Consumer Data Research Centre
An ESRC Data Investment
THE DATA CITY

A machine learning framework to estimate the innovation capability of business entities.

Michael Andreas Kämpf[1], Jack Lewis[2]
[1]University College London, [2]The Data City

Background and Motivation

Innovation is regarded as the main factor stimulating economic growth and is subsequently given a great emphasis from policymakers and business managers alike. However, decision making regarding innovativeness requires timely, firm-level granular and reliable data which is frequently not available. Conventional measures based on patenting data or surveys are not suitable to measure innovativeness in such a dynamic economic environment as of today.

Therefore, this paper addresses this area of research by proposing a machine learning framework for ranking different business entities according to their level of innovativeness. The framework uses entirely open data and is, therefore, establishing a fully reproducible procedure with multiple areas of application. Moreover, the proposed framework offers an alternative to the cost and time intensive innovation surveys by offering a methodology which is scalable, time efficient and does not require significant investments in neither data access nor hardware. As a mean to rank the innovativeness of a company, the framework proposes a methodology of generating a labelled dataset based on objective criteria. By proposing this methodology, the framework becomes relevant to numerous fields of application and research areas. The framework is hen used on UK companies to demonstrate its advantages by using various machine learning methods built on the framework.

The paper concludes with a review of the findings and further research directions to further develop the proposed framework.

Data and Methods

The data used in this research comes from two sources. First, the sponsor of the research, The Data City, has made website data from UK companies available for the purpose of this research. In this data included are also all inbound, outbound and domain internal links from a domain which will be used to generate the first part of the labelled training dataset. Second, a dataset of companies has been generated by publicly available data from InnovateUK.

To apply a supervised machine learning method to estimate the innovativeness of a company, a comprehensive labelled data set must be available. While there are top-level country innovativeness studies across the world both from academic institutions, governments and private research organizations, no comprehensive enough dataset on a company-granular level is published for the UK.

The main challenge of generating the labelled dataset is that it must cover all industry sectors according to their relative importance within the UK economy. The first approach was to solely focus on companies being part of a start-up accelerator program. While companies found through this approach are likely to be deemed innovative in one or another way, they will hardly cover every possible industry. Precedence from Germany proposes that Bosch, a well-established company, is the most innovative across the German economy (Egeln et al., 2018). As Bosch and similarly established companies will most likely not enter an incubator program, I propose to use a novel approach of mapping the innovativeness of companies based on their website linking to generate a more industry-balanced labelled datasets while ensuring that newly founded ventures without many website links are not negated either. Eventually, the labelled dataset generation is based on two approaches. First, URL link network analysis using the PageRank centrality measure to rank the importance of a node in the network. Second, InnovateUK investment data published by the UK government.

Key Findings

This paper introduced a new framework on measuring innovativeness on a company-level. The main advantage of the proposed method is the scalability and cost-efficiency as it does not require a prolonged timeline like traditional survey-based approaches.

Further, no specific hardware is required, and openly available Python libraries suffice for the generation of the labelled dataset and the training of the machine learning algorithm. The findings will be split into the two main objectives of this paper.

Generating a labelled dataset

Supervised machine learning tasks require labelled data to start with. However, often no such data is available for a narrow research objective. This paper introduced a framework to generate a labelled dataset based on a rigorous approach which is reproducible and scalable for any project and purpose. Moreover, it demonstrated a mean to evaluate different training sets and chose the most suitable one. Last but not least, additional guidance on how the labelled data can be validated was proposed.

One key advantage of the proposed framework is the cost-efficient generation of a labelled data set. While survey-based research requires several months to collect, clean, and analyse the data, the proposed approach provides the means to collect, clean and analyse similar data within less than a week. I do not indicate that survey-based methods are no longer valuable but rather that they may be combined with the comparably cheap and fast machine learning framework to enhance the overall accuracy. I propose that survey-based innovation research is done regularly such as every 5 years and the proposed machine learning framework on a more frequent, bi-annual basis. This would further establish a benchmark with which the machine learning framework as well as the survey can be compared to.

Scoring innovativeness

The machine learning algorithm achieved a high score and is promising. While it is undoubtable that the LinearSVC model outperforms the logistic regression, the overall performance, as usual, will depend on the training set. As pointed out earlier, Kinne and Lenz (2021) were able to use the extensive CIS survey data including firm-level granularity and, therefore, had a large sample of ranked companies to validate their models on. This paper, however, is limited by any extensive mean for validation. To circumvent this challenge partially, a manually generated validation set was created and with binary variables classified whether the expected prediction should be “innovative” or “not innovative”.

While the results may not be taken as certain due to the small size of the validation set, it nevertheless indicates that the model would achieve a comparably solid classification of about 70%. Whether the final scores are appropriable or not remains a question to be further elaborated. While 70% may sound low, the question on how accurate traditional approaches are, remains. Eventually, the proposed machine learning algorithm must not be compared to the potential 100% accuracy but rather to the existing approaches.