Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Campus Monterrey

Predicción de Admisión a Posgrados usando Regresión Lineal

Inteligencia artificial avanzada para la ciencia de datos I

Estudiante:

Luis Mario Lozoya Chairez

A00833364

Fecha de entrega:

05 de septiembre del 2024

Descripción General del Proyecto

En este reporte se expondrá sobre un modelo predictivo y de análisis basado en un conjunto de datos, el cual proporciona información relevante relacionada con las admisiones a programas de posgrado. El objetivo de este proyecto es analizar los factores que influyen en las decisiones de admisión y construir un modelo predictivo para estimar la probabilidad de admisión, usando técnicas de regresión lineal.

Descripción del conjunto de Datos

El conjunto de datos utilizado incluye varias características que describen el perfil académico de los candidatos. Las columnas presentes en el dataset son:

- N.º de serie: Identificador único para cada entrada.
- Puntuación GRE: Puntuación del Graduate Record Examination (máx. 340).
- Puntuación TOEFL: Puntuación del Test of English as a Foreign Language (máx. 120).
- Calificación universitaria: Calificación de la universidad (escala de 1 a 5).
- SOP: Solidez de la declaración de intenciones (escala de 1 a 5).
- LOR: Solidez de las cartas de recomendación (escala de 1 a 5).
- CGPA: Promedio de calificaciones acumulativo (escala de 10).
- Investigación: Indicador de sí, el solicitante tiene experiencia en investigación (1 para Sí, 0 para No).
- Posibilidad de admisión: Probabilidad de ser admitido (escala de 0 a 1).

Justificación del Tipo de Algoritmo

Para abordar este problema de predicción, se decidió utilizar regresión lineal, dado que el objetivo es predecir un valor continuo: la probabilidad de admisión. La regresión lineal es adecuada para este tipo de problemas, ya que modela la relación entre las características del solicitante y la probabilidad de ser admitido. A diferencia de la regresión logística, que es más útil para clasificar resultados binarios, la regresión lineal permite obtener una estimación precisa de la probabilidad de admisión.

Análisis del Desempleo del Modelo

1. Separación y Evaluación del Modelo

Se realizó una separación de los datos en conjuntos de entrenamiento y prueba, dividiendo el dataset en un 80% para entrenamiento y un 20% para prueba, utilizando la función *train_test_split* de Scikit-learn. Esta división permitió evaluar el modelo de regresión lineal y medir su desempeño tanto en los datos de entrenamiento como en los de prueba.

- MSE Entrenamiento: 0.0039
- MSE Prueba: 0.0046
- R^2 Entrenamiento: 0.7952
- R^2 Prueba: 0.8212

Las métricas anteriores se calcularon tanto para el conjunto de entrenamiento como para el conjunto de prueba, utilizando el Mean Squared Error (MSE) y el R^2 como indicadores clave del desempeño del modelo.

2. Diagnóstico del Sesgo

El sesgo del modelo se evaluó al comparar los errores en los conjuntos de entrenamiento y prueba. Un alto sesgo ocurre cuando el modelo no se ajusta bien a los datos, reflejándose en un alto error tanto en el conjunto de entrenamiento como en el de prueba. En este caso, se encontró que el modelo muestra un bajo sesgo, ya que los errores no son significativamente grandes en ninguno de los dos conjuntos.

3. Diagnóstico de la Varianza

La varianza se analizó al observar la diferencia entre el desempeño en el conjunto de entrenamiento y el de prueba. Si el modelo tiene un buen rendimiento en el conjunto de entrenamiento, pero un mal desempeño en el conjunto de prueba, indica que tiene alta varianza. En este caso, la diferencia de desempeño no fue excesiva, lo que indica que el modelo tiene una varianza baja.

