Identifying Public Sector Demand for Third-Party Data: An NLP Framework for Lead Scoring in U.S. Job Postings

Rory G. Quinlan[1*]

[1*]Economics Department, University of Pittsburgh, Pittsburgh, PA, USA.

Corresponding author(s). E-mail(s): RoryQuinlan@pitt.edu;

Abstract

Identifying data demand at the federal level in the public sector remains a challenge for data vendors and researchers, partly because of limited access to procurement records. This study uses U.S. federal job postings to propose a novel framework for estimating external data demand and vendor leads. By combining structured features such as agency size, seniority, and industry, along with natural language processing (NLP) of job descriptions, titles, and key duties, we train a logistic regression model to identify positions likely to purchase external data. The model not only classifies likely data buyers, but also generates a Data Buyer Score, allowing vendors to prioritize leads based on their likelihood of purchasing data. To address class imbalances, we applied the synthetic minority oversampling technique (SMOTE) during training. The model was evaluated using stratified 5-fold cross-validation, achieving a mean ROC-AUC of 0.86 and a precision of 88.1%. This framework provides a method for estimating third-party data demand from job postings, enabling real-time market intelligence for data vendors and researchers. This framework offers a foundation for demand estimation in a variety of markets via job postings.

Keywords: Public Sector Analytics, Natural Language Processing, Third-Party Data Demand, Market Intelligence, Job Postings

1 Introduction

In this digital and AI era, the demand for external data has surged, as reliable data form the foundation of decision-making across all sectors and industries. To keep pace with digital transformation, public agencies require access to unbiased external data to make and evaluate important policy decisions across areas such as fraud detection, patient record matching, and sentiment analysis. However, despite this rapidly growing demand, it remains unclear who within the government is responsible for purchasing such data. This lack of transparency in the market poses a significant challenge for data vendors seeking to identify potential government clients and public sector researchers.

Currently, many data vendors rely on primitive targeting methods, such as relying on prior business relationships, keyword matching, or attempting to access limited procurement databases, to identify potential data buyers. These approaches often fail to capture the full market potential because of linguistic complexity and miss emerging demand. Job postings have an advantage in viewing emerging trends as they reflect upcoming initiatives and the sector's evolving digital and data needs. These job postings offer an underutilized and overlooked signal of third-party data demand.

This study proposes a replicable framework to process U.S. public sector job postings and identify positions likely involved in purchasing external data. The framework utilizes job-structured features, such as agency, role seniority, and industry, with unstructured text (e.g., job descriptions and titles) to assign a Data Buyer Score indicating the likelihood of interaction with external data vendors.

Previous studies, such as those by Xu et al. [1], Senger et al. [2], and Lukauskas et al. [3] built a foundation for natural language processing (NLP) applications to job postings for job classification, skill extraction, and labor market forecasting. This study builds on previous research by incorporating additional structured features from job postings, applying this information to estimate demand for third-party data, and integrating a scoring system to prioritize future client targeting and outreach.

2 Literature Review

Research via job posting analysis has become increasingly viable as large-scale job boards host thousands of positions, each of which is a rich source of business information and needs embedded within the fields of descriptions, titles, and key duties. Prior work has focused on classification, skill extraction, and labor market forecasting. However, few studies have explored the public sector‘s external purchasing demand.

Xu et al. [1] introduced the Job Classification Text Corpus (JCTC), a dataset of 10,000 Chinese job postings. Using a combination of supervised learning and manual review, each posting was classified according to the standard job taxonomy, the China Grand Classification of Occupations (CGCO). The labeled dataset was used as training data to create an NLP model capable of classifying jobs based on their descriptions. This work highlights the value of job descriptions for classification and market insights, and establishes a foundation for NLP-driven job posting analysis.

Senger et al. [2] comprehensively reviewed current and past research methods for skill extraction and classification from job postings. Their review demonstrates the evolution of job-posting analysis. It began with simple keyword matching, continued

to machine learning models, and is being studied using deep learning algorithms. However, these methods focus on classifying job skills to analyze and forecast labor market trends, and do not estimate purchasing demand.

