Leveraging ML for Credit Default Prediction

11.03.2024

HOME CREDIT

Team Members

Riddhi Dhameliya

Reza Azarang

Sohail Zafar

Content

Agenda

1. Importance of accurate credit risk assessment

   The challenges faced by traditional methods.

2. ML - based approach

   Business Goal - Data Understanding

3. Technical Aspects of Our Model

   Real world example.

4. Credit Risk App Calculator

   live.

Credit Risk

Credit On Risk Or Not ?

Objectives

1. **Purpose of Assessment**
   Borrower will default.
2. **Risk Management**
   Determine appropriate interest rate on individuals.
3. **Impact on Financial Stability**
   Extend Credit on Individuals more likely to default.
4. **Regulatory Compliance**
   Financial institutions follow regulations.

Loan Repayment

- Borrower will be added into goodwill list.
- Credit limits got extended.
- Less interest rate.

Traditional Method

Challenges!

- Limited Predictive Power
- Reliance on Historical Data
- Subjectivity
- Inflexibility
- Biasness

ML Based Approach

1. Leverages ML Algorithms
2. Analyze vast amount data
3. Identify patterns
4. ML predictions

Business Goal

Goal

Home credit is an international consumer finance provider focusing on responsible lending primarily to people with little or no credit history.

Our goal is to make a model which can predict default of clients based on internal and external information that are available for each client.

As we have classification problem so here our goal is to improve AUC metric.

Data Understanding

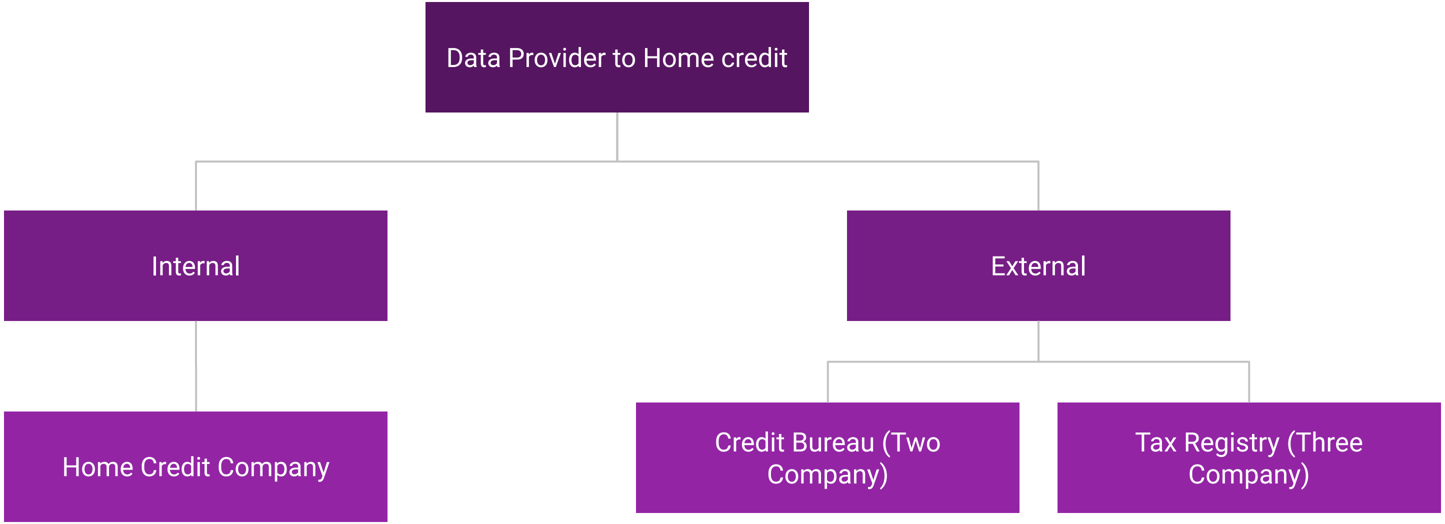
EDA

Data Reference:

