virtusa

Business Cipher Challenge Season 4

Home Credit Default Risk

'Modelling a system for financial inclusion of unbanked population'

TEAM 15

Upamanyu Mukherjee, *Data Analyst*

Susmitha H, *Data Analyst*

Surabhi Sawant, *Business Analyst*

Sri Visagan V, *Business Analyst*

SYMBIOSIS CENTRE FOR INTERNATIONAL EDUCATION

IIM KASHIPUR

JBIMS Jamnalal Bajaj Institute of Management Studies

IIM TRICHY

Case at a Glance

- Bringing financial inclusion for unbanked population with non-existent credit history
- Providing positive & safe experience for customers against untrustworthy lenders
- Analyzing the risk associated with each borrower & making educated decisions
- Identifying & acquiring customers with repayment abilities

Towards Financial Inclusion

$ 6.9 trillion	Global Lending Market *(Est. 2021)*	**36.9%**	Household Lending Market	**$ 1 trillion**	Potential Value Creation by AI in banking
14.8%	Compounded Annual Growth Rate	**52.3%**	Of household lending, Home Loans Market	**1.52**	Mobile Connections per user, globally
35%	**Western Europe** Largest Region	**4.2%**	CAGR for Personal Loans, Fastest Growing Segment	**59.5%**	Internet Penetration Rate, globally

Market Estimates

1.7 Bn	Adult unbanked population globally
190 Mn	**India** ,Adult unbanked population (Developing nation)
14.1 Mn	**US,** Adult unbanked population (Developed nation)
3 in 10	Unbanked adults between the ages of 15 and 24

Consumer Persona

Laborer wanting to purchase a motorbike

30- 35 yrs

- Married, single child
- Educated till secondary level
- Owns house

Sales Person looking for a home loan

25- 35 yrs

- Unmarried, working in a factory
- Completed Graduation
- Does not own a house

Driver wanting to purchase home appliances

40- 55 yrs

- Married & has dependents
- Educated till secondary level
- Owns house & second-hand car

Dataset Overview

application_train/ application_test

Details about each loan application
Main table broken into Train & Test, 1 row for 1 loan

bureau

All previous client credits provided by other financial institutions that have been reported to Credit Bureau

bureau_balance

Monthly balances of previous credits in bureau
One row for one month of a previous credit

previous_application

Data of previous applications for client loans who have loans in the application data

