Team Random Presentation

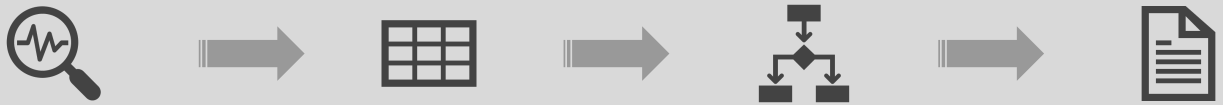
Dora Yuan, Mia Zhang, Tako Suzuki, York Fang, Yuan Liu

TABLE OF CONTENTS

01 Introduction

- **Background:**
  - Directed marketing: focus on targets that will be keener to specific product/services
  - Bank marketing: contact less but achieve higher number of clients subscribing the deposit
  - Dataset: direct marketing campaigns of a Portuguese banking institution
- **Goals:**
  - Build predictive models to predict the success of a contact
  - Rank the variables based upon the important level in the success of direct marketing campaigns

02 Process Overview



Understand Business and Data

Prepare Data

Build Models

Generate Insights

03 Data Pre-Processing

Part 1 Attribute Conversion

Part 2 Complexity Reduction

Part 3 Confirmation of Significance Difference

Attribute Conversion

Part 1

1. **Model Selection**
   - Naive Bayes
   - Decision Tree

2. **Binning (Numerical —> Categorical)**
   - Disadvantage: Lose accuracy
   - Purpose: Reduce overfitting, Process time

```
> result <- smbinning(df=df, y="y", x="balance")
> result$ivtable
  Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec GoodRate BadRate
1   <= -47   3193     166   3027      3193        166      3027 0.0706   0.0520  0.9480
2    <= 60   7628     594   7034     10821        760     10061 0.1687   0.0779  0.9221
3   <= 798  17577    1963  15614     28398       2723     25675 0.3888   0.1117  0.8883
4    > 798  16813    2566  14247     45211       5289     39922 0.3719   0.1526  0.8474
5  Missing      0       0      0     45211       5289     39922 0.0000      NaN     NaN
6    Total  45211    5289  39922        NA         NA        NA 1.0000   0.1170  0.8830
    Odds  LnOdds     WoE     IV
1 0.0548 -2.9033 -0.8820 0.0392
2 0.0844 -2.4716 -0.4503 0.0288
3 0.1257 -2.0737 -0.0524 0.0010
4 0.1801 -1.7142  0.3071 0.0394
5    NaN     NaN     NaN    NaN
6 0.1325 -2.0213  0.0000 0.1084
```

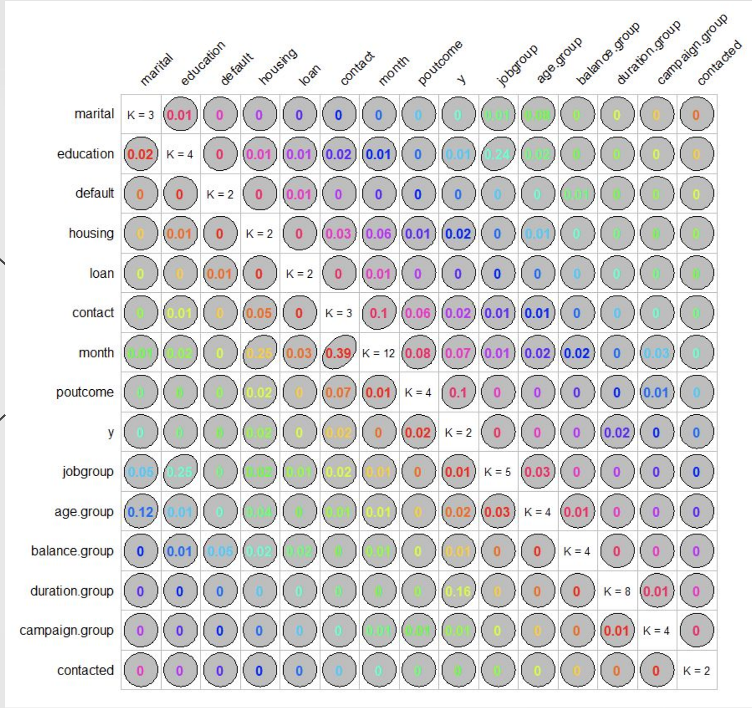
Complexity Reduction

Part 2

1. **Level Reduction - “Job”**
   - Clustering method with personal data attributes as the input

