**Trabajo practico N°4**

**Iván Martínez, Joaquín Reyes y Santiago Reinoso Sokol**

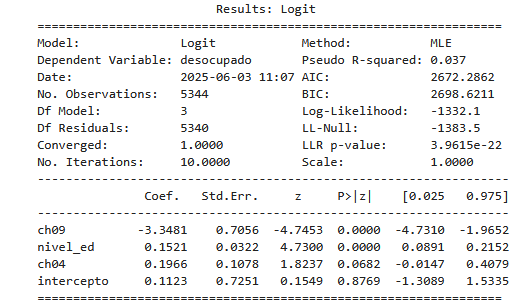
1. Tanto para la base 2004 como para 2024 vemos que la diferencia entre la media para la base de entrenamiento y la base de test es relativamente baja en la mayoría de las variables. Es decir, si bien hay diferencias en los valores del promedio estas son bastante similares, por lo que podríamos decir que la base de entrenamiento es representativa de la base de testeo.
2. Tabla:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
| Edad | -1117.443 | -5087.510 | -7310.923 | -6724.342 | -3362 |
| Edad2 |  | 426.321 | 653.929 | 610.846 | 610 |
| Educ |  |  | 1525.947 | 1646.343 | 1646 |
| mujer |  |  |  | -2218.556 | -3703 |
| Analfabetismo |  |  |  |  | -3362 |
| Nivel Ed |  |  |  |  | 823 |
| observaciones | 3224 | 3224 | 3224 | 3224 | 3224 |
| R^2 | 0 | 0 | 0.04 | 0.052 | 0.0516 |

Podemos ver que tanto los modelos 1 y 2 tienen un R^2 de 0. A partir de la variable Educación (entendido como máximo nivel educativo alcanzado, CH12), la cual tiene una relación positiva con el salario semanal aumenta R^2. Un mayor se ve al añadirse la variable de género (en este casó mujer), teniendo esta una relación negativa con respecto al salario semanal. Lo mismo ocurre con una de las variables escogidas (analfabetismo), esto significa que al aumentar este, el salario semanal disminuye.

Y por último el modelo 5, el más completo incluye la variable Nivel educativo (entendido esta vez como máximo nivel educativo completado), esto se distingue de educación (como se veía antes) ya que esta toma en cuenta simplemente los niveles educativos terminados, aunque esta diferencia parezca menor, podemos ver como al agregar esta variable y la consideración del analfabetismo, el modelo 5 toma mucha más significancia.

Cabe aclarar que en todos los modelos el número de observaciones fue el mismo y se puede ver una evolución progresiva, modelo a modelo, en el R^2.

5)

**Resultados de la regresión logit en el año 2004**

**Tabla

El contenido generado por IA puede ser incorrecto.**

**Resultados de la regresión logit en el año 2024**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | AUC score | Matriz de confusión |
| **Logit (2004)** | 0.937582 | 0,62 | **Texto  El contenido generado por IA puede ser incorrecto.** |
| **KNN (K=5) (2004)** | 0.938 | 0.5305 |  |
| **Logit (2024)** | 0.951973 | 0.56 |  |
| **KNN (K=5) (2024)** | 0.952 | 0.5 |  |

**Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.**

**Curva ROC Logit del año 2004**

**Gráfico, Gráfico de líneas, Histograma

El contenido generado por IA puede ser incorrecto.**

**Curva ROC KNN del año 2004**

**Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.**

**Curva ROC Logit del año 2024**

**Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.**

**Curva ROC KNN del año 2024**

Dependiendo el año podemos ver diferencias entre cada uno de los modelos, así que vamos a examinarlos por año:

Para el año 2004 vemos que tanto el modelo Logit como el modelo de vecinos cercanos tienen una accuracy (precisión o exactitud) muy cercana, por no decir igual, lo que nos dice que clasifican correctamente la misma proporción de observaciones, ósea que su proporción de la suma de true positives y true negatives sobre los positivos y negativos da muy parecido ((TP+TN) /(P+N)).

Por lo que pasaremos a la siguiente medida de precisión, la AUC, o Area under curve, que nos muestra el área bajo la curva de los scores predichos, en este caso tenemos un AUC score de 0,62 para logit y de 0,5305 para KNN. Esto nos dice que nuestro modelo logit suele discriminar más entre ocupados y desocupados, los que nos sirve, por que significa que el modelo no es totalmente aleatorio (como seria si es 0,5 o muy cercano) y que tenemos mejor capacidad de predicción, todo esto, puede notarse en las curvas ROC del año 2004, que representan justamente esta capacidad gráficamente, por lo que, en conclusión, el Modelo logit es mejor para predecir en el año 2004.

Por otro lado, para el año 2024, vemos que tenemos un caso similar, pero más reñido, donde vemos que hay una accuracy de 0,9519 para el modelo logit y una accuracy de 0,952 para KNN, por lo que podemos decir que su accuracy es la misma. En cuanto a la AUC, vemos que los modelos de logit y KNN tienen un score de 0,56 y 0,5 respectivamente, por lo que vemos que el modelo logit tiene una pequeña capacidad de discriminar entre ocupados y desocupados, mientras que en KNN vemos que esta capacidad no está, es totalmente aleatorio y, por lo tanto, no es viable usarla para predecir, como se puede ver en la curva ROC donde las líneas se superponen. Por lo tanto, el modelo de regresión logit es el mas ideal para predecir en el año 2024.

6) con el modelo logit, en el año 2004, pudimos predecir que 0,69 personas están desocupadas, por lo que en proporción a las 10 observaciones que tiene la base, sería un 6,9% de personas desocupadas.

En el año 2024, pudimos predecir que 1,46 personas están desocupadas, por lo que en proporción a las 41 observaciones que tiene la base, sería un 3,6% de personas desocupadas.