HOME CREDIT

# Credit Default Risk Analysis

Team 5 : Weifu Shi, Xiaorui Shen, Sizhe Fan, Yinghui Wei, Hang Zhang

# Overview of Presentation

1. Introduction
2. Dataset & Data Cleaning
3. EDA & Feature Selection
4. Model Selection & Results
5. Challenges (if needed)
6. Conclusion & Summary

# Executive Summary

**Dataset:** Home Credit Default Risk from Home Credit group

- International Non-Bank financial institution which operates in 10 countries ; Served over 124 million customers.
- Headquarter: Amsterdam, Netherlands.
- Total Assets: 25 billion euro dollars.

**Project Goal**:

- Whether the bank should accept the loan application or not?
- How capable of each applicant to repay a loan?
- What are the characteristics of those applicants with low debt paying ability?

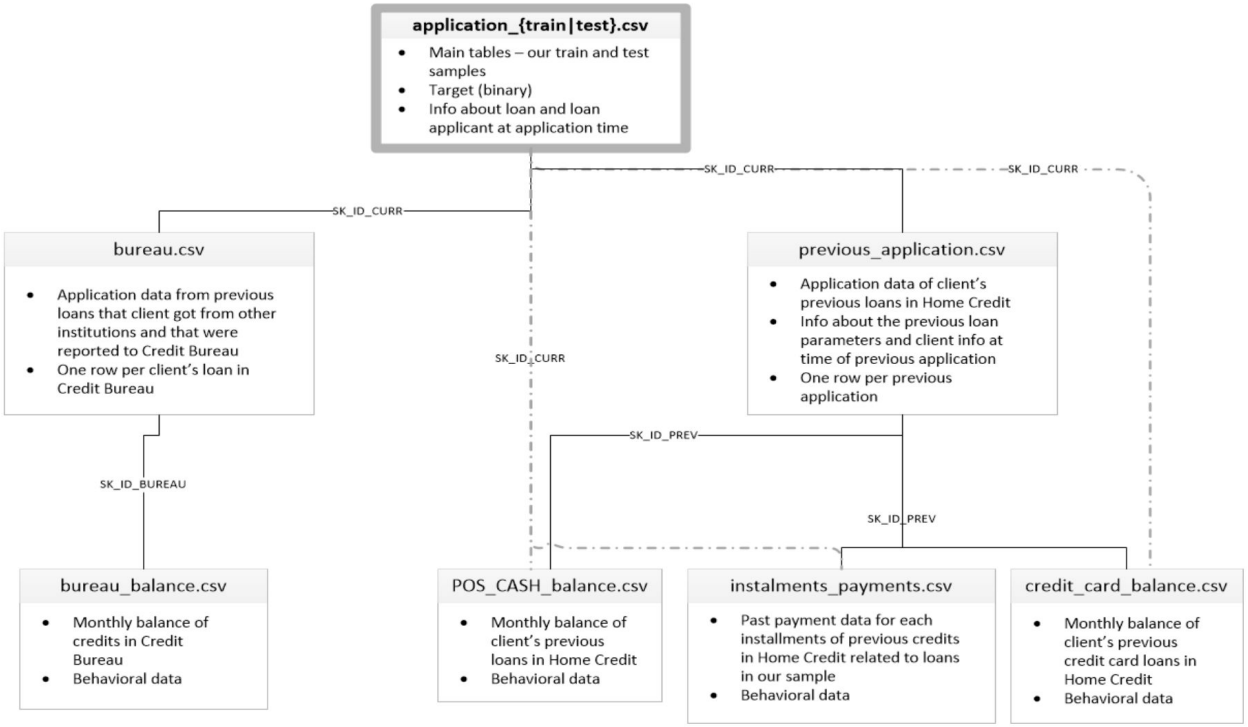
**Significance:**

- We will ensure that clients who are capable of repayment are not getting rejected.
- We will also provide a solvency benchmark for loan applicants.

# HomeCredit Mission Statement

" Our service are simple, easy and fast. Our responsible lending model empowers underserved customers with little or no credit history to access financing, enabling them to borrow easily and safely, both online and offline" --- HomeCredit

# Dataset




Train Dataset: 307511 rows  122 variables.
Other six files: 100 variables.
Two Main Types of Variables:

1. Users' Attribute (most variables)
    a. Family Status, Income, Education, Occupation, etc.
2. Loan's Attribute
    a. Credit, Annuity, Previous Application, Purpose of Loan, etc.

# Business Objective

1. Prediction
   - To predict the loan repaying capability of each applicant 0/1. To find out those who have difficulties of repaying the loan.
   - By studying the characteristics of applicants who can and who can't afford to pay back the loan, we can provide a helpful guidelines/advices for them to increase the repaying ability.
2. Customer Segmentation
   - Cluster these applicants into different groups.
   - Study the repaying capability of each group and provide a risk score for each group.
   - Home Credit can provide different levels of services/supervision/alarm settings to different groups.

