UNIVERSIDAD DEL VALLE DE GUATEMALA

Proyecto 2

Andre Marroquin Tarot- 22266

Minería de Datos

Explicación del método utilizado para obtener los conjuntos de entrenamiento y de prueba:

En este caso el método seleccionado para entrenar ambos conjuntos; se basa en agarrar el 20% del conjunto de datos para la prueba y el 80% para el entrenamiento. Se le aplica un random state 42 para que los datos sean replicables. Se divide en cuatro los datos para entrenar al modelo. Las 4 variables son X_train, X_test, y_train, y_test.

Selección de la variable respuesta y explicacion de las transformaciones:

La variable respuesta que se decidió elegir en este caso es ‘Revenue’ qué son los ingresos totales que obtuvo la película. Es de tipo numérico continuo. De ahí el resto del dataframe se tomaron como variables de X para ayudar a predecir, sin embargo se eliminaron 2 ya que no tenían ninguna relevancia con la predicción de los datos. Se eliminó Title y Description.

Transformación de Características Numéricas

- Imputación: Se reemplazan los valores faltantes con la mediana de la columna correspondiente. La mediana es buena opción para imputación ya que es menos sensible a los valores atípicos en comparación con la media.
- Estandarización: Se escalan las características para que tengan media cero y una desviación estándar de uno. Esto es fundamental para modelos que son sensibles a la magnitud de las características, como la regresión lineal o los modelos basados en distancias.

Transformación de Características Categóricas

- Imputación: Se reemplazan los valores faltantes por una etiqueta genérica como 'missing'. Esto asegura que el modelo pueda manejar datos faltantes sin descartar la fila entera.
- Codificación One-Hot: Se transforman las variables categóricas en vectores binarios, permitiendo que el modelo interprete adecuadamente estas características sin asumir un orden arbitrario.

Manejo de Valores Faltantes en la Variable Objetivo

Para la variable objetivo, que es 'Revenue (Millions)', se manejan los valores faltantes utilizando la estrategia de imputación más frecuente, que reemplaza los valores faltantes con el valor más comúnmente observado en esa columna. Esto prepara el vector objetivo para el entrenamiento sin descartar datos debido a la falta de ingresos reportados.

Aplicación de algoritmos seleccionados:

Se seleccionan tres modelos de aprendizaje automático diferentes:

- Random Forest: Un modelo basado en árboles que utiliza múltiples árboles de decisión para obtener una predicción más robusta y menos propensa al sobreajuste que un solo árbol de decisión. El modelo es útil para manejar una variedad de tipos de datos y relaciones, proporcionando una buena métrica de la importancia de las características.

- SVM (Support Vector Machine): Este modelo es efectivo en espacios de alta dimensión y en casos donde el número de dimensiones es mayor que el número de muestras. Utiliza un hiperplano para clasificar datos o, en el caso de regresión (SVR), para ajustar una función dentro de un umbral determinado.

- Regresión Lineal: Un modelo de aprendizaje automático que intenta predecir un valor objetivo mediante la combinación lineal de características de entrada. Es rápido de ejecutar y proporciona una base sólida para comparar el rendimiento de modelos más complejos.

Ajuste de Hiperparámetros
Para cada modelo, se definen hiperparámetros que serán optimizados para mejorar el rendimiento del modelo. Esto se hace a través de un proceso llamado ajuste de hiperparámetros, que busca encontrar la combinación de parámetros que produce el mejor resultado en función de una métrica específica.

Random Forest:

- n_estimators: El número de árboles en el bosque. Valores típicos pueden ser 100, 200, o 300.
- max_depth: La profundidad máxima de cada árbol. Se explora desde árboles sin restricción de profundidad hasta árboles con profundidades de 10, 20, o 30 niveles.
- min_samples_split: El número mínimo de muestras necesarias para dividir un nodo interno. Los valores a considerar son 2, 5 o 10.

SVM:

- C: Parámetro de regularización que ayuda a controlar el compromiso entre lograr un margen bajo y asegurar que la mayoría de los puntos de datos estén clasificados correctamente.
- gamma: Coeficiente para los kernels no lineales. Afecta la influencia de cada punto de datos individual.
- kernel: Tipo de función del núcleo a usar en el entrenamiento. Las opciones son 'rbf', 'poly', y 'sigmoid'.

Predicción de modelos:

Luego se predicen los 3 distintos modelos aplicados anteriormente algunos con hiperparametros para obtener su mejor rendimiento y una mejor predicción.

