



The use of AI in legal systems: determining independent contractor vs. employee status

Maxime C. Cohen¹ · Samuel Dahan²  · Warut Khern-am-nuai¹ · Hajime Shimao² · Jonathan Touboul³

Accepted: 13 February 2023

© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract

The use of artificial intelligence (AI) to aid legal decision making has become prominent. This paper investigates the use of AI in a critical issue in employment law, the determination of a worker's status—employee vs. independent contractor—in two common law countries (the U.S. and Canada). This legal question has been a contentious labor issue insofar as independent contractors are not eligible for the same benefits as employees. It has become an important societal issue due to the ubiquity of the gig economy and the recent disruptions in employment arrangements. To address this problem, we collected, annotated, and structured the data for all Canadian and Californian court cases related to this legal question between 2002 and 2021, resulting in 538 Canadian cases and 217 U.S. cases. In contrast to legal literature focusing on complex and correlated characteristics of the employment relationship, our statistical analyses of the data show very strong correlations between the worker's status and a small subset of quantifiable characteristics of the employment relationship. In fact, despite the variety of situations in the case law, we show that simple, off-the-shelf AI models classify the cases with an out-of-sample accuracy of more than 90%. Interestingly, the analysis of misclassified cases reveals consistent misclassification patterns by most algorithms. Legal analyses of these cases led us to identify how equity is ensured by judges in ambiguous situations. Finally, our findings have practical implications for access to legal advice and justice. We deployed our AI model via the open-access platform, <https://MyOpenCourt.org/>, to help users answer employment legal questions. This platform has already assisted many Canadian users, and we hope it will help democratize access to legal advice to large crowds.

Keywords Employment law · Artificial intelligence · Predictive models · Decision-support legal systems

Maxime C. Cohen, Samuel Dahan, Warut Khern-am-nuai, Hajime Shimao and Jonathan Touboul have contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

Much is at stake in determining whether a worker should be classified as an employee or as a contractor. This question strikes at the core of the structure of the economy, as it has implications for both employers' and workers' rights. As argued by the recent California Supreme Court decision, *Dynamex Operations West Inc. v. Superior Court of Los Angeles County*, "the question whether an individual worker should properly be classified as an employee or, instead, as an independent contractor has considerable significance for workers, businesses, and the public generally."¹

The question of worker status determination has come to the forefront of the public debate thanks to a string of cases regarding the status of gig workers. The gig economy, which enables flexible work arrangements for workers in service-based, on-demand environments (Huang et al. 2020), has seen tremendous growth in recent years. In the U.S., it has been estimated that 35% of the workforce participated in the gig economy in 2018, while this figure was estimated to be 50% in the UK in 2020 (Marius 2021). The value of this market has also grown tremendously, with a twenty-fold increase over the last decade and projected value of \$335 billion dollars in 2025 (Taylor et al. 2020).² Thus, issues surrounding the gig economy have attracted the attention of numerous academic researchers, particularly regarding the behavior of workers (e.g., Greenwood et al. 2017; Allon et al. 2018; Babar and Burtch 2020). For such companies, the issue of worker classification has critical implications for workers' rights. In fact, many gig workers are classified by their employers as independent contractors and therefore cannot receive minimum employment benefits, such as overtime and vacation pay.

Gig economy companies have faced at least 40 major legal challenges worldwide, as transportation and delivery drivers have tried to improve their rights. In a bold move, California voters passed Prop 22 in November 2020, the most expensive ballot initiative in Californian history, allowing gig companies such as Uber, Lyft, and DoorDash to continue classifying their workers as independent contractors while offering them a few additional benefits.³

The answer to this seemingly basic legal question is far from straightforward. Indeed, worker classification relies on multiple factors that are often correlated and may even depend on the nature of the work (Burosu 2021). Furthermore, it often implies an implicit precedence and the weighting of several work characteristics (Deakin 2020). In this paper, we approach this legal question from the perspective of algorithmic decision-making, using two complementary datasets: a compilation of Canadian and U.S. (California) cases concerning the issue of worker classification.

The application of artificial intelligence (AI) techniques in legal contexts has seen remarkable developments. AI can be used in dispute resolution (Cohen et al. 2021), such as determining the employment notice period following job

¹ 416 P.3d 1 (Cal. 2018). In this case, the issue at stake was the application of a statutory wage order.

² <https://www.pwc.es/es/publicaciones/digital/evaluacion-economia-colaborativa-europa.pdf>

³ <https://fortune.com/2020/11/04/prop-22-california-proposition-uber-lyft-gig-companies-workers-passes/>

termination (Dahan et al. 2020) and predicting asylum court decisions (Dunn et al. 2017). In the U.S., individualized automatic estimates of the likelihood of criminal recidivism provided by the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system were utilized for nearly a decade to assist judges and parole boards for sentencing decisions (Shimao et al. 2022). Law firms have also increasingly adopted AI models to provide legal services to their clients, particularly regarding tax and contract law (Kauffman and Soares 2020). However, the *computability of law*, namely the ability of algorithms to provide legal analysis, advice, or predictions, remains a subject of fierce debate, especially regarding adjudication. At one end of the spectrum, several scholars have claimed that law is on a path towards “legal singularity,” in which the law will eventually be complete with no genuinely grey areas and thus fully computable (Alarie 2016). Others have been more skeptical, arguing that certain features of legal reasoning are not consistent with AI, particularly because of the fundamental nature of legal reasoning, specifically the reflexivity of legal knowledge and the incompleteness of legal rules (Markou and Deakin 2020).

In this paper, we provide a quantitative account of the computability of law regarding the question of worker classification. To this end, we collected and analyzed two complementary datasets. The first dataset includes all Canadian court cases related to worker classification between 2001 and 2021, providing invaluable information on Canadian case law. The concept of worker classification is not unique to Canada. It is found in a functionally similar form, albeit with some important differences, in many legal systems (Deakin et al. 2005). To explore the potential application of algorithms in comparative laws, our second dataset consists of all relevant cases in the state of California between 1989 and 2020. Finally, we also gathered data from the real-world field deployment of our AI model via an open-access platform, <https://MyOpenCourt.org/>, to help workers and employers answer legal questions about employment law. This platform has been available to the public since June 2020. Our two datasets (as well as our field data) allowed us to investigate the validity and generalizability of our results.

In Sect. 2, we investigate the ability of machine learning (ML) algorithms to predict workers’ status. Using interpretable AI algorithms, we further decipher the structure of the resulting predictive methods and infer typical judges’ decision-making processes as well as the main factors that determine workers’ status. We found a surprising consistent judicial decision making in this domain, whereby even simple off-the-shelf algorithms could predict workers’ status with very high accuracy (around 90%). However, we will show that a number of cases were systematically misclassified by our algorithms. Isolating those cases allowed us to determine a subset of unconventional decisions that are based on some legitimate, or at least deliberate, judicial choices.

By combining the Canadian and Californian datasets, we propose using AI algorithms as a tool for comparative law. By using algorithms trained on a specific dataset to predict outcomes for the other dataset, we will showcase a striking and unexpected consistency between two vastly distinct jurisdictions. Thus, our results suggest that AI algorithms could provide a new systematic tool for comparative law.

Finally, we deploy our AI model in the field for the benefit of the public and to potentially assist self-represented litigants. Most research endeavours in this area have focused on technical aspects, such as measuring the accuracy of models in predicting case outcomes. However, our understanding of the legal and practical implications of AI-powered systems in legal contexts remains limited. We discuss the practical implications of this research, including the deployment of <https://MyOpenCourt.org/>, which is a direct-to-public platform based on our AI model, for helping users answer legal questions on employment law. Deploying this platform to predict workers' status along with clear explanations, allows us to directly help the public and reduce lawyers' workload. It is also one step forward in democratizing the access to legal advice and to justice.

2 Results

2.1 Worker classification and data analyses

2.1.1 Factors established by the case law

In September 2001, the Canadian Supreme Court delineated the contours of the legal determination of workers' status in the landmark case *671122 Ontario Ltd. v Sagaz Industries Canada Inc.* (hereafter, *Sagaz*). This text clearly expresses the position of the common law in determining whether a work relationship can be construed as an employer-employee relationship:

"In making this determination, the level of control the employer has over the worker's activities will always be a factor. However, other factors to consider include whether the worker provides his or her own equipment, whether the worker hires his or her own helpers, the degree of financial risk taken by the worker, the degree of responsibility for investment and management held by the worker, and the worker's opportunity for profit in the performance of his or her tasks."

We collected all 1,391 cases between 2001 and 2021 that referred explicitly to *Sagaz* (see the "[Methods](#)" section below for more details). Among these, 538 cases were related to the determination of an employee/contractor status and formed the basis of our dataset. *Sagaz* not only provides a reference for identifying the relevant judgements, but also outlines the following set of attributes to use in monitoring each case:

- Level of supervision of the work,
- Ownership of equipment,
- Ability to hire helpers,
- Degree of financial risk (level of financial investment, business expenses, and liability), and
- Opportunity for profit (in particular, considering whether the worker was paid via fixed wages, commissions, or a combination).

In addition to the above factors, which were explicitly stated in the Sagaz case, we also monitored other key elements in each case, including

- Whether the contract included exclusivity of services,
- Who sets the work hours,
- Where the work is performed, and
- Whether the worker was required to wear a uniform.

