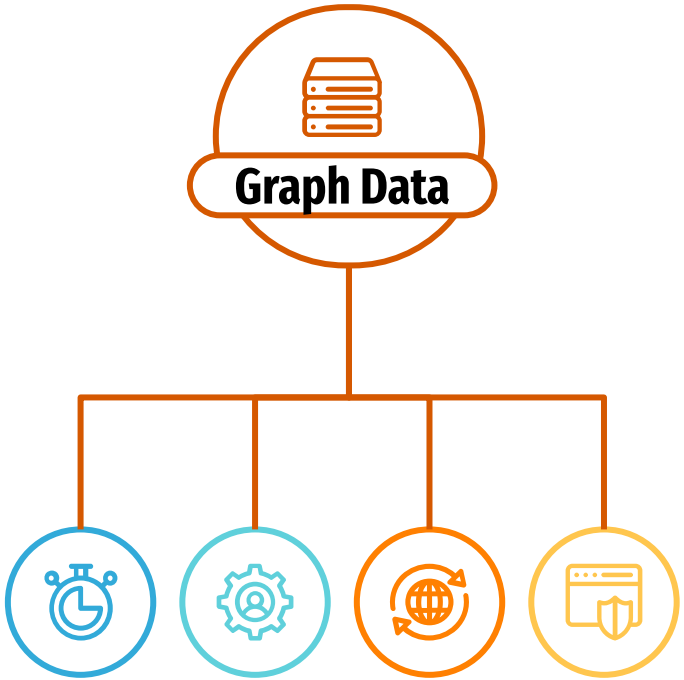
Link Prediction on Future Research Topics

- Science4cast competition 2021

Ying Ding
Aprajita Arora
Zeliang Liu
Abhimanyu Soni



AGENDA

Introduction

Data exploration & preprocessing

Feature Engineering

Model Evaluation & comparison

Conclusion

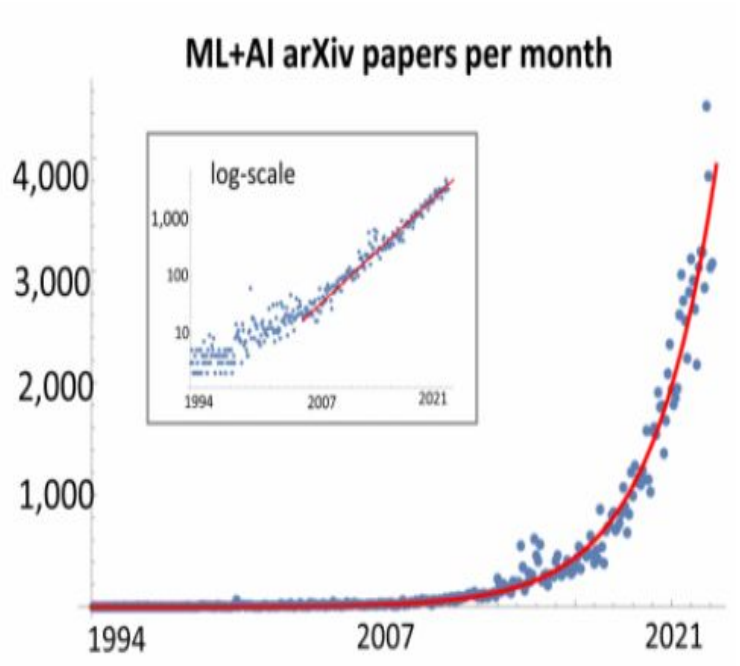
Introduction

In the field of AI&ML, the number of papers grows exponentially and doubles approximately every 23 months.

Researchers have to specialize in narrow sub-disciplines, making it challenging to uncover scientific concepts and connections beyond their own area of research.

To explore beyond the specialized areas, research ideas need to transcend the individual focus bubbles.

A tool that could offer such meaningful, personalized scientific ideas would open new avenues of research.



