CODERHOUSE

Estafas Laborales

¿Cómo reconocer una oferta laboral fraudulenta?

Autor:

David Ferere

Contenido

Contenido

Contexto 3
Audiencia y Objeto 3
Limitaciones 3
Metadata 4
Hipotesis 4
General 4
Especificas 4
Análisis Exploratorio 5
Valores Nulos 5
Análisis de Campo Locación 6
Análisis de Campo Industry 7
Análisis Campos Educación y Experiencia Laboral 7
Análisis Campo Logo de la Compañía Reclutadora 8
Metodología 9
Data Wrangling 9
Feature Engineering 10
Oversampling 10
Datasplitting 10
Feature Selection 10
Algoritmos de aprendizaje y selección de parámetros 11
Resultados 11
Modelo de Aprendizaje 11
Análisis de componente principales 12
Indice de la Silueta 12
BIAS y Varianza 13
Insights, recomendaciones y conclusiones 14
Insights 14
Recomendaciones 14
Conclusiones 14
Referencias consultadas 15

Contexto

La era Post Pandemia se ha visto marcada por una importante recesión económica y un aumento de los niveles de desempleo a nivel mundial, convirtiendo, sobre todo a las economías más vulnerables, en perfecto caldo de cultivo para las actividades delictivas como el robo, el fraude y las estafas.

De esta realidad no escapa el mundo laboral, en el que aparecen ofertas laborales fraudulentas con el fin de captar a las poblaciones más vulnerables, y tomar de ellos datos confidenciales, dinero y hasta redes de tráfico humano.

Audiencia y Objeto

El presente trabajo busca entender los distintos patrones y morfologías del fraude en las ofertas laborales, contestar la pregunta ¿Cómo reconocer una oferta laboral fraudulenta? Este estudio es de interés para aquella población en busca de una oportunidad laboral, así como investigadores del área de prevención de fraude.

Limitaciones

El Dataset a estudiar se encuentra en el idioma inglés lo cual puede representar algunas dificultades para aquellos analistas que no manejen el idioma, adicionalmente, es un dataset desbalanceado en términos del tipo de ofertas (registros) que se puede analizar, en su mayoría legitimas.

Metadata

En esta sección describimos los aspectos mas importantes del Dataset en estudio (Observar Imagen 1).

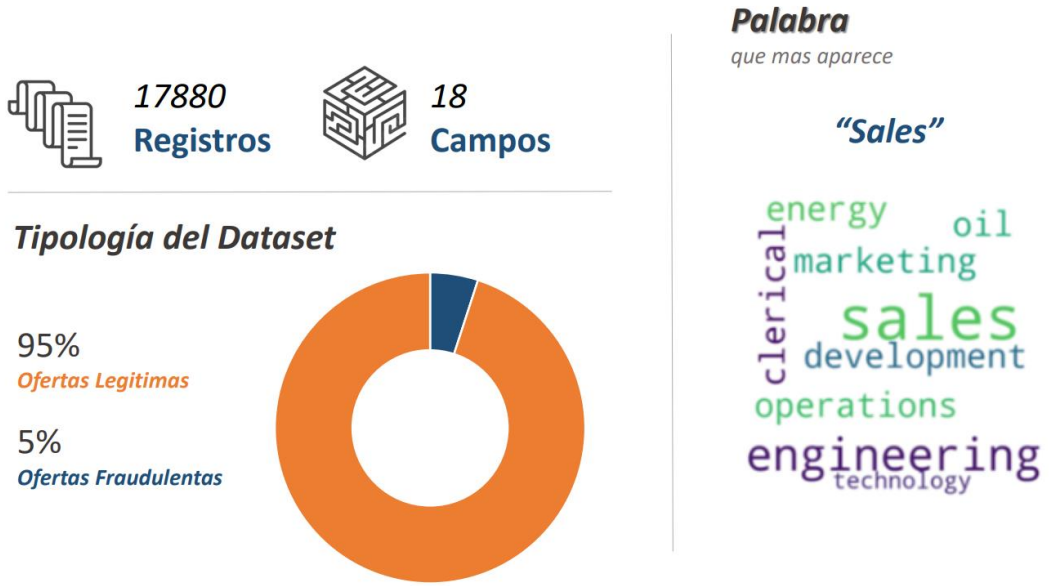

Imagen 1. Metadatos del Dataset.

