1042. DATA SCIENCE IN PRACTICE

FINAL PROJECT

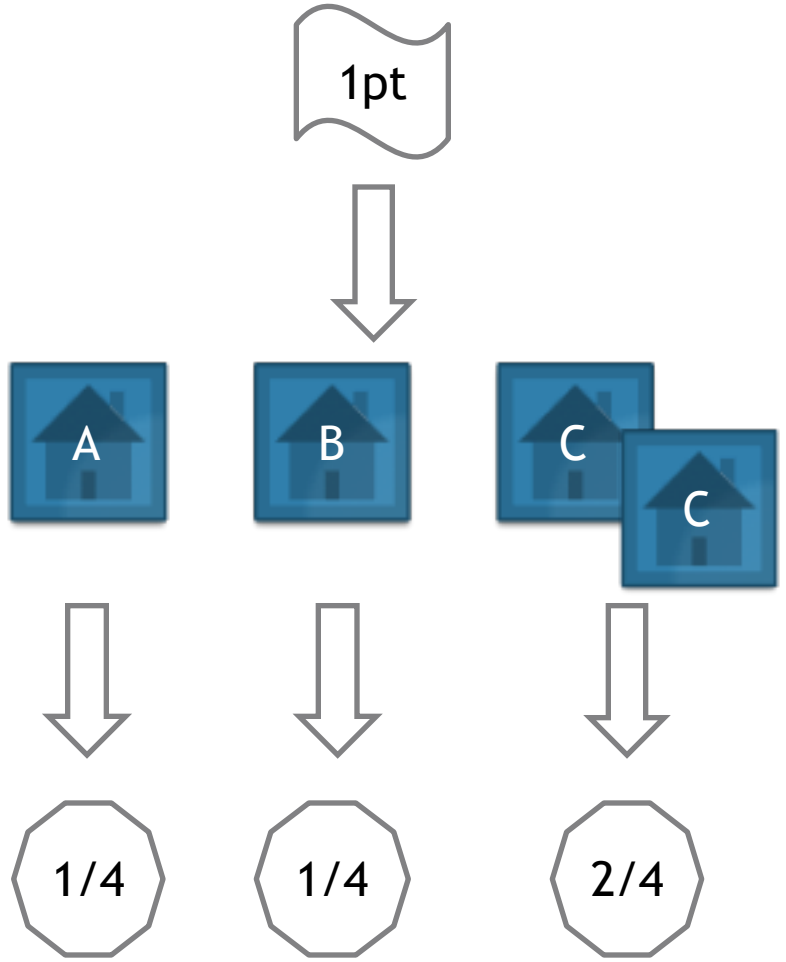
104753001 CasperHsia

Outline

- Input Data
- Goal
- Modeling & Evaluation
- Demo
- Future Work
- Q & A

Input Data

- Competition from KDD 2016
- Rank the affiliation



Input Data

- ~~Competition from KDD 2016~~
- I changed my dataset last night…
- STULONG – Longitudinal Study of Atherosclerosis Risk Factors
  - http://euromise.vse.cz/projects-en/index.php
- Competition from Discovery Challenge 2004