2. **Drop low impact variables**
   - “Pday”
   - “Previous”
   - “Default”
   - “Contacted”

Correlation Matrix



Top 5 variables:

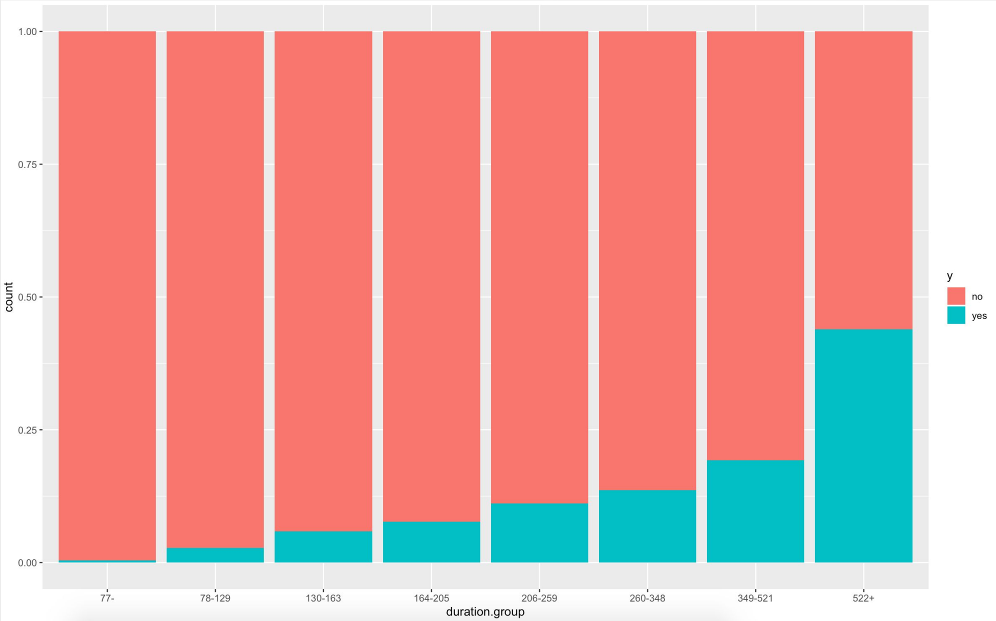
- poutcome: 0.02287
- duration.group: 0.02208
- housing: 0.01937
- contact: 0.01803
- loan: 0.00465

