What Job Is This Anyway?

Using LLMs to Classify USAJobs Data Scientist Listings

Author: Abigail Haddad Date: October 19, 2023

I'm a data scientist, not a {data analyst, business intelligence analyst, software engineer}

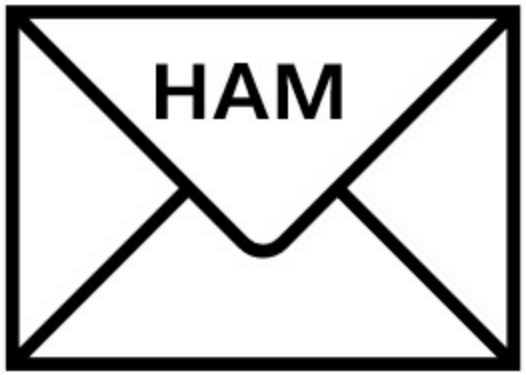
I'm a data scientist, not a {data analyst, business intelligence analyst, software engineer}

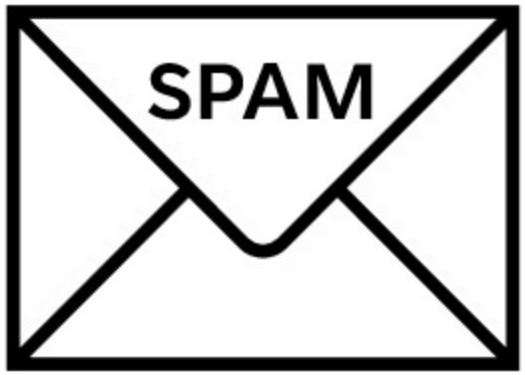
Solution: Text Classification

1. We can use LLMs.
2. This can help job seekers....
3. ..and enable internal government analysis and changes.

ChatGPT, can you just show me the actual data scientist job postings?

A detour to discuss classification problems





Assessing classification problems

Confusion Matrix

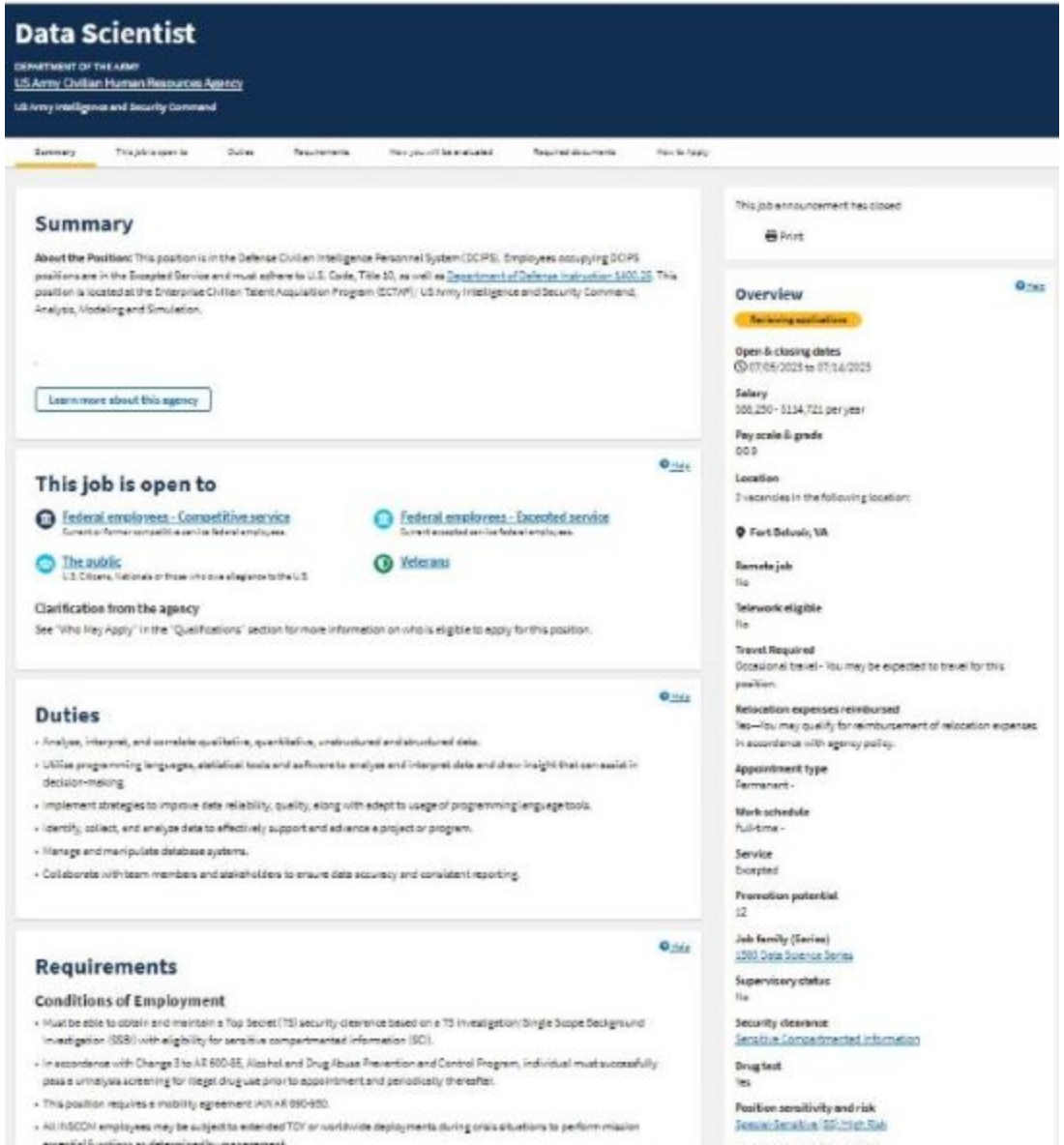
	Predicted Not Spam	Predicted Spam
Actual Not Spam	479	27
Actual Spam	24	470

