



Facultad de Estudios Superiores: Acatlán.

Diplomado en Ciencia de Datos



Proyecto:


La magia del cine (o de los datos)

Valencia Martínez Jesús Adolfo




La magia del cine (o de los datos)

Resumen


En el mundo, la información crece de manera constante y poco a poco  se convierte en valiosos activos para su mayor explotación. Hablando del arte, y más específicamente, en el llamado séptimo arte, este fenómeno no sería la excepción, la cantidad de producciones emerge y se va anexando a la base gigantesca de películas creadas por los humanos .

~~Por lo cual,~~ el presente reporte expone la aplicación de diversas metodologías de Ciencia de Datos  para obtener resultados de valor haciendo uso de los datos generados por la industria del cine.



La fuente

Los datos de insumo corresponden a una extracción dentro de un sitio web llamado Kaggle. Estos datos se actualizan diariamente y detallan la información de las películas que se encuentran en la página de reseñas de internet llamada TMDb. Esta página ~~que~~  información no solamente de películas, también de series y capítulos de series y, ~~si bien,~~ aunque la misma página ofrece una API para realizar extracción de información, la base que se encuentra en Kaggle contiene inforión únicamente de películas junto con otra información relevante, por lo cual, se opta por elegir esta última  para la extracción de información.

Los datos

Los datos  gan la información de un total de 750,318 películas y 20 características:

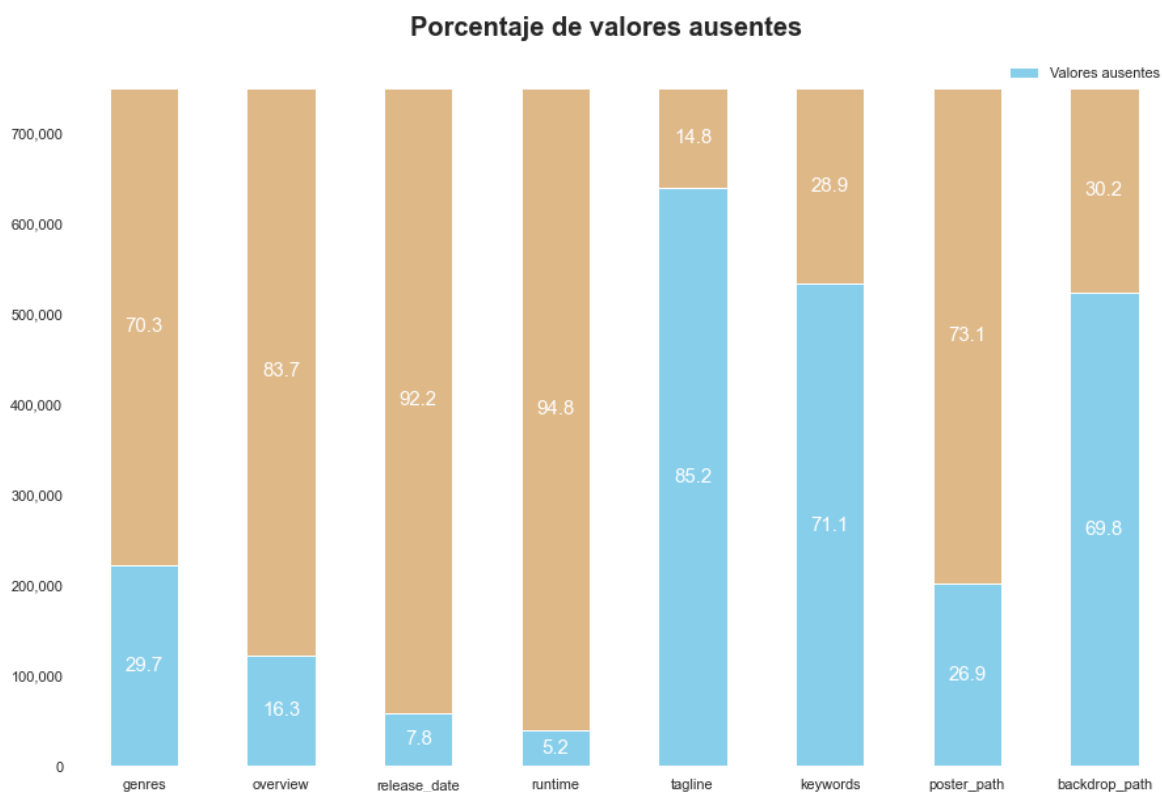
<i>Característica</i>	<i>Descripción</i>
<i>id</i>	Identificador único en la página de TMDb
<i>title</i>	Nombre de la película
<i>genres</i>	Géneros relacionados
<i>original_language</i>	Idioma original
<i>overview</i>	Resumen o sinopsis
<i>popularity</i>	Métrica de TMDb para medir la popularidad dentro del sitio web
<i>production_companies</i>	Compañías productoras
<i>release_date</i>	Fecha de estreno
<i>budget</i>	Presupuesto de producción
<i>revenue</i>	Ganancia reportada
<i>runtime</i>	Duración en minutos
<i>status</i>	Estado actual de la película (estrenada o no)
<i>tagline</i>	Breve frase alusiva a la trama de la película
<i>vote_average</i>	Calificación promedio que los usuarios le dan a la película
<i>vote_count</i>	Cantidad de calificaciones proporcionadas
<i>credits</i>	Nombre de los productores principales
<i>keywords</i>	Palabras clave sobre la película
<i>poster_path</i>	URL para obtener el poster de la película
<i>backdrop_path</i>	URL para obtener el backdrop de la película
<i>recommendations</i>	Recomendaciones a otras películas relacionadas

De las características anteriores, hay algunas que, para propósitos de aplicación, no  consideraron para su uso, por lo que  descartaron. Las variables que se descartaron fueron:

- 1 title
- 2 production_companies
- 3 credits
- 4 recommendations


Valores ausentes

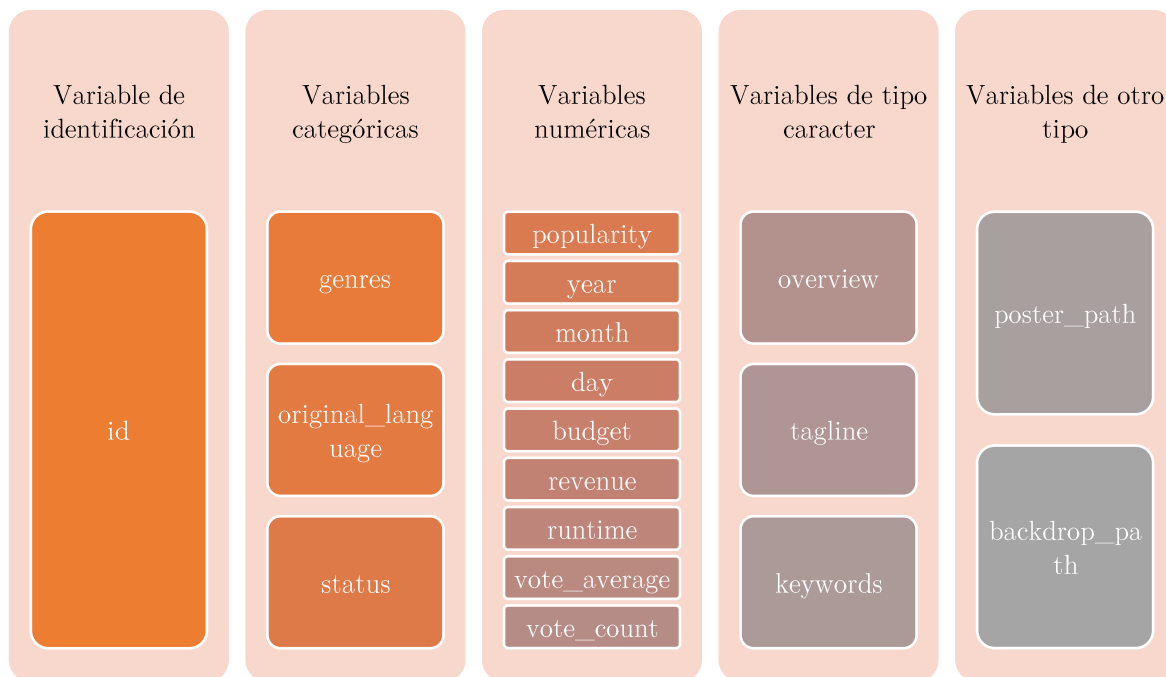
Realizando la exploración de la base de datos anterior descrita, se pudo observar que algunas películas no cuentan con información de algunas características. Las características que no cuentan con algunos registros se presentan en el siguiente gráfico, donde, a su vez, se detalla el porcentaje de valores ausentes respecto al total de registros:




Como podemos observar, existen variables que cuentan con una gran cantidad de valores ausentes, sin embargo, el tratamiento que se optó por realizar deriva de la naturaleza de su uso. En consecuencia, el tratamiento de estos valores no se realiza de manera general dentro de la primera exploración. Posteriormente, se describe el tratamiento de los mismos en cada caso de uso.