POS_CASH_balance

Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had

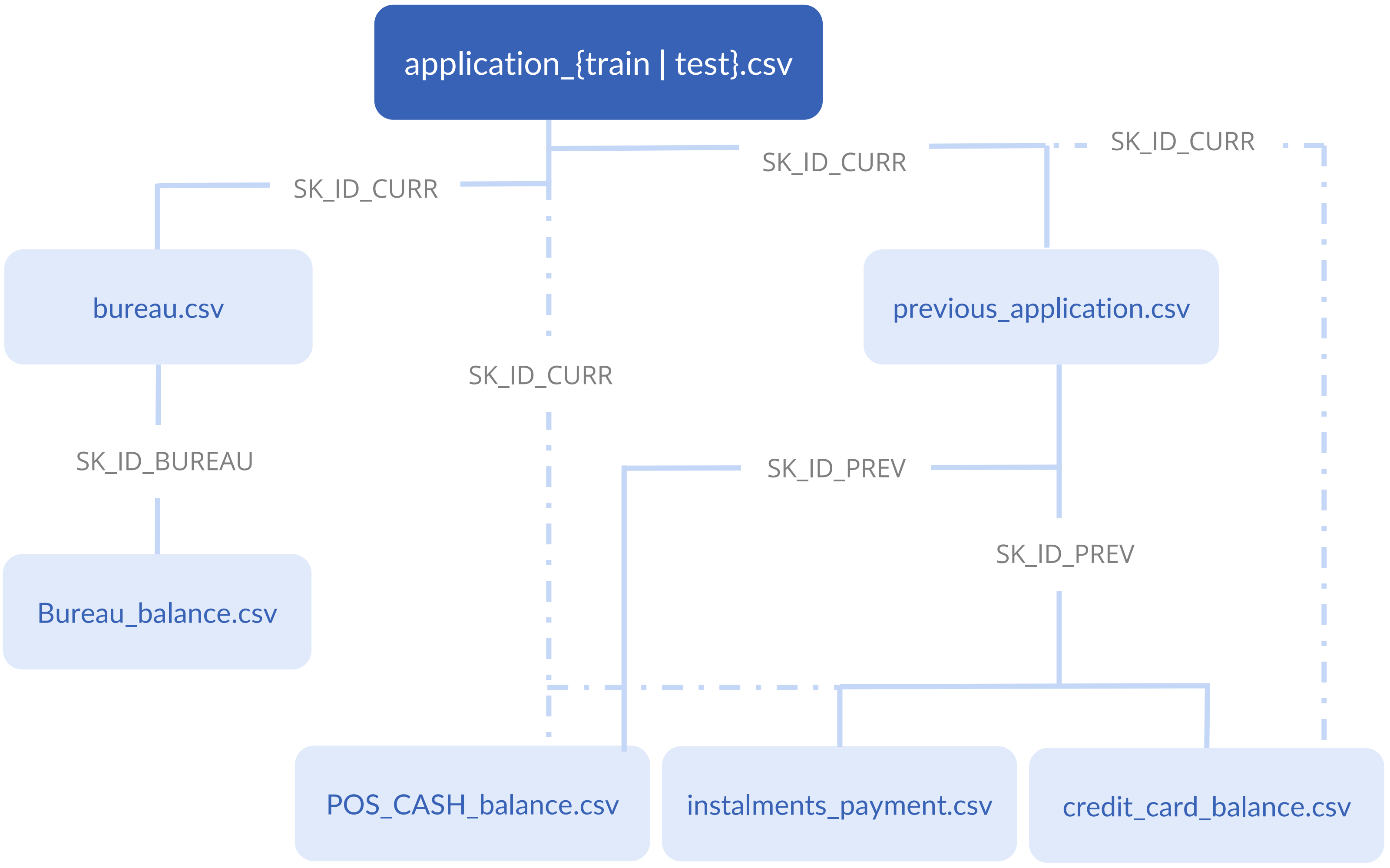
credit_card_balance

Monthly data about previous credit cards clients have had with Home Credit

instalments_payment

Data of payment history for previous loans at Home Credit, 1 row for every made & 1 for missed payment

Connecting Data Sources



Objectives

- The main objective is to identify the potential Defaulters based on the given data about the applicants.
- The probability of classification is essential because we want to be very sure when we identify someone as a Non-Defaulter, as the cost of making an error can be immense to the organization.

Performance Metric

- **ROC-AUC Score:** It works by ranking the probabilities of prediction of the positive class label and calculating the Area under the ROC Curve.
- **Confusion Matrix:** The confusion matrix helps us to visualize the mistakes made by the model on each of the classes, be it positive or negative. Hence, it tells us about misclassifications for both classes.

Constraints

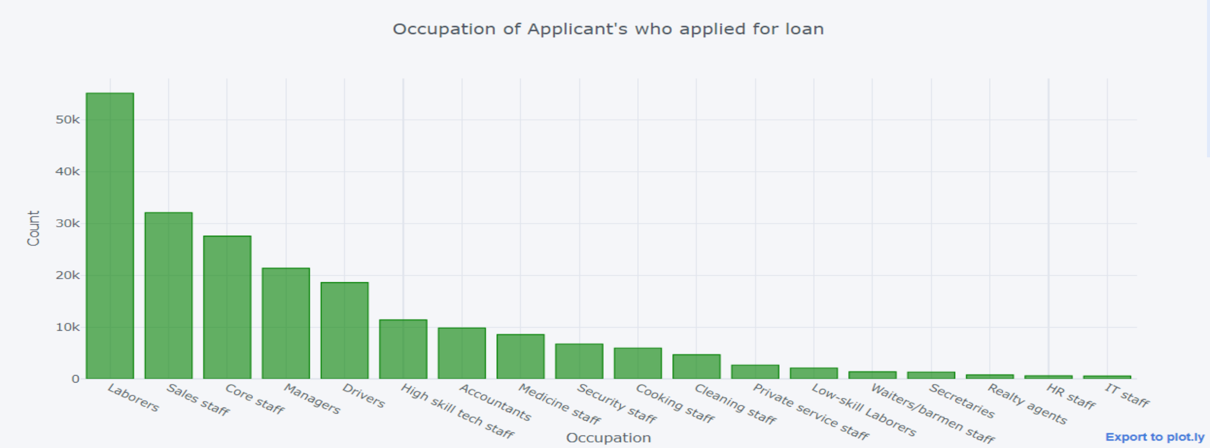
- No strict latency constraints.
- Predict the probability of capability of each applicant of repaying a loan.
- The cost of a mis-classification is very high.
- Interpretability is partially important.

