Murabei CONNECT

Modelos Ensemble:
Combinando Modelos para Melhorar Performance

Lucas Ribeiro, Luiz Freire, Valesca Silva, Wanessa Mendonça

Introdução

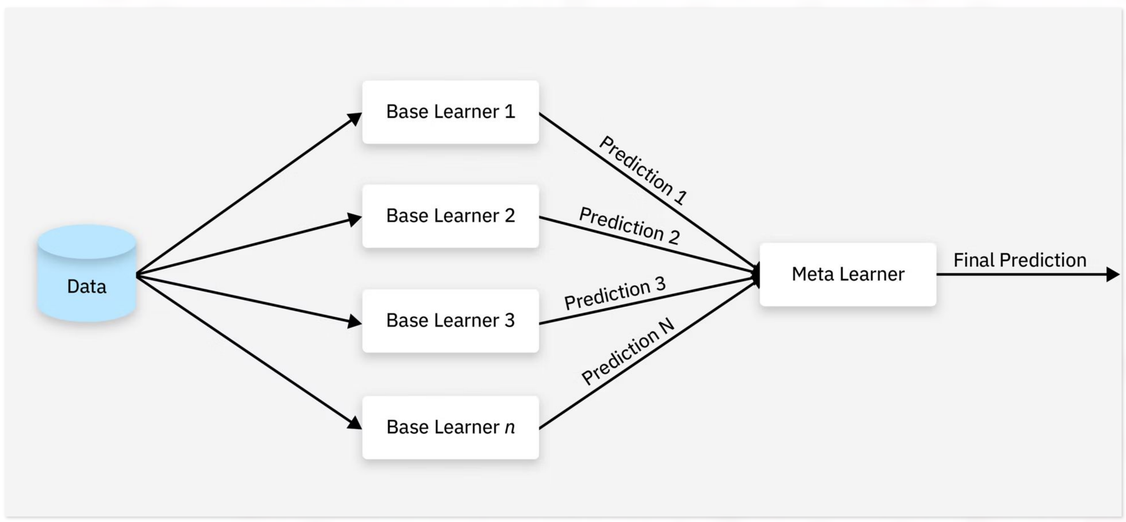
O modelo de *ensemble* é uma técnica que consiste em definir um conjunto de classificadores ou regressores que possuem particularidades e saídas específicas, as quais são agregadas utilizando técnicas como *bagging*, *stacking* e *boosting* em uma saída final. Os modelos de ensemble frequentemente superam modelos convencionais. Cada elemento do conjunto de modelos pode ser treinado com amostras diferentes, aumentando a diversidade dos modelos, o que traz uma séries de vantagens estatísticas e computacionais.

A principal vantagem estatística é a redução da variância das saídas. Ao treinar os modelos com subconjuntos e hiperparâmetros (ou mesmo técnicas) diferentes entre si, o ensemble se torna mais robusto aos vieses individuais de cada modelo a partir da agregação final do resultado, sendo especialmente vantajoso quando há variância muito grande na distribuição dos dados.

Dentre as vantagens computacionais, vale mencionar o paralelismo e a busca por um ótimo global, visto que os modelos podem ter pontos de partida diferentes e convergir em diferentes pontos, resultando em uma busca maior no espaço da solução da hipótese, mesmo que não haja garantia de máximo global.

Metodologia

Na técnica de Stacking, um modelo agregador (*meta-learner*) é treinado a partir das predições geradas por diversos modelos paralelos (*base learners*). Para evitar o *overfitting*, o *meta-learner* deve ser treinado exclusivamente com dados não vistos pelos modelos intermediários, utilizando técnicas como a validação cruzada. Neste tipo de técnica, é possível a combinação de diferentes tipos de algoritmos.



A técnica de Stacking organiza os modelos em camadas para otimizar a previsão final:

- **Arquitetura Hierárquica:** Diferente de outras técnicas, o Stacking não foca apenas no peso do erro, mas sim na inteligência coletiva. Ele utiliza uma camada de *base learners* (modelos heterogêneos) cujas saídas servem como entrada para um *meta-learner*, que aprende a decidir qual modelo é mais confiável para cada tipo de padrão.
- **Diversidade de Algoritmos:** A grande força desta abordagem reside na combinação de naturezas distintas. É possível unir a sensibilidade de uma Árvore de Decisão com a rigidez estatística de uma Regressão Logística e a flexibilidade de um SVM, permitindo que o modelo final capture nuances que um algoritmo isolado ignoraria.