Segmentación de variables

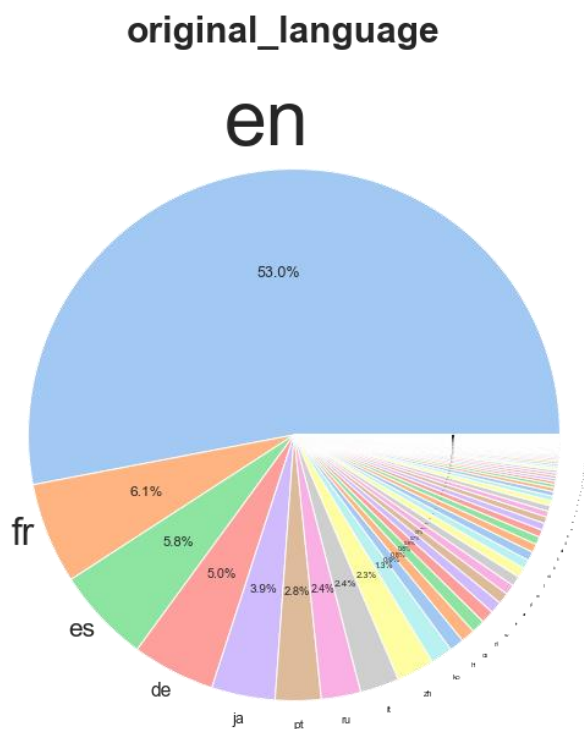
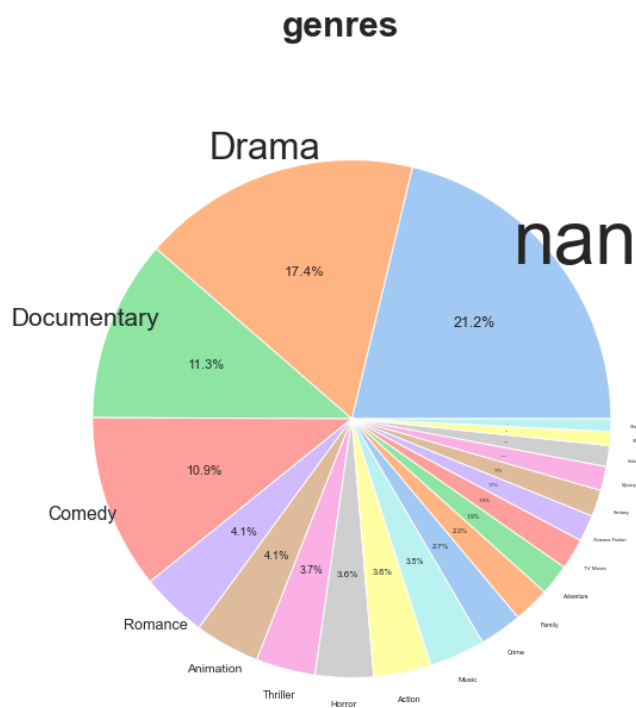
Una vez explicado el punto referente a los valores ausentes, las variables de nuestra base de datos se segmentaron de acuerdo a su tipo  describe su clasificación en la siguiente figura:



A continuación  se pueden apreciar algunos gráficos con que pretende resumir las variables de nuestra base:

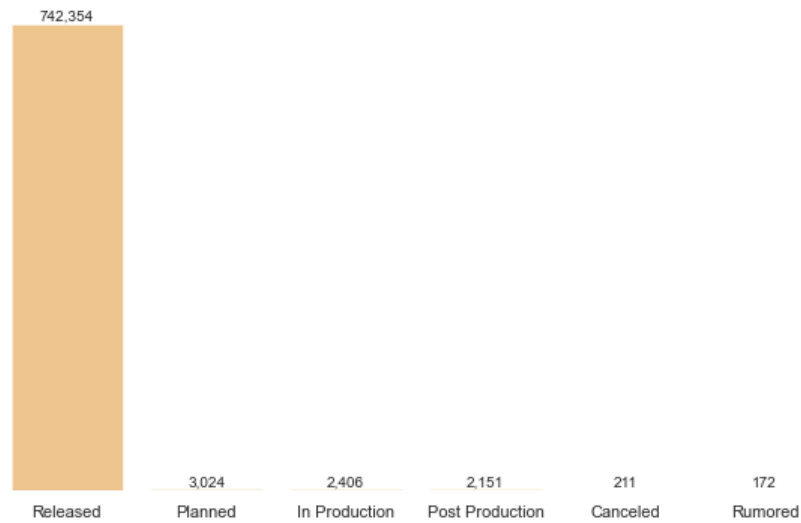
Variables categóricas

En la gráfica de pie de la derecha, podemos observar que la cantidad de valores ausentes juega un papel importante en la cantidad de géneros que se encuentran en nuestra información. Además, existen diversos géneros que si bien, existen en la base, su representación estadística es muy baja.



Para el caso de la cantidad del idioma original de la película, aunque no cuenta con valores ausentes, se puede apreciar la carga de información al idioma inglés y que varios idiomas no cuentan con una gran representación estadística.

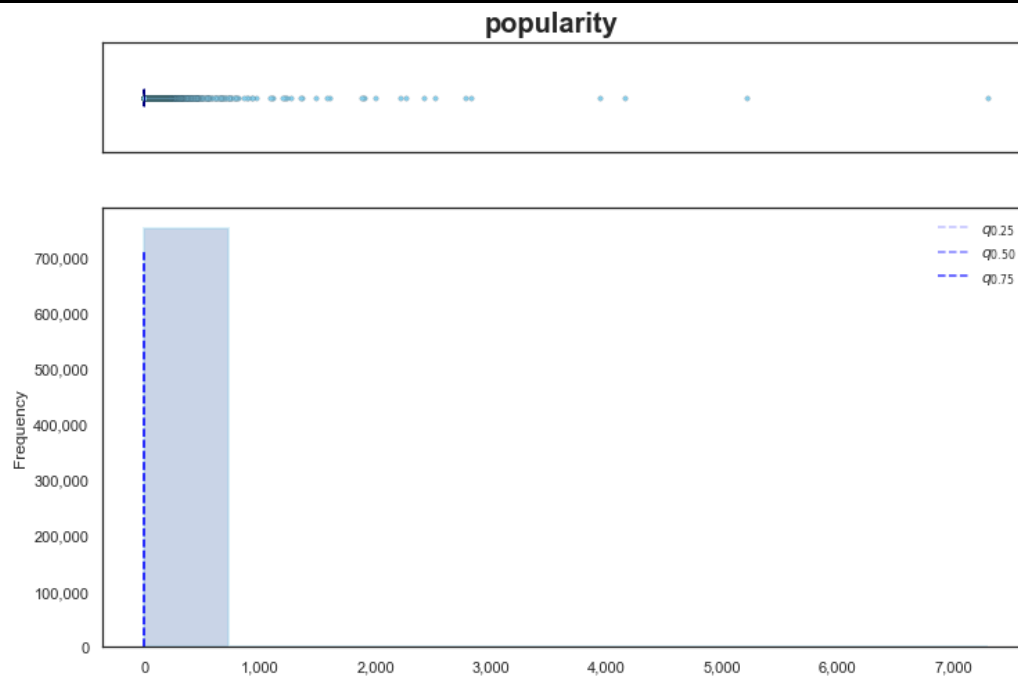
status

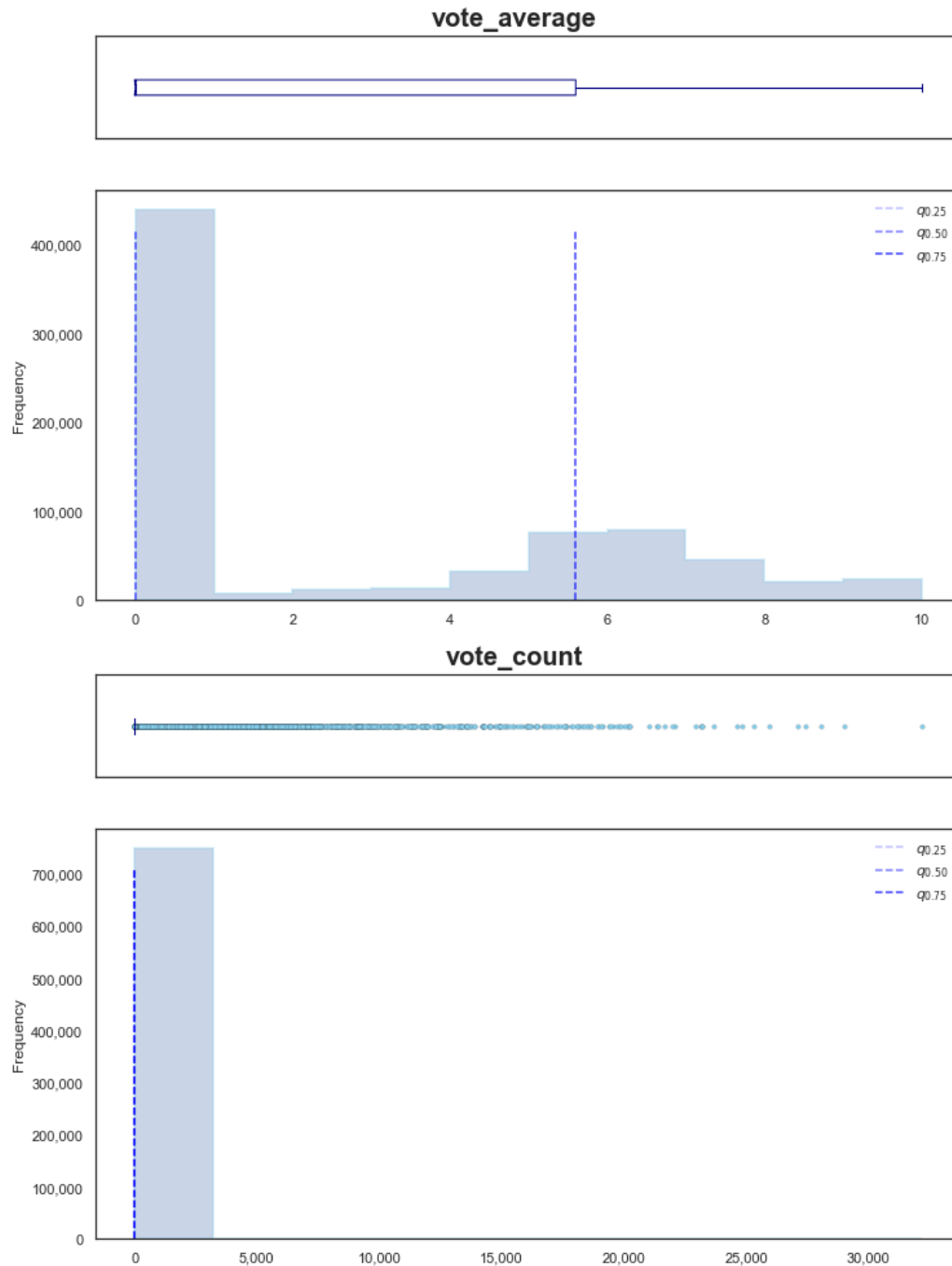


En el caso del estado de la película, es bastante evidente que la información prácticamente se basa en las películas que han sido estrenadas ya.

