DESARROLLO DEL CASO DE ESTUDIO DE LA AGENCIA DE MARKETING STERLING COOPER ADVERTISING

NATALIA GUZMAN
SUSANA BARRIENTOS
JAVIER BURGOS
FABIO ANDRES GONZALEZ

ANALÍTICA PARA LA TOMA DE DECISIONES

FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN
2023

Introducción

La selección de variables es un componente fundamental en el proceso de construcción de modelos de machine learning, ya que tiene un impacto significativo en la eficiencia y eficacia de los modelos predictivos. En este trabajo, exploramos la importancia de la selección de variables al abordar un problema de clasificación binaria en el contexto de recursos humanos: la predicción de la rotación de empleados (Attrition). Utilizamos un conjunto de datos que contiene información relevante sobre los empleados, incluidas características como la satisfacción laboral, la edad, el nivel educativo, el salario mensual y más.

El objetivo de este trabajo es comparar dos enfoques diferentes para la construcción de modelos predictivos. En primer lugar, aplicamos un proceso de selección de variables para identificar las características más relevantes para la predicción de la rotación de empleados. Luego, entrenamos un modelo de clasificación utilizando solo estas características seleccionadas. En segundo lugar, entrenamos un modelo de clasificación sin ningún proceso de selección de variables, utilizando todo el conjunto de datos disponible. Compararemos el rendimiento de estos dos enfoques en términos de exactitud y otras métricas de evaluación.

Este estudio tiene como objetivo proporcionar información valiosa sobre la importancia de la selección de variables en la construcción de modelos de machine learning y cómo puede afectar el rendimiento de los modelos en un escenario empresarial crítico como la retención de empleados. Los resultados de esta investigación pueden ser aplicables en la toma de decisiones estratégicas de recursos humanos y ayudar a las organizaciones a desarrollar estrategias efectivas para reducir la rotación de empleados y retener talento clave.

TABLA DE CONTENIDO

1. Exploración de datos
    a. Variables categóricas
    b. Variables numéricas
    c. Definición de variables con las cuales se continuará el proceso
2. Modelos
3. Resultados
4. Conclusiones

1. Exploración de datos.

Para comprender mejor la naturaleza de los datos, identificar patrones y detectar valores atípicos se procede a realizar un análisis exploratorio de la información suministrada.

Inicialmente se evidencia que la base de datos contiene 4.410 registros por variable, con 30 columnas, repartidas en 4 conjuntos de datos. El punto de partida para la unión de las bases de datos es el ID del empleado y la imputación para el tema de nulos se realiza mediante el reemplazo de datos NA por la mediana de la variable respectiva.

Attrition	Employee
Yes	711
No	3699
Total	4410

Tabla 1 Cantidad de registros

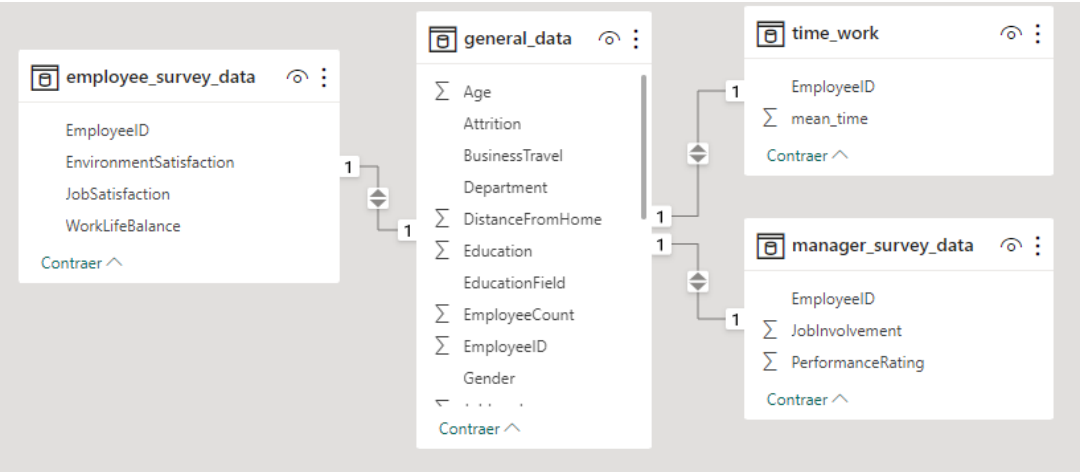

Figura 1. Diagrama de relación

Mediante el análisis de correlación entre variables y gráficas de asociación se realiza limpieza de variables poco aportantes al modelo base, terminando con una base de datos de 8 variables y 4.410 registros.

a. Variables categóricas:

Con la finalidad de tener una apreciación completa de la relación tres variables categóricas respecto a la variable objetivo, se realizó un diagrama de barras y una matriz de correlación, de dónde se determina que variables como **WorkLifeBalance** y **BusinessTravel** tienen -bajo nuestro criterio- una correlación fuerte, ello sumado a que la distribución de la variable objetivo (Renuncia o no) se acerca más a una distribución uniforme en **WorkLifeBalance** decidimos elegir **BusinessTravel**

De forma similar ocurre con las gráficas MaritalStatus y **PerformanceRaiting** en este caso la variable que consideramos que aportaría más es MaritalStatus. Por su parte la variable

Gender tiene una distribución uniforme respecto a la variable objetivo, es por ello que está no aportaría significativamente al modelo.

b. Variables numéricas:

En el análisis de las variables numéricas se usó gráficos dispersión y una matriz de correlación, ellos nos conducen a descartar las variables como **mean_time**, **DistanceFromHome**, **PercentSalaryHike** las cuales tienen una distribución uniforme y las variables Age, **YearAtCompany** por fuerte correlación con **TotalWorkingYears** y YearsWithCurrManager respectivamente.

NOTAN: Las variables que planteamos descartar tras su correspondiente análisis NO serán eliminadas salvo **mea_time,** con el objetivo de comparar sí luego de la evaluación de variables los modelos las descartan o no, y tras ello realizar el cambio si se considera necesario.

c. Definición de variables con las cuales se continuará el proceso

#	*Column*	*Non-Null Count*	*Dtype*
0	Age	4410 non-null	int64
1	Attrition	4410 non-null	object
2	BusinessTravel	4410 non-null	object
3	Department	4410 non-null	object
4	DistanceFromHome	4410 non-null	int64
5	Education	4410 non-null	int64
6	EducationField	4410 non-null	object
7	EmployeeCount	4410 non-null	int64
8	EmployeeID	4410 non-null	int64
9	Gender	4410 non-null	object
10	JobLevel	4410 non-null	int64
11	JobRole	4410 non-null	object
12	MaritalStatus	4410 non-null	object
13	MonthlyIncome	4410 non-null	int64
14	NumCompaniesWorked	4391 non-null	float64
15	Over18	4410 non-null	object
16	PercentSalaryHike	4410 non-null	int64
17	StandardHours	4410 non-null	int64
18	StockOptionLevel	4410 non-null	int64
19	TotalWorkingYears	4401 non-null	float64
20	TrainingTimesLastYear	4410 non-null	int64
21	YearsAtCompany	4410 non-null	int64

22	YearsSinceLastPromotion	4410 non-null	int64
23	YearsWithCurrManager	4410 non-null	int64

Tabla 2. Tabla resumen de variables finales.

2. Modelos

Diseño de solución propuesto
Inicialmente identificamos las variables categóricas del modelo ya que los modelos de regresión están diseñados para trabajar con variables numéricas. Después de esto, hicimos la codificación Dummy, ya que con este método se convierte las variables categóricas en múltiples variables binarias (0 o 1), una para cada categoría. Cada variable binaria representa la presencia o ausencia de una categoría específica.

Seguidamente, hicimos selección de variables, utilizando el método K-Best puesto que su objetivo es seleccionar las mejores características (variables o atributos) de un conjunto de datos para mejorar el rendimiento de un modelo de aprendizaje automático. Utilizamos este método con el criterio ANOVA, nos basamos en el análisis de la varianza para determinar si las variables eran significativas para este modelo. Después de aplicar el K-Best obtuvimos que estas variables eran las más significativas para el modelo: “Age”, “TotalWorkingYears”, “YearsWithCurrManager”, “mean_time”, “MaritalStatus_Single”.

	Age	***TotalWorkingYears***	***YearsWithCurrManager***	***mean_time***	***MaritalStatus_Single***
0	51	1.0	0	68.702	0
1	31	6.0	4	73.160	1
2	32	5.0	3	68.161	0
3	38	13.0	5	67.892	0
4	32	9.0	4	78.776	1

Tabla 3. Variables posteriores a la selección.

Después de seleccionar las variables significativas para el modelo, procedemos a aplicar técnicas de modelado para entrenar el modelo. Inicialmente utilizamos la regresión logística para este problema, ya que era un problema de clasificación.

Posteriormente, aplicamos **GradientBoostingClassifier**, ya que éste método está diseñado para resolver problemas de clasificación mediante la construcción de un conjunto de árboles de decisión, que se combinan de manera secuencial para mejorar la precisión predictiva del modelo. Así pues, entrenamos el modelo y nos dió un accuracy de 88.01% para el entrenamiento y 86.05% para la prueba.

