

Aug 2019

# Publicis Sapient

**Social Financial Lending**

Prepared by: Vishal Jindal  
Vishaljindal09@gmail.com



# TABLE OF CONTENTS

03	OBJECTIVE & ASSUMPTIONS
05 - 11	EXPLORATORY DATA ANALYSIS
12 - 13	FEATURE ENGINEERING
14 - 15	END TO END SOLUTION

---

# Project Objective

The objective of the report is to share an architectural diagram proposing end to end solution along with the Exploratory Data Analysis for the hackathon organized by Sapient.

The following steps have been taken to achieve the goal of this report:

- Understanding the structure of the dataset
- Performing the Exploratory Data Analysis
- Performing Feature Engineering
- End to End solution proposition

## Assumptions

It is assumed that the data collected is without any personal bias and anonymous for the privacy concerns.

For Target Column in data, 1 means repayment.

# Exploratory Data Analysis

*The following steps were implemented to conduct the EDA for the given dataset:*

- Data dimension and structure check
- Missing values check
- Univariate Analysis of the Data
- Bivariate Analysis with Target

## Data Dimension and Structure

Data dimension is 257512 x 122 with a Target column as Binary (0/1)  
Target column has ~8% hit or success rate

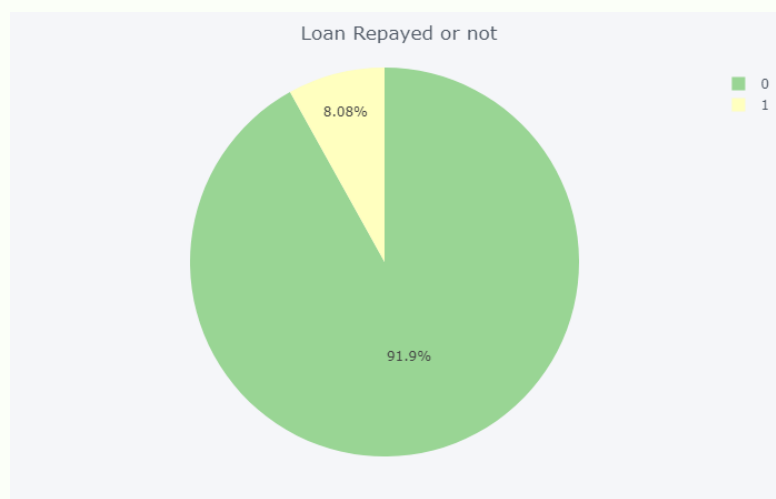
## Missing Value Check

In total 187 variables has missing values. Below is the screen shot of top 20 variables in terms missing value percentage.

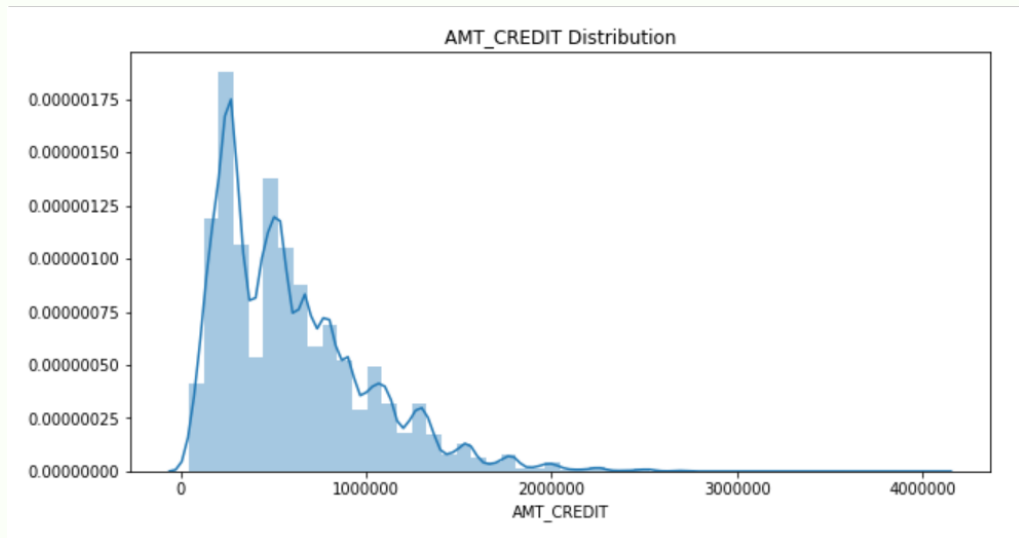
	Total	Percent
COMMONAREA_MEDI	179905	69.862764
COMMONAREA_AVG	179905	69.862764
COMMONAREA_MODE	179905	69.862764
NONLIVINGAPARTMENTS_MODE	178800	69.433657
NONLIVINGAPARTMENTS_MEDI	178800	69.433657
NONLIVINGAPARTMENTS_AVG	178800	69.433657
FONDKAPREMONT_MODE	176104	68.386716
LIVINGAPARTMENTS_MEDI	175973	68.335845
LIVINGAPARTMENTS_MODE	175973	68.335845
LIVINGAPARTMENTS_AVG	175973	68.335845
FLOORSMIN_MEDI	174748	67.860139
FLOORSMIN_MODE	174748	67.860139
FLOORSMIN_AVG	174748	67.860139
YEARS_BUILD_MEDI	171249	66.501367
YEARS_BUILD_AVG	171249	66.501367
YEARS_BUILD_MODE	171249	66.501367
OWN_CAR_AGE	169979	66.008186
LANDAREA_MODE	152869	59.363835
LANDAREA_AVG	152869	59.363835
LANDAREA_MEDI	152869	59.363835

