Bootcamp Data Science - Ejercicio semana 9 – Gerardo Rodríguez

Supón que estás trabajando en un proyecto de detección de fraude en transacciones bancarias y decides utilizar técnicas de ensamble para mejorar la precisión del modelo. Documenta en un archivo las siguientes respuestas:

1. ¿Qué ventajas podría tener el uso de un ensamble de modelos sobre un único árbol de decisión en este contexto?

Investigando identificó que es posible ejecutar el ensamble con un único modelo, pero pierde totalmente el propósito del mismo ensamble.

El objetivo es tener diferentes modelos y que entre ambos se potencien por medio de votos y/o promedios.

2. Si eliges utilizar la técnica de bagging y decides implementar un RandomForest, ¿por qué esta técnica podría ser útil en la detección de fraudes?

Son varias causas, los árboles de decisión corren el riesgo de sobreestimar, por lo que es conveniente usar esta técnica mediante bagging que usa muestreo por remplazo lo que esto reduce el riesgo de que el modelo este desbalanceado.

Situaciones estacionales pudiesen contenerse al no aglomerar toda la data en eventos como el fraude que son peculiares con respecto a la totalidad de los datos. Esto ya que Bagging reutiliza la data aleatoriamente en los diferentes árboles.

Los resultados de los diferentes árboles de decisión al final permitirán clasificar de diferentes formas los datos, por lo que al obtener resultados variantes, se realizará un sistema de votación que permitirá clasificar los valores basados en la cantidad de datos más preponderante (Voto duro) o mediante la probabilidad de aparición (voto suave). Esto permitirá clasificar utilizando lo mejor de los modelos utilizados.

3. Imagina que, después de implementar el Random Forest, decides añadir un modelo basado en boosting, como XGBoost. ¿Qué características específicas del boosting podrían mejorar aún más la precisión del modelo?

En definitiva sumar Boosting a nuestro Random Forest, empodera más el modelo completo por las siguientes causas:

La forma diferente de analizar los datos mediante análisis secuenciales y la ponderación de valores fallidos, enriquece la forma en que el modelo llega al dato más exacto al momento de clasificar y mitiga el riesgo de los sesgos de Random Forest.

Boosting también identifica los escenarios en que es difícil de identificar un posible fraude y en base a la historia mejora el modelo de manera continua si se ejecucta regularmente.

Investigando XGBoost, este en particular puede ponderar de manera especial aquellos escenarios en que existe un mayor riesgo de asumir costos, por ejemplo, enfocarnos en los fraudes más peligrosos

para el cliente y para el banco. XGBoost a diferencia de otras herramientas puede optimizar los recursos computacionales.

En resumen, ambos modelos, el bagging de Random Forest y Boosting en conjunto hacen un modelo mucho más poderoso y atinado para identificar fraudes.