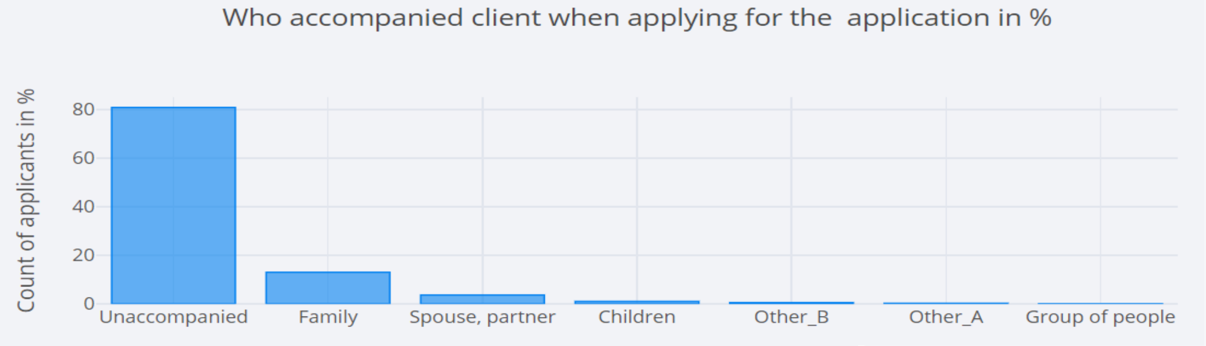
Key Observations from EDA

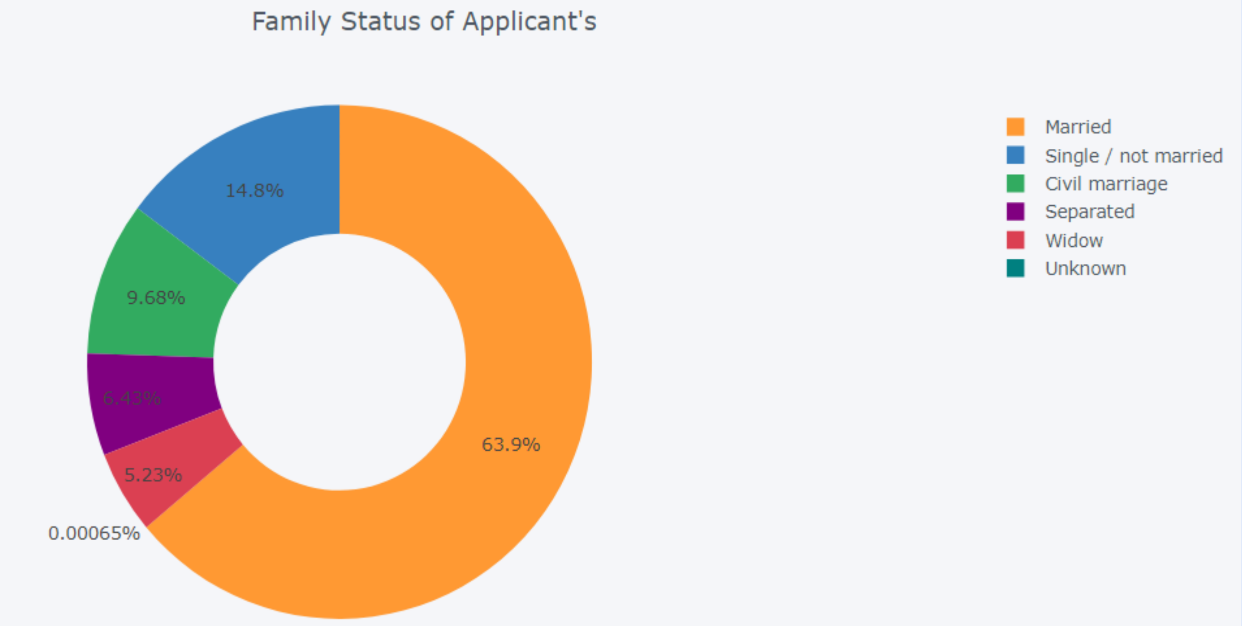
%/ Count	Consumer Characteristics
91.90%	People haven't repayed loan
44.80%	Cash loans
43.70%	Consumer loans
82%	Unaccompanied
12%	Family
90.50%	Cash loans
66%	Don't own a car
69.40%	Owns realty
51.60%	working
23.30%	Commercial associates
18%	Pensioners
63.90%	Married
14.80%	Single/ Not married
56K	Labourers
32K	Sales Staff
28K	Core Staff
71%	Secondary
24.30%	Higher Education
88.70%	House/ Apartment

