Modelo Predictivo de Rendimiento y Competitividad - PENSIONISSSTE

Javier Horacio Pérez Ricárdez

February 11, 2025

Introducción

El objetivo de este modelo es predecir la evolución de los afiliados al sistema PENSIONISSSTE en función de dos factores clave: el rendimiento de las inversiones y la comisión aplicada por la institución. Para lograr esto, se utiliza un modelo de **Gradient Boosting**, que es un método de aprendizaje supervisado basado en la combinación de modelos débiles (como los árboles de decisión) para hacer una predicción robusta y precisa.

En el análisis, se utilizan datos históricos de los rendimientos de PENSIONISSSTE y de otras dos AFOREs (Administradoras de Fondos para el Retiro) con el fin de comparar el rendimiento y sus efectos sobre el número de afiliados. Estos datos de rendimientos y comisiones, así como los de afiliados, se combinan con datos sintéticos generados para modelar el futuro comportamiento de las variables.

Las variables utilizadas para predecir el número de afiliados son:

- **Rendimiento de PENSIONISSSTE**: La tasa de rendimiento de las inversiones administradas por el fondo PENSIONISSSTE.
- **Comisión de PENSIONISSSTE**: El porcentaje que PENSIONISSSTE cobra por la gestión de los fondos de retiro.
- **Rendimiento de AFORE1 y AFORE2**: Tasas de rendimiento de otras AFOREs que se consideran para el análisis comparativo.

El modelo tiene como objetivo predecir el número de afiliados a PENSIONISSSTE en función de las variaciones de estas variables.

Descripción del Modelo

El algoritmo de **Gradient Boosting** construye un modelo fuerte mediante la combinación de múltiples modelos débiles (generalmente árboles de decisión) en un proceso secuencial. En cada paso, se ajusta un nuevo árbol para corregir los errores cometidos por el modelo anterior.

Formalmente, el proceso de *boosting* se puede describir de la siguiente manera:

Función Objetivo

El objetivo es minimizar la función de pérdida L entre las predicciones $\hat{y}_i$ y los valores reales y_i:

$$L(\hat{y}_i, y_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Donde $\hat{y}_i$ es la predicción del modelo para la i-ésima observación, y y_i es el valor real.

Gradient Boosting

El algoritmo de Gradient Boosting construye los árboles de decisión de manera secuencial. En cada iteración t, se agrega un nuevo árbol $h_t(x)$ a la predicción global $F_{t-1}(x)$. La actualización de la predicción global se realiza de la siguiente manera:

$$F_t(x) = F_{t-1}(x) + \eta h_t(x)$$

Donde:

- $F_t(x)$ es la predicción en la iteración t.
- $F_{t-1}(x)$ es la predicción del modelo en la iteración anterior.
- $h_t(x)$ es el modelo débil (árbol de decisión) ajustado en la iteración t.
- η es el parámetro de tasa de aprendizaje, que controla cuánto influye cada nuevo árbol en la predicción final.

La idea de Gradient Boosting es ajustar cada árbol $h_t(x)$ para reducir el error cometido por el modelo anterior, utilizando el gradiente de la función de pérdida con respecto a las predicciones anteriores:

$$\text{Gradiente} = -\frac{\partial L(F_{t-1}(x), y)}{\partial F_{t-1}(x)}$$

Este gradiente se utiliza para ajustar el siguiente árbol $h_t(x)$, y así reducir el error en cada paso.

Modelo de Predicción

En el caso del modelo predictivo utilizado, el modelo de Gradient Boosting se entrena para predecir la variable dependiente $y = \text{Afiliados\_PENSIONISSSTE}$ utilizando las características
$X = \{\text{Rendimiento\_PENSIONISSSTE}, \text{Comision\_PENSIONISSSTE}, \text{Rendimiento\_AFORE1}, \text{Rendimiento\_AFORE2}\}$.

El proceso de entrenamiento busca ajustar los parámetros del modelo para minimizar la diferencia entre las predicciones y los valores reales de afiliados. La función objetivo en este caso es el error cuadrático medio:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Métricas de Evaluación

Para evaluar el desempeño del modelo, se utilizan varias métricas:

- **MAE (Error Medio Absoluto)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **R^2 (Coeficiente de Determinación)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Donde $\bar{y}$ es la media de los valores reales y.

- **Validación Cruzada MAE**: La validación cruzada es utilizada para evaluar la generalización del modelo en diferentes subconjuntos de datos. En este caso, el MAE promedio se calcula utilizando 5 particiones de los datos.

Generación de Datos Sintéticos

Dado que no siempre se dispone de datos reales completos o históricos de las variables, se generaron datos sintéticos para modelar el comportamiento futuro de las variables. En este caso, se generaron valores de **Rendimiento de PENSIONISSSTE** y **Comisión de PENSIONISSSTE** de manera aleatoria dentro de rangos predefinidos, lo que permitió simular escenarios futuros y entrenar el modelo de predicción.

Conclusión

El modelo de Gradient Boosting es adecuado para predecir la evolución de los afiliados de PENSIONISSSTE a partir de las características de rendimiento y comisión, ya que es capaz de capturar relaciones no lineales y manejar interacciones complejas entre las variables. Las métricas de evaluación, como el MAE y el R^2, permiten medir la precisión y la capacidad de generalización del modelo. El uso de datos sintéticos, en este caso, ofrece una alternativa útil para probar el modelo cuando los datos reales no están disponibles.