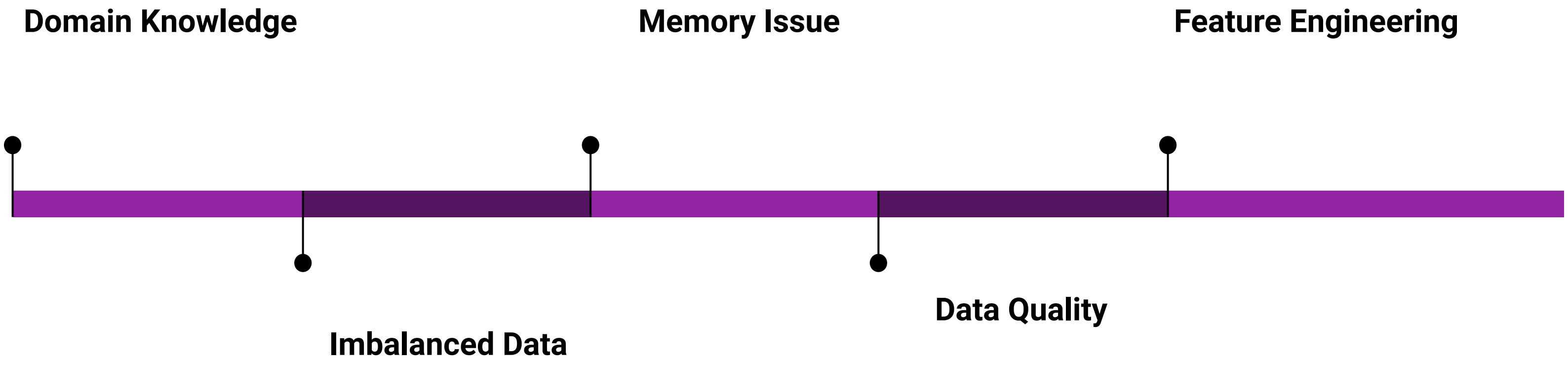
<u>Home Credit - Credit Risk Model Stability</u> | Kaggle

Clients	1.5 Million
Features	467
Numerical features	56%
Categorical features	44%
Missing Values	63%

Data Source:


Data Provider to Home credit
Internal
External
Home Credit Company
Credit Bureau (Two Company)
Tax Registry (Three Company)

Challenges


Domain Knowledge
Memory Issue
Feature Engineering
Imbalanced Data
Data Quality

Defining Use Cases

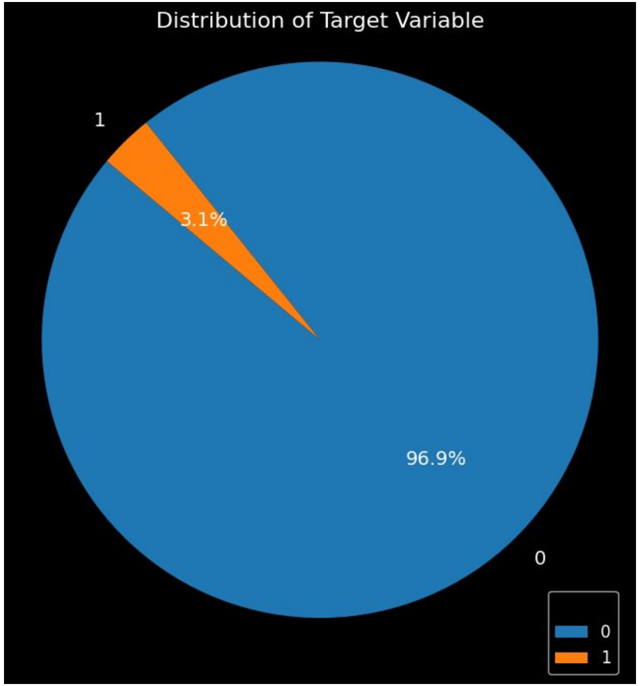
Case 1

A person having a Credit History.

Case 2

A person having no Credit History.

