DMS Accelerator

D4.6 Whitepaper v2

# Selection practices of European data incubators and accelerators: what are the predictors of applicants' success?

This project has received funding from the European Union's Horizon 2020 Programme under Grant Agreement N°. 825014.

| | |
|---|---|
| **Workpackage** | WP4 - Monitoring & Analysis |
| **Editor(s)** | Maria Priestley, Gefion Thuermer, Elena Simperl |
| **Responsible partner** | KCL |
| **Contributors** | All |
| **Status-Version** | v1 |
| **Due to** | 31/12/2021 |
| **Submission date** | 30/12/2021 |
| **EC Distribution** | PU |
| **Abstract** | One of the core functions of data-centric innovation programmes is the ability to attract, select and effectively allocate resources towards the most promising companies. Our study explores the selection practices of three publicly-funded data innovation programmes in Europe: DMS Accelerator, DataPitch and ODINE. Using administrative data from 725 open call applications and outcomes, we present a methodological approach for quantifying applicants' characteristics and identifying qualities that contributed to greater chances of acceptance. After evaluating attributes related to company maturity, team composition and application content, we found that the length of submissions and disciplinary diversity within teams contributed to higher chances of acceptance. However, several other indicators which were predicted to be important (e.g. specific linguistic cues and demographic diversity within teams) did not appear to influence selection outcomes. We discuss the implications of these findings and present recommendations for investigating selection practices through quantitative data. |

## Document History

| Version | Date | Comment |
|---|---|---|
| v0.1 | 10/12/21 | First draft prepared by KCL |
| v0.2 | 15/12/2021 | Draft revised by Zabala |
| v0.3 | 22/12/2021 | Draft containing inputs from other partners |
| v1 | 30/12/2021 | Final version |
| | | |

# Contents

## Executive Summary

Quantitative research into the selection practices of publicly-funded business support programmes is meaningful for two reasons. One aspect relates to the growing interest in the use of past applications data to develop automated screening systems that can predict which applicants are more likely to succeed, thus reducing the resources required for manual review. The second purpose relates to the detection of trends in past selection decisions for auditing purposes. EU-funded programmes are designed to benefit companies with specific profiles in terms of maturity, specialisation and demographic diversity. Quantitative methods for evaluating past decisions can therefore help programme managers to identify which characteristics are favoured in reality, enabling recommendations for a better allocation of public funds.

This whitepaper presents a quantitative investigation of open call applications and selection decisions from three EU-funded data innovation programmes that operated between the years of 2016 to 2021. Using data from 725 applications received by DMS Accelerator, DataPitch and ODINE, we present a methodological approach for quantifying unstructured aspects of application texts and team characteristics, and for measuring their explanatory power in predicting the decision outcome. We also discuss tools for obtaining demographic metrics from applicants' names where explicit data on equality, diversity and inclusion (EDI) are unavailable.

Based on our analysis, we find that the use of past data for predictive modelling is not yet a viable option. We attribute this to the difficulty of quantifying subjective selection criteria and the shortage of data that capture successful applicants. Despite this, we were able to detect statistically significant trends in the existing data which can be used to audit selection practices. We found that longer application answers and disciplinary diversity within teams consistently contributed to acceptance decisions. This suggests a strong preference for diverse expertise and companies that make the effort to submit informative responses at application stage.

Other metrics such as the linguistic content of applications, team size and demographic diversity were not consistently associated with selection outcomes. Of these, demographic diversity is worthy of further attention. While we were not able to detect any overt biases against teams containing under-represented groups such as women and non-European ethnicities, we were also unable to prove a selective preference for teams that were inclusive of this kind of diversity. We estimate that only 19% of individuals named inside applications were women, and that 19% came from ethnic minorities. Although these figures are comparable to previous estimates of diversity in data and AI entrepreneurs, our findings underscore the European Commission's ongoing concern with pursuing better representation of women and different ethnicities in industries affiliated with data and AI.

We conclude the paper with a summary of methodological implications and recommendations for other innovation programmes that may be considering the use of machine learning approaches for predictive or auditing purposes in the selection process.

# 1. Introduction

Business incubation and acceleration programmes are one of the instruments used by the European Commission to promote innovation and entrepreneurship among European startups and small- and medium-sized enterprises (SMEs). Compared to larger stakeholders, these nascent companies have been recognised as an important source of innovations that drive economic growth (De Marco et al., 2020), but they are also most vulnerable to financial, operational and resourcing challenges. One of the core functions of data incubation and acceleration programmes lies in their ability to select companies that are most likely to survive, yield financial returns and meet objectives that are meaningful in the data market.

Our study explores selection practices through the lens of publicly-funded data innovation programmes. Recent studies have detected divergences between the scope of EU-funded policy designs and the SMEs that are chosen to benefit from these initiatives (De Marco et al., 2020), leading to potential inefficiencies in budget allocation. It is therefore important to audit selection decisions to ensure that they align with the intended objectives of the funder.

EU-funded programmes that specialise in preparing data-centric startups and SMEs for the data economy have a unique set of aspirations regarding the characteristics of companies and their teams. Although these criteria are routinely monitored and reported through qualitative accounts and descriptive statistics in project documentation, the administrative data accumulating in innovation initiatives create the scope to use more rigorous quantitative approaches. This includes the potential to detect unconscious biases against under-represented gender and ethnic groups, with previous examples observed in selection decisions by social impact accelerators (Yang et al., 2020) and investors (Kleinert & Mochkabadi, 2021). Given that social and cultural heterogeneity is an important stimulant of corporate innovation (Brixy et al., 2020), it is advantageous for programme managers and funders to be able to detect and revert selection biases that could undermine diversity in the fields of data and AI.

While our interest in selection practices is motivated largely by auditing reasons, previous quantitative studies on applications data have pursued a different purpose. A substantial body of prior work relates to the development of decision-support systems that use a variety of attributes to generate predictions about which applicants are more likely to succeed. This avenue of research has been particularly explored in the areas of crowdsourced innovation (Hoornaert et al., 2017; Nagar et al., 2016) and entrepreneurial finance (Blohm et al., 2020). Similarly to publicly-funded incubation and acceleration programmes, these opportunities attract large numbers of submissions that need to undergo manual evaluation. Predictive systems are becoming an active area of research, as they can assist with filtering applications and reducing the costs of manual review under resource constraints. One of the outcomes of our study will be a discussion of the viability of data-driven tools based on past applications data for increasing the efficiency of selection practices.

In view of the potential benefits of modelling selection practices for auditing and predictive purposes, we propose and implement a methodology based on the applications data received by three EU-funded initiatives that worked with data-centric

startups and SMEs between the years of 2016 to 2021. These initiatives were Data Market Services Accelerator (DMS)[1], DataPitch[2] and Open Data Incubator Europe (ODINE)[3]. Using these examples, we provide a critical appraisal of open call applications as a data source, considering the kinds of information that are contained inside them and the pre-processing steps that can be used to convert such data into generalisable metrics. The goals of our paper are to:

- Derive operational definitions of the selection criteria that are advocated by publicly-funded data innovation programmes and suggest how these can be quantified from applications data.
- Verify whether the intended selection criteria are detectable empirically using data from selected and rejected applications.
- Measure the relative importance of different characteristics in predicting startups' success in being selected.

Based on our findings and experience of implementing the above steps, we will critically evaluate the utility of our methodological approach both in terms of predictive and auditing purposes.

The remainder of this paper is structured as follows. In Section 2, we discuss previous research into the selection criteria of business incubators in general, as well as the criteria that are specific to EU-funded programmes. We then discuss the motivations for modelling selection practices in a quantitative manner. In Section 3, we introduce the methodological challenges associated with open call data, including approaches for extracting meaningful attributes and measuring their predictive power. We demonstrate these tools in practice using a dataset of applications collated from three data innovation programmes. The results of our analysis are presented in Section 4, followed by a discussion in Section 5 of their practical implications for programme managers and funders.

# 2. Theoretical background

We begin this section by describing our study in the context of wider academic literature on the typical screening practices of European incubation and acceleration programmes, and those of data-centric programmes in particular. We then discuss why and how selection criteria have been modelled empirically in previous studies, followed by a summary of our hypotheses.

## Business incubators and accelerators

Business acceleration and incubation programmes have grown in popularity in the recent decades as instruments to support startups and SMEs. Although the field continues to change rapidly, with new types of programmes launching almost weekly (Bone et al., 2017), a number of general trends have been identified in the format and selection practices of these programmes. Incubators have established a reputation among scholars and policy-makers as a local economic development tool, while

[1] https://www.datamarketservices.eu/ [accessed 27/12/21]
[2] https://datapitch.eu/ [accessed 27/12/21]
[3] https://opendataincubator.eu/ [accessed 27/12/21]

accelerators have been more recognised in the private sector (Dempwolf et al., 2014; Cohen & Hochberg, 2014). Both types of programmes are alike in providing nascent firms with advice, services, financing (sometimes), and facilities to help them develop and launch their business. Below we discuss how acceleration and incubation programmes have been delineated by their specific objectives and service offerings, followed by a summary of the ways in which these aspects influence their selection practices.

