Recommender systems for career guidance

Trần Đức Thế, **Evgeny Cherkashin**, Viktoria Kopylova, Nikita Lukyanov

Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia,
eugeneai@icc.ru
National research Irkutsk state technical university (INRTU), Irkutsk, Russia,
kopylovika@mail.ru

AIIT-2021, October, 15

Introduction

Recommender systems (RS) are *decision support systems*. They are useful to support users with additional information and decision variants. We consider two RS in career guidance field.

Both systems were developed within master degrees at Institute for Information Technologies and Data Analysis, National research Irkutsk state technical university.

Relevance is justified by

- In the **countries with young population**, such as Việt Nam, the **demand of young specialists is high**. In fact, quite often graduates remain unemployed for a long time or do not work in originally mastered specialties.
- **University** institutions **are interested in rising quality of entrants**: attract attention of domain targeted students with high potentials to the existing courses, organize new courses for target student groups.

Interest assessment

In the tables, we present the result of interest assessment to career guidance and relevance of improving of the guidance system.

Table: Interest assessment to the carer guidance in Việt Nam, Hà Tĩnh city

Interest level	**Count of votes**	**Ratio, %**
Very interested	218	51.9
Relatively interested	155	36.9
Less interested	36	8.6
No interest	11	2.6
Total	**420**	**100**

Table: Questionary results of middle school students on need for improvement of the carer guidance system

Answer option	**Quantity**	**Ratio, %**
very necessary	272	64.8
necessary	145	34.5
no necessity	3	0.7
Total	**420**	**100**

Background

RS classification by solved problems

1. **Increasing** the **sales** of a product
   - a **number** of commodities sold
   - organizing **wider range** product sales
2. **Increasing** user **satisfaction** and/or loyalty
3. Better understanding **user needs**
4. **Better products offers** with respect to the user needs
5. Selection of sets of **products** for users **with a common properties**
   - "good" ones, and
   - product groups, having a common usability properties
6. Mining the classes of products, **structuring** the RS **product domain**, and
7. Generate a **continuity of recommendations** using the classification
8. **Rectifying user profile**, *e.g.* with targeted questioning
9. For the users having no goal to make a choice, but searching for an expert opinions
   - **Analysis of** other **users impact** on a choice
   - Formalizing opinions
   - Recommending opinions

General stages of RS design and implementation

RS development is being carried on according to the following scenario:

1. Investigation of the domain and defining the **users** and **objects** of recommendation.
2. Determine the **sources** of information.
3. Develop techniques for solving "**cold start**" problems.
4. Propose **methods and algorithms of comparing** objects and users, and their **mutual relations**.
5. Implement data structures, storage, analysis and synthesis algorithms.
6. Testing.

In both systems **users** are **middle school students**, **entrants**, and **high school students**. **Objects** are references to professions expressed as **universities and their departments**.

Sources of information: *collecting refinement data*

https://letstalkscience.ca/educational-resources/lessons/work-interests-and-holland-code

1. Application of John Holland's Six Interest Types Technique for **structuring** school students' **interests**. This requires 54 questions to be answered.
2. Relate local universities to the types by
3. obtaining students' ratings on universities

Then the resulting recommendations will be generated as Top-N from an interest type. Thus, the **cold start** problem has been solved mostly by applying **expert knowledge**.

Sources of information: *social network data*

Social networks are popular ways of self-expression for school students and entrants. User's properties are the profile data and set of group and channel subscriptions represented with **tags, keywords**. TSU[1] research showed that 'In Contact' (В Контакте) students profiles are closely related to educational interests.

1. Students of first three courses were asked their In Contact profile names. (**6270** from **7 institutes of INRTU**)
2. The profile data were obtained via In Contact API (**4325, 68.98%, 480 features**).
3. Data were filtered
   - Removed inaccessible 'private' profiles (**686**)
   - Removed content prohibited by law (**389** features)

Students' distribution among INRTU institutes was obtained by means of **parsing enrollment documents**. Enrollments will be related to the set of subscriptions and other profile characteristics (features).

This solves **cold start** problem as well.