Finally, factual elements, including year of judgement, court, province, industry, and duration of services, were also collected.

We followed a similar methodology in collecting data on U.S. cases. To ensure comparable volumes of cases and avoid heterogeneity between different jurisdictions, we focused our attention on the state of California, because of its sizable technology industry and its central role in the recent debate on the status of gig workers. The question of worker status can be approached from various legal angles, and there are many legal tests available depending on the issue at stake (e.g., tax, employment insurance, and social security). In this paper, we consider the leading court case *S. G. Borello & Sons, Inc. v. Department of Industrial Relations* (1989), which is largely considered to have set a landmark for worker classification. In this case, *S. G. Borello & Sons, Inc.*, a Gilroy grower, was found guilty of having failed to secure workers' compensation coverage for the 50 migrant harvesters of its cucumber crop. By placing expectations on worker's classification, this case has led to the identification of 11 largely correlated factors that determine worker classification, and that have become known as the Borello multifactor test for determining various dimensions of the level of the employer's control of the worker. In addition to the above Sagaz criteria, which are explicitly covered in the Borello test, the latter also includes other dimensions, notably the following:

- Whether the worker engages in an occupation distinct from that of the company or part of the company's regular business,
- The skill level required in the occupation,
- The duration of services and the degree of permanence of the work relationship, and
- The perception by one of the parties of the existence of an employer/employee relationship.

The overlap in criteria between the two jurisdictions allowed us to use a consistent scale for labeling Canadian and Californian cases, which was crucial in making fair comparisons. We collected and processed 810 cases that explicitly referenced the Borello case between 1989 and 2020. Among these cases, a total of 217 cases were relevant to the question of worker classification, and data were collected to ensure consistency with the Canadian dataset and to conserve as much information as possible for further use, leaving us with a total of 28 criteria including the Sagaz criteria (see the "[Methods](#)" section).

2.1.2 Case typology in Canada and California

Beyond opening the way to prediction, these datasets provide invaluable insights into the specifics of worker classification litigation history in Canada and California. Interestingly, we found that there was no significant imbalance in the court decisions, and comparable fractions of workers were classified as employees in California (60.4%) and Canada (61.2%),⁴ The overwhelming majority of cases corresponded to workers in the service industry (90.5% in Canada and 83% in U.S.), with brief work duration (in half of the cases, workers worked for less than a year in both jurisdictions, and in more than 80% of cases, workers worked for less than five years).⁵ In both datasets, the majority of workers could choose where to work (74.5% in Canada and 75.6% in California).

Other criteria were less balanced and showed less similarity between the two jurisdictions, likely stemming from differences in the work culture. In Canada, workers had relatively a low risk of loss in two out of three cases, which was more frequent than in California (43%). Workers did not own their tools in about 40% of Canadian cases versus 14.3% of cases in California, and high chances of profit (reflecting whether workers were paid fixed wages or commissions) were 43% in Californian cases compared with 5.7% in Canada.

One important aspect of investigating judicial decision-making is the characterization of the correlation between the different criteria, and whether some criteria weigh more than others. For example, the ownership of tools, as a concrete financial investment, should directly correlate with the risk of loss. We computed the correlation matrices between criteria in both datasets (see Fig. 1) and obtained similar correlation maps in Canada and California. Overall, except for the length of service that, quite unsurprisingly, was uncorrelated to other criteria, we found a high correlation between the different features in our data.

Interestingly, the correlation between the outcome and the features in the dataset provided the first assessment of feature importance and predictability. We found that outcomes in Canada and California had a strong correlation (50% and above) with seven common criteria of the 11 criteria, with lower correlations with the chance of profit (38% in Canada and 21% in California), uniform wearing requirements (both around 30%), and extremely low correlations with the length of service. The only difference was a lower correlation with the risk of loss in California (34%) compared with Canada (61%), potentially reflecting the already noted heterogeneity between the two datasets regarding this specific criterion.

In particular, we found that the outcome was correlated with whether the work was supervised or reviewed in California (74%). In Canada, the highest correlation

⁴ Note that Canada has a third category, the *dependant contractor* which is a worker that is not an employee but still considered economically dependent on the company (or individual) they work for. This category has not been widely used by Canadian courts, as we only found 16 such court outcomes and, thus, it was not considered in our analysis.

⁵ While the gig economy is more prone to job hopping, shorter work duration is a general trend in the workforce, see <https://www.forbes.com/sites/jeannemeister/2012/08/14/the-future-of-work-job-hopping-is-the-new-normal-for-millennials/> accessed December 2021; see also Pranaya (2014).

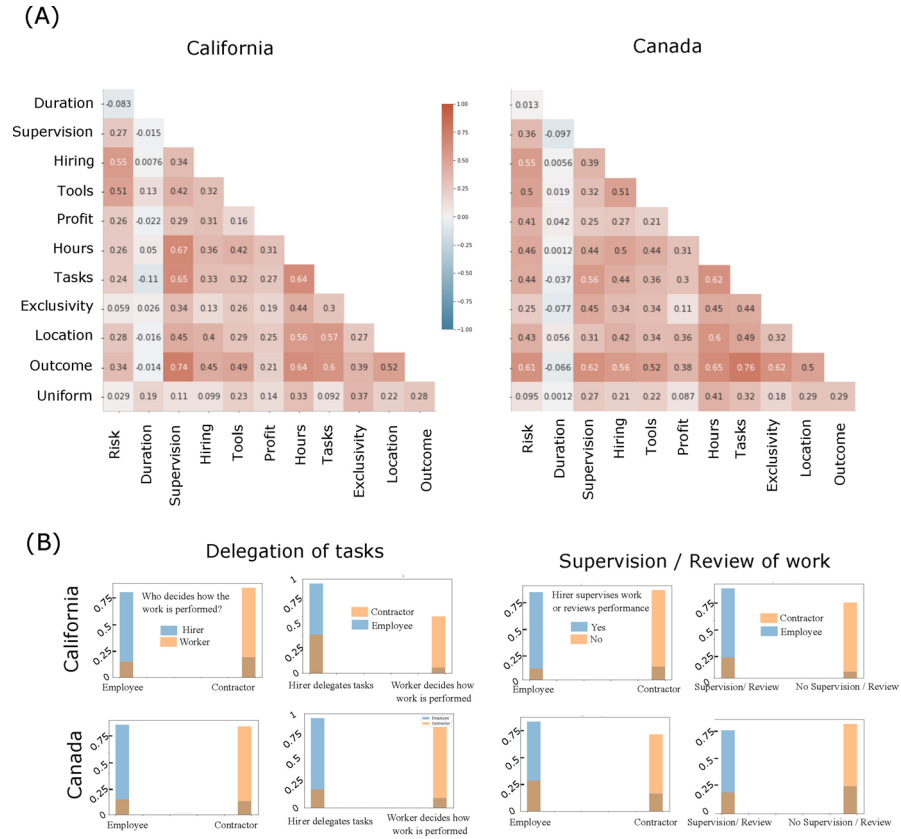


Fig. 1 Statistics of the Canadian and Californian datasets. (A) Correlation matrix between pairs of features and between features and outcomes for Canada (right) and California (left). Our data convention (see “Methods” section) consisted of coding feature scores with increasing values for decreasing levels of control (e.g., works involving a supervision were coded as 1, and works with no supervision were coded 2). (B) To appreciate the meaning of high correlations of factors with the worker’s status, we plotted the data distribution for the two most highly correlated features with the outcomes: delegation of tasks and supervision of work. Histograms of the outcomes as a function of the feature and of the distribution of the outcomes given a feature value showed strong imbalances

was with whether the worker was delegated specific tasks to complete the job or whether the worker had the ability to decide which task to perform to complete the work (76%). In fact, we observed that over 80% of the workers who either were delegated specific tasks to complete the job or were directly supervised (or reviewed) by their hirer were classified as employees. This indicated the likelihood of a high predictability of the decisions, with those two criteria constituting factors that determined the worker’s status. Indeed, we found that predicting the outcome directly based on delegation of tasks provided an astonishing 82% accuracy for the Californian dataset and 89% accuracy for the Canadian dataset. Similarly, predicting the outcome solely based on whether the work was reviewed or supervised yielded high accuracies of 88% in California and 81% in Canada.

Therefore, we observed that when we had information on the possible delegation of tasks and/or the supervision of work, highly consistent decisions were provided by the judges. However, in many situations, these dimensions of work were unspecified in the judgement. Information may be missing when it is absent in the source. For instance, a judge may not have found it relevant to specify whether a worker could wear a uniform or not, while in fact the worker was required to do so. Other features may be missing because they do not apply to a given work situation. Missing information may have two consequences: (i) likely hindering the predictive ability of the models, and (ii) potentially carrying information about the eventual outcome. We further examined the issue of missing data.

For instance, data on whether the work was supervised were absent in 40% of the cases (84 cases) in California and 17% of the cases in Canada, while the ability to delegate tasks was absent in less than 13% of the cases in both datasets. Another aspect of missing data could be an *a posteriori* correlation with the outcome: judges did not mention a particular feature of specific outcomes. We confirmed the absence of such a correlation (with virtually no correlation in the Canadian dataset and very low correlations in the Californian dataset. See the “[Methods](#)” section).