A Recent work in Applied Sciences [3], Enhancing Skills Demand Understanding through Job Ad Analysis. Skill demand was examined to understand the labor market by applying the NLP. However, these methods have been applied to federal job advertisements to identify the desire for technical capabilities, such as how "data literacy" appears in job descriptions. This study demonstrated the technical requirements of the federal government. However, it focuses on internal labor needs rather than on external procurement demand.

While job-posting analysis has been applied to tasks such as classification and skill extraction, no existing framework has focused on identifying third-party data demands in the public sector. Moreover, existing NLP models rarely capture nuanced signals and distilling results in binary or categorical classifications instead of using a continuous probability range. This study addresses this gap by introducing a novel, interpretable framework that integrates structured and unstructured features from job postings to estimate the likelihood that a public sector role will act as a third-party data buyer.

3 Methods

3.1 Initial Data Collection

The dataset used in this study was constructed from scraped data using the USAJobs.gov public API. According to USAJobs (2025), the federal government is the largest employer in the U.S. USAJobs is known as the primary hiring site for the U.S. federal government, making it an ideal sampling frame for analyzing emerging data demand in the federal public sector. The dataset is a one-time scrape from April 2, 2025. The API was queried using a selected set of words in the platform's "Areas of interest" search field, the initial search filter for surface jobs associated with the user types. This set of words selected to surface job titles balances ensuring jobs with the potential for external data purchasing while maintaining the variation in job types in the dataset. A full list of the selected "Areas of interest" words is available in Appendix A. "Areas of interest" was tagged in each posting to capture how the job was retrieved.

The query yielded 5,829 job postings, each with a unique job ID to ensure non-duplication. Each job record includes structured fields, such as agency, job series, and location, as well as unstructured text, such as duties, qualifications, and job descriptions.

3.2 Buyer Labeling and Feature Engineering

A base set of positive cases—explicit third-party buyers—is required to train the model to detect likely buyers in the future. We identified "explicit buyers" by applying keywords and fuzzy matching to the job description and key duties text to generate this training set. Key phrases such as "data acquisition," "third-party data licensing," and "data procurement" were used to identify explicit data buyers. To reduce false negatives, fuzzy matching was used to capture linguistic alternatives such as "data

licensing" or "external data use" that may not appear in exact form from our keyword set but still indicate explicit data purchasing. (See Appendix B for the full set of matching terms.)

A random sample of 100 postings was manually reviewed to assess the labeling strategy. This review confirms the high accuracy of keyword-matching strategies in identifying explicit data buyers.

Additional features were engineered to support future modeling using keyword matching; these features include:

- Seniority, a binary variable equal to 0 for non-seniors and 1 for seniors, was determined by keyword matching of the job title text. Positions containing words such as "lead," "chief," "senior," or "director" were flagged as senior positions. (See Appendix C for the full mapping).
- Agency size is a categorical variable based on the agency scale with three possible values: large, medium, and small. This feature was created collaboratively with AI-assisted prompts, which use a known institutional scale to identify large agencies based on their number of employees. Large agencies included the Department of Defense, Department of Veterans Affairs, and Department of Treasury. Medium agencies included the Department of Transportation and Department of Commerce (See Appendix C for full mappings).
- Industry is a categorical variable that takes six values (finance, marketing, medical, security/tech, policy, and other). Industry was determined via keyword matching on agency, job titles, and "Areas of interest." (See Appendix B for the full mapping). These industries were selected to reflect the common industries that data vendors currently serve. Including these industries allows vendors to better grasp the overall third-party data market in the federal government and could yield valuable insights into future market analyses using this framework.
- The use case is a categorical variable that can take four values (fraud detection, sentiment analysis, patient matching, and ad targeting). Use cases are the application categories of sales data determined via keyword matching in the job description text. (See Appendix C for the full mapping). Tying use-cases to individual job postings directly links possible clients with marketable data services.

  Other supporting features were also engineered and are available in Appendix D. However, they are not discussed here because they did not play a role in the final model.

