Machine Learning to Predict Loan Defaults

December 4th, 2021

Dakshin Kannaan, Siddharth Iyer, Santino Luppino, Zheng Li, Dimitri Niles

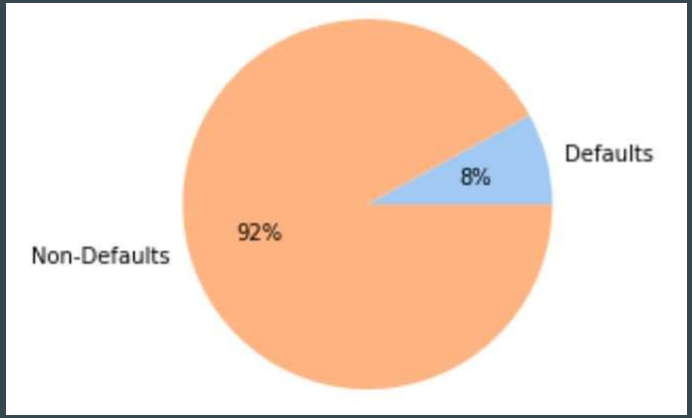
Overview

- Objective: Develop machine learning methods to forecast loan defaults for risk optimization problems
- This is a classification problem
- Models:
  - Clustering
  - Random Forest
  - Support Vector Machine
  - Neural Network
  - Logistic Classification

Exploratory Data Analysis

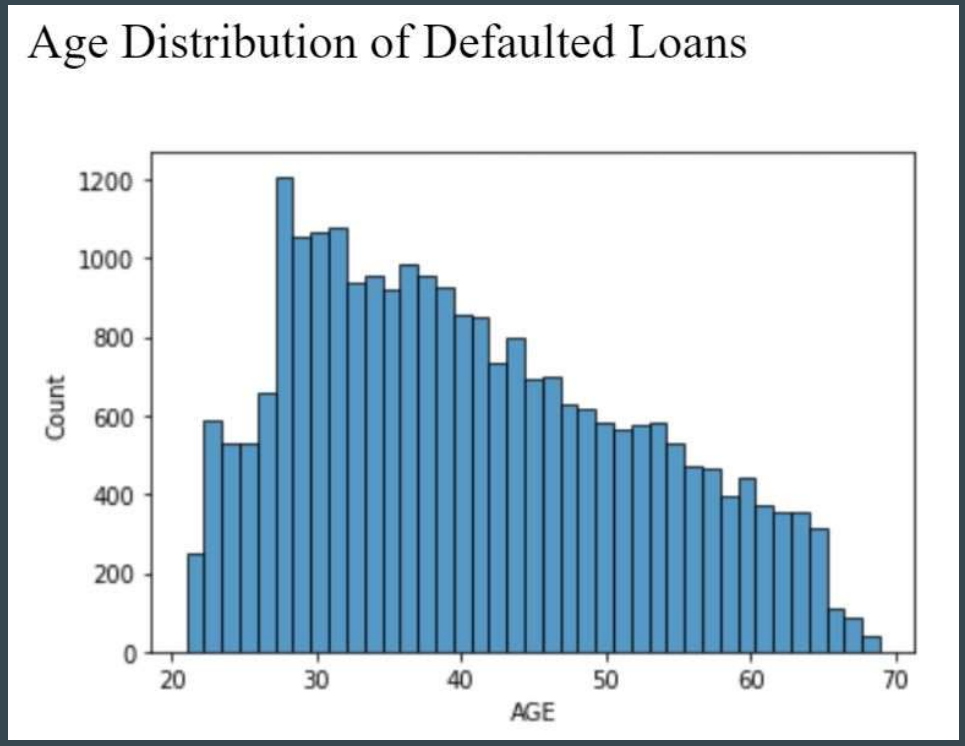
- Reduced predictors from 122 to 15
  - Then using PCA reduced further
- Many columns seemed irrelevant
  - SK_ID_CURR held the IDs of each loan
  - FLAG_DOCUMENT columns
  - FLAG_MOBIL or FLAG_EMAIL