[1] Tomsk State University, Tomsk Russia

Object and user comparison

Pearson correlation was applied to the user ratings of universities and the correlation matrix were produced, $r_{u,j}$ is a rating of u-th user given for i-th university.

$$sim(i,j) = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}}$$

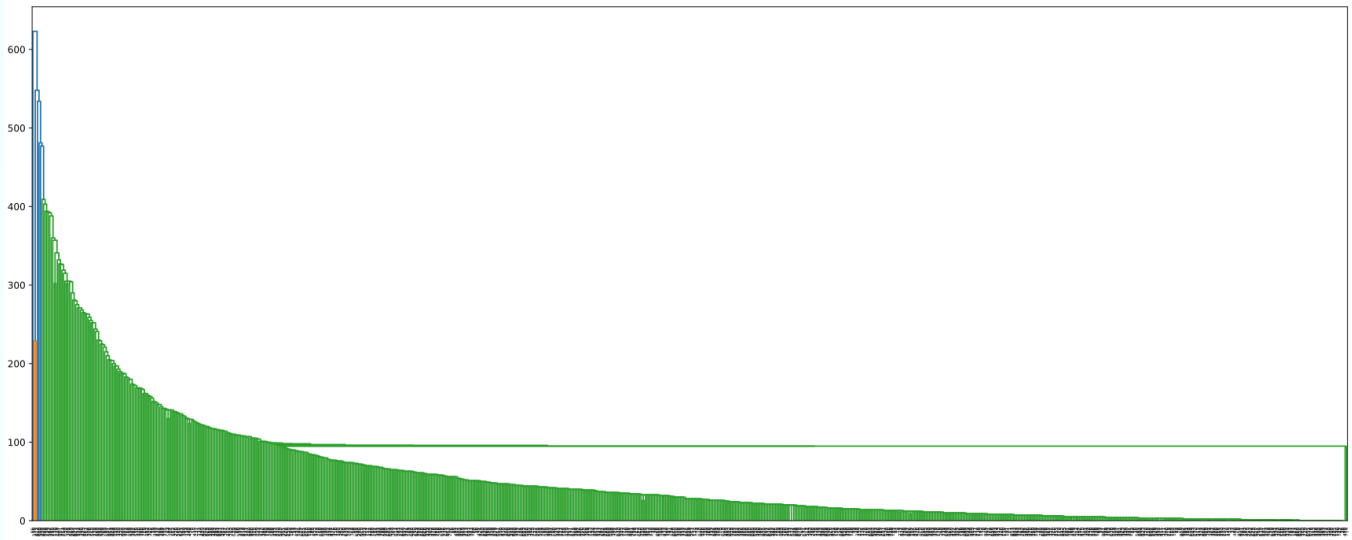
The correlation matrix for the universities is obtained similarly.

Name_Uni	Academy of Finance	An College of Culture and Arts	Asian University of Industrial Fine Arts	Bac Giang University of Agriculture and Forestry:	Banking Academy	Banking University of Ho Chi Minh City	Binh Duong University of Economics and Technology	Can Tho University	Can Tho University of Medicine and Pharmacy	Centr Constructio Universit
Name_Uni										
Academy of Finance	1.000000	-0.042206	-0.061459	-0.266602	-0.135564	-0.134517	0.121472	0.212420	0.453148	-0.05242
An College of Culture and Arts	-0.042206	1.000000	0.266658	0.097767	0.248917	-0.013975	0.044716	-0.084712	0.178497	0.09798
Asian University of Industrial Fine Arts	-0.061459	0.266658	1.000000	-0.109866	0.090173	0.096562	-0.063758	-0.401843	-0.025557	-0.67279
Bac Giang University of Agriculture and Forestry:	-0.266602	0.097767	-0.109866	1.000000	-0.042479	-0.056320	-0.086769	0.375120	-0.230425	0.29353

