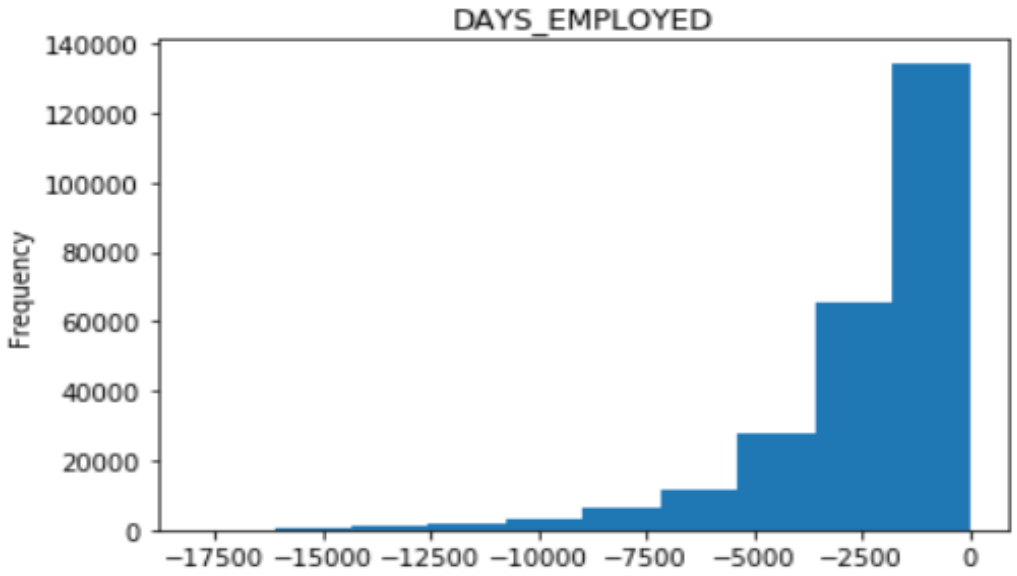
DATASET PREPARATION (MISSING VALUES)

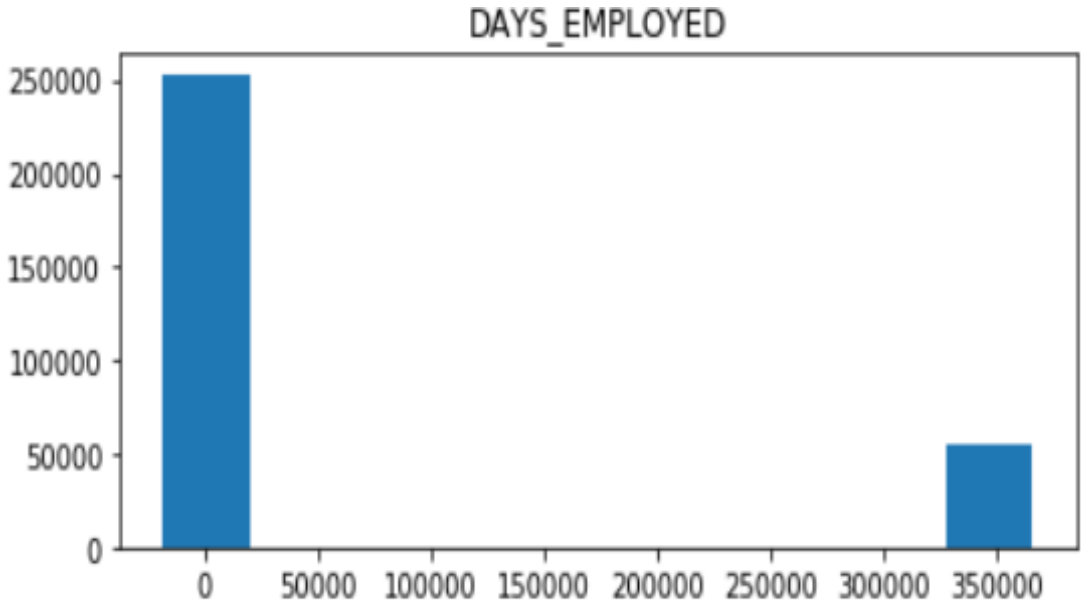
Attr
DATA
NA
NA
DATA
NA
DATA



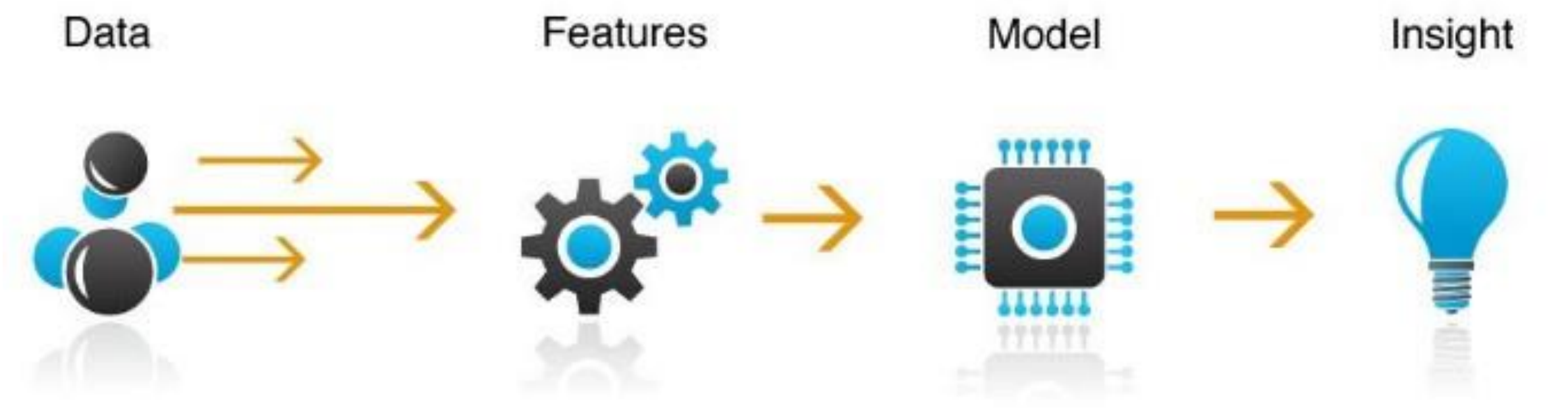
Attr	Miss_Tag
DATA	0
NA	1
NA	1
DATA	0
NA	1
DATA	0