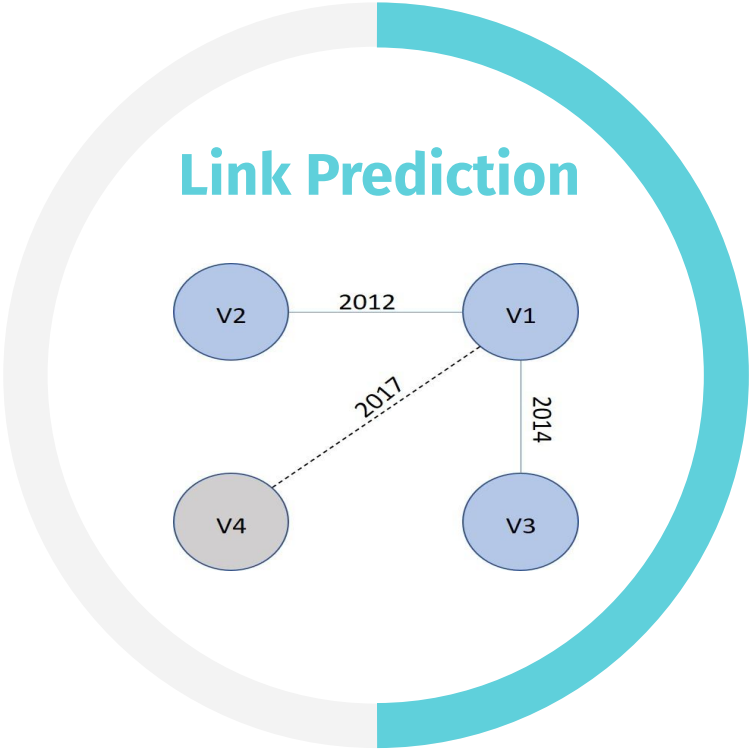
Science4cast Dataset (2021 IEEE BigData Cup Challenges)

64,719

The dataset has a network of 64719 AI concepts (nodes) and we need to predict future research topics.

7,652,945

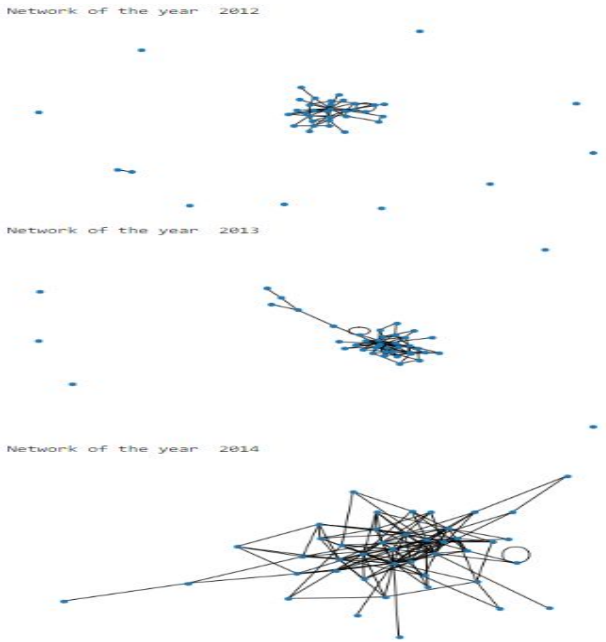
By the end of 2017, there are 7,652,945 edges. If two nodes are connected means these two AI topics are researched together.



1994- 2017

The datasets has the timestamp between 1994 and 2017. It will be used as training datasets. Later, it is used to predict the topics for 2020.