Confirmation of Significance Difference

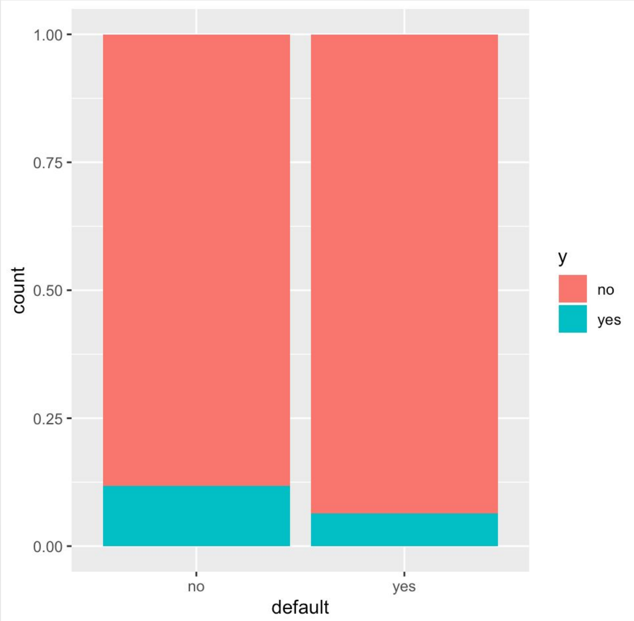
Part 3

Explore the correlation within each variable

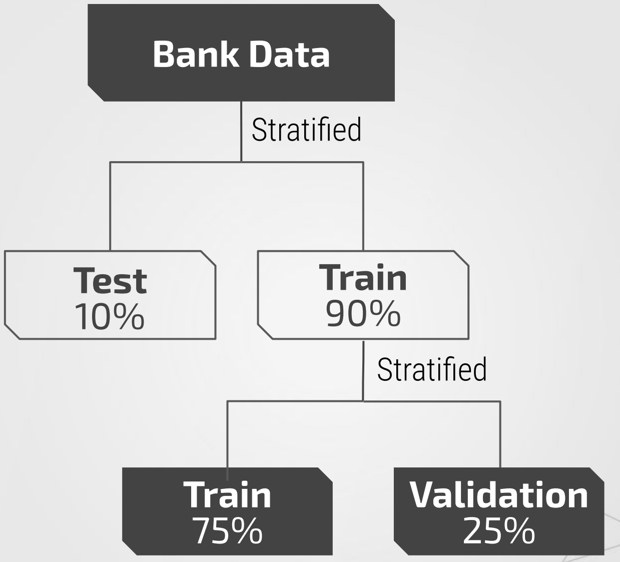
“Duration” - Keep



“Default” - Toss

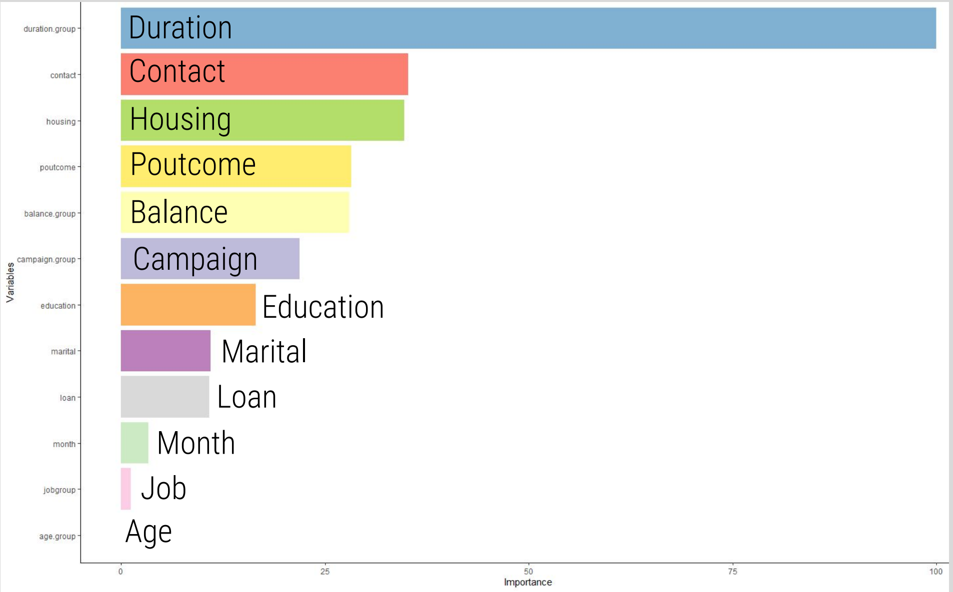


04 Modeling: Data Partition

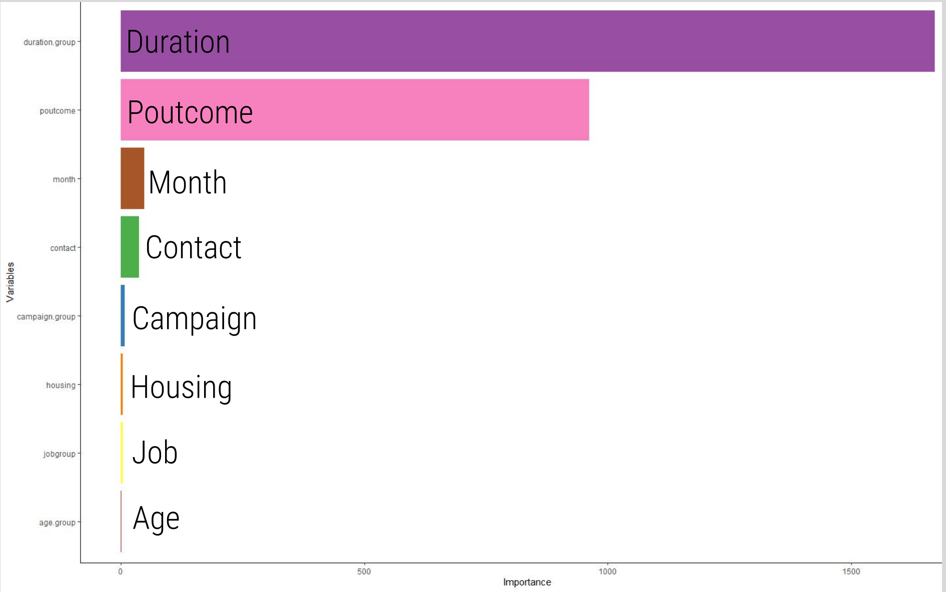


Model Comparison: Variable Importance

Naive Bayes



Decision Tree



Model Comparison: Accuracy

Naive Bayes

```
> confusionMatrix(nb.pred, validate.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  8579  698
       yes  403  492

               Accuracy : 0.8918
                 95% CI : (0.8856, 0.8977)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 0.002921

                  Kappa : 0.413

 Mcnemar's Test P-Value : < 0.0000000000000022

            Sensitivity : 0.9551
            Specificity : 0.4134
         Pos Pred Value : 0.9248
         Neg Pred Value : 0.5497
             Prevalence : 0.8830
         Detection Rate : 0.8434
   Detection Prevalence : 0.9120
      Balanced Accuracy : 0.6843

       'Positive' Class : no
```

Accuracy: 0.8918

VS.

Decision Tree

```
> confusionMatrix(dtree.pred, validate.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  8784  825
       yes  198  365

               Accuracy : 0.8994
                 95% CI : (0.8934, 0.9052)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 0.00000007894

                  Kappa : 0.369

 Mcnemar's Test P-Value : < 0.0000000000000022

            Sensitivity : 0.9780
            Specificity : 0.3067
         Pos Pred Value : 0.9141
         Neg Pred Value : 0.6483
             Prevalence : 0.8830
         Detection Rate : 0.8635
   Detection Prevalence : 0.9447
      Balanced Accuracy : 0.6423