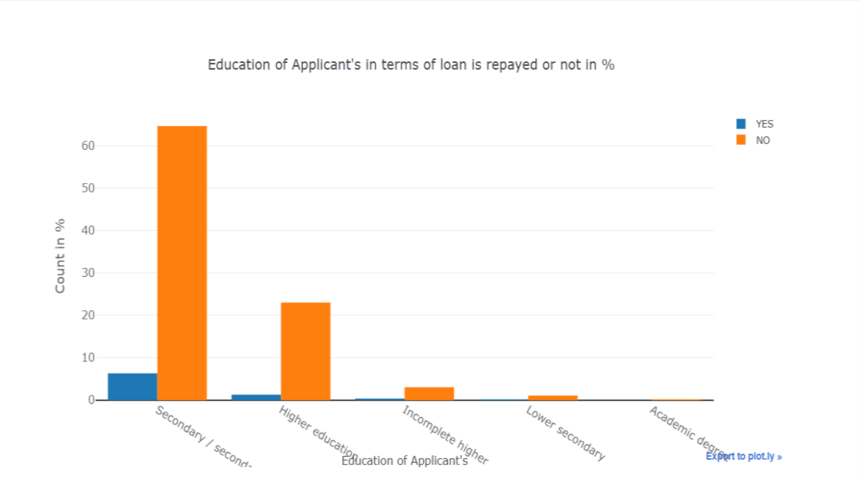
%/ Count	Consumer Characteristics
4.83%	with parents
68K	Business Entity Type 3
56K	XNA
38K	Self-employed
46%	Working- loan repaid
5%	Working- loan not repaid
43%	Credit & Cash Offices
17.10%	Cash
66.70%	Did not request for insurance during the previous application
160	Repeat Clients
155	New Clients
98	Repeat Clients
10	New Clients

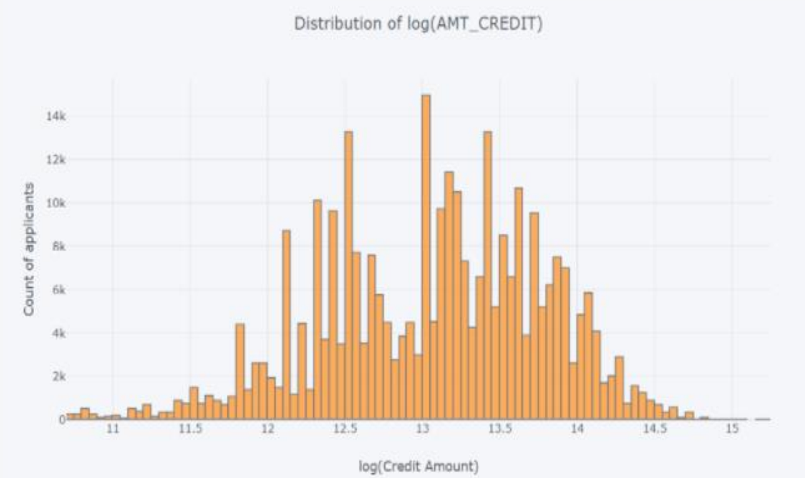
Exploratory Data Analysis

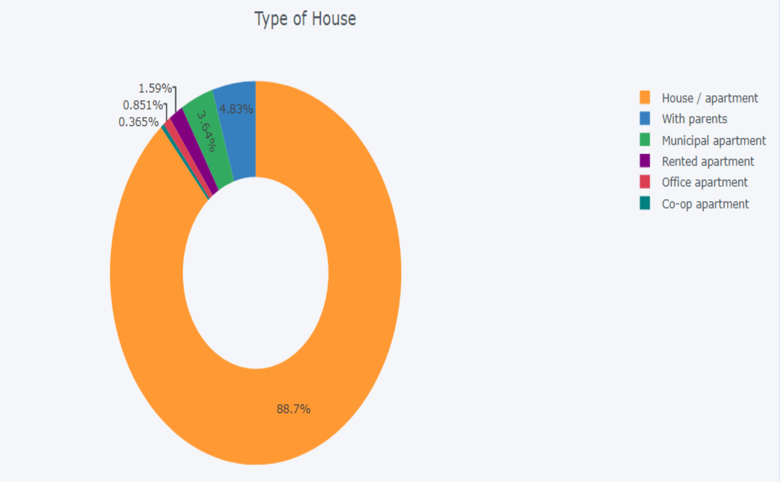












Ensemble Models

- Bagging Algorithms
- Boosting Algorithms

Why Bagging

- It reduces complexity of the model that overfit the training data
- When features size is too large model suffers from the "Curse of dimensionality"

