Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Actividad 7

Regresión Logística

Integrantes:

Rodrigo Ruiz Teodoro- A01730322

Natalia Cedillo Hernández - A01660022

Elena Nivón Hernández - A01174666

Jarlyn Loza Pacheco - A0176943

José Jaime Ponce de León - A01552256

Profesores:

Rigoberto Cerino Jiménez

Candy Yuridia Alemán Muñoz

Juan Manuel Ahuactzin

Alfredo García Suárez

18 / 10 / 2023

Procesamiento de Nulos y Outliers

Para la parte de preprocesamiento de los datos se inició con la detección de valores nulos y outliers. Donde, en el caso de los valores nulos se aplicaron conteos de NAs por dataframe y con los outliers se aplicó la identificación de los mismos por rango intercuartílico. Para ambos casos no se identificaron valores de este tipo en la base de datos, lo que implica que el conjunto no requería ningún proceso de limpieza o tratamiento adicional.

Correlación Logística

Comparación de resultados

A continuación, se presenta una tabla que refleja las métricas de diferentes combinaciones de variables independientes para predecir resultados en cada una de las variables independientes. Se puede observar cómo las variables independientes afectan los valores de las predicciones, reflejados en las métricas utilizadas (precisión, exactitud, sensibilidad y F1). Por tal motivo, fue importante tomar en consideración la selección de variables independientes en función de la mejora del rendimiento de los distintos modelos.

Filtros aplicados por variable.

Profession.

```
# Aplicar la condición y asignar etiquetas
base1=df[(df["Profession"]=="Mechanical_engineer") | (df["Profession"]=="Software_Developer")]
base1
```

State.

```
# Aplicar la condición y asignar etiquetas
base2=df[(df["STATE"]=="Andhra_Pradesh") | (df["STATE"]=="Maharashtra")]
base2
```

Income.

```
# Aplicar la condición y asignar etiquetas
media_ingreso = df['Income'].mean()
print(media_ingreso)
df['Ingreso Alto/Bajo'] = df['Income'].apply(lambda x: 'Ingreso Alto' if x > media_ingreso else 'Ingreso Bajo')
df.head(5)
```

Experience.

```
media = df['Experience'].mean()

#Dividir la columna en dos basándonos en la mediana
df['Experience'] = df['Experience'].apply(lambda x: 'exp_alta' if x > media else 'exp_baja')
df
```

```
#Reemplazar 'exp_alta' por 1 y 'exp_baja' por 0 en la columna 'experience'
df['Experience'] = df['Experience'].replace({'exp_alta': 1, 'exp_baja': 0}).astype(float)
df
```

Current Job Yrs.

```
mediana = df['CURRENT_JOB_YRS'].median()

# Dividir la columna en dos basándonos en la mediana
df['Tiempo_trabajo'] = df['CURRENT_JOB_YRS'].apply(lambda x: 'altos' if x > mediana else 'bajos')
df
```

Tabla 1. Comparación de modelos

Variables Dependientes	Variables Independientes	Precision Score	Accuracy Score	Recall	F1 Score
Profession 1	Income, Age, Experience	0.5757	0.5242	0.5259	0.5078
Profession 2	Current_Job_Yrs, Age, Experience, Current_House_Yrs	0.559352 51798561 15	0.5494969 165855241	0.50836 120401 33779	0.522696 01100412 65
State 1	Income, Experience	0.504553 44304201 95	0.5074714 903657098	0.59653 916211 29326	0.549595 44501048 85
State 2	Income,	0.519595	0.5184821	0.59668	0.554376

	Current_Job_Yrs, Age	6454121307	077467558	3640161901	1751683144
Income 1	Experience, Age	0.5046	0.5046	0.5351	0.4888
Income 2	Experience, Age, Current_Job_Yrs, Current_House_Yrs	0.4986	0.5018	0.4366	0.5310
Experience 1	Risk_Flag, Age, Income, Current_House_Yrs	0.5217	0.5215	0.0567	0.6738
Experience 2	Age, Income, Current_House_Yrs	0.5209	0.5202	0.0240	0.0458
Current_Job_Yrs 1	Age, Experience	0.6984	0.6984	0.6502	0.7306
Current_Job_Yrs 2	Age, Experience, Current_House_Yrs, Risk_Flag	0.6596	0.6948	0.6450	0.7282

Profession 1

Para el primer modelo respectivo a la variable dependiente de profesión, se analizó creando un filtro donde se tomarán en cuenta las instancias "Mechanical_engineer" y "Software_Developer", con las variables independientes de Income, Age, Experience, las cuales son numéricas. Con esto, el modelo arrojó un nivel de precisión significativo (0.5757), esto indica la capacidad del modelo para realizar predicciones sólidas en relación a los resultados deseados. Además, la exactitud, se encuentra en un rango moderado, con un valor de 0.5242 lo cual indica que, el modelo mantiene un nivel aceptable de precisión en las predicciones en general. Por otro lado, los valores de recall y F1, sugieren la habilidad del modelo para predecir una cantidad de datos positivos así como un equilibrio entre la precisión

y la sensibilidad de las predicciones por lo que, este modelo puede considerarse efectivo.