Variables numéricas

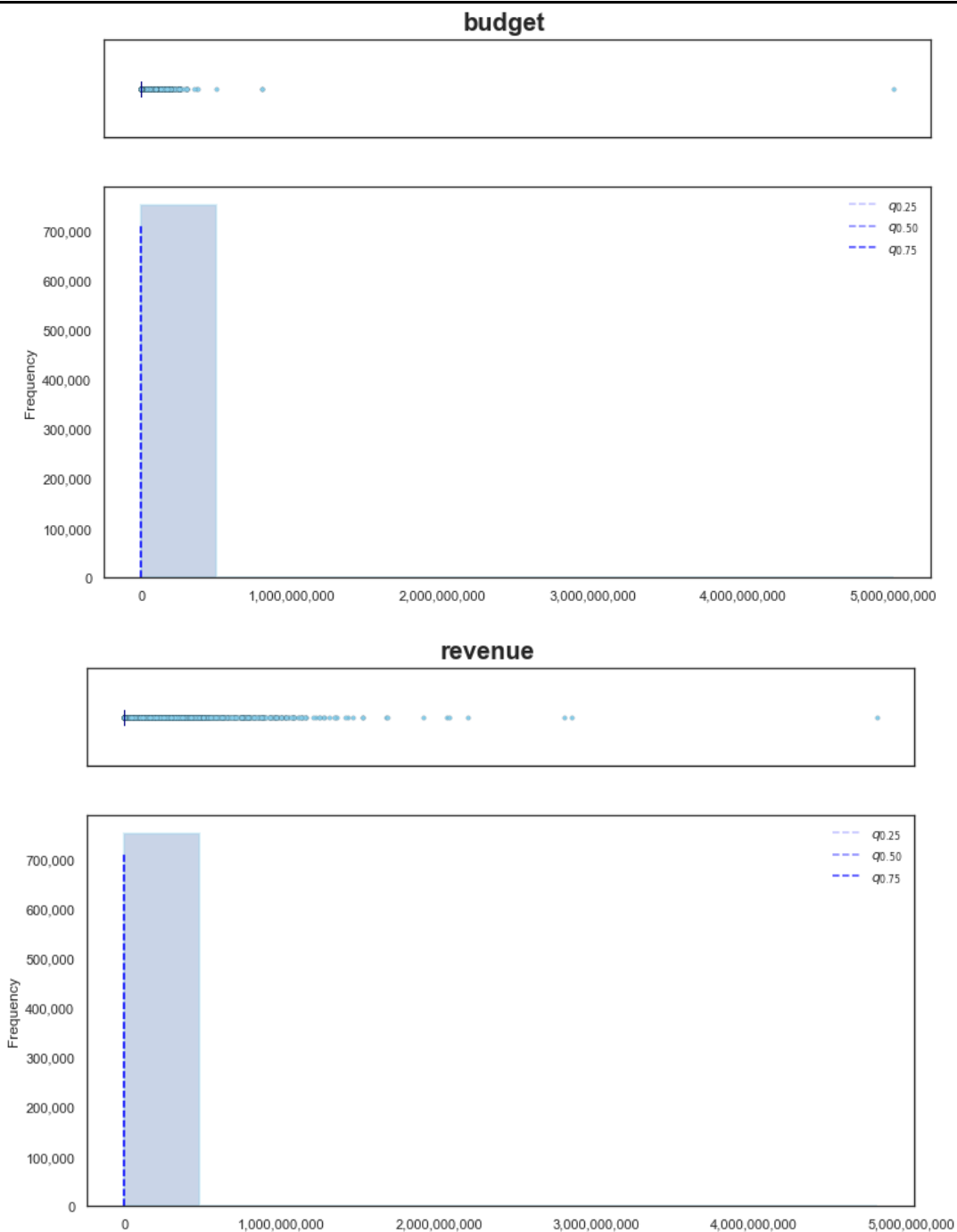
Referentes a opiniones del usuario



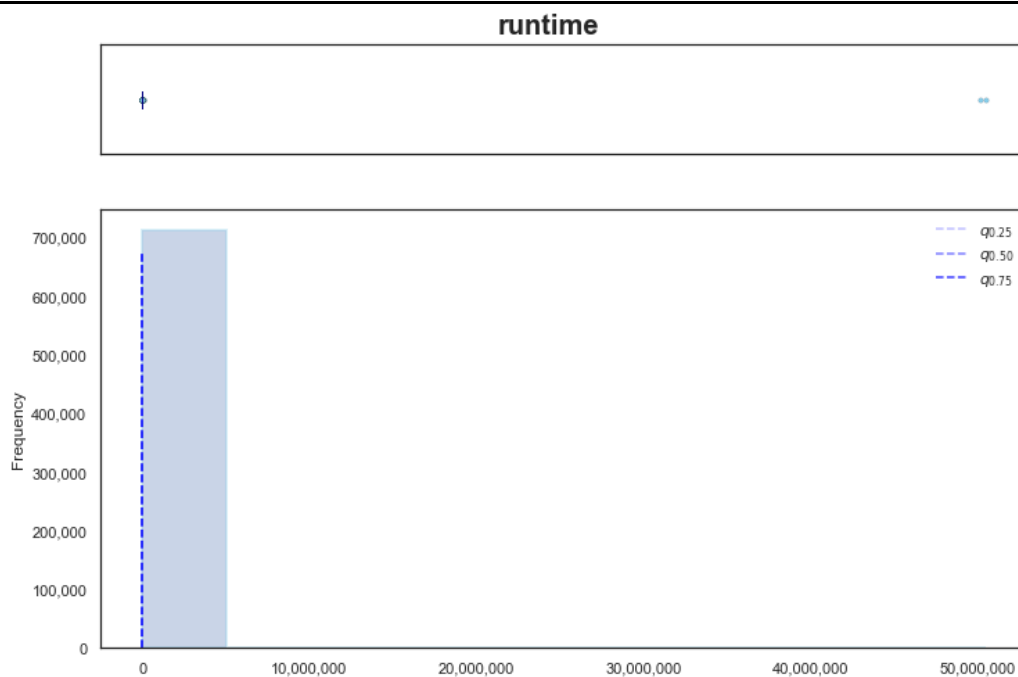


Podemos observar que, en estas variables, se carga mucho la distribución al 0 derivado a que existe una cantidad considerable de películas que, si bien, no cuentan con valores ausentes, no cuentan con calificación por lo tanto, esto influye en la calificación promedio y en la métrica de popularidad.

Referentes cifras monetarias

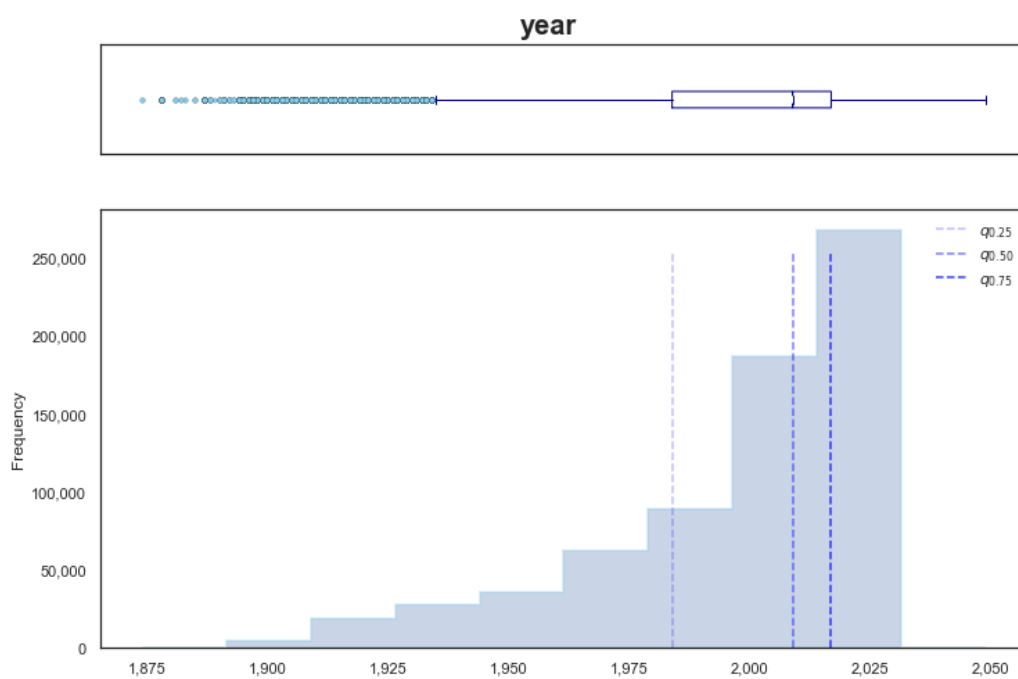


El comportamiento en estas variables es similar a las anteriores, si bien, no hay valores ausentes, su carga hacia el 0 se debe a que se registra la información 0 al no contar con información

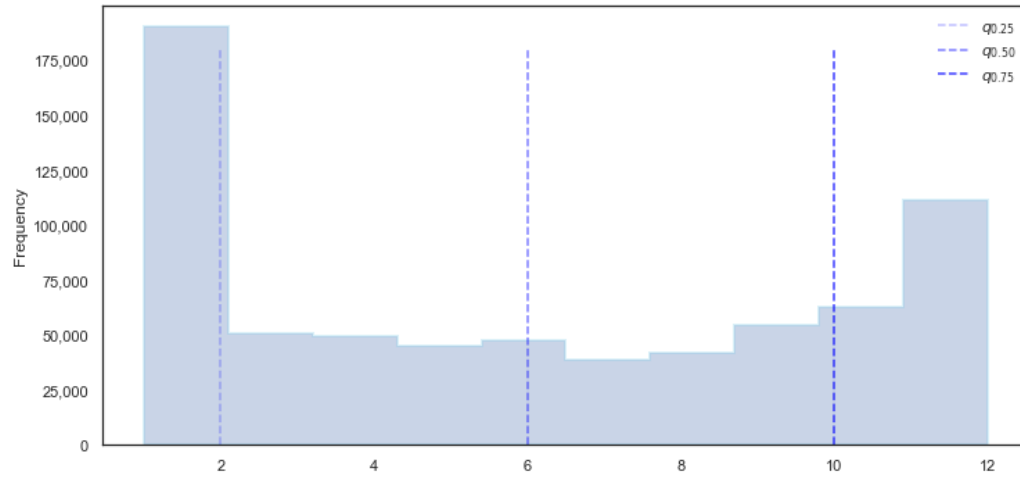
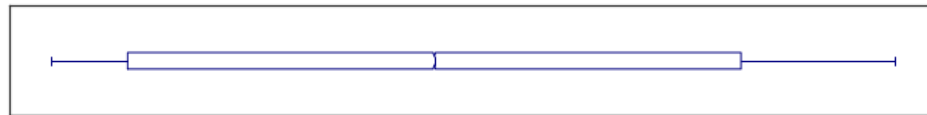


En este gráfico se puede apreciar la evidencia de un claro valor atípico o bien, una captura errónea que provoca un sesgo enorme en la distribución.