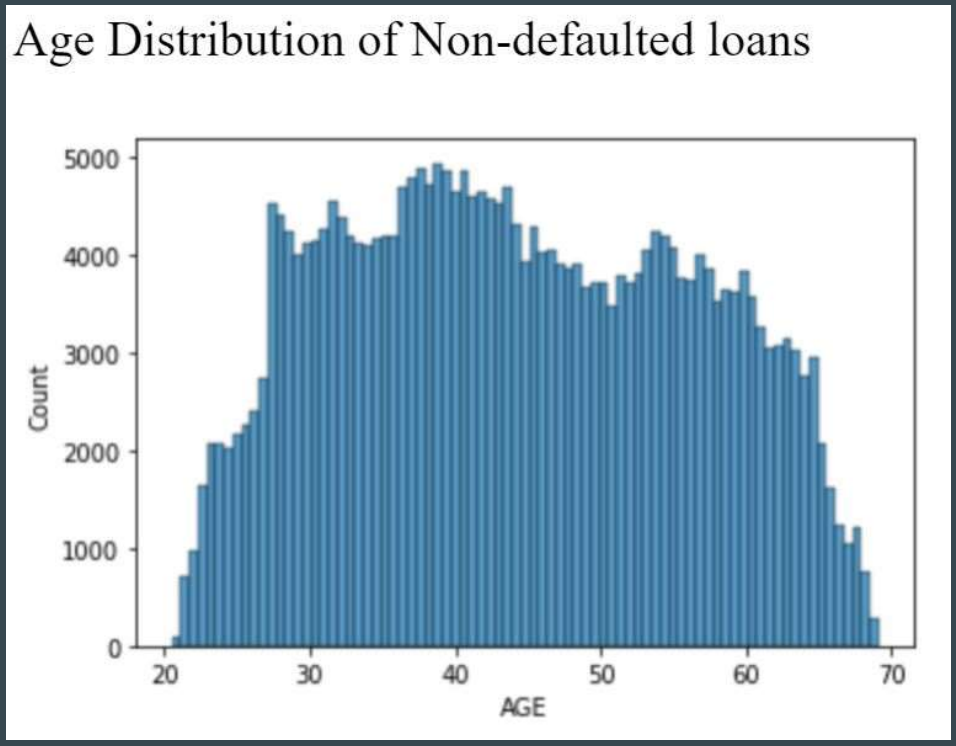
Very unbalanced dataset...model will try to classify everything as Non-default