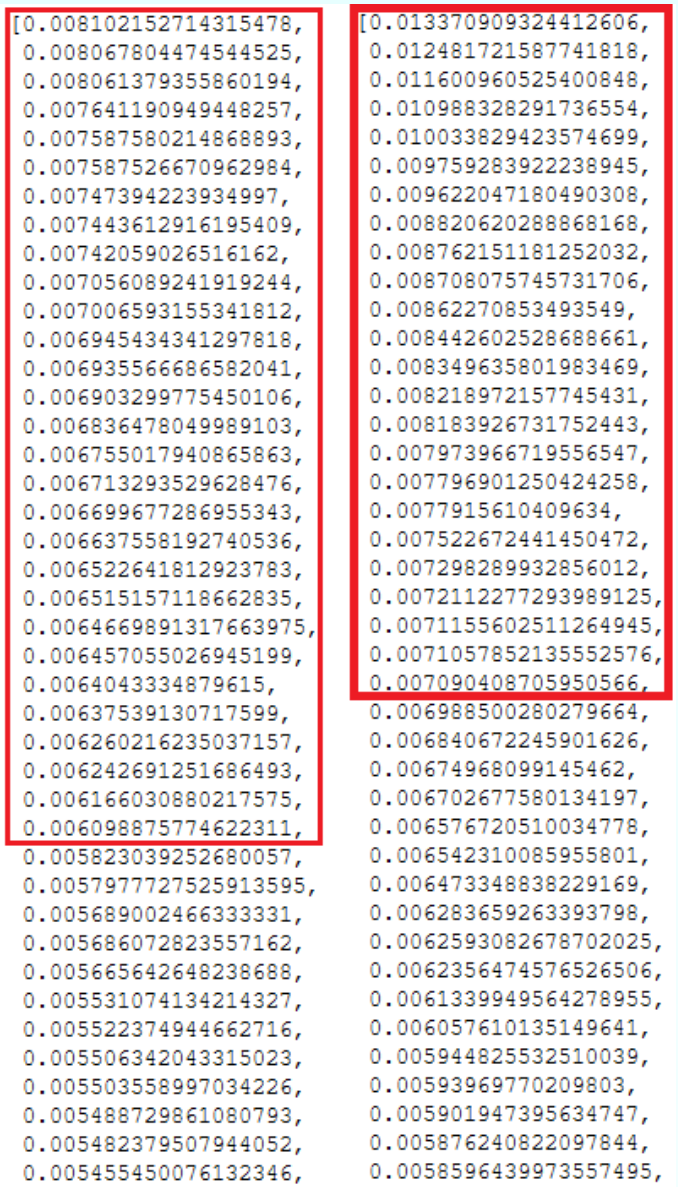
Relating INRTU institutes with students

Experiments were carried on to figure out the best feature set to be used in train set. Namely

- Induce a taxonomy of subscription group by application of a hierarchical clustering.
- Explore informative values of the groups as features by application of principal component analysis.
- Deduce the values with construction of a decision tree.