- Replace with Median for Quantitative Variables
- Replace with Mode for Qualitative Variables
- Remove the Column with 90% Missing value
- Remove Records with Missing Value
- Create an Identification Variable to Mark Missing Values



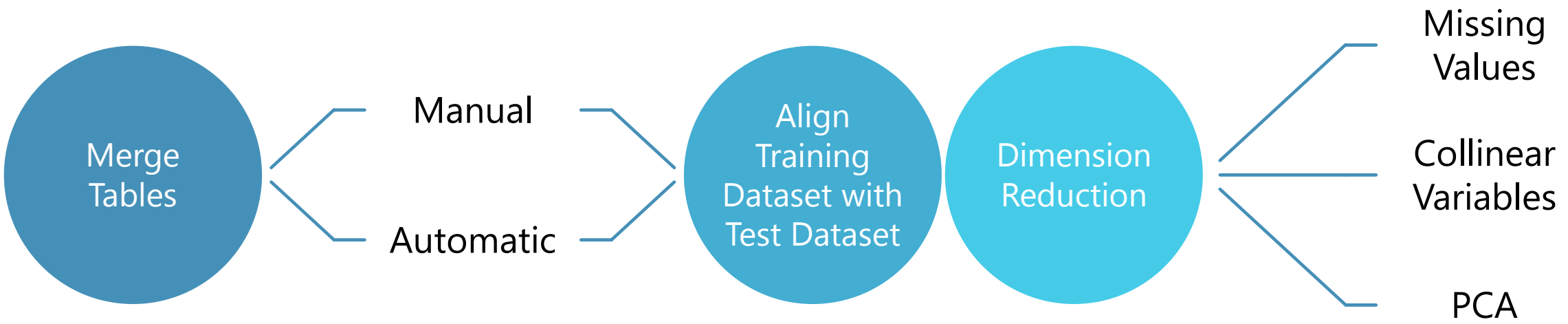


DATASET PREPARATION (OUTLIERS)
REPLACE WITH MEDIAN

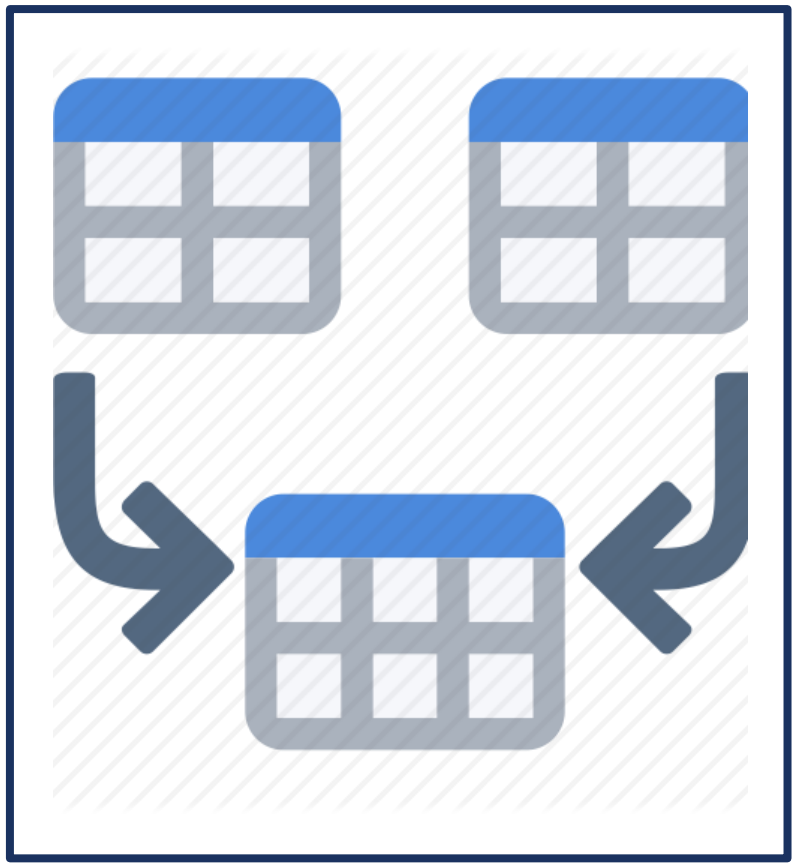


FEATURE ENGINEERING

FEATURE ENGINEERING



MERGING TABLE BY FEATURE ENGINEERING



- Manual Feature Engineering
  - Merge Categorical Attributes
    - Count, Count Norm
  - Merge Numerical Attribute
    - Count, Mean, Max, Min, Sum