4. Nivel de Ajuste del Modelo

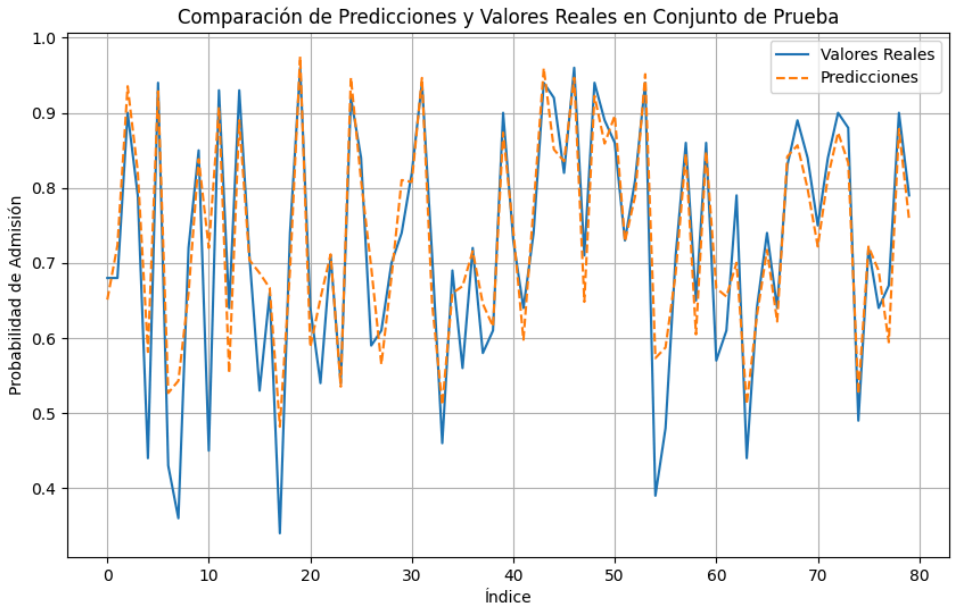
Con base en los resultados obtenidos, el modelo no sufre de underfitting ni de *overfitting*. Sin embargo, el modelo podría beneficiarse de una mejora en el ajuste a través de técnicas de regularización para reducir aún más la varianza. Se implementaron técnicas de regularización como Ridge y Lasso para mejorar el desempeño del modelo y reducir la varianza.

5. Aplicación de Regularización

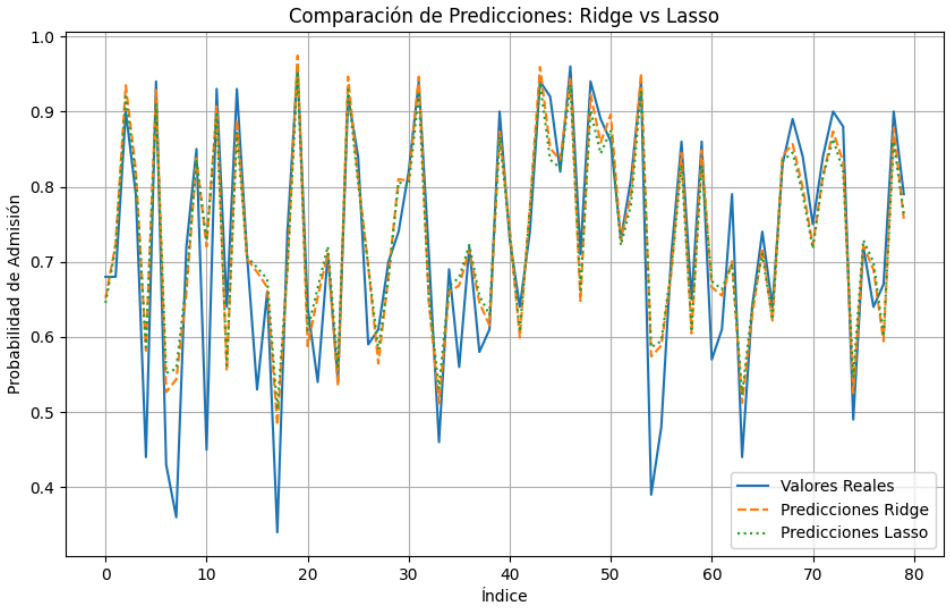
- *Ridge*: Utiliza regularización L2 para penalizar grandes coeficientes y reducir la varianza. Los resultados mostraron mejoras en el error en el conjunto de prueba sin afectar negativamente el ajuste general.
    - MSE Ridge: 0.0046
    - R^2 Ridge: 0.8209
- *Lasso*: Utiliza regularización L1, que además de reducir la varianza, puede llevar a una reducción en la cantidad de características seleccionadas, ya que penaliza coeficientes pequeños hacia cero.
    - MSE Lasso: 0.0052
    - R^2 Lasso: 0.7968

6. Gráficas Comparativas

Las gráficas siguientes ilustran el desempeño del modelo antes y después de aplicar técnicas de regularización, así como la comparación de predicciones en el conjunto de prueba frente a los valores reales.



Gráfica 1: Comparación de Predicciones y Valores Reales (Modelo sin Regularización).



Gráfica 2: Comparación de Desempeño con Ridge y Lasso.

Estas gráficas respaldan los análisis del modelo en términos de sesgo, varianza y ajuste, mostrando cómo la regularización ayuda a mejorar el desempeño general del modelo.

Conclusión

En resumen, el modelo de regresión lineal implementado logró predecir con precisión la probabilidad de admisión a programas de posgrado, y mediante la aplicación de regularización, se logró reducir aún más la varianza sin sacrificar el ajuste. El análisis detallado de sesgo y varianza demostró que el modelo tiene un buen desempeño y puede mejorarse con técnicas avanzadas como la regularización Ridge y Lasso. Este análisis proporciona una base sólida para futuras mejoras en la predicción de admisiones, utilizando datos adicionales o explorando otros algoritmos de aprendizaje automático más avanzados.