## Univariate Analysis

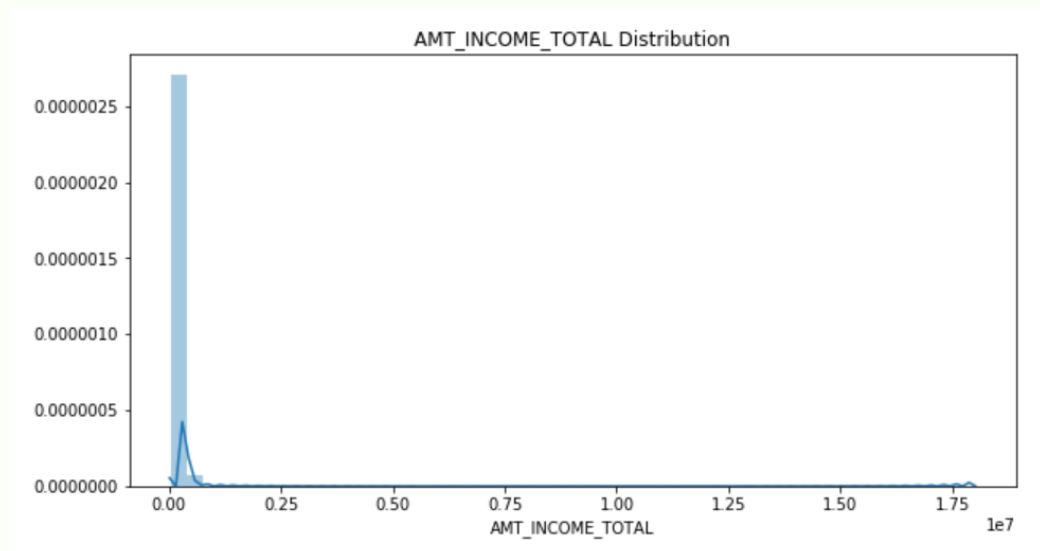
- *Target*



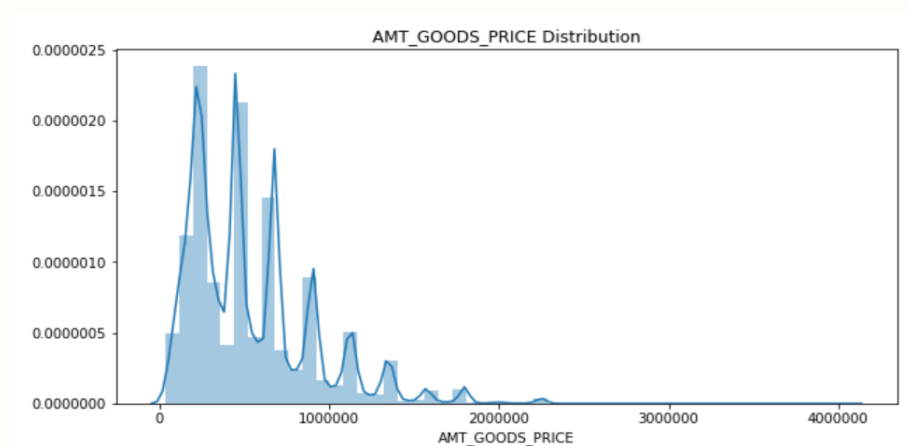
- *AMT\_CREDIT*



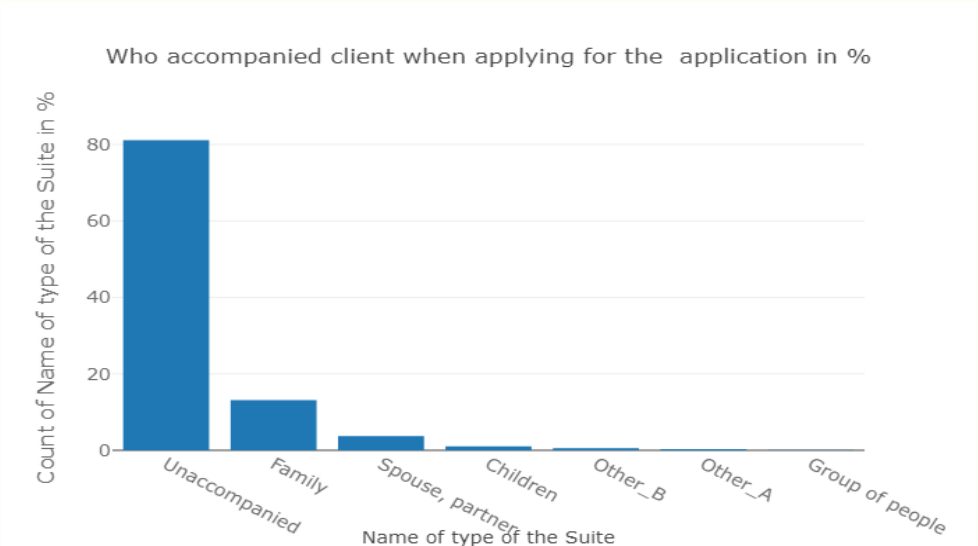
- *AMT\_INCOME\_TOTAL*



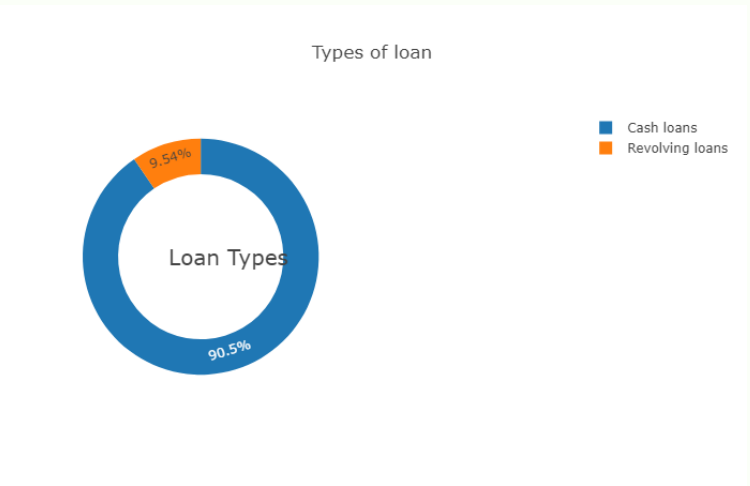
- *AMT\_GOODS\_PRICE*



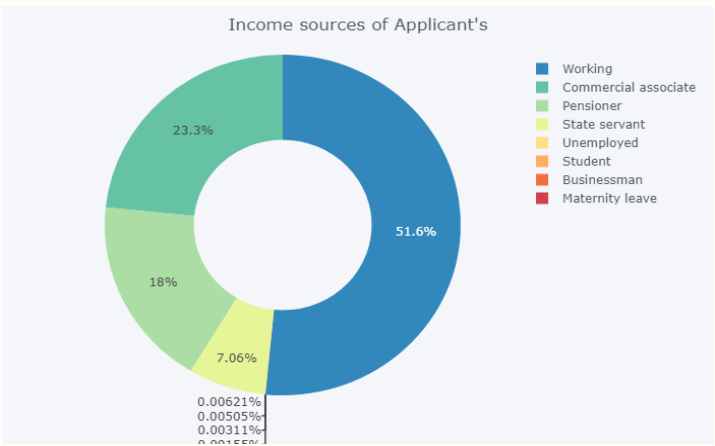
- NAME\_TYPE\_SUITE



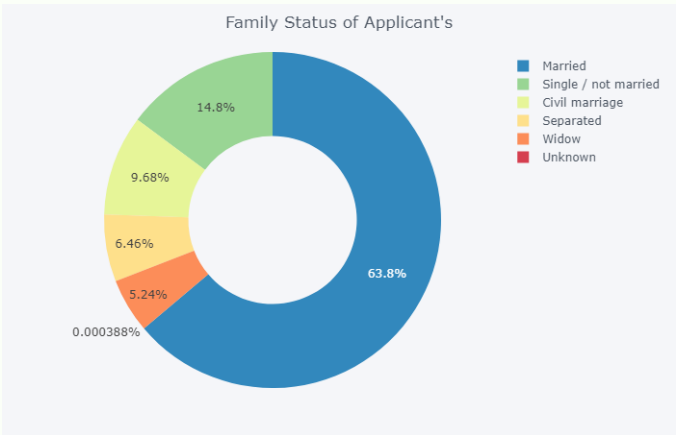
- NAME\_CONTRACT\_TYPE