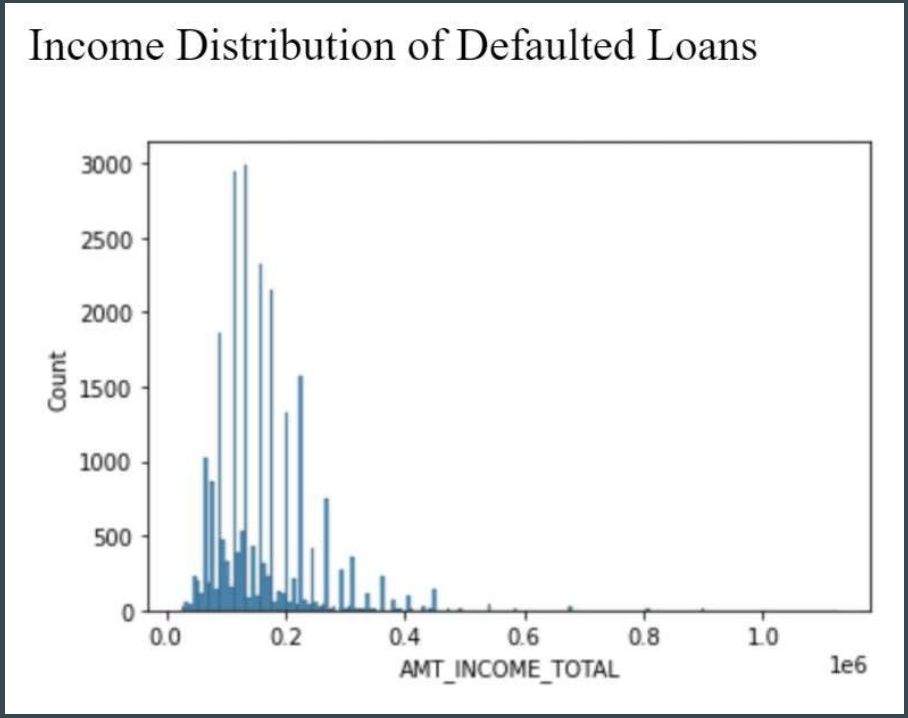
TARGET	NAME_CONTR	CODE_GENDEI	FLAG_OWN_C	FLAG_OWN_R	CNT_CHILDRE	AMT_INCOME	AMT_CREDIT	AMT_ANNUITY	NAME_TYPE_S	NAME_INCOM	NAME_EDUCA	NAME_FAMILY	NAME_HOUSI	REGION_POPU	AGE
1	0	1	0	1	0	202500	406597.5	24700.5	6	7	4	3	1	0.018801	25.920548
0	0	0	0	0	0	270000	1293502.5	35698.5	1	4	1	1	1	0.003541	45.9315069
0	1	1	1	1	0	67500	135000	6750	6	7	4	3	1	0.010032	52.1808219
0	0	0	0	1	0	135000	312682.5	29686.5	6	7	4	0	1	0.008019	52.0684932
0	0	1	0	1	0	121500	513000	21865.5	6	7	4	3	1	0.028663	54.6082192
0	0	1	0	1	0	99000	490495.5	27517.5	5	4	4	1	1	0.035792	46.4136986

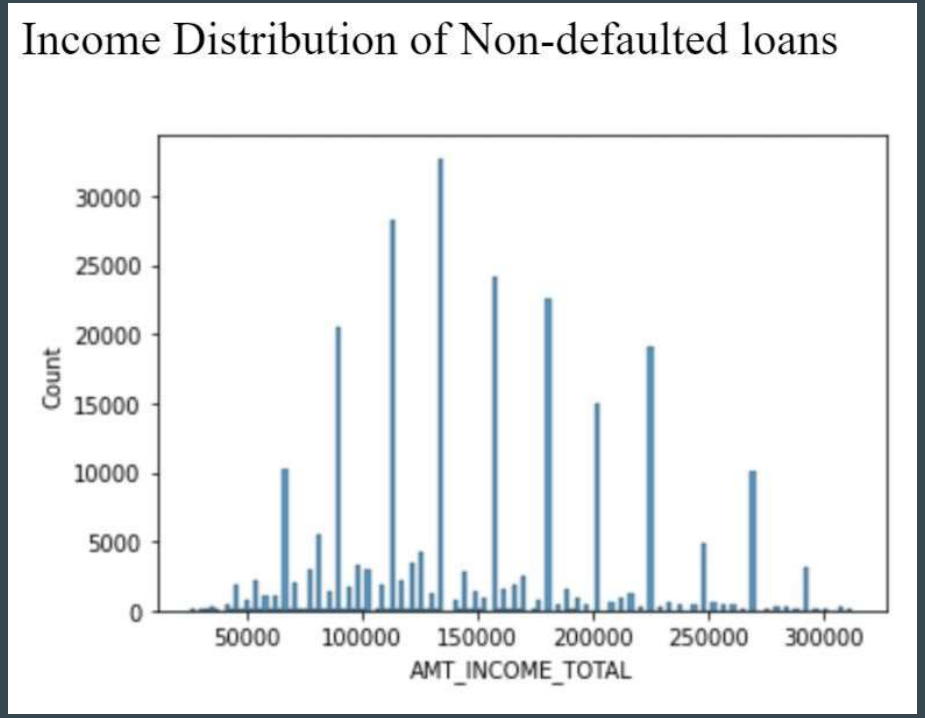
Age Distribution





Income Distribution





Are women better? Yes

Men are more likely to default

Male Default Rate: 0.10166590161426058

Female Default Rate: 0.07004736280903613

ANNUITY PAYMENT/AVAILABLE CREDIT:

Male: 0.05256824575168550 5

Female: 0.05599976395069611

K-Means Clustering

- Both clusters have most of the data concentrated into a singular cluster
- The two different methods of scaled and not scaled do not have a significant effect
- Some of the features that are included in cluster one are Children Count, Realty Status, and Family Status
- Age stood alone as its own cluster in both methods

Not Scaled:

data.NAME_CONTRACT_TYPE 1	data.CODE_GENDER 1	data.FLAG_OWN_CAR 1
data.FLAG_OWN_REALTY 1	data.CNT_CHILDREN 1	data.NAME_FAMILY_STATUS 1
data.NAME_HOUSING_TYPE 1	data.REGION_POPULATION_RELATIVE 1	data.AMT_CREDIT 2
data.AMT_INCOME_TOTAL 3	data.AMT_ANNUITY 4	data.AGE 5
data.NAME_TYPE_SUITE 6	data.NAME_INCOME_TYPE 6	data.NAME_EDUCATION_TYPE 6

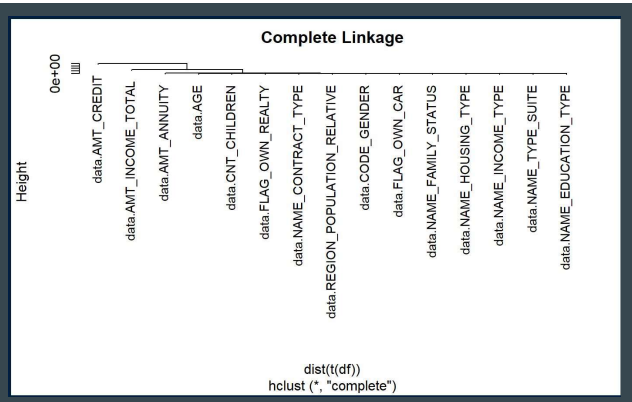
Scaled:

data.AGE 1	data.AMT_INCOME_TOTAL 2	data.NAME_TYPE_SUITE 3
data.NAME_INCOME_TYPE 3	data.NAME_EDUCATION_TYPE 3	data.NAME_CONTRACT_TYPE 4
data.CODE_GENDER 4	data.FLAG_OWN_CAR 4	data.FLAG_OWN_REALTY 4
data.CNT_CHILDREN 4	data.NAME_FAMILY_STATUS 4	data.NAME_HOUSING_TYPE 4
data.REGION_POPULATION_RELATIVE 4	data.AMT_CREDIT 5	data.AMT_ANNUITY 6

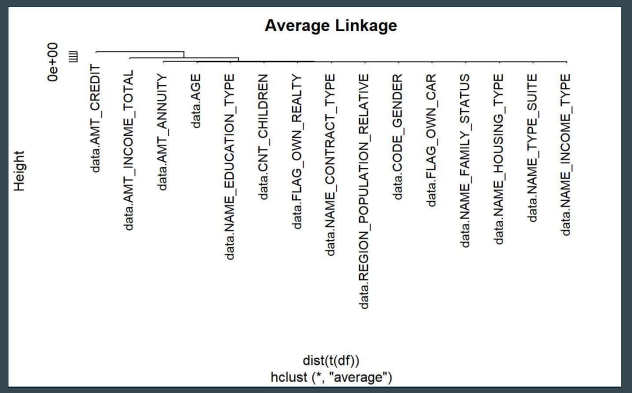
Hierarchical Clustering

- The dendrograms for each of the methods are roughly the same
- Credit and Income being in their own branches, and the rest of the data is on its own branch
- Clusters for all 3 methods will be similar since the composition of the tree is the exact same.
- Trees by default have a height of 4