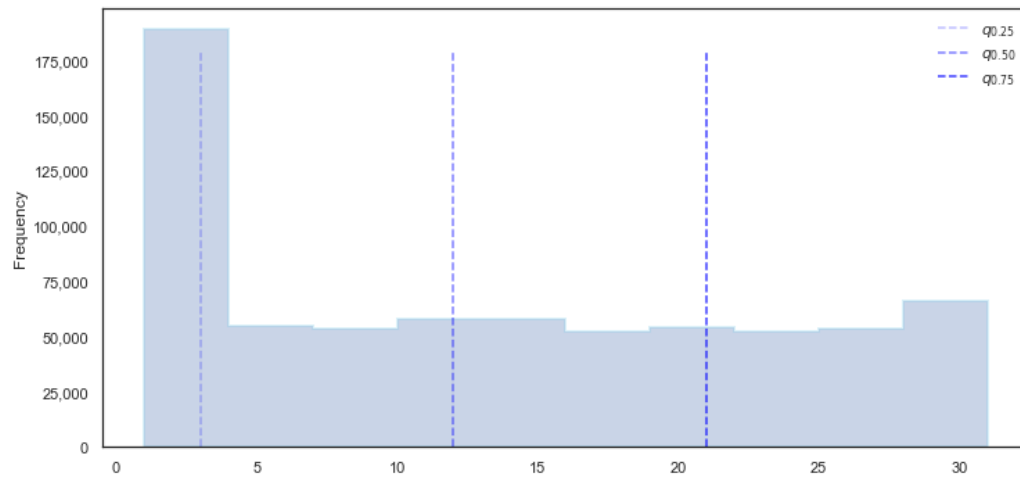
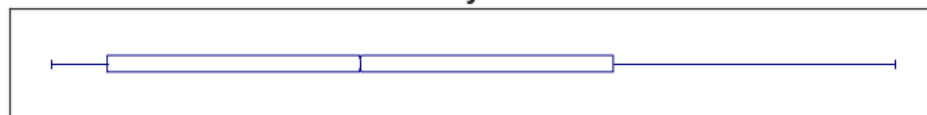
Las siguientes, son las distribuciones de la fecha de lanzamiento de las películas en la base:



month



day

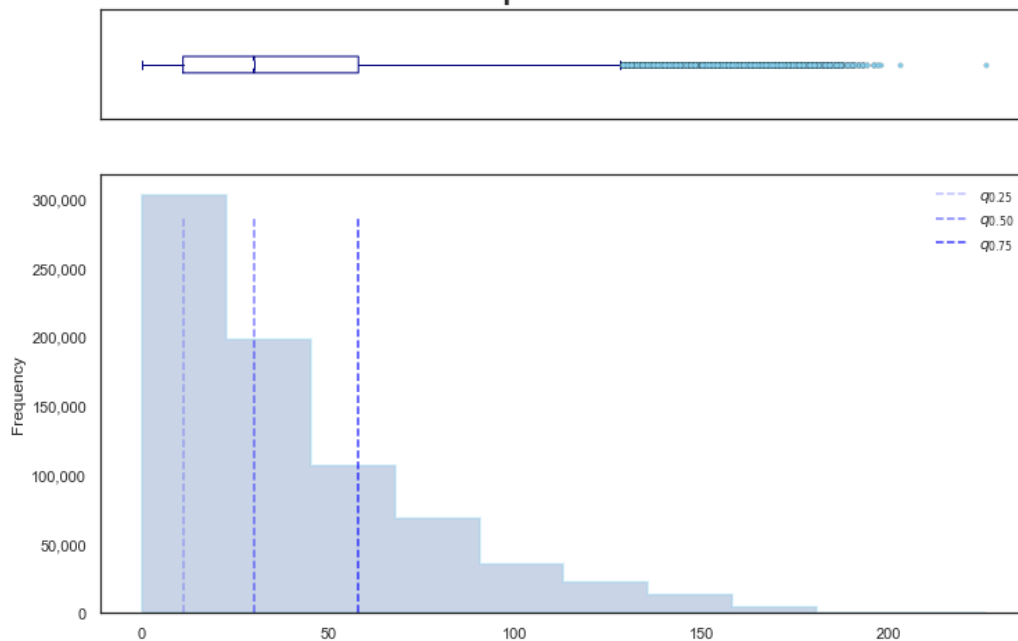


Variables de tipo carácter

overview

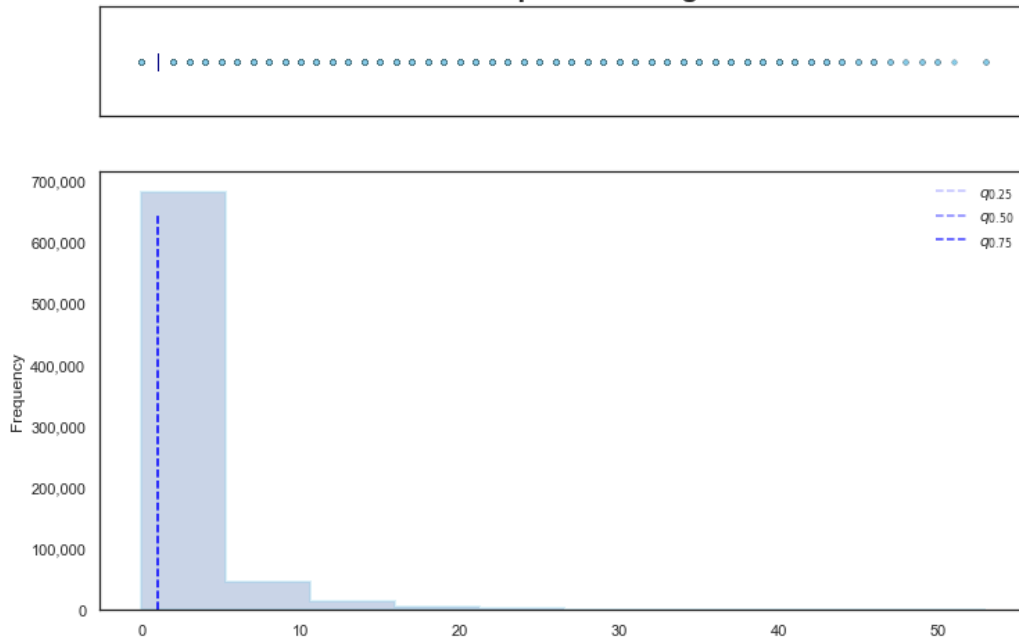


Cantidad de palabras : overview



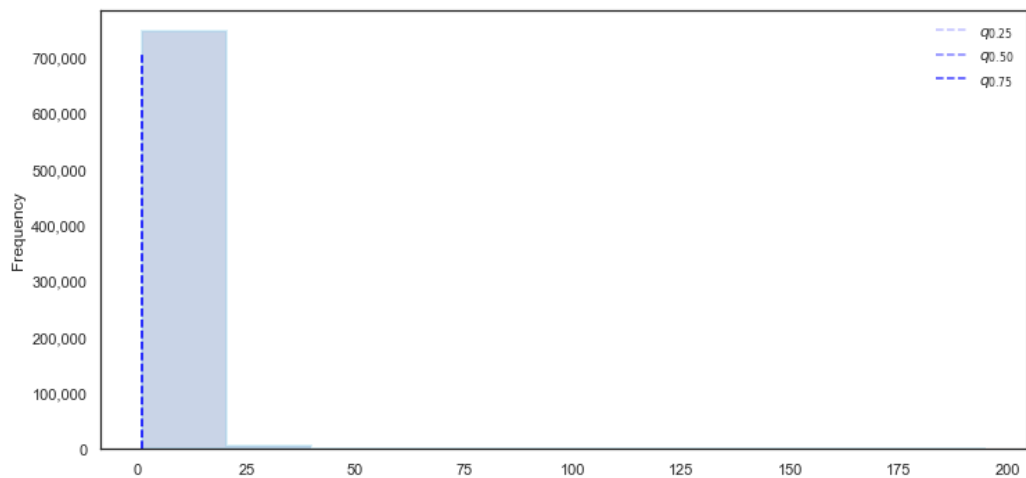


Cantidad de palabras : tagline





QUESTION



Como conclusiones del análisis exploratorio, podemos mencionar que bien, se cuenta con una gran cantidad de películas, muchas de ellas cuentan con información faltante. Por esta razón, el uso de determinadas películas no será relevante para futuros análisis. A continuación, se presentan las metodologías aplicadas con las consideraciones correspondientes para el aprovechamiento de nuestros datos.