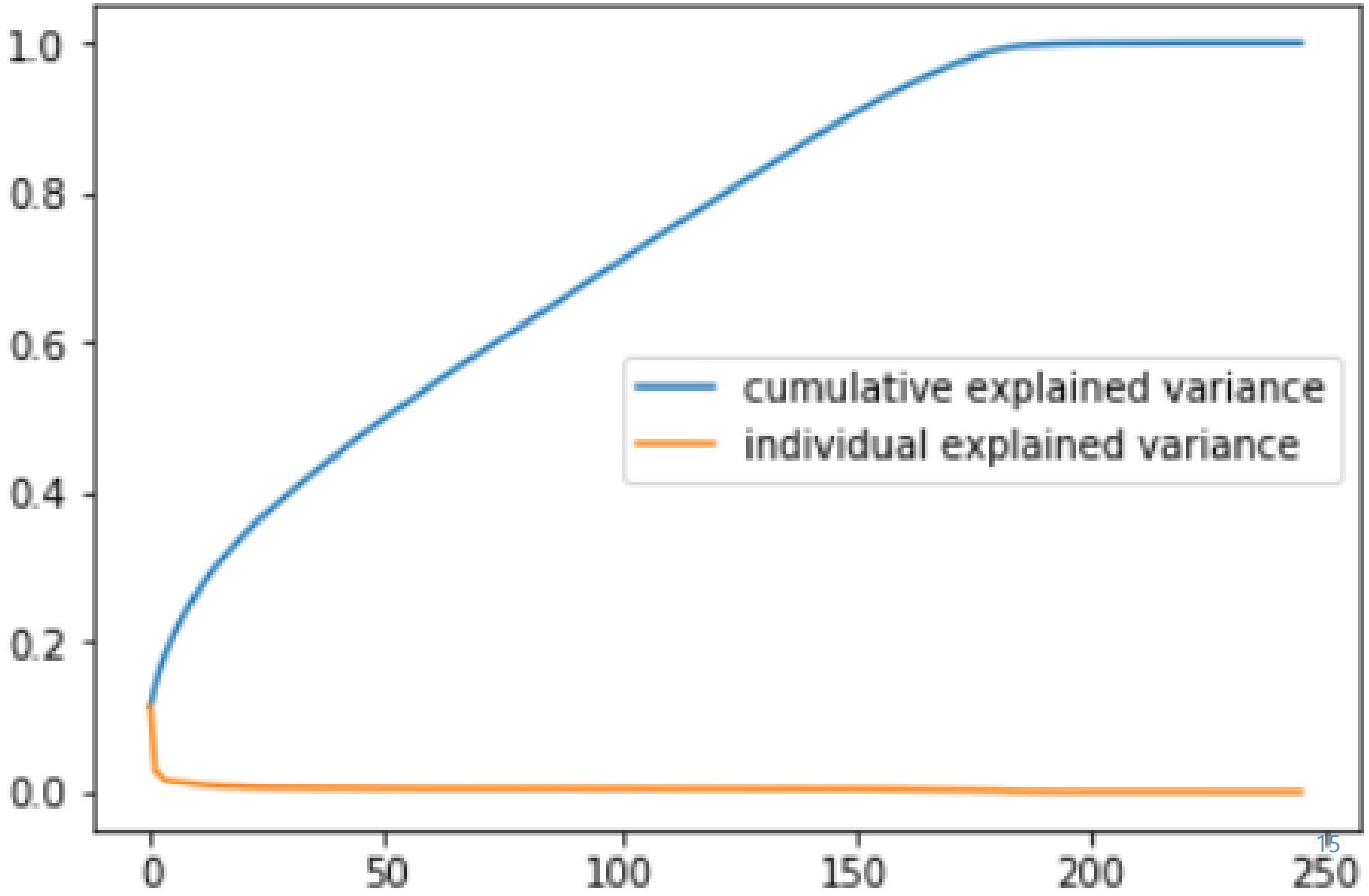
1143

- Automatic Feature Engineering
  - Technique: deep feature synthesis
  - Python Library: Featuretools