Even in the absence of a direct correlation with the outcome, we carefully handled the missing information. We tested multiple ways of handling missing data in the sequel, including (i) simply considering ‘missing’ as a new category for each class, (ii) inputting the missing fields with the median value over the entire data by default, or (iii) relying on a more sophisticated method based on predicting the most likely value by using the observed features (For such predictions, we used a random forest predictor. See the “[Methods](#)” section). Our results were based on the latter methodology, but all methods yielded similar results.

2.2 Prediction results and inferring judicial reasoning

We then used AI to further decipher judicial reasoning and test the predictability of the worker classification outcome. We used three distinct approaches to train and test our algorithms—Canadian data only, Californian data only, and combined data—which provided us with nine accuracy estimates, corresponding to (i) the ability of the algorithms to predict cases within a given jurisdiction, (ii) the ability of the algorithms to generalize from one jurisdiction to the other, and (iii) the possibility of enhancing generalizations by combining different jurisdictions.

Based on our observation of a high correlation between the outcome variable with both the delegation of tasks and the ability to hire employees, we first considered a simple decision tree. This ML method was used to identify an optimal flow-chart-like structure for predicting the outcome based on the features, whereby the algorithm at each step of the flow-chart determined the optimal binary decision that best reduced the so-called *impurity* of the subsets, defined as the level of heterogeneity of the outcome values. To avoid over-fitting our relatively small datasets, we used shallow decision trees with a depth of three. This yielded the following nine accuracy values, which were computed using a ten-fold cross-validation (for more

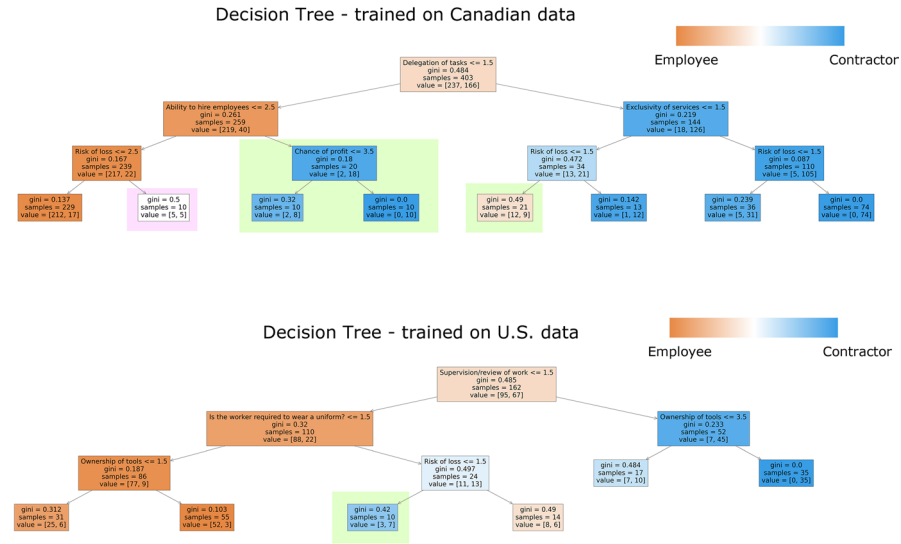


Fig. 2 Decision trees, along with the impurity level (Gini index), the number of samples in each node, and the total number of cases of each type. Orange: employee, Blue: contractor, with gradient indicating the purity level (i.e., certainty of the classification). Top: tree trained using Canadian data, Bottom: tree trained using Californian data. The differences relative to a naive algorithm based only on the root of the tree are highlighted in green; a fully ambiguous situation is highlighted in pink for the Canadian dataset

details, see the “Methods” section with an assessment on the dependence of accuracy on the number of cross validations). We obtained the following results:

Train \ Test	Canada	U.S.	All
Canada	0.85	0.82	0.82
U.S.	0.88	0.84	0.82
All	0.88	0.82	0.82

Among the most relevant observations, we noted that predicting new Canadian worker classifications from the algorithm trained on Canadian cases achieved a cross-validation accuracy score of 88%, but, more interestingly, the remarkable score of 82% for all cases. In the same vein, training the algorithm on U.S. cases led to a 84% prediction accuracy in the U.S. cases and a 88% accuracy in the Canadian cases. Training on all cases did not improve the result further.

The decision tree algorithms, because of their flow-chart structure, allowed for a deeper understanding of judicial reasoning. In Fig. 2, we depict the decision trees generated by the algorithms. In addition to the (already observed) importance of the ability to delegate tasks in Canada, we observed that the tree was able to capture two main corrections to the naive, uni-dimensional perspective. For instance, workers who did not have the ability to delegate tasks but could hire employees were often considered contractors, particularly when there was a high chance of profit (blue branches in Fig. 2 arising from the leftmost split). However, workers who could

delegate tasks but worked exclusively for the hirer and had a low risk of loss were classified as employees.

Similarly, the tree trained on the U.S. data identified supervision of work as the main criterion (as previously observed), but it had a single refinement relative to a purely binary decision: workers who were supervised but could hire employees were more likely to be considered contractors.

We concluded that the simplest AI algorithms could achieve a high predictive accuracy within a given jurisdiction, and, more strikingly, between two different, potentially vastly distinct, jurisdictions. Moreover, the decision tree provided a finer, yet realistic vision of the criteria used by judges to determine workers' status. Before we delve into a deeper analysis of these observations, we discuss the outcome of alternative, more complex AI algorithms to determine whether nonlinear or non-binary decisions could further improve prediction accuracy. We note that ML algorithms crucially rely on having access to a good sample of labeled data for training purposes. Such a condition is true for all studies that rely on supervised ML models.

First, we tested random forests, which combined multiple decision trees and used them as an ensemble for prediction. The trees were different because they were constructed based on random splits of the data (bagging, or bootstrap aggregation) and featured randomness (trees were selected from a random subset of features instead of using a greedy algorithm that found the best feature for a split). Random forests, with a maximum of 15 trees and a maximum depth of four, yielded an improved precision as reported below:

Train \ Test	Canada	U.S.	All
Canada	0.90	0.82	0.86
U.S.	0.90	0.86	0.89
All	0.91	0.80	0.9

An estimate of the relative importance of each feature in the dataset was provided by estimating the statistics of the decrease in *impurity* in which their split resulted. We found a quantitative confirmation of the previous observations: for both the Canadian and Californian datasets, we found similar importance levels, except for a switch between supervision of work and delegation of tasks. In the Canadian dataset, delegation of tasks was the most important feature with 26% decreased impurity, preceding exclusivity of services (16%), ability to set work hours (13%), supervision of work (12%), and risk of loss (10%). In the Californian dataset, supervision of work accounted for an averaged of 23% in decreased impurity, preceding ability to set work hours (22%), exclusivity of services (14%), whereas all other features were below 10%.

Other classification algorithms were also applied to further analyze our ability to predict workers' status, including logistic regression, *k*-neighbors classifier, support-vector classification (SVC), Gaussian process classifier, and the AdaBoost or XGBoost classifiers (see the "Methods" section). They all provided predictions with similar (or better) accuracy but were less conducive to interpretation. However, an interesting conclusion was that most cases were misclassified by decision trees,

and random forests were also misclassified by other algorithms. We thus turned our attention to these misclassified cases.

2.3 Misclassified cases and adaptive judicial decisions

Of 755 cases (in both Canadian and Californian datasets) and seven classification algorithms, we observed that 612 (81%) cases were correctly classified by all algorithms, showing a high degree of consistency in capturing judicial behavior regardless of the algorithm used. Interestingly, 5.7% of the cases (43 cases) were misclassified by all algorithms, some even providing predictions with high degrees of confidence. For instance, for the 2012 *Pluri Vox Media Corp. v. Canada*, all algorithms predicted that the worker, Martin Reesink, was a contractor with high confidence levels (99.1% confidence for the logistic regression, 98.8% for SVC, and 100% for random forest or XGboost). This prediction was likely associated with the unique position of the worker, Martin Reesink, who was in fact the director of Pluri Vox. Many of the attributions associated with this leadership position, in particular his ability to hire an independent contractor in his work with Pluri Vox, was not considered to imply that he himself was an independent contractor.

In fact, these algorithmically misclassified cases suggested a slight deviation from the application of the legal precedent, as they did not always signal inconsistent or biased decision-making patterns, in which judges made diverging decisions based on factors that were unrelated to the merits of the case. It depended on the cause of the divergence. For example, many situations may have presented unusual circumstances that warranted deviations from rigid algorithmic decisions based only on a few aspects of the case. However, it was also possible that the judges were affected by legally irrelevant factors such as behavioral biases, possibly leading them to ignore the circumstances of the case (Chen and Egel 2017). In our assessment, we identified three patterns of algorithmic errors and divided the misclassified cases into three groups according to the cause of the deviation.

