Classification for Credit Card Default

Team 2-15

Lydia Zhao, Michael Uhrig, Aaron Liu, Leyi Wen

Business Problem

- **Business Problem**: Credit card default payment prediction studies
- **Purpose**: to evaluate the performance of machine learning methods on credit card default payment classification
- **Evaluation Metrics**: accuracy and type II error rate
- **Importance**: to improve and ease the process of credit card default detection and therefore help the banking system in decision making

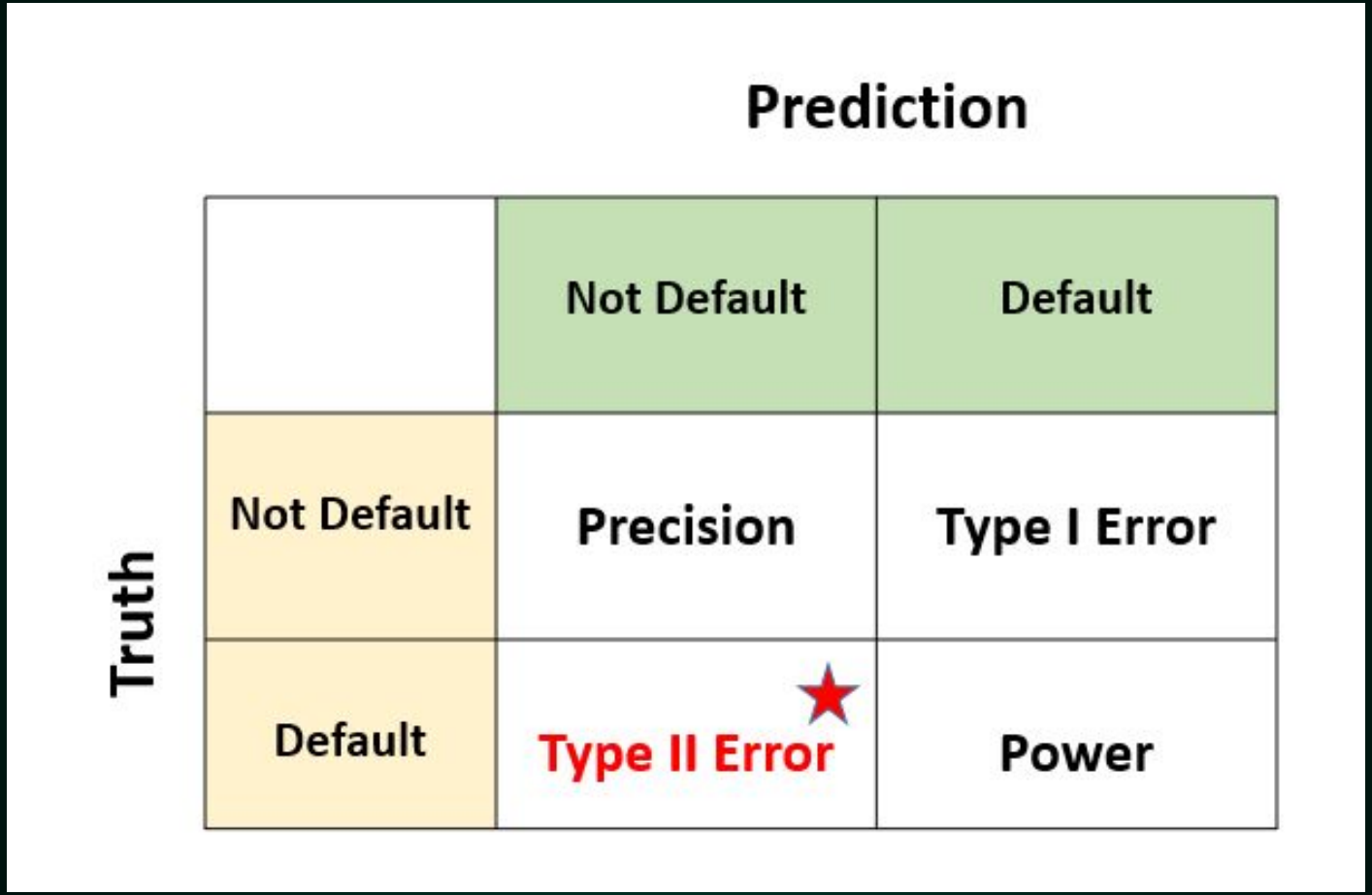
Results Overview

- **Hypothesis Test**: Whether clients would default on their credit card or not

  H_0 : not default; H_1: default

- **Objective of prediction :** 1) High classification accuracy; 2) Low Type II error

- **Overall Result**: Random forest with equal sample size set gives the lowest type II error rate of 32.4%


Prediction
Not Default
Default
Truth
Not Default
Precision
Type I Error
Default
Type II Error
Power

Data Exploration

The original dataset has 30,000 instances and 23 attributes. Collected in Taiwan, 2005.

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
2198	1000000	2	1	1	47	0	0	0	-1	0	0
27558	730000	1	3	1	56	0	0	0	0	0	0
26548	520000	2	3	1	54	0	0	0	0	0	0
6913	590000	1	1	1	63	0	0	0	-1	0	0
28143	430000	2	5	2	44	0	0	0	0	0	0
11558	480000	2	1	2	33	0	0	0	0	-1	0
27499	500000	1	3	1	55	2	0	0	0	0	0
14554	450000	1	2	2	30	0	0	0	0	0	0
26228	610000	2	1	2	38	0	0	0	0	0	0
10781	330000	1	1	2	44	0	0	0	0	0	0
28625	340000	2	2	2	31	5	4	4	3	2	0
23759	500000	2	2	1	52	0	0	0	0	0	0
5925	340000	2	2	2	31	4	3	2	-1	-1	-1
29042	460000	2	2	1	50	0	0	0	0	0	0

Data Exploration

BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month
964511	983931	535020	891586	927171	961664	50784	50723	896040	50000	50000	50256	0
746814	374028	351588	86927	66111	38491	20500	16500	3000	2000	2000	5000	0
653062	671563	689627	706864	383160	294641	28500	30500	30000	15000	15000	0	0
630458	646770	693131	324522	358774	369685	28000	61115	325000	40000	20000	51000	0
626648	586825	547667	504474	462640	420585	20659	20421	16943	15634	14933	15131	0
621749	550102	475386	384078	397682	399659	22024	21008	17134	400046	17003	4436	0
613860	512526	334227	145482	125936	91382	37300	11000	4500	4000	4000	100000	1
610723	555086	497132	514249	462666	472480	20200	18000	25135	432130	17000	20000	1
608594	624475	632041	516575	454845	456596	26868	22375	17221	15300	16000	18000	0
604019	605943	439854	404157	370686	294348	19001	14000	12000	15000	7000	7000	0
589654	581775	572677	384060	304508	247178	3000	0	0	1000	4320	287982	1
588000	277559	288835	281810	273700	269552	17559	36500	20000	36000	20000	80013	0
581775	572677	384060	304508	247178	228349	0	0	1000	4320	287982	8007	1
581319	552144	523423	493548	429966	427216	19141	19141	20141	14642	31549	15522	1