Podemos observar como aspecto de mayor importancia, que nuestro Dataset se encuentra desbalanceado, es importante destacar que los metadatos se extrajeron a través de un Análisis Exploratorio.

Otro aspecto importante es que la mayoría de los campos en el dataset son del tipo objeto, seguido por valores numéricos enteros, estos valores numéricos son binarios (1, 0), basados en valores verdadero o falso que cumpla cada registro como característica.

Hipotesis

El análisis del Dataset se baso en responder primeramente las siguientes hipótesis:

General

¿Cómo reconocer una oferta laboral fraudulenta?

Especificas

1. ¿Existe relación entre presencia de logo de la compañía reclutadora y la legitimidad de la oferta laboral?
2. ¿Las ofertas laborales según su legitimidad se enfoca en un público con una determinada formación académica?
3. ¿Existen patrones de palabras para identificar una oferta fraudulenta o legitima?

Análisis Exploratorio

En esta sección nos enfocamos en entender la morfología de cada uno de los campos y como sus distintos valores podían estar condicionados al valor del campo target “Fraudulent”, es decir el campo donde su valor nos indica si una oferta laboral es fraudulenta o no.

Valores Nulos

Primero realizamos una exploración del nivel de relevancia que tienen cada uno de las campos de nuestro Dataset.

Para ello se extrae la cantidad de registros con valores nulos en dicho campo, de esta forma hallamos que campos como Salario y Departamento cuentan con un 84% y 65% de valores nulos, respectivamente, lo que puede indicarnos que no sean campos de relevancia para alimentar nuestro algoritmo de aprendizaje.

Por otro lado, nos encontramos con campos como has_questions y telecommuting, referente a si la oferta laboral cuenta preguntas y ofrece trabajo a distancia, que cuentan con 0% de valores nulos, esto implica que pueden ser campos de alta relevancia para nuestro modelo de aprendizaje.

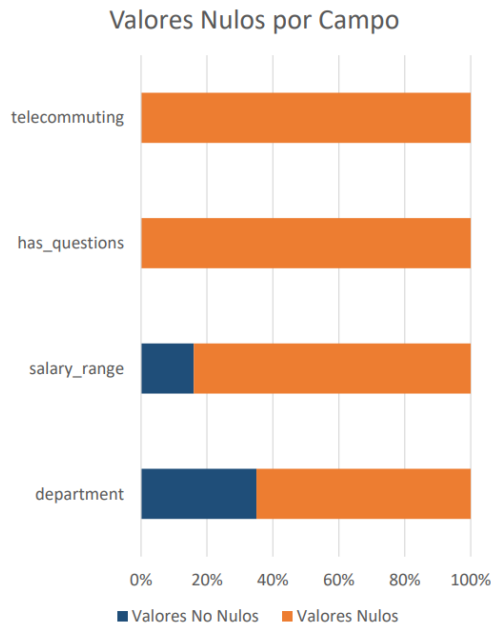

Imagen 2. Cantidad de valores nulos por campo.

Análisis de Campo Locación

Para el caso del campo location nos puede indicar cierto nivel de relevancia debido a que 98% de los registros cuentan con este campo.

El Treemap de abajo se observa las top 10 locaciones presentes en las ofertas laborales Fraudulentas. En la que destaca en primer lugar, con 92 ofertas, Houston – Texas – USA (Observar Imagen 3).

Imagen 3. Ofertas laborales Fraudulentas por Locación.