Principal Component Analysis



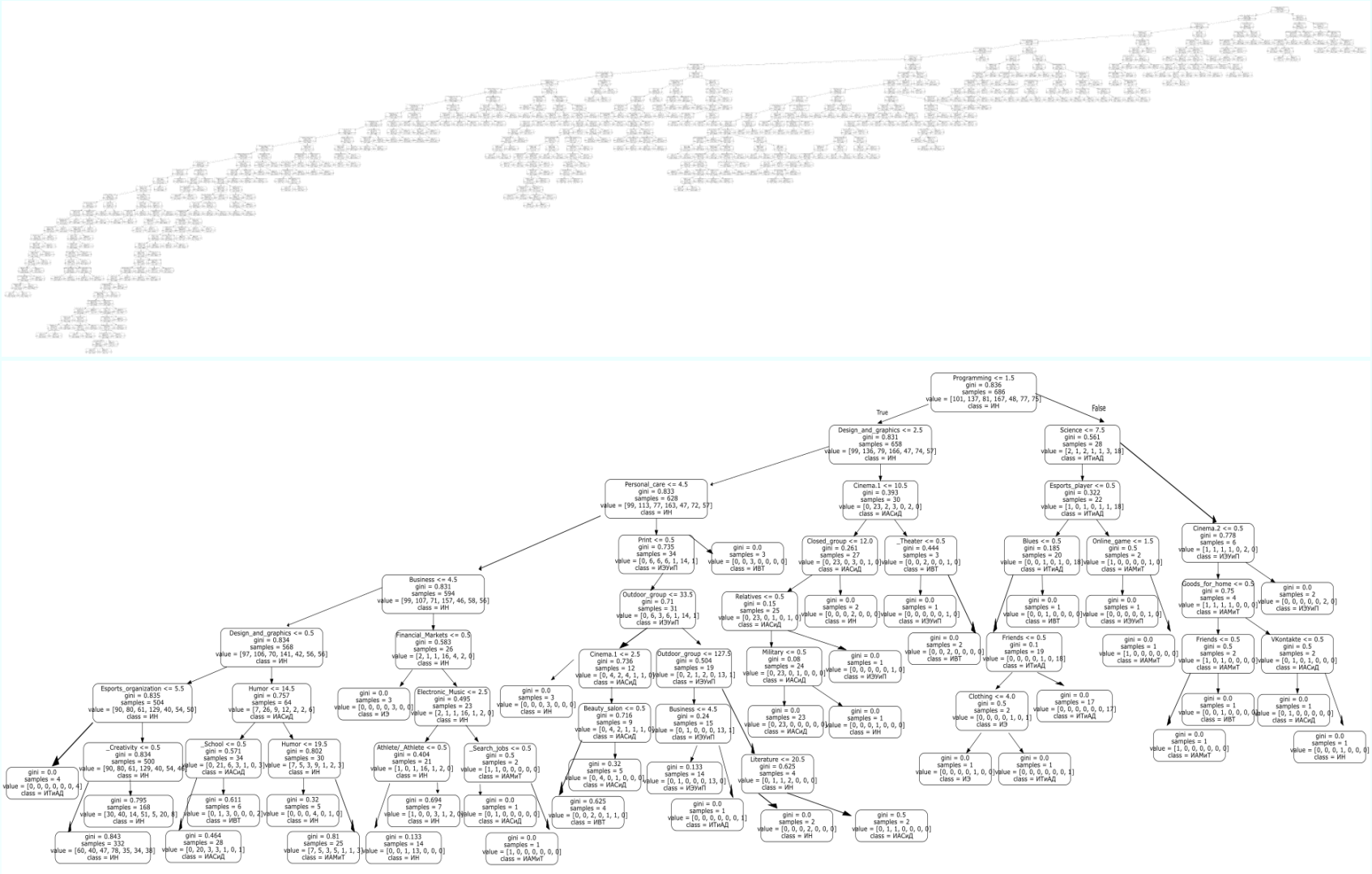
```
[0.008102152714315478,
 0.008067804474544525,
 0.008061379355860194,
 0.007641190949448257,
 0.007587580214868893,
 0.007587526670962984,
 0.00747394223934997,
 0.00744361291619540 9,
 0.00742059026516162,
 0.007056089241919244,
 0.007006593155341812,
 0.006945434341297818,
 0.006935566686582041,
 0.006903299775450106,
 0.006836478049989103,
 0.006755017940865863,
 0.006713293529628476,
 0.006699677286955343,
 0.006637558192740536,
 0.006522641812923783,
 0.006515157118662835,
 0.006466989131766397 5,
 0.006457055026945199,
 0.0064043334879615,
 0.00637539130717599,
 0.006260216235037157,
 0.006242691251686493,
 0.006166030880217575,
 0.006098875774622311,
 0.005823039252680057,
 0.005797727527513595,
 0.005689002466333331,
 0.005686072823557162,
 0.005665642648238688,
 0.005531074134214327,
 0.005522374944662716,
 0.005506342043315023,
 0.005503558997034226,
 0.005488729861080793,
 0.005482379507944052,
 0.005455450076132346,
```

```
[0.013370909324412606,
 0.012481721587741818,
 0.011600960525400848,
 0.010988328291736554,
 0.010033829423574699,
 0.009759283922238945,
 0.009622047180490308,
 0.008820620288868168,
 0.008762151181252032,
 0.008708075745731706,
 0.00862270853493549,
 0.008442602528688661,
 0.008349635801983469,
 0.008218972157745431,
 0.008183926731752443,
 0.007973966719556547,
 0.007796901250424258,
 0.0077915610409634,
 0.007522672441450472,
 0.007298289932856012,
 0.007211227729398912 5,
 0.007115560251126494 5,
 0.007105785213555257 6,
 0.007090408705950566,
 0.006988500280279664,
 0.006840672245901626,
 0.0067496809914546 2,
 0.006702677580134197,
 0.006576720510034778,
 0.006542310085955801,
 0.006473348838229169,
 0.006283659263393798,
 0.006259308267870202 5,
 0.006235647457652650 6,
 0.006133994956427895 5,
 0.006057610135149641,
 0.005944825532510039,
 0.0059396977020980 3,
 0.005901947395634747,
 0.005876240822097844,
 0.005859643997355749 5,
```

The chosen features by the red rectangles are the sets with the same informative 'capabilities'.