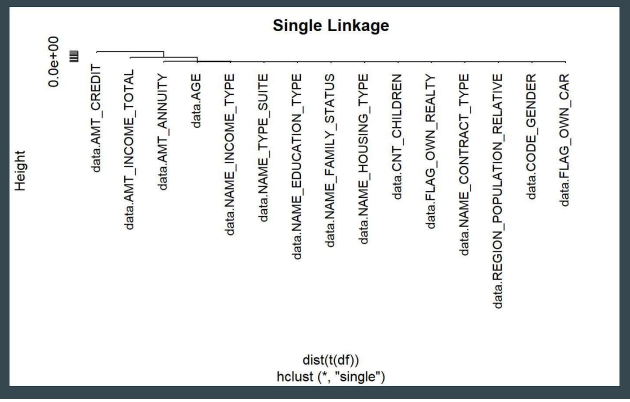
Complete:



Average:



Single:



Hierarchical Clustering (cont.)

- Age is its own cluster in all three of the methods
  - Could be explained by the composition of the variable compared to the other data
- Income Type, Income Total , Credit, and Annuity are also stand alone features
- All of the other features are contained within the first cluster
- Hierarchical Clustering appears to be the better model considering this model gives more consistent results regarding clustering than the k-means clustering

Complete:

data.NAME_CONTRACT_TYPE 1	data.CODE_GENDER 1	data.FLAG_OWN_CAR 1
data.FLAG_OWN_REALTY 1	data.CNT_CHILDREN 1	data.NAME_FAMILY_STATUS 1
data.NAME_HOUSING_TYPE 1	data.REGION_POPULATION_RELATIVE 1	data.AMT_INCOME_TOTAL 2
data.AMT_CREDIT 3	data.AMT_ANNUITY 4	data.NAME_TYPE_SUITE 5
data.NAME_INCOME_TYPE 5	data.NAME_EDUCATION_TYPE 5	data.AGE 6

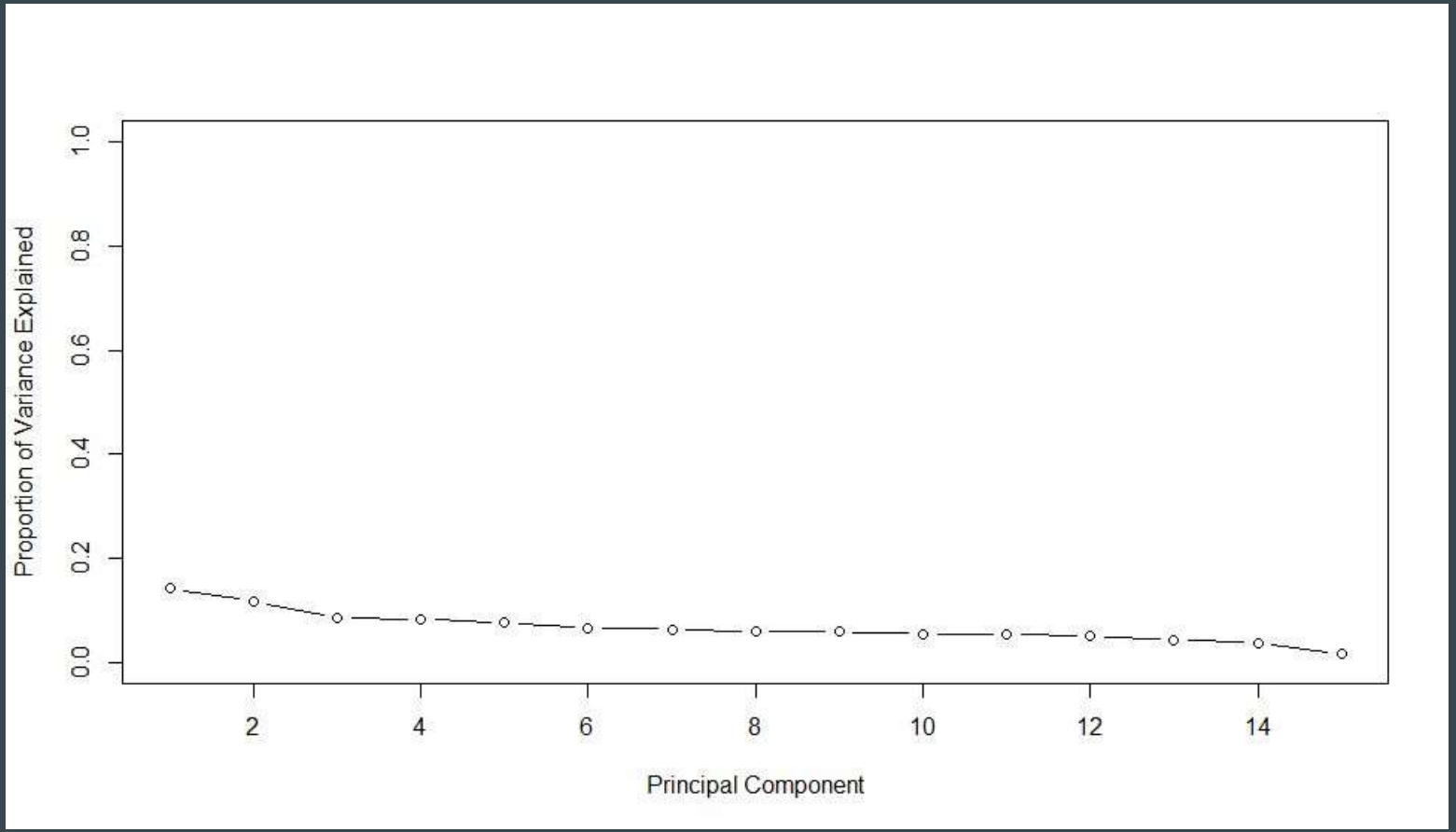
Average:

data.NAME_CONTRACT_TYPE 1	data.CODE_GENDER 1	data.FLAG_OWN_CAR 1
data.FLAG_OWN_REALTY 1	data.CNT_CHILDREN 1	data.NAME_EDUCATION_TYPE 1
data.NAME_FAMILY_STATUS 1	data.NAME_HOUSING_TYPE 1	data.REGION_POPULATION_RELATIVE 1
data.AMT_INCOME_TOTAL 2	data.AMT_CREDIT 3	data.AMT_ANNUITY 4
data.NAME_TYPE_SUITE 5	data.NAME_INCOME_TYPE 5	data.AGE 6

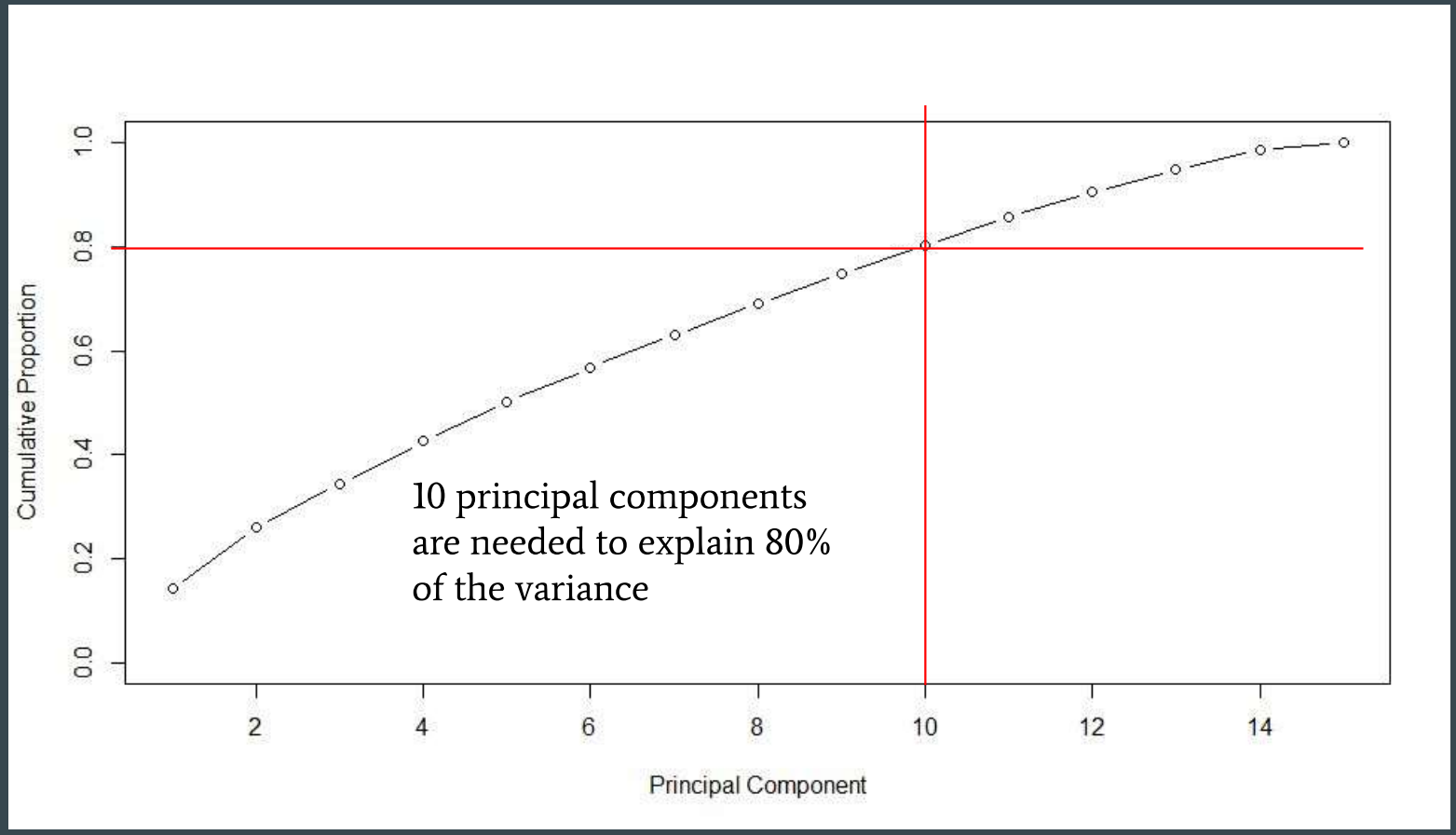
Single:

data.NAME_CONTRACT_TYPE 1	data.CODE_GENDER 1	data.FLAG_OWN_CAR 1
data.FLAG_OWN_REALTY 1	data.CNT_CHILDREN 1	data.NAME_TYPE_SUITE 1
data.NAME_EDUCATION_TYPE 1	data.NAME_FAMILY_STATUS 1	data.NAME_HOUSING_TYPE 1
data.REGION_POPULATION_RELATIVE 1	data.AMT_INCOME_TOTAL 2	data.AMT_CREDIT 3
data.AMT_ANNUITY 4	data.NAME_INCOME_TYPE 5	data.AGE 6

