**Proyecto 1: Modelo de Clasificación**

Cecilia Abigail Monterroso

Universidad Francisco Marroquín

Facultad de Ciencias Económicas

Machine Learning Models

1. **Introducción**

Este reporte contiene un resumen de los puntos más importantes en el desarrollo de un modelo de clasificación para la predicción de valor binario que indica si un nominado gana o no un Oscar, basado en la información histórica de las ceremonias del ultimo siglo. A través de este análisis se pretende visualizar si los factores de sexo y etnicidad influencian fuertemente la elección de los ganadores de los premios de la academia de tal manera que se pueda hacer una predicción principalmente basado en estos valores.

Se hizo uso de dos algoritmos, KNN y SGD, los cuales se entrenaron a través de pipelines que contienen un pipe de preprocesamiento de variables categóricas y el algoritmo de clasificación respectivo. Para la optimización de estos se utilizo un RandomizedSearch y un GridSearch que resulto en un modelo KNN con 93% de precisión, aunque bastante desbalanceado con un recall de 10% con un f1 de 18%. El modelo SGD fue altamente deficiente y no dio ningún resultado relevante. 125

1. **Dataset**

Se escogió un dataset en kaggle que describe a los nominados a los Oscars desde 1927 hasta el 2020 en el que se encuentran 10,396 observaciones y 9 columnas con la siguiente información:

* Year\_film (int): El año en que la película fue nominada
* Year\_ceremony (int): El año en que se llevo a cabo la ceremonia de premios
* Ceremony (int): El número de ceremonia
* Category (object): La categoría en la que la persona fue nominada
* Gender (object): Sexo de la persona nominada
* Name (object): Nombre de la persona nominada
* Race (object): Etnicidad de la persona nominada
* Film (object): Nombre de la película por la que la persona fue nominada
* Winner (object): Verdadero, ganó un Oscar. Falso, no ganó

La variable winner será el target para futuro análisis. Considerando el bajo número de ganadores se hizo un gráfico de pie en el que se observa como únicamente el 23% de las observaciones describen el perfil de los ganadores (Ver Figura 1). De igual forma se utilizó la entropía de shanon para verificar el nivel de incertidumbre de la variable objetivo, dando como resultado de 0.21, que para propósitos de este análisis se considera suficiente para el desarrollo de los modelos.

Por otra parte, al hacer un análisis de la estadística descriptiva y una matriz de correlación se concluyó que todas las columnas numéricas están correlacionadas (Ver tabla 5), por lo que para beneficio del modelo se decidió dejar únicamente la variable year\_ceremony.

Dentro del análisis de variables categóricas se encontró que el dataset contiene 4,834 valores únicos de películas, así como 94 distintas categorías en la premiación. Concerniente a las características de los nominados se describe su sexo, ya sea femenino o masculino, y se listan 4 distintas etnicidades Hispano, asiático, afrodescendiente y blanco

El perfil predominante en las nominaciones es el del hombre blanco. Siendo que el genero masculino representa el 80% de las observaciones, así como la etnicidad blanca abarca aproximadamente un 95%. El menor porcentaje de nominados son los hispanos con un 0.38%. Profundizando en la etnicidad de los nominados, se puede ver a través de una gráfica de violín que la diversidad de la premiación ha ido en aumento desde los 2000, especialmente en las personas afrodescendientes que tienen una mayor media en comparación a las otras minorías representados en este dataset.

Imagen que contiene Gráfico

Descripción generada automáticamente

**Figura 2.** Distribución de etnicidad en nominados a través del tiempo

1. **Metodología**

Para la optimización del modelo se seleccionaron las metricas de precisión y f1; ya que la primera es de gran importancia aumentar el número de verdaderos positivos para encontrar un patrón en los ganadores, mientras que la segunda es especialmente útil para evaluar modelos basados en datasets poco balanceados como el que se ha escogido para este análisis.