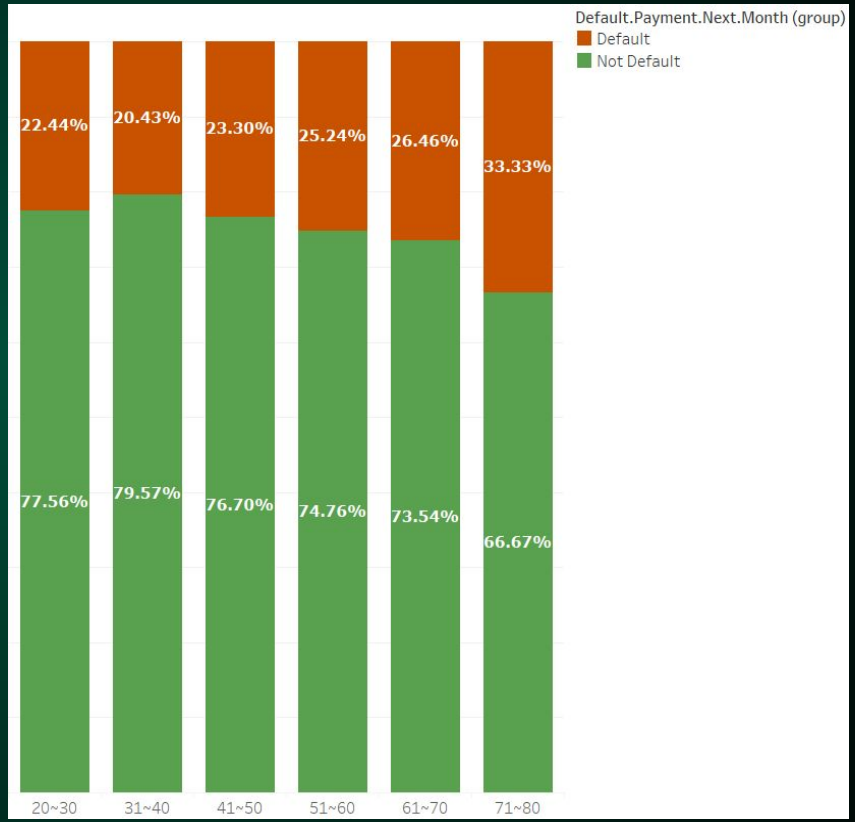
The total proportion of defaults in the data is 22.1%, which is 6,636 out of 30,000.

Exploratory Analysis of Data

Payment Status by Age

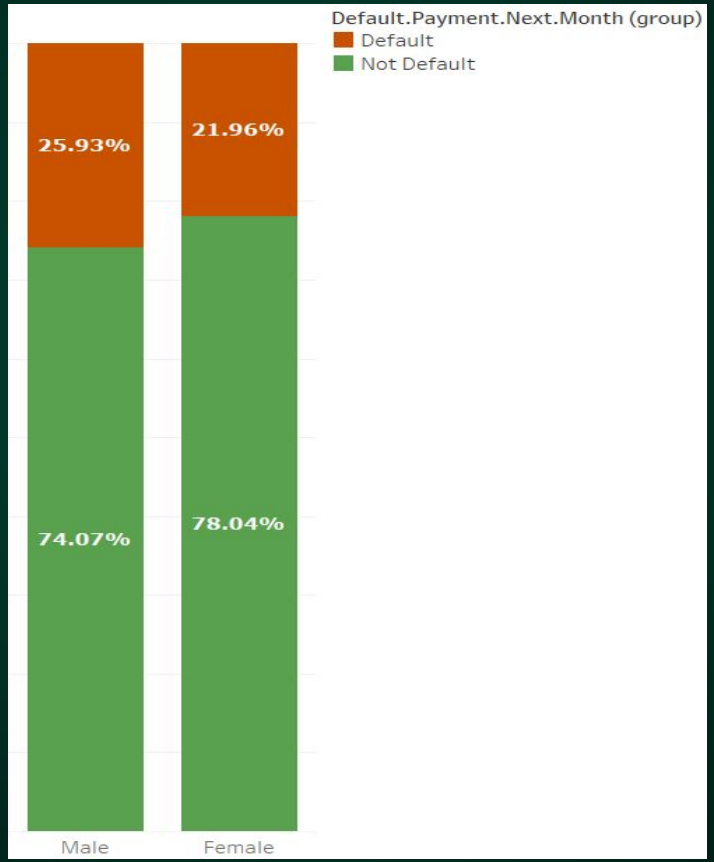
```
summary(data$AGE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.00   28.00   34.00   35.49   41.00   79.00
```

- Lowest default rate for age group 31~40.
- Older customers have a higher chance of default.