2221

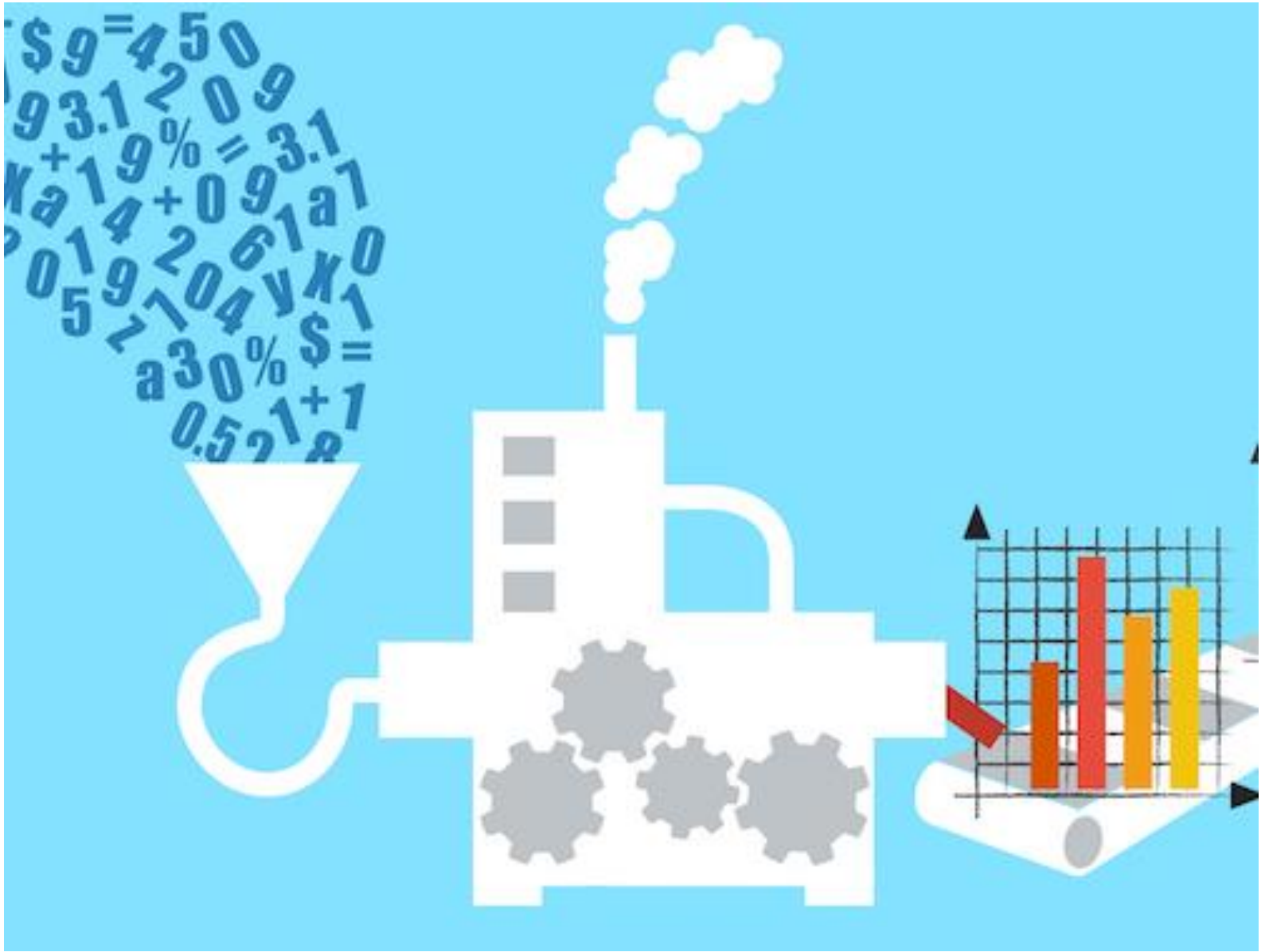
DIMENSION REDUCTION

- Missing Values over 90%
- Collinear Variables
- PCA

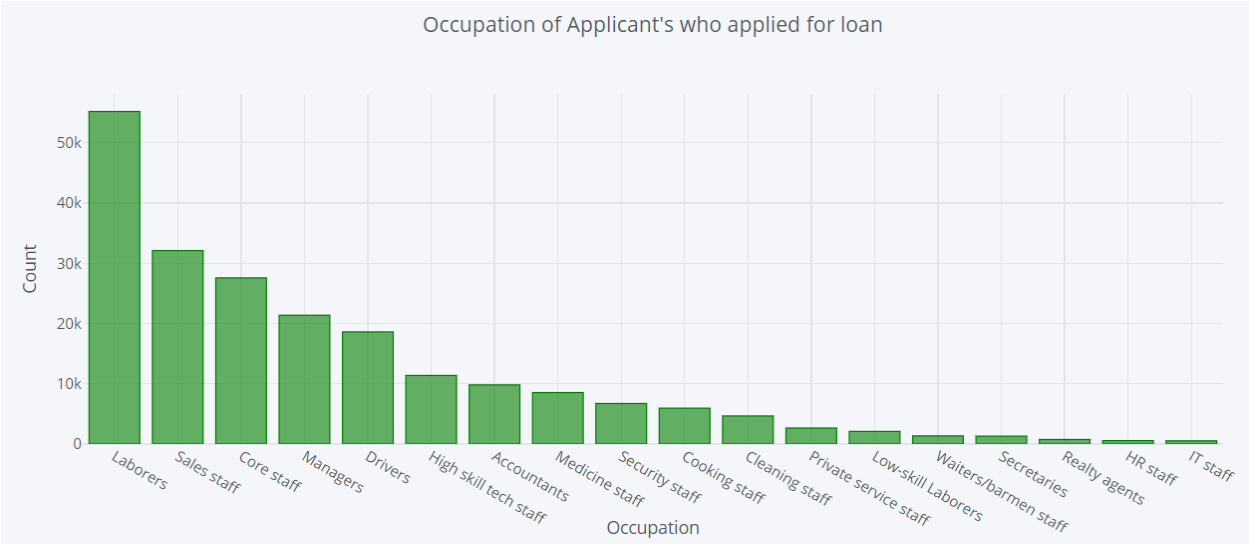


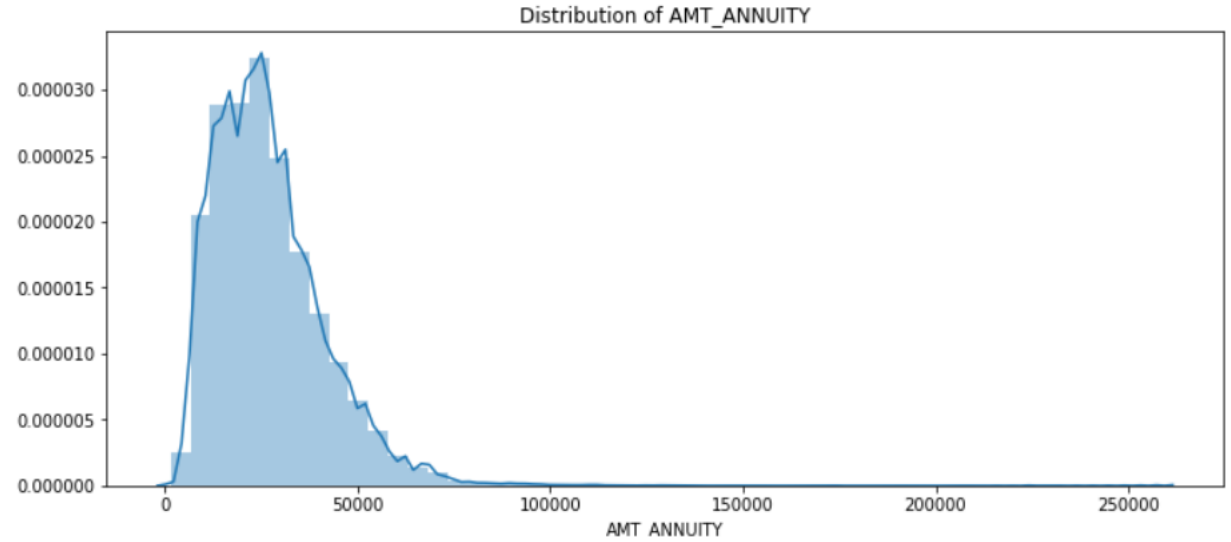
Removed Columns	Baseline	Manual Feature Engineering	Automatic Feature Engineering
Missing Values over 90%	0	0	0
Collinear Variables	134	224	592
PCA	65	NA	NA
Total Num of Col Bef Rem	333	1143	2221
Total Num of Col aft Rem	134	919	1629