Profession 2

Para el primer modelo respectivo a la variable dependiente de profesión, al igual que con el primero, se analizó creando un filtro donde se tomarán en cuenta las instancias "Mechanical_engineer" y "Software_Developer", con las variables independientes de Current_Job_Yrs, Age, Experience, Current_House_Yrs, las cuales son numéricas. Se pudo observar que los resultados fueron bastante similares, ya que, de acuerdo a los resultados, el modelo demuestra un nivel de precisión sólido y con capacidad para realizar predicciones con una buena exactitud. En este caso de la exactitud, el modelo mantiene un nivel de precisión moderado. En este caso el recall, aunque no es óptimo, muestra algo de eficacia y por parte del F1, refleja la capacidad del modelo para destacar casos positivos. Si bien este modelo puede ser clasificado como satisfactorio, es verdad que puede presentar mejoras en la identificación de casos positivos por lo que, es importante considerar las métricas de rendimiento al momento de evaluar un modelo.

State 1

Para la variable dependiente de State, la cual fue filtrada para encontrar a los estados de "Andhra_Pradesh" y "Maharashtra", con las variables independientes, Income y Experience. De acuerdo al análisis se puede encontrar que, la precisión es moderada en las predicciones, que la exactitud puede considerarse aceptable en términos generales y que el recall es lo que destaca particularmente en términos de análisis en comparación de las otras métricas. Además, el F1 demuestra un equilibrio entre la precisión y la capacidad de detectar casos positivos lo cuál puede ser relevante en aplicaciones donde la detección precisa es prioritaria por lo que ayuda a detectar casos relevantes.

State 2

En el caso de State 2, el cuál fue filtrado de la misma manera que el State 1, con la única variable independiente de Income, Current_Job_Yrs y Age, el modelo demuestra una precisión moderada, indicando la capacidad de realizar predicciones con un nivel razonable de exactitud. La métrica de exactitud sugiere que el modelo mantiene una precisión aceptable en términos generales. Sin embargo, es la destacada sensibilidad la que resalta en este análisis, sugiriendo que el modelo es capaz de identificar correctamente la gran mayoría de

los casos positivos, lo que es esencial en muchas aplicaciones. Además, el puntaje F1, muestra un equilibrio efectivo entre la precisión y la capacidad del modelo para detectar casos positivos. Este modelo presenta un rendimiento sólido, especialmente en la detección de casos positivos, lo que puede ser crucial en escenarios donde la identificación precisa es fundamental.

Income 1

Para la variable dependiente de Income 1, con las variables independientes Experience, Age. El modelo muestra una precisión moderada, lo que sugiere que es capaz de realizar predicciones con un nivel aceptable de exactitud. La métrica de exactitud, indica que el modelo mantiene un nivel de precisión moderado en general. No obstante, es relevante destacar que la sensibilidad la cual indica que el modelo puede identificar adecuadamente una proporción considerable de casos positivos. A pesar de la moderada precisión, el puntaje F1, señala un equilibrio razonable entre la precisión y la capacidad del modelo para detectar casos positivos. En conjunto, el análisis de "Income 1" resalta la efectividad del modelo en la identificación de casos positivos, lo que puede ser crucial en aplicaciones donde la detección precisa es fundamental.

Income 2

Por otro lado, para la variable dependiente de Income 1, con las variables independientes Age, Income y Current_House_Yrs. Se observa que el modelo presenta una precisión moderada, lo que sugiere su capacidad para realizar predicciones con un nivel razonable de exactitud. La métrica de exactitud indica un nivel de precisión moderado en general.Por otro lado, la sensibilidad indica que el modelo puede identificar algunos casos positivos de manera adecuada, pero no logra capturar todos. Por último, se destaca el puntaje de F1, que muestra un equilibrio efectivo entre la precisión y la capacidad del modelo para detectar casos positivos. En conjunto, este análisis sugiere que una eficacia razonable en la identificación de casos positivos, aunque hay margen para mejorar la sensibilidad. La elección de las variables independientes y la optimización del modelo podrían ser áreas clave para mejorar aún más su rendimiento (como el caso de Income 1).