There are no explicit set of features, describing varieties of data for PC1 and PC3.

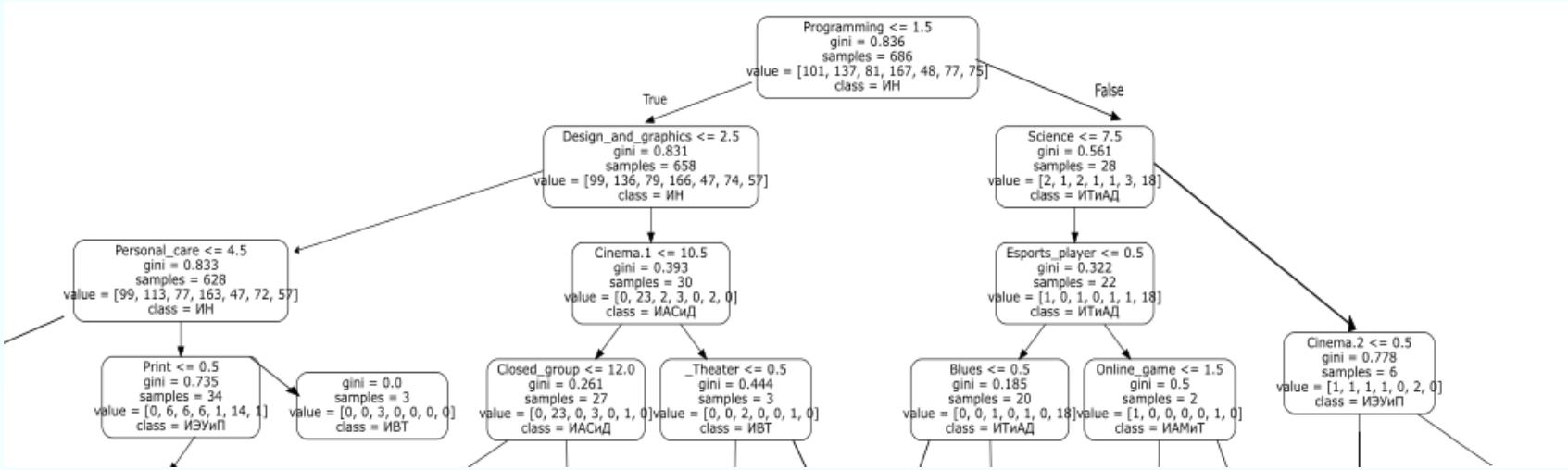
Decision trees



Decision trees

The top-13 features describe only 20% of phenomena.

- Programming
- Design & graphics
- Personal care
- Cinema.1 (probably novel, documental)
- Science
- Sports player
- Print, Theater, Blues, Online games, Cinema.2 (probably fantastic)



