

Detección de patrones multivariados en ofertas fraudulentas mediante análisis de datos mixtos y sistema de scoring interpretable con visualización interactiva

INTEGRANTES:

Jharold Alonso Mayorga Villena

CONTEXTO

- **Origen del Dataset**

El dataset proviene de Kaggle, una plataforma reconocida por alojar conjuntos de datos para análisis y competencias de ciencia de datos.

- **Importancia del Problema**

- Las ofertas de trabajo fraudulentas son un problema creciente en el mercado laboral, afectando tanto a candidatos como a empresas legítimas.
- Estas ofertas pueden llevar a estafas financieras, robo de identidad o explotación laboral.

- **Composición del Dataset**

- Datos mixtos: Incluye información estructurada (salario, ubicación, tipo de empleo) y no estructurada (descripciones, requisitos).
- Etiquetas claras: Cada oferta está clasificada como "real" (0) o "fraudulenta" (1).

- **Variables Clave**

- Características numéricas: Salario, longitud de la descripción.
- Características categóricas: Tipo de empleo (full-time, part-time), nivel de experiencia.
- Texto libre: Descripción del puesto, requisitos, beneficios.



Software engineer in Austin, Texas
2,702 results

Software Engineer
CDS Global, Inc.
Austin, TX, US
Preferred skills and experience. Proven successful experience making decisions on variety of tasks requiring discretion, judgment,...

Sr. Software Engineer in Test
UFCU
Austin, TX, US
Eight years of quality assurance experience; may substitute related job qualifications such as production/customer support or busi...

Sr. Software Engineer
Main Street Hub
Austin, TX, US
The Engineering Team is on a mission to build the absolute best software in the local space. 7+ years in software development ...

Experienced Installation Software Engineer
Emerson Automation Solutions
Austin, TX, US
The engineer is responsible for installing new sites, including application...

Job description

Summary

Seniority Level
Mid-Senior level

Industry
Outsourcing/Offshoring

Employment Type
Full-time

Job Functions
Engineering

Problema

El problema central de este análisis es la gran cantidad de datos mixtos provenientes de diversas fuentes (textos, numéricos, categóricos, etc.) que dificulta un análisis profundo y eficiente debido a su estructura compleja.



Características del dataset: El dataset está compuesto por datos mixtos (textos, datos numéricos, categóricos) que incluyen información sobre ofertas de trabajo, como salario, tipo de empleo, ubicación, requisitos, entre otros.



Desafíos del análisis: Los datos no están estructurados de manera convencional, lo que hace difícil aplicar técnicas tradicionales de análisis de datos.



Consecuencias: No se realiza un análisis exhaustivo de este tipo de datos, lo que limita la capacidad de detectar patrones y hacer predicciones efectivas.

Objetivo

El objetivo principal de este proyecto es detectar patrones de ofertas fraudulentas mediante el análisis de variables correlacionadas entre sí, con el fin de realizar un análisis más profundo y encontrar patrones recurrentes que permitan identificar ofertas fraudulentas con mayor precisión.



Realizar un análisis exploratorio de los datos para identificar patrones preliminares en las variables.



Implementar un sistema de scoring interpretable que permita clasificar las ofertas en fraudulentas y no fraudulentas basándose en patrones detectados.



Correlacionar variables que influyen en la probabilidad de fraude, como tipo de empleo, salario, ubicación, y longitud de las descripciones.



Visualizar los resultados de manera interactiva, permitiendo al usuario explorar diferentes escenarios y detectar patrones de fraude.

1

Hipótesis 1: ¿El tipo de empleo (tiempo completo, medio tiempo, contrato) tiene un impacto significativo en la probabilidad de que una oferta sea fraudulenta?