3.3 Model Selection and Training

To determine the likelihood of a job posting representing a data buyer, a logistic regression model with both the aforementioned engineered features and text-based inputs from job postings was used. The text fields— job description, duties, and title—were combined and then vectorized for n-gram extraction (unigrams to trigrams); these are short phrases related to data purchasing behavior.

Logistic regression was selected as the model because of its interpretability, rapid computational power, and ability to generate probabilities. The rapid computation of the logistic regression model enables it to process thousands of relevant jobs active on USAJobs at any given time, thereby ensuring a practical and efficient tool that is sufficient for reuse and expansion. The probabilities logistic regression generates can be used as "scores" that can later be used to prioritize outreach by their likelihood of being a data buyer. Given the model's strong performance in cross-validation and out-of-sample testing and the aforementioned advantages that underscore its practical integration, more complex and computationally intensive methods were not pursued.

Data buyers represent a small minority of the overall positions; therefore, a machine learning model is incentivized to overpredict the majority class (non-buyers), which is catastrophic for our research question, as it would never predict that a position would be a data buyer. To eliminate this, we addressed the class imbalance during model training using the synthetic minority oversampling technique (SMOTE). SMOTE creates synthetic examples of the minority between existing points using k-nearest neighbors (n=5), doubling the minority class to better balance the training set and improve the model's ability to distinguish between buyers and non-buyers.

4 Results

4.1 Model Evaluation Metrics

The dataset was split into training and testing sets with an 80/20 split, and a five-fold stratified cross-validation was implemented. The model was evaluated using standard metrics including precision, recall, F1-score, and ROC-AUC. The performance metrics are listed in Table 1. The model achieved a mean ROC AUC of 0.8575 ($\pm$ 0.0101), indicating a strong overall discriminatory power. A mean precision of 88.1% ($\pm$ 5.44%) indicates that when a job posting was a predicted data buyer, it was correct almost nine out of ten times. This performance is especially useful for commercial vendors whose targeted outreach for each lead is costly, and where minimizing false positives (unlikely leads) protects their bottom line. The low standard deviations in the cross-validation metrics suggested that the model exhibited minimal overfitting, thereby ensuring its generalizability.

However, the model's recall was 17.6% ($\pm$ 2.97%), reflecting a known limitation in data with high class imbalances, particularly when paired with logistic regression modeling. Despite implementing SMOTE to address this imbalance, the model remains conservative in predicting data buyers.

Although this relatively low recall indicates that many potential buyers may remain undetected, the model's precision, computational speed, and stable cross-validation performance make it a strong candidate for scalable lead identification and prioritization.

4.2 Important Features

A feature importance analysis revealed that the model relied primarily on vectorized text to determine likely data-buying behaviors. Figure 1 shows the top 20 most

Table 1 Cross-validated model performance metrics (mean and standard deviation across 5 folds).

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Mean	0.8672	0.8810	0.1763	0.2927	0.8575
Std	0.0045	0.0544	0.0325	0.0478	0.0101

important features based on the absolute values of their coefficients. The most influential features included unigrams such as "management," "acquisition," "procurement," "contracts," and "licensed", which strongly align with some of our initial labeling of explicit data buyers. This strong alignment, but with additional insights, shows that the model has captured strong data buying signals from our initial keyword mapping from the training data, and also developed its own textual insights, showcased by the presence of features such as "stakeholders," "program service," and "commercial".

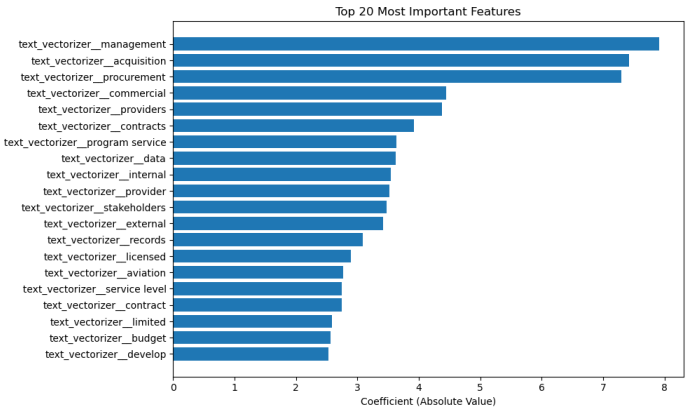

Fig. 1 Feature importance scores based on the absolute values of the model's coefficients. Higher absolute coefficients indicate greater influence on the model's predictions.