Generación de matrices de confusión y explicación de resultados obtenidos:

Matriz de Confusión:

- Baja (Low): De las observaciones reales bajas, 55 fueron predichas correctamente como bajas, pero 20 fueron incorrectamente predichas como medias y 2 como altas.
- Media (Medium): De las observaciones reales medias, 37 fueron correctamente identificadas como medias, mientras que 5 fueron predichas como bajas y 15 como altas.
- Alta (High): De las observaciones reales altas, 52 fueron correctamente predichas como altas, 12 fueron predichas como medias y 2 como bajas.

Métricas de Rendimiento:

- MAE (Error Absoluto Medio): 38.74, indica en promedio qué tan alejadas están las predicciones del modelo de los valores reales.
- MSE (Error Cuadrático Medio): 4534.27, pone más énfasis en los errores grandes debido al cuadrado de las diferencias.
- RMSE (Raíz del Error Cuadrático Medio): 67.34, proporciona una medida de la magnitud de los errores en las mismas unidades que los datos.
- MAPE (Error Porcentual Absoluto Medio): 9204.95%, muestra que, en promedio, las predicciones pueden estar alejadas en términos porcentuales. Este valor muy alto podría ser indicativo de outliers o errores en el cálculo.
- R2 (Coeficiente de Determinación): 0.556, indica que aproximadamente el 55.6% de la variabilidad en los datos reales es explicada por el modelo.

La matriz muestra una capacidad razonable del modelo para clasificar correctamente las observaciones en las categorías baja y alta, aunque presenta más dificultades con las observaciones medias. Los errores, especialmente el MAPE alto, dice que puede haber casos donde las predicciones del modelo están muy desviadas de los valores reales.