Por su parte en la Imagen 4 se muestra el top 10 locaciones presentes en las ofertas laborales Legitimas. En este caso destacan las ciudades de Londres y Nueva York en primer y segundo lugar, con 716 y 638 registros, respectivamente.

Imagen 4. Ofertas laborales Legitimas por Locación.

Análisis de Campo Industry

Otro campo de interés para nuestro modelo de aprendizaje es el industry, dicho campo cuenta con solo un 27% de valores nulos, lo cual, puede otorgarnos relevancia en el caracterización de los tipos de ofertas laborales.

En la gráfica de barras de abajo (Imagen 5) podemos observar los top 10 sectores de la industria que aparecen para las ofertas laborales fraudulentas. En la cual podemos destacar el sector de energía, salud, marketing y finanzas.

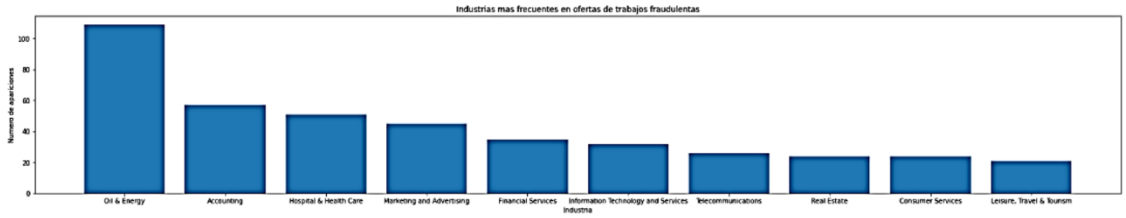

Imagen 5. Principales industrias de las ofertas laborales fraudulentas.

Para el caso de las ofertas laborales legitimas destacan principalmente los sectores de tecnología, servicios de internet e información y educación.

En el análisis del campo industry podemos observar que las ofertas fraudulentas, presentes en el Dataset de estudio, se enfocan más en los sectores de profesiones administrativas mientras las ofertas legitimas se centran en el sector de tecnología (Imagen 6).

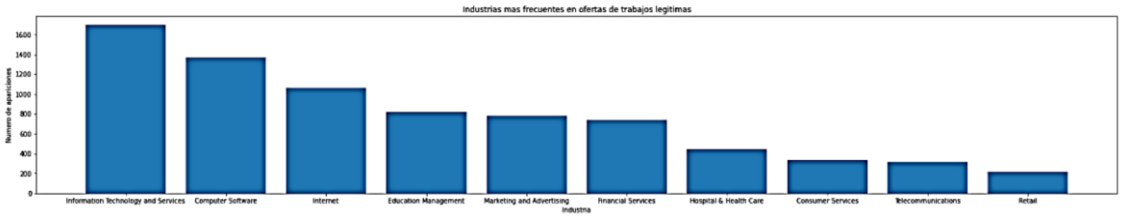

Imagen 6. Principales industrias de las ofertas laborales legitimas.

Análisis Campos Educación y Experiencia Laboral

En esta sección analizamos la correlación que existe entre el nivel de formación y la experiencia laboral que presenta como requerimiento cada tipo de oferta de trabajo.

Para el caso de las ofertas laborales legitimas aproximadamente un 30% exigen tener una licenciatura y un 21% tener una experiencia de un senior medio.

Por su parte, las ofertas fraudulentas en un 20% exigen una educación secundaria y un nivel inicial de experiencia laboral.

De esta observación, presentada en los heatmaps de abajo (Imagen 7 e Imagen 8), se deduce que las ofertas legitimas se enfocan en reclutar personas con una formación media laboral y académica, mientras las fraudulentas en una formación inicial o básica.