5 Applications

This framework is designed to help data vendors and research economists estimate real-time demand for third-party data within the U.S. federal government. By scoring and assigning use-cases to job postings, the framework enables data vendors to prioritize outreach toward the agencies and positions most aligned with the data types they provide.

For example, a vendor specializing in patient record matching datasets could filter the output list for job postings that match the patient record matching use case and then sort it by score (high to low), enabling the vendor to direct outreach efforts towards the most likely potential customers that match their business strength. This strategy provides outreach effort structures by generating possible leads by position

and order to contact these leads. Scores include more nuances in outreach strategies by including a scale instead of a binary indicator of the buyer.

In addition to structuring the outreach for vendors, this framework allows vendors and researchers to assess market dynamics over time. By running the model over previous periods in the USAJobs API, users can create panel data to enable researchers to track which agencies are expanding their data teams, which use cases are most in demand, how these evolutions compare across industries, and how data-buying positions evolve across hiring seasons or policy shifts.

In short, this framework allows the transformation of publicly available federal job postings into structured data that can create an actionable source of market intelligence and demand estimation.

6 Discussion

This framework provides a novel and interpretable method for uncovering third-party data demand in the U.S. federal sector. By engineering structured features from postings and deriving text signals via NLP processing from job descriptions and titles, we trained a logistic regression on the surface patterns of data purchasing behavior across agencies, industries, and positions. However, these findings rely on several key assumptions: that the intent of job postings to purchase data is a reliable signal of actual procurement, and that keyword and fuzzy matching techniques can approximate buyer classification to train the model adequately. Although these assumptions introduce limitations, they also create a structured framework that enables scalable lead generation and classification across thousands of job postings. This section discusses the observed market patterns, offers vendor-client targeting strategies, the methodology's limitations, and future directions for expanding, validating, and improving the framework.

6.1 Patterns in Public Sector Demand

The model revealed several patterns of interest regarding how third-party data demand is distributed across agencies, industries, and positions. Positions most frequently identified as potential buyers often included language related to "program management," "procurement," and "contracting," suggesting that data buyers are not limited to traditional "data scientist" or specialist data roles. This model demonstrates that buying data can be combined with existing contracting roles. Vendors could benefit from broadening their outreach and targeting strategies to include more generalist contracting positions.

Additionally, smaller agencies had a significantly higher percentage of data purchasing positions than large agencies. This phenomenon could reflect smaller agencies' reliance on external contracts, possibly due to leaner staffing or limited internal resources. A high proportion of finance, marketing, and policy analysis positions were flagged as data buyers. These aligned well with fraud-detection and sentiment-analysis use cases. Vendors have rich demand density in these use cases and industries. Additionally, many agencies can be mapped to industries; for instance, the treasury department is primarily associated with the finance industry, and the main use of

financial institutions is fraud detection. With this knowledge, vendors can personalize agency outreach, specifically by framing their pitch around each agency's needs.

6.2 Limitations

This study has several limitations that warrant consideration. First, training labels for explicit data buyers were generated via keywords and fuzzy matching. Although validated through a manual review, some notable false positives were repressed before future modeling. These false positives include jobs whose duties resemble those of data buyers, but are not known to purchase data. The most notable false-positive title was pharmacy technician. Pharmacy technicians deal with third-party insurance claims and often have duties listed that include words such as "records," "inventory," and "operations." Similar words trigger the fuzzy matching algorithm at a .85 threshold. While these false positives represent a minority, future human reviews and narrowing of keyword-mapping lists could remedy this problem. Notably, these false positives mostly come from within the medical industry, as many positions include similar words, such as pharmacy technician positions.

Second, this approach likely misses subsets of true data buyers by assuming that data purchasing activity is explicitly listed in the job description (for the training set). This labeling assumption likely affected the model's recall. However, the precision of this method remained high.