Artificial Neural Nets

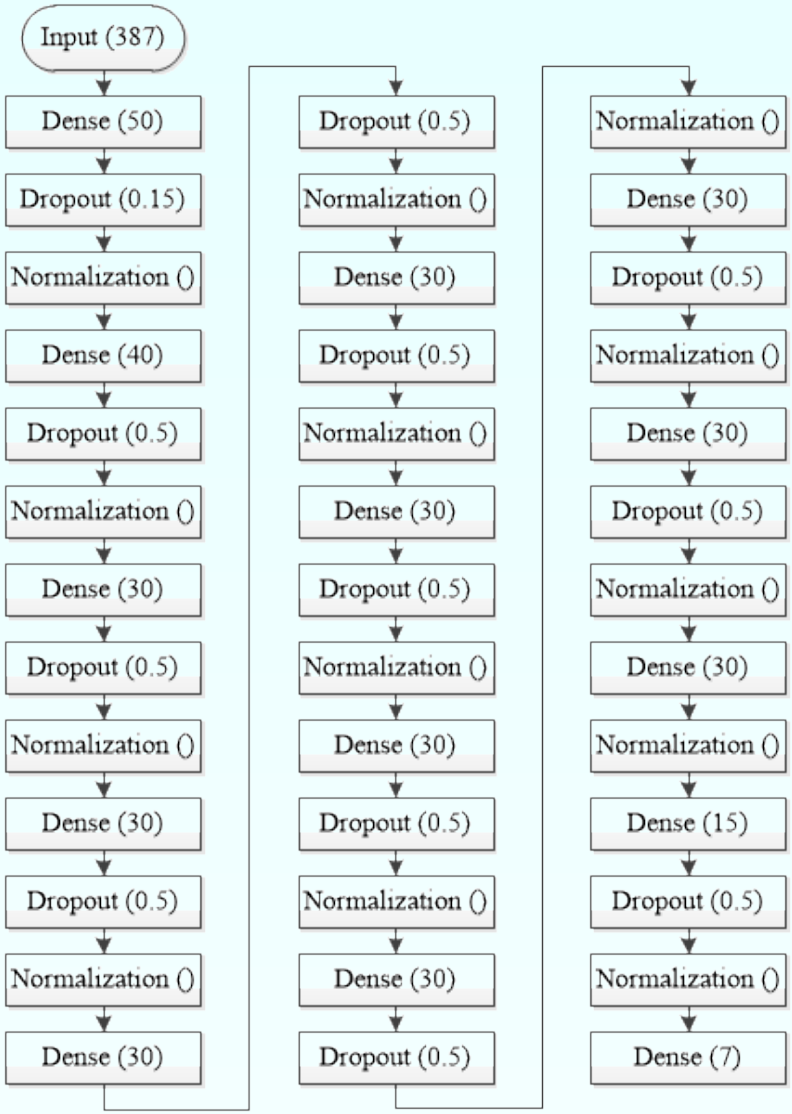
The training results

Best precision: 23.96%
Best model: FastForestOva
Training time: 84 008 sec
Models explored: 14

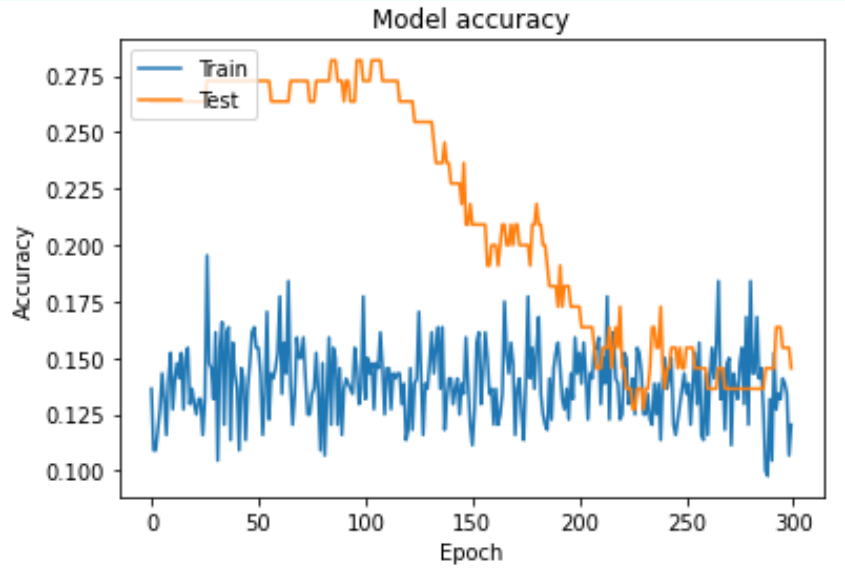
Library: ML.NET

```
C:\Users\Виктория\Downloads\test.csv
|     Trainer                              MicroAccuracy  MacroAccuracy  Duration #Iteration                    |
Время обучения завершено без обученных моделей.
|     Trainer                              MicroAccuracy  MacroAccuracy  Duration #Iteration                    |
Время обучения завершено без обученных моделей.
|     Trainer                              MicroAccuracy  MacroAccuracy  Duration #Iteration                    |
Время обучения завершено без обученных моделей.
|     Trainer                              MicroAccuracy  MacroAccuracy  Duration #Iteration                    |
|1    AveragedPerceptronOva                       0,2078        0,1442    1528,3          1                    |
|2    SdcaMaximumEntropyMulti                     0,2037        0,1309    1542,8          2                    |
Измерения массива превысили поддерживаемый диапазон.
|3    SymbolicSgdLogisticRegressionOva            0,1777        0,1320     684,5          3                    |
|4    FastTreeOva                                 0,2074        0,1564    1225,4          4                    |
|5    LinearSvmOva                                0,2104        0,1406     377,8          5                    |
|6    LbfgsLogisticRegressionOva                  0,2070        0,1382   66430,9          6                    |
|7    SgdCalibratedOva                            0,1801        0,1347     309,2          7                    |
|8    FastForestOva                               0,2396        0,1735    1287,6          8                    |
Выдано исключение типа "System.OutOfMemoryException".
|9    AveragedPerceptronOva                       0,2010        0,1292    3138,0          9                    |
|10   LinearSvmOva                                0,2046        0,1381     820,0         10                    |
|11   FastForestOva                               0,2360        0,1672    1248,0         11                    |
Недостаточно места на диске.


==============================================Experiment Results=============================================
-------------------------------------------------------------------------------------------------------------
|                                                  Summary                                                  |
-------------------------------------------------------------------------------------------------------------
|ML Task: multiclass-classification                                                                         |
|Dataset: C:\Users\Виктория\Downloads\test.csv                                                              |
|Label : col0                                                                                               |
|Total experiment time : 84009,7908504 Secs                                                                 |
|Total number of models explored: 14                                                                        |
-------------------------------------------------------------------------------------------------------------

|                                             Top 5 models explored                                         |
-------------------------------------------------------------------------------------------------------------
|     Trainer                              MicroAccuracy  MacroAccuracy  Duration #Iteration                    |
|1    FastForestOva                               0,2396        0,1735    1287,6          1                    |
|2    FastForestOva                               0,2360        0,1672    1248,0          2                    |
|3    LinearSvmOva                                0,2104        0,1406     377,8          3                    |
|4    AveragedPerceptronOva                       0,2078        0,1442    1528,3          4                    |
|5    FastTreeOva                                 0,2074        0,1564    1225,4          5                    |
-------------------------------------------------------------------------------------------------------------
```