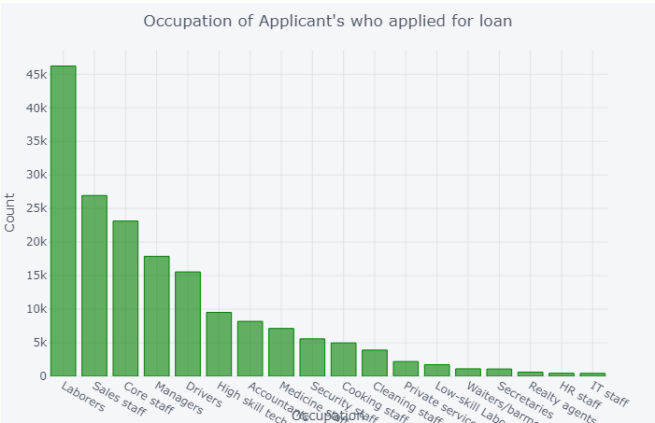
- NAME\_INCOME\_TYPE



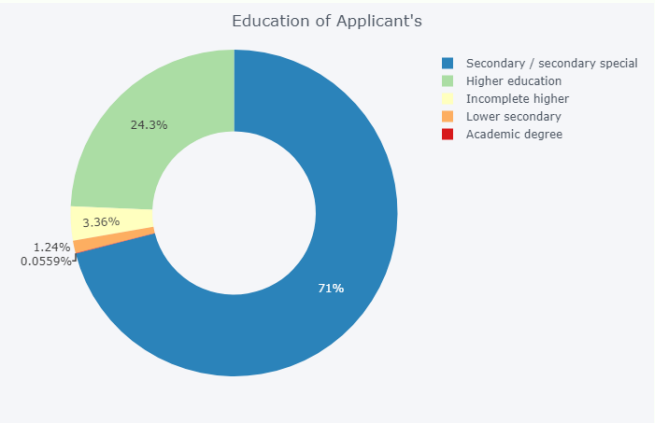
- NAME\_FAMILY\_STATUS



- OCCUPATION\_TYPE

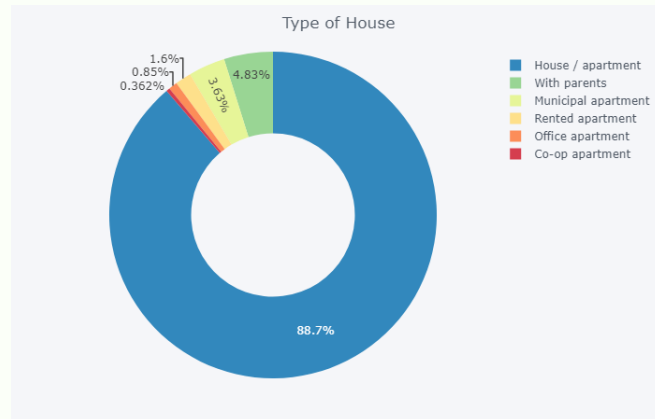


- NAME\_EDUCATION\_TYPE





- *NAME\_HOUSING\_TYPE*

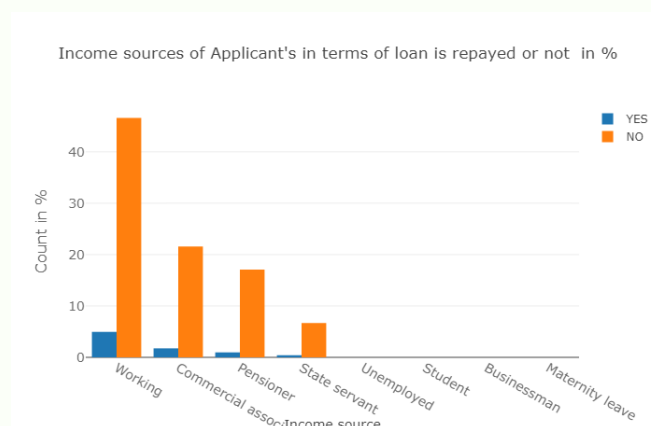


- *ORGANIZATION\_TYPE*

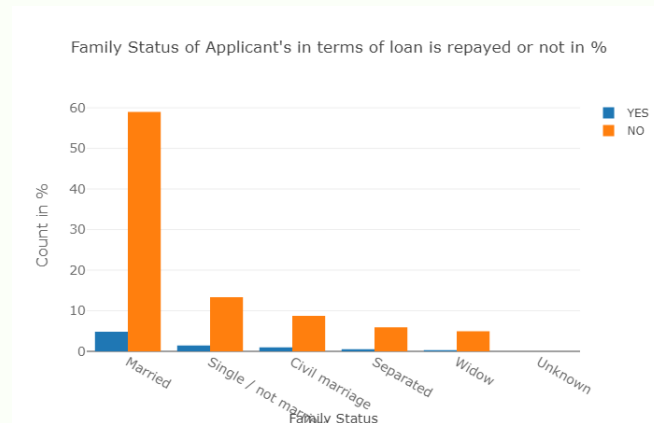