Modelaje supervisado

El incentivo principal del uso del modelaje supervisado se basa en la capacidad de predecir con ayuda de las variables que ya contamos. Por lo cual, el objetivo que se planteó y se trabajó para este apartado es el siguiente:

“Predecir si una película es buena o mala”

Este problema puede ser un problema de clasificación, ya que se basa en la predicción de alguna categoría, por lo cual, se aplicaron diversos algoritmos de Machine Learning que atacan de manera eficaz nuestro objetivo:



Variable objetivo

Para realizar nuestro modelo de clasificación es necesario definir cuando una película es buena y cuando es mala. Por lo cual, el criterio que se tomó para esta segmentación se basa en la intuición natural sobre esta estratificación:

- Una película es buena si sobrepasa  la calificación promedio de 6.5.

Variables predictoras

Por otro lado, la elección de las variables que determinarían el que una película sea buena o mala obedece a la idea de poder ocupar variables que se encuentren disponibles incluso para una película recién estrenada. Por ejemplo, el ocupar como predictora la ganancia de una película no podría ser tan viable ya que es un dato que no necesariamente se puede obtener o bien, el dato no podría ser tan verídico.

Teniendo en cuenta lo anterior, las variables que se consideran para la construcción del modelo son las siguientes:

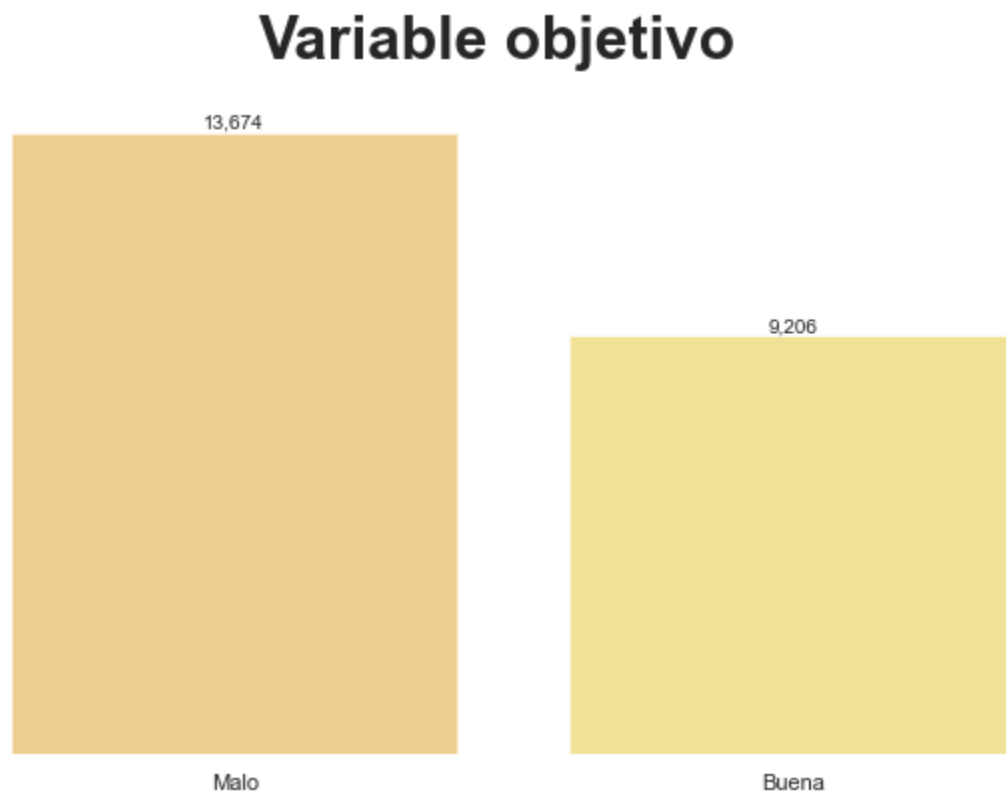
Año de lanzamiento
Mes de estreno
Presupuesto
Duración
Géneros relacionados
Sinapsis
Palabras clave

En consecuencia de lo anterior, es necesario realizar el filtrado de la información para obtener películas que cuenten con la información requerida. Por lo cual, se realiza lo siguiente:

- Filtrado de películas con si cuentan con sinapsis
- Filtrado de películas que si cuentan con información del presupuesto
- Filtrado de películas que si cuentan con información de calificaciones
- Filtrado de películas que cuentan con información de géneros relacionados

Una vez realizado el filtrado, se obtiene un base de datos con una cantidad total de 22,880 películas.

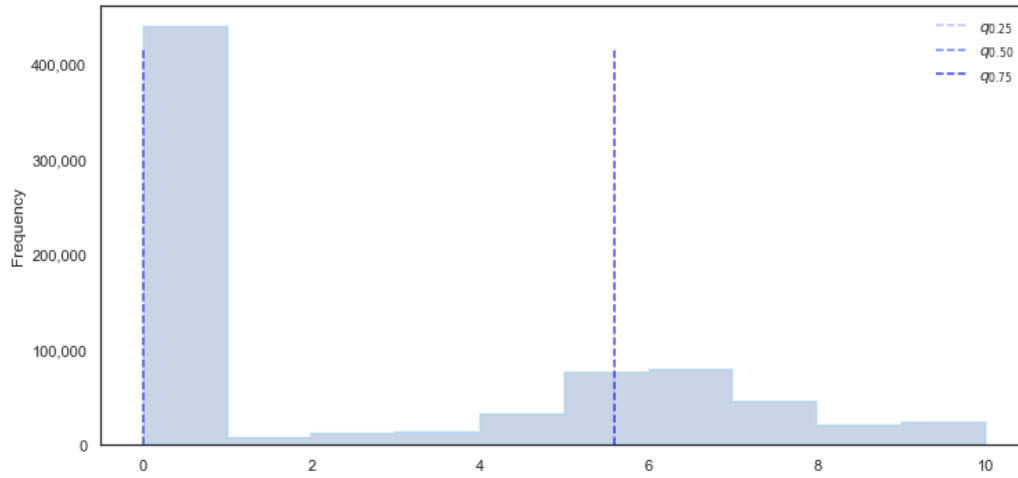
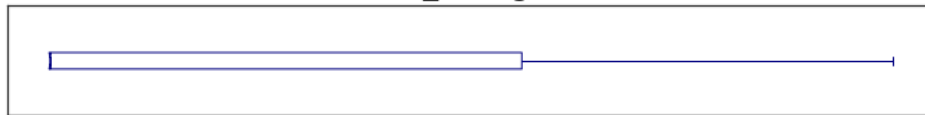
En el siguiente gráfico podemos ver la cantidad de películas, que, de acuerdo a nuestra segregación, se clasifican como una película buena o mala:



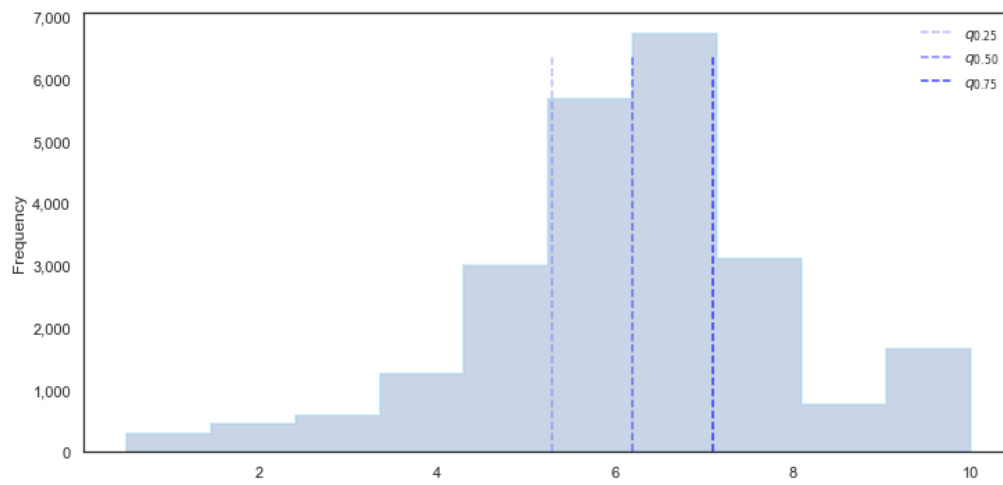
Adicional, se muestran los cambios significativos que sufren las distribuciones al realizar los filtrados.



vote_average



vote_average



Posterior, para poder entrenar nuestros modelos y generar métricas de desempeño, se realiza la división de nuestro set de datos en dos sets: el de entrenamiento y el de validación. Las proporciones se realizaron como sigue:



Finalmente, como consideraciones adicionales para poder realizar los entrenamientos se realizaron los siguientes pasos:

- Normalización de la variable de género y
- Normalización de la variable de idioma
- Limpieza de las variables de tipo carácter
- Word embedding de las variables de tipo carácter
- Escalamiento de variables

Cabe mencionar que estos últimos pasos se realizan dentro del set de entrenamiento, ya que se busca que estos pasos preserven la estructura general y se apliquen de manera independiente al set de validación. Así, en caso de requerir generar una predicción de una muestra adicional, los pipelines de datos pueden funcionar de manera correcta y se puede generar la nueva predicción.