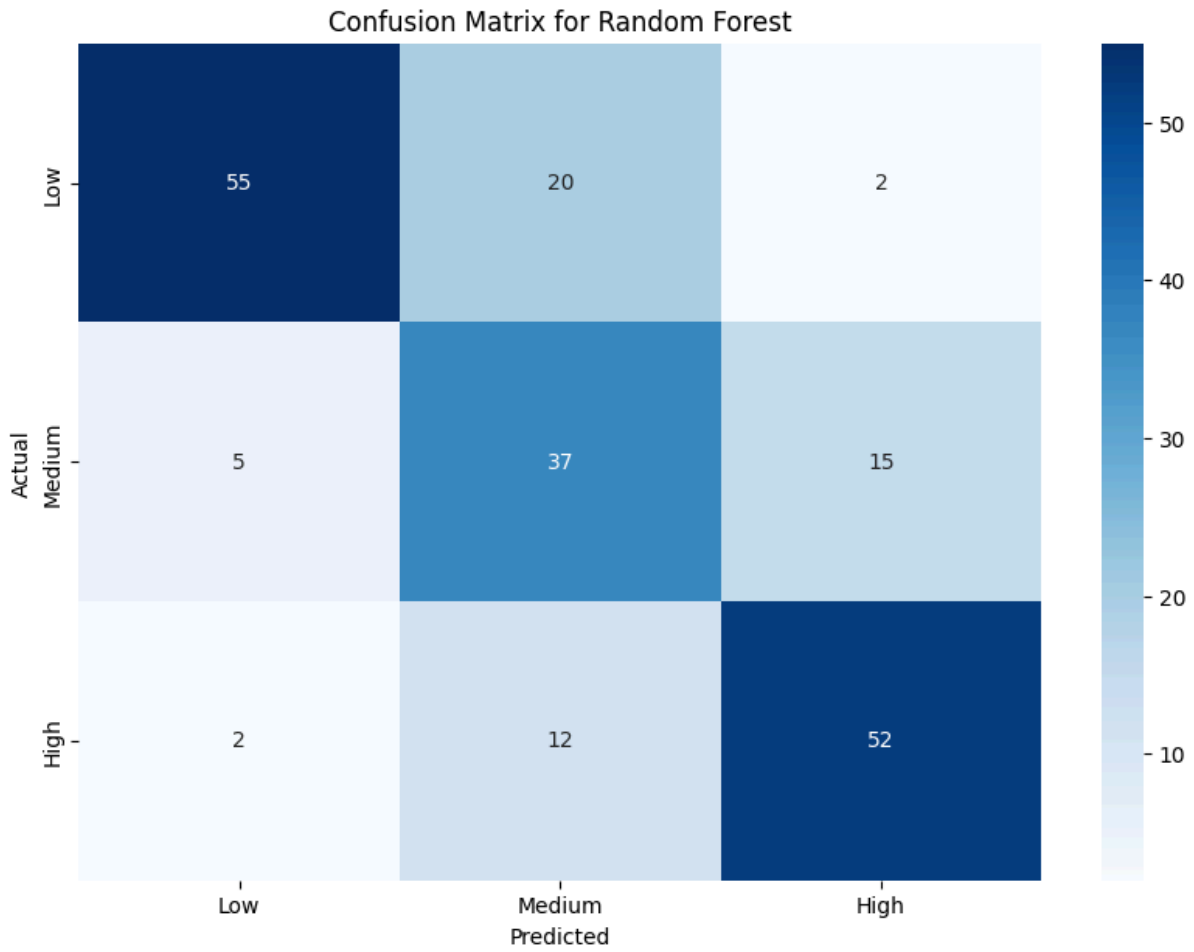

Gráfico 1: matriz confusión Random forest

Matriz de Confusión:

- Baja (Low): De las observaciones reales bajas, 62 fueron correctamente clasificadas como bajas, mientras que 13 fueron incorrectamente clasificadas como medias y 2 como altas.
- Media (Medium): De las observaciones medias, 40 fueron correctamente identificadas como medias, 4 fueron clasificadas incorrectamente como bajas y 13 como altas.
- Alta (High): De las observaciones altas, 47 fueron correctamente predichas como altas, 17 como medias y 2 como bajas.

Métricas de Rendimiento:

- MAE (Error Absoluto Medio): 39.43, un error promedio entre las predicciones y los valores reales.
- MSE (Error Cuadrático Medio): 4677.01, se enfoca en los errores grandes al elevar al cuadrado las diferencias.
- RMSE (Raíz del Error Cuadrático Medio): 68.39, proporciona una perspectiva de la magnitud de los errores en las mismas unidades que los valores observados.

- MAPE (Error Porcentual Absoluto Medio): 8770.11%, indica errores grandes en términos porcentuales, lo cual dice que algunas predicciones están significativamente desviadas.
- R2 (Coeficiente de Determinación): 0.543, muestra que cerca del 54.3% de la variabilidad en los datos observados es explicada por el modelo.

Logra un alto grado de precisión al identificar correctamente la mayoría de las observaciones bajas y altas, aunque enfrenta algunos retos al clasificar las observaciones medias. Las métricas de rendimiento, incluyendo un MAE de 39.43 y un RMSE de 68.39, indican un error moderado en las predicciones. Aunque el MAPE es elevado, reflejando errores significativos en algunos casos, el coeficiente R2 de 0.543 sugiere que el modelo es capaz de explicar más de la mitad de la variabilidad observada en los datos.