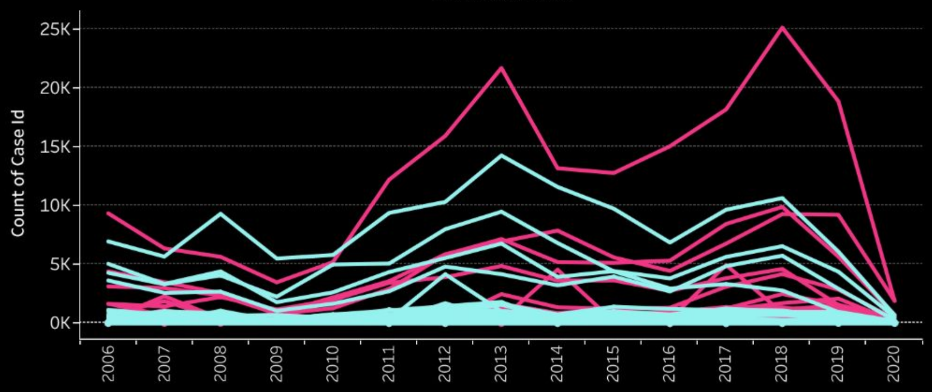
Distribution of Target



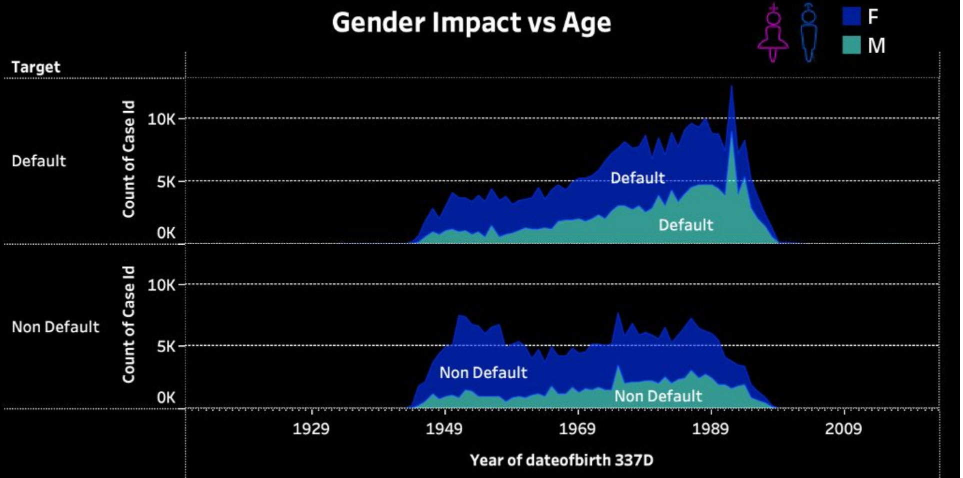
0 = Non Default
1 = Default

Tableau Dashboard

Interestrate Impact On First Due Date

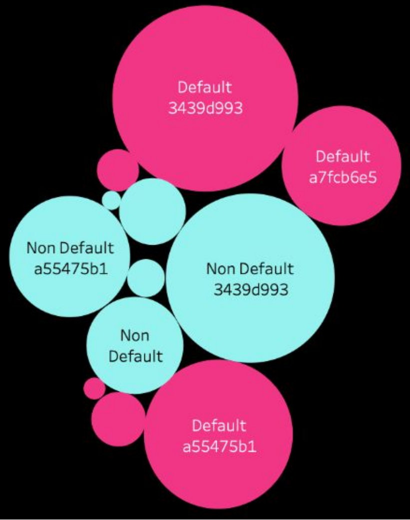


Gender Impact vs Age

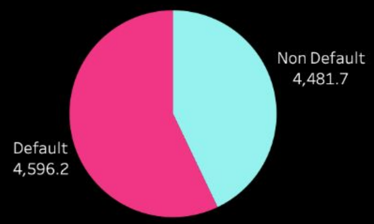


Total Debt
$ 53.16 B

Marital Status



Monthly Annuity



Income type Impact

Count of Case Id: 1.201 – 105.289

incometype 1044T	Target: Non Default	Default
EMPLOYED	48.283	77.148
HANDICAPPED_2	1.389	1.985
HANDICAPPED_3	1.201	1.572
OTHER	2.209	1.453
PRIVATE_SECTOR_EMPLOYEE	74.679	105.289
RETIRED_PENSIONER	89.704	53.200
SALARIED_GOVT	59.925	72.128
SELFEMPLOYED	3.316	4.668

Modeling

Predict that if the user will be default or not!