       'Positive' Class : no
```

Accuracy: 0.8994

Decision Tree-Test

```
> confusionMatrix(dtree.test, test.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  3902  391
       yes   90  137

               Accuracy : 0.8936
                 95% CI : (0.8842, 0.9024)
    No Information Rate : 0.8832
    P-Value [Acc > NIR] : 0.01478

                  Kappa : 0.3148

 Mcnemar's Test P-Value : < 0.0000000000000002

            Sensitivity : 0.9775
            Specificity : 0.2595
         Pos Pred Value : 0.9089
         Neg Pred Value : 0.6035
             Prevalence : 0.8832
         Detection Rate : 0.8633
   Detection Prevalence : 0.9498
      Balanced Accuracy : 0.6185

       'Positive' Class : no
```

Accuracy: 0.8936

TAKE-HOME MESSAGES

What we did:

- Attribute conversion
- Complexity reduction
- Cross-validation

What we found:

- Effective approach to data preparation for modeling
- Reliable model in predicting the bank marketing campaign outcomes
- Duration has the highest influence over whether the clients deposit or not
    - Longer duration —> higher success rate
- Systematic approach to improving model accuracy

Application:

- Such knowledge can be used by managers to increase the call time or segmenting audience with a specific goal of focusing more on clients who have previously deposited

Fun Facts

Day 1

hooray!!!

Day 2: To be a foodie

York Fang

Master of Business Analytics at University of California, Irvine - The Paul Merage School of Business

Yuan Liu

MS in Business Analytics at The Paul Merage School of Business, UC Irvine

Takako Suzuki

Data Analyst | Master of Science in Business Analytics @ UCI Merage | Experienced with Python, R, SPSS, Google Analytics

THANK YOU !