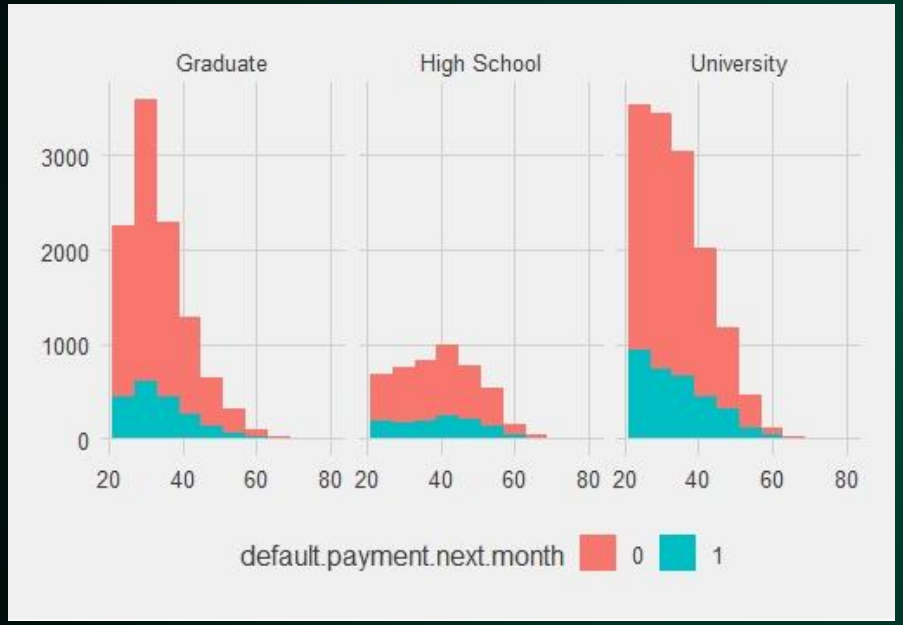
Exploratory Analysis of Data

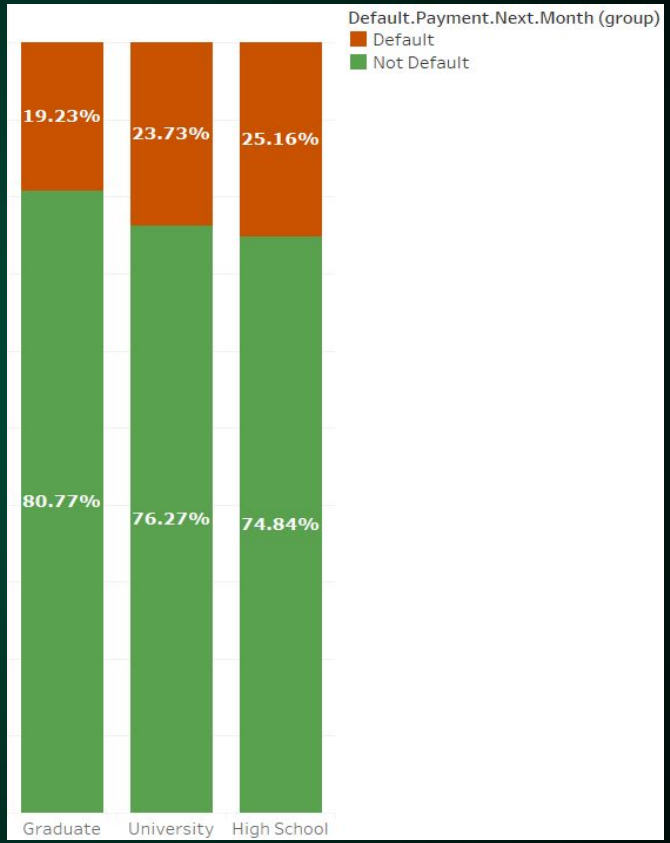
Payment Status by Gender



Exploratory Analysis of Data

Payment Status by Education Level

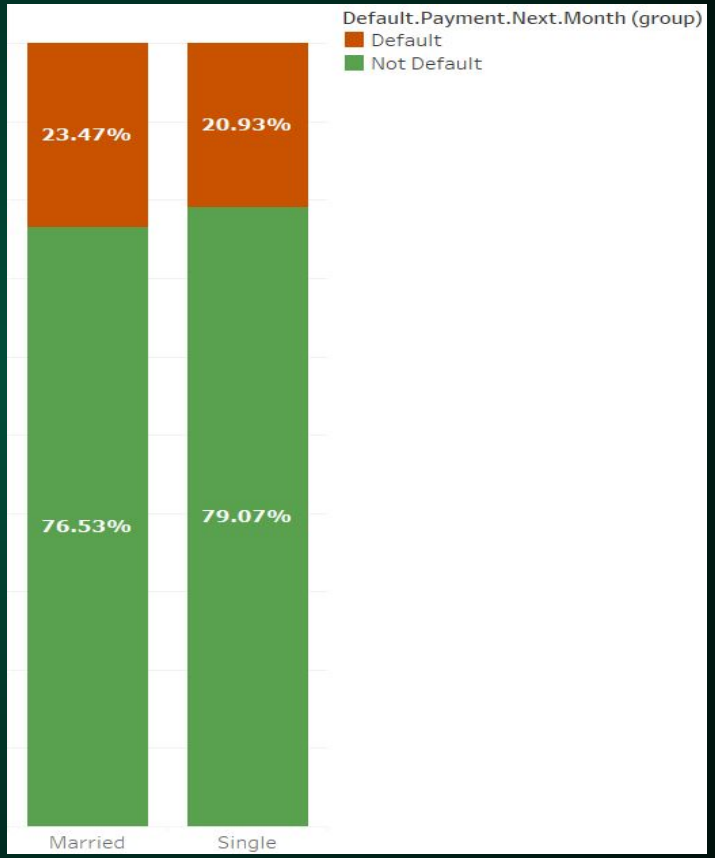




Exploratory Analysis of Data

Payment Status by Marital Status

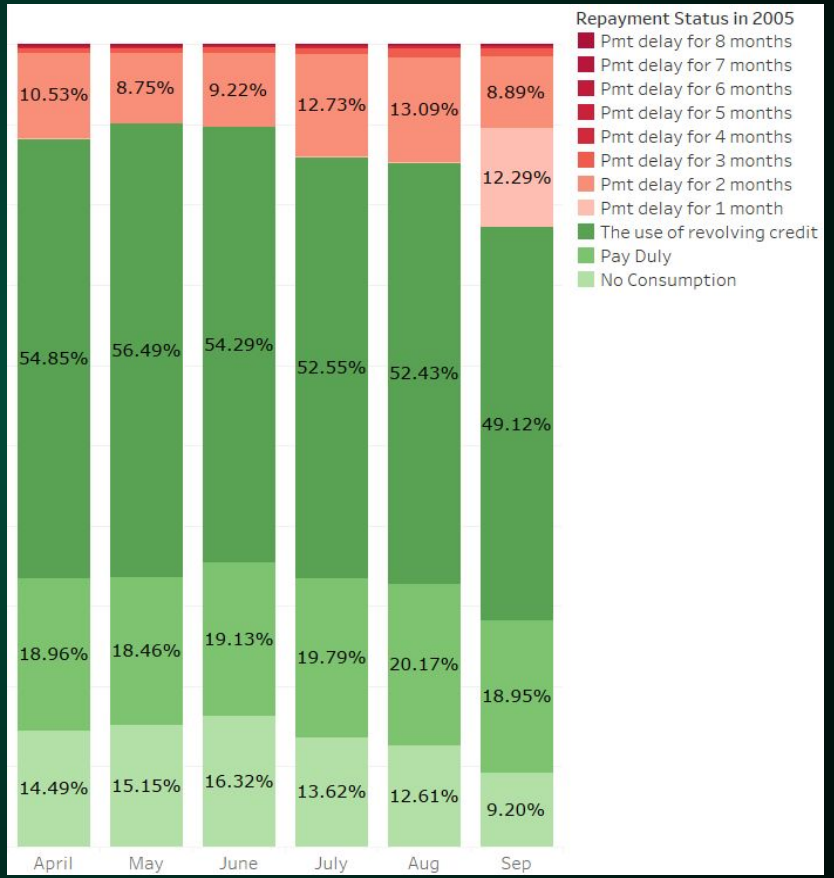
- Marital Status is mostly divided into categories “Married” and “Single”.



Exploratory Analysis of Data

Payment Status Overview

- Higher delayed payment rate in July and August
- Large percent of delayed payment for one month in September
    - Shorter time interval between billing date and data collection date.



Data Preprocessing & Preparation

Step 1: Data cleaning

- dropping groups that have insignificant amount of data:

```
data$EDUCATION[data$EDUCATION== 4] = 0
data$EDUCATION[data$EDUCATION== 5] = 0
data$EDUCATION[data$EDUCATION== 6] = 0
data$MARRIAGE[data$MARRIAGE== 3] = 0
myData = myData[!myData$A > 4,]
data = data[!data$EDUCATION == 0,]
data = data[!data$MARRIAGE == 0,]
```

Data Preprocessing & Preparation

Step 2: Factorize categorical variables

```
# Factorize Categorical variables: gender, education, marital_status, default
factor_vars <- c('SEX','EDUCATION','MARRIAGE','default.payment.next.month')
data[factor_vars] <- lapply(data[factor_vars], function(x) as.factor(x))
```

```
$ default.payment.next.month: Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 2 ...
$ paystate                  : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 1 2 2 2 1 .
$ genderNew                 : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 2 2 1
$ educationNew              : Factor w/ 3 levels "Graduate","High School",..: 1 2
$ maritalNew                : Factor w/ 2 levels "Married","Single": 1 1 1 1 2 1 2
$ AGE.group                 : Factor w/ 6 levels "(20,30]","(30,40]",..: 3 4 4 5 2
```

Data Preprocessing & Preparation

Step 3: Add a new variable named “paystatus”:

- Paystatus <= 0 ⟶ pay duly in last 6 months
- Paystatus > 0 ⟶ pay delayed in last 6 months

Step 4: Split data:

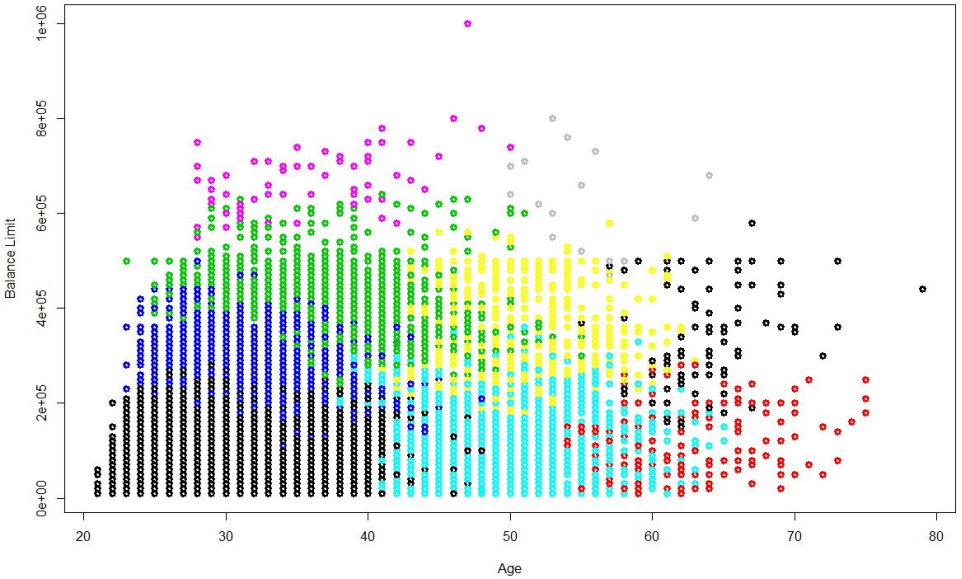
- 80%: training set
- 20%: test set

Model Selection

- **Clustering**

- Logistic Regression
- Support Vector Machines
- Decision Tree
- Random Forest

Clustering



	Colors								
	black_left	black_right	blue	green	grey	pink	red	teal	yellow
Default Rate	0.2486	0.2376	0.1503	0.1391	0.2286	0.0781	0.3415	0.2587	0.1727
PAY_0	0.1633	0.2277	0.2407	0.2542	0.2571	0.0781	0.1338	0.1835	0.2996
PAY_2	0.1698	0.2871	0.2558	0.2718	0.2571	0.1250	0.1444	0.1977	0.3445
PAY_3	0.1652	0.2673	0.2515	0.2745	0.2000	0.2031	0.1479	0.1914	0.3282
PAY_4	0.1591	0.2574	0.2428	0.2511	0.1429	0.2031	0.1338	0.1857	0.3177
PAY_5	0.1591	0.2277	0.2355	0.2347	0.2000	0.2031	0.1479	0.1759	0.2977
PAY_6	0.1640	0.2376	0.2412	0.2458	0.2286	0.2344	0.1761	0.1853	0.3006
BILL_AMT1	40,044	96,134	69,530	88,624	170,202	159,969	49,674	42,302	72,232
BILL_AMT2	38,671	91,476	66,928	83,889	167,809	147,768	47,075	40,887	68,955
BILL_AMT3	36,685	88,023	63,839	82,531	170,697	140,464	43,688	38,650	67,754
BILL_AMT4	33,675	88,241	59,377	78,142	151,714	134,647	39,288	34,335	62,759
BILL_AMT5	31,288	82,127	56,026	73,069	131,064	136,260	36,449	31,840	58,760
BILL_AMT6	30,297	79,916	54,191	69,957	132,028	140,080	34,843	30,797	55,560
PAY_AMT1	4,056	9,623	8,140	11,048	20,106	21,502	3,193	4,330	10,455
PAY_AMT2	3,870	5,839	8,214	14,621	17,930	20,016	3,703	4,120	11,882
PAY_AMT3	3,492	8,491	7,534	11,883	24,533	29,656	3,076	3,374	10,195
PAY_AMT4	3,206	4,779	7,287	10,899	11,553	23,805	3,441	3,338	9,707
PAY_AMT5	3,265	8,632	7,114	10,679	15,749	33,480	2,827	3,132	9,277
PAY_AMT6	3,277	5,101	7,932	13,229	17,378	25,985	3,662	3,324	10,312
EDUCATION	1.768	2.099	1.748	1.574	1.257	1.391	1.799	2.111	1.781
MARRIAGE	0.360	0.277	0.395	0.468	0.800	0.672	0.356	0.221	0.566
SEX	1.626	1.238	1.561	1.649	1.086	1.563	1.472	1.598	1.423

Model Selection

- Clustering

- **Logistic Regression**
- Support Vector Machines
- Decision Tree
- Random Forest