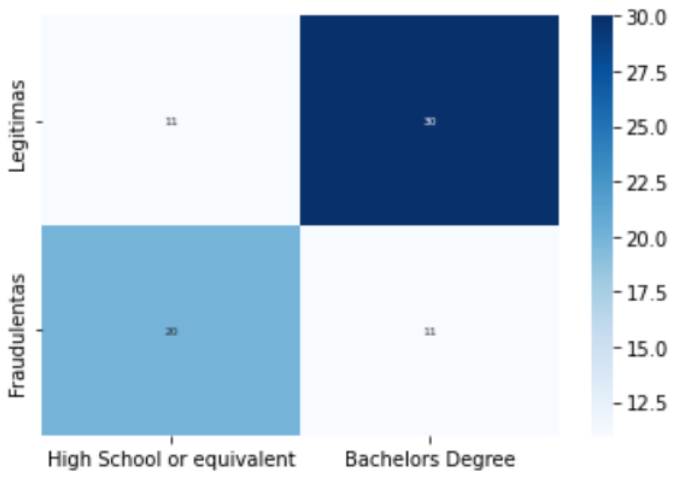

Imagen 7. Nivel de Educación Requerida por tipo de Oferta Laboral.

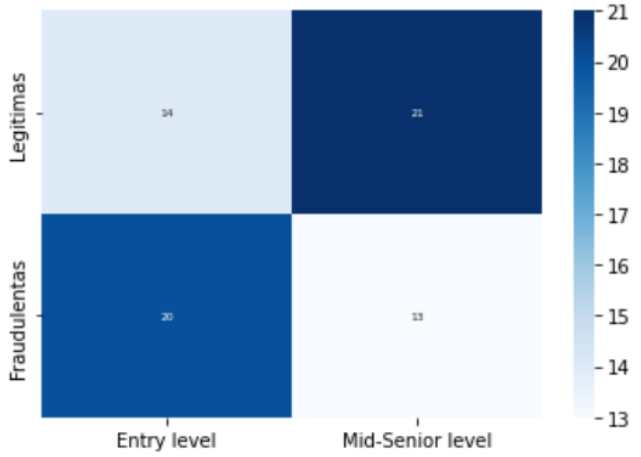

Imagen 8. Nivel de Experiencia Laboral Requerida por tipo de Oferta Laboral.

Análisis Campo Logo de la Compañía Reclutadora

De los gráficos de anillo al lado derecho, podemos observar que el 67% de las ofertas fraudulentas no muestran un logo de la empresa reclutadora, mientras el 82% de las ofertas legitimas si lo hacen (Observar Imagen 9).

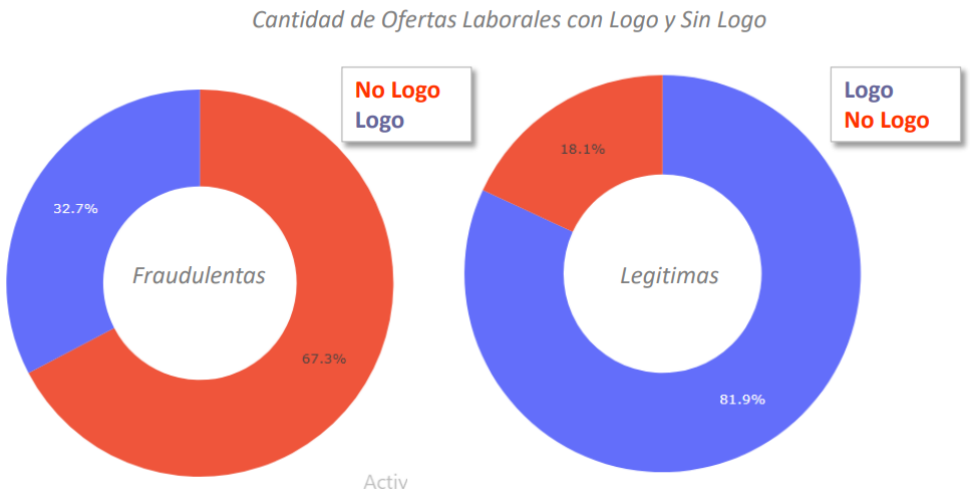

Imagen 9. Ofertas Laborales con y sin Logo.