* 1. Preprocesamiento

Se creo un pipeline para automatizar el preprocesamiento de datos categóricos. Se hizo uso de un imputador (SimpleImputer) para manejar los valores faltantes en el dataset, siendo estos reemplazado por un valor genérico que no altere significativamente el posterior procesamiento. Por consiguiente, también se utilizó un preprocesador (OneHotEncoder) para crear una columna binaria para cada valor único de cada una de los features para la correcta interpretación de estos datos al momento de procesar los modelos requeridos.

Agregado al pipeline se aplico un LabelEncoder a la variable objetivo para codificar las etiquetas categóricas (verdadero y falso) en valores numéricos para que sea posible el procesamiento por los algoritmos de aprendizaje.

* 1. Clasificador KNN

Un algoritmo de aprendizaje supervisado y no paramétrico. Para su uso se creó un nuevo pipeline en el que se integró el pipeline de preprocesamiento para variables categóricas y el clasificador para su fase de entrenamiento.

Para evaluar el desempeño del modelo inicial se realizó un reporte de clasificación que lanzo los resultados que se observan en la tabla 1. El modelo se muestra bastante deficiente para la predicción de ganadores, con un f1 de 39% y un recall de 27%.

precision recall f1-score support

0 0.82 0.97 0.89 5608

1 0.71 0.27 0.39 1669

accuracy 0.81 7277

macro avg 0.76 0.62 0.64 7277

weighted avg 0.79 0.81 0.77 7277

**Tabla 1.** Reporte de clasificación KNN modelo 1

Se hizo uso de un RandomizedSearch para analizar distintos parámetros para n\_neighbors y leaf\_size, que luego fueron utilizados en un Grid Search para la búsqueda de los mejores hiperparametros que optimizaran el modelo de clasificación. Asimismo, se escogió el error promedio como estrategia de refiting ya que se quiere mejorar la confiabilidad del modelo y reducir la posibilidad de un falso positivo o un falso negativo.

* 1. Clasificador SGD

Un algoritmo de aprendizaje supervisado y paramétrico. Igualmente se hizo uso de un nuevo pipeline para el preprocesamiento y el entrenamiento del modelo. Los resultados de la evaluación del modelo 1 se muestran en la tabla 2. Se puede ver un peor desempeño del modelo inicial en comparación con el resultado del algoritmo de KNN.

precision recall f1-score support

0 0.82 0.92 0.87 2431

1 0.49 0.27 0.34 688

accuracy 0.78 3119

macro avg 0.65 0.59 0.61 3119

weighted avg 0.74 0.78 0.75 3119

**Tabla 2.** Reporte de clasificación SGD modelo 1

Se utilizó la misma técnica al usar el RandomizedSearch para filtrar los mejores modelos por intervalos de los parámetros alpha, max\_iter y eta0para después aplicarlos a un Grid Search. Asimismo, se agregó early\_stopping = True para que reducir el tiempo de entrenamiento si se muestran indicios de que las metricas no están mejorando.

1. **Análisis de Resultados**

Se escogió la matriz de confusión como método de comparación del desempeño de ambos modelos.

* 1. Clasificador KNN

La matriz de confusión de los resultados se muestra en la figura 2. Gracias a la optimización del modelo se logró conseguir un 92% de precisión, y como se puede ver en la matriz, los falsos positivos son mínimos. Se considera un resultado adecuado para propósitos de este proyecto, ya que se deseaba optimizar la precisión de predicción con acceso a pocas variables para evaluar su impacto. No obstante, se logró un modelo bastante desbalanceado a costa de un recall de 10% con un f1 de 18%.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Figura 3.** Matriz de confusión de predicciones (KNN)

* 1. Clasificador SGD

A pesar de las diferentes variaciones que se escogieron para optimizar los hiperparametros, el modelo no predice ningún valor positivo. Por lo tanto, su pobre desempleo da como resultado 0 en todas las metricas de evaluación pertinentes. Para este escenario no se puede utilizar este modelo para sacar alguna opinión concluyente sobre el comportamiento de las variables utilizadas.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Figura 4.** Matriz de confusión de predicciones (SGD)

