Universidad Nacional Autónoma de México
Facultad de Ciencias
Taller de modelación II | 6009
Proyecto 1 - IMDb
Flores Callejas Arturo
30/04/2023

Planteamiento.

A partir de la base de datos pública gestionada por IMDb (Internet Movie Database), se busca generar un modelo capaz de clasificar de forma binaria las diferentes películas existentes en la base de datos.

Las categorías de clasificación se darán a partir de la calificación media otorgada por el público a cada una de las películas, diferenciando entre una "Película exitosa" y una "Película no exitosa". Donde el modelo, a partir de las características inherentes de cada película (tales como género, país de origen, lenguaje, etc.) y sin conocer la calificación media, deberá asignarlas correctamente en dichas categorías.

Para ello, se debe establecer una clara población objetivo, en este caso particular, se buscan películas *"relevantes"* a la fecha con una calificación en la plataforma de IMDb sobre el *"punto de corte"* establecido.

Dichos conceptos de *"relevancia temporal*" y *"punto de corte de calificación"* fueron propuestos bajo criterio estadístico y un análisis de la problemática. De tal forma que la población objetivo definida para este proyecto está dada por:

1. Únicamente títulos con formato de película.
2. Únicamente títulos con fecha de creación posterior a 1980.

Análisis exploratorio.

Durante esta fase del proyecto se estudió a profundidad las bases de datos, para posteriormente seleccionar las variables de interés y posible utilidad para el modelo, ya sea introduciéndolas directamente al modelo, o generando nuevas variables a partir de ellas.

1. **Estudio de la base de datos.**
   En este primer estudio, se identificó la información relevante para el objetivo del proyecto proporcionada por cada una de las diferentes tablas de datos.
   (a) **tconst**: Idenficador alfanumérico único de la película.
   (b) **title**: Título del filme por ubicación.
   (c) **region**: Región del título.
   (d) **language**: Lenguaje del título.
   (e) **titleType**: Tipo de formato del título.
   (f) **primaryTitle** : Nombre más popular del filme.
   (g) **isAdult** : Indicador booleano para películas para adultos o para el público general.
   (h) **startyear** : Año de lanzamiento del filme.

(i) **runtimeMinutes** : Duración del filme en minutos.
(j) **genres** : Géneros asociados al filme con máximo de 3 géneros.
(k) **directors** : Director o directores del filme.
(l) **writers** : Escritor o escritores del filme.
(m) **averageRating** : Calificación media en IMDb para el filme.
(n) **numVotes** : Número de votos en IMDb para el filme.
(o) **nconst** : Identificador alfanumérico único de cada persona del elenco.
(p) **primaryProfession** : Profesiones asociadas a cada persona del elenco con un máximo de 3 profesiones.

2. **Restricción a *target*.**
Se restringen los datos a únicamente la población objetivo, es decir, se realiza una selección de datos dados los criterios previamente mencionados: que sea un filme tipo película y que su estreno haya sido posterior a 1980.
Una vez generada esta restricción, se aplica al resto de conjunto de datos para recopilar la información de dicha población objetivo, en este caso, se obtuvieron $211,383$ registros *target*.

3. **Conteos y verificación de registros.**
Una vez identificadas las variables de interés, es importante cerciorarse de que estas sean de utilidad estadística para el modelo, verificando que existan suficientes registros para la población objetivo y que el porcentaje de valores nulos sea lo más bajo posible *(aceptable debajo del* 10%*)*, de lo contrario se deberán tratar dichos datos.
Tras realizar las consultas pertinentes, verificamos que la mayoría de las variables son aptas por cantidad de registros y porcentaje de valores nulos para entrar al modelo, sin embargo, en algunas se debe realizar un tratamiento específico de los valores nulos.