Third, the model relied heavily on the vocabulary present in job descriptions. While this allows for strong overall pattern detection, it may be limited by agency-specific language or position title standards. For example, a data analyst at the Centers for Disease Control and Prevention is likely to have very different daily tasks than those at the Treasury Department, despite having the same title. As such, positions could vary across agencies or in the future as the data market continues to evolve rapidly. Finally, and most importantly, we do not have access to true procurement records; this model estimates demand through intent in job descriptions rather than through actual purchases.

7 Future Work

Several possible changes could strengthen this framework and provide market insights. Integrating human oversight ensures that model outputs make intuitive sense and recognize market shifts by introducing new keyword mapping to keep the framework relevant. Incorporating temporal scraping of USAJobs postings to uncover market trends over time could be especially useful when evaluating policy changes or hiring seasons. Additionally, access to true purchasing data could be used to cross-validate and assess the discrepancy between published intent and actual purchasing in the market, as well as to improve model performance.

Future work for vendors could lead to dashboard ranking by data-buyer score and customizable industry, use cases, and agency filters. Researchers can apply the same framework to other job boards, possibly focusing on sections of the private sector where the data market lacks transparency.

8 Conclusion

This study presents a novel open-source framework for identifying third-party data demand in the U.S. federal sector by analyzing job postings through natural language processing and structured positional features. This approach generates a Data Buyer Score for each position, enabling data vendors, researchers, and market analysts to prioritize the roles most likely to engage in data purchasing.

Unlike prior job classification or skill extraction models, this framework is explicitly designed to detect latent signals of data purchasing intent, which suggests that it could play a role in data purchasing, where these patterns are reflected in the data buyer scores. The model balanced precision and practical scalability by combining interpretable machine learning with engineered job features and text-based signals. Its high cross-validation precision and stable performance indicate that it is a strong candidate for demand estimation. An openly available code base ensures that the tool can be reused, expanded, and most importantly, adapted as the data market continues to evolve rapidly.

As external data continues to play an increasingly critical role in public sector decision making, understanding this emerging demand has become increasingly valuable. This framework provides a foundation for more structured market analysis and client targeting. The framework can be expanded to monitor shifts in public sector data demand over time. Future work may consider improving labeling strategies, developing more complex and computationally intensive models, or estimating demand in the private sector or at lower levels in the public sector, such as state or local governments.

Acknowledgements. We would like to thank Editage (www.editage.com) for editing. We also acknowledge the use of OpenAI to support the classification of agency size, generating keywords, language refinement, and code debugging; the author conducted all decisions, validation, and interpretations.

Declarations

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Clinical trial number: not applicable

8.1 Author Contribution Statement

The author is solely responsible for the conceptualization, collection, analysis, interpretation, and manuscript preparation.

8.2 Funding

No funding was received to support the preparation of this manuscript.

8.3 Conflict of Interest

The author declares no competing interests.

8.4 Data Availability

The analysis scripts and study findings are also publicly available at:
https://github.com/RoryQo/Public-Sector-Data-Demand_Research-Framework-For-Market-Analysis-And-Classification

Additionally, the model developed in this study has been integrated into a python package available for reuse via PyPI (https://pypi.org/project/data-buyer-toolkit/).

References

[1] Xu, H., Gu, C., Zhou, H., Kou, S., & Zhang, J. (2017). *JCTC: A Large Job Posting Corpus for Text Classification*. Retrieved May 15, 2025.

[2] Senger, E., Zhang, M., Van Der Goot, R., & Plank, B. (2024). *Deep Learning-based Computational Job Market Analysis: A Survey on Skill Extraction and Classification from Job Postings*. Retrieved May 15, 2025.

[3] Lukauskas, M., Šarkauskaitė, V., Pilinkienė, V., Stundžienė, A., Grybauskas, A., & Bruneckienė, J. (2023). Enhancing Skills Demand Understanding through Job Ad Segmentation Using NLP and Clustering Techniques. *Applied Sciences*, *13*(10), 6119. https://doi.org/10.3390/app13106119

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[5] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 57–61).