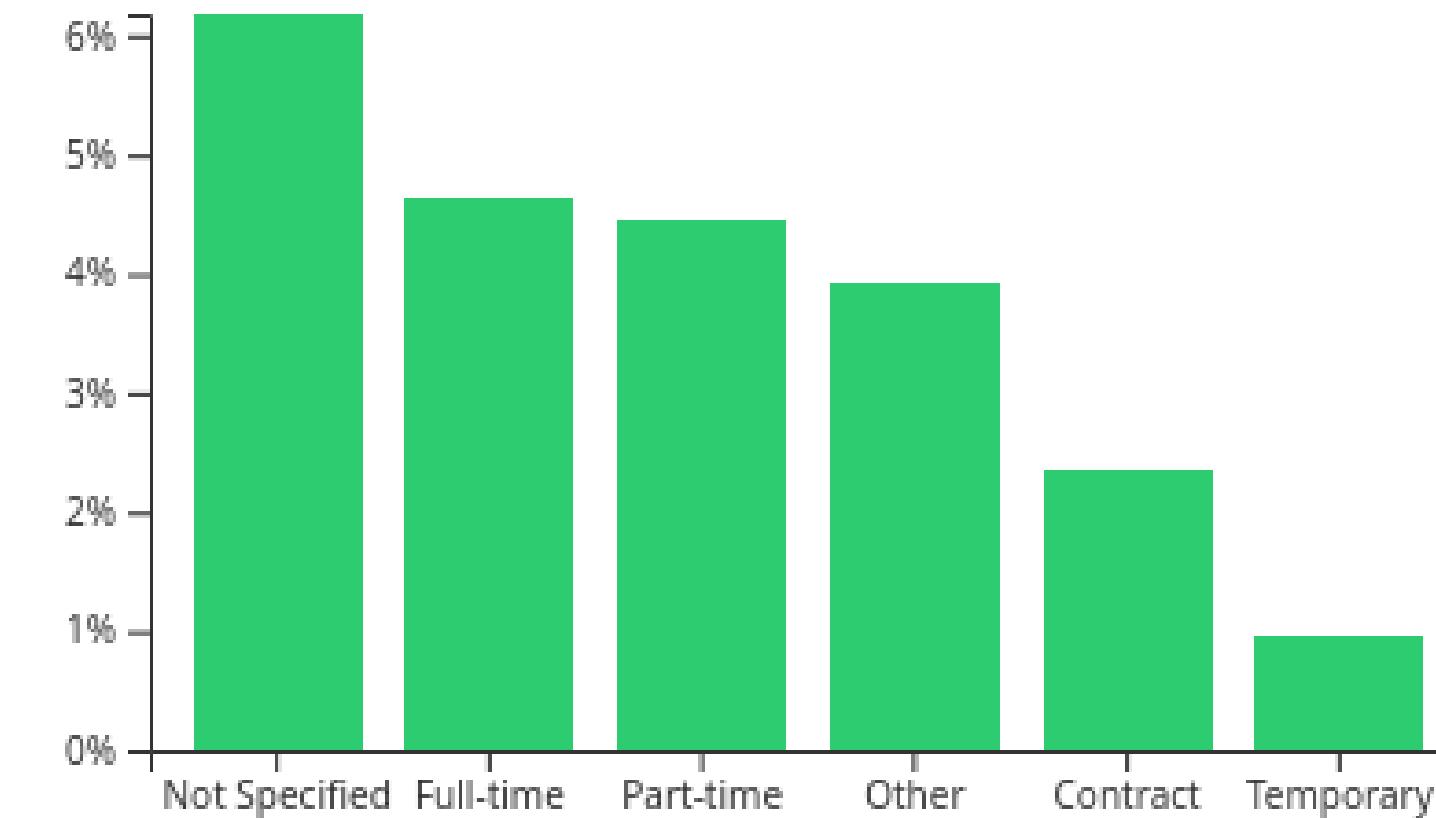
1. **Hipótesis:** Las ofertas de trabajo de medio tiempo o contrato son más propensas a ser fraudulentas que las ofertas a tiempo completo.

2. **Análisis:** En el gráfico se observa que los trabajos de part-time tienen una tasa de fraude considerablemente más alta en comparación con trabajos full-time y contract. Esto puede deberse a que las ofertas freelance o part-time son más fáciles de crear sin mucha verificación, lo que aumenta el riesgo de fraude.