DIMENSION REDUCTION

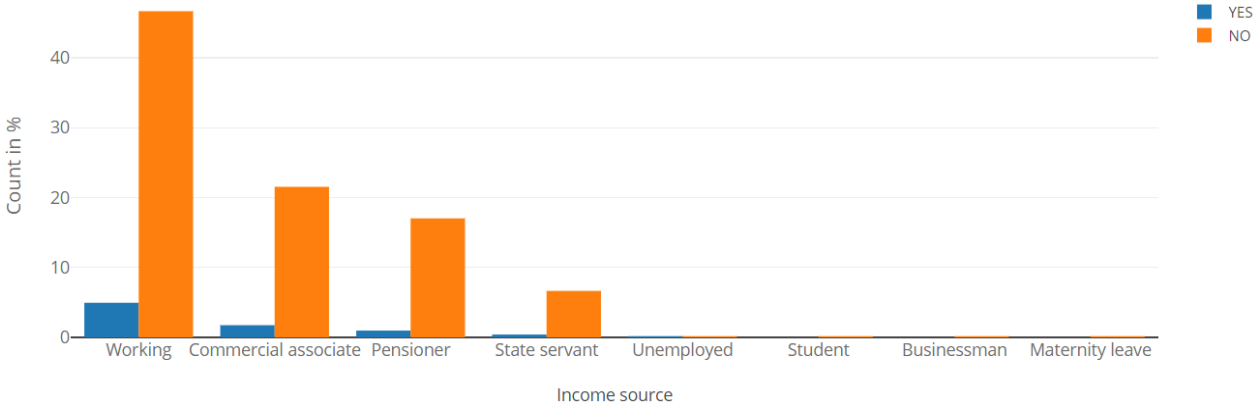


EXPLORATORY DATA ANALYSIS

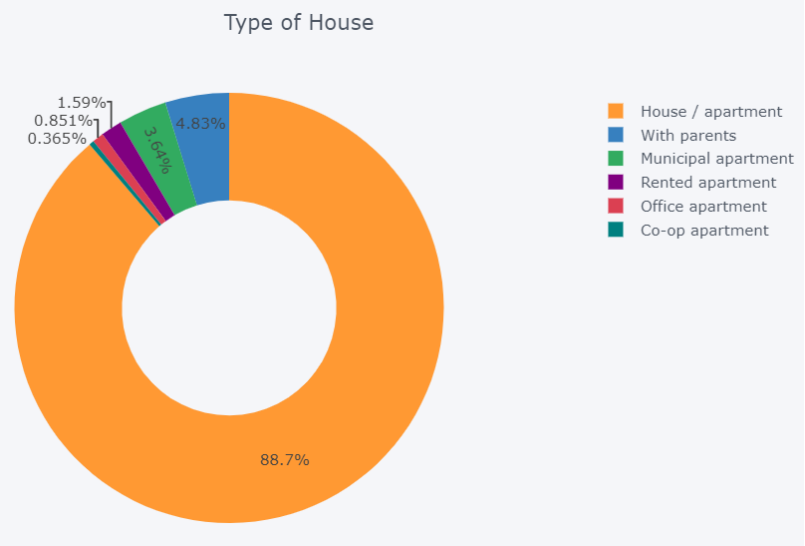


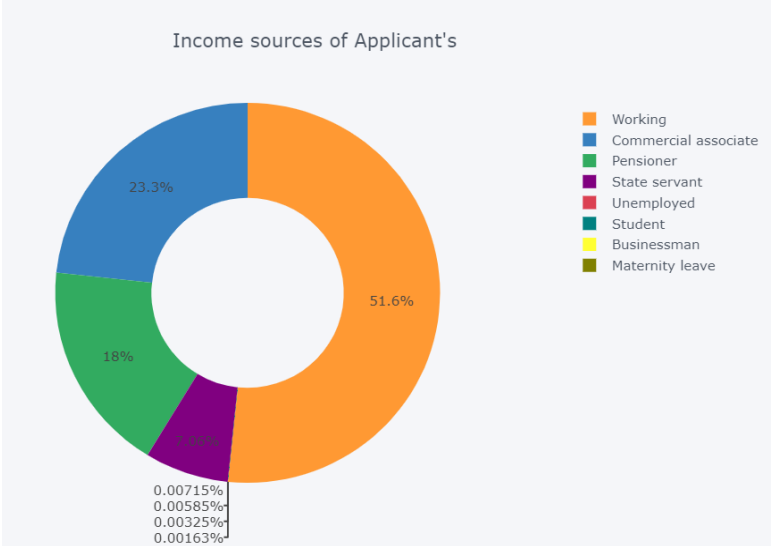


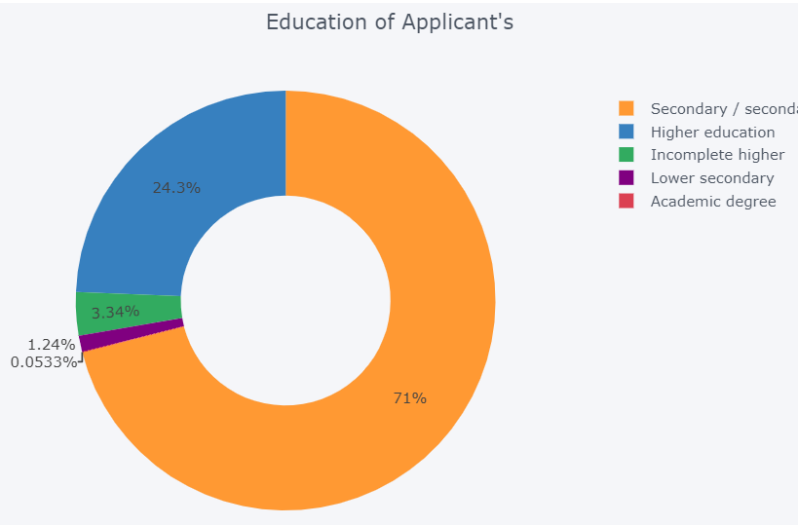
EXPLORATORY DATA ANALYSIS

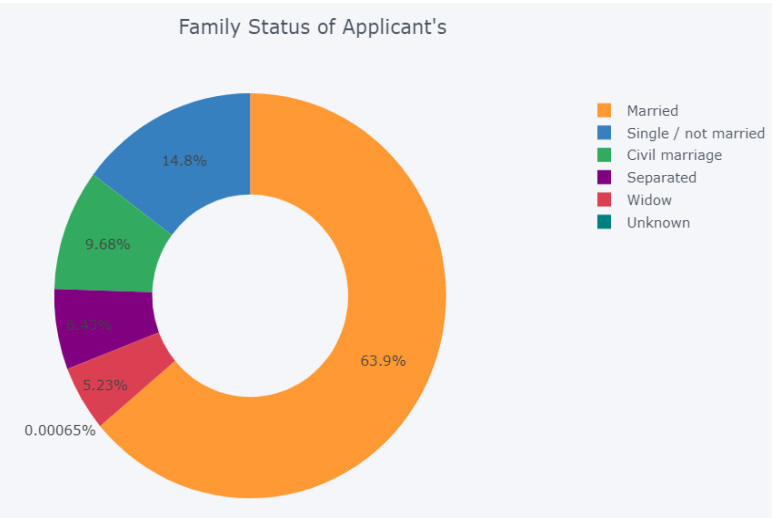


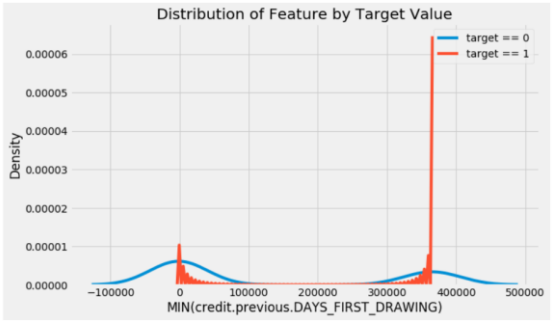
EXPLORATORY DATA ANALYSIS (PIE CHART)