De este modo podemos indicar que existe una correlación negativa entre el campo has_logo y la legitimidad de las ofertas laborales, esto implica que es muy probable que una oferta fraudulenta no cuente con un logo de la empresa reclutadora (Observar Imagen 10).

	has_company_logo	fraudulent
has_company_logo	1.000000	-0.261971
fraudulent	-0.261971	1.000000

Imagen 10. Correlación entre tipo de oferta laboral y existencia del logo de la compañía reclutadora.

Metodología

En esta sección realizamos una breve descripción de los pasos tomados para el diseño y desarrollo de nuestro modelo de aprendizaje.

Data Wrangling

En este primero paso realizamos un arreglo del dataset, comenzando por la sanitización de los valores nulos y luego el agrupamiento de los datos, por ejemplo en el caso del campo Department agrupamos los datos acorde al tipo de departamento que se definía.

Feature Engineering

Una vez eliminado los datos nulos, se procedió a tomar los campos que podrían tener mayor incidencia en el valor del campo target (Fraudulent), para ello se realizo un etiquetado a través del método Label Encoder de los campos Tipo de Empleo, Experiencia Requerida, Educación, Función e Industria, luego para el caso de los campos Salario, Locación y Beneficios se transformaron en valores binarios 1-0 (True – False), basados en si cada registro tenia estos campos con valores nulos o no nulos. Posterior a ello se procedió a eliminar el resto de campos del dataset.

En una segunda evaluación se realizo nuevamente un Feature Engineering pero esta vez se agregaron los campos Departamento y Perfil de la Compañía Reclutadora.

Oversampling

Debido a que el 95% de las ofertas en nuestro Dataset son legítimas, se debió realizar un balanceo de los tipos de datos, creando datos sintéticos de ofertas laborales fraudulentas, de modo que el 50% del Dataset estuviera conformado por ofertas fraudulentas y otro 50% legitimas.

Datasplitting

Luego para la evalución de nuestro modelo se dividio el Dataset en 80% Datos de entrenamiento y 20% datos de pruebas, estos valores fueron escogidos basados en las recomendaciones hechas en trabajos anteriores de modelos de aprendizajes y la bibliografía consultada.

Feature Selection

En esta sección se aplicaron 3 métodos de selección para determinar previamente al desarrollo de nuestro modelo de aprendizaje, cuales campos tienen mas influencia en la definición del valor del campo target, los tres métodos usados fueron: Forward Selection, Backward Selection y Stepwise Regression. Para los tres métodos se determino que los siguientes campos son los tienen mayor influencia en el valor target.

1. is_profile (Perfil de la Compañía).
2. Industry (Sector de la Industria).
3. has_company_logo (Logo de la Compañía).
4. Department (Departamento de la Oferta Laboral).
5. Function (Funciones de la Oferta).
6. employment_type (Tipo de empleo – Tiempo completo o medio tiempo).
7. has_questions (Realizan preguntas).
8. is_benefits (Describen beneficios).
9. required_experience (Experiencia requerida).
10. Salary (Definen salario).
11. Telecommuting (Ofrecen trabajo remoto).

Algoritmos de aprendizaje y selección de parámetros

Dado que el objetivo de nuestro modelo de aprendizaje es determinar si una oferta laboral es Legitima o Fraudulenta, nos encontramos bajo un escenario de clasificación, por ello se seleccionaron los algoritmos Decision Tree Classifier, Random Forest Classifier y N Vecinos Cercanos.

Para la elección de los mejores parámetros para cada uno de los algoritmos de aprendizajes se tomaron en cuenta 2 metodos: Grid Search y Random Search. Vale acotar que para cada algoritmo se realizaron dos evaluaciones, la primera con el dataset resultante del primer Feature Engineering y la segunda con el del segundo Feature Engineering, esto con el fin de comparar cuales de los dos dataset arrojaban mejores resultados para los parámetros de nuestros modelos de aprendizaje (ver sección de Resultados).