Tipo de Empleo vs. Fraude

Tasa de Fraude por Tipo de Empleo



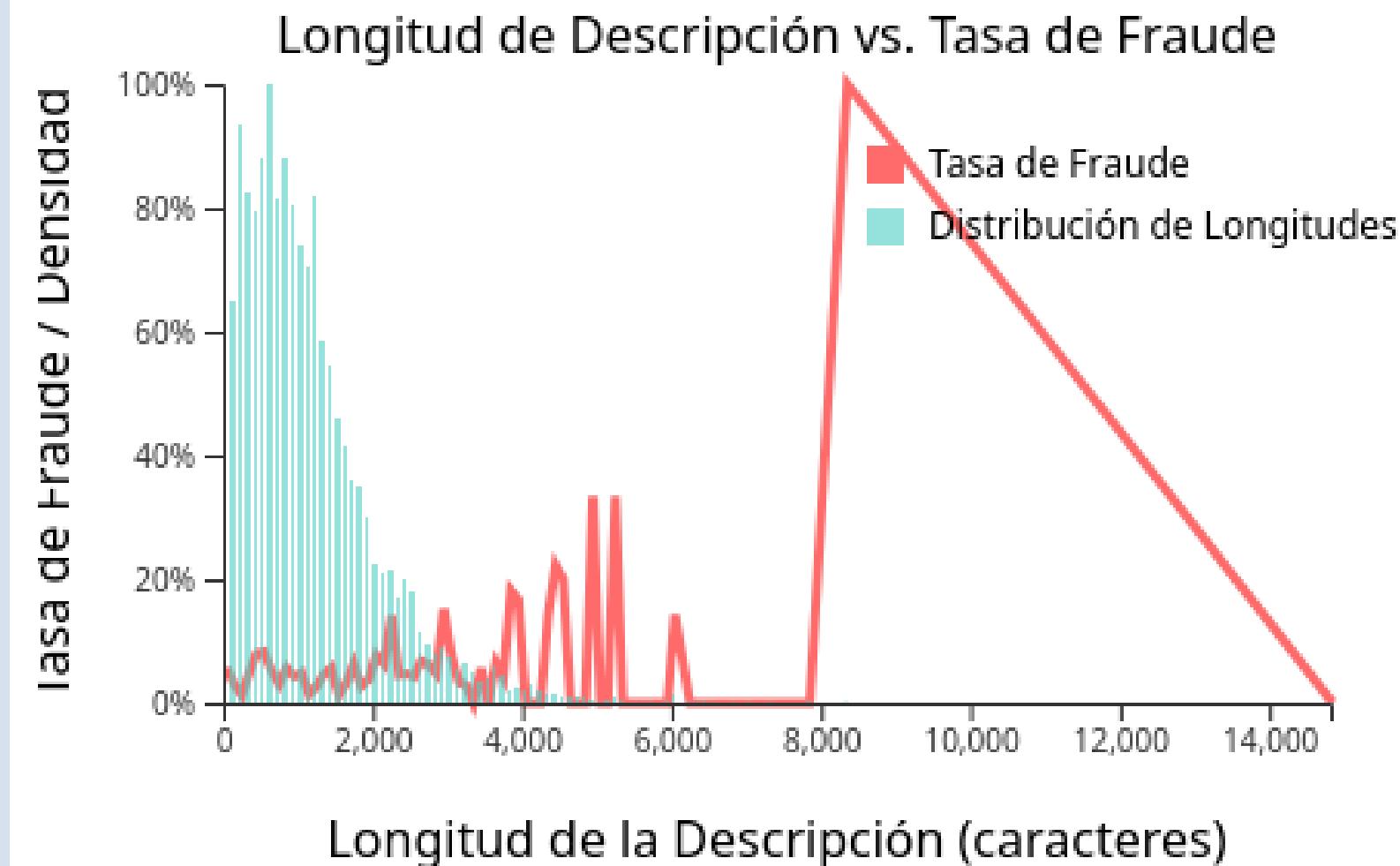
2

Hipótesis 2: ¿Las ofertas fraudulentas tienden a tener descripciones y requisitos más vagos o menos detallados en comparación con las ofertas legítimas?

1. **Hipótesis:** Las ofertas fraudulentas presentan descripciones más genéricas y menos detalladas. Esto se debe a que los estafadores buscan atraer a una gran cantidad de candidatos sin brindar información clara ni específica sobre el trabajo.

2. **Análisis:** Las ofertas fraudulentas suelen tener descripciones más cortas o muy largas, lo que sugiere que los estafadores emplean textos ambiguos o extensos para captar rápidamente la atención de candidatos, sin ofrecer detalles claros y específicos sobre el trabajo.

III Longitud Descripción vs. Fraude



3

Hipótesis 3: Las ofertas de trabajo con requisitos menos específicos (educación, experiencia o habilidades) tienen una mayor tasa de fraude en comparación a las trabajos reales?

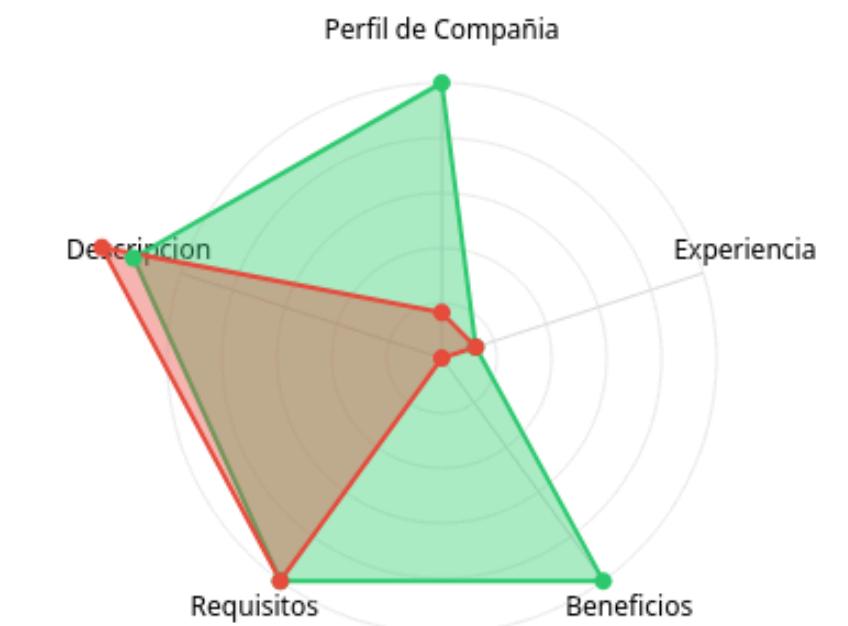
1. **Hipótesis:** Hipótesis: Las ofertas fraudulentas están diversas según la experiencia que se pide, posiblemente debido a diferencias en las regulaciones laborales, o a la falta de requerimientos en diferentes trabajos.

2. **Análisis:** El mapa revela que algunas regiones tienen una alta prevalencia de fraudes por porcentaje según la experiencia requerida, lo que puede estar relacionado con una menor regulación en esas áreas.

Especificidad de Requisitos

- Ofertas Reales
- Ofertas Fraudulentas

Especificidad de Requisitos: Reales vs. Fraudulentas



GRACIAS