Experience 1

Para iniciar los modelos usando la variables dependiente de “Experience” primero se sacó la

media de años de experiencia de los usuarios, en caso de que la variable fuera mayor a la media se le colocó “alto”, en caso de ser lo contrario se colocaría “bajo”, de esta manera la variable se haría de tipo dicotómica. Después de ya haber determinado la variable dependiente pasamos a las independientes, que en este caso fueron las de Risk Flag, Age, Income y Current_House_Yrs, ya que encontramos similitudes al momento de analizar la experiencia.

Analizando los resultados brindados por los modelos, podemos observar que tanto la precision como la exactitud tiene un porcentaje de 52%, en cambio el recall tiene un porcentaje de 5% este siendo el más bajo, todo lo contrario al F1 score donde se obtuvo un porcentaje del 67% estando por encima de la precisión y la exactitud.

Experience 2

Para el segundo modelo usando la variable dependiente de “Experience” se decidieron cambiar las variables independientes, en este caso se usaron Age, Income, Current_House_Yrs, esto para observar si cambian en algo los resultados.

En este caso se logró observar que en el caso de la precisión y la exactitud se mantuvieron en 52%, los únicos cambios que se tuvieron fue en el recall que tuvo un 2% de exactitud y en F1 score se tuvo 4%, a lo que se puede concluir que para la variable “Experience” el primer modelo tuvo mejores resultados.

Current_Job_Yrs 1

Otra variable que se utilizó fue la de “Current Job Yrs” en donde primero se obtuvo la mediana, después se cambió la columna para que en caso de que el valor fuera mayor a la mediana se colocara un “altos” y en caso de ser lo contrario se colocará un “bajos”.

Como variables independientes se usaron Age y Experience, ya que se encontró cierta similitud al momento de relacionarlas con la dependiente.

Los resultados obtenidos utilizando estas variables están arriba del 60% de eficacia, en el caso de la precisión y exactitud, ambos se encuentran con un 69%, mientras que el en recall tiene un porcentaje del 65%, el resultado más alto fue el de F1 score el cual tuvo un porcentaje de 73%

Current_Job_Yrs 2

Para el segundo modelo usando la variable de Current_Job_Yrs se cambiaron las variables

independientes agregando dos variables más terminando usando Age, Experience, Current_House_Yrs y Risk_Flag.

Usando este modelo se lograron obtener los siguientes resultados; una precisión del 65%, exactitud del 69%, un recall de 64% y una F1 score del 72%. En el caso de los modelos con esta variable se nota que la eficacia es de las más altas, aun cuando se cambian las variables independientes, esto nos indica que la forma en que predice es más eficaz a comparación de otras.

Conclusión

En la actividad, se ha identificado que el mejor modelo de clasificación tiene como variable dependiente "Current_Job_Yrs" y como variables independientes "Age" y "Experience". Este modelo ha arrojado los siguientes valores de métricas de evaluación:

- Precision Score: 0.6984
- Accuracy Score: 0.6984
- Recall: 0.6502
- F1 Score: 0.7306

El modelo "Current_Job_Yrs 1" destaca por varios motivos:

- Alta Exactitud: El modelo tiene un Accuracy Score de 0.6984, lo que significa que el 69.84% de sus predicciones son correctas en el conjunto de datos de prueba.
- Balance entre Precision y Recall: El modelo tiene un Precision Score de 0.6984 y un Recall de 0.6502. Estos valores están bastante cerca el uno del otro, lo que indica que el modelo es capaz de hacer predicciones positivas con una alta precisión mientras que también es capaz de identificar la mayoría de las instancias positivas.
- Puntuación F1: El F1 Score de 0.7306 es un indicador que combina tanto la precisión como la sensibilidad en una sola métrica. Un F1 Score alto indica un equilibrio efectivo entre ambas métricas, lo que es esencial para clasificar de manera efectiva en un conjunto de datos.

Esto sugiere que el modelo es efectivo por su capacidad para realizar predicciones precisas y al mismo tiempo ser capaz de identificar un alto porcentaje de instancias positivas en los datos. Su equilibrio entre Precision, Recall y F1 Score lo convierte en una opción sólida y

efectiva para clasificar y predecir datos relacionados con la variable "Current_Job_Yrs".