In their taxonomy of business support programmes, Dempwolf et al. (2014) identify a number of differences between accelerators and incubators. Accelerators have traditionally been characterised by the intention to quickly move startups from one stage to the next, while incubators have sought to develop longer-term capabilities that constitute self-sustaining, mature businesses. In line with this, accelerators have traditionally followed a fixed-term cohort-based structure lasting several months (typically 3), while incubators offered longer-term support on a case by case basis. According to previous observations (Dempwolf et al., 2014; Bone et al., 2017), it was rare for incubators to invest directly in their startups, while accelerators were more likely to provide funding in exchange for an equity stake in the business. When it came to the selection process, incubators were found to target applicants from the local community, while accelerators additionally considered firms at national and international levels. Although these distinctions between accelerators and incubators have been identified, few academic studies have examined them in field-specific and funder-specific contexts, such as the recent EU-funded data initiatives in our paper. Based on our experience of data-centric innovation initiatives funded by the European Commission, we intuit that the previously established definitions of acceleration and incubation are less clear-cut in this context.

Data-centric acceleration and incubation programmes share many similarities with each other. They typically offer a comprehensive training package and mentorship in traditional business functions such as entrepreneurship, fundraising and marketing, alongside more specialised topics such as data skills, data protection (GDPR), standardisation and Intellectual Property Rights (IPR). In both types of programmes, the services address short-term business acceleration as well as longer-term sustainable innovation goals, the latter of which would typically be characteristic of incubators in traditional settings. Both types of programmes also tend to adopt a cohort structure with open calls being used to recruit applicants, a practice that is traditionally more common in accelerators rather than incubators. Moreover, both types of data-centric programmes tend to adopt an online delivery format, where the applicants and service providers are decentralised across different regions and countries, which contrasts with traditional incubators that typically target local companies on-site.

We therefore suggest that some of the characteristics that were traditionally used to distinguish between acceleration and incubation programmes in the past (e.g. cohort structure, objectives, mode of service delivery) have become merged in the context of data initiatives. The main feature that delineates data-centric incubation programmes from accelerators is their provision of technical infrastructure and practical collaboration structures with other stakeholders. The cost of these collaborative innovation endeavours is typically accompanied by an equity-free grant from the funding body (De Marco et al., 2020), meaning that data-centric incubators (rather than accelerators) are the ones that are more likely to provide funding. Collectively, these aspects of

EU-funded data innovation programmes create multiple areas of transference between traditional understandings of business acceleration and incubation, making it difficult to map specific data initiatives to either one or the other archetype in traditional business literature. We will therefore draw interchangeably from relevant parts of the earlier literature relating both to incubation as well as acceleration programmes.

Previous empirical research into the screening practices of European (Aerts et al., 2007) and American (Lumpkin & Ireland, 1988) incubators has observed that screening criteria can be roughly divided into three groups: those based on financial strength, personal characteristics of the management team, and market factors. Each of these criteria are operationalised using specific metrics. As shown in Table 1, financial ratios may be assessed using liquidity, profitability and asset utilisation, while the management team may be evaluated based on their technical, managerial and financial skills, and market fit assessed using the company's stage of development, uniqueness of product and business plan. While American incubation policy has tended to favour financial criteria that are likely to lead to rapid success (e.g. companies entering the stock market), European incubators orient their assessment towards longer-term development by prioritising aspects related to the management team and market fit. In their study of 107 European incubators, Aerts et al. (2007) found that market factors were considered most important by 61% of incubators, while the management team was prioritised by 27%, and only 6% prioritised financial factors during selection. The remaining portion of incubators (only 6%) demonstrated balanced preferences that were split more equitably between the three types of criteria. This latter group of incubators also demonstrated the healthiest outcomes in terms of lower tenant failure rates.

## The innovation programmes in our study

To our knowledge, there have been no prior studies that systematically mapped the screening criteria of data acceleration and incubation programmes. As stated in the Introduction, our study focuses on three EU-funded innovation programmes: DMS, DataPitch and ODINE. According to their documentation, each of these programmes aspired to a balanced screening approach. We describe each programme in more detail below, followed by a summary of their selection criteria.

**DMS Accelerator** operated between 2018 - 2021. Its objective was to help data-centric SMEs and startups overcome the challenges of entering and operating in the European data market. The programme consisted of free training and personalised support services in relation to fundraising, acceleration, promotion, data skills, standards and legal issues. The services were delivered to a total of 150 startups and SMEs split across three cohorts, working with 50 companies over a duration of 6 months each year. Unlike other initiatives, DMS was unique in that it did not offer seed funding to the participating companies.

**DataPitch** operated between 2017 - 2020. The programme brought together corporate and public-sector organisations that had data with startups and SMEs that worked with data. Within this, the larger organisations proposed challenges to which the startups could apply. Alternatively, startups could apply on a range of topical challenges, provided they had an organisation that would share data with them to address it. In addition to the central focus on data sharing, the programme included other support

services similar to DMS, such as data skills, entrepreneurship, standards and legal issues. The programme supported a total of 47 companies during two cohorts, with each company being given an equity-free grant of up to €100K.

**ODINE** operated between 2016 - 2019. The programme supported startups and SMEs in creating new business ventures based on open data. During a 6-month acceleration period, the participants received mentorship, business and data training, dedicated events, media promotion, introductions to investors and access to an international network. 57 companies were supported in total, with each company being given an equity-free grant of up to €100K.

Information about the evaluation criteria for each programme was available through their published documentation[4]. All three programmes reported screening their applicants on balanced criteria that included financial qualities, team characteristics and market factors. For example, in the dimension of finances, ODINE evaluated companies' financial plans, DataPitch considered their growth and budgets, and DMS screened them based on stages of development and investment plans. In regards to team characteristics, all three programmes assessed the skills and experience of the core team. For market factors, the programmes were alike in evaluating the companies' value proposition and market opportunity. Additionally, ODINE and DataPitch enquired about their specific uses of data, while DMS asked companies to explain how they would benefit from this particular programme in order to understand their needs and assess their fit. Table 1 summarises the selection criteria of each of our three programmes alongside the traditional screening factors identified by Aerts et al. (2007).

Table 1. Evaluation criteria of traditional business incubators and the three data-centric incubators in our study.

| | Traditional screening factors (Aerts et al., 2007) | DMS | DataPitch | ODINE |
|---|---|---|---|---|
| **Financial ratios** | Liquidity, profitability, asset utilisation, price earnings, debt utilisation. | How they would use the investment round. | Realistic forecast, convincing growth, clear & appropriate budget. | Financial plan and appropriateness of budget. |
| **Team characteristics** | Age, Sex, Technical Skills, Management Skills, Financial Skills, Marketing Skills, Aggressiveness/ Persistence, | Background of core team. | Skills, capacity, commitment & understanding of finances. | Team skills - interdisciplinarity & past experience. |

[4] Copies of the documentation can be found in the data repository that accompanies this study.

| | | | | |
|---|---|---|---|---|
| | References from Others. | | | |
| **Market factors** | Current Size, Growth Rate, Uniqueness of Product/Service, Marketability of Product/Service, Written Business Plan. | Stage of development (scaling, validating, establishing).<br><br>Company vision, USP, how they will benefit from the programme. | Strength & novelty of idea, quality of data value chain, outputs.<br><br>Value proposition, market opportunity, expected impacts. | Strength & novelty of idea, Use of open data.<br><br>Value proposition, market opportunity, expected impact. |

## The distinguishing features of European data incubators

While many of the above criteria are held in common with traditional incubators, there are a number of distinguishing features specific to data incubators that are worth discussing. Compared to commercially oriented incubators that rely on companies' financial success in order to sustain themselves, those sponsored by the European Commission are motivated by broader-socio-economic goals. These encompass several aspects, namely: skills, entrepreneurial capacity and social diversity.

With regard to upskilling, publicly funded initiatives tend to offer more comprehensive training curricula compared to privately funded accelerators (Clarysse et al., 2015). The uptake of EU-funded programmes therefore relies on companies that are willing to commit substantial time and resources to training their staff. In the present case, these curricula address traditional business functions such as entrepreneurship and fundraising in addition to data-centric skills such as data science, standardisation, GDPR and legal strategy. Rather than focusing purely on short-term profitability, the latter trainings contribute towards long-term, sustainable data innovation. To ensure that selected candidates have the capacity to draw value from these resources, the programmes in our analysis screened applicants using questions about the data value chain and the benefits which the company wished to receive from the programme's specific service offering.