Resultados

Modelo de Aprendizaje

En la imagen a continuación (Imagen 11) podemos observar un cuadro resumen de los resultados obtenidos de los parámetros de cada uno de los algoritmos de aprendizajes evaluados y para el caso del Dataset 1 y Dataset 2.

Modelo	ROC AUC Score	accuracy	recall	f1-score	support
KNN I	0.97	0.91	0.93	0.89	0.91
KNN II	0.98	0.94	0.98	0.90	0.94
Random Forest I	0.91	0.83	0.82	0.85	0.84
Random Forest II	0.92	0.87	0.86	0.88	0.87
Decision Tree I	0.90	0.83	0.86	0.78	0.82
Decision Tree II	0.89	0.85	0.89	0.80	0.84

Imagen 11. Resultados de parámetros para cada algoritmo de evaluación.

Se puede determinar que el algoritmo que otorga mejor accuracy y ROC AUC Score es el de N Vecinos cercanos, y el que otorga peor performance el de Decision Tree. Adicionalmente, la evaluación de nuestro modelo bajo el uso del Dataset II muestra un mejor performance que el caso del Dataset I.

Luego para el método de búsqueda de los mejores parámetros de los algoritmos de aprendizaje usados, observamos que el método Random Search reduce de forma importante los tiempos de búsquedas de los valores de los parámetros (Observar Imagen 12).

Modelo	Grid Search	Random Search
KNN	224.00	80.0
Random Forest	147.00	90.0
Decision Tree	4.95	5.3

Imagen 12. Tiempos ejecución Grid Search vs Random Search.

Análisis de componente principales

En el análisis de PCA obtuvimos que al reducir nuestro dataset a dos componentes principales se perdia aproximadamente el 70% de información, el cual se observa en la imagen 13, donde no existe una división apropiada de nuestro dataset que concentre de forma correcta cada tipo de dato.

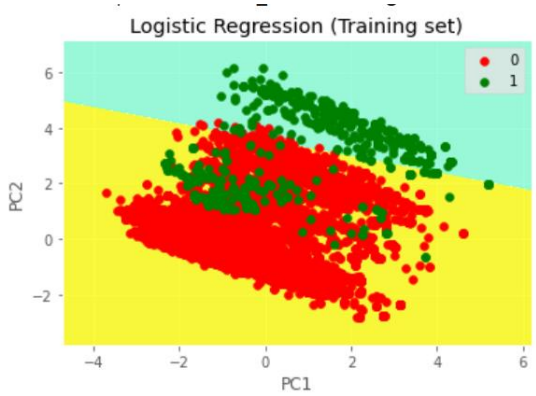

Imagen 13. Reducción del Dataset de entrenamiento en dos componentes.

Indice de la Silueta

Con el análisis de este índice se busca determinar cual es el numero de clusters mas eficiente en el que puede ser divido nuestro dataset, y de modo que todos los clusters logren el valor del kmeans. En nuestro caso se encontró que 8 es el numero de clusters óptimos para nuestro Dataset, como se puede observar en la imagen 14, donde se logra el mayor score de 0.18.

