

Curso DATA SCIENCE

Entrega Trabajo Proyecto Final

Análisis de cambios de trabajo de Data Scientist

INTEGRANTES

Juan Ignacio Gonzalez

Juan Manuel Osses

Juan Manuel Colombain

Diciembre 2021

Contenido

[1. Descripción del caso de negocio 2](#_Toc89406020)

[2. Objetivos del modelo 3](#_Toc89406021)

[3. Descripción de los datos 3](#_Toc89406022)

[4. Hallazgos encontrados en los datos 4](#_Toc89406023)

[5. Algoritmo elegido 5](#_Toc89406024)

[6. Métricas de desempeño del modelo 5](#_Toc89406025)

[7. Iteraciones de optimización 6](#_Toc89406026)

[8. Métricas finales del modelo optimizado 6](#_Toc89406027)

[9. Conclusiones 6](#_Toc89406028)

# Descripción del caso de negocio

Una empresa dedicada a dictar cursos de Data Science cuenta con datos de las personas que participaron de los mismos. A partir de los datos recolectados y con el fin de sumarlos como recursos propios, se genera la necesidad de identificar aquellos individuos que están predispuestos al cambio laboral.

# Objetivos del modelo

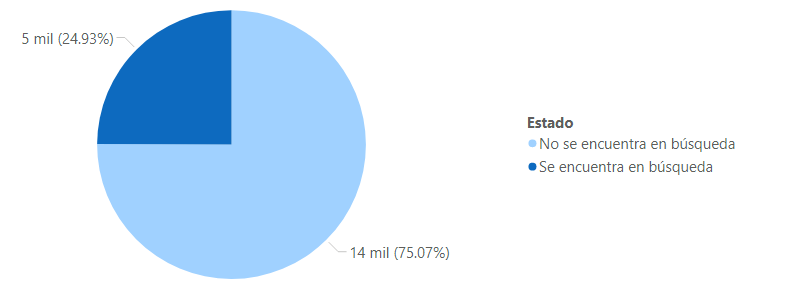
* Identificar cuáles son los factores que llevan a una persona a cambiar de empresa.
* Predecir la probabilidad de que los candidatos estén dispuestos a cambiar de puesto basado en los factores antes mencionados.
* Crear un modelo de Machine Learning capaz de predecir esta probabilidad.

# Descripción de los datos

* city: Codigo de ciudad. Variable categórica nominal
* city\_development\_index: Indice de desarollo de la ciudad. Variable numérica continua
* gender: Genero del candidato. Indice de desarollo de la ciudad. Variable categórica nominal
* relevent\_experience: Si el candidato tiene experiencia relevante. Variable categórica nominal
* enrolled\_university: Tipo de universidad en la cual se encuentra cursando. Variable categórica nominal
* education\_level: Nivel de educacion del candidato. Variable categórica ordinal
* major\_discipline: Disciplina. Variable categórica nominal
* experience: Años de experiencia del candidato. Variable numérica discreta
* company\_size: Tamaño de la compañia. Variable categórica ordinal
* company\_type: Sector al que se dedica la compañia. Variable categórica nominal
* last\_new\_job: Años de diferencia entre el ultimo trabajo y el nuevo. Variable numérica discreta
* training\_hours: Horas de entrenamiento completados. Variable numérica discreta
* target: Variable a supervisar. 1=En búsqueda de cambio de trabajo / 0=No se encuentra en búsqueda de cambio de trabajo

# Hallazgos encontrados en los datos

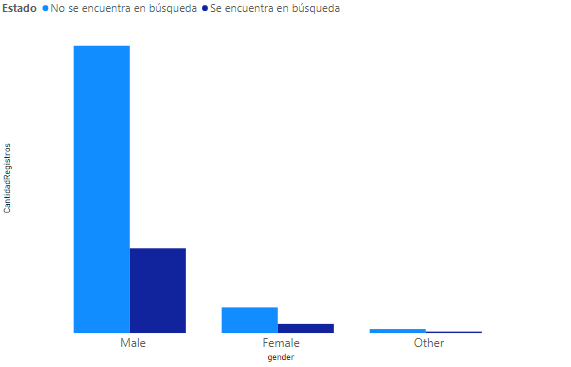
* Análisis de balanceo de datos: Se detecta que cerca del 75% de los datos a analizar, no se encuentran en búsqueda activa. Esto resulta importante al momento de analizar las métricas del modelo



* Análisis de inscripción universitaria según experiencia y variable target: Se observa que aquellas personas sin inscripción universitaria y con experiencia relevante no son propensos al cambio laboral



* Análisis por género: La distribución según género entre quienes están en búsqueda del cambio laboral y las que no, se visualiza con una relación similar



# Algoritmo elegido

Debido a que el modelo posee los datos necesarios para realizar un entrenamiento supervisado y de clasificación, se opta en primera instancia por Random Forest.

# Métricas de desempeño del modelo

Matriz de confusión para la evaluación del desempeño en las predicciones

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicciones** | |
|  |  | **VERDADERO** | **FALSO** |
| **Reales** | **VERDADERO** | 3871 | 410 |
| **FALSO** | 846 | 601 |

Métricas analizadas:

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Recall | Precision | F1 Score |
| 77.90% | 40.77% | 59.05% | 48.24% |

El accuracy (predicciones correctas sobre las totales) del modelo se considera con un valor aceptable.

Pero analizando el recall y precision (se calculan tomando como referencia el resultado de las predicciones para cada valor de la variable target) se corrobora que el modelo no es óptimo.

# Iteraciones de optimización

A partir del análisis realizado en la exploración de datos, se opta por aplicar oversampling con el objetivo de balancear los datos para optimizar el entrenamiento del mismo

# Métricas finales del modelo optimizado

Matriz de confusión:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicciones** | |
|  |  | **VERDADERO** | **FALSO** |
| **Reales** | **VERDADERO** | 3515 | 762 |
| **FALSO** | 741 | 3586 |

Métricas resultantes

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Recall | Precision | F1 Score |
| 83.00% | 83.00% | 82.00% | 83.00% |

# Conclusiones

Se ha logrado obtener un modelo capaz de predecir si un individuo se encuentra en búsqueda de un cambio de trabajo. A partir de esto se podría aplicar el mismo a futuras encuestas o incluso simular datos para identificar los atributos que deben reunir para focalizar la búsqueda considerando los mismos.