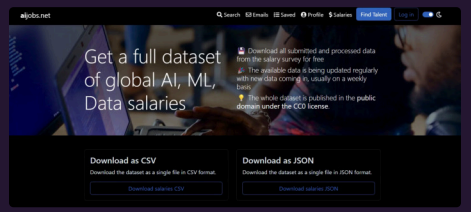
Modelo Predictivo de Machine Learning

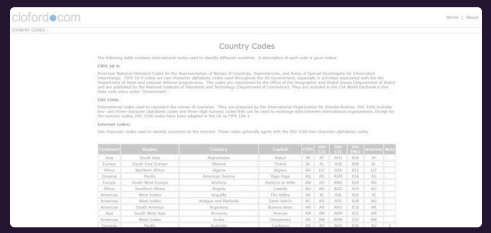
Salarios de puestos de trabajo relacionados con el Data Science

Tamara Acedo - Bootcamp de Data Science

Obtención y Tratamiento de Datos (I)

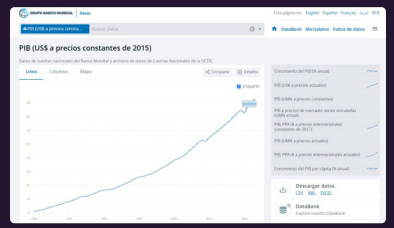


Salarios

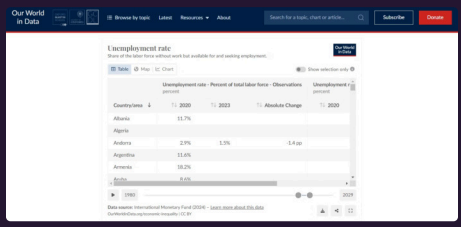


Códigos ISO

(Web Scrapping)

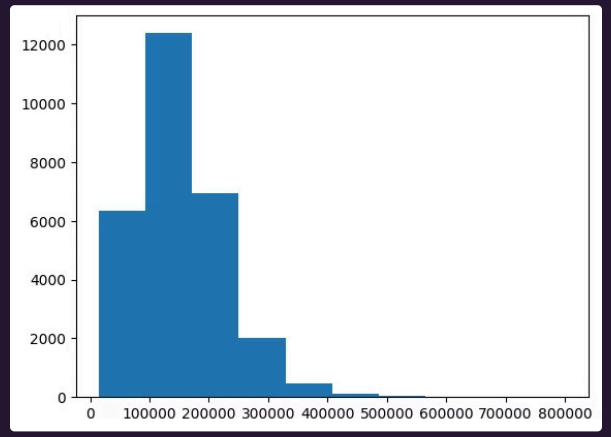


PIB



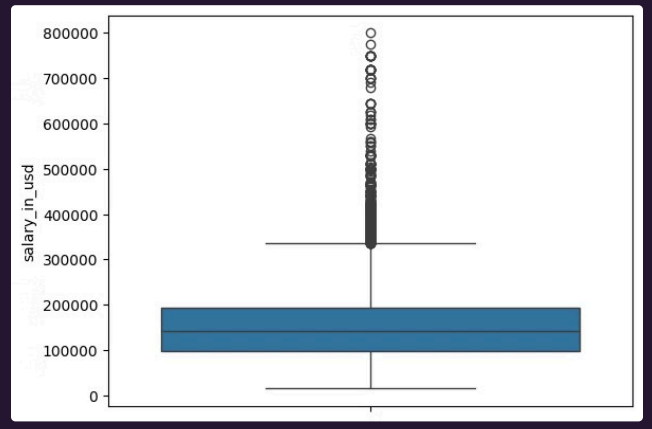
Tasa de desempleo

Obtención y Tratamiento de Datos (II)