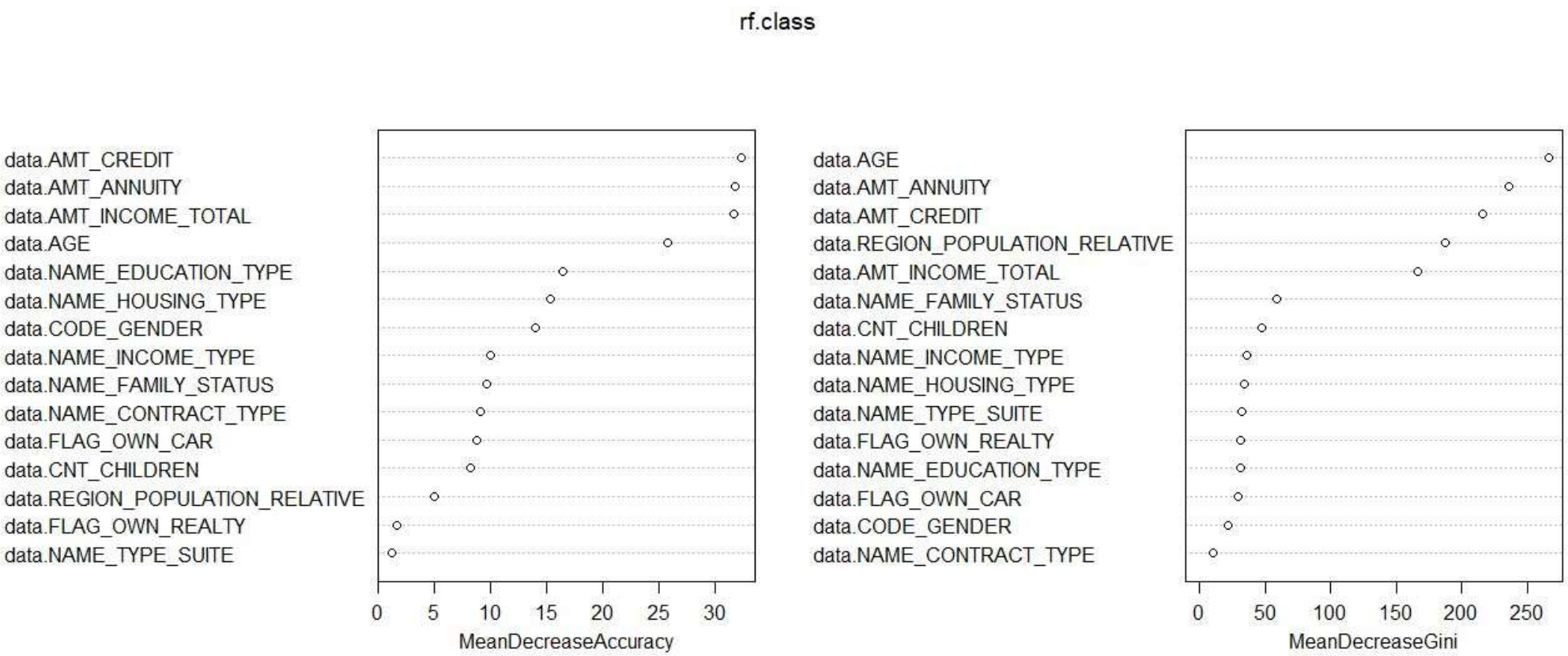
Principal Component Analysis - I



Principal Component Analysis - II



RF Feature Importance



4 Models for Classification Prediction

1 represents default loans 0 represents paid loans

Logistic

SVM:

- Radial Basis Kernel

Random Forest:

- 1000 trees with 5 variables for each tree

Neural Network:

- 1 hidden layer, 4 nodes, 300 epochs

3 Ways to Train and Test the Models:

- Unbalanced train with unbalanced test
- Balanced train with balanced test.
- Balanced train with unbalanced test

* “Unbalanced” represents the raw data’s ratio of default to paid loans which is about 1 to 9
* “Balanced” represents the modified data’s ratio of default to paid loans which is about 1 to 1

Unbalanced Train with Unbalanced Test

	0	1
0	2300	200

Accuracy: 92%

Balanced Train with Balanced Test

Logistic	0	1
0	742	479
1	507	772

60.6%

SVM	0	1
0	794	529
1	453	724

60.7%

RF	0	1
0	740	485
1	507	768

60.3%

NN	0	1
0	1060	1039
1	187	214

51.0%

Balanced Train with Unbalanced Test

Logistic	0	1
0	1345	74
1	956	125

58.8%

SVM	0	1
0	1505	82
1	796	117

64.9%

RF	0	1
0	1366	61
1	935	138

60.2%

NN	0	1
0	954	51
1	1343	152

44.2%

Next Step

- More factors
- More data to train the models
- More complicated models