## Bivariate Analysis

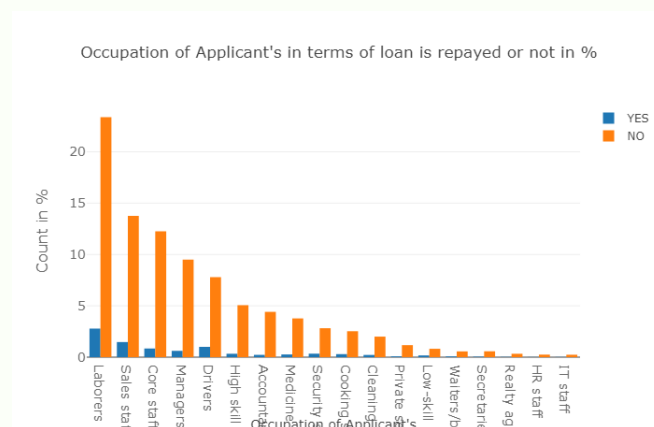
- *NAME\_INCOME\_TYPE VS TARGET*



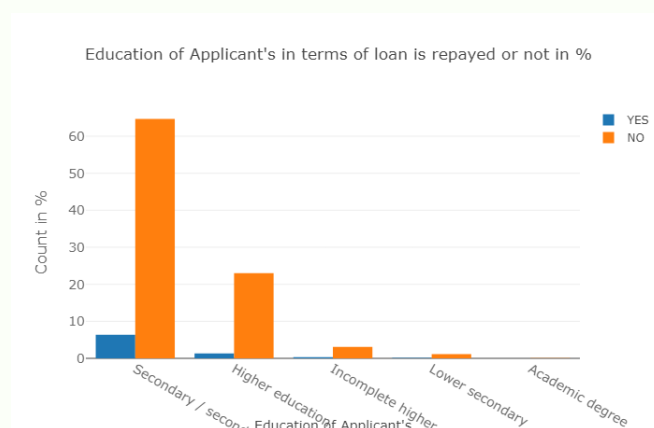
- **NAME\_FAMILY\_STATUS VS TARGET**



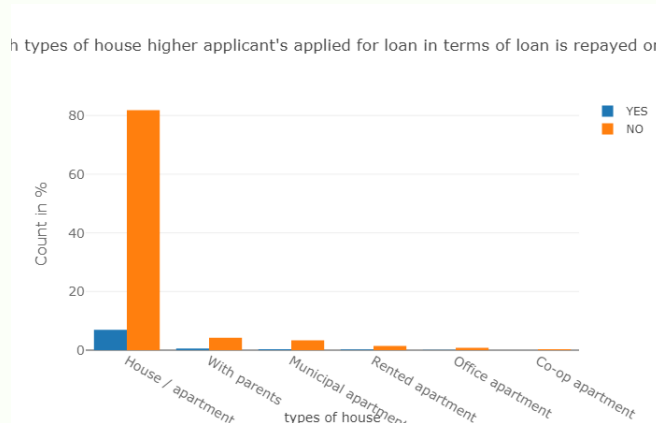
- **OCCUPATION\_TYPE VS TARGET**



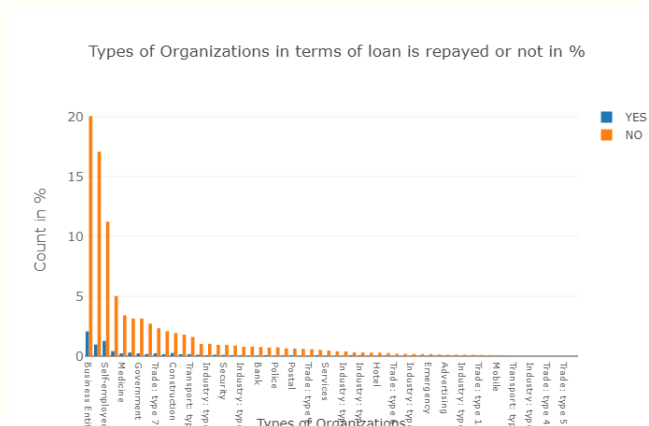
- **NAME\_EDUCATION\_TYPE VS TARGET**



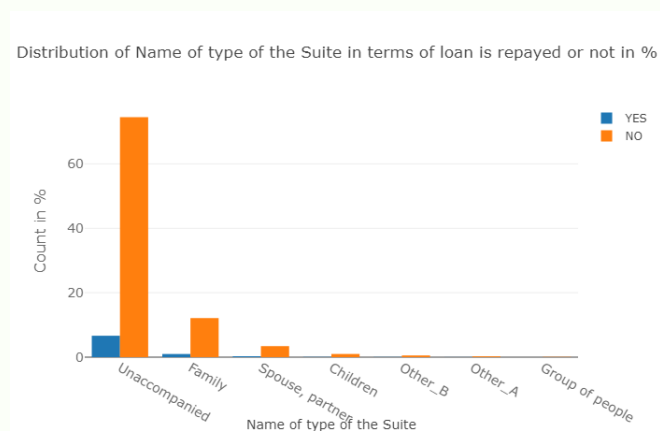
- **NAME\_HOUSING\_TYPE VS TARGET**