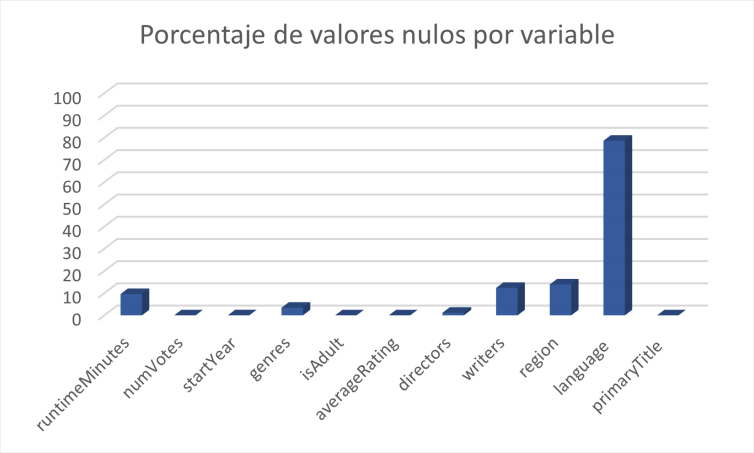


Figure 1: Valores nulos

Como se puede observar en la figura 1, la columna *language* tiene un porcentaje de valores nulos cercano al 80%, y a pesar de parecer un porcentaje inaceptable para el modelo, al estudiar más de cerca la bases de datos, podemos observar que los únicos registros no nulos en dicha columna, corresponden al lenguaje original de cada del título, mientras que los valores nulos se refieren a las traducciones de cada título.
Por lo que a pesar de tener un alto porcentaje de valores nulos, esta variable puede ser de utilidad al momento de explorar a fondo los lenguajes de origen de cada película, y por ende, de interés para el modelo.

Construcción de la ABT.

La tabla ABT (*Analytical base table*) es una tabla plana usada para la construcción de modelos predictivos en aprendizaje supervisado. La tabla ABT contiene una sola columna representando al sujeto u objeto objetivo y el resto de columnas, o mejor llamadas variables estadísticas, describen las características del mismo.

En la ABT existen dos tipo de clasificación general de variables estadísticas :

- **Variables categóricas**: Variables que describen una propiedad cualitativa, es decir, cualidades del objeto de estudio. Dentro de esta clasificación, se pueden diferenciar variables categóricas nominales (sin orden) y ordinales (con orden establecido).
- **Variables cuantitativas** : Tal como su nombre lo indica, representan numéricamente características del objetivo, lo que permite realizar cálculos y estadísticas con dichos datos. Pueden ser de carácter continuo o discreto.

Selección y creación de variables.

Esta fase del proyecto se dedicó seleccionar, modificar o crear variables estadísticas para posteriormente construir la tabla ABT.
Para ello, la variable de identificación que se usó fue *tconst*, mientras que las variables iniciales que se eligieron para formar la ABT son:

- **prymaryTitle**
- **isAdult**
- **runtimeMinutes**
- **numVotes**

Mientras que las variables modificadas y creadas son:

Variables categóricas.

- **decade**: Variable modificada de *startyear* donde se agrupan los años desde 1980 por decadas, es decir 80´s, 90's, 00's, 10's y 20's.
- **genres1**: Variable modificada de *genres* donde se muestra el primer género o género principal de la variable original.
- **principalcrew**: Variable creada a partir de *profession* como vector de combinaciones binarias para cada profesión ejercida en las películas.
- **departments**: Variable creada a partir de *profession* como vector de combinaciones binarias para cada departamento de trabajo en las películas.
- **othercrew**: Variable creada a partir de *profession* como vector de combinaciones binarias para parte extra del staff en las películas.
- **continent code**: Variable creada a partir de *region* y su respectivo continente donde se muestra el continente principal donde se presentó la película
- **top director**: Variable creada a partir de *directors* y *averageRating* donde se muestra un indicador 1 si el director participó en alguna de las películas con mejor rating y 0 en otro caso.

