Propuesta modelos para tesis

Árboles de Decisión y Métodos de Ensamble (Random Forest, XGBoost)

Los Árboles de Decisión son modelos intuitivos que clasifican las instancias mediante una serie de preguntas jerárquicas sobre sus características, dividiendo los datos recursivamente hasta llegar a una decisión (hoja del árbol). Aunque un solo árbol de decisión puede ser propenso al sobreajuste, su verdadero poder en la detección de fraude se manifiesta a través de los métodos de ensamble, que combinan múltiples árboles para mejorar la robustez y la precisión.

Random Forest (RF):
Este algoritmo construye una multitud de árboles de decisión durante el entrenamiento. Cada árbol se entrena con una muestra aleatoria (con reemplazo) del conjunto de datos original y considera sólo un subconjunto aleatorio de características en cada división. La predicción final se realiza por mayoría de votos (para clasificación) o promediando las salidas de todos los árboles (para regresión). Esta doble aleatorización ayuda a reducir significativamente el sobreajuste y mejora la generalización del modelo.

XGBoost (Extreme Gradient Boosting):
Es un algoritmo de boosting (potenciación) basado en árboles que ha ganado gran popularidad por su eficiencia y alto rendimiento. XGBoost construye árboles de forma secuencial, donde cada nuevo árbol se entrena para corregir los errores residuales cometidos por los árboles anteriores. Incorpora técnicas avanzadas como la regularización (para prevenir el sobreajuste), el manejo eficiente de valores faltantes y la paralelización, lo que lo convierte en una herramienta muy potente. Se le considera a menudo como el "estado del arte en la evolución de los algoritmos basados en árboles".

En el contexto de las plataformas de delivery, estos modelos son altamente aplicables para: Identificar reglas de decisión complejas y no lineales que son indicativas de fraude. Por ejemplo, un modelo podría aprender una regla como: "Si la cuenta del cliente es muy reciente y el valor del pedido es inusualmente alto y la dirección de entrega no coincide con patrones históricos y el método de pago ha sido asociado previamente con actividad fraudulenta entonces la probabilidad de fraude es alta".

Manejar eficazmente la naturaleza mixta de los datos comunes en transacciones de delivery, que a menudo incluyen tanto variables categóricas (p. ej., tipo de restaurante, categoría de producto) como numéricas (p. ej., monto del pedido, tiempo de entrega).

Proporcionar una medida de la importancia de las características, lo que ayuda a los analistas a comprender qué factores son los más influyentes para predecir el fraude. Esta información es invaluable para refinar estrategias de prevención y para comunicar los hallazgos a las partes interesadas.

Los datos típicos para estos modelos son similares a los de la Regresión Logística, pero los ensambles de árboles pueden manejar de forma nativa un mayor volumen y complejidad de características sin necesidad de transformaciones extensas. Estudios los han aplicado a la detección de fraude en solicitudes de tarjetas de crédito utilizando predictores como código postal, género, edad, ingresos, historial crediticio, etc..

Las ventajas son numerosas:
Random Forest:
Es robusto frente al ruido y al sobreajuste, maneja bien los datos faltantes de forma inherente y proporciona estimaciones de la importancia de las características. Algunas investigaciones han encontrado que RF es el algoritmo más efectivo en sus comparativas.
XGBoost: Ofrece una precisión predictiva muy alta, es rápido en ejecución (especialmente con grandes conjuntos de datos), maneja valores perdidos de manera sofisticada e incluye regularización para controlar el sobreajuste. Varios estudios concluyen que XGBoost supera a Random Forest en términos de precisión.

Entre las limitaciones, los árboles de decisión individuales son muy propensos al sobreajuste si no se podan o controlan. Aunque los ensambles como RF y XGBoost mitigan esto, pueden volverse más como "cajas negras" en comparación con un solo árbol, aunque la información sobre la importancia de las características ayuda a paliar esta opacidad.

La capacidad de los ensambles de árboles para manejar datos heterogéneos y relaciones no lineales de forma nativa, junto con su robustez inherente y la valiosa información que proporcionan sobre la importancia de las características, los convierte en herramientas extremadamente versátiles y efectivas para una amplia gama de escenarios de fraude en plataformas de delivery. A menudo superan a modelos lineales más simples en rendimiento predictivo. Los datos de fraude en delivery son inherentemente complejos, con interacciones sutiles y no lineales entre múltiples variables (p. ej., la combinación de una dirección IP anónima, un pedido de alto valor realizado desde una cuenta nueva y enviado a una dirección de entrega inusual es mucho más sospechosa que cada uno de estos factores considerados aisladamente). Random Forest y XGBoost son particularmente buenos para capturar estas interacciones complejas sin requerir una ingeniería de características manual exhaustiva. La información sobre la importancia de las características que estos modelos generan no solo es útil para la interpretabilidad del modelo en sí, sino que también puede guiar el desarrollo de futuras estrategias de prevención y facilitar la comunicación de los hallazgos a los stakeholders del negocio. Su rendimiento consistentemente alto en diversos benchmarks de detección de fraude los establece como una opción inicial sólida, o incluso como el modelo principal, para muchas plataformas. Por lo tanto, para numerosas plataformas de delivery, la implementación de un modelo Random Forest o XGBoost bien ajustado y mantenido puede ofrecer un excelente equilibrio entre rendimiento predictivo, interpretabilidad relativa (a nivel de la importancia de las características) y facilidad de implementación.