- **ORGANIZATION\_TYPE VS TARGET**



- **NAME\_TYPE\_SUITE VS TARGET**



# Feature Engineering

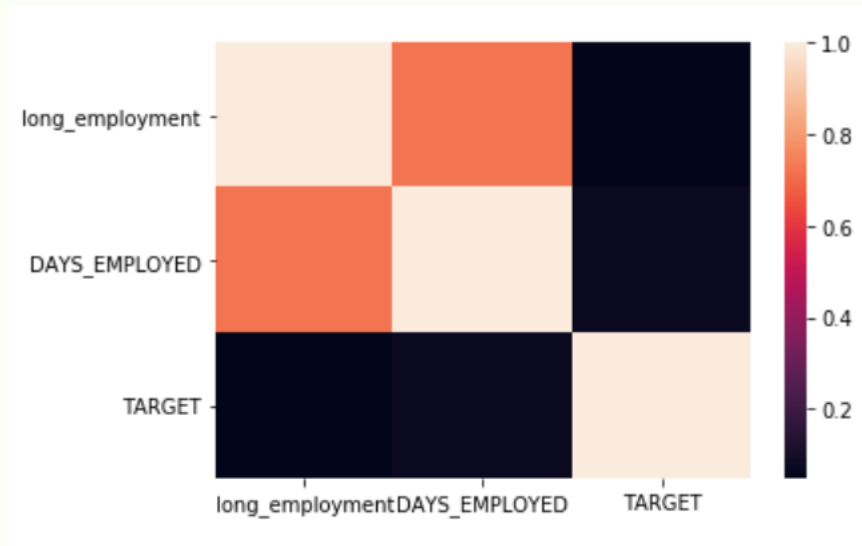
- Ration Features VS Target*

TARGET	1.000000
external_sources_weighted	0.235028
external_sources_mean	0.222133
external_sources_nanmedian	0.218030
external_sources_max	0.197127
external_sources_min	0.185437
external_sources_sum	0.173338
credit_to_goods_ratio	0.069384
car_to_birth_ratio	0.048457
days_employed_percentage	0.043187
credit_to_annuity_ratio	0.031420
car_to_employ_ratio	0.030803
phone_to_birth_ratio	0.029440
income_per_child	0.024110
children_ratio	0.019914
income_per_person	0.015136
annuity_income_percentage	0.013535
payment_rate	0.012255
income_credit_percentage	0.009177
credit_to_income_ratio	0.008026
phone_to_employ_ratio	0.006792

