DEMANDA DE ENERGÍA EN ESPAÑA

CoderHouse Data Science

8 de Agosto, 2023

Weigandt, Herman

Tabla de contenido

1. Descripción del caso de negocio

El mercado de energía es uno de los mercados más dinámicos de los últimos tiempos por varios aspectos. Por un lado, siempre hablando del mercado mayorista, la energía producida no puede almacenarse por lo que su producción, comercialización y distribución se realiza en tiempo real día a día. Por otra parte, los distintos países y regiones han ido adaptando sus matrices energéticas en consideración de la conciencia ecológica (huella de carbono y energías renovables), la independencia energética y el crecimiento del consumo, como asimismo el cambio de preferencias de los consumidores tanto por los nuevos sistemas de climatización, el uso de dispositivos electrónicos y la movilidad sustentable.

Atendiendo a esto y en especial consideración de que la energía producida no puede acumularse a gran escala, se considera primordial la predicción de la cantidad a producir/cantidad demandada, para evitar desperdicios que llevaran a su encarecimiento, al igual que la subproducción que puede llevar a la sobreexigencia del sistema energético o la falta de su servicio, el cual en muchos casos es de vital importancia para la refrigeración de alimentos u otros casos como el funcionamiento de equipamiento indispensable como lo es el hospitalario.

Sobre lo anterior y destacando la regionalidad del mercado mayorista de energía eléctrica, se aborda en el presente trabajo el mercado español, considerando cuatro regiones y estudiando tanto variables climáticas como de producción y comercialización de la misma.

2. Tabla de versionado.

Sólo existe una versión.

3. Objetivos del modelo.

Con base en todas las variables que presenta el dataset relevado, se propone estudiar: ¿Cuáles son las razones (variables) que determinan la demanda de energía eléctrica?, para de ésta manera predecir la misma a efectos de cubrirla de la manera más eficiente posible.

Partiendo de esta pregunta objetivo, resulta de interés identificar la mínima cantidad de variables que la determinan y así encontrar un modelo de predicción que permita saber con la mayor anticipación posible, el comportamiento de nuestra variable principal.

4. Descripción de los datos

Este conjunto de datos contiene 4 años de consumo eléctrico, generación, precios y datos meteorológicos para España. Los datos de consumo y generación se recuperaron de ENTSOE, un portal público para datos de operadores de servicios de transmisión (TSO). Los precios de liquidación se obtuvieron del TSO español Red Electric España. Los datos meteorológicos se compraron como parte de un proyecto personal de Open Weather API para las 5 ciudades más grandes de España y se publicaron, como dominio público, en el sitio Kaggle y se puede acceder en el siguiente Link.

A continuación se provee de una breve descripción de las variables:

Dataframe clima:

```
#   Column  Descripción           Non-Null Count   Dtype
---  ------                        --------------   -----
 0   dt_iso   Fecha                 178396 non-null  object
 1   city_name Ciudad               178396 non-null  object
 2   temp      Temperatura (K°)      178396 non-null  float64
```

3	temp_min	Temp. minima (K°)	178396	non-null	float64
4	temp_max	Temp. máxima (K°)	178396	non-null	float64
5	pressure	Presión atmosférica	178396	non-null	int64
6	humidity	Humedad	178396	non-null	int64
7	wind_speed	Vel. del viento	178396	non-null	int64
8	wind_deg	Dir. Del viento	178396	non-null	int64
9	rain_1h	Lluvia en la últ hs	178396	non-null	float64
10	rain_3h	Lluvia en las últ. 3hs	178396	non-null	float64
11	snow_3h	Nieve en las últ. 3 hs	178396	non-null	float64
12	clouds_all	Niebla	178396	non-null	int64
13	weather_id	Codificador	178396	non-null	int64
14	weather main	Codificador	178396	non-null	object
15	weather_description	Descripción	178396	non-null	object
16	weather_icon	Ícono	178396	non-null	object

Dataframe producción y consumo:

#	Column	Descripción	Non-Null Count	Dtype
0	time		35064 non-null	object
1	generation biomass		35045 non-null	float64
2	generation fossil brown coal/lignite		35046 non-null	float64
3	generation fossil coal-derived gas		35046 non-null	float64
4	generation fossil gas		35046 non-null	float64
5	generation fossil hard coal		35046 non-null	float64
6	generation fossil oil		35045 non-null	float64
7	generation fossil oil shale		35046 non-null	float64
8	generation fossil peat		35046 non-null	float64
9	generation geothermal		35046 non-null	float64
10	generation hydro pumped storage aggregated		0 non-null	float64
11	generation hydro pumped storage consumption		35045 non-null	float64
12	generation hydro run-of-river and poundage		35045 non-null	float64
13	generation hydro water reservoir		35046 non-null	float64
14	generation marine		35045 non-null	float64
15	generation nuclear		35047 non-null	float64
16	generation other		35046 non-null	float64
17	generation other renewable		35046 non-null	float64
18	generation solar		35046 non-null	float64
19	generation waste		35045 non-null	float64
20	generation wind offshore		35046 non-null	float64
21	generation wind onshore		35046 non-null	float64
22	forecast solar day ahead		35064 non-null	float64
23	forecast wind offshore eday ahead		0 non-null	float64
24	forecast wind onshore day ahead		35064 non-null	float64
25	total load forecast	Cant. Pronosticada	35064 non-null	float64
26	total load actual	Cant. Real	35028 non-null	float64
27	price day ahead		35064 non-null	float64
28	price actual		35064 non-null	float64

En principio se realizó un análisis de todas las variables. Se detecto que las columnas “forecast wind offshore day ahead” y “generation hydro pumped storage aggregated” se encuentran vacías y que la columna de cantidad real consumida (26) presenta 36 datos faltantes y algunas otras presentar varios

valores iguales a 0, las cuales no se detallan en honor a la brevedad y que las mismas, según se explicará más adelante, no son relevantes.

El dataset contiene 46 variables con 35064 registros para el dataframe de producción y consumo y 178396 registros para el dataframe de clima de lo cual hay que destacar que respecto del clima hay 1 registro por cada ciudad relevada para un mismo momento, mientras que relativo al mercado de energía hay un registro para cada momento y la cantidad consumida no se encuentra discriminada por ciudad.

En el primer relevamiento de las variables, y en vistas de nuestro objetivo se realiza un primer descarte de variables por no considerarse pertinentes y se las excluye del EDA y de los modelos. Las variables excluidas son:

- Las variables del dataframe de producción y consumo, exceptuando las variables de cantidad consumida (total load actual) y de precio (Price actual).

Del total de las 46 variables, si quitamos las mencionadas arriba, nos quedamos con 19 variables, de las cuales se tienen :

- 8 variables categóricas
- 11 variables numéricas (cuantitativas)

Entre las categóricas hay algunas ordinales y otras nominales.

A continuación se detallan los principales resultados del EDA, de las demás variables estudiadas.