Data exploration

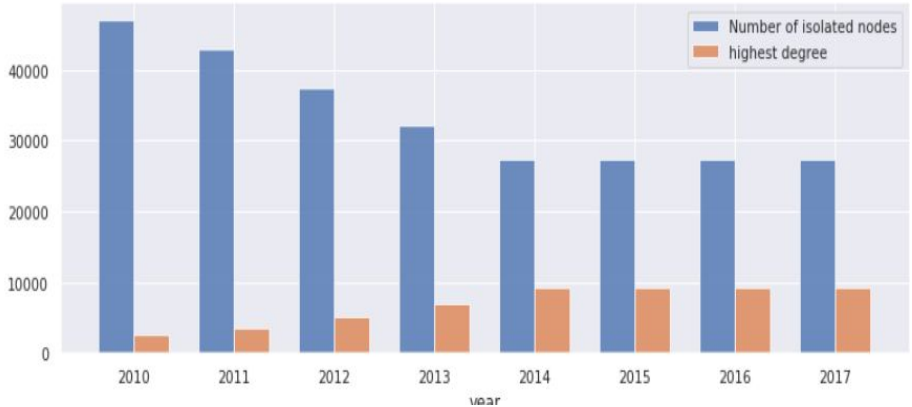


Visulisation of network

Decrease in the number of Isolated Nodes

Exponential Increase in Node degree

Incomplete Training Data (2014-2017)



Isolated vertices vs highest degree

Feature Engineering

01 Degree Based Feature

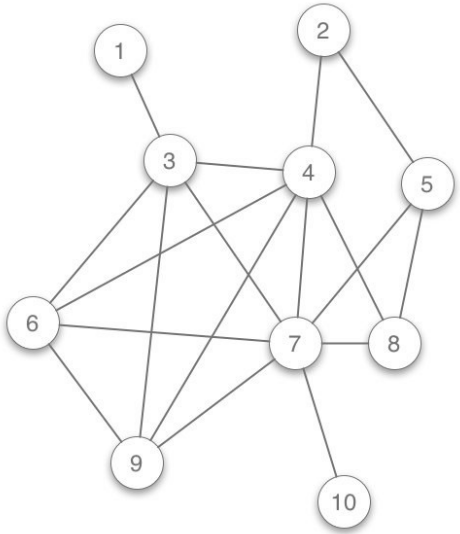
For nodes (v1, v2), we compute individual degrees, sum and difference of degrees

02 Common Neighbor

Common Neighbors is defined as the number of vertices that are among the intersection of their one-hop neighborhood

03 Average Neighborhood Degree

Average Neighbor Degree of vertex u is defined as the average of vertex degrees of neighbors of u



04 Jaccard Similarity

Jacard similarity is a measure to show how two vertices in the graph are similar

05 Page Rank

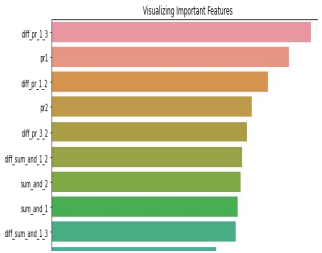
Assigns a score to each vertex based on its link. The score can be considered as a measure of importance in the network

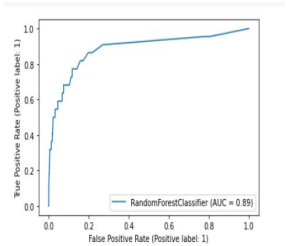
06 Shortest Path

Shortest path is the minimum distance between two vertices in a graph.

Model Evaluations

Random Forest





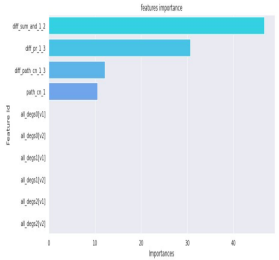
Models

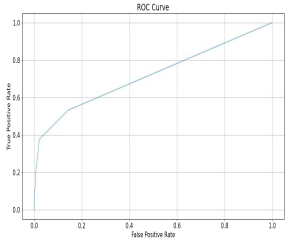
Consists of small tress. The most important features is the page rank

Performance

The model has a score of 0.89 with 100 estimators

Gradient Boosting





Models

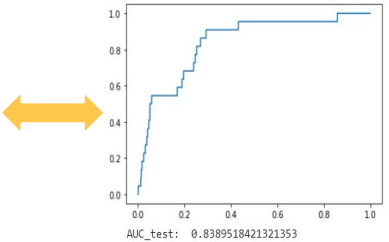
LightGBM and Catboost are the gradient boosting methods.