1. The first group consisted of “simple” cases that were misclassified because of highly unusual factual or procedural circumstances. Consider, for instance, the 2011 *Atlantic Council of Canada v. Minister of National Revenue*, which set out to classify the status of an intern. In this case, by considering all factual aspects of work dependence, the algorithms predicted that the worker was an employee. Indeed, the intern was directly supervised, had no ability to hire employees or to delegate tasks, was required to work exclusively for the council, did not own his tools, and did not bear any risk of loss. All these aspects, in regular work relationships, are highly determinant of employee status and not contractor status. However, interns occupy a category that is outside the dichotomy of employee and contractor. Their level of dependence is in fact relative, as the work is highly temporary and part of training. As a result, the judges considered that the obligations of the council to the intern were closer to those they had to contractors. Another case where algorithms could not make an accurate prediction is the 2014 *Morris Meadows Country Holidays and Seminars Ltd. v. Minister of National Revenue*.

This case considered the classification of 25 people working in a meeting, dining, and sleeping facility. In this case, the judge found that 24 of them were employees, while a single person, who in all respects shared the same features as the other workers, was in fact a contractor because of the unpredictable and unreliable nature of the functions performed by this worker, who was “spasmodically called upon once in a while to do a bit of work for an indeterminate time.” Of course, this specific determination by the judge was unpredictable, particularly when our algorithms lent themselves to a few criteria that did not include the regularity of the work, and thus predicted that all 25 workers were employees.

2. The second group included “difficult” or “borderline” cases where the worker’s status was not straightforward. These cases presented features that strongly supported the classification of both an employment relationship and an independent contractor status. For instance, in the 2013 *Atlantic Council of Canada v. Minister of National Revenue*, the worker was classified by the court as an independent contractor because he had a separate carpentry business, and the relationship was not exclusive. However, the carpenter was supervised (i.e., was told what to build), was provided with tools, needed permission to leave early, and was given occasional building maintenance work. This situation included a subordinate relationship, but because the worker owned his own business, it clearly bore aspects of both contractor and employee. Ultimately, the court decided that the non-exclusive nature of the relationship prevailed over the subordinate relationship. Similarly, we highlighted two additional notable cases, *S.C. t/a Hollywood Limousine Service and T. (N.), Re* (Case 195 in our database and hereafter), and *S.C. t/a Hollywood Limousine Service and T. (N.), Re* (Case 590). Both cases involved a topical policy issue, namely, whether gig workers are contractors or employees (although the policy dimension was not formally addressed in those cases). All AI models were wrong in both cases—but for different reasons. In Case 590, the algorithms classified the worker as an employee, whereas in Case 195, the worker was classified as an independent contractor. These results may indicate the inability of ML to solve difficult cases involving unusual circumstances, especially when they involve novel policy issues such as classifying gig workers.⁶ In fact, in Case 195, which involved a limousine business where the workers could set their own hours, the results may have been skewed by the ambiguity of the relationship. On one hand, the workers did not own their tools, and the relationship was exclusive, which indicated employment. On the other hand, the drivers had a high degree of autonomy, which in this case apparently prevailed over the other factors. As in Case 590, while the facts indicated an independent contractor relationship, the court decided that the unusual disciplinary

⁶ Note that this issue is currently being debated in Canadian courts. In *CUPW v. Foodora Inc.*, the Ontario Labour Relations Board found that “Foodora couriers are dependent contractors and must be treated as such under the Act. As the evidence bears out, couriers more closely resemble employees than independent contractors.” In *Heller v. Uber Technologies Inc.*, 2021 ONSC 5518, Justice Perell at the Ontario Superior Court of Justice certified the \$400 million class action lawsuit where the workers submitted that they had been misclassified as independent contractors. The judge decided, in para 192, that there was at least some basis for concluding that some or all the drivers may have been misclassified.

power exercised by the hiring person reflected a high degree of supervision and, thus, an employment relationship. The notion of discipline is not new to lawyers, but it was not included in the Sagaz criteria and thus was not taken into account in our prediction algorithms.

3. Finally, the third group included cases that deviated from the norm because the judges were influenced by extra-judicial variables. These cases could be described as “reflexive” situations insofar as the outcomes were context dependent, reflecting unique societal circumstances. These cases were difficult to predict because they did not have strong algorithmic features. In fact, many of these cases had unique factual features that had influenced the judges to adjust the application of the legal test to the specific circumstances of the focal case. For instance, we found several cases (including *Beaudoin et 9154–4635 Quebec Inc., Re and Wilcom Systems Ltd. v. Minister of National Revenue*) where the judge disregarded the particular nature of the work relationship (which was typical of an employment relationship) based on the fact that it had been the worker’s choice to be a contractor. These are quite unusual decisions in Canada, because the worker’s intent is in principle non-determinative (in contrast to Californian cases where the worker’s intention is explicitly part of the Borello test). In other cases, we noted that the judges had tweaked the test and argued that some factors (which were usually relevant) were non-determinative because of the industry or the nature of the work. For instance, in Decision No. 598/12, the AI algorithm classified the worker as an employee, but the court decided upon a contractor status even though most features of the case indicated employment (supervision, exclusivity, and delegation of tasks). The judge found that these elements were not determinative and claimed that they were a product of statutory requirements or industry practice. In the same vein, the case *Morris Meadows Country Holidays and Seminars Ltd. v. Minister of National Revenue* considered the situation of a worker hired as a municipality treasurer and was later appointed Chief Administrative Officer. Many of the features of the relationship indicated the existence of a contractor relationship, which led the AI models to classify this worker as a contractor. However, the judge argued that many of the factors (which usually indicate a contractor relationship)—including the ability to hire, setting work hours, and delegation of tasks—were not determinative in this case because of the management nature of the position.

2.4 AI-powered legal aid: <https://MyOpenCourt.org/>

In addition to deciphering legal decision-making, algorithms seem to be able to provide a broader access to justice by delivering personalized predictions and identifying similar historical situations. To harness this potential, we deployed our predictive algorithms as a direct-to-public (DTP) legal help platform at <https://MyOpenCourt.org/>. This initiative was in response to various crucial needs in the current economy. We next discuss the deployment and implications generated from this real-world field implementation.

2.4.1 Access to justice

Disputes between employers and workers regarding employment status have been increasing over the past years. As discussed, gig economy firms have faced at least 40 major legal challenges around the world. These cases have been brought by gig economy workers seeking to access basic rights, such as minimum wages and paid sick days. Recent evidence showed that national courts fined Uber Eats, Glovo, Just Eat, and Deliveroo € 733 m for misclassifying 60,000 couriers. In addition, the COVID-19 pandemic has exacerbated labor market frictions between workers and employers, presumably due to incentives misalignment between the contracting parties (Smith 2020).

However, recent evidence has also shown that many claims were not successful and that gig firms used their resources to fight back (Moyer-Lee and Contouris 2021). In California, Uber and Lyft spent more than 200 million USD on a referendum campaign to exempt their businesses from the newly introduced AB5 law under which their drivers were categorized as employees. Furthermore, as noted in the pending Canada Supreme Court case, *Uber Technologies Inc. v. Heller*, a classic tactic includes mandatory arbitration clauses that consist of forcing those with a grievance to pay costly administration and filing fees in preliminary stages of the litigation.

The misclassification of workers is not a new problem that emerged from the proliferation of gig firms. It rose to the forefront of the public debate because of the modern magnitude and transnational nature of the problem. Moreover, those who are (mis)classified as independent contractors cannot receive benefits such as minimum wage, overtime pay, and medical leave. Many of these workers happen to be in precarious financial situations and are often unable to hire a lawyer for legal advice. In fact, it was observed that the number of people forced to represent themselves has ballooned over the last twenty years (Fragomeni et al. 2020), and the COVID-19 pandemic has exacerbated the problem of the lack of access to justice (Dahan and Liang 2020). This issue affects not only low-income populations; today, many middle-class individuals find themselves financially unable to benefit from the justice system. While there are many possible solutions for increasing the access to justice, one alternative is through legal aid technology. In particular, a new generation of DTP AI tools have emerged with the goal of providing basic legal help. Open AI technology has the potential to determine whether someone has a legitimate legal claim and to help litigants for whom litigation is otherwise out of reach.

2.4.2 <https://MyOpenCourt.org/>

Our platform, which is currently intended for the Canadian market and based on AI models calibrated using Canadian court cases, offers various tools in addition to the worker classification algorithm. In particular, it includes a tool for estimating severance packages, which we recently developed (Dahan et al. 2020). Launched in June 2020, the platform has already been accessed by 30,000 unique users. This platform includes the AI model presented in this paper (as well as other tools) not only to serve the public but also to test our results and collect new data that extend beyond

Fig. 3 Example of one question on <https://MyOpenCourt.org/>



Employee or Contractor Classification Tool

Question 5 of 18

Can you hire other people to work with/for you?

No
 Yes, with the approval of my supervisor
 Yes, without approval
 Does not apply

Back Next

court cases. Our tool is intended for both workers and employers, with the goal of providing a quick first assessment of the status of the law and a personalized assessment of the situation at no cost, thus helping democratize the access to justice.

In practice, the classification tool gathers information about the focal situation by asking a series of multiple-choice questions that correspond to the features used by the AI model. It takes less than five minutes for a user to complete all the questions and receive an assessment of the situation (see Fig. 3 for an illustration of one of the questions asked). We also collect demographics data (e.g., age and gender) for statistical purposes. These variables are not used by the AI model, and the data collection is fully anonymized.