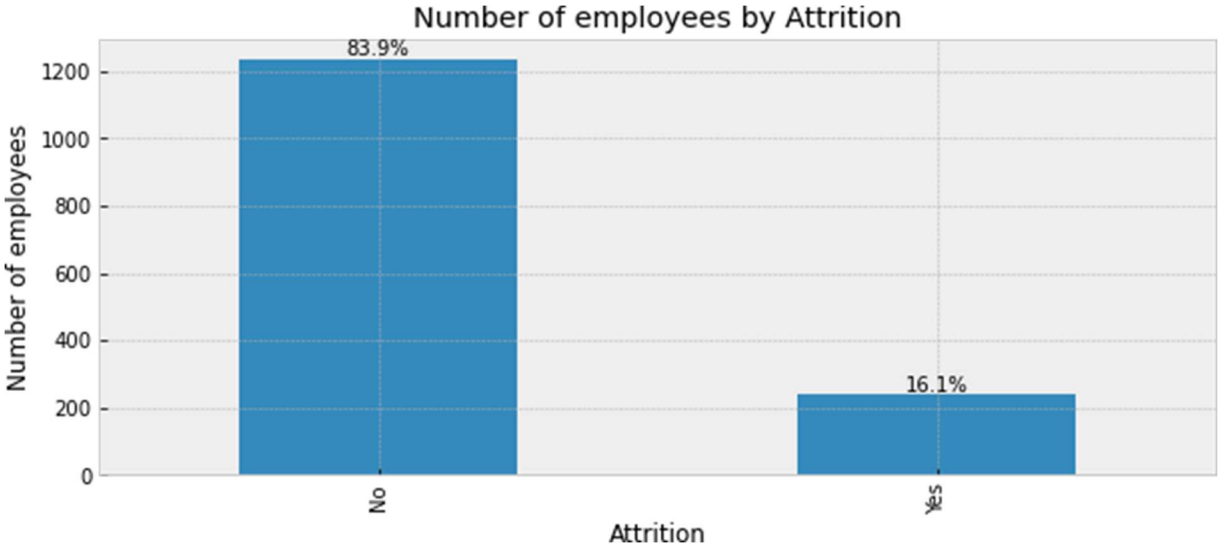
5. EDA: Exploratory Data Analysis

En el análisis exploratorio se estudiaron las distribuciones de las variables numéricas originales y de las variables categóricas.

Entre los resultados que se destacan, se observó que no hay diferencias importantes en las distribuciones de las variables ordinales debidas al género de los individuos, ni tampoco debidas al Attrition. El 60 % de los participantes del estudio son hombres y el 40% mujeres.

Entre las variables categóricas, la variable objetivo, Attrition, se pudo observar que la mayoría (el 83,9%) de empleados advirtieron algún grado de desgaste laboral, mientras que el 16,1% restante fueron los empleados que no presentan esta característica. Se muestra el desbalance del dataset respecto de esta variable en la imagen a continuación.

Figura 5.1 **Número de empleados por Attrition**



Entre los resultados del análisis exploratorio, se puede resumir que se cuenta con información de 3 departamentos de la empresa IBM, Research & Development, Sales y Human Resources. La distribución de edades es normal, con el 68% de los participantes del estudio en el rango de edades de 28.0 a 46.0, y una media de 37 años.

El 81% de los empleados raramente viaja o no lo hace directamente. Solo el 18,8 % viaja frecuentemente. Más del 70% tienen formación en ciencias naturales (Life Sciences) o medicina (Medical). El 30% restante poseen se dividen en Marketing, técnicos (Technical Degree), recursos humanos (Human Resources) y otros. Entre los roles laborales, se observaron 9 tipos distintos. El 71% de los empleados ha indicado que no realiza horas extras.

Los empleados casados representan casi la mitad del dataset (45.8%), el 54% restante se divide en 32% solteros y 22.2% divorciados.

Del análisis bivariado, no se observaron correlaciones que permitieran escribir una variable en función de otra. Si bien se observaron coeficientes de correlación altos, la relación entre los campos mostró gran variabilidad, por lo que no se consideró suficiente para descartar alguna variable adicional del análisis.

Al observar el ingreso mensual y el nivel de educación, se encontró que hay muy pocos empleados con alto nivel de estudio, en general poseen un nivel medio y no se advirtió una dependencia del ingreso respecto al nivel de estudio del empleado.

Tampoco se observó que el ingreso fuera una variable que determinará el grado de participación laboral del empleado. Respecto al Attrition, o desgaste laboral, se vió que aquellos que respondieron "Yes", poseen bajos ingresos en general, aunque existen algunos casos de ingresos altos.

La distancia desde la casa, no se observó como un factor determinante de Attrition.

Altos coeficientes de correlación no siempre implica una real correlación. Del análisis con coeficientes de correlación superior a 0,7; observamos que existen reales correlaciones entre las variables MonthlyIncome, JobLevel y TotalWorkingYears. Que tiene sentido cuando pensamos que el ingreso mensual suele ser mayor cuanto mayor es el nivel de responsabilidad en el trabajo, y para acceder a estos altos puestos de seniority también se requiere haber trabajado una cierta cantidad de años; cuantos más años de trabajo haya tenido una persona más chance tendrá de haber accedido a puestos de mayor responsabilidad y por ende, de mayor sueldo. A continuación aquellas variables que presentaron una fuerte correlación según lo antes explicado, acompañadas de su coeficiente de correlación:

Tabla 5.1 **Tabla de variables con índice de correlación mayor a 0,7.**

Corr Coef	Variable 1	Variable 2
0,95	MonthlyIncome	JobLevel
0,78	TotalWorkingYears	JobLevel
0,77	TotalWorkingYears	MonthlyIncome
0,77	PerformanceRating	PercentSalaryHike
0,77	YearsWithCurrManager	YearsAtCompany
0,76	YearsInCurrentRole	YearAtCompany
0,71	YearsWithCurrManager	YearsInCurrentRole
0,68	TotalWorkingYears	Age

El resto de las variables muestran una falsa correlación, dada por que cuanto mayor edad tienen los participantes mayor es la participación en la segunda variable analizada, en cualquier caso es una dependencia de la cantidad de años trabajados.

Por otro lado, a pesar de no encontrar correlación entre las variables EnvironmentSatisfaction, JobSatisfaction y RelationshipSatisfaction, conceptualemnte las primeras dos pueden resumirse en la última, RelationshipSatisfaction. Por este motivo se decidió sólo considerar RelationshipSatisfaction como variable del problema a analizar y descartar EnvironmentSatisfaction y JobSatisfaction.

Existen otras variables como las tasas de ingreso por hora, diario y mensual, la tasa de incremento salarial y el nivel de opciones sobre acciones, que las se excluyen del análisis por no presentar relevancia.

Finalmente, las variables utilizadas en los modelos de ML implementados fueron 26. Se muestran en la siguiente tabla:

1) Numéricas

Age	DistanceFromHome	Education	JobInvolvement	JobLevel	RelationshipSatisfaction
MonthlyIncome	NumCompaniesWorked	PerformanceRating	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance
YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager		

2) Categóricas (que debemos transformar en numérico para el análisis)

Attrition	BusinessTravel	Department	EducationField
Gender	JobRole	MaritalStatus	OverTime

6. Algoritmo Elegido.

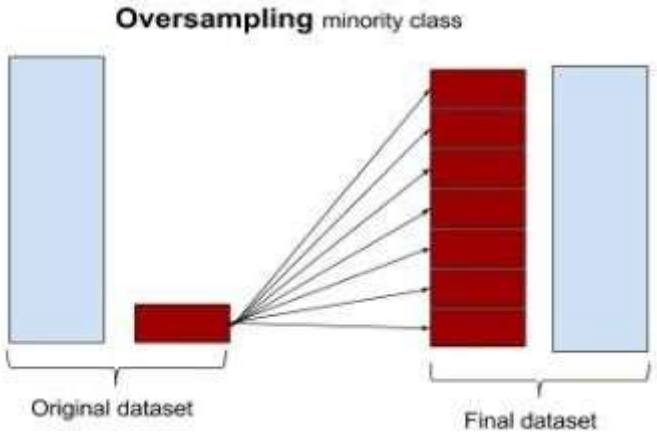
A fin de poder aplicar los modelos de ML, se generó una codificación a las variables categóricas, es decir, se les asignó un valor numérico, mediante el algoritmo de *labelencoder*.

Los modelos utilizados para predecir la variable Attrition fueron los siguientes:

1. Árbol de Decisión
2. Random Forest
3. Regresión Logística
4. KNN
5. XGBoost

En un primer lugar se testearon los modelos con el dataset completo, filtrando las variables que se mencionaron en la sección anterior, tomando una división de Train/Test con una relación de 70/30.

Luego del primer análisis se aplicó la técnica de **oversampling** al dataset, con el objetivo de compensar el desbalance de la variable, y disminuir la posibilidad de overfitting observada en el estudio de los primeros modelos (árbol y random forest).



Con la técnica de oversamplig se construyó una división entre Test/Train con el mismo porcentaje para los modelos, con la diferencia que ahora la cantidad de positivos y negativos en la variable objetivo es la misma.

Finalmente se utilizó una técnica de optimización de parámetros, sobre tres de los modelos implementados, el árbol de decisión, la regresión logística y el XGBoost.