- **top actor**:Variable creada a partir de *prymaryProfession*, *averageRating* y *knownfortitles* donde se muestra un indicador 1 si el actor participó en alguna de las películas con mejor rating y 0 en otro caso.
- **top actress**: Variable creada a partir de *prymaryProfession*, *averageRating* y *knownfortitles* donde se muestra un indicador 1 si la actriz participó en alguna de las películas con mejor rating y 0 en otro caso.

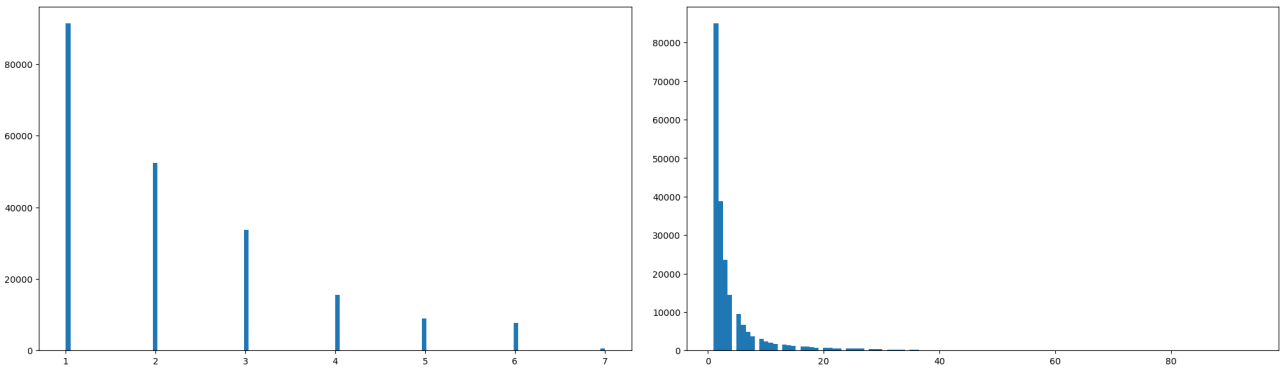
Variables numéricas.

- **cnt genres**: Variable creada a partir de *genres* donde se muestra el conteo de géneros por cada película.
- **cnt types**: Variable creada a partir de *types* donde se muestra el conteo de tipos por cada peícula.
- **cnt region**: Variable creada a partir de *region* donde se muestra el conteo de regiones en que se presentó cada película.
- **cnt titles**: Variable creada a partir de *title* donde se muestra el conteo de los distintos títulos para cada película.
- **target**: Variable objetivo creada a partir de *averageRating* como indicador booleano, con 1 como promedio de califación mayor o igual a 8 y 0 en cualquier otro caso.
- **cnt continents**: Variable creada a partir de *region* y su respectivo continente donde se muestra el conteo de los distintos continentes donde se presentó cada película.

Análisis estadístico de la ABT.

Esta fase del proyecto fue dedicada a realizar un estudio estadístico, tanto univariado como multivariado, a la tabla ABT generada en la fase previa, comprobando la normalidad de las variables numéricas y el nivel de correlación entre variables categóricas y entre variables numéricas. Esto con el objetivo de mejorar el poder predictivo que las variables pueden otorgarle al modelo, ya sea ajustándolas, tranformándolas o eliminándolas, de manera que el modelo alcance su rendimiento óptimo.

Tras sustituir los valores nulos tanto en las columnas categóricas como en las numéricas, se realizó una breve representación con histogramas para analizar el comportamiento de las variables numéricas.