Besides selecting startups who are committed to and interested in the data economy, these programmes also targeted startups and SMEs that were in their earlier stages of development. The reason for this is that despite being at the forefront of science and research in terms of publications, Europe still lacks the entrepreneurial capacity to translate this into corporate innovation, growth, and jobs[5]. To address these challenges, EU-funded initiatives target startups and SMEs that are struggling to transition their product or service from the laboratory into the marketplace. This means selecting companies that are in their earlier or pre-revenue stages. In the new Horizon Europe

[5] https://www.eca.europa.eu/lists/ecadocuments/ap19_06/ap_sme_en.pdf [accessed 09/11/21]

work programme, this has been implemented through consistent indication of expected incoming and achieved technology readiness levels for programmes like these.[6]

In addition to exerting a sustainable impact on the innovation capacity and commercialisation of early-stage ventures, the European Commission is keen to support social diversity within the data economy. A critical area of focus has been the mitigation of inequalities in labour markets, where women are currently under-represented in leadership roles and STEM careers (science, technology, engineering, mathematics) (Ortiz et al., 2020). These concerns were part of the Horizon 2020 programme under which all three of our incubators were sponsored[7]. However, it was difficult to measure progress towards these goals because none of our programme had a routine for collecting equality, diversity and inclusion (EDI) data from applicants. An attempt was made during the first open call of DMS to consider the gender diversity of teams at application stage. However, this question was subsequently rephrased to address social responsibility in a more general sense, so as not to discourage applications from predominantly male teams (D2.3).

EDI data capture and disclosure is a challenge that is not unique to our context, continuing to pervade a range of other public and private initiatives[8]. Gender diversity has been proven to lead to beneficial outcomes in Responsible Research and Innovation (RRI) projects through a large variety of studies (Nielsen et al., 2018). However, AI teams remain predominantly male and white (Hagendorff, 2020), leading to a variety of issues when products developed by such teams remain unchecked, and are taken to markets that are more diverse than their developers (D'Ignazio & Klein, 2020).

It is reasonable to expect that the objectives and pressures discussed above may reasonably contribute to the selection process of publicly-funded data incubators in Europe. In addition to traditional evaluation criteria related to company finances, team characteristics and market factors, these programmes are implicitly expected to align with the goals of the European Commission. This includes selecting companies that demonstrate sufficient commitment to upskilling their workforce and generating economic value from the specialised support services offered by the programme, as well as those that have the capacity to produce positive social impact by striving for gender equality and ethnic representation within their teams.

## Modelling the selection process

The multitude of considerations discussed above present a range of signals that evaluators need to assess when selecting applicants. Prior research from entrepreneurial finance has found that being faced with such complex tasks can cause decision makers to suffer from “analysis paralysis” that influences their ability to conduct thorough assessments and select the best possible candidates (Huang, 2018). Suboptimal decisions of this kind have previously been attributed to evolved heuristics

[6] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2021-2022/wp-13-general-annexes_horizon-2021-2022_en.pdf [accessed 06/12/21]

[7] https://ec.europa.eu/info/sites/default/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-gender-equality_2019.pdf [accessed 21/11/21]

[8] https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-EDI-EvidenceReviewInternational.pdf [accessed 21/11/21]

that help people to process large volumes of information at speed, including local bias, overconfidence, and loss aversion (Maxwell et al., 2011). The advent of big data and data-driven approaches is making it easier to detect these human biases in entrepreneurial contexts and to override them using machine intelligence tools (Blohm et al., 2020).

One of the prerequisites of data-driven approaches is the extraction of quantitative metrics from corporate information. While some informational cues are available in the form of structured data (e.g. size of company, prior funding, revenue), other cues of a more subjective nature are embedded in unstructured text responses. Previous research from the fields of management and organisational science has sought to extract the behavioural and cognitive characteristics of entrepreneurs using various natural language processing (NLP) techniques. For example, Short et al.(2010) developed word dictionaries for capturing companies' innovativeness, proactiveness and competitiveness based on the language used in CEO's letters to shareholders. Others have used more generalised tools such as the Linguistic Inquiry WordCount (LIWC) (Pennebaker et al., 2015) to measure applicants' cognitive processes and personal concerns (Nagar et al., 2016). Chae & Goh (2021) used IBM Personality Insights (IBM PI) to map entrepreneurs' tendencies towards introversion and extraversion based on the data contained in their tweets, while Blohm et al.(2020) used the Twitter API to analyse the emotional valence of posts shared by entrepreneurial ventures. Although a variety of text-based metrics have been used in previous studies, we test only a small subset of traits, namely innovativeness and competitiveness, as these can be mapped onto the expected selection criteria of our programmes.

The specific context of our research also requires us to consider metrics that relate to social diversity. While it is still uncommon to find equality, diversity and inclusion (EDI) data to be openly reported by entrepreneurial ventures, numerous studies have been able to rely on the names of individuals as a proxy for extracting demographic metrics. For example, in their descriptive study of data and AI entrepreneurs, Chae & Goh (2021) used machine learning tools to predict the likely genders and ethnicities of entrepreneurs based on their first names and surnames, while Chang & Fu (2021) used similar techniques to examine ethnic inequalities in the field of mathematics. Our methodological approach will combine these metrics with other explanatory variables that have been more commonly used in prior models of selection in entrepreneurial settings.

## Research questions

Our primary research question is:

- Is it possible to automatically predict the success of applications from the data contained in them?

Within the scope of European data innovation programmes in particular, we will use the model to test the following hypotheses:

- H1: Companies in their earlier stages (based on the number of employees) will have a higher chance of being selected than larger more mature companies.

- H2: Diverse teams (in terms of disciplines, gender and ethnicity) will have a higher chance of being selected than non-diverse teams.
- H3: Companies with greater innovativeness, competitiveness and programme fit will have a higher chance of being selected.

In the next section, we describe our methodological approach for obtaining quantitative features to test these hypotheses. We will also discuss some of the challenging characteristics of open call data that needed to be resolved before using them in a predictive model.

# 3. Methodology

The data for our study come from three innovation programmes: ODINE, DataPitch and DMS. The applications collected by them varied in their data format, information contained and acceptance rates. Below we discuss the steps that were taken to collate the information into a single dataframe, to extract generalisable metrics that were meaningful to our hypotheses, and to build the model.

The anonymised data and reproducible code underlying this study are openly accessible via an online repository (Priestley et al., 2021).

## Ethics statement

Our study involves secondary use of personal data to make inferences about the demographic composition of company teams. Ethical clearance to process the data in this way was granted by the BDM Research Ethics Panel at King's College London.

## The dataset

Our study relies on heterogeneous sources of data that required a number of preparatory steps before relevant attributes could be extracted. This included ensuring that the applications were in machine-readable format and that they contained all the information necessary for analysis. Table 2 summarises the numbers of cases that were subsequently used in our study.

The applications data at DMS and DataPitch were available in machine-readable csv format. However, applications from ODINE were received in the form of PDF documents. We converted these files to csv format using the "pdfplumber" package in Python.

ODINE originally had 57 accepted and 1116 rejected applications. Some information was lost due to unavailable documents and text encodings that could not be rendered. We randomly sampled the rejected applicants down to four times the number of acceptances, in order to create a class distribution that was similar to the other two programmes and to ensure a manageable sample size for manual pre-processing. We also removed borderline cases that were marked as "probably accept" but were subsequently rejected by the programme. After removing further cases with missing data, we were left with 44 accepted and 155 rejected applications from ODINE.

DataPitch originally contained 47 successful and 192 unsuccessful applicants. As with ODINE, we removed cases with missing data and those that did not consent to secondary use of their data, leaving us with 45 accepted and 149 rejected applicants.

DMS data originally consisted of 3 cohorts. However we chose to include only cohorts 2 and 3 as the application format during these cohorts was most informative and amenable for analysis. These data originally included 100 successful and 332 unsuccessful cases, which were reduced down to 96 successful and 236 unsuccessful after removing cases with missing data.

It is worth noting that earlier studies of a similar kind retained cases with missing data and incorporated metrics about the completeness of applications into their model (Nagar, 2016). We chose not to do this in our study because several of our predictor variables rely on values that could only be obtained from completed application answers. Moreover, our automated approach to extracting structured data from PDF documents at ODINE introduced the risk of losing some attributes as a result of the data extraction process rather than the fault of the applicants themselves, meaning that completeness would not be a reliable signal in our study.

