Feature Ranking via Eigenvector Centrality

Team 35:

- Ivan Baybuza
- Aleksandr Belov
- Mariia Kopylova
- Miron Kuznetsov

Skoltech
Skolkovo Institute of Science and Technology

Introduction

- Dataset: 165k clinical records by 55k patients
- Source of data: Moscow Cardiology center
- Records parsed to 88 features (used OHE, text processing)
- 1.6×10^7 values in matrix X
- Matrix is sparse, only 13% are not NaN

Skoltech
Skolkovo Institute of Science and Technology

Problem statement

- Target variable is “fibrillation” (“фибрилляция предсердий”)
- Trying to find predictors of this diagnosis

➔ Project by Dmitrii Dylov with Moscow Cardiology Center

➔ Doctors have many different scales (i.e. on the right)

Число присваиваемых баллов (максимальный суммарный балл для каждого пациента – 9)	**Признак**
1 балл	Женский пол
1 балл	Возраст от 65 до 74 лет
2 балла	Возраст 75 лет и старше
1 балл	Артериальная гипертония

Skoltech
Skolkovo Institute of Science and Technology

Main idea from paper [1]

Graph-based method for feature selection:

$$X = \{x^{(1)}, ..., x^{(n)}\}$$

$G = (V, E)$ → Adjacency matrix A

$$a_{ij} = \varphi(x^{(i)}, x^{(j)})$$

ECFS algorithm

$$f_i = \frac{\left|\mu_{i,1} - \mu_{i,2}\right|^2}{\sigma_{i,1}^2 + \sigma_{i,2}^2}$$

$$\Sigma(i,j) = max\left(\sigma^{(i)}, \sigma^{(j)}\right)$$

$$m_i = \sum_{y \in Y} \sum_{z \in x^{(i)}} p(z,y) log\left(\frac{p(z,y)}{p(z)p(y)}\right)$$

$$k = (f \cdot m^\top)$$

$$A = \alpha k + (1 - \alpha)\Sigma$$

where $a_{ij} = \varphi(x^{(i)}, x^{(j)})$ is a pairwise potential term

ECFS algorithm

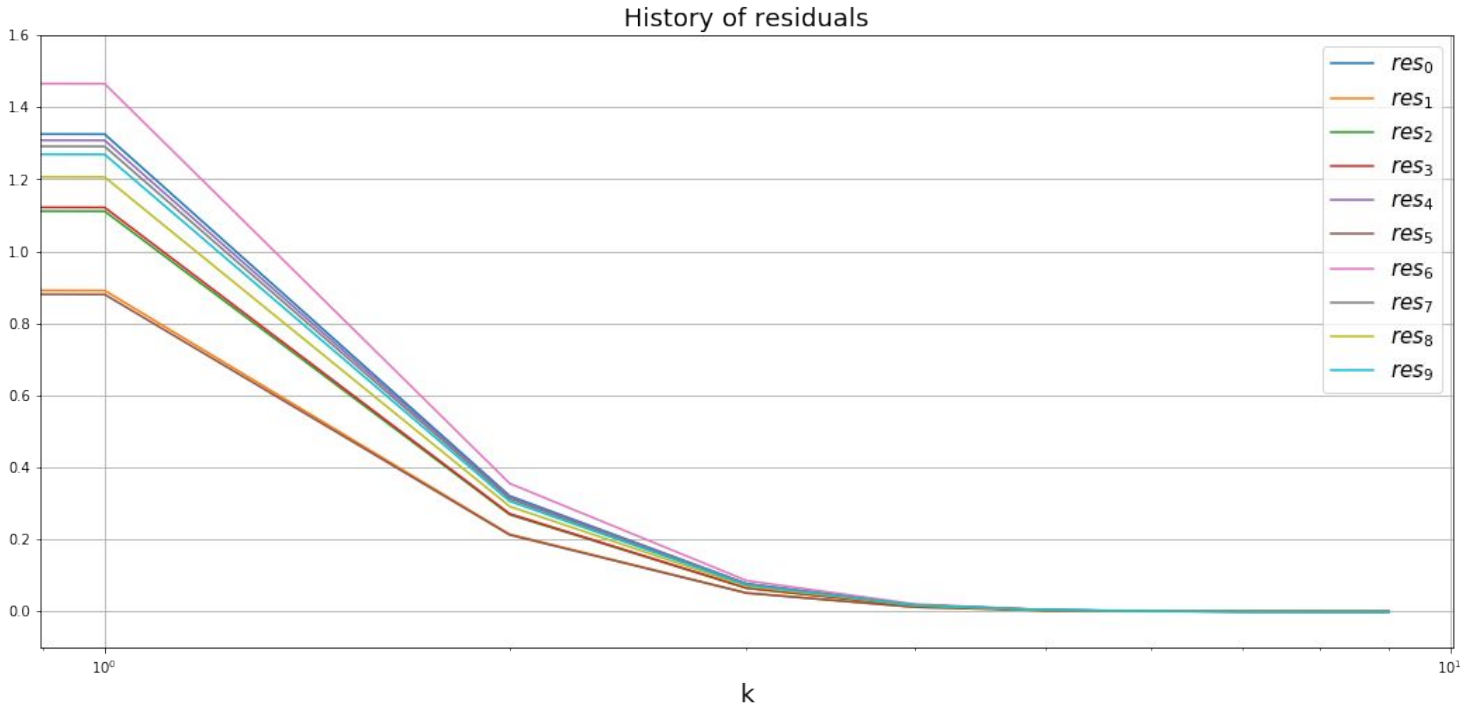
Skoltech
Skolkovo Institute of Science and Technology