Variable Selection by Logistic Regression

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.743e-01  1.883e-01  -1.987 0.046873 *
LIMIT_BAL                 -6.463e-07  1.824e-07  -3.544 0.000394 ***
AGE                        6.560e-03  6.560e-03   1.000 0.317342
PAY_0                      4.264e-01  2.026e-02  21.043  < 2e-16 ***
PAY_2                     -3.648e-02  2.311e-02  -1.578 0.114467
PAY_3                     -3.976e-03  2.505e-02  -0.159 0.873863
PAY_4                     -4.682e-02  2.725e-02  -1.718 0.085775 .
PAY_5                      2.737e-02  2.901e-02   0.943 0.345562
PAY_6                     -4.950e-02  2.414e-02  -2.051 0.040291 *
BILL_AMT1                 -4.021e-06  1.257e-06  -3.198 0.001382 **
BILL_AMT2                  1.463e-06  1.698e-06   0.862 0.388897
BILL_AMT3                  1.898e-06  1.526e-06   1.244 0.213633
BILL_AMT4                 -4.815e-07  1.520e-06  -0.317 0.751425
BILL_AMT5                  1.376e-06  1.670e-06   0.824 0.409922
BILL_AMT6                 -9.772e-07  1.314e-06  -0.744 0.457179
PAY_AMT1                  -1.130e-05  2.519e-06  -4.486 7.25e-06 ***
PAY_AMT2                  -8.385e-06  2.245e-06  -3.735 0.000188 ***
PAY_AMT3                  -2.265e-06  1.928e-06  -1.174 0.240213
PAY_AMT4                  -3.526e-06  2.015e-06  -1.750 0.080068 .
PAY_AMT5                  -2.349e-06  1.996e-06  -1.177 0.239289
PAY_AMT6                  -1.515e-06  1.409e-06  -1.075 0.282248