Table 2. Data samples that were used from each programme.

| | ODINE | DataPitch | DMS | Study total |
|---|---|---|---|---|
| **Accepted** | 44 | 45 | 96 | 185 |
| **Rejected** | 155 | 149 | 236 | 540 |
| | | | | Total = 725 |

## The metrics

Our innovation programmes share several similarities in their intended selection criteria (e.g. prioritising novelty, competitiveness, market fit and team composition). While some questions were common across all programmes (e.g. the number of employees), other characteristics were recorded using different types of questions, and often in free text format. This presented us with the challenge of operationalising clearly-defined signals that could be collated and compared between programmes. Table 3 summarises the attributes which were extracted. Our methodological approach to obtaining each attribute is discussed below in greater detail.

Table 3. Variables included in the model.

| Metric | Method | Mean or % | SD | Min | Max |
|---|---|---|---|---|---|
| **Field diversity** | Manual assessment of team description. | 1 - 71%<br>0 - 29% | | 0<br>(no) | 1<br>(yes) |
| **Gender diversity** | Automatically deduced from team names. | 1 - 37%<br>0 - 63% | | 0<br>(no) | 1<br>(yes) |
| **Ethnic diversity** | Automatically deduced from team names. | 1 - 37%<br>0 - 63% | | 0<br>(no) | 1<br>(yes) |

| | | | | | |
|---|---|---|---|---|---|
| **Programme fit** | Computed text similarity to programme description. Subsequently converted to Z-score. | 0.31 | 0.09 | 0.06 | 0.55 |
| **Innovativeness** | Proportion of words linked to innovativeness. Subsequently converted to Z-score. | 0.02 | 0.01 | 0.00 | 0.06 |
| **Competitiveness** | Proportion of words linked to competitiveness. Subsequently converted to Z-score. | 0.01 | 0.01 | 0.00 | 0.04 |
| **Number of employees** | Application answer. | 6.75 | 7.50 | 0 | 100 |
| **Wordcount** | Automatically calculated, excluding stopwords. Subsequently converted to Z-score. | 313.51 | 175.75 | 27 | 951 |

## Team composition metrics

Our assessment of team composition consisted of three metrics: field diversity, gender diversity and ethnic diversity.

*Field diversity* was assessed by manually reading the description provided by each applicant in response to questions about team experience (e.g. “List the core members of your team and their skills and experience”). We created a new column to indicate if at least two different disciplines or areas of expertise were represented by the team members. The response was coded using binary values, where companies that met the condition were labelled with “1” and those that did not with “0”. For example, a company where all team members were developers would be labelled with “0”, while a team containing a computer scientist and a manager would be labelled with “1”.

*Gender diversity* was assessed by parsing the team’s first names through the “gender_guesser”[9] package in Python, which classifies names into the categories of male, female or androgynous. We created a new column where teams with at least one name parsed as female would be marked as gender-diverse (labelled “1”) and teams that did not have names that were parsed as female marked with “0”. Teams that were predicted to consist entirely of women also met our condition for gender diversity, since this is an under-represented group in the fields of data and AI (Chae & Goh, 2020). We recognise that our approach to capturing gender diversity is somewhat crude, and that continuous measures based on self-reported percentages of female and non-binary employees may be more appropriate. However, it was not possible to obtain such metrics from the present data, where team names were typically provided for only a

[9] https://pypi.org/project/gender-guesser/ [accessed 12/11/21]

few key team members rather than the entire company, and very few respondents voluntarily specified their company's gender composition without being prompted.

*Ethnic diversity* was assessed by parsing the team's names through the "ethnicolr"[10] package in Python. Middle names were ignored, and cases where the first name was absent were based only on the surname. Recent uses of ethnicolr have included evaluations of ethnic representation in the fields of data and AI entrepreneurship (Chae & Goh, 2020) and mathematics (Chang & Fu, 2021). Several versions of the ethnicolr model are available, trained on US Census data, Florida Voter registration data and Wikipedia data (Sood & Laohaprapanon, 2018). We used the Wikipedia version of the model, which offers a granular separation between various European, Asian and African regions, with precision and recall scores averaging 73% (Sood & Laohaprapanon, 2018). Although this is not ideal, we do not expect the misclassified ethnicities to have a significant impact on the results, as the error rate is likely to be similar across accepted and rejected applicants. In order for a team to be classified as ethnically diverse, we imposed the condition that they must have at least one team member whose name is classified as Asian or African. Teams where all named staff were classified in only one of these groups also met our condition for ethnic diversity, as they contributed to pluralism in the predominantly European context of our study. Our resultant metric of ethnic diversity was captured in a binary variable ("1" for yes, "0" for no). As with the case of gender, there is scope for developing more nuanced signals of ethnic diversity in future given better availability of EDI data.

## Market factors

Our evaluation of market factors encompassed the application's fit to the programme, innovativeness and competitiveness. Each of these metrics was based on the language contained in open-ended application answers[11]. Before extracting the metrics as described below, each company's written answers to various questions were concatenated into a single text string. We tokenised, stemmed and removed stopwords from each entry using the "nltk" package in Python.

*Programme fit* was computed using the cosine similarity[12] between the application text and a supporting document that described the service offering of the programme to which they applied. In the case of DataPitch, which had multiple challenges, the service description was joined with a description of the specific challenge to which the company applied.

*Innovativeness* and *competitiveness* metrics were calculated to reflect the proportion of words inside the application that were related to each of these two entrepreneurial constructs. The calculation was based on word dictionaries developed by Short et al.(2010), where innovativeness was captured using 77 single-word items (e.g. "discover", "create", "new"), and competitiveness contained 53 single-word items (e.g. "achievement", "challenge", "exploit").

[10] https://ethnicolr.readthedocs.io/ [accessed 12/11/21]
[11] A list of the application fields is included in the data repository that accompanies this study.
[12] Cosine similarity is a metric that calculates the distance between two arrays of word counts from both documents (Thada & Jaglan, 2013).

Each of the language-based metrics above could be affected by differences in the structure of application forms and the types of information requested (e.g. some programmes ask more about innovation, while others assess fit to the programme). We dealt with this issue by converting the text metrics into standardised Z-scores[13] within the applicants' respective programmes.

### Control variables

The *number of employees* was included to reflect each company's level of maturity.

We also included the *wordcount* of each application (excluding stopwords) as a measure of effort made by the applicant. As with the other text metrics, the wordcount was standardised using Z-scores to accommodate differences in the length and structure of application forms between programmes.

We acknowledge that our analysis does not capture all relevant control variables. For example, at DMS, the evaluators additionally considered the year in which the company was founded, while DataPitch and ODINE also considered revenue. We excluded these metrics because such data were not routinely collected by all three programmes. Although these traits would be useful to consider in subsequent research, prior research has not found a statistically significant relationship between company age and financial attributes with their ability to perform open innovation activities that are meaningful to the European Commission's objectives (De Marco et al., 2020).

## The model

We used a series of logistic regression models to analyse the relationship between the eight predictor variables listed in Table 3 (diversity metrics, programme fit, innovativeness etc.) and the binary outcome of each application (accepted or rejected).

First, we trained the model on a combined dataset containing the applications from all three programmes in our study. We then investigated these initiatives separately. As ODINE and DataPitch datasets were quite small and followed a similar selection process, we combined their data together for one of the models. The other model was built using only the DMS data, which had a sufficient number of successful cases to be used on its own. In each of these cases, we built several versions of the model using different samples to accommodate class imbalances in the dataset (number of accepted vs. rejected cases). We also checked to ensure that there was no collinearity among the explanatory variables.

To evaluate the predictive power of the models, they were trained and evaluated using an 80/20 train/test split, which allowed us to see how each model performed on unseen data. In the next subsections we detail our approach to managing imbalanced training data and evaluating the model performance.

[13] Z-scores offer a generalisable scaling approach that captures an item's position in relation to the rest of its population, producing a number that reflects its distance from the average in terms of standard deviations (Wang & Chen, 2012).

## Dealing with imbalanced training data

The essential nature of the selection process creates an imbalance in the proportion of accepted and rejected applicants. In our programmes, the original rate of acceptance ranged from around 5% to 26%, meaning that the event of interest (success) was underrepresented in the datasets. This aspect of the open call data has important methodological implications, since there is less information for a machine learning model to learn about the rare successes, creating output that is biased towards the majority class (unsuccessful applicants). Such models are also hard to evaluate because they can achieve high overall accuracy simply by assuming that all cases belong to the majority class. Issues of this kind are common in other research areas that focus on rare events, such as fraud detection and medical diagnosis (Branco et al., 2016). A number of solutions have been proposed to construct and evaluate models from imbalanced data. The simplest and most effective strategies have included resampling the data prior to analysis, which is the approach we chose to adopt.

We trained each of our models on several partitions of the data. As a reference point, we began by using the original training dataset that had an imbalanced distribution of classes. Then, we created resampled versions of the training data, each of which contained all of the original successful cases and one third of the unsuccessful cases, producing distributions that were more balanced. Three random samples of unsuccessful applicants were sufficient to capture all of the available data. Figure 1 visualises our sampling approach. By investigating the model performance using different samples, we were able to evaluate its performance and stability under different training conditions and to identify the best sampling approach.