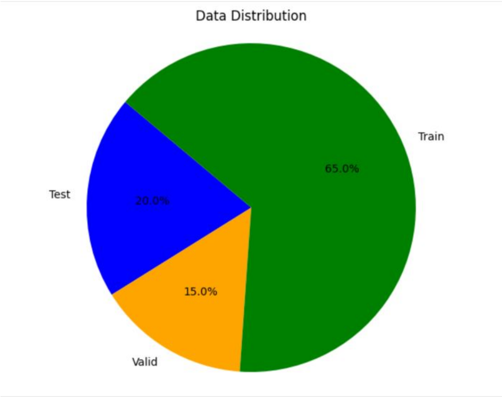
Data

Data was the tough one!

- A lot of features (~500)
- A lot of Categorical Features (~200)
- with one-hot-encoding (~90,000)
- A lot of System Memory required (400+GB)
- A lot of Missing Values

Splitting Data:

- Train Data: 65%
- Validation Data: 15%
- Test Data: 20%



Metrics

Accuracy

Accuracy = How many percent we are predicting correctly.

But here not working!

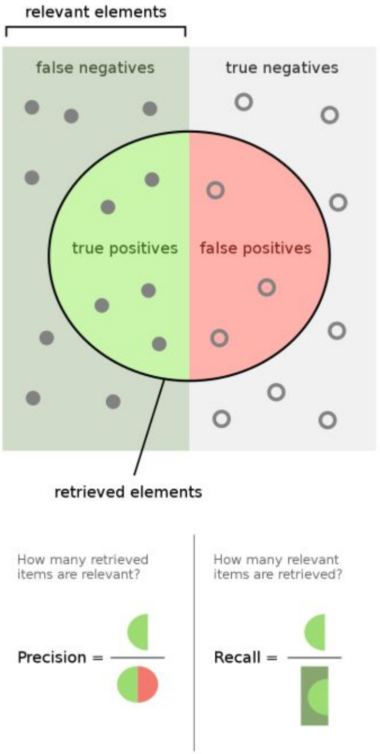
Saying always NO is ~97% accurate!

- Using AUC instead!

AUC

Tolerant against imbalanced data.

- Precision (quality)
- Recall (quantity)
- F-Score (both)



- high precision/quality:
  - If the model is saying YES, it is trustful
- high recall/quantity:
  - The model finds YESes as much as possible
- high certainty:
  - When is saying YES, is very sure about it, if is saying NO, is very sure about it.