Cabe aclarar que originalmente se construyeron los modelos buscando optimizar el resultado de la métrica Accuracy, que luego se entendió que no es la generalmente utilizada para el estudio de variables con desbalance como la que se plantea estudiar. Se buscó entonces aquel modelo que permita conocer la respuesta sobre la variable Attrition en el que se alcance la sensibilidad y especificidad más alta, esto significa maximizar el

área de la curva roc (Sencibilidad vs Especificidad o bien TruePositiveRate vs FalsePositiveRate).

Se detallan los modelos implementados:

1. Árbol de Decisión:

Se realizó un análisis de la métrica Accuracy respecto de la profundidad del árbol, a fin de conseguir una profundidad que no presente signos de overfitting, es decir, que el Accuracy de set test diverja del valor obtenido para set train.

Con el mismo criterio se buscó la profundidad del árbol, pero teniendo en cuenta el parámetro class_weight del modelo de árbol de decisión que permite indicar que queremos utilizar una clase balanceada (class_weight='balanced).

En el primer caso, se adoptó una profundidad max_depth = 4 (arbol_1), y en el segundo con el parámetro class_weight='balanced, se obtuvo una profundidad de arbol max_depth = 3 (arbol_2).

Ambos modelos fueron también entrenados con el set de datos con oversamplig y testeados con el set Test original. Los resultados se presentan con la denominación arbol_1b y arbol_2b respectivamente.

Finalmente el árbol de decisión se optimizó mediante la BÚSQUeda de los parámetros 'max_depth' (profundidad del árbol), 'criterion' (criterio), 'splitter' (división de selección), 'max_features' (máxima cantidad de variables para la decisión) y 'ccp_alpha' (parámetro de complejidad), de forma tal que maximice el área de la curva roc.

2. Random Forest:

Al igual que para los árboles de decisión, se realizó un análisis del accuracy para conocer la cantidad de árboles que se deberían aplicar en el modelo, y también se estudió la métrica con la cantidad de estimadores.

Primero se estudió la variación del accuracy con la cantidad de estimadores, a partir de lo cual se adoptaron 9 estimadores (n_estimadores = 9).

Al realizar el análisis de la métrica respecto a la profundidad del árbol (max_depth), se observó una rápida divergencia entre los resultados para Train y Test, por lo que se recurrió al parámetro class_weight, para tener en cuenta la característica desbalanceada de la variable. Finalmente, se adoptaron los parámetros: n_estimadores = 9, max_depth = 3.

Se construyeron dos modelos de random forest con estos parámetros y tomando en cuenta que la cantidad de parámetros elegidos por el forest para la selección de una hoja, sea seleccionada mediante la técnica log2 (max_features = "log2").

El primer forest, se denominó forest_1, y el SEGÚNDO forest_2, el cual sólo difiere del primero por haber tenido en cuenta el desbalance de la variable Attrition (class_weight="balanced").

No se ejecutó el módulo de optimización sobre este modelo, ni sobre el dataset con oversampling, ya que se consideró que posee un comportamiento similar al árbol y añade complejidad.

3. Regresión Logística:

Se aplicó el modelo de regresión logística con los parámetros de scikit-learn por defecto, al que denominamos *LogReg*. Este modelo permite estimar con cierta probabilidad la categoría de una variable binaria (en este caso el Attrition) en función de una serie de variables.

El modelo fue entrenado con el set de datos con oversamplig y testeado con el set Test original, los resultados se presentan con la denominación *LogReg_b*. También se ejecutó la optimización de los parámetros del modelo (*Regresión Logística optimizada*): 'penalty' (penalidad por mala clasificación), 'solver' (algoritmo de optimización), 'max_iter' (máxima cantidad de iteraciones para convergencia).

4. K-Nearest-Neighbor (knn):

Se buscó hacer la predicción de la variable Attrition por medio del comportamiento de los vecinos más cercanos, en el espacio multidimensional compuesto por otras 25 variables, usando los 7 vecinos mas cercanos, como métrica de distancia se adoptó la técnica de minkowski con parámetro p = 5.

No se ejecutó el modelo knn sobre el set con oversampling, ni tampoco se buscó optimización de parámetros.

5. XGBoost:

Se seleccionó también el modelo XGBoost, uno de los modelos más utilizados en la actualidad para objetivos de clasificación como el presente.

