Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Inteligencia artificial avanzada para la ciencia de datos I

Módulo 2

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Grupo 101

Gilberto André García Gaytán - A01753176

Profesor:

Jorge Adolfo Ramírez Uresti

Reporte: Análisis del Modelo de Random Forest en el Dataset de Jugadores de Fútbol

Repositorio de Github

Justificación de la selección del dataset

El dataset elegido contiene información detallada sobre jugadores de fútbol de las cinco principales ligas. Las características incluyen datos demográficos, detalles del contrato, valor del jugador y más. Esta riqueza de características permite que el algoritmo de ML tenga suficientes datos para aprender y hacer predicciones precisas. Además, dado que los datos provienen de las cinco principales ligas, hay una buena diversidad y generalidad en los datos, lo que ayuda a demostrar que el modelo no solo se ajusta a una liga o conjunto de jugadores en particular, sino que puede generalizar a través de diferentes ligas y jugadores.

Separación y evaluación del modelo con conjuntos de prueba y validación (Train/Test/Validation)

El dataset fue dividido en tres conjuntos: entrenamiento, validación y prueba. Las divisiones son:

- Conjunto de entrenamiento: 1566 muestras
- Conjunto de validación: 523 muestras
- Conjunto de prueba: 523 muestras

Este método de división permite entrenar el modelo en una porción sustancial de los datos, validar y ajustar el modelo en un conjunto separado, y finalmente probar el rendimiento del modelo en un conjunto de prueba que no ha visto antes.

Visualización de la separación:

```
22  # Split data into training, validation, and test sets
23  X = data.drop(columns=['price'])
24  y = data['price']
25  X_temp, X_test, y_temp, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
26  X_train, X_val, y_train, y_val = train_test_split(X_temp, y_temp, test_size=0.25, random_state=42)
27
```

Diagnóstico y explicación

Grado de Bias o Sesgo: Bajo.
El error en el conjunto de entrenamiento es relativamente bajo, lo que indica que el modelo se ajusta bien a los datos de entrenamiento.

Grado de Varianza: Medio.
Hay una diferencia notable entre el error en el conjunto de entrenamiento y el conjunto de validación, pero no es excesivamente grande. Esto indica una varianza moderada.

Nivel de ajuste del modelo: Ligeramente Overfitting.
Dado que el sesgo es bajo y la varianza es media, el modelo se ajusta adecuadamente a los datos, pero hay una ligera tendencia al overfitting. Sin embargo, no es severo.

Uso de técnicas de regularización o ajuste de parámetros

Para abordar el ligero overfitting observado, se realizaron ajustes manuales en los parámetros del modelo de Random Forest. Los parámetros ajustados incluyen:

- n_estimators: 150
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2

Al utilizar estos parámetros, se buscaba reducir la complejidad del modelo y, por lo tanto, prevenir el overfitting. Sin embargo, las métricas mostraron que no hubo una mejora significativa en el rendimiento del modelo después de estos ajustes.

Conclusión

A lo largo de este análisis, se exploró y procesó un dataset detallado de jugadores de fútbol de las cinco principales ligas. Utilizando un modelo de Random forest, se pudo identificar un sesgo bajo y una varianza moderada en las predicciones. Aunque el modelo original mostró un ligero overfitting, los esfuerzos de ajuste manual de parámetros no dieron lugar a mejoras significativas en su rendimiento general.

Este resultado pone de manifiesto la importancia de la experimentación en el aprendizaje automático. Aunque el Random forest es un modelo poderoso, no siempre es el mejor para cada conjunto de datos o problemas. Además, el ajuste manual de parámetros, aunque útil, no garantiza siempre mejoras en el rendimiento. La selección de características, otros modelos, o incluso la ingeniería de características podrían proporcionar mejoras significativas en futuros análisis.

Finalmente, es crucial recordar que el aprendizaje automático es un proceso iterativo. A medida que se dispone de más datos o se adquiere una comprensión más profunda del problema, es esencial volver a visitar y ajustar el modelo para garantizar que sigue siendo relevante y preciso en sus predicciones.