Input Data

- Column
  - ISTAV, VZDELANI, ZODPOV, TELAKTZA, AKTPOZAM, DOPRAVA, DOPRATRV
  - KOURENI, DOBAKOUR, ALKOHOL, VINO, LIHOV, PIVOMN, VINOMN, LIHMN
  - SYST1, DIAST1, SYST2, DIAST2, CHLST, TRIGL, PIVO, BMI
- Missing Value

```
 1 ISTAV,VZDELANI,ZODPOV,TELAKTZA,AKTPOZAM,DOPRAVA,DOPRATRV,KOURENI,DOBAKOUR,ALKOHOL,VINO,LIHOV,PIVOMN,VINOMN,LIHMN,SYST1,DIAST1,SYST2,DIAST2,CHLST,TRIGL,PIVO,BMI,death
 2 1,2,3,3,2,1,5,5,10,2,11,12,1,5,8,120,70,120,75,260,131,0,1,Y
 3 1,4,1,1,2,3,6,5,10,2,11,12,2,5,8,155,90,160,90,301,169,1,1,Y
 4 1,2,3,3,2,3,6,5,10,2,11,12,2,5,8,125,85,120,80,272,199,1,1,Y
 5 1,3,3,3,1,3,6,5,10,3,11,13,2,5,7,140,75,130,75,270,137,1,1,Y
 6 1,3,1,1,2,3,5,5,10,2,11,13,1,5,7,130,85,130,85,232,246,0,0,Y
 7 1,4,3,3,2,0,0,5,10,2,11,12,2,5,8,140,80,140,80,199,129,1,0,Y
 8 1,4,0,1,2,3,6,3,9,2,11,13,1,5,7,170,105,155,110,232,131,0,1,Y
 9 1,1,5,0,1,0,0,3,10,1,12,13,11,11,11,170,100,190,100,227,127,0,0,Y
10 2,1,3,3,2,3,6,4,10,3,11,12,2,5,8,140,75,135,80,0,299,1,1,Y
11 1,2,3,2,2,3,5,1,0,2,11,12,2,5,8,160,100,170,100,227,90,1,1,Y
12 1,1,3,3,1,1,5,4,10,3,11,12,3,5,8,115,80,120,85,236,79,1,0,Y
13 1,2,3,4,1,3,6,5,10,3,12,13,2,4,7,150,90,150,90,232,117,1,1,Y
14 3,1,3,2,2,1,5,5,10,3,12,13,3,4,7,115,85,115,85,202,135,1,1,Y
15 1,3,2,3,2,3,7,5,10,2,11,13,2,5,7,130,80,150,90,179,196,1,0,Y
```

Input Data

- 713 people
- 23 features
- 124 death

Goal

- Predict death or alive

Modeling & Evaluation

- libSVM for R
  - N-fold cross validation
  - alive as positive

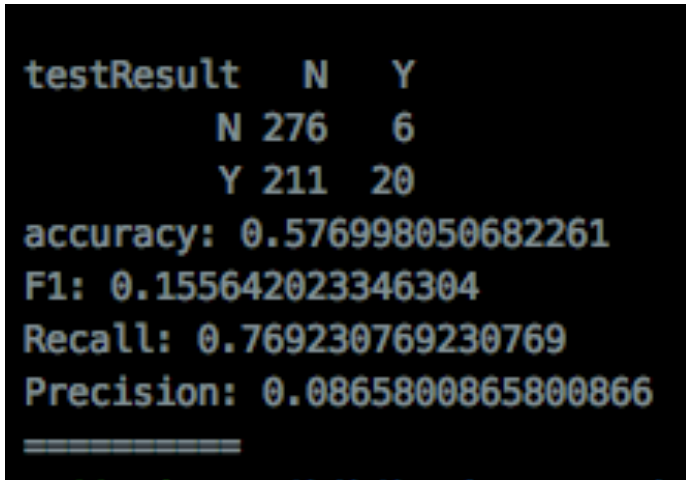
  - How about death as positive?
  - 214:499

```
testResult  N  Y
         N 62  9
         Y  0  0
accuracy: 0.873239436619718
F1: 0.932330827067669
Recall: 1
Precision: 0.873239436619718
==========

testResult  N  Y
         N 62  9
         Y  0  1
accuracy: 0.875
F1: 0.932330827067669
Recall: 1
Precision: 0.873239436619718
==========
```

Modeling & Evaluation

- Control the training data
  - death as positive
  - 50% death
  - 50% alive



Modeling & Evaluation

- Feature selection
  - ZODPOV (type of the job)
  - KOURENI (smoke frequency)
  - DOBAKOUR (smoke age)
  - LIHMN (drink spirits)
  - SYST1 (Systolic blood pressure)
  - DIAST1 (Diastolic blood pressure)
  - CHLST (cholesterol)
  - PIVO (drink beer)

```
testResult   N   Y
         N 299   7
         Y 188  19
accuracy: 0.619883040935672
F1: 0.163090128755365
Recall: 0.730769230769231
Precision: 0.0917874396135266
==========
```

Modeling & Evaluation

- Other features

```
testResult   N   Y
         N 316  15
         Y 171  11
accuracy: 0.637426900584795
F1: 0.105769230769231
Recall: 0.423076923076923
Precision: 0.0604395604395604
==========
```

Demo

- https://summer.shinyapps.io/finalProject/

Future Work

- Iteration feature selection
- Complete the visulization
- Try other models

Q & A

- Thanks for your attention!!
- And please don’t ask too much.