Entre sus parámetros se adoptó objectiv = "binary: logistic", es decir una técnica de regresión logística para clasificación binaría, ya que la variable a predecir es del tipo "yes/no". Se seleccionaron 10 estimadores, del mismo orden que los random forest aplicados, y una profundidad de 6 niveles y una parámetro de aprendizaje tìpico learning_rate = 0.01. Los resultados se presentan con la denominación *XGboost*.

Con los mismos parámetros se ejecutó el modelo XGBoost sobre el set Train con oversampling y luego sobre el set Test original (*XGboost_b*).

Finalmente, se buscó la optimización de los parámetros (*XGboost_sg*): max_depth' (máxima profundidad), 'n_estimators' (cantidad de estimadores), 'learning_rate' (tasa de aprendizaje), 'gamma' (reducción mínima requerida para hacer una división)

Dados los resultados de las métricas, que se presentan en la siguiente sección, el algoritmo que mejor performance tuvo respecto al área bajo la curva roc, es la **Regresión Logística con optimización de parámetros**.

7. Métricas de Desempeño del Modelo.

A continuación se presentan los resultados obtenidos

- **Modelos con división del dataset Test/Train en 70/30**:

Métrica \ Modelo	arbol_1	arbol_2	forest_1	forest_2	LogReg	knn	XGboost
Accuracy	0.861678	0.809524	0.863946	0.759637	0.843537	0.836735	0.836735
Precision	0.500000	0.350650	1.000000	0.297300	0.277780	0.176470	0.351350
Recall	0.114750	0.442620	0.016390	0.016390	0.081970	0.049180	0.213110
ROC_curve	0.691390	0.696270	0.700630	0.707940	0.686500	0.510760	0.710960

- **Modelos con dataset considerando la técnica Oversampling**:

Métrica \ Modelo	arbol_1b	arbol_2b	forest_1b	forest_2b	LogRegb	knn_b	XGboost_b
Accuracy	0.646259	0.811791	NaN	NaN	0.673469	NaN	0.786848
Precision	0.176870	0.333330	NaN	NaN	0.221480	NaN	0.296300
Recall	0.426230	0.360660	NaN	NaN	0.540980	NaN	0.393440
ROC_curve	0.617470	0.641780	NaN	NaN	0.641780	NaN	0.620530

- **Modelos optimizados**:

Métrica \ Modelo	arbol_ optimizado	Regresión Logística optimizada	XGboost_ optimizado
Accuracy	0.82337	0.866848	0.847826
Precision	0.22222	0.666670	0.600000
Recall	0.03333	0.366670	0.200000
ROC_curve	0.66523	0.796750	0.795830

8. Iteraciones de Optimización.

No se realizaron iteraciones sobre los modelos optimizados.

9. Métricas finales del Modelo Optimizado.

El modelo considerado óptimo SEGÚN los resultados de las métricas obtenidas, fue la Regresión Logística, para la cual se obtuvo:

Métrica \ Modelo	Regresión Logística optimizada
Accuracy	0.866848
Precision	0.666670
Recall	0.366670
ROC_curve	0.796750