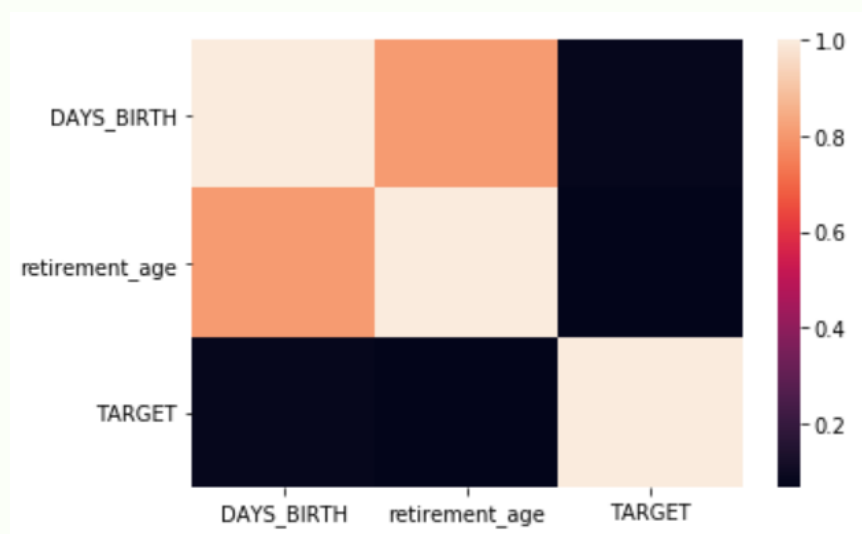
- Aggregation Features VS Target*

TARGET	1.000000
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_REG_CITY_NOT_WORK_CITY_mean_EXT_SOURCE_1	0.090862
CODE_GENDER_NAME_EDUCATION_TYPE_OCCUPATION_TYPE_REG_CITY_NOT_WORK_CITY_mean_EXT_SOURCE_2	0.088821
CODE_GENDER_NAME_EDUCATION_TYPE_OCCUPATION_TYPE_REG_CITY_NOT_WORK_CITY_mean_EXT_SOURCE_1	0.086734
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_EXT_SOURCE_1	0.084575
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_EXT_SOURCE_2	0.082998
OCCUPATION_TYPE_mean_EXT_SOURCE_1	0.077959
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_REG_CITY_NOT_WORK_CITY_mean_ELEVATORS_AVG	0.077639
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_OW_N_CAR_AGE	0.073659
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_YEARS_BUILD_AVG	0.073553
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_AMT_REQ_CREDIT_BUREAU_YEAR	0.073016
OCCUPATION_TYPE_mean_EXT_SOURCE_2	0.071982
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_AMT_CREDIT	0.071908
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_APARTMENTS_AVG	0.071615
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_NONLIVINGAREA_AVG	0.070728
CODE_GENDER_NAME_EDUCATION_TYPE_mean_EXT_SOURCE_1	0.070388
NAME_EDUCATION_TYPE_OCCUPATION_TYPE_mean_BASEMENTAREA_AVG	0.070027
CODE_GENDER_ORGANIZATION_TYPE_mean_EXT_SOURCE_1	0.070021
CODE_GENDER_NAME_EDUCATION_TYPE_max_OW_N_CAR_AGE	0.066638
CODE_GENDER_REG_CITY_NOT_WORK_CITY_mean_CNT_CHILDREN	0.065892
CODE_GENDER_NAME_EDUCATION_TYPE_max_AMT_ANNUITY	0.064714
CODE_GENDER_NAME_EDUCATION_TYPE_max_AMT_CREDIT	0.063963
CODE_GENDER_NAME_EDUCATION_TYPE_mean_EXT_SOURCE_2	0.061696
CODE_GENDER_NAME_EDUCATION_TYPE_sum_OW_N_CAR_AGE	0.061361
CODE_GENDER_ORGANIZATION_TYPE_mean_DAYS_REGISTRATION	0.053078
CODE_GENDER_ORGANIZATION_TYPE_mean_AMT_ANNUITY	0.052510
OCCUPATION_TYPE_mean_DAYS_EMPLOYED	0.050454
CODE_GENDER_ORGANIZATION_TYPE_mean_AMT_INCOME_TOTAL	0.050054
OCCUPATION_TYPE_mean_AMT_ANNUITY	0.048634
CODE_GENDER_REG_CITY_NOT_WORK_CITY_mean_AMT_ANNUITY	0.047807
CODE_GENDER_REG_CITY_NOT_WORK_CITY_mean_DAYS_ID_PUBLISH	0.042131
OCCUPATION_TYPE_mean_DAYS_REGISTRATION	0.034296
OCCUPATION_TYPE_mean_CNT_CHILDREN	0.018418
OCCUPATION_TYPE_mean_EXT_SOURCE_3	0.007847