Because of comparable accuracy of the different algorithms we considered, we chose to use a decision tree in the platform. This choice allowed us to provide users with clear explanations about the main aspects of the work relationship that determined the prediction outcome. These explanations were generated through a dictionary that expressed in simple language the reason for each split of the tree associated with the focal situation. Moreover, we leveraged the richness of our dataset to provide users with a list of the most relevant past court cases. More specifically, among the cases in the decision tree, we identified those that were closest in terms of industry and duration of work (see Fig. 4 for an example of the tool's output). Based on this information, we further developed an application that automatically generates a letter draft. For this purpose, we wrote a template letter that was automatically populated using the explanations behind the decision, and cited the most relevant case law that was identified. This letter is generated free of charge at the request of the user.


2.4.3 MyOpenCourt users and data

Between June 2020 and June 2021, 322 people used the Employee or Contractor tool to make a prediction (i.e., 322 unique visitors completed all the questions and received a prediction outcome). The following is a brief summary of these users.

- 84.5% of the users were workers seeking their own classification, while 15.5% were employers.

Fig. 4 Example of the output from <https://MyOpenCourt.org/>

MyOpenCourt predicts:

 **You are a contractor**

What this means for you

As a contractor you are not entitled to health benefits, minimum wage, paid time off, or overtime pay. The employer usually does not withhold EI or taxes and the contractor can work for others.

Disclaimer

Your prediction is not legal advice. Instead, it is an 85% accurate AI generated answer based on relevant Canadian case law.

Why our tool made this prediction

The main elements of your inputs that led the algorithm to make this prediction are:

- Although your work is closely supervised, you have the ability to decide what tasks are to be done.
- You own most of the tools essential to your job.
- Since 2010, 25 cases presented in court had exactly the same profile. Among these cases, judges decided in 76.0% of the cases that the worker was an independent contractor.

Related Cases

[Margaret Ann Keast and Karen H. McNeill o/a Toucan Communications v Ange Irafasha](#)

The worker was determined to be a contractor. The case took place in Ontario in 2017 and was decided by the Ontario Labour Relations Board.

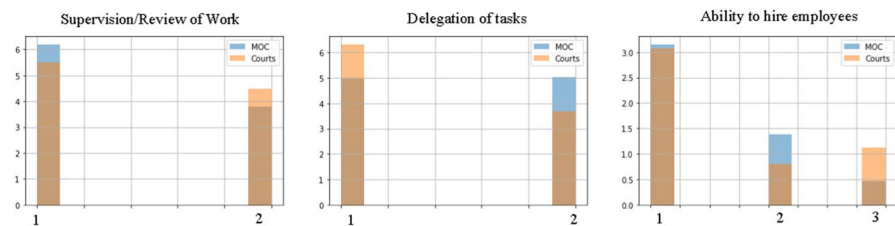


Fig. 5 Comparison of the distributions of three features (supervision/review of work, delegation of tasks, and ability to hire employees) between the Canadian courts dataset and the data from <https://MyOpenCourt.org/> (MOC)

- Among the users who provided gender identification (92.8% of the total users), 56.3% identified themselves as male.⁷
- The majority of the users (72.4%) inquired about the situation of their current job (as opposed to a past job).
- The prediction algorithm determined that the worker's status was employee in 70.8% of cases.

Comparing the data we collected from our public platform relative to the court cases offers an opportunity to examine whether court cases differ significantly from generic situations settled outside the judicial system. Investigating the distribution of features for both datasets (Canadian court cases and <https://MyOpenCourt.org/>) did not reveal any strong bias in the distribution of the relevant features, as we can see in Fig. 5.

⁷ The answer options to the question "What is your gender?" were man, woman, non-binary, other, and prefer not to disclose.

In summary, our study did not end with the development of an AI model with a strong prediction accuracy for workers' status. We further deployed our model in the field by offering an open-access platform that can help users predict the outcome of their own situations, along with clear, tailored explanations. We hope that our platform will help a large number of individuals and, ultimately, democratize the access to justice and to legal advice.

3 Discussion

The classic legal dispute of “employee vs. independent contractor” has become an important issue in many parts of the world, partly because of the ubiquity of the gig economy. When such disputes arise, workers are often asymmetrically affected because of the resources and time required for resolution. In some cases, these disputes end up in court, consuming scarce public resources.

One possible way of addressing this asymmetry is the development of AI-based tools to automate and understand the reasoning that underlies legal outcomes. In this work, our ultimate achievement was the deployment of an open-access direct-to-public legal help tool, <https://MyOpenCourt.org/>, to provide a modicum of legal assistance for workers and small companies. Such open-access tools may have great implications for the access to justice insofar as many individuals have very limited access to legal services (Macfarlane and Sullivan 2021). From a judicial standpoint, advanced AI tools may also help courts to screen and process cases more efficiently. In the criminal justice arena, this approach has already been taken by several countries. In the context of civil cases, the academic literature has shown that AI models can capture legal reasoning and accurately classify cases. Nevertheless, the application of AI models to employment law cases, such as the classification of workers, has received limited attention. As part of an attempt to bridge this gap, this paper has offered a data-driven study on the implications of using an algorithmic model to classify workers' employment status. We trained our model using all previous cases decided by Canadian and Californian courts. We then presented the potential implications for judicial decision-making and discussed the field deployment of our AI model via <https://MyOpenCourt.org/>.

More specifically, we evaluated the feasibility of developing an AI model to determine whether a worker is an employee or an independent contractor. To do so, we collected all relevant Canadian and Californian court cases between 2002 and 2021, and we collaborated with a team of lawyers to carefully annotate and pre-process the data by generating the features needed as input to the AI model. We found that supervised ML models, even simple off-the-shelf methods, could consistently classify court cases at an out-of-sample accuracy of 82–96%. These findings partially shed light on the debate surrounding the computability of law (Deakin and Markou 2020). Skeptics have argued that certain features of legal reasoning are not consistent with ML, particularly the reflexivity of legal knowledge and the incompleteness of legal rules (Markou and Deakin 2020).

This study provided strong evidence that some legal questions can be supported by AI models. First, while we argued that many aspects of legal reasoning

have algorithmic features that lend themselves to automation, there are still limitations to computability, notably if certain conditions are not met. Second, we argued that algorithms are reflexive, and thus can capture, or at least identify, the reflexive features of the law. In the same way that the goals and priorities pursued by lawyers and judges are important, the goals and priorities that are incorporated in the development of algorithms matter. Therefore, in the legal context, AI could play a powerful role in assisting decision makers and workers, provided that the algorithm is carefully crafted to respect the reflexive nature of the law. Our study empirically tested the computability of employment status determination, a foundational issue in many areas of law, including tax, employment, and tort law. With regard to this specific issue, it has been argued that while the classification test used by courts is essentially algorithmic (because it follows a specific list of rules and criteria), there is no easy solution to classifying a new case. For example, when a new type of social or economic category is presented to the court for classification (e.g., a gig worker), judges often propose a number of alternative solutions that are tested against each other (Markou and Deakin 2020).

While the critiques of ML in law have merits, the current study attempted to challenge the assertions. We argue that legal reasoning is not always forward looking and that legal ML is not always backward looking. More specifically, our study provides empirical evidence that novel situations are not as frequent as one might think. More specifically, our models show that a large number of cases present strikingly similar facts. In addition, we show that ML can help identify new facts along with the emergence of a new precedent. Similar to ML, judges and lawyers also apply “past” facts to future cases. Finally, even when confronted with novel facts, such as the context of the gig economy, a well-crafted reflexive AI can act as a detector of novel facts by spotting unconventional cases and atypical situations. A non-adjudicating ML is not designed to predict every case; it is designed to evolve and to be retrained regularly with new data. Unfortunately, it is hard to come up with a clear segmentation of when ML techniques should be used and when they should not. That being said, ML techniques often work well for fact-intensive questions of law that generates a compact universe of cases and a list of factors that are most relevant to the court when deciding the outcome like in this paper (see Alarie et al. (2018)). Also, there is an emerging consensus in the legal community that ML should not be used for adjudication (although already being used in some countries, as discussed in Markou and Deakin (2020)). The same applies to areas of law that are fast changing, to cases that involve a high level of subjectivity, and to cases where judges are prone to biases (e.g., family or criminal cases, see Park (2020)).

By combining two different datasets (Canadian and Californian court cases), we leveraged AI algorithms as a tool for use in comparative law. By using algorithms trained on one dataset to predict outcomes of the other dataset, we found surprisingly strong consistency between two vastly distinct jurisdictions. In addition, we observed a remarkable consistency in judicial decisions among several AI algorithms. We ultimately chose to use a simple decision tree that provided a highly desirable interpretability dimension to explain the rationale for the prediction. We also carefully examined missclassified cases using our AI algorithms to gain insights

into adaptive judicial decisions, and we proposed the division of misclassified cases into three categories.

Finally, we conclude our paper by discussing the deployment of our AI model via an open-access platform called <https://MyOpenCourt.org/>. This platform aims to serve the public and help self-litigants answer questions about employment law. During the first 15 months, 322 people used the Employer or Contractor tool offered in this platform (and more than 20,000 people used the platform, which includes several legal aid tools) to predict whether they would be considered employees or independent contractors. We believe that this type of platforms can help the community and democratize the access to legal advice and to justice.