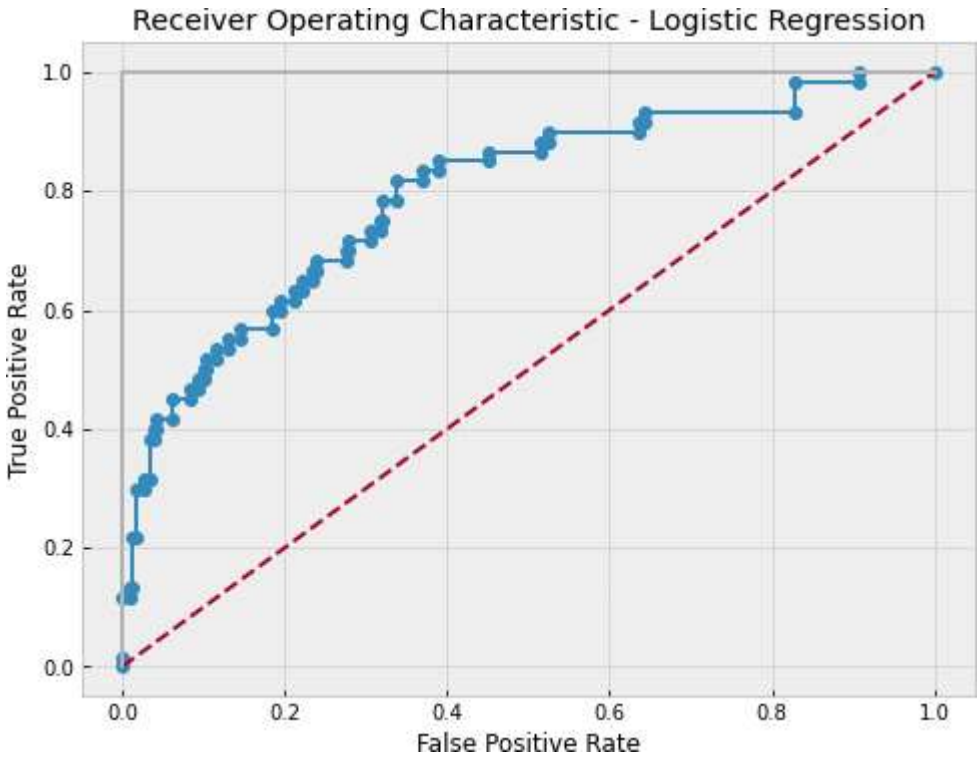


Figura 9.1 **Curva roc que resulta del modelo de regresión logística optimizado.**

10. Futuras líneas

En el presente trabajo, no se realizaron iteraciones sobre la optimización de los modelos. A futuro, pueden ejecutarse las mismas y evaluar si se consigue obtener una mejor respuesta tanto de la métrica del área de la curva roc, como del Accuracy, Precision y Recall.

Pueden evaluarse también otras métricas, en particular para el modelo de Regresión Logística.

También, queda como futuras propuesta ejecutar la optimización del modelo knn, y random forest, para completar la comparativa entre modelos.

Resulta fundamental ampliar el dataset, si fuera posible obtener mayor cantidad de registros serviría para un mejor entrenamiento de los datos, ya que resulta bastante pequeño y particularmente afectado por el desbalance de clases de la variable a categorizar.

Una vez que se tiene un modelo definido, y óptimo, permitiría utilizarlo para predecir o bien clasificar a aquellos empleados que estén sufriendo desgaste laboral y poder actuar antes de una posible baja laboral, que puede representar costos de formación en la empresa.

11. Conclusiones:

De la descripción de los datos disponibles podemos conocer la composición de los empleados de la empresa IBM. Los principales resultados que los describen se destacan: el 68% de los empleados son personas entre 28.0 y 46.0 años, que por lo general viven cerca del trabajo, poseen entre 5 y 10 años en la empresa y por lo general tiene el resto de las métricas tienen distribuciones similares para hombres y mujeres.

La variable objetivo, desgaste laboral (Attrition), se encuentra desbalanceada, es decir existen muchos casos en un respuesta y muy pocos en otra, a saber el 83,9% de empleados advirtieron algún grado de desgaste laboral, y respondieron 'Si', mientras que el 16,1% restante respondieron 'No'.

El dataset recibido para el análisis propuesto contaba con 34 variables, sólo 26 se utilizaron en los 5 modelos de Machine Learning (ML). Los modelos implementados fueron el Árbol de Decisión, el Random Forest, la Regresión Logística, el KNN y el XGBoost.

Se utilizó una subdivisión 70/30 del dataset, para los set de Train/Test respectivamente, se ejecutaron los modelos también utilizando la técnica de Oversampling y finalmente se seleccionaron tres modelos para ser optimizados.

La optimización del modelo de Regresión Logística fue el algoritmo que consiguió el mayor valor de área bajo la curva roc, métrica que se seleccionó como objetivo para seleccionar el modelo que mejor puede predecir si un empleado sufre o no, desgaste laboral. El modelo obtuvo una exactitud (accuracy) del 86,68%, una precisión (precisión) o proporción entre el número de predicciones correctas respecto al total del 66,67%, una sensibilidad (Recall), casos positivos correctamente identificados del 36,67% y un área de la curva roc igual a 0,7968.