Dora Yuan

Marketing Analyst Specialist at Evan Paul Auto Capital

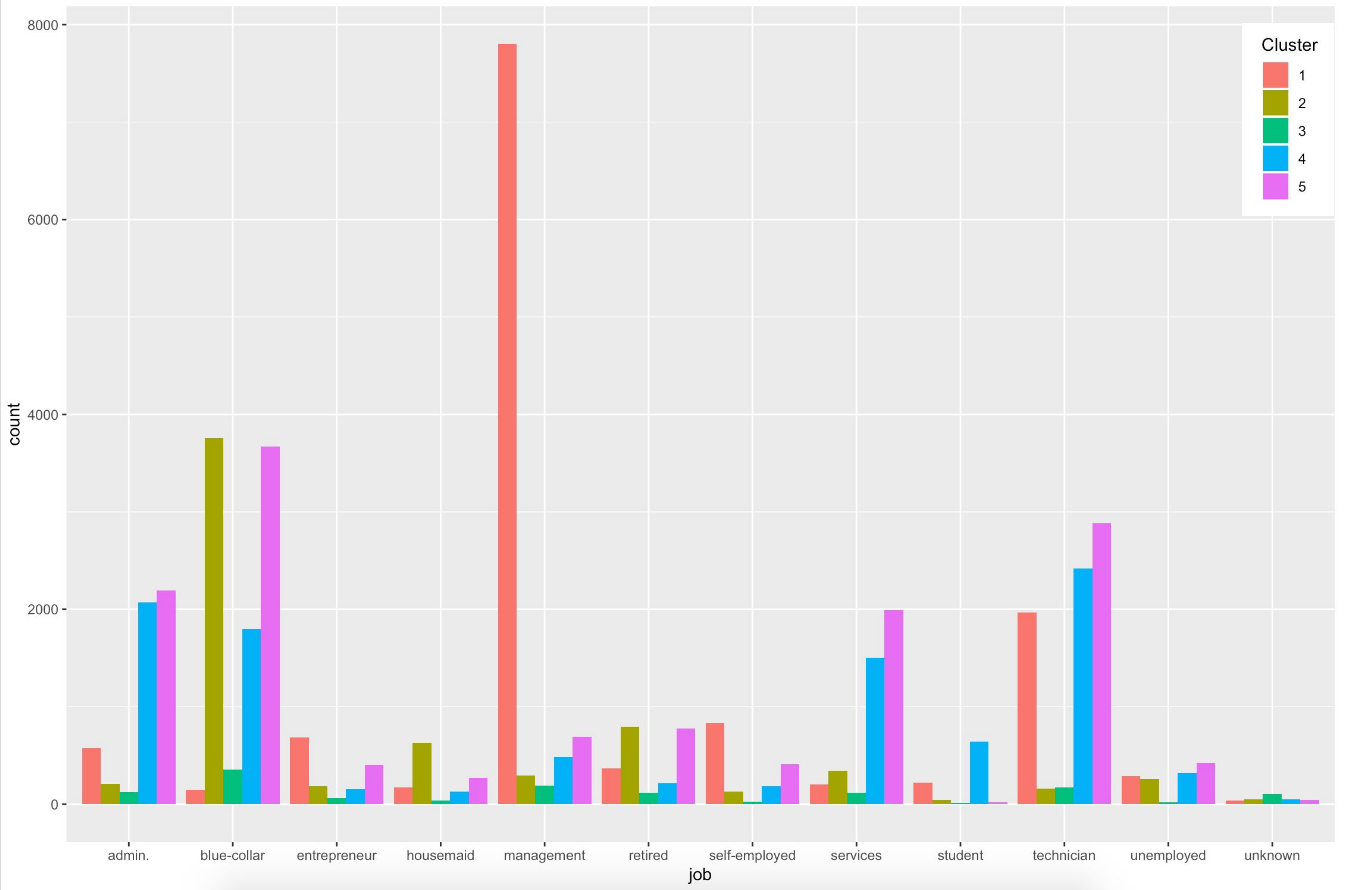
Mia (Chuyan) Zhang

MSBA 20' at UC, Irvine | Student Ambassador

Supplementary Information

Level Reduction via Clustering

Part 2



Bank client data:
Age, marital, education, default, balance, housing, loan

K-mean clustering; Compare distribution

Job group 1:
entrepreneur, management, self-employed

Job group 2:
blue-collar, housemaid, retired

Job group 3:
unknown

Job group 4:
student

Job group 5:
admin, services, technician, unemployed