Computer Perception

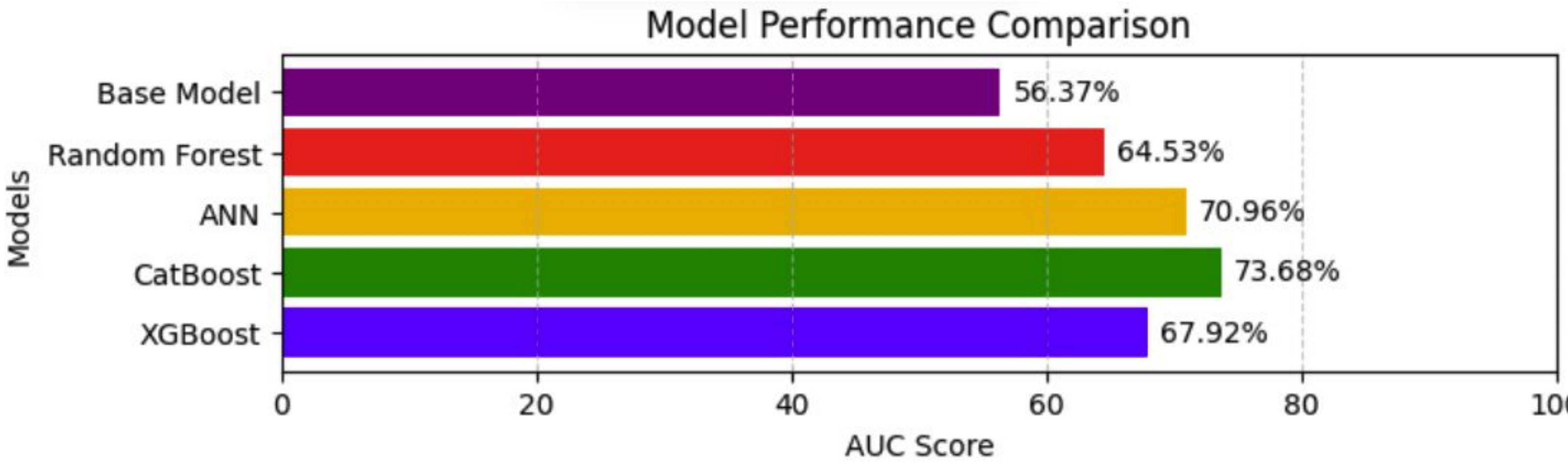
- Most important categorical features by Cramer's:

cols	importance	cols_desc
min_name_4527232M	1.41	Name of employer.
max_incometype_1044T	1.39	Type of income of the person
min_incometype_1044T	1.39	Type of income of the person
max_name_4527232M	1.36	Name of employer.
min_registaddr_zipcode_184M	1.26	Registered address's zip code of a person.
min_relationshiptoclient_642T	1.21	Relationship to the client.
min_relationshiptoclient_415T	1.21	Relationship to the client.
riskassesment_302T	1.19	Estimated probability that the client will def...
requesttype_4525192L	1.18	Tax authority request type.
min_contaddr_zipcode_807M	1.14	Zip code of contact address.

- Most important numerical features by PCA:

cols	importance
DPD of client with tolerance.	3.2
Number of instalments paid before due date in ...	2.8
Monthly annuity amount.	2.6
Next month's amount of annuity.	2.0
Number of applications associated with the sam...	2.0
Number of applications made by the client in t...	1.9
Number of applications associated with the sam...	1.8
Number of applications made in the last 30 day...	1.6
Number of applications with the same employer ...	1.5
Number of applications associated with the sam...	1.4