Assessing classification problems

Derivative metrics

Metric	Value
Recall (True Positive Rate)	95%
Specificity (True Negative Rate)	94%
Precision (Positive Predictive Value)	94%
NPV (Negative Predictive Value)	95%

Back to USAJobs Listings



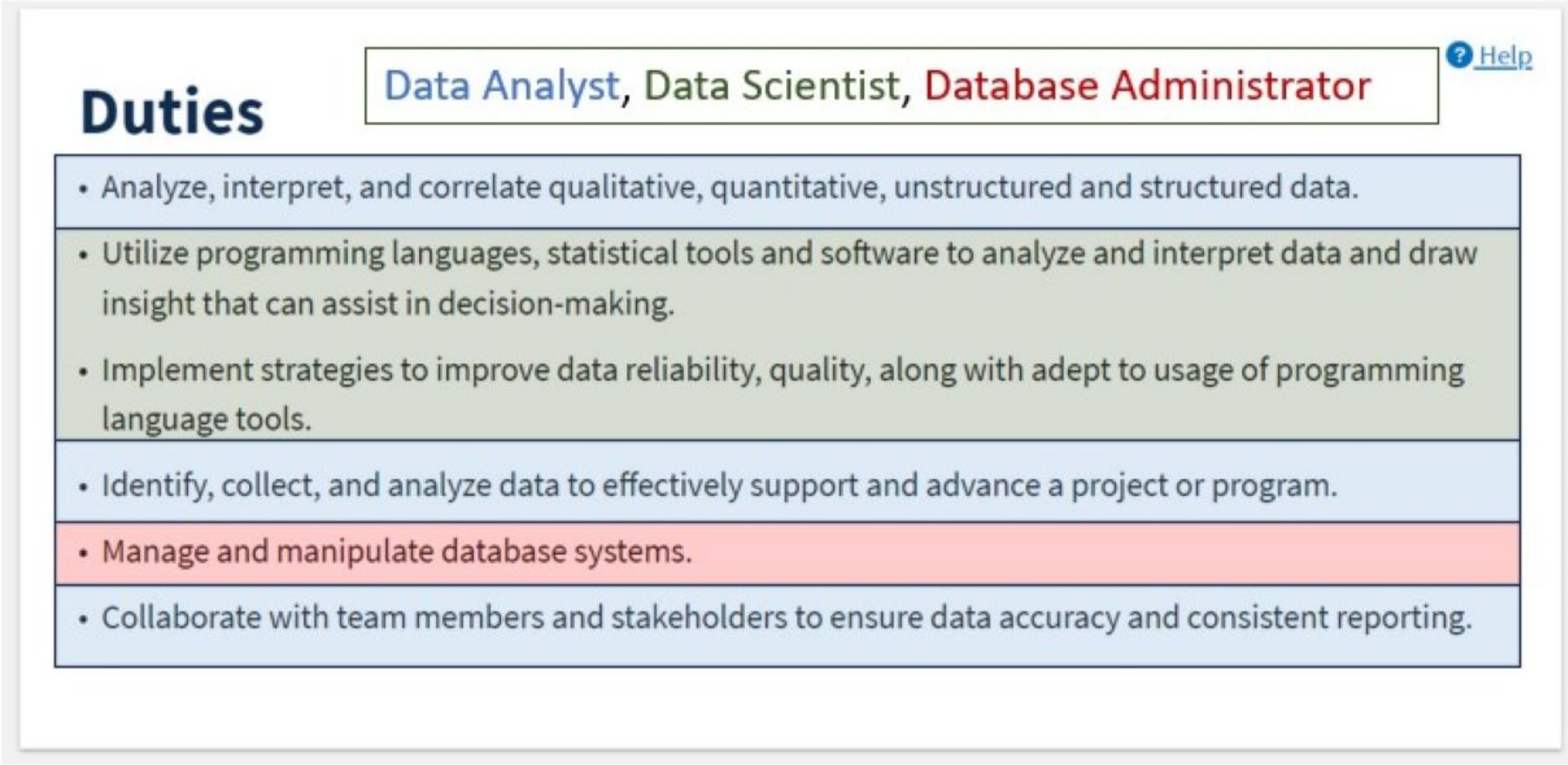
Example 'Duties' Text

Help

Duties

- Analyze, interpret, and correlate qualitative, quantitative, unstructured and structured data.
- Utilize programming languages, statistical tools and software to analyze and interpret data and draw insight that can assist in decision-making.
- Implement strategies to improve data reliability, quality, along with adept to usage of programming language tools.
- Identify, collect, and analyze data to effectively support and advance a project or program.
- Manage and manipulate database systems.
- Collaborate with team members and stakeholders to ensure data accuracy and consistent reporting.

Example 'Duties' Text with Highlights



We can't overcome text ambiguity problems with LLMs. If the information isn't there, it's not there.

My MVP Project Workflow

1. Pulled 843 1560 Data Scientist job listings.
2. Dropped those with no duty list via length/keyword filter.
3. Multi-label classification via GPT-3.5 3a. Assess consistency scores/inter-rater reliability
4. Analysis of results 4a. BERT encoding of labels, clustering. 4b. GPT labels of clusters 4c. Word cloud 4d. Ad hoc validation of results

What can we actually get from this?

What can we actually get from this?

- A view of the variety of different roles that are under 1560
- A prototype job labeling system for applicants
- A way to highlight jobs that might be mislabeled or challenging to hire for