We note that the test data used to evaluate each model still retained the original imbalanced distribution of successful and rejected cases, in order to reflect how the model would perform on distributions observed in the real-world.

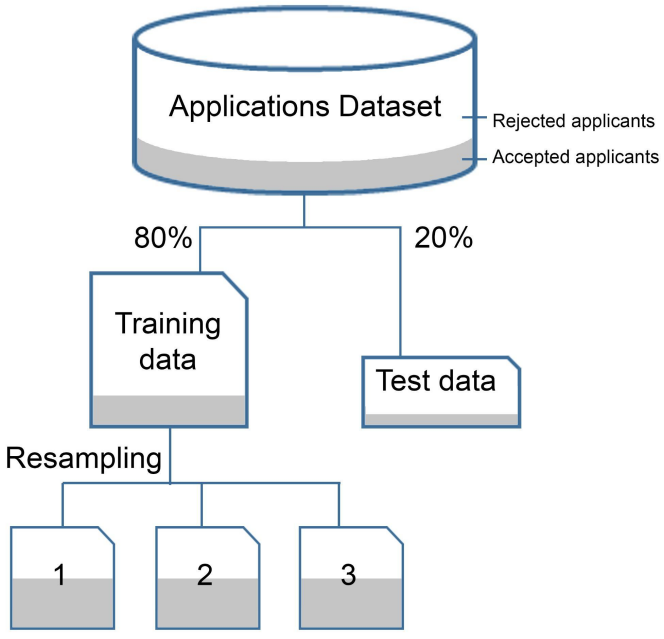


Figure 1. Division of the dataset.

## Model evaluation

We evaluated the performance of each of our models using metrics of accuracy, sensitivity (recall), specificity, precision, F-measure and area under the curve (AUC). As described in Itoo et al.(2020), these metrics can be calculated using the rates of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates as follows:

Accuracy = (TP +TN) / (TP + FP + TN +FN)
Accuracy is defined as the ratio of the total number of predicted decisions that are correct.

Sensitivity (recall) = TP / (TP + FN)
Sensitivity is the proportion of true acceptances that were correctly predicted as positive (i.e. True Positive rate).

Specificity = TN / (FP + TN)
Specificity is the proportion of true rejected applicants that are classified negative (i.e. True Negative rate).

Precision = TP / (TP + FP)
Precision is the proportion of predicted acceptances that are actually correct.

F-measure = 2 x (sensitivity x precision) / (sensitivity + precision)
F-measure captures a harmonic mean of sensitivity and precision scores to capture the trade-off between them. A score of 1 is considered optimal.

AUC = ½ x (sensitivity + specificity)
AUC is a measure that describes the degree to which the model is capable of differentiating between the classes (accepted vs. rejected).

Although each of the metrics above are relevant to evaluating and balancing the costs and benefits of a predictive model, we consider Sensitivity and F-scores to be the most useful in our use case. Sensitivity helps to counter the risk of rejecting strong applications, whereas the F-score helps to evaluate the trade-off between accurately detecting positive cases while considering the downside of having to review misclassified negative cases.

# Results

Our first regression was built using the combined dataset of startup applications from all three programmes. Four models were constructed in total, one based on the original imbalanced distribution of accepted and rejected applicants, and the other three models were built on smaller subsamples where the classes were balanced more equitably.

In every case, we found that the model was significantly better than chance at predicting selection decisions ($p < 0.001$). When trained on the original imbalanced distribution of accepted and rejected applicants, the model's overall accuracy was 0.69. However, closer inspection of the evaluation metrics showed that this model had a sensitivity of 0.05, meaning that it could detect only 1 out of 20 applicants that were selected in reality. The remaining three models that were trained on balanced samples

of data performed much better in terms of sensitivity. Of these, the model trained on sample #3 (shown in Table 4) achieved the highest sensitivity, estimated at 0.57. However, this was done at the expense of precision (0.37), meaning that although the model could accurately predict approximately 6 out of 10 successful applicants, it would also shortlist a large portion of those that would subsequently be rejected. Accordingly, the F1 score for sample #3 was also low (0.45), but it offered the best trade-off between sensitivity and precision when compared to the other three models.

This best performing model (#3 in Table 4) contained two variables that were flagged as important in predicting the decision outcome. Specifically, field diversity and wordcount were statistically significant at 0.05 and 0.01 levels respectively, meaning that there was a less than 5% chance that the difference in these variables could be due to chance. The odds ratio for field diversity was 1.93, with a 95% Confidence Interval of [1.05, 3.57]. This suggests that, when the other predictor variables were held constant, applicants with interdisciplinary teams were almost twice as likely to be selected compared to teams where everyone specialises in the same field. Similarly, the odds ratio for wordcount was 2.04 [1.42, 2.95]. This means that a one unit increase in the wordcount's Z-score (or the number of standard deviations above the mean) would translate to a 104% increase in the applicant's odds of being selected; in the current context, one standard deviation would be equivalent to being in the top 16% of their cohort in terms of wordcount.

Among the three models with poorer performance, field diversity remained significant in all three models to at least a $p<0.1$ level of statistical significance, while wordcount was significant in all three of them at the 0.01 level, suggesting that these two predictors were reasonably stable across different training samples. Among the models with poorer sensitivity, two additional predictors emerged as being statistically significant. Programme fit was significant in balanced model #1, with an odds ratio of 1.54 [1.06, 2.25] ($p <0.05$), suggesting that applicants whose submissions were textually similar to their programme's service description were more likely to be selected. In terms of effect size, a one unit increase in the metric's Z-score (equivalent to being in the top 16% of their cohort in terms of text similarity) would translate to a 54% increase in the applicant's odds of being selected. Innovativeness also emerged as a significant predictor in one of the models (#2), with an odds ratio of 0.75 [0.57, 0.99] ($p <0.05$). This implied that more frequent usage of words related to innovativeness (e.g. "discover", "create", "new") reduced the applicants' chances of being selected, where a one unit increase in the Z-score (equivalent to being in the top 16% of their cohort in terms of innovativeness) would translate to a 25% reduction in the applicant's odds of being selected.

Additional models were built to investigate potential differences between the selection practices of the two incubators and DMS. The model coefficients are reported in Appendix A. Similarly to the findings based on the combined dataset, wordcount remained as a relatively stable, significant and positive predictor of selection in both types of programmes. However, field diversity was statistically significant in only one of the models based on DMS data, and not significant in any of the models built using the incubators' data. This may be due to the smaller sample sizes used in these programme-specific models.

Of the four models that were built exclusively with DMS data, only one sample revealed an additional significant predictor, with programme fit having an odds ratio of 1.69 [0.97, 2.97] ($p < 0.1$). On the other hand, models based on the two incubators (DataPitch & ODINE) highlighted a number of additional predictors. Three of the four models showed a positive relationship between ethnic diversity and the odds of being selected, suggesting that, when the other variables were held constant, applicants with ethnically diverse teams were 2 to 3 times more likely to be selected compared to companies where all team members were predicted to be European. Three of the models additionally showed that the number of employees was positively associated with the decision outcome, where every additional team member increased the odds of being selected by around 7 - 13%. One of the models revealed a negative association between linguistic indicators of programme fit and innovativeness with the decision outcome, with odds ratios of 0.64 and 0.65 respectively ($p < 0.1$). A further negative association was revealed for competitiveness by two of the models, with odds ratios of 0.72 and 0.62 ($p < 0.05$). These negative associations suggest that seemingly desirable linguistic qualities in applications can sometimes be associated with negative selection outcomes.

When it came to the predictive power of programme-specific models, the results were mixed. Models built with the two incubators' data performed poorly, with the highest precision score being just 0.40. On the other hand, the models built exclusively with DMS data achieved precision scores of up to 0.79, surpassing the performance of even the overall combined model that was built on a much larger dataset. This difference in performance may be attributable to the fact that the DMS model was the only one that was built using a dataset from a single programme, making it less vulnerable to the noise introduced by multiple programmes that may have had different preferences in terms of the ideal startup archetype. However, given the small sample sizes of the training and test data, and the instability of the model coefficients under different sampling conditions, any findings regarding predictive power and differences between programmes should be treated with caution.

From the collective findings observed above, we can conclude with reasonable confidence that wordcount and field diversity are likely to have contributed to increased chances of being selected. Other potentially significant variables may relate to the linguistic content of applications in terms of their programme fit, innovativeness and competitiveness; however, the significance and directionality of these influences was ambiguous and differed by sample. Similarly, team characteristics such as ethnic diversity and the number of employees were not stable in their significance across different training samples, but both were found to have a positive effect in the few models where significance was detected.