En las siguientes gráficas podemos ver algunos cambios que se suscitan al realizar los filtrados y estos últimos pasos:

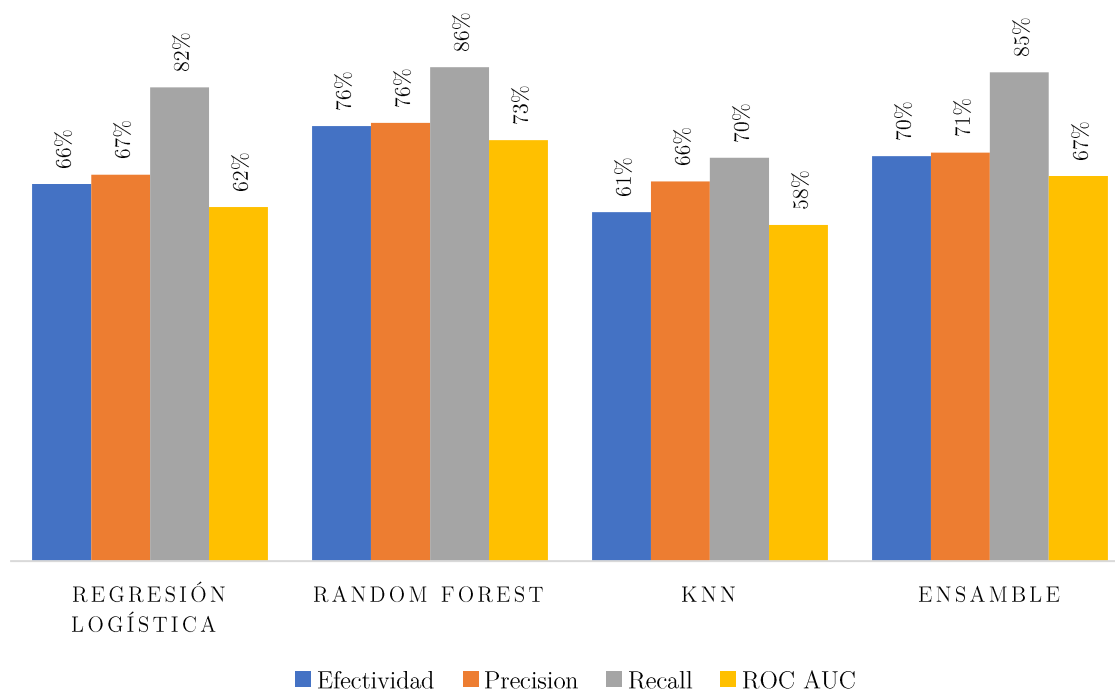
Resultados

Continuación, se exponen los resultados de los entrenamientos de los modelos anterior propuestos. Se entrenaron dos modelos para cada tipo, uno correspondiente a un modelo con parámetros default y otro obtenido al realizar el tuning de parámetros, eligiendo aquellos parámetros que lograron el mejor desempeño en el conjunto de validación usando Cross-Validation. Adicional, se consideraron dos métodos diferentes de word embedding, el método TF-IDF y el método Doc2Vec. Por lo cual, se tienen los resultados de los entrenamientos ocupando cada tipo de word embedding.

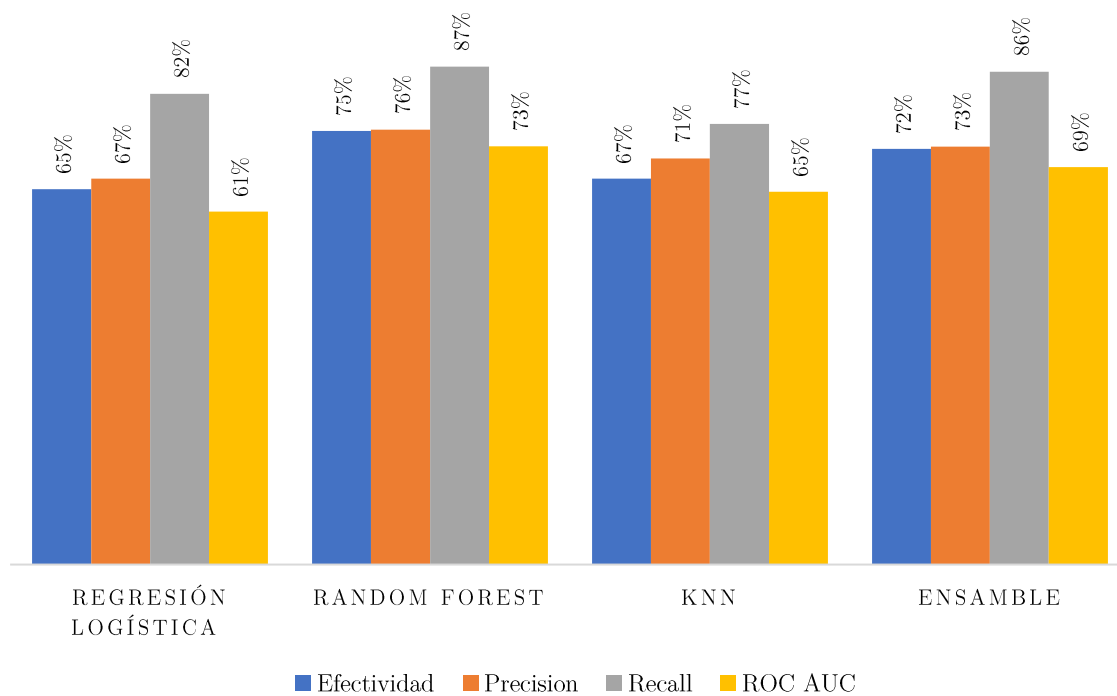
Es importante mencionar que se optó por elegir 100 features resultantes en cada word embedding y que las métricas expuestas fueron obtenidas evaluando los modelos en el conjunto de validación.

Modelos ocupando TF-IDF.

	Modelos default			
	Regresión logística	Random Forest	KNN	Ensamble
Efectividad	0.6558	0.7563	0.6070	0.7041
Precision	0.6722	0.7621	0.6602	0.7102
Recall	0.8240	0.8591	0.7014	0.8500
ROC AUC	0.6159	0.7320	0.5847	0.6695

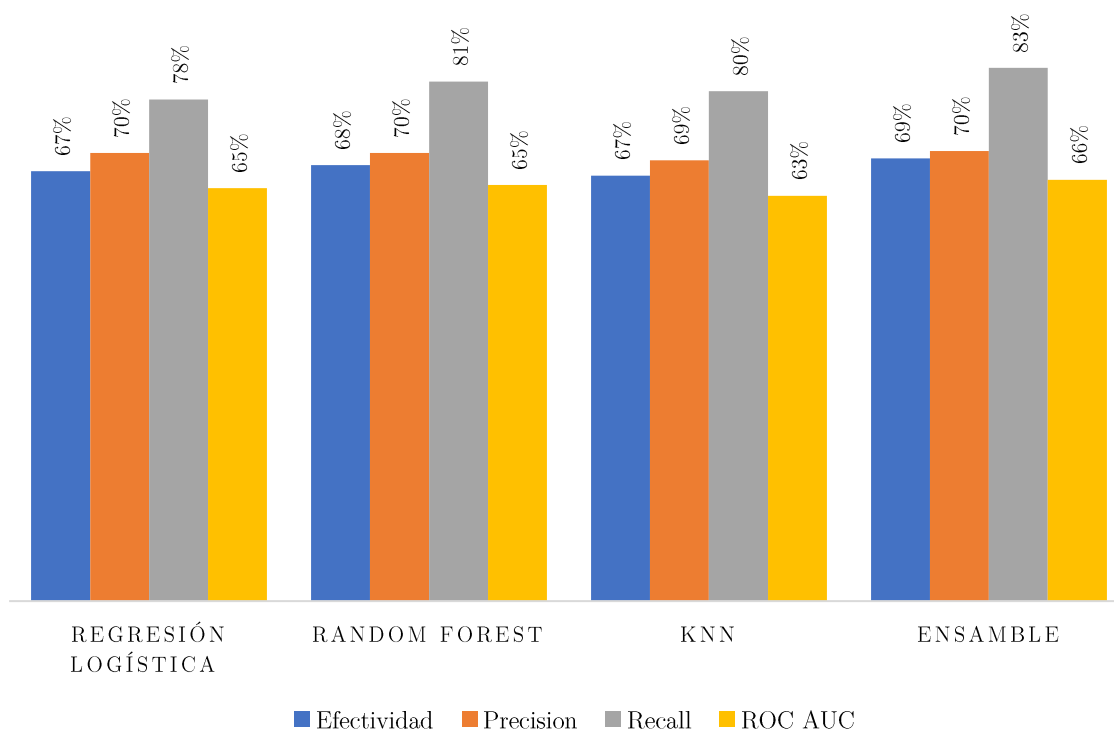


	Mejores modelos			
	Regresión logística	Random Forest	KNN	Ensamble
Efectividad	0.6528	0.7540	0.6709	0.7229
Precision	0.6709	0.7564	0.7062	0.7267
Recall	0.8188	0.8658	0.7664	0.8571
ROC AUC	0.6135	0.7275	0.6483	0.6911
Parámetros	C = 3.4568	max_depth = 50 min_samples_split = 4 n_estimators = 200	leaf_size = 50 n_neighbors = 50	

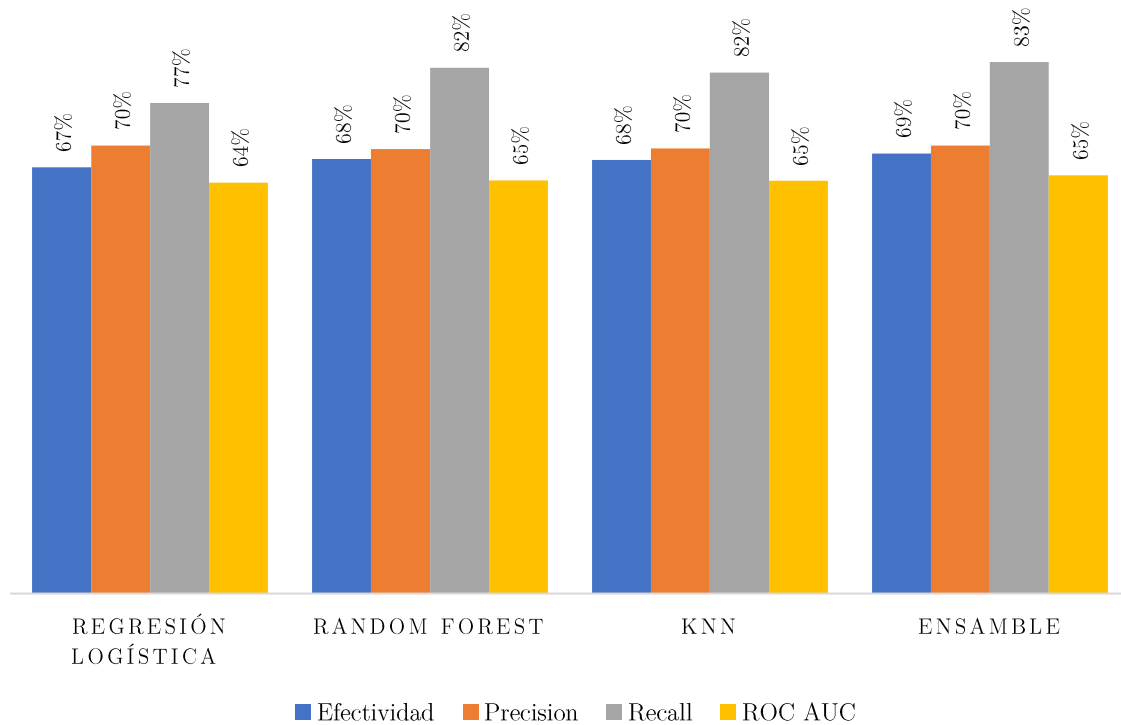


Modelos ocupando Doc2Vec.