Our research contributes to the literature and the practice of using AI in legal contexts. We are among the first to discuss the potential implications of leveraging AI for civil employment cases, and we hope that this study will stimulate further discussion on this topic.

4 Material and methods

4.1 Data collection

4.1.1 Cases collection

All Canadian cases in the Westlaw database that referred explicitly to the Sagaz case were collected,⁸ leading to a total of 1,225 cases. Among these cases, a significant number were not related to the determination of workers' status. After selecting only the relevant cases, we retained 538 cases. Similarly, Californian data were also obtained from the Westlaw database by collecting all cases citing the Borello case. Between 1989 to 2021, we found 810 such cases. After filtering out irrelevant cases that were not related to the employee vs. contractor question, the dataset was reduced to 217 relevant cases.

4.1.2 Feature identification

Important aspects of developing an effective dataset are to identify a set of prominent quantifiable factors and design objective scales that permit the reduction of any given case to a set of standardized descriptors. To determine the relevant variables to be extracted from the cases, we thoroughly considered the factors that were brought up in the Sagaz seminal case (Canada) and in the Borello case (California).

The Borello case and its associated Borello test cover a variety of factors that include the six factors outlined in the Sagaz case. Thus, these factors formed a strong basis for our comparative analysis. They included the following: level of control of the employer, equipment property, ability to hire helpers, financial risk, opportunity

⁸ <https://signon.thomsonreuters.com>

Table 1 Summary of the observed variables in the Canadian dataset

	Type	Min	Max	Median	Observations
Industry	Categorical	1	4	2	538
Supervision/review of work	Categorical	1	2	1	445
Delegation of tasks	Categorical	1	2	1	481
Where the work is performed	Categorical	1	2	1	447
Is the worker required to wear a uniform?	Categorical	1	2	2	193
Length of service	Continuous	0	38	1	505
Ownership of tools	Ordinal	1	5	2	487
Ability to hire employees	Ordinal	1	3	1	354
Chance of profit	Ordinal	1	6	1	510
Risk of loss	Ordinal	1	3	1	487
Exclusivity of services	Ordinal	1	3	1	360
Who sets the work hours	Ordinal	1	4	2	436
Outcome	Categorical	1	2	1	538

for profit, and responsibility for investment. These common factors were estimated on a common scale for both the Canadian and the Californian cases.

To quantify each of these factors, we based our scales on a legal analysis of how judges interpreted and appreciated these factors. This led us to narrow each subjective criterion into a set of graded scales. For instance, the “worker’s opportunity for profit in the performance of his or her tasks” brought up in the Sagaz case, which could formally cover many aspects of a worker’s experience, was generally estimated by the judges by considering the way that the worker was paid. Workers paid via fixed wages have no opportunity for profit, whereas workers who receive commissions have a clear opportunity for profit. We thus designed a four-point scale ranging from receiving only wages to being paid only on commissions to quantify the opportunity for profit of a worker, including intermediate cases depending on the proportion of fixed salary. We applied the same methodology to all six Sagaz criteria, which allowed us to make this subjective and multifarious dimension more quantifiable, as well as ensure consistency across legal researchers.

In addition to the variables related to worker classification, we also collected other factual data on each case, namely, the year, industry, length of service, province (in Canadian cases), and the court, as well as other indicators that were present in some cases, including the exclusivity of services or whether the worker was required to wear a uniform. Altogether, 12 common factors were collected in both databases and used for prediction (see Table 1 for Canadian cases and Table 2 for Californian cases). These variables were of three types: (i) categorical variables, such as the industry sector (1: finance, 2: services, 3: manufacture, 4: agriculture), (ii) binary information about the case (e.g., whether or not the hirer supervised the tasks of the worker or provided performance reviews), and (iii) ordinal variables that were used to evaluate more finely some factors, such as the opportunity for profit discussed above, or, for instance, who sets the work hours, which were split into four levels from the highest level of control (1: hirer sets the number of hours and when

Table 2 Summary of the observed variables in the Californian dataset

	Type	Min	Max	Median	Observations
Outcome	Categorical	1	2	1	217
Supervision/review of work	Categorical	1	2	1	130
Ability to hire employees	Categorical	1	3	2	91
Exclusivity of services	Categorical	1	3	3	110
Ownership of tools	Continuous	1	5	3	175
Is the worker required to wear a uniform?	Continuous	1	2	1	84
Who sets the work hours	Continuous	1	4	1	157
Delegation of tasks	Continuous	1	2	1	188
Chance of profit	Continuous	1	5	3	185
Risk of loss	Continuous	1	3	2	123
Where the work is performed	Continuous	1	2	1	166
Length of service	Categorical	0	25	1	76

the work is done) to the lowest level of control (4: worker sets the number of hours and when the work is done), through intermediate levels (2: hirer sets the number of hours, but not when the work is done, or 3: worker sets the number of hours but not when the work is done). After defining and scaling all the variables, we began the annotation process.

4.1.3 Features and scales

The following is a summary of the features and scales used in this study. The *Industry* variable was labeled as follows: 1: finance, 2: services, 3: manufacturing, and 4: agriculture. *Length of service* was rounded to the closest duration in years. *Delegation of tasks* referred to the level of management of the worker: 1: Hirer delegates most tasks to the worker, and 2: Worker decides how to do most or all the tasks and what tasks need to be completed. *Who sets the work hours* was based on a four-point scale: 1: Hirer sets the number of hours and when the work is done, 2: Hirer sets the number of hours but not when the work is done, 3: Worker sets the number of hours but not when the work is done, and 4: Worker sets the number of hours and when the work is done. *Supervision of work* assessed whether the hirer supervises the tasks of the worker and/or gives performance reviews (1: Yes, 2: No). *Ability for the worker to hire employees* was based on a three-point scale: 1: No, 2: Yes with the hirer's approval (or allowed to hire but does not in practice), and 3: Yes on their own. *Ownership of tools* was scored on a five-point scale: 1: Hirer owns or provides all the tools, 2: Hirer owns or provides tools that are essential to performing the job, 3: Hirer and worker both provide a substantial amount of tools, 4: Worker owns or provides tools that are essential to performing job, and 5: Worker owns or provides all the tools. *Chance of profit* referred to how the worker is paid: 1: Only fixed wages (i.e., paid hourly or per task), 2: Mostly fixed wages with a chance of commission, 3: Mix of a wage and a commission, where it is unclear which is predominant, 4: Mostly commission, with a chance of other incentives, 5: Only commission, with

a chance of incentives, and 6: Only commission, without bonuses (no possibility to earn additional compensation). *Risk of loss* was scored on a three-point scale: 1: No risk of loss for the worker, 2: Some risk of loss for the worker (pay for some business expenses), and 3: High degree of risk of loss for the worker (pay for business expenses/have money invested/liable for losses). *Exclusivity of services* referred to the amount of work done for the hirer: 1: Worker works exclusively for the hirer, 2: Worker does some work for this hirer as well as for others (or for themselves) or worker is allowed to work for other hirers, and 3: Worker does the majority of work for other hirers or for themselves. *Where the work is performed* referred to whether the work premises were provided or instructed by the hirer: 1: Yes, and 2: No. Finally, *Is the worker required to wear a uniform*: 1: Yes, and 2: No.

4.1.4 Data collection

Our databases were collected by a team of caseworkers trained in law. First, to ensure consistency, we trained caseworkers in Sagaz and its various criteria, and they completed exercises on a few common sample cases to assess the consistency of data collection variables. The 448 Canadian cases were shared among the caseworkers, who read each case thoroughly and answered a series of questions that were considered when determining the employer's level of control based on the scales determined above. For each feature annotated, the caseworker was given a set of multiple-choice questions, including the option of not answering if the case did not consider or report any indication on a specific factor. The caseworker answered each question by reading the facts of the case and the judge's legal analysis.

4.1.5 Missing data

As shown in Tables 1 and 2, many cases contained missing information. There were missing values either because a specific criterion was not applicable to the situation, or it was irrelevant to the judgement. The presence of missing values may have been an issue because they could have potentially correlated with the decision of the judge who wrote the case. To test whether missing data were correlated with the outcome, we constructed a binary database of missing values, where zeros indicated that information was present in the original database and one indicated when information was missing. We computed the correlation matrix of this database, focusing on correlations with the outcome. We found very low correlation coefficients, all of which were below 0.09 in absolute value in the Canadian dataset and were slightly higher in the U.S. dataset regarding three features (21% with delegation of tasks, 15% with the ownership of tools and where the work is performed, and 12% with exclusivity of services). As expected, we also observed strong correlations between different missing data fields (e.g., missing information about the risk of loss showed high correlations with the absence of data on chance of profit or ownership of tools) (Fig. 6).