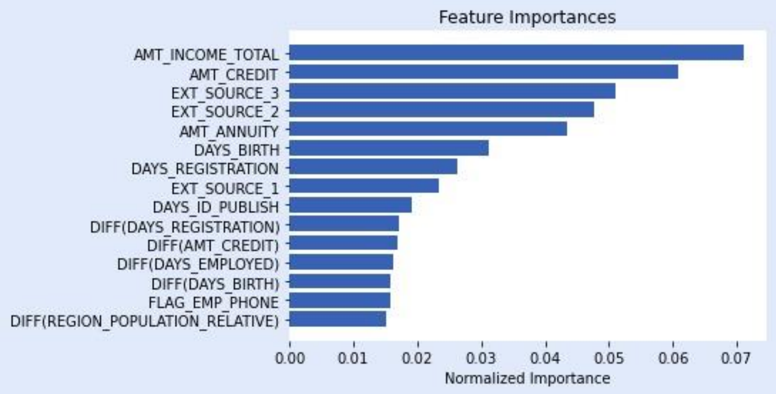
  E.g., Random Forest

Why Boosting

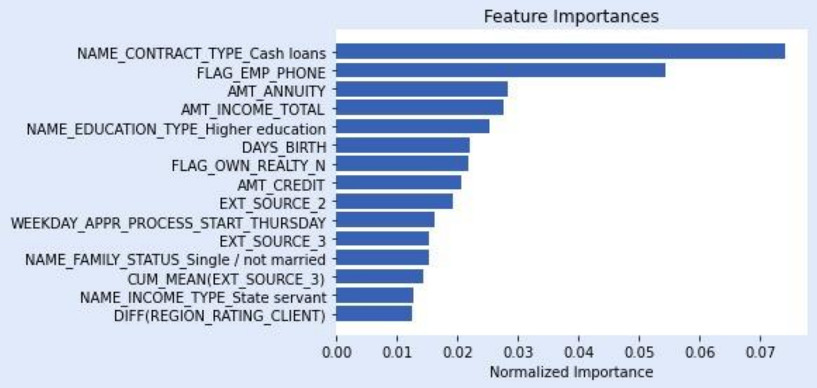
- Boosting algorithms are used when the training data is underfitting
- When feature size is too small, it ends up with higher variance

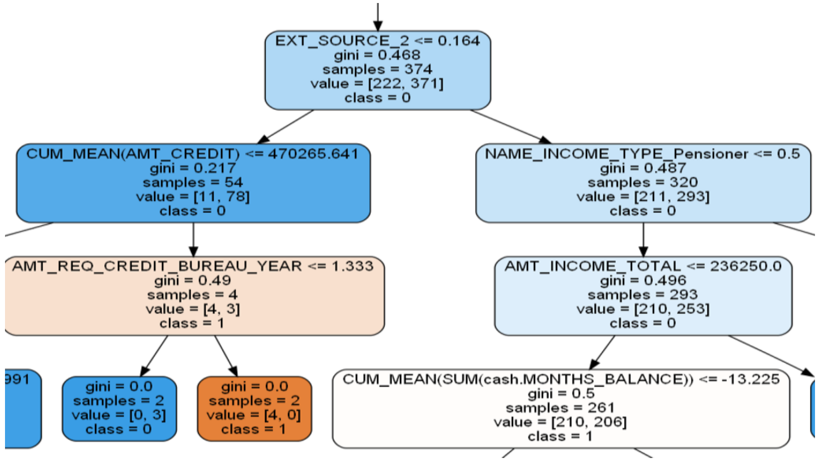
  E.g., XGBoost

LightGBM



XGBoost





Model	Acc	Rec	F1	Prec	Spec
LightGBM	0.77	0.74	0.76	0.79	0.80
RandomForest	0.74	0.76	0.74	0.73	0.73
XGBoost	0.74	0.74	0.74	0.74	0.75

Model Selection & Evaluation

- LightGBM
- Random Forest
- XGBoost

Analysis

- LightGBM scores higher in most of the parameters
- Both LightGBM & XGBoost models predict the decision more accurately even for random validation dataset
- XGBoost is selected considering the feature importance & consumer persona

Recommendations

- Models can be made in deep neural networks with the help of algorithms like LIME etc., considering the high-dimensionality of dataset
- Developing model considering various job-types
- To reduce total cost of cloud, trained models can be used for use cases defined

Virtusa Business Cipher Season4

Date of birth

dd-mm-yyyy

DAYS_BIRTH

Highest education

Secondary / secondary special

NAME_EDUCATION_TYPE

Marital status

Married

NAME_FAMILY_STATUS

Gender ○ Female ○ Male

CODE_GENDER

Annual income

Enter your income

AMT_INCOME_TOTAL

Have work phone ○ Yes ○ No

FLAG_EMP_PHONE

Thank You!