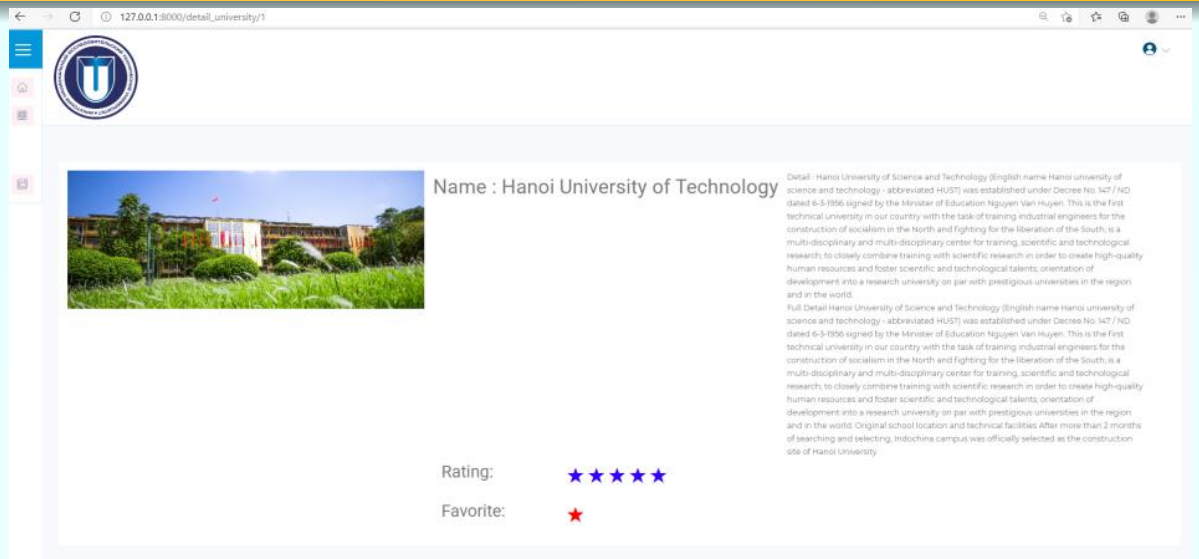
Resulting neural model



- One multilayer neural network.
- Recommending **four** directions instead **one**
- Precision: 86% (testing set) 22% (learning set)
- Library: Keras (Python)