O Stacking possui como vantagens o fato de ser uma técnica extremamente robusta para reduzir o viés e a variância simultaneamente, extraindo o melhor desempenho de cada arquitetura individual.

Além disso, possui um alto grau de sofisticação através da escolha do meta-modelo e do ajuste de hiperparâmetros, garantindo que a combinação final seja superior à simples média das previsões individuais.

Resultados

Base de Dados:

O conjunto de dados utilizados para analisar a técnica foi o *Credit Card Fraud Detection* que contém informações sobre as transações realizadas com cartões de crédito em setembro de 2013 por titulares de cartões europeus, disponibilizado na plataforma Kaggle.

- **Cenário:** Este conjunto de dados apresenta transações ocorridas em dois dias, com 492 fraudes em um total de 284.807 transações.
- **Natureza:** O conjunto de dados é altamente desbalanceado, com apenas 0,172% das transações sendo fraudes, e contém apenas variáveis de entrada numéricas, que são o resultado de uma transformação PCA.

Modelagem:

No modelo proposto foi utilizado um modelo Logit como *base learner* e um Gradient Boosting como *meta-learner*. A base foi separada em 70% para treino e 30% para teste. No treinamento foi aplicado um *grid search* para encontrar os melhores parâmetros.

O modelo alcançou um desempenho excepcional, com KS de 0,99 e AUC de 0,97 na base de teste. Na prática, a arquitetura funciona como um "filtro estatístico" de duas etapas:

- **O Filtro (*Base Learner* - Logit):** Processa as 28 variáveis latentes (V1 a V28) e o valor da transação (Amount), estimando uma probabilidade inicial a partir desses dados.
- **O Refino (*Meta-Learner* - Gradient Boosting):** Este modelo foca exclusivamente na probabilidade estimada na etapa anterior, calibrando o veredito final e reduzindo erros residuais.

Interpretação de Resultados:

Como o *Meta-Learner* refina a saída do Logit, entender os coeficientes da Regressão Logística nos dá a chave do comportamento do modelo:

- **Intercept (-8,36):** Representa o "viés conservador". Dado que fraudes representam apenas 0,172% dos dados, o modelo é calibrado para assumir que uma transação é legítima por padrão, exigindo evidências fortes para mudar de opinião.
- **Detectores de Fraude (Coefs Positivos):** As variáveis V4 (0,64) e V22 (0,53) são os principais gatilhos. Quanto maiores seus valores, maior a propensão do modelo em sinalizar uma suspeita.
- **Inibidores de Fraude (Coefs Negativos):** As variáveis V14 (-0,62) e V10 (-0,50) atuam como "atenuadores". Valores altos dessas variáveis ajudam a confirmar a legitimidade da transação.
- **Amount (0,0004):** O impacto do valor financeiro é residual. Isso indica que a detecção da fraude neste dataset está muito mais ligada ao comportamento (variáveis V) do que ao montante transacionado.

Nota sobre Expansão: Caso fossem utilizados diferentes modelos como Base Learners (ex: Random Forest e SVM juntos ao Logit), o Feature Importance do Gradient Boosting seria a ferramenta ideal para identificar qual desses modelos está entregando a predição mais confiável para o resultado final.

Conclusões

A aplicação da técnica de Stacking demonstrou ser altamente eficaz para o desafio de detecção de fraudes em cartões de crédito. Ao combinar a interpretabilidade da Regressão Logística com o poder de refino do Gradient Boosting, o modelo alcançou um equilíbrio raro entre precisão e explicabilidade.

Bibliografia

MUREL, Jacob; KAVLAKOGLU, Eda. What is ensemble learning? IBM, [S. l.], [202-?]. Disponível em: https://www.ibm.com/think/topics/ensemble-learning. Acesso em: 25 fev. 2026.

SCIKIT-LEARN. sklearn.ensemble.StackingClassifier. [S. l.], c2007-2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html. Acesso em: 2 mar. 2026.