	Modelos default			
	Regresión logística	Random Forest	KNN	Ensamble
Efectividad	0.6727	0.6824	0.6658	0.6930
Precision	0.7013	0.7013	0.6897	0.7047
Recall	0.7850	0.8132	0.7978	0.8344
ROC AUC	0.6461	0.6515	0.6345	0.6596



Mejores modelos				
	Regresión logística	Random Forest	KNN	Ensamble
Efectividad	0.6670	0.6802	0.6785	0.6885
Precision	0.7014	0.6958	0.6966	0.7010
Recall	0.7679	0.8230	0.8154	0.8317
ROC AUC	0.6431	0.6464	0.6460	0.6546
Parámetros	C = 8.2972	criterion = 'entropy' min_samples_leaf = 3 n_estimators = 50	leaf_size = 100 n_neighbors = 10	

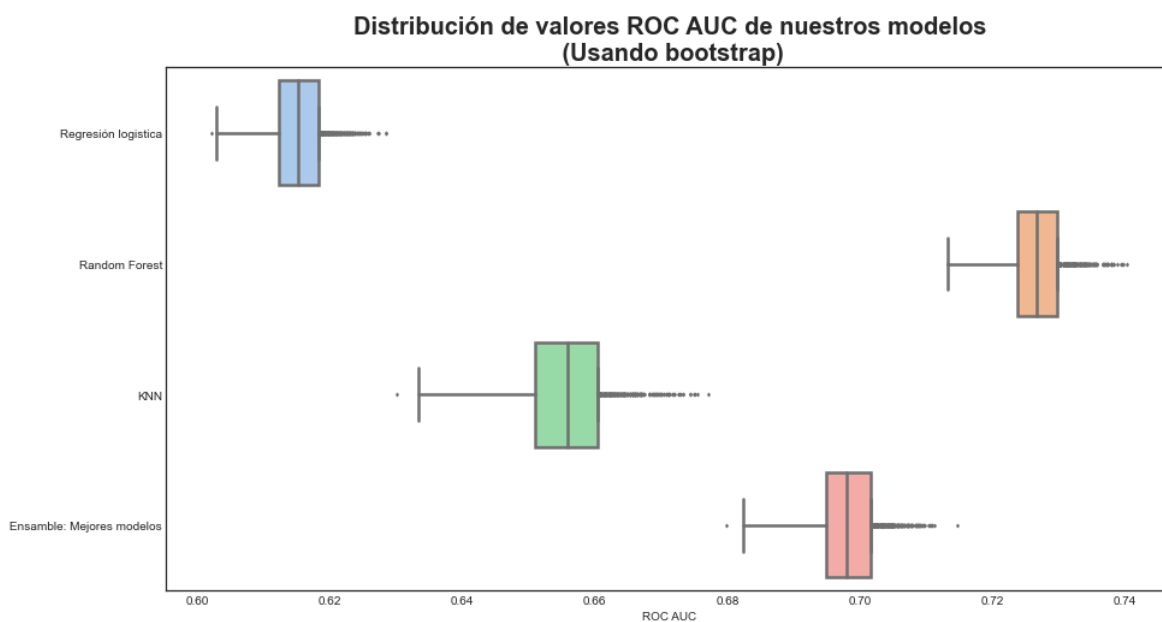


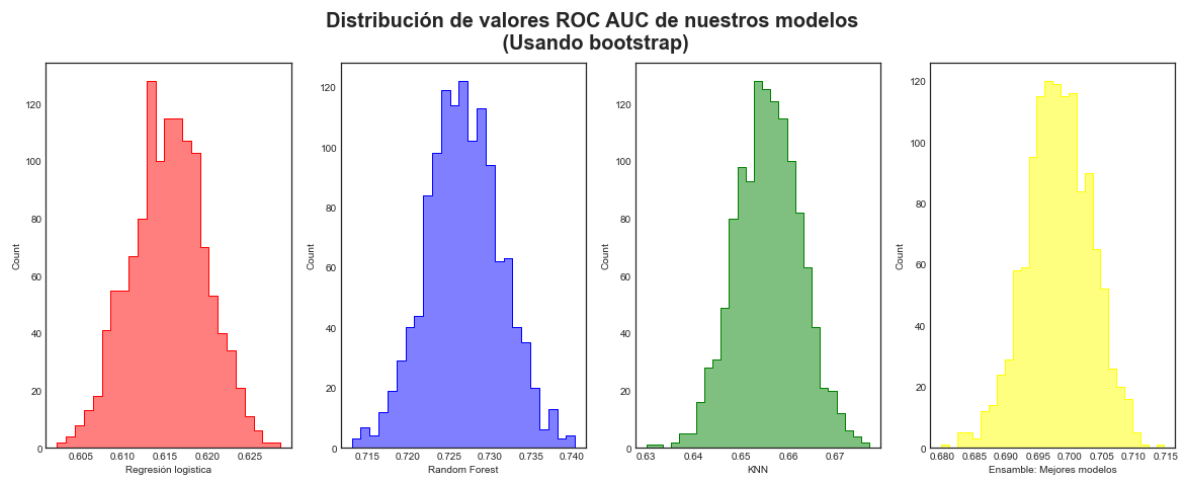
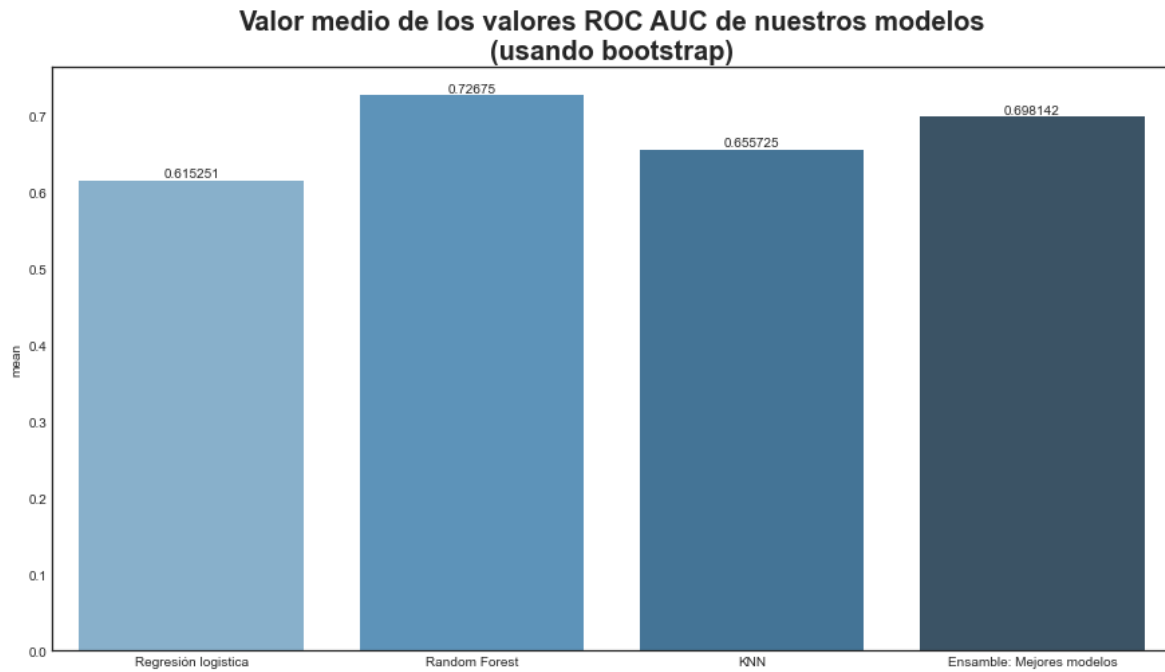
Estabilidad

A manera de verificar que nuestros modelos estén brindando métricas confiables y poder generar una elección de modelo, se realizó un proceso de bootstrap con los modelos obtenidos de hacer el tuning de parámetros.

Para este proceso, solamente se hizo bootstrap con los mejores modelos usando el embedding de TF-IDF, ya que los modelos usando Doc2Vec aparecen de que brindan peores resultados, el proceso de entrenamiento es considerablemente mayor al de TF-IDF, por lo cual, realizar bootstrap con este método de embedding resultaría en un tiempo excesivo de cómputo.

Los siguientes resultados recuperan el valor ROC AUC de cada modelo elegido realizando un total 1,250 iteraciones bootstrap que tomaron alrededor de 9 hrs de cómputo:





Por lo tanto, de estos resultados, podemos concluir que las métricas de validación, en efecto, son estables y es posible elegir nuestro modelo final.