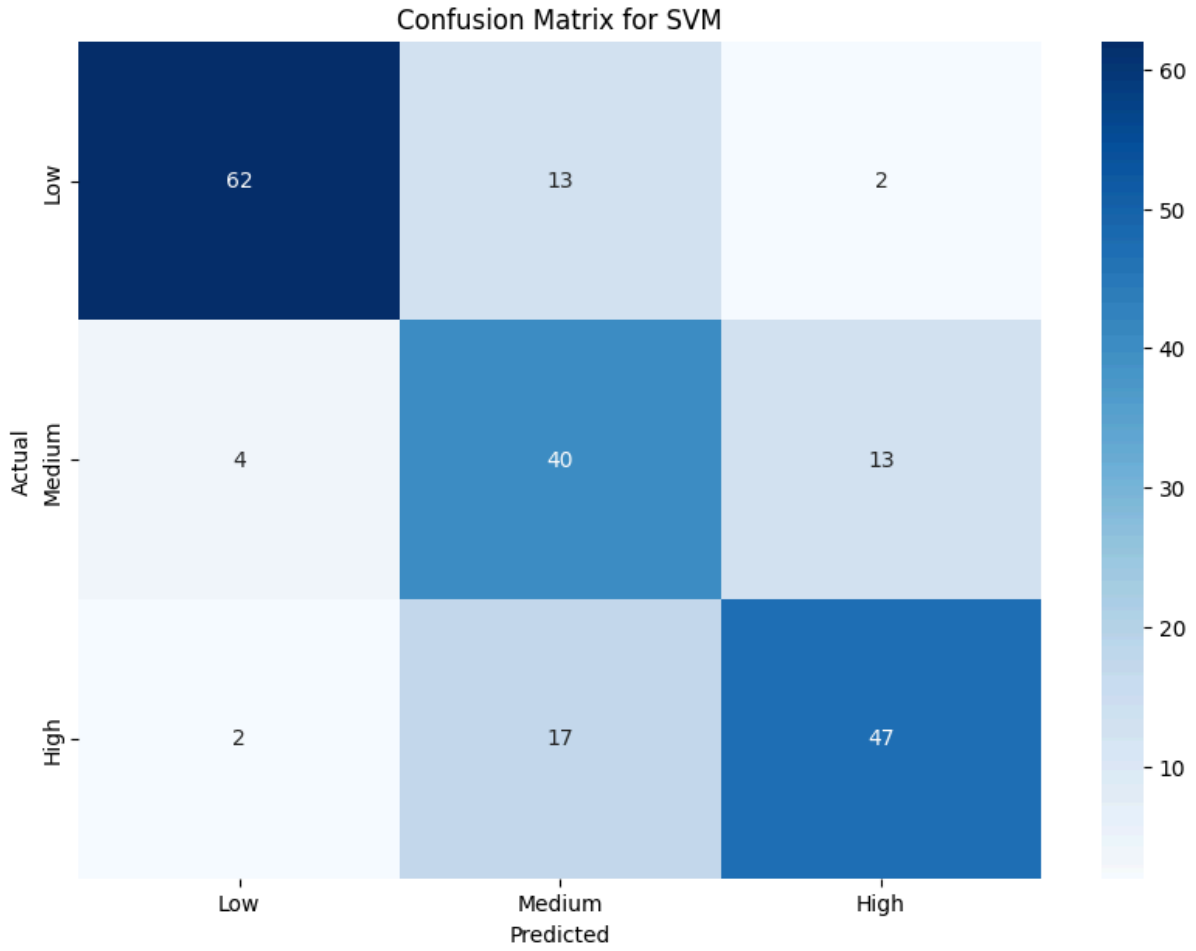

Gráfico 2: matriz confusión SVM

Matriz de Confusión:

- Baja (Low): De las observaciones reales bajas, 43 fueron correctamente clasificadas como bajas, 32 como medias y 2 como altas.

- Media (Medium): De las observaciones medias, 31 fueron correctamente clasificadas como medias, mientras que 5 fueron incorrectamente clasificadas como bajas y 21 como altas.
- Alta (High): De las observaciones altas, 57 fueron correctamente predichas como altas, con 7 predicciones incorrectas como medias y 2 como bajas.

Métricas de Rendimiento:

- MAE (Error Absoluto Medio): 38.74, sugiere un error promedio modesto entre las predicciones y los valores reales.
- MSE (Error Cuadrático Medio): 3442.97, indica que los errores grandes tienen un peso considerable en el modelo.
- RMSE (Raíz del Error Cuadrático Medio): 58.68, proporciona una perspectiva de los errores en las mismas unidades que los datos.
- MAPE (Error Porcentual Absoluto Medio): 14966.44%, muestra una variabilidad extremadamente alta en algunos casos, lo que podría indicar problemas significativos en la capacidad predictiva del modelo en ciertas observaciones.
- R2 (Coeficiente de Determinación): 0.664, este es el valor más alto entre los modelos examinados, lo que indica que la Regresión Lineal explica una mayor proporción de la variabilidad de los datos.

En general, el modelo de Regresión Lineal parece ser eficaz, especialmente en la categoría alta, aunque el alto MAPE destaca la presencia de predicciones extremadamente inexactas en ciertos casos. La mayor capacidad explicativa, según el R2, sugiere que este modelo maneja mejor la variabilidad de los datos en comparación con los modelos anteriores.