paystateYES               -1.261e+00  5.320e-02 -23.698  < 2e-16 ***
genderNewMale              1.263e-01  3.538e-02   3.569 0.000358 ***
educationNewHigh School   -7.999e-02  5.487e-02  -1.458 0.144886
educationNewUniversity    -3.539e-02  4.068e-02  -0.870 0.384334
maritalNewSingle          -1.576e-01  4.053e-02  -3.888 0.000101 ***
AGE.group(30,40]           4.375e-04  7.021e-02   0.006 0.995028
AGE.group(40,50]           2.570e-02  1.283e-01   0.200 0.841239
AGE.group(50,60]          -9.791e-02  1.934e-01  -0.506 0.612769
AGE.group(60,70]          -7.512e-02  3.000e-01  -0.250 0.802265
AGE.group(70,80]           9.110e-01  6.841e-01   1.332 0.182926
---
```

Variables Selected :

- LIMIT_BAL : credit level
- PAY_0 : repayment status in Sep, 2005
- PAY_2 : repayment status in Aug, 2005
- PAY_4: repayment status in Jun, 2005
- PAY_6 : repayment status in Apr, 2005
- BILL_AMT1: amount of bill statement in Sep, 2005
- PAY_AMT1: amount of previous payment in Sep, 2005
- PAY_AMT2: Amount of previous payment in Aug, 2005
- PAY_AMT4 : Amount of previous payment in Jun, 2005
- educationNew: education level
- Paystate: delayed payment status
- genderNew: Gender
- maritalNew: marital status

Model Selection

- Clustering

- Logistic Regression
- **Support Vector Machines**
- Decision Tree
- Random Forest

Support Vector Machines

Linear

Cost = 1
Accuracy = 81.1%
Type II error = 75%

actual	predict 0	1
0	4398	106
1	997	332

Radial