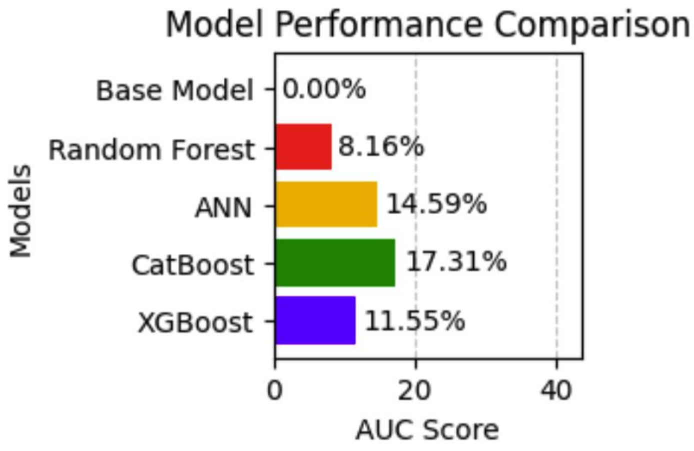
Results



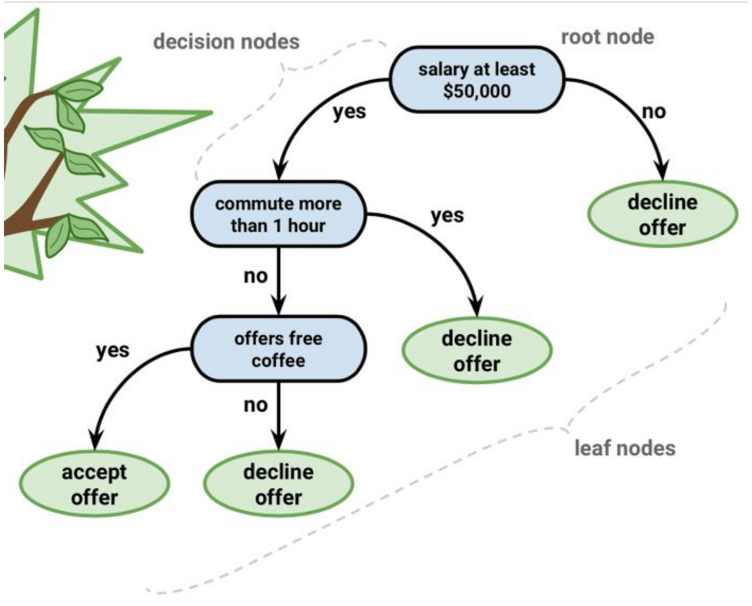
Model Performance Comparison
Base Model
Random Forest
ANN
CatBoost
XGBoost
56.37%
64.53%
70.96%
73.68%
67.92%
Models
0
20
40
60
80
100
AUC Score

Baseline Model & Comparison

- A very basic model which we can create in a short time.
- Required to compare the improvement between models.

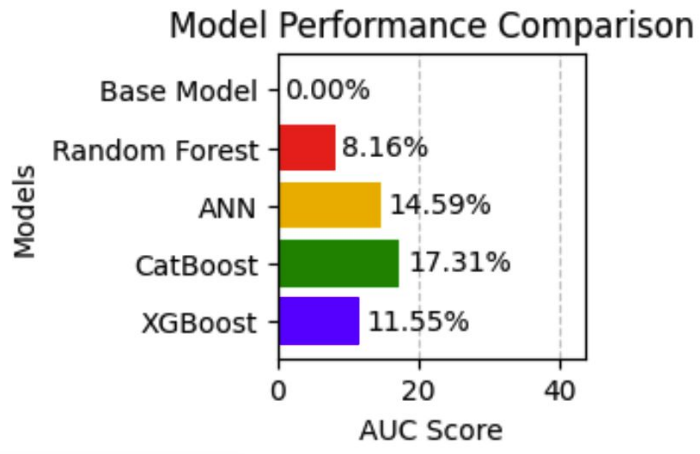


The Base Line here is a Decision Tree.



Cat Boost

- ➔ Handling Categorical Features
- ➔ Very fast, also Support for GPU
- ➔ Robustness to Overfitting

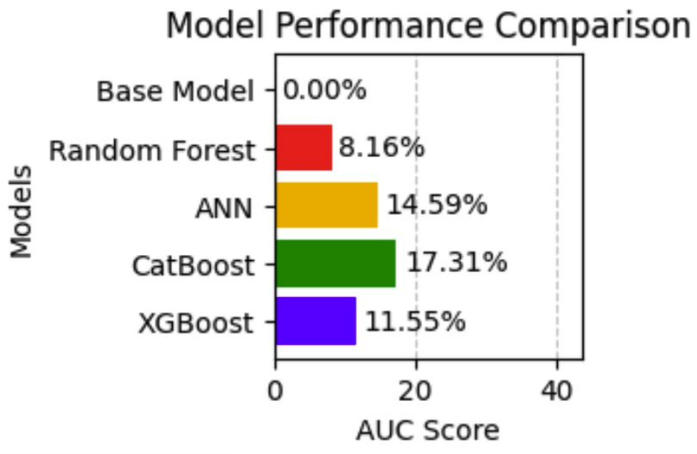


- over 17% better than base model
- nearly 74% AUC score

ANN

➔ Good with some data
➔ But not this data!

- over 14% better than base model
- But not as good as CatBoost!



Future Works

- Ensemble
- Combining different models

Thank You