To mitigate the impact of missing data on the prediction outcome, we used three different pre-processing approaches. We obtained similar results for all three approaches, including imputing missing data by the median value of the entire

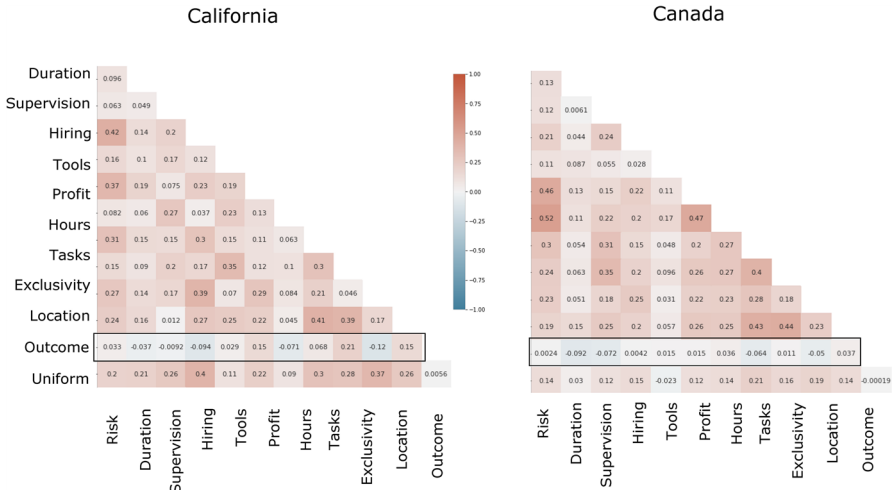


Fig. 6 Correlation matrix of the missing data including the outcome. We found very low correlations in the Canadian dataset, and low, yet slightly higher, correlations in the Californian dataset

dataset, as well as relying on more sophisticated methods where the missing data were coded as new categorical features, whereas the missing numerical variables were iteratively imputed (Van Buuren and Groothuis-Oudshoorn 2011).

In the main text of this paper, we report our results for a dataset in which the missing data were replaced by a predicted value based on the remaining features, using a random forest predictor, with 15 trees of depth less than three. Our findings showed that all methodologies provided similar accuracy levels and managerial insights.

4.2 Algorithmic worker classification

4.2.1 Naive data analyses

We began the analyses by computing the correlation matrix of the dataset, as shown in Fig. 1 in the “Results” section. This figure indicates the difficulty in identifying the different factors, given that high correlations exist between the features. Moreover, the correlation analysis led us to identify a strong correlation between the outcome variable and some features, which we examined more carefully by plotting several bi-variate distributions.

4.2.2 Absence of clear clusters

Based on our observations of strong correlations between factors as well between some factors and the outcome, we investigated whether the cases labeled employees were *organically different* from the cases classified as independent contractors. We relied on clustering algorithms to identify the potential existence of

Fig. 7 Samples in two-dimensional PCA space with k -means clustering for the Canadian dataset (two clusters)

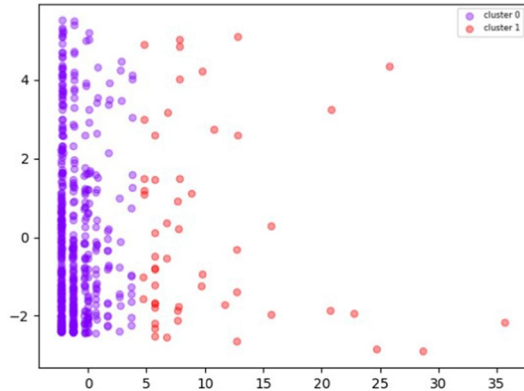


Table 3 Adjusted rand index between clusters and outcome and silhouette score

Model	Adj. rand index	Silhouette score
k -means	-0.010	0.583
k -means	-0.014	0.466
Affinity propagation	0.063	0.144
Mean shift	-0.008	0.521
Spectral clustering	-0.006	0.778
Spectral clustering	0.000	0.058

clusters in the dataset, and then we considered whether these clusters explained the employer/contractor division.

The results of the k -means clustering algorithm (Ahmad and Dey 2007) with two clusters are depicted in Fig. 7, where the axes are the two principal components from the principal component analysis (PCA). For illustration purposes, we focused our clustering analyses on the Canadian dataset (which included a larger number of observations). Similar qualitative conclusions were drawn from the results of the analysis of the Californian dataset.

Although the two clusters in Fig. 7 may visually appear to be well separated, they scored quite low in the classical measures of cohesion and separation of clusters (e.g., a silhouette score of 0.226), indicating that the two clusters were not well separated. This conclusion was confirmed using alternative clustering algorithms, including affinity propagation, the mean shift method, and spectral clustering, all of which yielded moderate to low scores (see Table 3). Thus, these results suggested that there was no clear separation of the categories within the dataset.

We also found that the goodness-of-fit of clusters did not necessarily indicate that the clusters were aligned with the outcomes (i.e., that one cluster represented the “employee” cases and the other represented the “independent contractor” cases). To address this discrepancy, we computed the adjusted rand index (ARI) for each clustering method (Rand 1971). The k -means algorithm with $k = 2$ yielded the highest ARI (0.666). Given that the outcome was binary (and thus a random guess translated

Table 4 Prediction accuracy of supervised models. Accuracy is evaluated by the average of a ten-fold cross validation, for various combinations of training/test sets

Model	Can/Can	Cali/Cali	Can/Cali	Cali/Can	All/Can	All/Cali
Logistic regression	0.927	0.883	0.811	0.914	0.915	0.811
Random forest classifier	0.896	0.860	0.816	0.898	0.909	0.802
K-Neighbors classifier	0.852	0.827	0.788	0.884	0.885	0.788
SVC	0.904	0.813	0.805	0.909	0.909	0.806
Gaussian process classifier	0.919	0.810	0.792	0.911	0.911	0.792
AdaBoost classifier	0.860	0.770	0.810	0.911	0.911	0.811
XGB classifier	0.880	0.847	0.778	0.904	0.904	0.778

Note: Can for Canada, Cali for California, and All for the combined dataset

Table 5 Prediction accuracy, F1 score, and AUC for all algorithms trained and tested on the Canadian dataset. Accuracy was evaluated by the average of a three-fold cross validation

Model	Accuracy	F1	AUC	3-,Accuracy	3-,F1	3-,AUC
Logistic regression	0.919	0.872	0.917	0.957	0.965	0.951
Random forest classifier	0.904	0.846	0.924	0.963	0.968	0.956
K-Neighbors classifier	0.830	0.712	0.881	0.909	0.928	0.897
SVC	0.926	0.885	0.918	0.959	0.967	0.953
Gaussian process classifier	0.919	0.872	0.916	0.954	0.962	0.947
AdaBoost classifier	0.889	0.831	0.926	0.957	0.965	0.956
XGB classifier	0.933	0.895	0.926	0.946	0.956	0.939

Note: We compare the entire dataset (538 cases) with learning and testing on data with no more than three missing variables (labeled 3-, 466 cases)

into 0.5), an ARI of 0.666 indicated a modest match between the clusters and the actual outcome.

The results reported in this section strongly suggest that unsupervised learning algorithms are not capable of generating a satisfactory performance in terms of classifying employees versus independent contractors. In other words, a strong overlap in features was found in both classes.

4.2.3 Supervised learning models

Our analysis of the judicial reasoning, which was reported in the “[Results](#)” section of this paper, was performed using a decision tree. Specifically, we used the scikit learn function, *DecisionTreeClassifier*, with a maximal depth of three, the greedy algorithm (Gini index), and a minimum number of samples per split or per leaf equal to ten to avoid overfitting.

Several alternative algorithms were used to confirm the results and to examine whether significantly better outcomes would emerge when more sophisticated methods were applied. In particular, we considered a logistic regression, random forest

Table 6 CV score when varying the number of folds

Folds	2	3	5	10
Training samples	269	358	430	484
Logistic regression	0.916	0.926	0.929	0.922
Random forest classifier	0.929	0.937	0.926	0.931
<i>K</i> -Neighbors classifier	0.898	0.896	0.888	0.889
SVC	0.920	0.926	0.920	0.924
Gaussian process classifier	0.914	0.924	0.926	0.917
AdaBoost classifier	0.918	0.931	0.914	0.926
XGB classifier	0.911	0.931	0.931	0.931

(with 15 estimators and a maximal depth of four), *K*-nearest neighbor, support vector classifier (both a linear kernel and with probability), Gaussian process algorithm, Adaboost, and XGBoost. All models were evaluated based on a three-fold cross validation, and all hyperparameters were set at their default value based on Python's sklearn package. Table 4 reports the prediction accuracies of various training/test combinations for the seven algorithms that we considered. Table 5 shows the Canadian dataset and presents the accuracy, *F1* scores, and area under the curve (AUC) for the entire dataset and when the samples with more than three missing variables were dropped. Interestingly, removing these samples improved the accuracy of all the models we considered. In this setting, the accuracy exceeded 90% for all models, and we even reached an out-of-sample prediction accuracy of 96% in the random forest model.

The results shown in Table 4 indicate that the performance of the supervised ML models for the Can/Can and Cali/Cali combinations ranged between 0.81 and 0.927, where the logistic regression for the Can/Can setting yielded the highest performance (92.7%). Table 6 shows the variation in model performance with respect to the number of folds included in the cross-validation procedure, which provided insights into the effect of varying the sample size on the prediction performance. We found that a linear model, such as logistic regression, performed better when more observations became available for testing purposes. The variations in the performance of ensemble-based models (e.g., random forest, AdaBoost, XGBoost) appeared to be consistent across different values of the folds. Another observation worth mentioning is that the performance of the linear model was reduced relative to the non-linear models when more data were available for testing purposes.