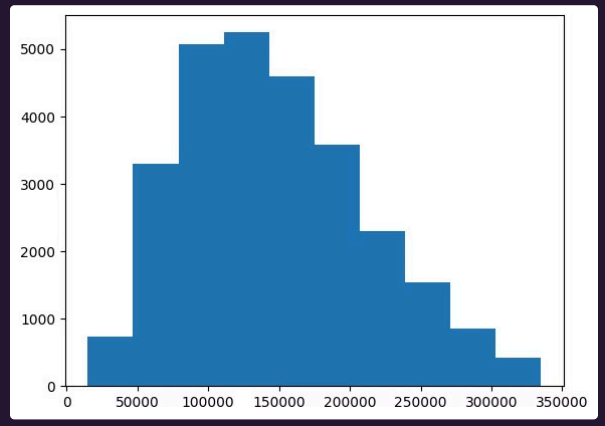


Distribución del target

ASIMETRÍA POSITIVA



Boxplot del target



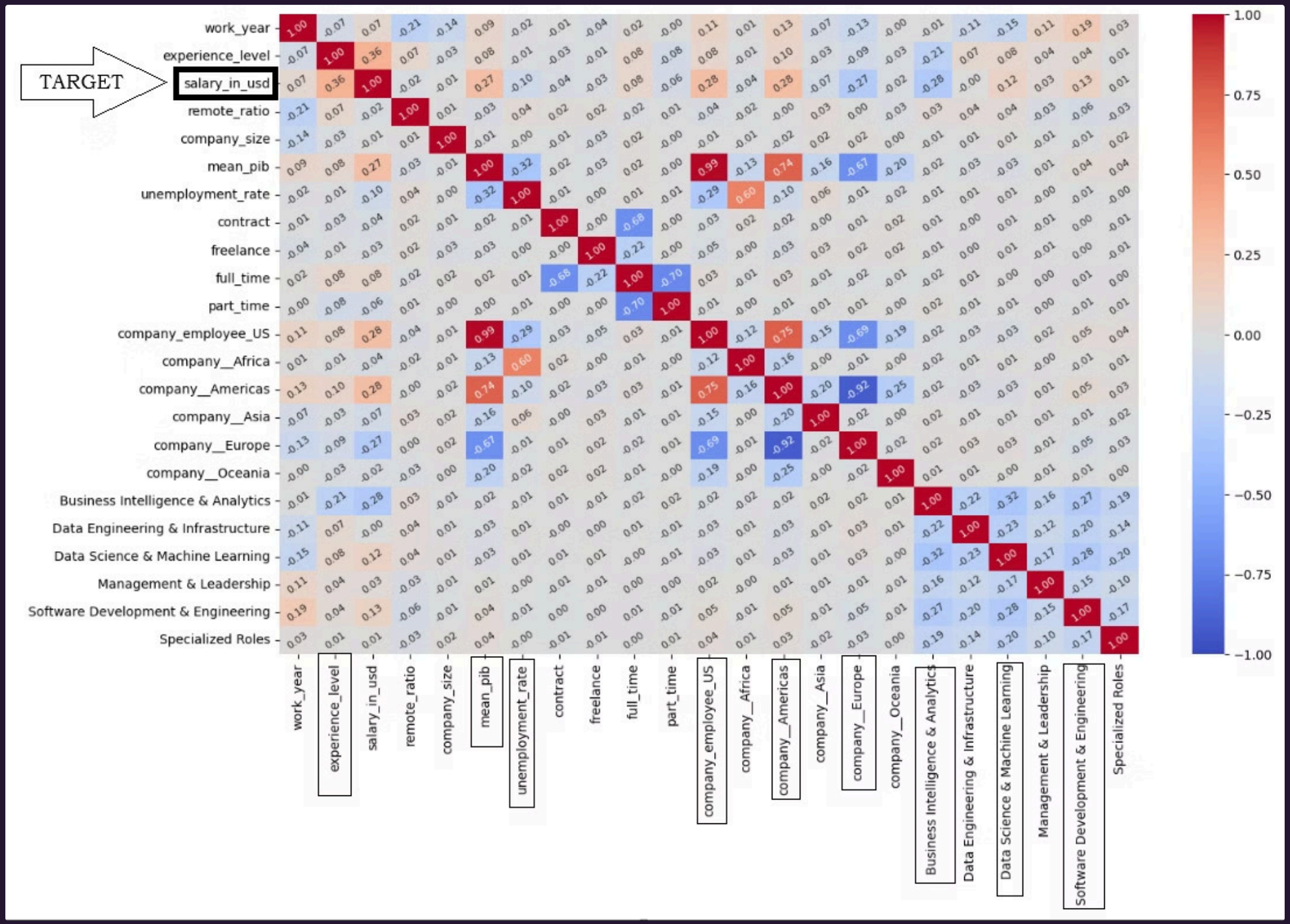
Distribución del target sin outliers

Posibles alternativas: Transformación logarítmica, elevar a 1/2 ...

Obtención y Tratamiento de Datos (II)

- **Eliminación de duplicados:** No mejoraban los resultados obtenidos y disminuía el volumen de los datos
- **Tratamiento de variables categóricas:**
  - Experiencia del empleado
  - Tamaño de la compañía
  - Tipo de empleo
  - Localización de la sede de la compañía y la residencia del empleado
  - Título del empleado: Agrupación por temática

Variables predictoras



Comparación y Selección del Mejor Modelo Supervisado

- **Modelos hiperparametrizados**: *Grid Search* y *Randomized Search*
- **Datos normalizados***: *StandardScaler* y *MinMaxScaler*

Modelo	R^2	MAE
Decision Tree Regressor	**0.265**	**42,927.276**
LightGBM	**0.281**	43,313.334
XGBoost	0.280	43,332.837
Random Forest Regressor	0.279	43,355.165
SVR	0.100	48,510.455

* *Excepto para los modelos basados en árboles de decisión*

Comparación y Selección del Mejor Modelo

DECISION TREE REGRESSOR

- **Profundidad máxima del árbol** (*max_depth*): Cuántas veces puede dividirse el árbol de decisiones (desde 2 hasta 7) ===> 7
- **Nº mínimo de datos en cada rama final** (*min_samples_leaf*): Número mínimo de datos que debe tener cada hoja (15, 20, 25) ===> 15
- **Criterio de división** (*criterion*): Usado el "absolute_error"

0.264576	$42927.28
R^2	MAE

Aplicación - Streamlit

1

Interfaz Intuitiva

Hemos desarrollado una aplicación web con Streamlit que permite a los usuarios interactuar fácilmente con el modelo predictivo.

2

Cálculo de Salarios

Los usuarios pueden ingresar sus datos de perfil y obtener una estimación de su salario

Próximos Pasos y Conclusiones

Próximos pasos:

Una vez entrenado nuestro modelo y en base a los resultados obtenidos, el objetivo es mejorar la capacidad predictiva del mismo disminuyendo el error absoluto medio (MAE) y aumentando el coeficiente de determinación (R^2).

Conclusiones:

A pesar de todos los análisis, ajustes de parámetros y esfuerzo depositado en seleccionar las mejores variables con técnicas de **feature engineering**, los resultados de los modelos no fueron del todo satisfactorios. Así pues, opino que no solo es crucial elegir y ajustar bien los modelos, sino que la **calidad de los datos con los que entrenamos y validamos** es igual o incluso más importante.