- A variable that might be predictive for research and analysis

What this can't be

- The one (or more) "true" label
- A solution to ambiguous duties sections


Machine Learning Engineer
Policy Analyst
Systems Analyst
Data Governance Specialist
Supervisor
Data Analyst
Health Data Analyst
HR Analyst
Database Developer
Operations Research Analyst
Statistician
Intelligence Analyst
Computer Scientist
Artificial Intelligence Specialist
Data Engineer
Database Administrator
Geospatial Analyst
Program Analyst
Business Intelligence Analyst
Consultant
Mathematical Modeler
Data Mining Specialist
Data Architect
Financial Analyst
Statistical Analyst
Business Analyst
Data Strategist
Software Engineer
Systems Engineer
Data Manager
Data Governance Manager
Data Scientist
Data Visualization Specialist
Research Scientist
Public Health Analyst
Economist
Project Manager
Research Analyst
Software Developer
Quantitative Analyst
Program Manager
Technical Advisor
Emergency Management Specialist

Cluster	%	Top 5 Titles
Data & Project Management	34	Data Scientist, Project Manager, Data Analyst, Program Manager, Supervisor
Core Data Science & Engineering	22	Data Scientist, Data Analyst, Machine Learning Engineer, Data Engineer, Statistical Analyst
Data Science & Research	17	Data Scientist, Data Analyst, Research Scientist, Data Engineer, Database Administrator
Data Strategy & Specialization	16	Data Analyst, Data Strategist, Geospatial Analyst, Supervisory Data Scientist, Chief Data Scientist
Data Analysis & Research Intelligence	10	Data Analyst, Data Scientist, Business Intelligence Analyst, Statistical Analyst, Research Analyst