Conclusiones

Las métricas obtenidas hacen evidente que los modelos ocupando Doc2Vec no brindan resultados que avalen la preferencia en su uso. Por lo que, si nos enfocamos en los modelos usando TF-IDF, el modelo Random Forest demuestra tanto en métricas bootstrap como en entrenamientos sencillos, una eficacia notoria comparando contra los demás modelos.

Por lo tanto, la elección de nuestro modelo final es un Random Forest con los siguientes parámetros:

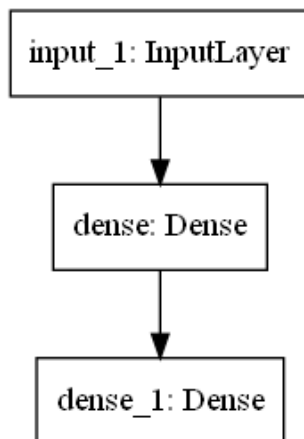
- `max_depth = 50`
- `min_samples_split = 4`
- `n_estimators = 200`

Haciendo uso de este modelo, podemos obtener una eficacia en nuestra clasificación de aproximadamente 75% y un valor de ROC AUC de aproximadamente 72%.

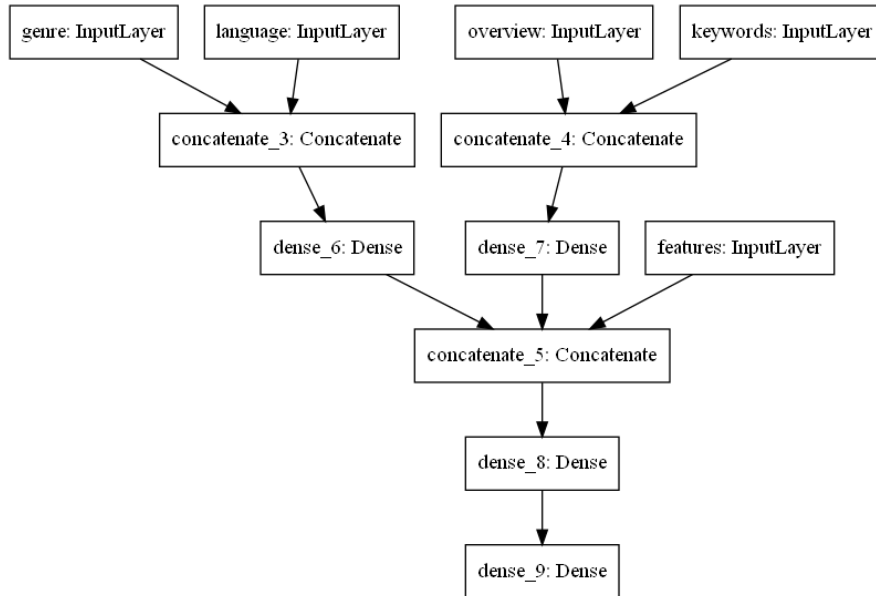
Redes neuronales

A manera comparativa, también se entrenaron tres tipos de redes neuronales. Todas se entrenaron con un early stopping de 50 épocas sobre la efectividad en el set de validación y las estructuras son las siguientes:

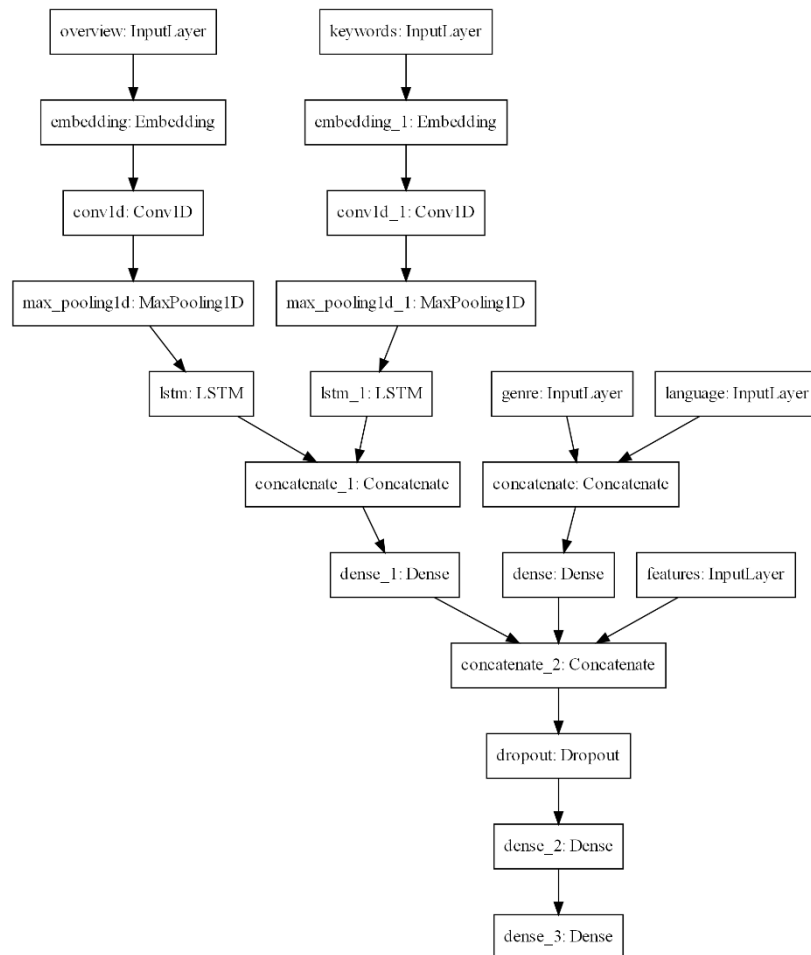
Estructura 1:



Estructura 2:



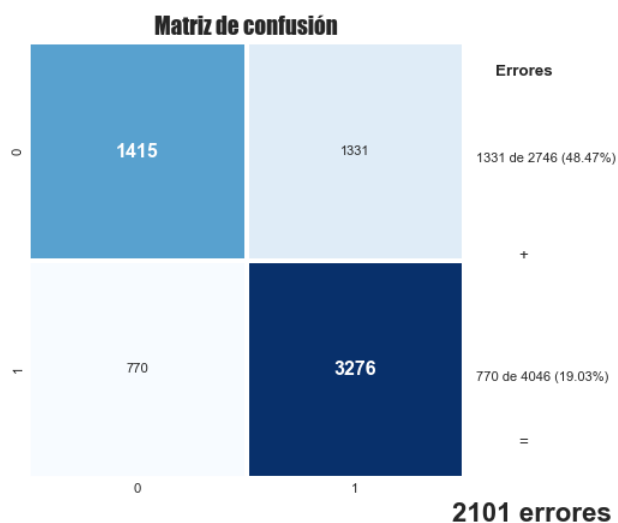
Estructura 3:



Resultados

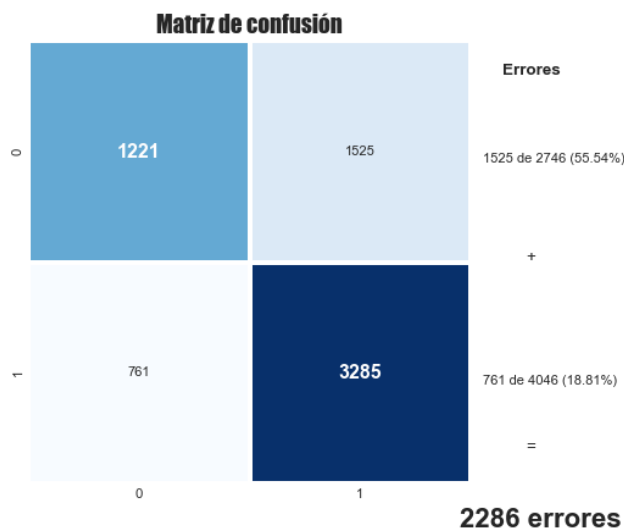
Redes neuronales (Estructura 1)

Efectividad: 69.067% Precision: 71.109%
Recall: 80.969% ROC AUC: 0.66249



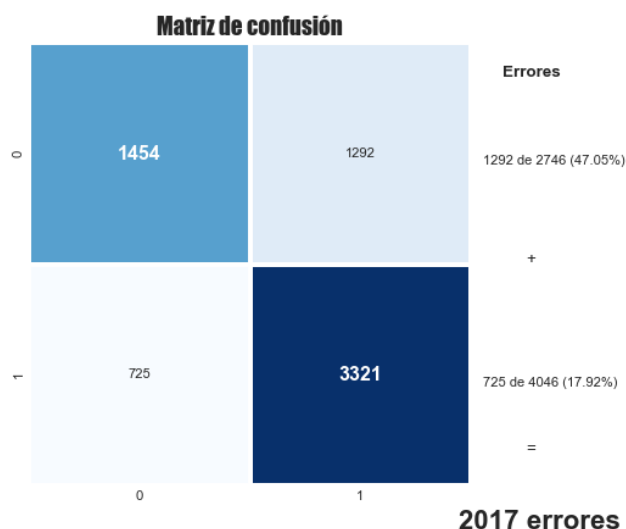
Redes neuronales (Estructura 2)

Efectividad: 66.343% Precision: 68.295%
Recall: 81.191% ROC AUC: 0.62828



Redes neuronales (Estructura 3)

Efectividad: 70.303% Precision: 71.992%
Recall: 82.081% ROC AUC: 0.67515



Comentarios finales

Los resultados de las redes anterior entrenadas muestran un buen desempeño (sobre todo la última estructura) dando pauta a un modelo incluso mejor que algunos modelos que anteriormente habíamos expuesto, sin embargo, las métricas obtenidas aun no son mejores que el modelo Random Forest que se había elegido.

A pesar de esto, las redes neuronales brindan posibilidades de establecer un sin fin de parámetros que podrían incluso superar las métricas del modelo que elegimos, sin embargo, para esto, se requieren hacer muchísimas pruebas exhaustivas en busca de esos parámetros.

Por lo tanto, aunque el uso de redes neuronales puede resultar beneficioso para ciertos tipos de problemas, en este caso, su uso no es tan eficiente si comparamos contra otros modelos menos complejos.