Además, aplicamos **RandomForestClassifier**, ya que al igual que el método anterior, usa árboles de decisión para entrenar el modelo. A diferencia de **GradientBoostingClassifier**, cada árbol se entrena de manera independiente para clasificar los datos en función de las

características disponibles en su conjunto de datos. Cuando se realiza una predicción, todos los árboles en el conjunto emiten una predicción y la clase que obtiene la mayoría de votos (en problemas de clasificación) se considera la predicción final del modelo. Después de haber entrenado el modelo con este método, nos resultó un accuracy de 95,18% para el entrenamiento y 86.36% para la prueba.

3. Resultados

Metricas confusion_matrix
Precision: 0.7333333333333333
Recall: 0.07482993197278912
Especificidad: 0.9945578231292517
F1 score: 0.13580246913580246

733 Verdaderos Negativos (personas que dijeron que no renunciarían y no lo hicieron).
145 Verdaderos Positivos (personas que dijeron que renunciarían y lo hicieron).
22 Falsos Positivos (personas que dijeron que no renunciarían pero renunciaron).
0 Falsos Negativos (ninguna persona dijo que renunciaría).

El valor de precisión es 0.7333, lo que significa que aproximadamente el 73.33% de las predicciones positivas son correctas.

El valor de recall es bajo, 0.0748, lo que indica que el modelo puede tener dificultades para detectar correctamente los casos de renuncia.

Metricas regresion logistica sin selección de variables

LogisticRegression	**RandomForestRegressor**
Accuracy (Train): 84.89%	Accuracy (Train): 94.51%
MSE entrenamiento: 0.15	MSE entrenamiento: 0.0074282312925170076
MAE entrenamiento: 0.15	MAE entrenamiento: 0.048582766439909296
R2 entrenamiento: 0.12	R2 entrenamiento: 0.9451635230062987
Accuracy (Test): 14.220607806319164%	Accuracy (Test): 56.38320441037127%
MSE prueba: 0.11520812670373928	MSE prueba: 0.05858061224489795
MAE prueba: 0.24830583345103788	MAE prueba: 0.14264172335600908
R2 prueba: 0.14220607806319163	R2 prueba: 0.5638320441037127

RandomForestRegressor supera significativamente a LogisticRegression en términos de rendimiento en el conjunto de entrenamiento y mantiene un mejor rendimiento en el conjunto de prueba en términos de exactitud, MSE, MAE y R2. Sin embargo, es importante tener en cuenta que el rendimiento en el conjunto de prueba para ambos modelos es mejor que en el conjunto de entrenamiento, lo que puede indicar un desequilibrio en los datos o la necesidad de una mayor optimización de los hiperparámetros.

Metricas regresion logistica con selección de variables (SelectKBest)

Exactitud en el entrenamiento: 0.848								
Exactitud en el conjunto de validacion: 0.841								
RandomForestClassifier					GradientBoostingClassifier			
Train - Accuracy : 0.858843537414966					Train - Accuracy : 0.8801020408163265			
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.99	0.92	2962	0	0.89	0.98	0.93	2962
1	0.79	0.16	0.27	566	1	0.80	0.33	0.47	566
Test - Accuracy : 0.8514739229024944					Test - Accuracy : 0.8605442176870748			
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.99	0.92	737	0	0.87	0.97	0.92	737
1	0.73	0.15	0.25	145	1	0.68	0.28	0.40	145

En términos de exactitud, el modelo GradientBoostingClassifier tiene el mejor rendimiento tanto en el conjunto de entrenamiento como en el de validación.

4. Conclusiones

- Se identificaron las variables más relevantes para predecir la rotación de empleados. Estas variables pueden incluir aspectos como la satisfacción laboral, el nivel educativo, el salario mensual y la cantidad de años trabajados en la empresa. Comprender estas variables es crucial para tomar medidas preventivas.

- Al realizar la comparación entre los modelos de entrenamiento y validación con y sin selección de variables, se evidencia que el modelo que se apoya de la selección de variables cuenta con mejores métricas. Esto indica que la selección de variables ha ayudado al modelo a mejorar su capacidad predictiva y a reducir la complejidad al enfocarse en las características más relevantes para el problema en cuestión.

- Se evaluaron las métricas de rendimiento de los modelos, como la exactitud, el MSE (Mean Squared Error), el MAE (Mean Absolute Error) y el R2 (Coeficiente de Determinación). Estas métricas proporcionan información sobre qué tan bien se desempeñan los modelos en términos de predicción.

- El mejor método para el entrenamiento del modelo fue GradientBoostingClassifier.

- Es importante destacar que el modelo RandomForestClassifier superó significativamente a la Regresión Logística en términos de rendimiento en el conjunto de entrenamiento y en el conjunto de prueba. Esto sugiere que el uso de árboles de decisión puede ser más efectivo en este problema de clasificación.

- La selección de variables es un paso crucial en la construcción de modelos de machine learning. En este estudio, se demostró que al seleccionar cuidadosamente las variables más relevantes, se puede mejorar significativamente el rendimiento de los modelos de clasificación.