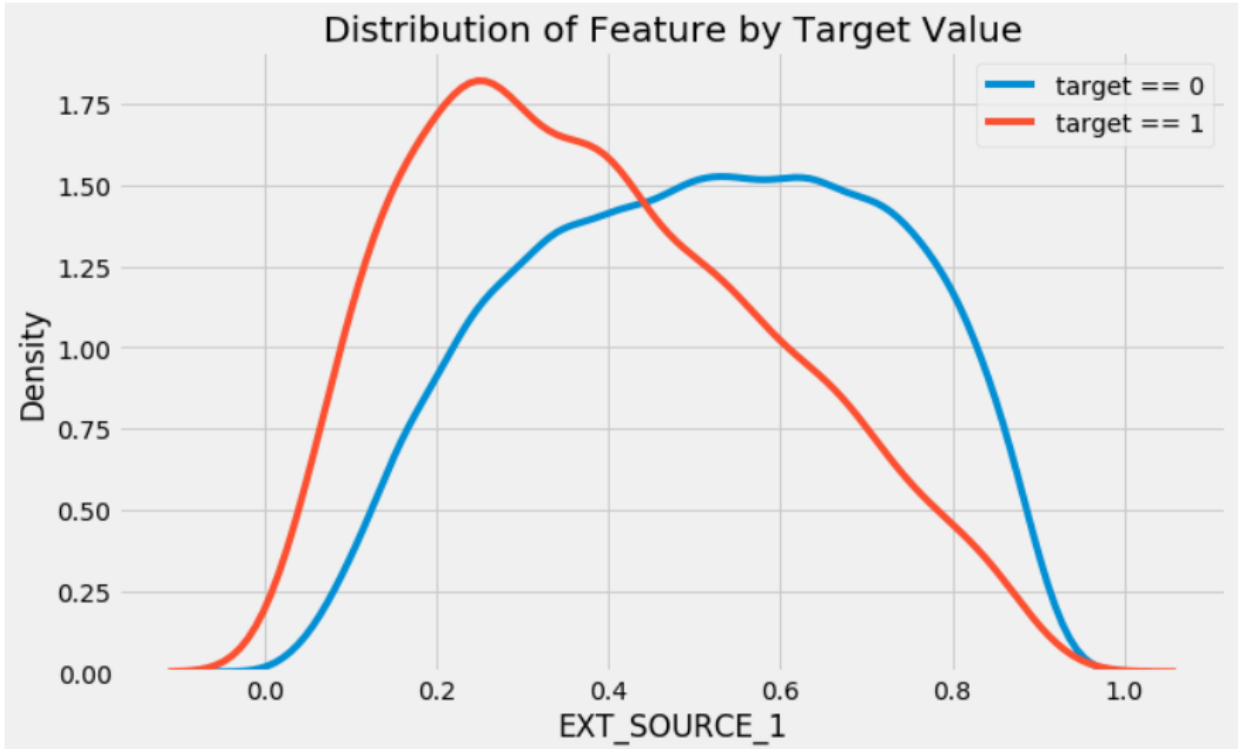


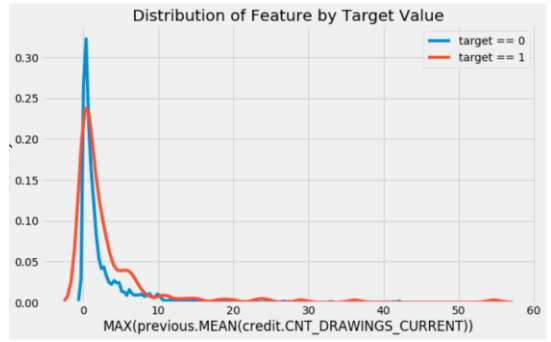










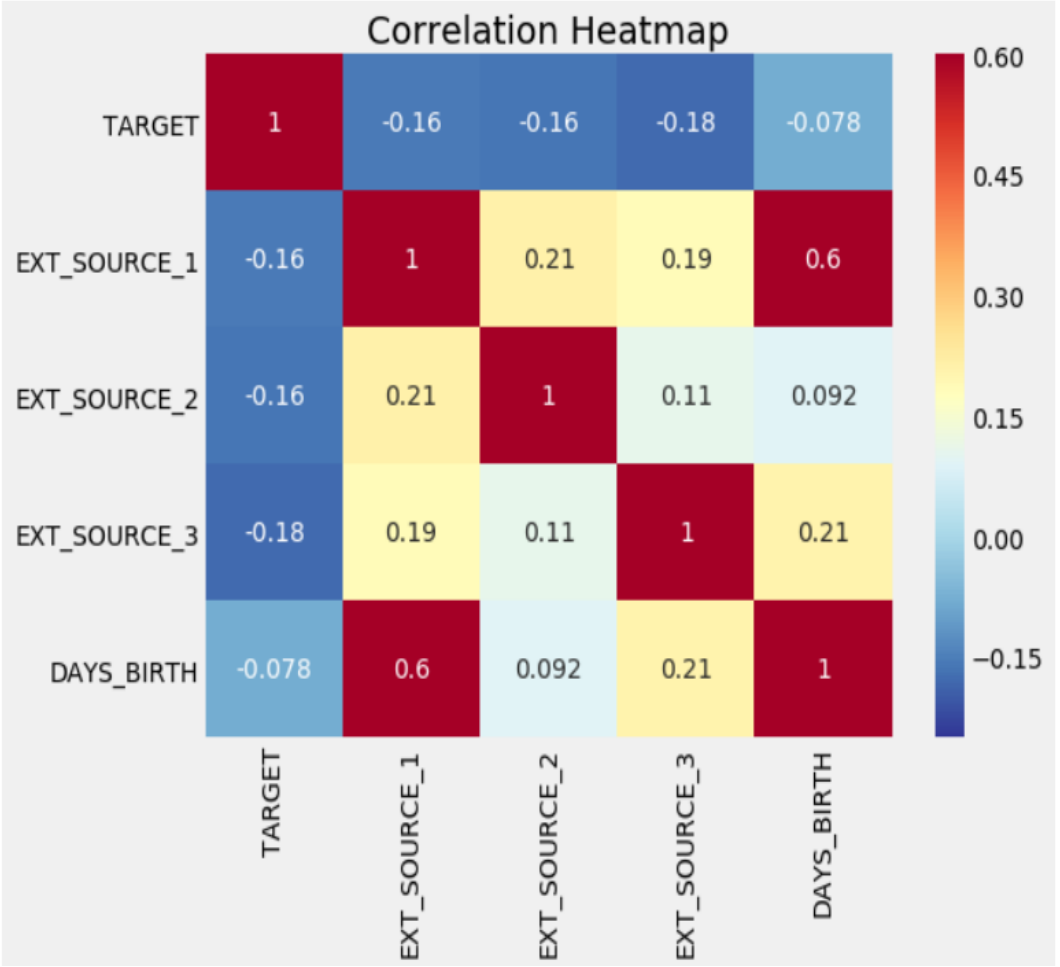


EXPLORATORY DATA ANALYSIS (KDE PLOT)

EXPLORATORY DATA ANALYSIS
(Correlation with Target – Before & After Feature Engineering)

TARGET	1.000000
EXT_SOURCE_3	0.178919
EXT_SOURCE_2	0.160472
EXT_SOURCE_1	0.155317
DAYS_BIRTH	0.078239
DAYS_EMPLOYED	0.074958