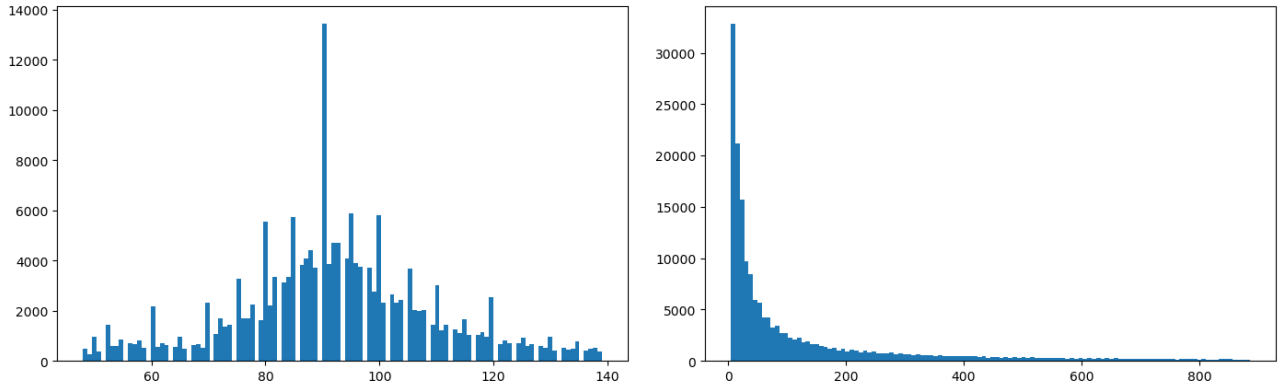
(a) cnt continents (b) cnt titles

Figure 2: Histograma inicial de variables.

Donde se encontró que

- Las variables *cnt titles*(fig 2b) y *cnt region* presentan un comportamiento exponencial.
- Las variables *numvotes* y *runtimeminutes* necesitan tratamiento de outliers pues no permiten analizar correctamente su comportamiento.
- Las variables *cnt genres*, *cnt types*, *cnt continents* (fig 2a) y *target* se agrupan en un número muy pequeño de valores, por lo que es más conveniente tratarlas como variables categóricas.

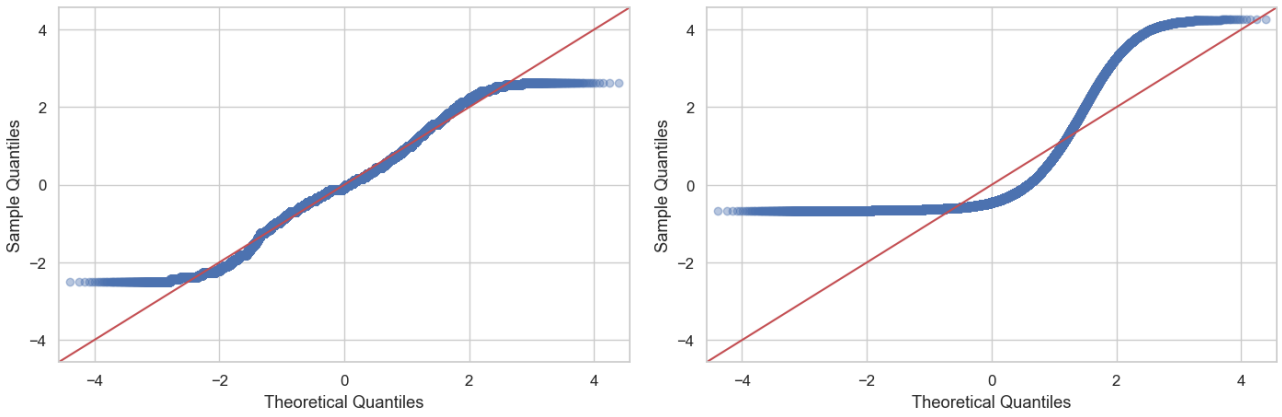
Terminado el primer análisis de variables numéricas, se cambia el tipo de variable a categórica de las variables antes mencionadas y se eliminan los outliers para una mejor visualización de las variables *numvotes* y *runtimemunites*. Sin las outliers, se puede observar que la variable *runtimeminutes* tiene una distribución cercana a la normal, mientras que *numvotes* aparenta un comportamiento exponencial.



(a) runtimeMinutes (b) numVotes

Figure 3: Histogramas sin outliers.