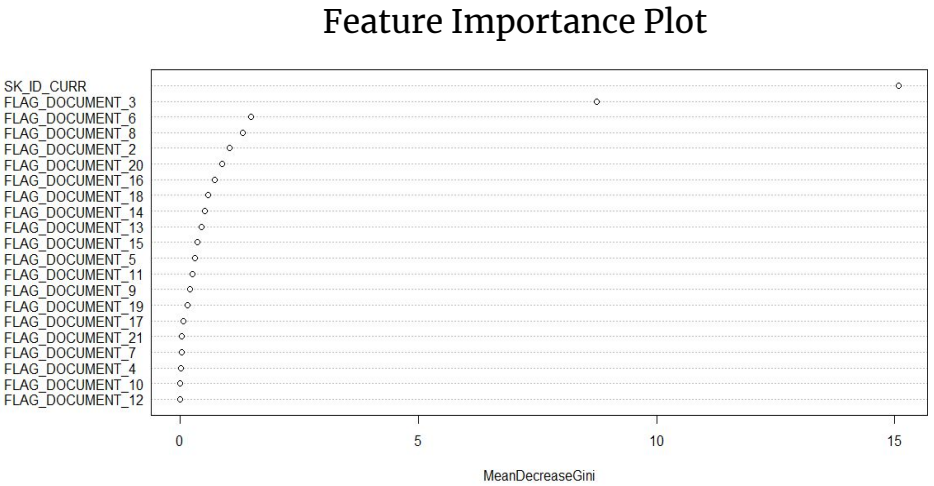
## Dashboard Links

# Data Pre-processing

1. Features selection:
   - Screening out the most relevant variables that can maximize the accuracy of the prediction algorithm
   - Filtering out the most irrelevant variables that might create bias and variance to the algorithms.
2. Decision on Missing Values
   - Replace N/A with 0 (except non numeric columns)
   - Combine “documents 1-21” into “Number of Document Submitted.
3. Data Cleaning:
   - Missing value > 50%
   - 0 variance
   - Highly correlated (correlation > 80%)
   - Fill left numeric NAs with 0
4. Data imbalance:
   - Under-sampling with ratio 1:2
5. Data Transformation:
   - Box-Cox Transformation: 2 variables are normalized: AMT_INCOME_TOTAL (lambda=-0.1), AMT_CREDIT(lambda=0.2)

# Use random forest to help choose features

- Find out the most important document among 21 of them.



## Support Vector Machine (accuracy: 70%)

- Builds a non-probabilistic model that assigns new examples to one category or the other.
- Efficiently performs a non-linear classification implicitly mapping their inputs into high-dimensional feature spaces.

## Logistic Regression (accuracy: 68%)

- Uses logistic (sigmoid) function to find the relationship between variables.
- Linear regression is not suitable for classification problem because it's unbounded, and this brings logistic regression into picture.

# Logistic regression vs SVM

```
Call:
summary.resamples(object = resamp)

Models: SVM_Radial, SVM_Linear, logistic
Number of resamples: 10

Accuracy
               Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM_Radial 0.6733333 0.6904167 0.6925000 0.6920000 0.6991667 0.7083333    0
SVM_Linear 0.6866667 0.6975000 0.7016667 0.7038333 0.7120833 0.7200000    0
logistic   0.6816667 0.6995833 0.7033333 0.7025000 0.7045833 0.7200000    0

Kappa
               Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM_Radial 0.1575931 0.1787391 0.2007194 0.1958759 0.2117758 0.2335766    0
SVM_Linear 0.1988636 0.2145520 0.2393827 0.2344196 0.2518538 0.2631579    0
logistic   0.2118294 0.2484155 0.2515677 0.2510291 0.2539336 0.2921348    0
```

Use SVM because of its highest accuracy

## SVM results

Sensitivity = 0.21

Specificity = 0.94

```
     1    0
1  839  427
0 3161 7573
```

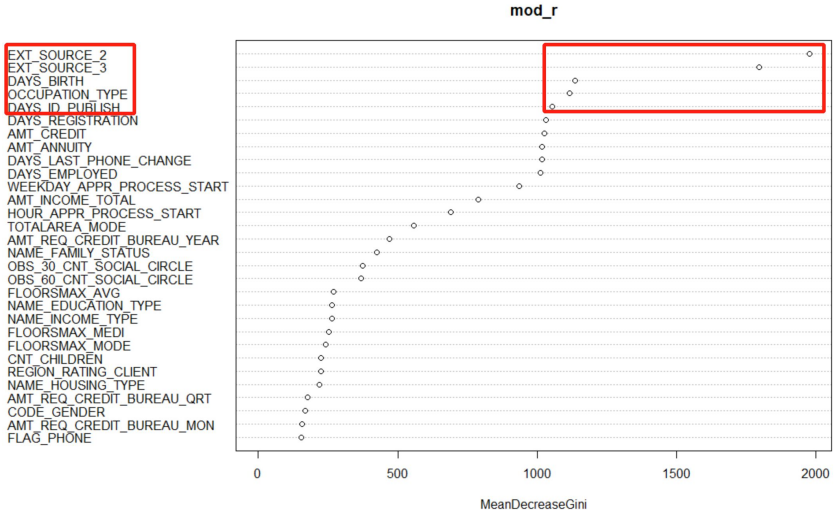
```
    1     0
 1762 10238
> 10238/(1762+10238)
[1] 0.8531667
```