Spot Checking Title Sets

Data Scientist, Data Analyst, Software Engineer

Help

Duties

- Conducts data mining, analysis, modeling, development of algorithms, and computer program design activities with knowledge software development methodologies, the necessary mathematics, coding languages, and analytics tools.
- Make recommendations regarding the implementation of existing technologies and techniques related to data science.
- Recommends usage and access control policies and systems in collaboration with system security design staff and stakeholders in continuous improvement processes impacting data quality, performance enhancements, and overall user experience.

Data Scientist, Data Analyst, Business Intelligence Analyst, Machine Learning Engineer, Artificial Intelligence Specialist, Data Engineer

Duties

Leads projects using advanced analytics techniques such as machine learning, natural language processing, robotic process automation, and artificial intelligence, to include the research, design, development, deployment, and enhancement of enterprise-wide analytics solutions.

Provides advice, guidance, and support to junior staff on advanced analytics, data access, data quality, data governance and storage, and related areas.

Identifies key deficiencies in project requirements and/or proposed approaches or data access issues which are significant to the design, execution, and reporting out of advanced analytics projects.

Uses expert level knowledge to utilize common data science tools, including scripted languages (such as R, Python, SQL, and JavaScript), Integrated Development Environment and analytics platforms, open-source solutions, commercial off-the-shelf tools and hardware-based capabilities.

Independently designs and develops new data requirements and approaches to knowledge discovery to solve known business problems and/or demonstrate patterns, trends, relationships and anomalies of particular interest or concern.

Researches, designs, and develops user-centric interactive reports, dashboards, and visualization solutions that provide business insights to decision makers using Microsoft Power BI and other tools.

Develops and maintains effective working relationships with staff at the OIG and other agencies in order to share information and learn the latest developments in analytical tools and techniques to efficiently assist the OIG with mission-related work.

Acts as a subject matter expert for data analytics as it pertains to both predictive and prescriptive analysis and ensure alignment with industry best practices.

Design, develop, implement, and manage methods for capturing and converting structured and unstructured data for reporting and analysis.

Data [Analyst, Architect, Scientist Manager]

Help

Duties

- Duties are at full performance grade.
- Conducts data analytics to assess data requirements, establishes data architecture, identifies shortfalls and gaps, troubleshoots data issues, and recommends ways to mitigate risk.
- Manages and collaborates with data analytics staff to develop, analyze, and recommends new methodologies to integrate various data sources derived from discrete or conjoined systems.
- Manages and collaborates with data analytics staff to develop, analyze, and recommends new methodologies to integrate various data sources derived from discrete or conjoined systems.
- Collaborates with senior leadership and other stakeholders on complex data problems, makes sound recommendations for solution sets, and implements data decisions based on scientific data principles, affordable solutions, and stakeholder input.

What's better than ad hoc assessment?

Systematic assessment!

Derivative Metrics Revisited

Metric Name	Acronym	Definition
Recall	TPR	The percentage of actual positives that the model correctly identified.
Specificity	TNR	The percentage of actual negatives that the model correctly identified.
Precision	PPV	Out of all the instances labeled as positives by the model, the percentage that are actual positives.
Negative Predictive Value	NPV	Out of all the instances labeled as negatives by the model, the percentage that are actual negatives.

Assessment Example with Synthetic Data

Label	TPR (Recall)	TNR (Specificity)	PPV (Precision)	NPV
Data Scientist	23%	90%	26%	88%
Project Manager	33%	91%	35%	90%
Data Analyst	32%	90%	33%	90%
Program Manager	24%	90%	25%	90%
Supervisor	28%	90%	27%	90%

But also, are current labels perfect?

What Can We Do With This?

- If you're a federal job seeker or researcher/analyst, let's talk!
- You might have a text classification problem -- and it might be multi-label.
- Importance of structured outputs and assessment.

Necessary Technical Improvements

- Labeling/assessment
- Revisit clustering
- More structured LLM responses (possibly with the Marvin library)