1. **Conclusión**

Aunque las variables de etnicidad y sexo si tienen cierta influencia sobre la decisión de los ganadores, aisladas no son suficiente para determinar la probabilidad de ganar un oscar. A pesar de que el modelo de KNN dio una alta precisión, este fue poco balanceado y por tanto no se recomendaría como un modelo completamente fiable para otros propósitos. Se concluye que para ganar un oscar se necesitan de otras características que pueden ser más relevantes para la elección.

Se realizan las siguientes recomendaciones para futuro análisis:

* Trabajar con un dataset mucho más balanceado para evitar sesgos debido a la falta de observaciones en algunas clases. Se recomienda hacer una técnica de oversampling o undersampling.
* Complementar con un mayor número de variables que tengan un mayor impacto en la elección de ganadores.
* Un modelo de regresión lineal puede mostrar de manera más clara la influencia de

1. **Appendix**

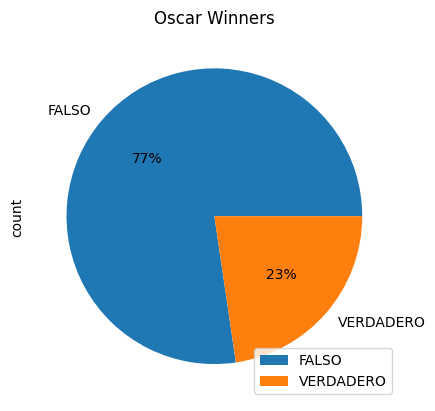
Thuwarakesh, Thuwarakesh, By, Strategy, P. in, & General, P. in. (2023, June 27). *How do you handle imbalanced dataset*. The Analytics Club. <https://www.the-analytics.club/how-to-check-if-data-is-imbalanced/>

GfG. (2022, February 10). *Plotting different types of plots using factor plot in Seaborn*. GeeksforGeeks. <https://www.geeksforgeeks.org/plotting-different-types-of-plots-using-factor-plot-in-seaborn/>

*What is the K-nearest neighbors algorithm?*. IBM. (n.d.). <https://www.ibm.com/topics/knn>

*Custom refit strategy of a grid search with cross-validation*. scikit. (n.d.). <https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html>

*F1 score in Machine Learning: Intro & Calculation*. V7. (n.d.). <https://www.v7labs.com/blog/f1-score-guide>



**Figura 5.** Matriz de confusión de clasificación KNN modelo 1

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Figura 5.** Matriz de confusión de clasificación KNN modelo 1

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Figura 6.** Matriz de confusión de clasificación SGD modelo 1

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Imagen 1.** Modelo optimizado de clasificación KNN

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Imagen 2.** Modelo optimizado de clasificación SGD

precision recall f1-score support

0 0.80 1.00 0.89 2431

1 0.92 0.10 0.18 688

accuracy 0.80 3119

macro avg 0.86 0.55 0.53 3119

weighted avg 0.82 0.80 0.73 3119

**Tabla 3.** Reporte de clasificación KNN optimizado

precision recall f1-score support

0 0.78 1.00 0.88 2431

1 0.00 0.00 0.00 688

accuracy 0.78 3119

macro avg 0.39 0.50 0.44 3119

weighted avg 0.61 0.78 0.68 3119

**Tabla 4.** Reporte de clasificación SGD “optimizado”

| **year\_film** | **year\_ceremony** | **ceremony** |
| --- | --- | --- |
| year\_film | 1.000000 | 1.000000 | 0.999983 |
| year\_ceremony | 1.000000 | 1.000000 | 0.999983 |
| ceremony | 0.999983 | 0.999983 | 1.000000 |
|  |  |  |  |

**Tabla 5.** Tabla de correlación de variables númericas