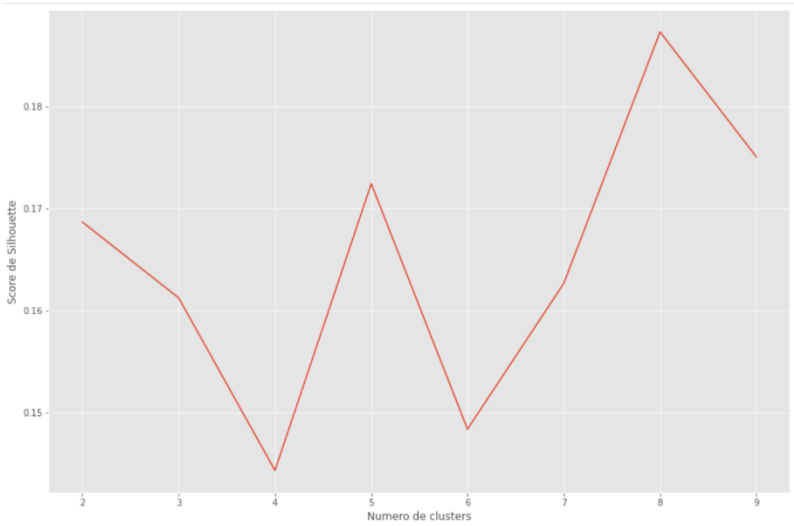

Imagen 14. Score de Silueta por número de Clusters.

BIAS y Varianza

Podemos observar que para el caso del algoritmo Decision Tree el bias es de 0.13 y una varianza de 0.072.

Para el caso de N Vecinos Cercanos obtuvimos un bias 0.13 y una varianza de 0.039, y por último para el algoritmo de Random Forest se obtuvo bias de 0.13 y varianza de 0.037 (Observar Imagen 15).

Lo cual implica que la variabilidad de nuestro modelo para varios dataset de entrenamientos y el margen de error entre los valores predichos y reales son bajos, sin embargo, el algoritmo que presenta un bias y una varianza más baja es el Random Forest, esto no implica que sea el mejor modelo, se debe encontrar un balance entre estos valores debido que una varianza y un bias muy bajos podría indicar un caso de overfitting. Dado esto se elegiría preliminarmente el algoritmo Decision Tree para el diseño y despliegue de nuestro modelo de aprendizaje.

Modelo	BIAS	Varianza
KNN	0.136	0.039
Random Forest	0.135	0.037
Decision Tree	0.138	0.072

Imagen 15. BIAS y Varianza del Modelo de Aprendizaje.

Insights, recomendaciones y conclusiones

Insights

- ✓ El campo con mas relevancia es el has_logo.
- ✓ Existe una correlación, no relevante, entre un nivel de formación básica y la reclusión de ofertas laborales fraudulentas.
- ✓ Las ofertas laborales fraudulentas enfocan el proceso de reclusión en los sectores administrativos, mientras las legitimas en el sector de tecnología.
- ✓ El campo de la locación no es de relevancia, acorde al análisis exploratorio.

Recomendaciones

La primera observación que se desprende de nuestro análisis es referente al desbalanceo de tipos de ofertas laborales que presenta el Dataset. De ello debemos tomar en cuenta la creación de datos sintéticos para completar los datos faltante necesarios para alimentar nuestro modelo de aprendizaje, otra opción presente sería el enriquecimiento de los datos con fuentes externas, que nos permita complementar los datos faltantes.

Conclusiones

- ✓ En términos del BIAS y la varianza el algoritmo Random Forest presenta el mejor performance.
- ✓ Acorde a los parámetros de ROC AUC score, recall, accuracy y support el algoritmo KNN es el que presenta el mejor performance.
- ✓ El análisis de los componentes principales no fue concluyente, debido a que para dos PCA se perdia el 70% de la información del Dataset.
- ✓ El proceso de búsqueda de los mejores parámetros de los algoritmos de aprendizaje mas eficiente es el Random Search, debido a que logro reducir el tiempo de búsqueda en mas de la mitad.
- ✓ El valor de los parámetros del ROC AUC y accuracy se logro mejorar con la inclusión de los campos department y company_profile.
- ✓ Bajo todos los escenarios de análisis nuestro modelo de aprendizaje logro predecir con una precisión mayor al 80%.

Referencias consultadas

- ✓ Kaggle: https://www.kaggle.com/
- ✓ Stackoverflow: https://stackoverflow.com/
- ✓ GeeksforGeeks: https://www.geeksforgeeks.org/
- ✓ W3Schools: https://www.w3schools.com/
- ✓ Masters in Data Science: www.mastersindatascience.org