The feature importance associated with the random forest method are reported in the “Results” section. They appeared to be consistent with other estimates of feature importance that were extracted from those algorithms. For instance, the estimated coefficients of the logistic regression for re-normalized data with one-hot encoding using the Canadian dataset, reported in Table 7, allowed us to identify exclusivity of services, delegation of tasks, and supervision of work as the most important features. Another view of feature importance can be obtained through the Shapley values (Lundberg and Lee 2017) of the variables used by the XGBoost algorithm (see Fig. 8 for the same dataset). This metric has been used in recent literature to

Table 7 Estimated coefficients of logistic regression using one-hot encoding of binary variables (indicated by the presence of a number after the feature label)

	Coefficient
Supervision/review of work_1	-0.754
Supervision/review of work_2	0.594
Delegation of tasks_1	-1.194
Delegation of tasks_2	1.476
Where the work is performed_1	-0.106
Where the work is performed_2	0.076
Is the worker required to wear a uniform?_1	-0.119
Is the worker required to wear a uniform?_2	0.311
Length of service	-0.105
Ownership of tools	0.582
Ability to hire employees	0.433
Chance of profit	0.235
Risk of loss	1.308
Exclusivity of services	2.493
Who sets the work hours	0.379

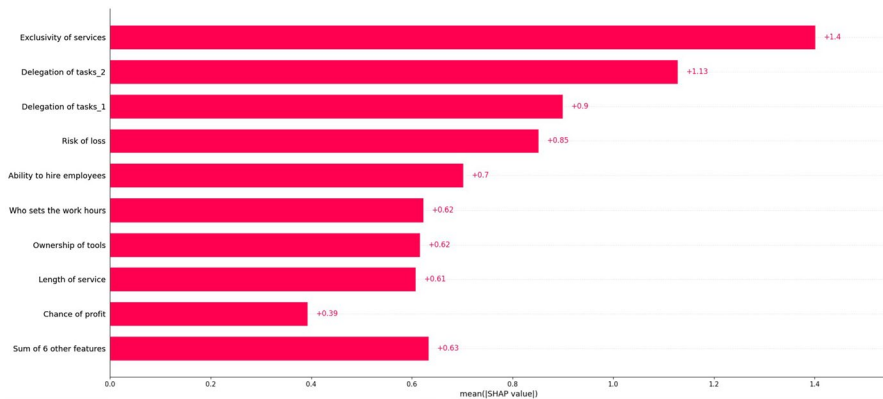


Fig. 8 SHAP importance of features for the XGBoost classifier

examine the importance of features in predictive analytics (e.g., Adulyasak et al. 2020; Satopää et al. 2021). The Shapley values in our case revealed a similar ranking of feature importance, with exclusivity of services ranking first, followed by delegation of tasks and risk of loss. Interestingly, exclusivity of services was not among the top features in the random forest feature importance analysis.

4.2.4 Analysis of misclassified cases

We combined the results of all the algorithms and determined that a large number of misclassified cases were misclassified by most—if not all—algorithms. We extracted those cases and systematically analyzed the decision and details of each

case in order to identify the source of the misclassification outcome. Our results (see the “[Results](#)” section for all the details) indicated that those cases had either (i) very unusual factors or procedural circumstances, (ii) features related to both employee and contractor relationships, or (iii) decisions that were influenced by extra-judicial variables.

Acknowledgements We would like to thank the team at the Conflict Analytics Lab and the Scotiabank Centre for Customer Analytics at Queen’s University, particularly our JD students, who compiled and structured the data (Holly Grosdanis, Solinne Jung, and Brian Tsui). We are also grateful to Brandon Loehle and Max Saunders, who supported this project for almost two years. Finally, we thank Art Cockfield, Yuri Levin, David Saunders, Stephen Thomas, Kevin Banks, Joshua Karton, Dan Sfedj, and Dean Mark Walter for their support.

Funding The Project is funded by Social Sciences and Humanities Research Council of Canada, Grant no. New Frontier.

References

- Adulyasak Y, Benomar O, Chaouachi A, Cohen M, Khern-am-nuai W (2020) Data analytics to detect panic buying and improve products distribution amid pandemic. Available at SSRN 3742121
- Ahmad A, Dey L (2007) A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl Eng* 63(2):503–527
- Alarie B (2016) The path of the law: towards legal singularity. *Univ Toronto Law J* 66(4):443–455
- Alarie B, Niblett A, Yoon AH (2018) How artificial intelligence will affect the practice of law. *Univ Toronto Law J* 68(s1):106–124
- Allon G, Cohen M, Sinchaisri WP (2018) The impact of behavioral and economic drivers on gig economy workers. Available at SSRN 3274628
- Babar Y, Burtch G (2020) Examining the heterogeneous impact of ride-hailing services on public transit use. *Inf Syst Res* 31(3):820–834
- Burosu SH(2021) UBER drivers: employee, worker or independent contractor? <https://www.lexology.com/library/detail.aspx?g=cc848bb3-b06d-404c-ad29-80846a4f892a>
- Chen DL, Eigel J (2017) Can machine learning help predict the outcome of asylum adjudications? In: Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, pp 237– 240
- Cohen MC, Dahan S, Rule C (2021) Conflict analytics: when data science meets dispute resolution. Forthcoming in *Management Business Review*
- Dahan S, Liang D (2020) The case for AI-powered legal aid. *Queen’s LJ* 46:415
- Dahan S, Touboul J, Lam J, Sfedj D (2020) Predicting employment notice period with machine learning: promises and limitations. *McGill Law J/Revue de droit de McGill* 65(4):711–753
- Deakin S (2020) Decoding employment status. *King’s Law J* 31(2):180–193
- Deakin S, Markou C (2020) *Is law computable? Critical perspectives on law and artificial intelligence*. Bloomsbury Publishing Plc, London
- Deakin S, et al (2005) *The comparative evolution of the employment relationship*. CiteSeer
- Dunn M, Sagun L, Şirin H, Chen D (2017) Early predictability of asylum court decisions. In: Proceedings of the 16th Edition of the international conference on artificial intelligence and law, pp 233– 236
- Fragomeni B, Scarrow K, MacFarlane J (2020) Tracking the trends of the self-represented litigant phenomenon: Data from the national self-represented litigants project, 2018/2019. Technical report, National Self-Represented Litigants Project
- Greenwood B, Burtch G, Carnahan S (2017) Unknowns of the gig-economy. *Commun ACM* 60(7):27–29
- Huang N, Burtch G, Hong Y, Pavlou PA (2020) Unemployment and worker participation in the gig economy: evidence from an online labor market. *Inf Syst Res* 31(2):431–448
- Kauffman ME, Soares MN (2020) AI in legal services: new trends in AI-enabled legal services. *SOCA* 14:223–226

- Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Curran Associates Inc, Red Hook
- Macfarlane J, Sullivan C (2021) Tracking the trends of the self-represented litigant phenomenon: data from the National Self-represented Litigants Project, 2019–2021. Working paper
- Marius M (2021) The gig economy: Global, regional, local realities. *Intell Econ Devel Caribbean*, 25
- Markou C, Deakin S (2020) Ex machina lex: exploring the limits of legal computability. *Is Law Comput? Crit Perspect Law Artif Intell*, 31
- Moyer-Lee J, Contouris N (2021) Taken for a ride: Litigating the digital platform model. Technical report, The International Lawyers Assisting Workers (ILAW) Network
- Park J (2020) Your honor, AI. *Harv Int Rev* 41(2):46–48
- Pranaya D (2014) Job-hopping—an analytical review. *Int J Res Bus Manag* 2(4):67–72
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66(336):846–850
- Satopää VA, Salikhov M, Tetlock PE, Mellers B (2021) Bias, information, noise: the BIN model of forecasting. *Manag Sci* 67(12):7599–7618
- Shimao H, Khern-am-nuai W, Kannan K, Cohen MC (2022) Strategic best response fairness in fair machine learning. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, pp 664–664
- Smith R (2020) Independent contractors and COVID-19: working without protections. <https://www.nlp.org/publication/independent-contractors-covid-19-working-without-protections/>
- Taylor J, Deng XN, Joshi K (2020) Exploring the role of ais scholars in framing the concept of “contractor” and “employee” in the gig economy. In: AMCIS 2020 TREOs, p 4
- Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. *J Stat Softw* 45(1):1–67

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Maxime C. Cohen¹ · Samuel Dahan²  · Warut Khern-am-nuai¹ · Hajime Shimao² · Jonathan Touboul³

✉ Samuel Dahan
samuel.dahan@queensu.ca

Maxime C. Cohen
maxime.cohen@mcgill.ca

Warut Khern-am-nuai
warut.khern-am-nuai@mcgill.ca

Hajime Shimao
hajime.shimao@mcgill.ca

Jonathan Touboul
jtouboul@brandeis.edu

¹ Desautels Faculty of Management, McGill University, 1001 Rue Sherbrooke O., Montreal, QC H3A 1G5, Canada

² Faculty of Law, Queen's University, 128 Union St, Kingston, ON K7L 2P1, Canada

³ Department of Mathematics, Brandeis University, 415 South Street, Waltham, MA 02453, USA