Cost = 1, Gamma = 1
Accuracy = 80.7%
Type II error = 71.1%

actual	predict 0	1
0	4315	189
1	936	393

Model Selection

- Clustering

- Logistic Regression
- Support Vector Machines
- **Decision Tree**
- Random Forest

Model Selection

- Clustering

- Logistic Regression
- Support Vector Machines
- Decision Tree
- **Random Forest**

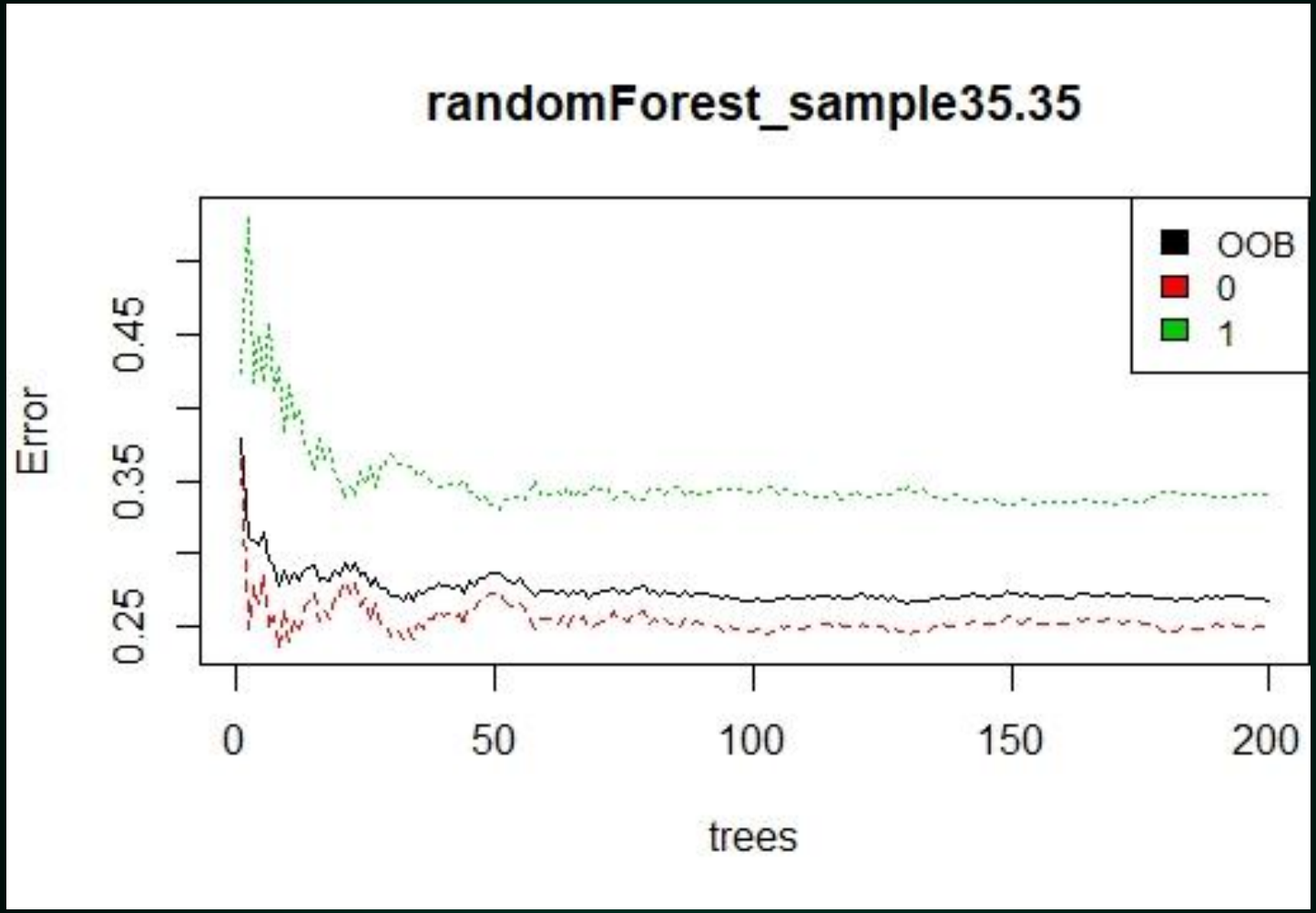
Random Forest

No Sample Size Specified:

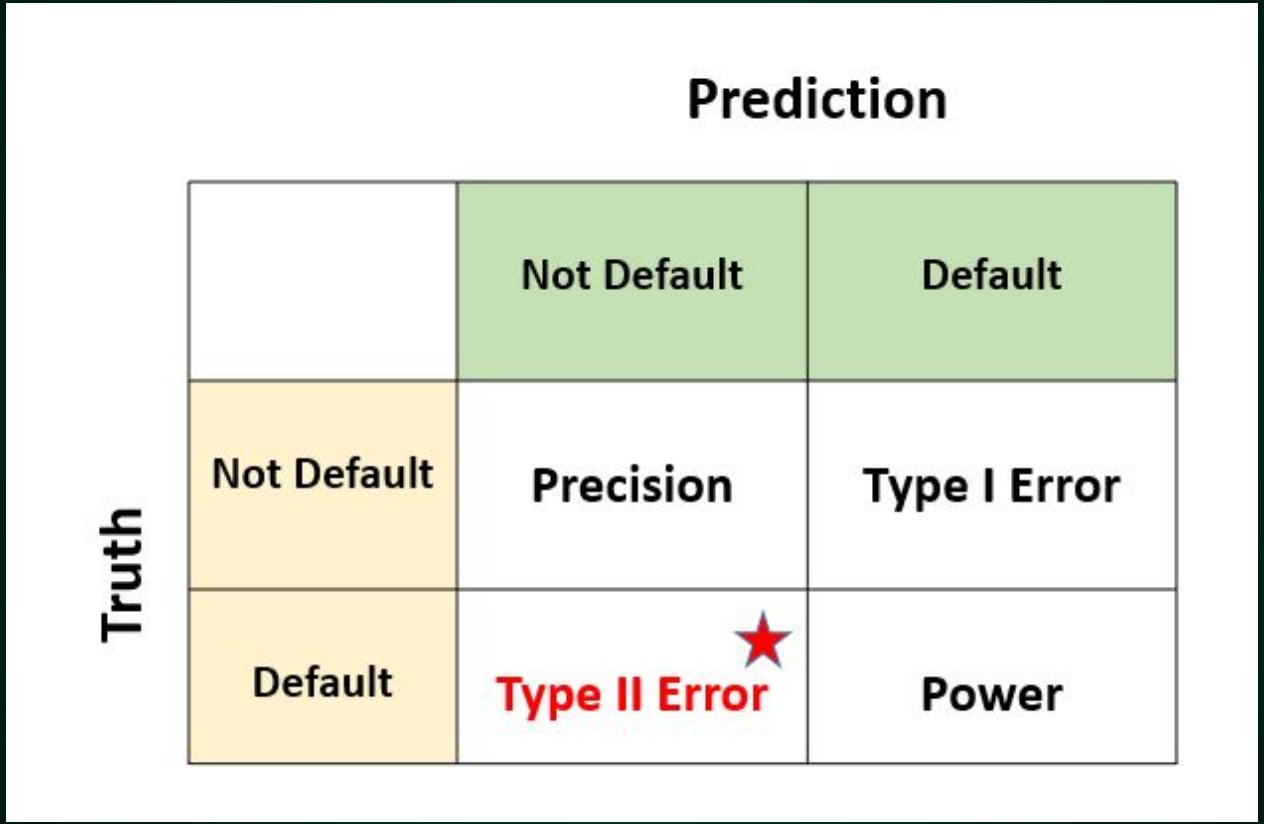
actual	preds 0	1
0	4261	243
1	845	484

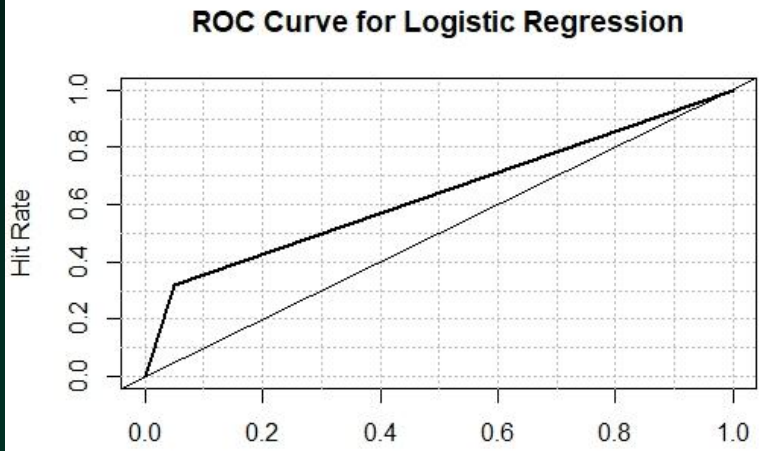
With Sample Size Set to (35,35):

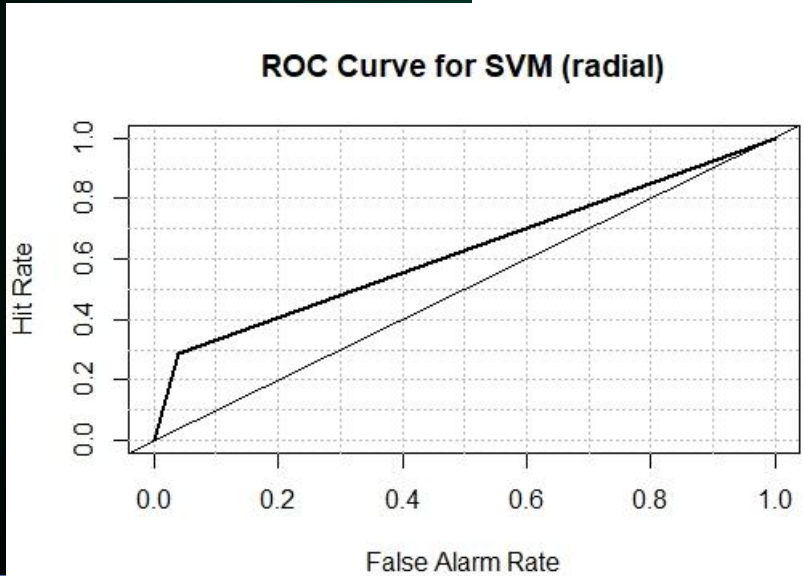
actual	preds 0	1
0	3410	1094
1	449	880

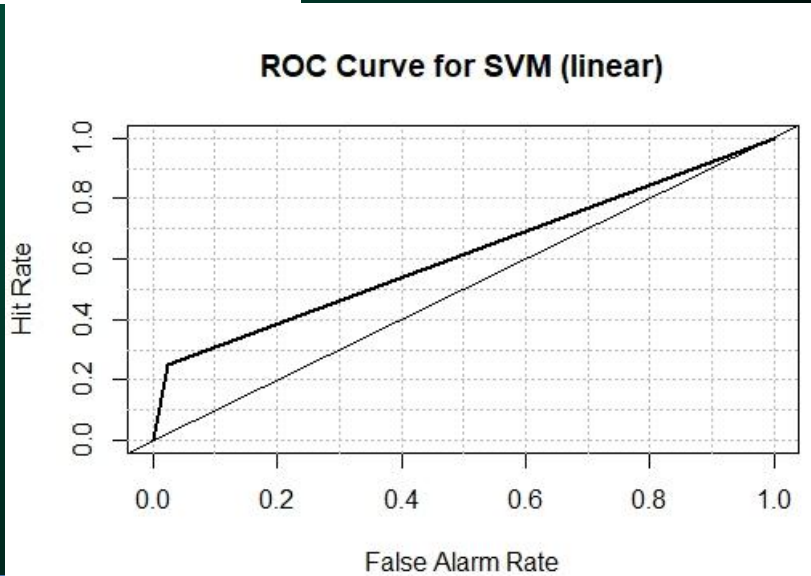

randomForest_sample35.35
OOB
0
1
Error
0.25
0.35
0.45
0
50
100
150
200
trees

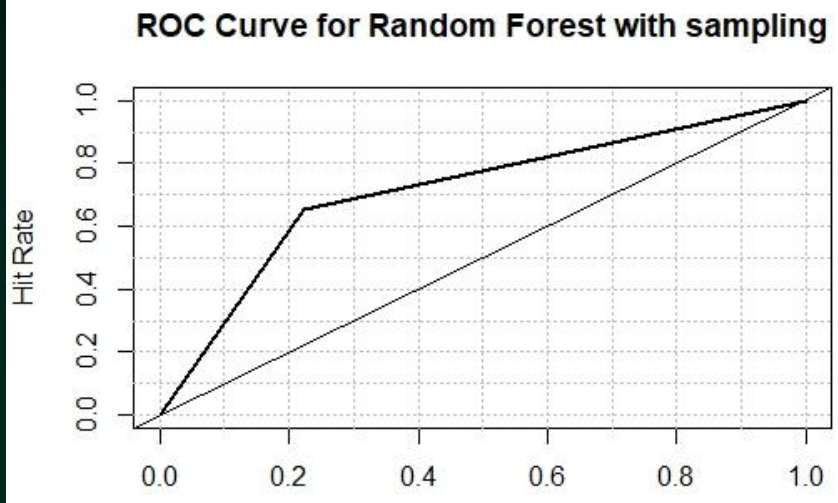
Model Comparison

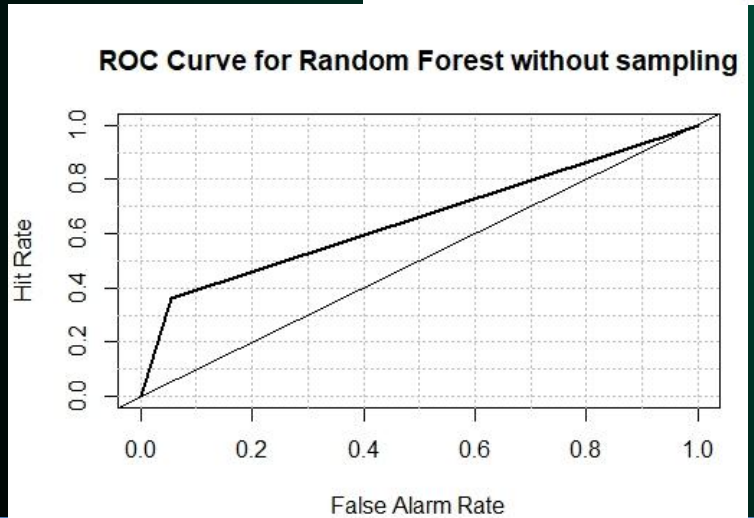


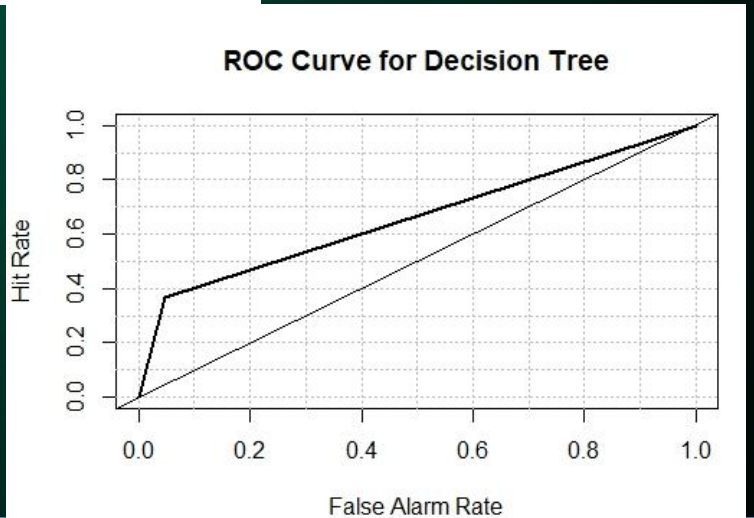

ROC Curve for Logistic Regression
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for SVM (radial)
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for SVM (linear)
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for Random Forest with sampling
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0

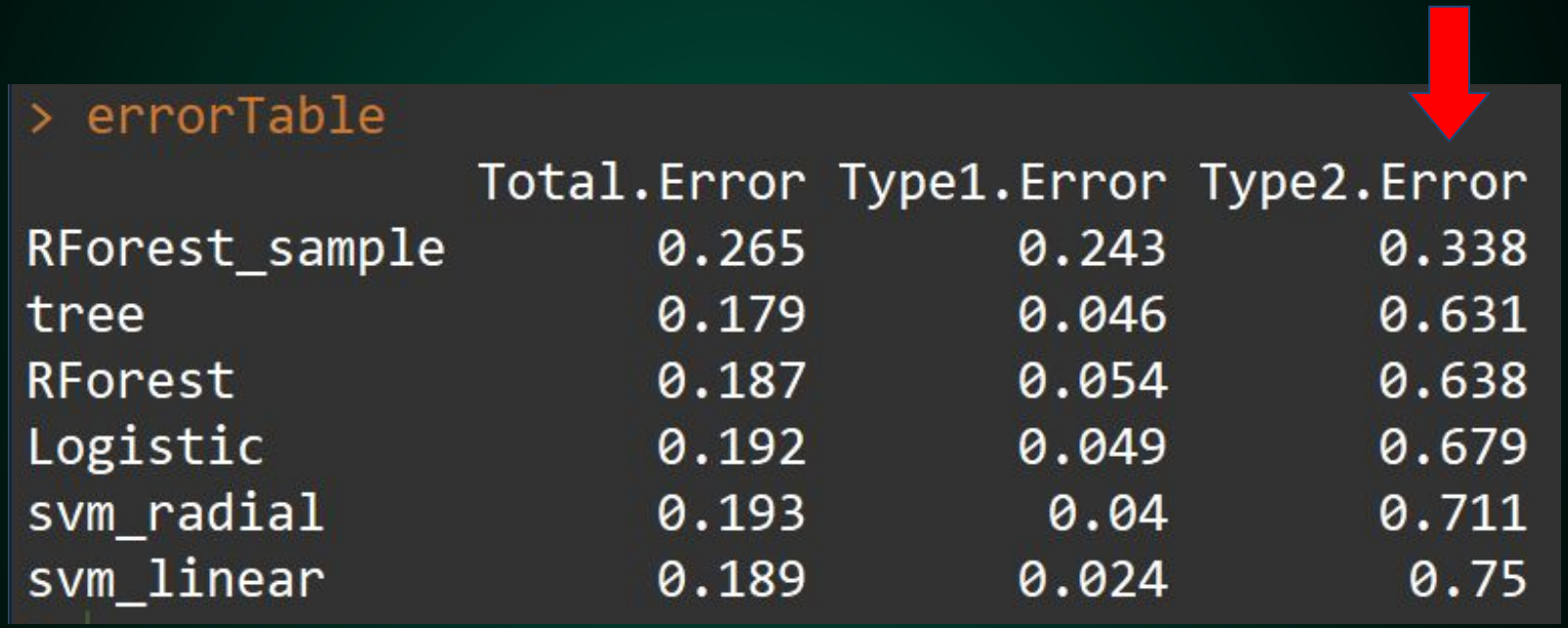

ROC Curve for Random Forest without sampling
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for Decision Tree
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0

Accuracy Comparison