Descriptive statistics of the underlying data showed that approximately 37% of teams included individuals from under-represented demographic groups. At a more granular level of analysis, we estimate that approximately 19% of the people named inside applications were women and 81% were men. In terms of ethnicities, 81% of the individuals' names were classified as European, 11% as Asian and 8% as African. These estimates were similar across all three programmes.

In terms of company size, the average number of employees was 7 per company. 84% of applicants fell within the European Commission's definition[14] of micro-enterprises (< 10 staff), while a further 15% came from small (< 50 staff) and 1% from medium-sized enterprises (< 250 staff).

Table 4. Logistic regression results for the combined training data from all three programmes. Coefficients are reported in the form of Odds Ratios (OR) alongside significance levels at ***p<0.01, **p<0.05, *p<0.1. 95% Confidence intervals are reported in brackets for the best fitting model.

| Variable | Original training data (N = 580) | Balanced sample #1 (N = 289) | Balanced sample #2 (N = 289) | Balanced sample #3 (N = 288) |
|---|---|---|---|---|
| (intercept) | 0.15 *** | 0.49 ** | 0.40 *** | **0.43** *** [0.24, 0.78] |
| Field diversity | 1.89 ** | 1.72 * | 1.81 * | **1.93** ** [1.05, 3.57] |
| Gender diversity | 1.07 | 1.27 | 1.06 | **1.02** [0.60, 1.71] |
| Ethnic diversity | 1.24 | 1.11 | 1.42 | **1.23** [0.73, 2.08] |
| Programme fit | 1.22 | 1.54 ** | 1.14 | **1.18** [0.83, 1.66] |
| Innovativeness | 0.84 | 0.82 | 0.75 ** | **0.90** [0.69, 1.18] |
| Competitiveness | 1.04 | 0.91 | 1.12 | **1.09** [0.84, 1.42] |
| Number of employees | 1.01 | 1.00 | 1.02 | **1.01** [0.98, 1.05] |
| Wordcount | 1.84 *** | 1.65 *** | 1.90 *** | **2.04** *** [1.42, 2.95] |
| Evaluation metrics | | | | |
| Accuracy | 0.69 | 0.55 | 0.55 | **0.59** |
| Sensitivity | 0.05 | 0.38 | 0.50 | **0.57** |
| Specificity | 0.95 | 0.62 | 0.57 | **0.60** |
| Precision | 0.29 | 0.29 | 0.32 | **0.37** |
| F1 score | 0.08 | 0.33 | 0.39 | **0.45** |
| AUC | 0.60 | 0.56 | 0.60 | **0.61** |

[14] https://ec.europa.eu/growth/smes/sme-definition_en [accessed 25/12/21]

# Discussion

Our study sought to investigate the selection practices of European data incubators and accelerators through a quantitative analysis of their open call data. Using the applications received by three initiatives: DMS, DataPitch and ODINE, we extracted generalisable metrics of applicants' characteristics, and used them to build binary logistic regression models. This technique enabled us to see which attributes pertaining to company maturity, team composition and application content contributed to selection outcomes while holding the other variables constant. We evaluated the models based on their predictive power in terms of detecting successful applicants, as well as generating insights that can be used to audit selection practices.

## The predictive power of the model

One of the aims of our study was to find out whether it is possible to automatically predict the success of applications from the data contained in them. This was motivated by recent interest in the use of automated decision-making tools as a means to improve resource allocation during review (Blohm et al., 2020; Hoornaert et al., 2017; Nagar et al., 2016). Based on the specific data and explanatory variables used in our study, we found that it was possible to obtain a reasonable degree of sensitivity in detecting successful applicants (up to 79% in the case of DMS). However, this is arguably too low to be used in a real-world context, where entrepreneurial success is so rare that discarding promising companies through automation would be too risky. Moreover, the sensitivity of our best model came at the cost of precision, meaning that a substantial portion of unsuccessful applicants would also be shortlisted by the algorithm, eliminating much of the intended efficiency of adopting a decision-support system in the first place.

Given the limited representativeness of our data, and the instability of the model under different sampling conditions, it is difficult to draw general conclusions about the efficacy of predictive models of selection for data innovation programmes. It could be that larger training data and the inclusion of other variables could increase the performance of the model. Previous researchers who were successful in developing predictive models in the fields of investment and innovation contests adopted a more exploratory approach by testing many explanatory variables (Blohm et al., 2020; Hoornaert et al., 2017; Nagar et al., 2016). Among the important predictors identified by these studies, statistically significant metrics included crowd feedback, social media, prior funding gained and linguistic dimensions derived using LIWC. Unlike this prior work, our study was guided by a much smaller set of explanatory variables that were informed by our specific research context and hypotheses. However, future studies may consider exploring additional metrics as a means to increase the explanatory power of the model. As we will discuss below, any subsequent use of the algorithm for predictive purposes would still need to be preceded by an interrogation of the model parameters to ensure that it does not replicate undesirable biases from past data.

## Using the model for auditing purposes

In addition to considering the predictive power of the model for assessing new applicants, we were also interested in auditing its parameters to see how the patterns

observed in past data align with the specific objectives of publicly-funded data innovation programmes in Europe. Our first hypothesis was that companies in their earlier stages (based on the number of employees) would have a higher chance of being selected than larger more mature companies. We found no evidence to support this hypothesis, which may be attributed partly to the fact that a majority of the applications already came from small early-stage companies. Nonetheless, several of the models that were built exclusively on the incubators' data (DataPitch & ODINE) revealed a positive relationship between team size and the companies' odds of acceptance. Although this goes in the opposite direction of our initial hypothesis regarding a preference for less mature companies, our result resonates with earlier research findings where the number of employees was positively associated with SMEs' engagement in open innovation activities (De Marco et al., 2020). This also makes sense in light of our next finding, where disciplinary diversity was associated with greater odds of being selected. Given that larger teams have greater chances of capturing more than one area of expertise, it is reasonable to expect that some minimum number of team members is beneficial to achieving diversity.

Our second hypothesis was that diverse teams had higher chances of being selected compared to non-diverse teams. We considered three indicators here, including the diversity of disciplinary backgrounds, genders and ethnicities. Only one of these items was strongly confirmed by our model, where having some diversity of fields almost doubled the odds of being selected, thus confirming a preference for multidisciplinary teams. This is in line with previous research which highlighted cognitive and cultural heterogeneity as meaningful contributors to corporate innovation (Brixy et al., 2020). Besides this, we found weak evidence in favour of ethnic diversity during selection decisions, which was limited specifically to the two incubators. Moreover, no significant evidence was found to show that gender diversity was favoured during selection decisions. One possible explanation for this could be that rather than considering surface-level demographic characteristics, the evaluators in our programmes prioritised merits that were reflected by other explanatory variables related to the application content.

Although the selection decisions in our study appeared to be generally neutral in regards to the demographic composition of teams, the underlying data indicated that around 19% of the people named inside applications were women and 81% were men. This imbalance is somewhat discouraging, but it is an improvement from other recent estimates where women were found to represent just 11% of entrepreneurs in the fields of data and AI (Chae & Goh, 2020). In terms of ethnicities, 81% of the individuals' names in our data were classified as European, 11% as Asian and 8% as African. These estimates are comparable to those of Chae & Goh (2020), who found that 72% of data and AI entrepreneurs were likely to be white, with 19% being Asian, 6% African and 3% Hispanic. Based on our overall findings, we interpret that there was no explicit selection bias in our programmes against teams that included under-represented demographic groups, but this does not alleviate the fact that issues of representation and diversity continue to persist in the ventures served by these programmes.

Besides considering team characteristics, one of our hypotheses was that companies whose application content demonstrates greater innovativeness, competitiveness and programme fit will have a higher chance of being selected. The evidence in support of

this hypothesis was partial and weak. Our signal of innovativeness was based on the presence of certain words (e.g. “discover”, “create”, “new”) in the application text. Although the significance of this predictor was not consistent across different training samples, several of the models showed a negative association between the use of words related to innovativeness and the startups’ odds of being selected. Competitiveness (based on the presence of words such as “achievement”, “challenge”, “exploit”) also showed a tentative negative association in some of the models built with the incubators’ data. Programme fit, evaluated using the text similarity between applications and the programme’s service description, had ambiguous results, with a positive association detected in some of the models and a negative effect in others.

The ambiguity of our findings for the above linguistic traits could be explained by several reasons. Firstly, our approach to constructing balanced training data resulted in small sample sizes that represented different parts of the population, leading to differences in the model parameters. Moreover, it is possible that our generalisable metrics for capturing innovativeness, competitiveness and programme fit were unreliable in capturing the specific qualities that make companies unique and competitive in reality. Prior studies based on innovation contest submissions similarly found that few of the tested linguistic cues significantly contributed to explanations of success, with the effects being unstable across different samples of data (Nagar et al., 2016). Although there is uncertainty about the significance of specific linguistic qualities, the authors found the wordlength of submissions to be a consistent predictor of success (Nagar et al., 2016). Our findings support these earlier observations.