TARGET	1.000000
EXT_SOURCE_3	0.178919
EXT_SOURCE_2	0.160472
EXT_SOURCE_1	0.155317
bureau_DAYS_CREDIT_mean	0.089729
client_bureau_balance_MONTHS_BALANCE_min_mean	0.089038
client_bureau_balance_MONTHS_BALANCE_count_mean	0.080193
bureau_CREDIT_ACTIVE_Closed_count_norm	0.079369
DAYS_BIRTH	0.078239
bureau_CREDIT_ACTIVE_Active_count_norm	0.077356

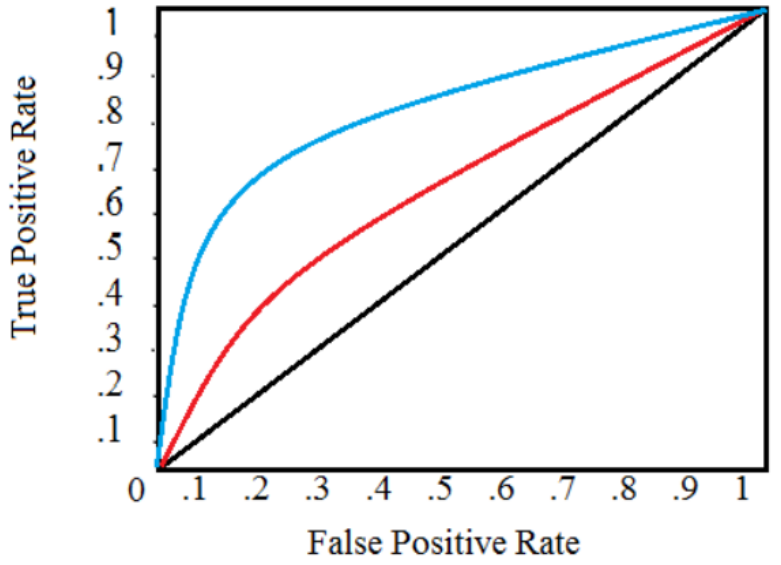


PREDICTIVE MODELING

MODELING METHODS AND EVALUATION RESULTS

ROC-AUC Assessment Score

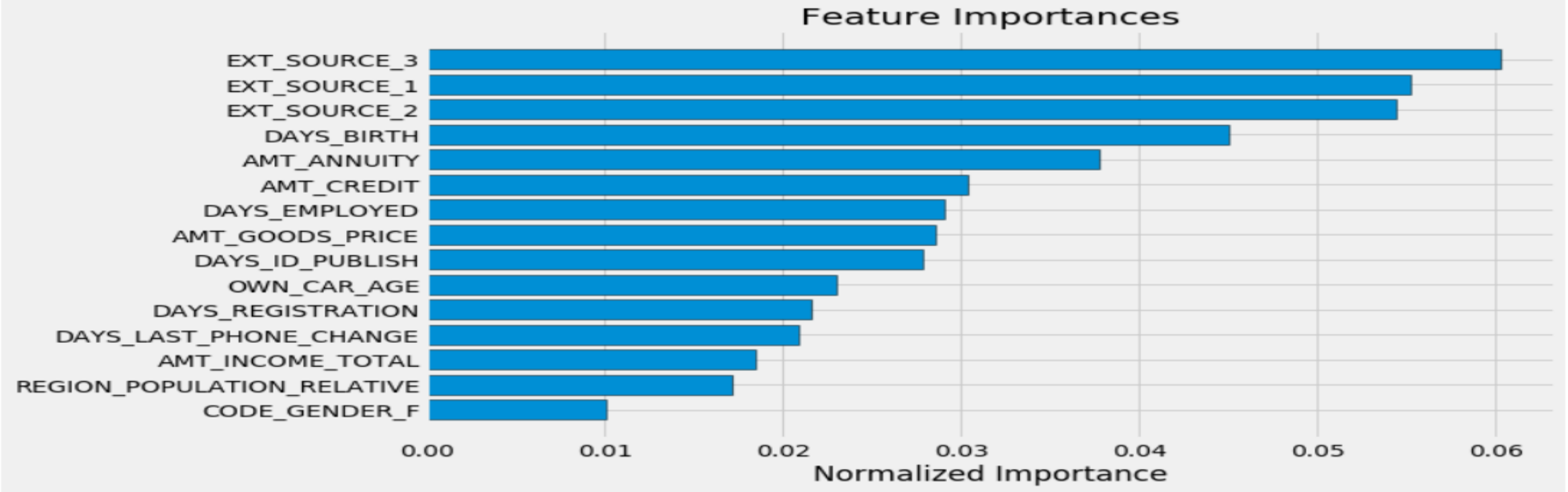
1st Model: Logistic Regression	• Score of 0.68035
2nd Model: Random Forest	• Score of 0.67508
3rd Model: Application Table and Light Gradient Boosting	• Score of 0.74533
4th Model: Manual Feature Engineering and Light Gradient Boosting	• Score of 0.77445
5th Model: Automated Feature Engineering and Light Gradient Boosting	• Score of 0.74169



SUMMARY

- Problem
- Dataset
- Approach
  - Data Preparation
  - Feature Engineering
  - Exploratory Data Analysis (EDA)
  - Modeling
- Result
  - EDA Plots
  - Modeling Accuracy Table

THANK YOU



FEATURE IMPORTANCE

ALIGN TRAINING DATASET WITH TEST DATASET

Train Dataset

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6

Test Dataset

Col 1	Col2	Col 3	Col 5	Col 6	Col 4

DATASET PREPARATION (MISSING VALUES)

- Create an Identification Variable to Mark Missing Values

Attr
DATA
NA
NA
DATA
NA
DATA

Attr	Miss_Tag
DATA	0
NA	1
NA	1
DATA	0
NA	1
DATA	0