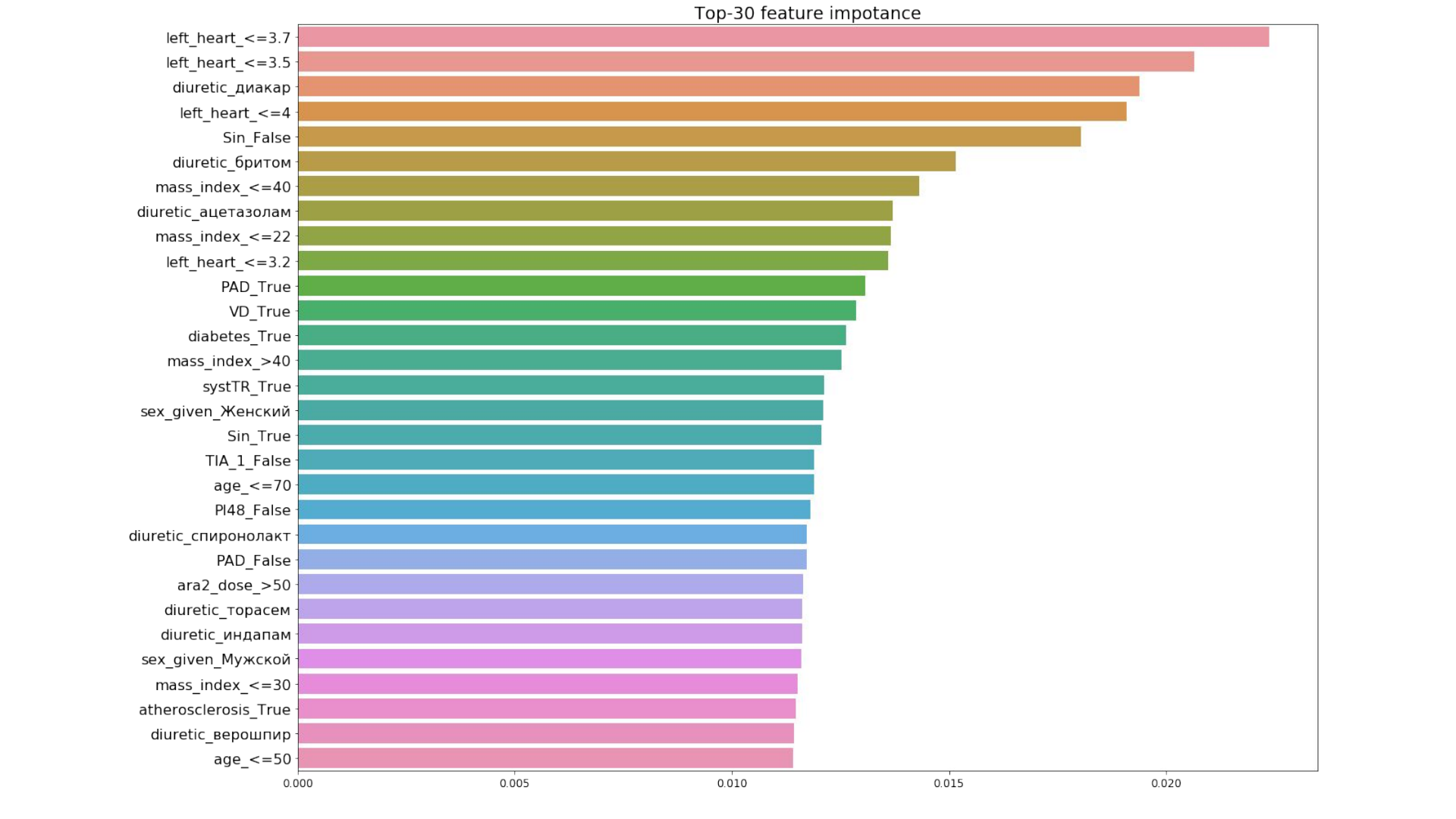
Desired rank vector is eigenvector of matrix A associated to max eigenvalue.

⇒ Use the Power method !!!

$$A^n e \to v_1$$

10 iterations is more than enough




Top-30 feature impotance
left_heart_<=3.7
left_heart_<=3.5
diuretic_диакар
left_heart_<=4
Sin_False
diuretic_бритом
mass_index_<=40
diuretic_ацетазолам
mass_index_<=22
left_heart_<=3.2
PAD_True
VD_True
diabetes_True
mass_index_>40
systTR_True
sex_given_Женский
Sin_True
TIA_1_False
age_<=70
PI48_False
diuretic_спиронолакт
PAD_False
ara2_dose_>50
diuretic_торасем
diuretic_индапам
sex_given_Мужской
mass_index_<=30
atherosclerosis_True
diuretic_верошпир
age_<=50
0.000
0.005
0.010
0.015
0.020

Skoltech
Skolkovo Institute of Science and Technology

Obtained results

The main predictors of Atrial fibrillation is:

1) Large size of the left heart *(> 3.2)*
2) Large Mass index of a patient *(2 degree of obesity and more)*
3) An absence of Sinus tachycardia
4) A presence of diabetes

Skoltech
Skolkovo Institute of Science and Technology

Conclusion

- The investigated *Eigenvector Centrality* method may be applicable to this type of problems and could give some unobvious information
- The results should be delivered to the specialists in Cardiology Center to be considered by them

Skoltech
Skolkovo Institute of Science and Technology

References

1. Ranking to Learn: Feature Ranking and Selection via Eigenvector Centrality by Giorgio Roffo and Simone Melzi (2017)
2. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients
3. Entity Embeddings of Categorical Variables by Cheng Guo and Felix Berkhahn (2016)
4. https://github.com/albellov/nla_project