## Random Forest (accuracy: 66%)

- Use a 1:1 ratio of different classes as train
  - Running capability
  - Extreme imbalance of dependent variable
- Train model class error
  - 0-34%, 1-31%
- Recall rate 70%
  - True positive/(true positive + false negative)
- Specificity 66%

# RF cont.

## Feature importance plot

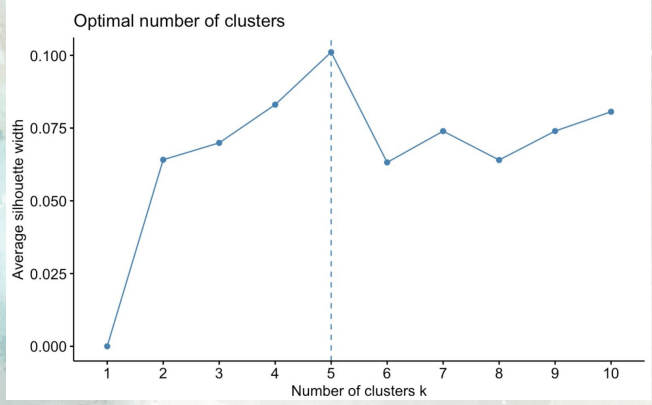



- Predict test data
  - 2727 applications may have repay difficulty (3927 in fact)
  - Others have the repay ability

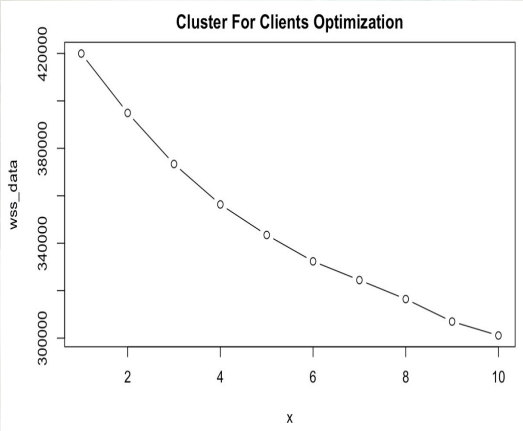
# Customer Segmentation & K-Means Clustering

- We randomly sample 10,000 customers and implement K-Means unsupervised ML algorithms. In terms of Silhouette and WSS method, The optimal number of cluster would be five.
- As shown above, HomeCredit Group should set different levels of alarm call. For example, Cluster 3 is considered as high risk group and should be alerted immediately: Old age, low income and low employment day.

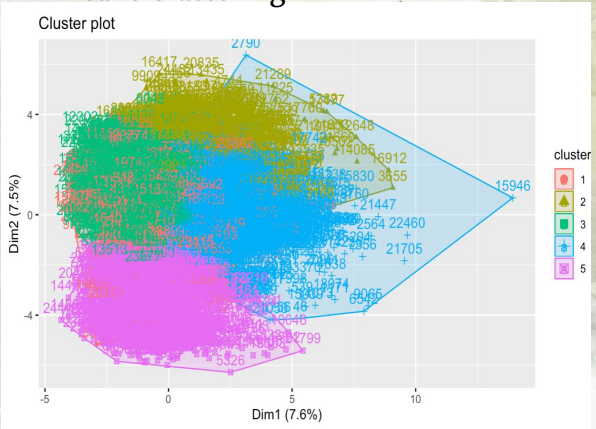
## WSS Method




## Silhouette Method




## K-Means Clustering




## Cluster Summary Table

| cluster <int> | count <int> | age <dbl> | mean_income <dbl> | mean_credit <dbl> | mean_annuity <dbl> | mean_day_employed <dbl> | mean_goods_price <dbl> |
|---|---|---|---|---|---|---|---|
| 5 | 2433 | 41.98440 | 215909.3 | 928289.2 | 39066.20 | 2258.77928 | 823085.9 |
| 2 | 548 | 37.60316 | 208505.4 | 595935.2 | 29658.20 | 1308.31387 | 528488.5 |
| 4 | 1425 | 38.27470 | 151630.1 | 498098.5 | 24791.55 | 1814.84912 | 434755.4 |
| 1 | 4480 | 36.46601 | 135654.4 | 380274.9 | 20961.68 | 1565.54442 | 328432.7 |
| 3 | 1114 | 58.97523 | 129303.2 | 512812.8 | 22217.96 | 53.34022 | 451454.9 |

# Conclusion and future work

For now:

- Two supervised models
- One unsupervised model.
- They have their own strengths in different directions.

For the future steps

- build a final model that integrates all the advantages of the model to achieve better results

# Thank you!

MSBA Cohort A TeamS
Weifu Shi, Xiaorui Shen, Sizhe Fan, Yinghui Wei, Hang Zhang