Based on the predictors of success identified in our model, both the significant and insignificant ones, we found only a partial alignment between the selection patterns detectable in past data and the socio-economic objectives that are advocated by the European Commission. There were only two explanatory variables in which we had statistical confidence. These variables included the length of application texts and the field diversity within teams, demonstrating a strong preference for diverse expertise within teams and companies that make the effort to submit informative responses at application stage. However, a number of other variables which we expected to be important did not appear to exert a reliable influence on selection decisions. For instance, while we did not detect any significant negative biases against gender-diverse and ethnically-diverse teams, neither were we able to show that the selection process consistently favoured these qualities. Moreover, we were not able to demonstrate a unanimous preference for teams that used language related to innovativeness in their applications, or those whose word usage was similar to descriptions of the programme offering.

Our findings resonate with other recent research that has detected divergences between the scope of EU-funded policy designs and the SMEs selected by these support programmes (De Marco et al., 2020). Our study reinforces the authors’ recommendation to incorporate stronger assessments of companies’ innovativeness during selection decisions (e.g in terms of knowledge and science base metrics). Additionally, we suggest that the demographic diversity of teams is an area that would benefit from concrete monitoring and support mechanisms within the selection practices of EU-funded data innovation initiatives.

## Methodological limitations and implications

Our study has a number of limitations that temper the findings and provide opportunities for future work. As discussed above, the main challenge in our study was related to small sample sizes in the data that were used to train and test the model. Although we sought to make the findings as representative as possible by drawing on data from several innovation programmes, the sparse availability of data (especially on selected applicants) led to a model that was unstable under different training samples. Moreover, we aggregated data from different programmes which, despite having similar overarching goals, may have tacitly idealised different startup profiles, potentially introducing noise into the dataset.

The experience of developing and evaluating our model's utility for predictive and auditing purposes highlighted a number of additional methodological issues and insights that may apply to other similar studies. This includes data sampling approaches, ways of quantifying the characteristics of applications and the design of the model itself. Each challenge is discussed below.

Given the competitive nature of innovation programmes, our data had an uneven distribution of classes (approximately 26% successful applicants, 74% unsuccessful). This was problematic for building a machine learning model, since the algorithm would have more opportunities to learn about unsuccessful cases compared to successful ones, making it less effective at detecting successful candidates. Similar issues with biased data have been addressed by previous researchers through resampling approaches to balance the distribution of classes, which subsequently improved model performance (Itoo & Singh, 2020; Branco et al., 2016). Likewise, we found that balancing the data through undersampling helped to increase the sensitivity of our models for detecting successful applicants.

Another challenge posed by our applications data was related to the unstructured nature of the information contained in them. While some metrics, such as the number of employees or wordcount, were relatively easy to obtain, others required additional pre-processing steps. For example, our interest in the demographic composition of teams in terms of gender and ethnic diversity could be satisfied only by making assumptions based on individuals' names, since demographic data were not captured by default. Moreover, some of the selection criteria are inherently subjective (e.g. novelty of idea, competitive edge, market opportunity), without a standardised approach to obtaining proxy metrics. Our study used only a few of the possible ways of modelling these characteristics. In the next section, we suggest some additional metrics from prior studies that could help to improve the model.

Lastly, it is meaningful to consider re-evaluating the design of our study. Our present analysis focused on evaluators' decisions as the dependent variable. However, as we observed in the results, these past decisions do not necessarily capture all of the objectives that the European Commission wishes to promote. Although this finding is valuable in itself for auditing reasons, the use of such models for predictive purposes may benefit from a different research design. As we discuss in the next section, it may be worthwhile to model the outcome based on honest signals of success or engagement achieved by startups.

### Future work

The challenges encountered by our study present a number of avenues for future work. This includes re-evaluating and enriching the variables that are included in the model, as well as increasing the size of the dataset.

With regard to the variables, both the independent and dependent variables could be improved. For the independent variables, it would be meaningful to explore additional company characteristics that were found to be meaningful in prior research (e.g. social media metrics, linguistic content of applications, prior funding gained) (Blohm et al., 2020; Hoornaert et al., 2017; Nagar et al., 2016). For the dependent variable, instead of relying on evaluators' decisions, it is worth exploring alternative measures of the applicants' current or future success. Typical examples could include the amount of funding subsequently attained by a startup, or their survival rates (Blohm et al., 2020). Other meaningful success metrics could be more specific to the goals of public funders and programme managers. For example, De Marco et al. (2020) measured the applicants' engagement in open innovation activities through indicators such as intellectual asset protection and external relations. Another potentially useful outcome metric for programme managers could be the startups' participation or engagement with the programme itself[15], making it possible to identify startups that are most likely to make efficient use of publicly funded services. A subsequent avenue for research here could involve verifying whether companies that were selected and engaged with the programme went on to achieve greater success in terms of survival or funding.

In addition to incorporating specific variables into the model, it would also be beneficial to augment the currently available data by incorporating cases from other similar programmes. At the aggregate level, this may help to increase the level of statistical confidence in the audited variables and, where applicable, raise the predictive power of the model. Some of the emerging data-centric innovation programmes could reasonably choose to adopt selection processes that proved effective in previous programmes[16]. In doing so, new initiatives may also wish to replicate or build upon the quantitative methodology presented here for auditing the selection process. This will make it possible to demonstrate the actualised (rather than intended) patterns of selection, to draw comparisons with past programmes, and help programme managers and public funders to track improvements in selection practices over time.

## Recommendations

Based on our findings, we make a number of recommendations for programme managers and researchers who are considering the use of data-driven approaches to examine and/or enhance selection practices.

Firstly, we would recommend **defining the objective of the study** from the outset. The desired purposes could include auditing past decisions, or to predict and enhance future outcomes. Depending on which of these intentions is given priority, the methodological approach needs to be adapted accordingly.

[15] At DMS, such metrics were routinely collected for the purposes of selecting participants for the annual bootcamp.

[16] The screening approaches and application forms used by our programmes are available in the data repository that accompanies this study.

If the applications data are being used for auditing purposes, we recommend the following:

- **Collect data about the demographic composition of applicant teams**. Although we were able to show that tenuous proxies based on individuals' names can be used to derive such metrics, it would be better to collect explicit self-reported data about the equality, diversity and inclusion (EDI) characteristics of teams. Progress towards this goal is already being made by recent data innovation programmes such as MediaFutures[17], where EDI data are systematically gathered as part of the open call applications.
- **Use multivariate analysis** to measure the effect sizes of specific traits while controlling for other variables. This is particularly useful for ensuring that high-level indicators (e.g. percentages of under-represented groups) do not overlook other important aspects of high quality applications (e.g. length and content).

If the applications data are being used for predictive and/or decision-support purposes, we would recommend researchers to:

- Incorporate the above recommendations to **audit the model parameters before using the solution**. This is important for ensuring that there are no embedded biases that go against the objectives of the funder.
- Instead of relying on evaluators' decisions (accept/reject), **consider modelling outcomes based on honest signals of success** such as the subsequent revenue, survival rate or service engagement achieved by applicants.
- **Decide whether it is sufficient to evaluate applications based only on their merit, or if the system needs to actively boost demographically diverse teams**.
- **Balance the training data** to improve the model's ability to detect rarer high-quality applicants.

## Conclusion

Despite the limitations of our study and the numerous areas for improvement, our findings provide various methodological and empirical insights. Firstly, we have shown that predictive modelling of selection practices in European data innovation programmes is a challenging task, where the use of past data for the purposes of building decision-support systems is still unlikely to be a feasible option. Part of this relates to a shortage of context-specific data about selected applicants, with another issue presented by the risk of recreating potentially undesirable patterns from historical data. Indeed, one of the core contributions of our work relates to auditing past decisions to ensure that they align with the intended objectives of programme managers and public funders.

Some of the selection criteria which we sought to verify were poorly supported by our data (e.g. linguistic indicators of programme fit, innovativeness and competitiveness). However, we learnt that there was a consistent preference for longer application

[17] https://mediafutures.eu/opencall/ [accessed 15/12/21]

answers and interdisciplinary teams. Moreover, we did not detect any negative biases against teams containing women and non-European ethnicities, but we were also unable to detect a convincing preference for teams that were inclusive of this kind of diversity. Based on our observations, we provided a number of methodological recommendations to help other programmes to monitor and audit their selection practices, helping to ensure ongoing alignment with the European Commission's objectives.