Performance

The models have ROC scores below 0.8 but the models converge faster

Model Evaluations

Fully connected neural networ (FNN)



Model

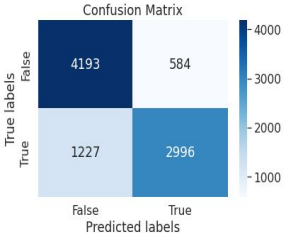
4 fully connected layers are constructed. Each euron is linked to the previous layer

Performance

The FNN has a score of 0.839 with 100 batch size and small learning rate

Graph Neural Network (GNN)

4 stacked GCN layers;
1 channel GCN layer;
2 1D Conv layers……



Model

The framework is based on the DGCNN network [5] of a research paper

Performance

The GNN has he accuracy of models at 79.87% and precision score of 87.78%

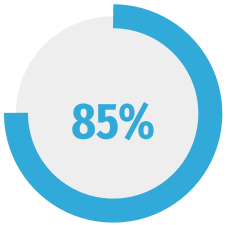
Model Comparisons

Best performance

Random forest has the highest AUC_ROC score of 0.89

Worst Performer

Catboost perform the worst among all the models



Neural Network vs Gradient Boosting

The average score of neural networks is 0.849, higher than 0.7415 of gradient boosting methods

Models	Accuracy	AUC_ROC
Random Forest	0.9931	0.890
Catboost	0.9942	0.703
LightGBM	0.9926	0.780
FNN	0.8244	0.839
GNN	0.7987	0.859

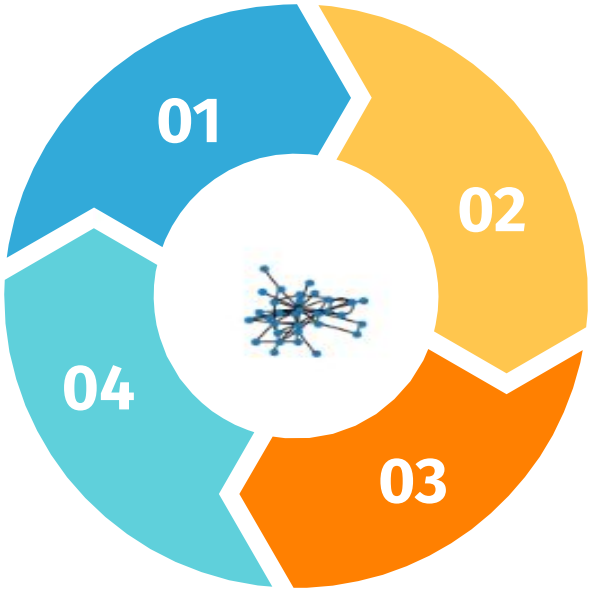
Conclusion

Data

The challenge is to make predictions of future research paper topics using link predictio.

Models

Total 5 models has been implemented and compared with the evaluation metrics of AUC_ROC



Features

A variety of features have been calculated including Jaccard and Page Rank

Improvements

Increase the training iterations, add more features and model layers

References

Science4Cast Challenge : https://www.iarai.ac.at/science4cast/

[1] Joao P. Moutinho, Bruno Coutinho, Lorenzo Buffoni 2022. Network-Based link prediction of scientific concepts. DOI: https://doi.org/10.48550/arXiv.2201.07978

[2] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6):1150–1170, 2011.

[3] Muhan Zhang, Yixin Chen, Link Prediction Based on Graph Neural Networks. DOI:https://arxiv.org/abs/1802.09691

[4] Gamal Crichton, Yufan Guo, Sampo Pyysalo, Anna Korhonen, Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches, BMC Bioinformatics, 2018. DOI:https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2163-9

[5] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, “An end-to-end deep learning architecture for graph classification,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11782