Ahora, el objetivo es verificar la normalidad en distribución de las distintas variables numéricas, para ello empleamos **Gráficos Q-Q**, un método gráfico usado en estadística para diagnosticar diferencias entre dos distribuciones de datos, en este caso, compararemos la distribución de cada una de nuestras variables numéricas con una distribución normal.



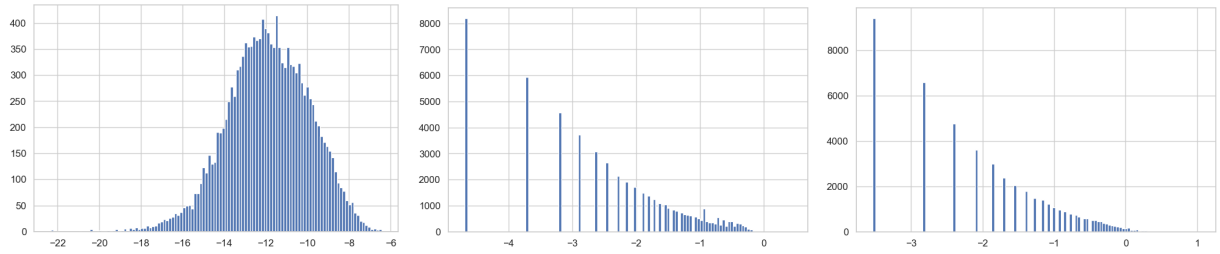
(a) runtimeMinutes (b) numVotes

Figure 4: Gráficas Q-Q.

Tal como era de esperarse gracias a la figura 3, la variable *runtimeminutes* es muy parecida en distribución (excepto en las colas) a una variable normal (4a), sin embargo para el resto de variables no ocurre lo mismo.
Dado que por cuestiones de simplicidad nos interesa que el resto de variables también se distribuyan similar a una normal, entonces les debemos aplicar algún tipo de transformación de potencia. En este caso particular, de las transformaciones que se probaron, la que mejor resultados dió fue

$$log((x - x_{mean})/(x_{var}))$$

Ya que aplicada esta transformación a *numVotes*, *cnt titles* y *cnt region* obtuvimos:



(a) numVotes (b) cnt region (c) cnt titles

Figure 5: Histogramas con transformación.

Donde de puede apreciar claramente un mejor comportamiento de la variable *numvotes* (más cercano al de una distribución normal), y en las otras dos restantes hay una mejor visualización del comportamiento de los datos a pesar de no ser normal. Conclusión después reforzada cuando posteriormente utilizamos las pruebas estadísticas de normalidad **Shapiro-Wilk**, **Anderson-Darling** y **Kolmogorov-Smirnov**.

Una vez analizado el comportamiento individual de cada cada variable (análisis univariado), se debe estudiar como se comportan entre ellas y cuán relacionadas están unas con las otras.
Para poder estudiar esta correlación se decidió realizar un **análisis de componentes principales** (PCA por sus siglas en inglés) para la reducción de variables en el modelo y el cálculo de los **coeficientes de correlación de Spearman**.

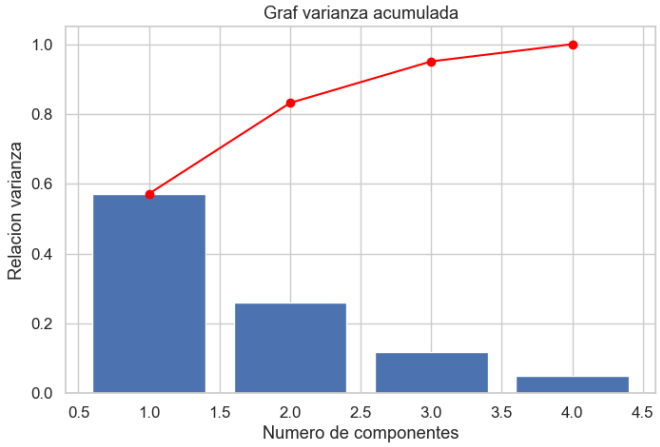


Figure 6: PCA

El análisis de correspondencia múltiple nos arrojará la significancia de cada variable para el modelo de acuerdo a la varianza acumulada por cada variable, mientras que el Coeficiente de Spearman con rango $[-1, 1]$, nos dirá si la correlación entre dos variables es positiva (valor cercano a 1), negativa (valor cercano a -1) o nula (cercano a 0).