```
> AccTable
                Accuracy Power Precision
tree               0.821 0.369     0.702
RForest            0.813 0.362     0.664
svm_linear         0.811  0.25     0.758
svm_radial         0.807 0.289     0.678
Logistic           0.808 0.321     0.661
RForest_sample     0.735 0.662     0.446
```

Error Rate Comparison



```
> errorTable
                Total.Error Type1.Error Type2.Error
RForest_sample        0.265       0.243       0.338
tree                  0.179       0.046       0.631
RForest               0.187       0.054       0.638
Logistic              0.192       0.049       0.679
svm_radial            0.193        0.04       0.711
svm_linear            0.189       0.024        0.75
```

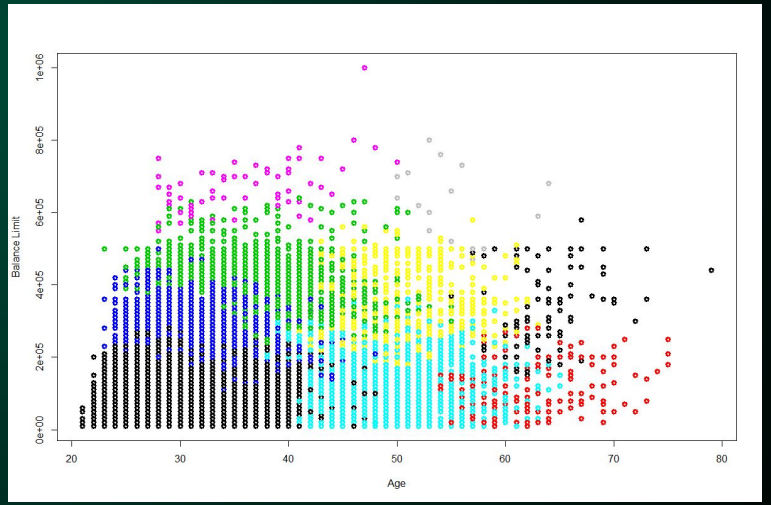
Recommendation

Customer segmentation

True defaults and credit level per customer

Female Married Graduate Age 50 Balance 200,000	Male Single Graduate Age 26 Balance 200,000	Female Single Graduate Age 33 Balance 520,000
Male Single High School Age 55 Balance 20,000	Male Married University Age 41 Balance 230,000	Female Single University Age 41 Balance 260,000
Male Married High School Age 72 Balance 50,000	Male Married University Age 61 Balance: 80,000	Female Married University Age 60 Balance:80,000

- Consider customer profile as criteria
- Combine classification model with customer segmentation



Conclusion

1. Overall, our analysis suggested that for modeling of credit card risk, the best prediction result is achieved by Random Forest with equal sample size

2. Trade-off between overall accuracy rate and more aggressive prediction on who actually default when predicted as “default”

3. Random Forest might not be intuitive and not easy to be implemented by banks

4. Dataset is from year 2005 covered for only 6 months

Appendix 1

- Dependent Variable
  - Default.payment.next.month: Default Payment (1=yes, 0=no)
- Independent Variables
  - LIMIT_BAL: Amount of given credit in NT dollars
  - SEX: Gender (1=male, 2=female)
  - EDUCATION: (1=graduate school, 2=university, 3=high school)
  - MARRIAGE: Marital status (1=married, 2=single, 3=others)
  - AGE: Age in years
  - PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, and so forth.)
  - PAY_2: Repayment status in August, 2005 (scale same as above)
  - PAY_3: Repayment status in July, 2005 (scale same as above)
  - PAY_4: Repayment status in June, 2005 (scale same as above)
  - PAY_5: Repayment status in May, 2005 (scale same as above)
  - PAY_6: Repayment status in April, 2005 (scale same as above)

Appendix 2

- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

Backup Slide 1 -- Tree Pruning

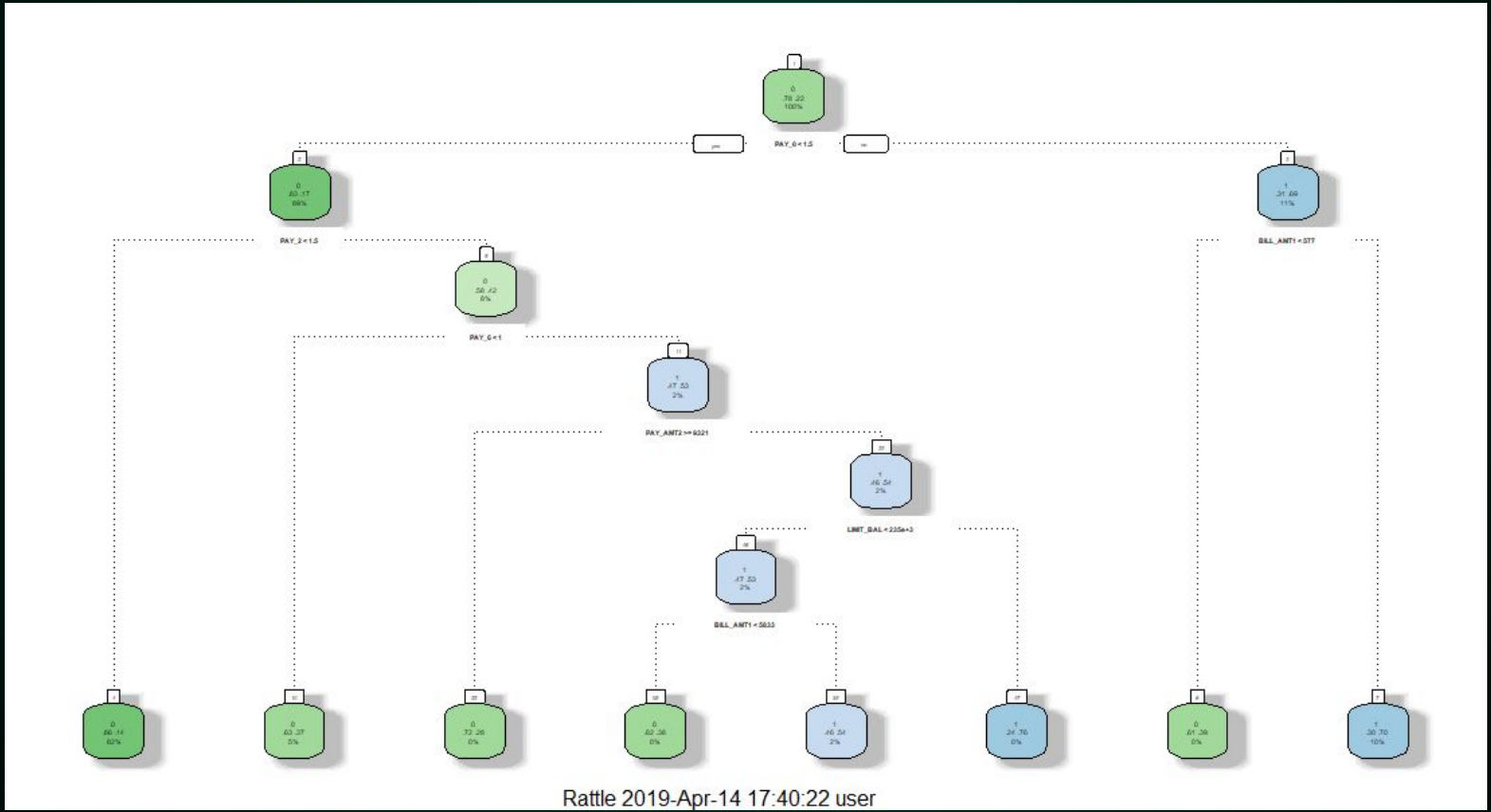
- No Tree pruning:

actual	preds 0	1
0	4296	208
1	839	490

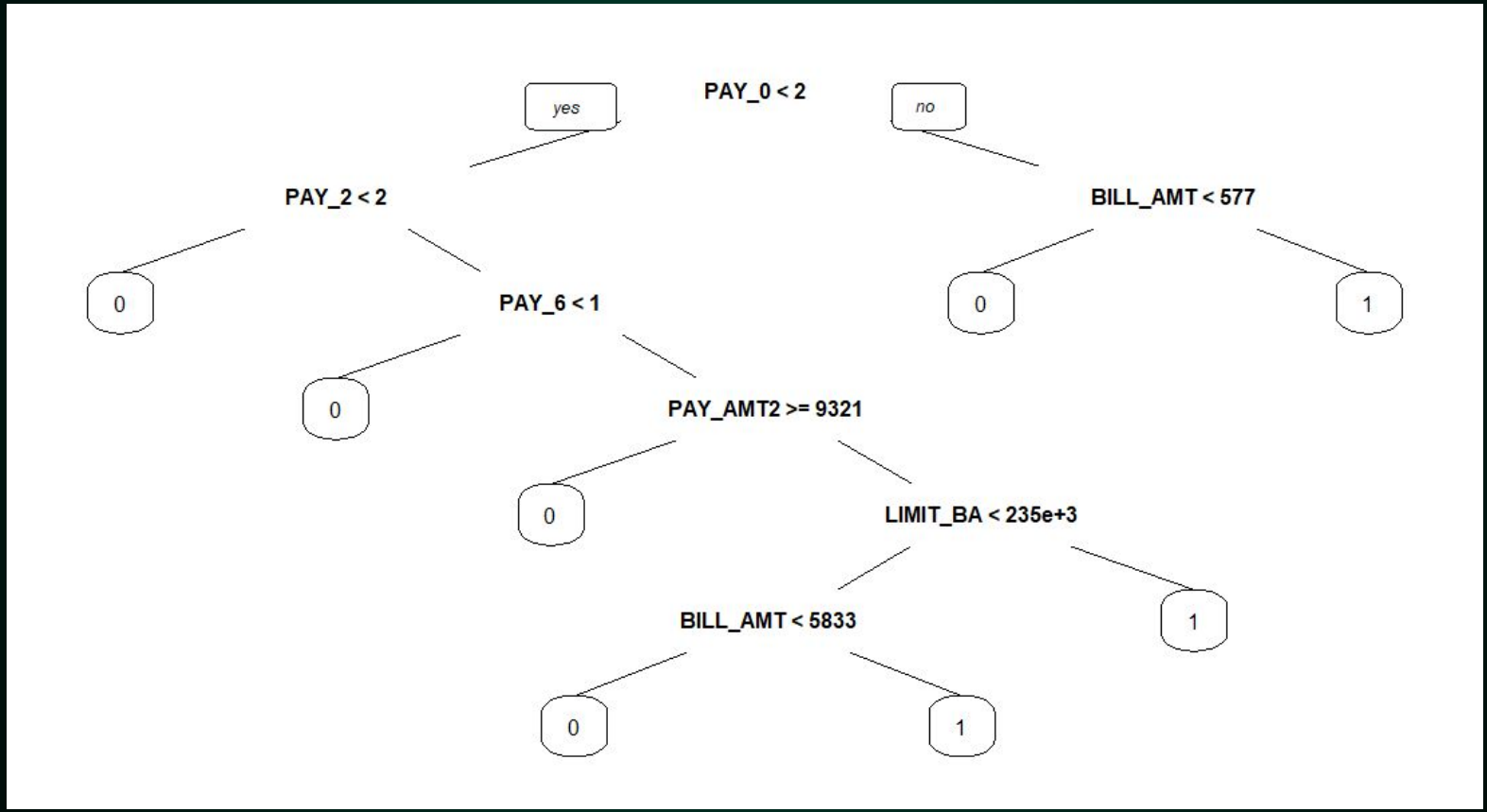
With Tree pruning:

actual	preds 0	1
0	4299	205
1	842	487

Backup Slide 2 -- Tree Plot



Backup Slide 2 -- Tree Plot



Backup Slide 3 -- Outlier

1) Retain data representativeness
2) Accuracy dropped when outliers are removed