## References

Aerts, K., Matthyssens, P. and Vandenbempt, K., (2007). Critical role and screening practices of European business incubators. *Technovation*, 27(5), pp.254-267.

Blohm, I., Antretter, T., Sirén, C., Grichnik, D., & Wincent, J. (2020). It's a Peoples Game, Isn't It?! A Comparison Between the Investment Returns of Business Angels and Machine Learning Algorithms. *Entrepreneurship Theory and Practice*, 1042258720945206.

Bone, J., Allen, O., & Haley, C. (2017). Business incubators and accelerators: The national picture (No. 2017/7). BEIS Research paper.

Branco, P., Torgo, L. and Ribeiro, R.P., (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2), pp.1-50.

Brixy, U., Brunow, S., & D'Ambrosio, A. (2020). The unlikely encounter: Is ethnic diversity in start-ups associated with innovation?. *Research Policy*, 49(4), 103950.

Chae, B. K., & Goh, G. (2020). Digital Entrepreneurs in Artificial Intelligence and Data Analytics: Who Are They?. *Journal of open innovation: technology, market, and complexity*, 6(3), 56.

Chang, H. C. H., & Fu, F. (2021). Elitism in mathematics and inequality. *Humanities and Social Sciences Communications*, 8(1), 1-8.

Clarysse, B., Wright, M. and Van Hove, J., (2015). A look inside accelerators. London: Nesta.

Cohen, S. and Hochberg, Y. (2014). Accelerating Startups: The Seed Accelerator Phenomenon. Available at SSRN: https://ssrn.com/abstract=2418000 or http://dx.doi.org/10.2139/ssrn.2418000

De Marco, C. E., Martelli, I., & Di Minin, A. (2020). European SMEs' engagement in open innovation When the important thing is to win and not just to participate, what should innovation policy do?. *Technological Forecasting and Social Change*, 152, 119843.

Dempwolf, C. S., Auer, J., & D'Ippolito, M. (2014). Innovation accelerators: Defining characteristics among startup assistance organizations. *Small Business Administration*, 1-44.

D'Ignazio, C., & Klein, L. (2020). Data Feminism. MIT press. https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/3

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hoornaert, S., Ballings, M., Malthouse, E. C., & Van den Poel, D. (2017). Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time. *Journal of Product Innovation Management*, 34(5), 580-597.

Huang, L. (2018). The role of investor gut feel in managing complexity and extreme risk. *Academy of Management Journal*, 61(5), 1821-1847.

Itoo, F. and Singh, S., (2020). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, pp.1-9.

Kleinert, S., & Mochkabadi, K. (2021). Gender stereotypes in equity crowdfunding: the effect of gender bias on the interpretation of quality signals. *The Journal of Technology Transfer*, 1-22.

Lumpkin, J. R., & Ireland, R. D. (1988). Screening practices of new business incubators: the evaluation of critical success factors. *American Journal of Small Business*, *12*(4), 59-81.

Maxwell, A. L., Jeffrey, S. A., & Lévesque, M. (2011). Business angel early stage decision making. *Journal of Business Venturing*, 26(2), 212-225.

Nagar, Y., De Boer, P., & Garcia, A. C. B. (2016). Accelerating the review of complex intellectual artifacts in crowdsourced innovation challenges.

Nielsen, M., Bloch, C., & Schiebinger, L. (2018). Making Gender Diversity Work for Scientific Discovery and Innovation. https://doi.org/10.1038/s41562-018-0433-1

Ortiz, L. P., Díez, Á. S., & Apaolaza, A. I. V. (2020). Employment quality and gender equality. an analysis for the european union. *Regional and Sectoral Economic Studies*, 20(2), 5-18.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. [accessed 29/11/21]

Priestley, M., Thuermer, G. and Simperl, E. (2021). Selection practices of European data incubators and accelerators [Source Code]. *CodeOcean*. https://doi.org/10.24433/CO.7674311.v1

Short, J.C., Broberg, J.C., Cogliser, C.C. and Brigham, K.H., (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods*, 13(2), pp.320-347.

Sood, G., & Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. arXiv preprint arXiv:1805.02109.

Thada, V., & Jaglan, V. (2013). Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4), 202-205.

Torgo, L., Branco, P., Ribeiro, R.P. and Pfahringer, B., (2015). Resampling strategies for regression. *Expert Systems*, 32(3), pp.465-476.

Wang, Y., & Chen, H. J. (2012). Use of percentiles and z-scores in anthropometry. In *Handbook of anthropometry* (pp. 29-48). Springer, New York, NY.

Yang, S., Kher, R., & Newbert, S. L. (2020). What signals matter for social startups? It depends: The influence of gender role congruity on social impact accelerator selection decisions. *Journal of Business Venturing*, 35(2), 105932.

# Appendix A

Tabes 5 and 6 present logistic regression model results that were based on specific data samples limited to DMS and the two incubators (DataPitch and ODINE).

Table 5. Logistic regression results with training data only from DMS. Coefficients are reported in the form of Odds Ratios (OR) alongside significance levels at ***p<0.01, **p<0.05, *p<0.1. 95% Confidence intervals are reported in brackets for the best fitting model.

| Variable | Original training data (N = 265) | Balanced sample #1 (N = 140) | Balanced sample #2 (N = 140 ) | Balanced sample #3 (N = 139) |
|---|---|---|---|---|
| (intercept) | 0.23 *** | **0.73** [0.27, 2.03] | 1.13 | 0.37 ** |
| Field diversity | 1.71 | **1.15** [0.44, 3.00] | 1.22 | 2.84 ** |
| Gender diversity | 0.93 | **1.11** [0.49, 2.51] | 0.98 | 0.89 |
| Ethnic diversity | 0.92 | **0.76** [0.34, 1.71] | 0.66 | 1.27 |
| Programme fit | 1.38 | **1.12** [0.63, 2.00] | 1.37 | 1.69 * |
| Innovativeness | 0.95 | **0.93** [0.61, 1.42] | 0.96 | 1.01 |
| Competitiveness | 1.12 | **0.93** [0.60, 1.45] | 1.33 | 1.15 |
| Number of employees | 1.01 | **1.04** [0.98, 1.09] | 1.00 | 1.02 |
| Wordcount | 1.81 *** | **3.26** *** [1.65, 6.41] | 1.94 ** | 1.48 |
| Evaluation metrics | | | | |
| Accuracy | 0.78 | **0.67** | 0.64 | 0.64 |
| Sensitivity | 0.26 | **0.79** | 0.74 | 0.79 |
| Specificity | 0.98 | **0.63** | 0.60 | 0.58 |
| Precision | 0.83 | **0.45** | 0.42 | 0.43 |
| F1 score | 0.40 | **0.58** | 0.54 | 0.56 |
| AUC | 0.82 | **0.77** | 0.79 | 0.83 |

Table 6. Logistic regression results with training data from DataPitch and ODINE. Coefficients are reported in the form of Odds Ratios (OR) alongside significance levels at ***p<0.01, **p<0.05, *p<0.1. 95% Confidence intervals are reported in brackets.

| Variable | Original training data (N = 314) | Balanced sample #1 (N = 154) | Balanced sample #2 (N = 154) | Balanced sample #3 (N = 154) |
|---|---|---|---|---|
| (intercept) | 0.11 *** | 0.28 *** | **0.29 *** [0.12, 0.74]** | 0.52 |
| Field diversity | 1.37 | 1.94 | **0.90 [0.40, 2.06]** | 0.84 |
| Gender diversity | 0.89 | 1.20 | **0.95 [0.46, 1.95]** | 0.75 |
| Ethnic diversity | 2.13 *** | 2.05 * | **3.21 *** [1.53, 6.72]** | 1.72 |
| Programme fit | 0.77 | 0.95 | **0.64 * [0.39, 1.06]** | 0.77 |
| Innovativeness | 0.89 | 0.99 | **0.68 * [0.46, 1.00]** | 1.03 |
| Competitiveness | 0.72 ** | 0.62 ** | **0.79 [0.54, 1.17]** | 0.76 |
| Number of employees | 1.07 ** | 1.05 | **1.13 ** [1.01, 1.26]** | 1.08 ** |
| Wordcount | 1.85 *** | 1.77 ** | **2.15 *** [1.30, 3.54]** | 1.65 ** |
| Evaluation metrics | | | | |
| Accuracy | 0.80 | 0.56 | **0.57** | 0.61 |
| Sensitivity | 0.00 | 0.40 | **0.40** | 0.33 |
| Specificity | 0.98 | 0.59 | **0.61** | 0.67 |
| Precision | 0.00 | 0.19 | **0.19** | 0.19 |
| F1 score | 0.00 | 0.26 | **0.26** | 0.24 |
| AUC | 0.48 | 0.49 | **0.45** | 0.45 |