- *Employment Features VS Target*



- *Age Features VS Target*

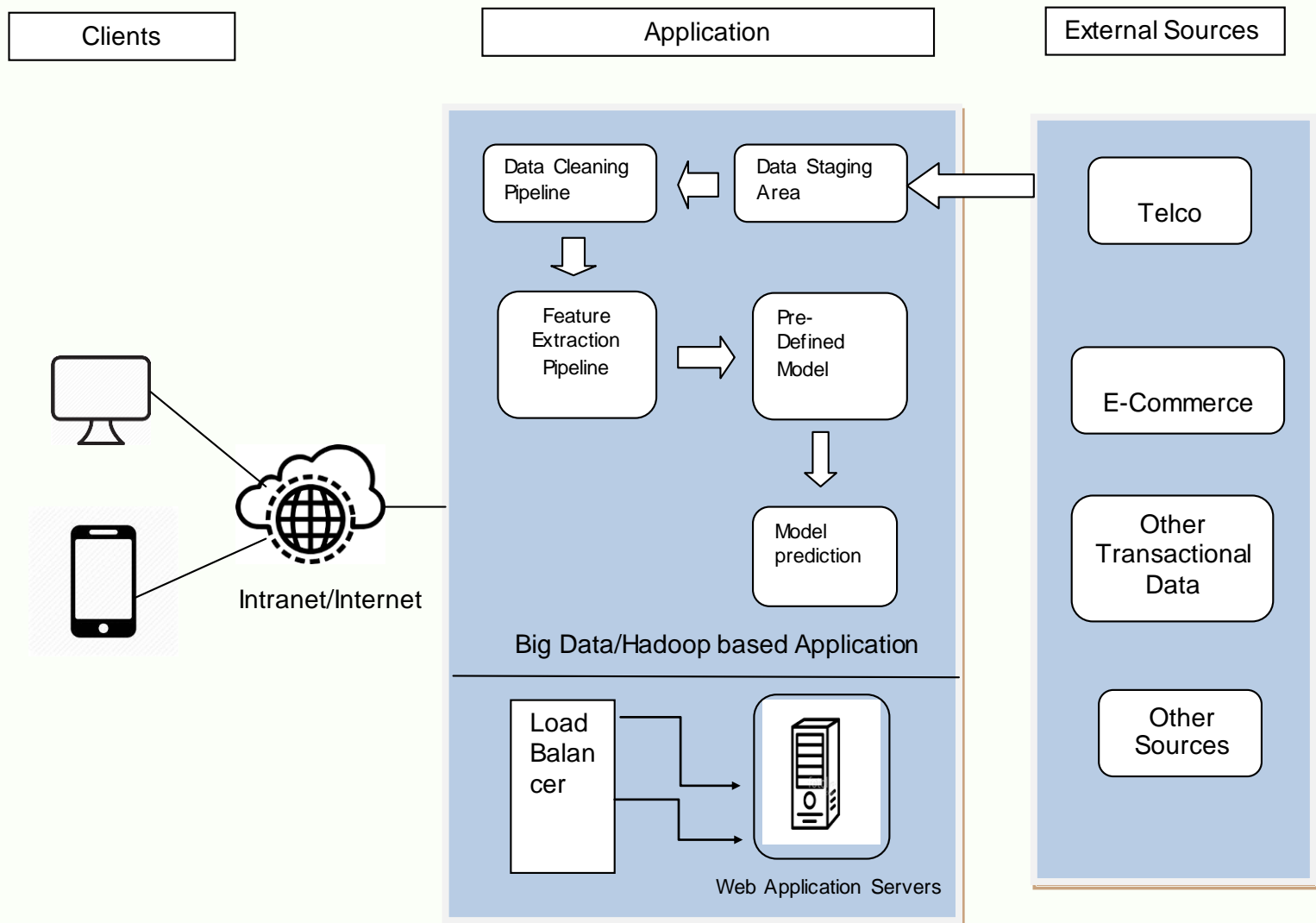


- *Mixed Features VS Target*



TARGET	1.000000
credit_per_child	0.032618
credit_per_person	0.022912
credit_per_non_child	0.020068
child_to_non_child_ratio	0.019420
income_per_non_child	0.012004
cnt_non_child	0.011364

# End To End Solution

- Figure below illustrated the proposed end to end solution to the requirement.



- The data from different sources like Telco, E-Commerce, and Supermarkets Chains. Etc. will be received on the staging Area of the application. It can be

- 
- 
- real time or end of Day data*
- *Data then will be processed by different pipelines namely*
    - *Data Cleaning*
    - *Feature Extraction*
  - *After the raw data is prepared by the pipelines the data will be churned though the Pre-defined Model*
  - *Model will output the probability/Target for the required individual.*
  - *The model output will be exposed to the client via Web application.*
  - *The application will accept a Unique Identification(PAN/ADHAAR) for an Individual who is seeking a Loan*
  - *The application will seek the model output data for the individual and show it real time.*

## **Scalability and fault-tolerance**

*As describer application will be the backbone of entire loan process, scalability and fault tolerance becomes an importance aspect to look into*

- **For the scalability aspect the backend application will be based on Big Data Platform. Based on future use new servers can be added or removed from the cluster.**
- **For the Web application we will be using the load balancer. This helps us to keep the web servers scalable as and when required based on the traffic on the web application**
- **With the Hadoop 2.0 fault tolerance of the system is also now addressed. With an automatic failover along with RAS Enterprise Namenode servers**
- **Load Balancer with multiple web servers also helps keeping the web application fault tolerant.**