[6] USAJobs. (2025). *About USAJOBS: Working for the Federal Government*. Retrieved from https://www.usajobs.gov/Help/about/

Appendix A. Initial API Search Keywords

The following keywords were used in the 'Areas of Interest' field of the USAJobs API to surface relevant job postings:

- data
- contract
- analyst
- machine learning
- marketing
- acquisition
- finance
- security
- tech
- purchasing
- statistics
- math
- data scientist
- research
- economist

Appendix B. Data Buyer Identification Phrases

B1. Exact Match Phrases (Direct Search)

The following phrases were used to identify explicit data buyers via exact text matching:

- data acquisition
- data procurement
- procure data
- purchase data
- buy data
- acquiring data
- data sourcing
- data licensing
- external data acquisition
- third-party data
- data vendor
- data provider
- data contracts
- contracting data
- data subscriptions
- vendor management
- external data
- commercial data

B2. Fuzzy Match Signal Phrases (Expanded)

Extended phrases captured via fuzzy matching (threshold = 85%):

- data assets
- data commercialization
- procurement of data
- external data sources
- data aggregators
- data monetization
- sourcing external data
- partner data
- data purchasing agreements
- data ingestion
- subscription data
- data acquisition strategy
- data buying
- external datasets
- external partnerships
- data sharing agreements
- data acquisition channels

- third-party data sources
- sourcing data providers
- managing data vendors
- data reseller
- external data vendors
- contracted data

Appendix C. Feature Engineering Keywords and Mappings

C1. Seniority Identification

Seniority was determined based on the presence of these keywords in job titles:

- senior
- lead
- chief
- principal
- director
- head

C2. Agency Size Classification

Agencies were classified as:

Large Agencies:

- Department of Defense
- Department of Veterans Affairs
- Department of the Treasury
- Department of Homeland Security
- Department of Health and Human Services
- Department of Justice
- Department of the Army

Medium Agencies:

- Department of Transportation
- Department of Commerce
- Department of Agriculture
- Department of Energy
- Department of the Interior
- National Aeronautics and Space Administration (NASA)

Small Agencies: All others.

C3. Industry Classification

Industry categories were assigned based on these keyword matches:

- Finance: finance, financial, account, budget

- Marketing: marketing, communications, advertising
- Medical: medical, pharmacy, nurse, health, clinical
- Security/Tech: cyber, security, information technology, IT, data scientist, software, tech
- Policy: policy, regulation, legislative, analyst, compliance
- Other: default when no match occurred

C4. Use Case Mapping

Use cases were derived from text matches to:

- Fraud Detection: fraud, eligibility, verification, audit, compliance
- Sentiment Analysis: sentiment, public opinion, media monitoring, engagement, communication
- Patient Matching: patient match, interoperability, record linkage, EHR, health record
- Ad Targeting: audience segmentation, targeting, ad performance, campaign data

C5. Generalist Role Identification

Generalist titles were determined using fuzzy matching against:

- Contract Specialist
- Grants Officer
- Grants Specialist
- Budget Officer
- Administrative Officer
- Operations Coordinator
- Program Coordinator
- Project Coordinator
- Procurement Specialist
- Procurement Analyst
- Communications Specialist
- Public Affairs Officer
- Public Information Officer
- Community Outreach Coordinator
- Health IT Coordinator
- Program Specialist
- Program Manager
- Business Operations Specialist

Appendix D. Feature Engineering

This appendix outlines the additional set of engineered features.

- **IsExplicitDataJob**: Indicates whether the job title includes direct references to data or analytics roles (e.g., 'data scientist', 'statistician', 'intelligence analyst').
- **IsGeneralistRole**: Indicates whether a job has a generalist or administrative title (e.g., 'Contract Specialist', 'Grant Officer') using fuzzy matching against a curated generalist role list.
- **SearchKeywords**: The original keyword(s) that triggered the inclusion of a job in the API search (e.g. 'data', 'marketing', 'contract').

Each of these features was incorporated into the labeled dataset. These features help enable market analysis of third-party data demand in federal job postings.