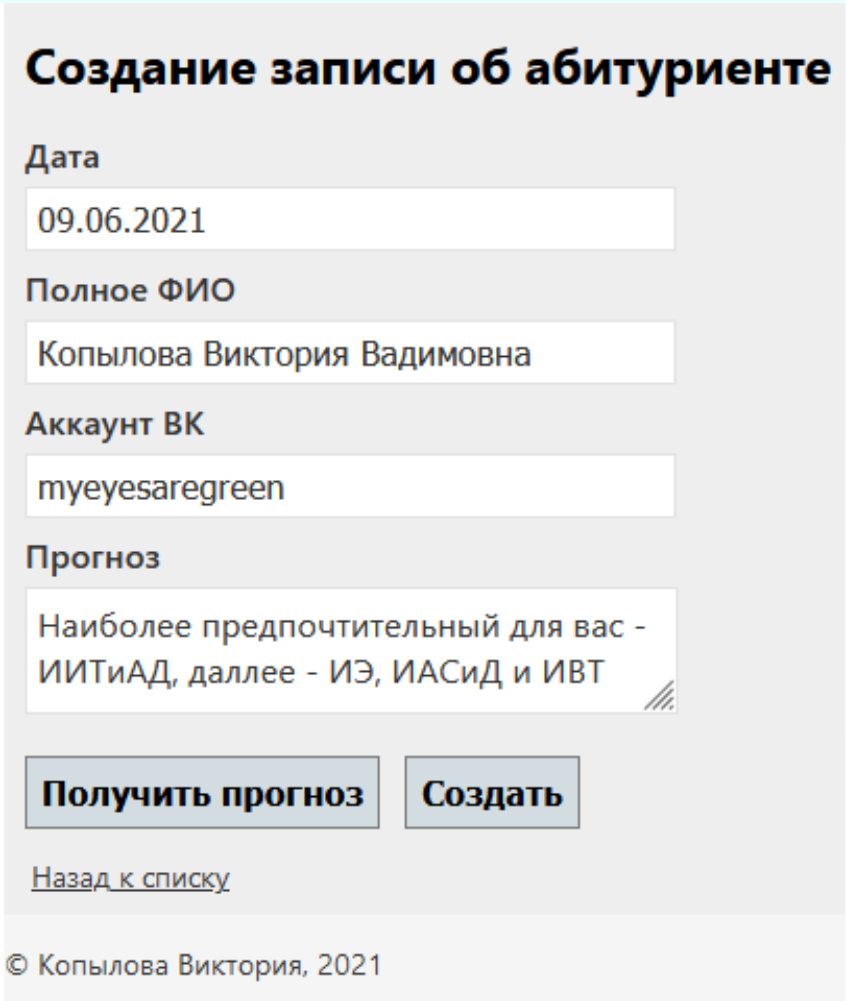


Testing


127.0.0.1:8000/detail_university/1
Name : Hanoi University of Technology
Rating:
Favorite:


Recommended University for University of Social Sciences and Humanities - Hanoi National University are:
University of Social Sciences and Humanities - Hanoi National University
Binh Duong University of Economics and Technology
University of Technology - The University of Da Nang
Labor - Social University (Hanoi Campus)
Nguyen Tat Thanh University
University of Education - Hue University
University of Economic and Technical Industry
Polytechnic University, National
University of Science - Thai Nguyen University

Testing



Stages of recommendation generation

1. Access In Contact account via API
2. Load subscriptions and profile data
3. Conversion to features
4. Apply neural net for classification
5. Print output

Discussion

Ha Tinh RS provides top-6 universities and was not measured on required criteria (precision, recall). It is intended to use as a web pages on the site of a Việt Nam school, and its purpose is in interpretation of the J. Holland's technique.

For the INRTU RS some analysis of the results was carried out.

- The wrongly classified (out of 86% precision) test set contained 9 students, which stopped studying, 3 study economics at IT institution, and for 7, no explanation were found.
- There are institutes having too wide set of specialties even contradictory to their generic intention. *E.g.* **IIT&DA** (our institute) have economics as a direction, **IAM&T** does not limit itself with aircraft engineering, engineering general and computer graphics.

The TGU's hypothesis on well correlated set of subscriptions and faculty direction was **not confirmed**, so a deep learning have to be used to obtain 'good' model.

Conclusion

We presented results of two master degree projects. Both are finished, and present somewhat different approaches to solve problems in a common domain of career guidance.

1. Authors carried out all necessary stages for producing techniques for RS software design.
2. The proposed techniques were implemented, using popular web frameworks.
3. Analysis of the results with different level of detail has been provided.
4. Recommender system MVP-s were constructed for introduction into school career guidance.

Thank You for attention!