En la figura 6, se representan 4 variables donde

1. *runtimeminutes*
2. *numvotes*
3. *cnt region*
4. *cnt titles*

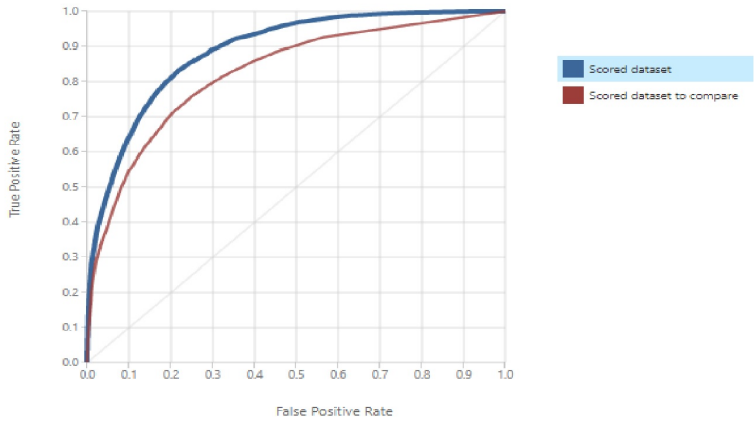
De modo que la variable menos significativa según los resultados de nuestra prueba PCA es la variable *cnt titles*. Lo cual se corrobora posteriormente al calcular los coeficientes de correlación de Spearman, donde el coeficiente de Spearman para la variable *cnt titles* y *cnt region* es 0.86, es decir, existe una fuerte correlación positiva entre ellas.

Con este conocimiento, podríamos remover la variable *cnt titles* para entrar al modelo y dejar únicamente *cnt region* pues el aporte al modelo es prácticamente el mismo. Sin embargo, al no contar con un número exagerado de variables, se ha decidido mantenerla para el modelo.

El resto de coeficientes de Spearman muestran una nula o muy baja correlación entre las demás variables.

Modelación.

Para esta fase modelación, es decir, generar el modelo, se utilizó la herramienta para machine learning de Microsoft llamada **Azure ML studio**, donde se introdujo la ABT construida, se seleccionaron la variables a emplear y el conjunto de datos de validación y entrenamiento.



(a) Curva ROC.

True Positive	False Negative	Accuracy	Precision
1880	3679	0.914	0.688
False Positive	**True Negative**	**Recall**	**F1 Score**
854	46229	0.338	0.453

(b) Resultados.

Figure 7: Resultados del modelo.

Tras probar y comparar resultados de diferentes algoritmos, se decidió dos algoritmos para el modelo puesto que fueron los que mejores resultados arrojaron. El primero, un árbol de decisión de dos clases y el segundo, una red neuronal también de dos clases. Como se puede observar en la figura 7a, la red neuronal (representada por la linea azul) fue un más acertada en comparación con el árbol de decisión

(representado por la linea roja).

Para visualizar el modelo en Azure ML, puede acceder al siguiente link.

Validación del modelo.

Como se puede observar en la figura 7b, el modelo obtuvo unos resultados destacados, con un *accuracy* del 91.4% sobre el conjunto de validación establecido en un 25% de la cantidad de registros originales, es decir, $52,711$ registros.

Se espera que el modelo sea capaz de asignar satisfactoriamente los objetos, en este caso películas, tras introducir los datos necesarios.