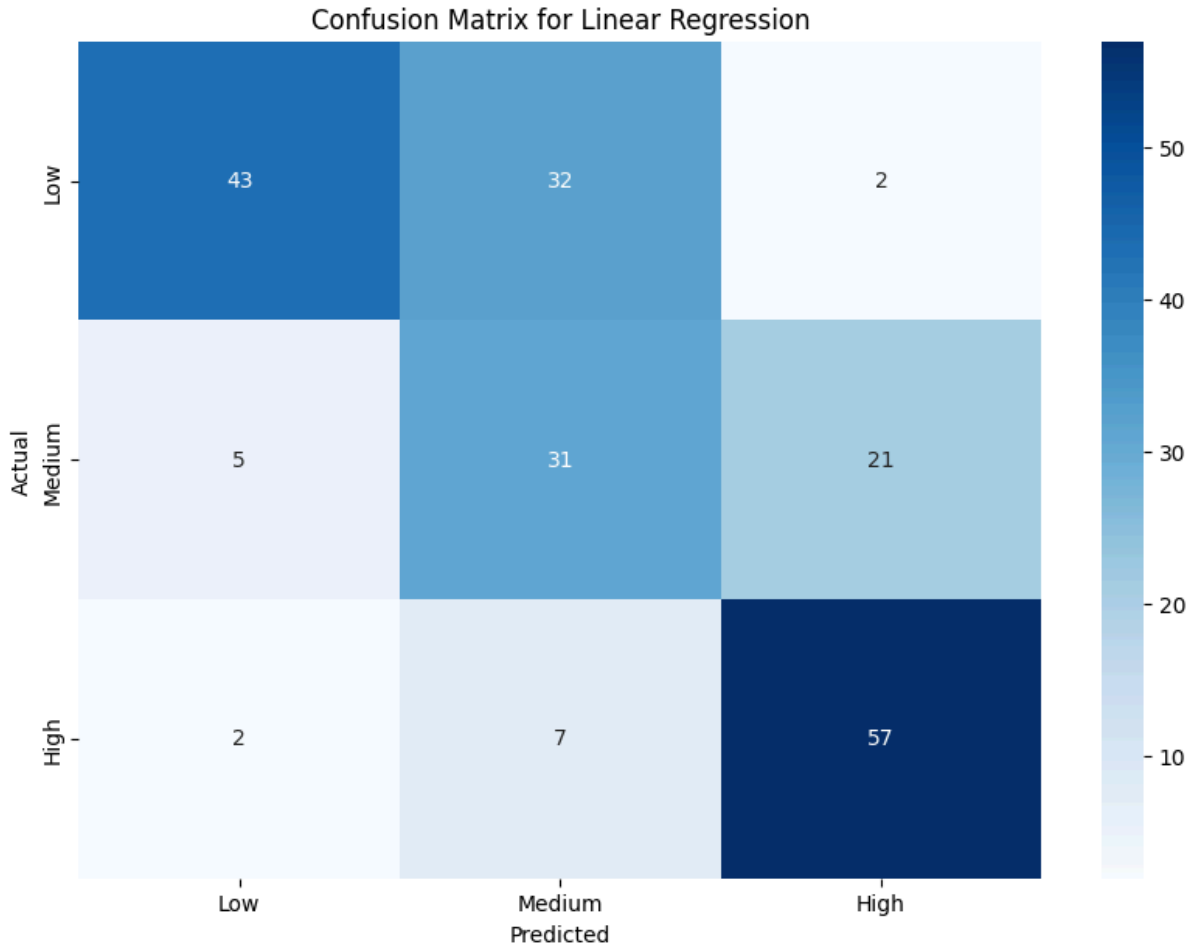

Gráfico 3: matriz confusión Regresión Lineal

Conclusiones Generales:

Después de varios intentos de mejorar los resultados de las predicciones de los 3 modelos, se llegó a la conclusión que los resultados obtenidos son la mejor predicción que se pudo llegar a tener. A pesar de que no predice del todo bien ya que puede ser por factores de los datos en sí entre otros, es la mejor predicción obtenida en todas las distintas pruebas y predicciones realizadas. Se aplicaron hiperparametros se derivaron columnas, se eliminaron algunas pero aun así se obtuvo el mejor resultado con el proceso establecido de último.

La Regresión Lineal tiene la mejor capacidad explicativa, pero su alto MAPE indica problemas con valores atípicos. Random Forest ofrece un equilibrio razonable en la clasificación de los extremos, aunque requiere mejoras en la categoría media. SVM, por su parte, es comparable en rendimiento a Random Forest pero necesita reducir las grandes desviaciones en sus predicciones. La elección del modelo más adecuado depende del

contexto específico y del balance deseado entre precisión general y manejo de valores atípicos.

Referencias:

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297. https://link.springer.com/article/10.1007/BF00994018

Montgomery, D. C., & Peck, E. A. (2019). Introduction to linear regression analysis. John Wiley & Sons. https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG.GeoffreyViningz-lib.org.pdf

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324