Métodos de Ensemble - Introdução com Scikit-Learn

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Métodos de Ensemble

Introdução com Scikit-Learn

O objetivo deste módulo é apresentar os fundamentos dos métodos de ensembles de algoritmos de machine learning, sobre o ponto de vista conceitual e prático, usando a biblioteca scikit-learn para consolidar melhor os conceitos. Exploraremos bases de dados nos segmentos de EdTech e Saúde e iremos desenvolver projetos aplicando cada um dos métodos de ensemble apresentados.

Agenda

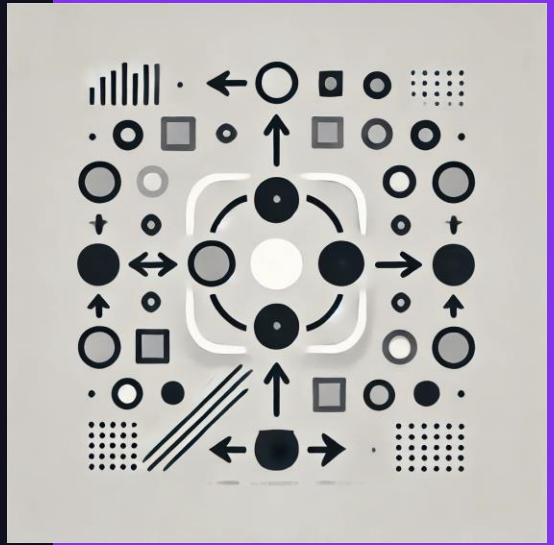
- O que são métodos de ensemble
- Casos de uso
- Desafios
- Visão geral dos principais métodos de ensemble
- Bagging – Bootstraping Aggregation
- Projetos Práticos - Bagging
- Boosting
- Projetos Práticos - Boosting
- Stacking
- Projetos Práticos – Stacking
- Voting
- Projetos Práticos - Voting
- O uso de métodos de ensemble em competições de Data Science

O que são métodos de ensemble

Métodos de ensemble são técnicas de aprendizado de máquina que combinam as previsões de múltiplos modelos para melhorar a performance geral do modelo final em comparação com a de qualquer modelo individual.

A premissa fundamental por trás dos métodos de ensemble é que, ao combinar diferentes modelos, é possível compensar as fraquezas de um modelo com as forças de outros, resultando em um desempenho mais robusto, preciso e generalizável.

Na essência, os métodos de ensemble buscam agrupar modelos de base (preditores fracos) gerando um modelo mais robusto (preditor forte).

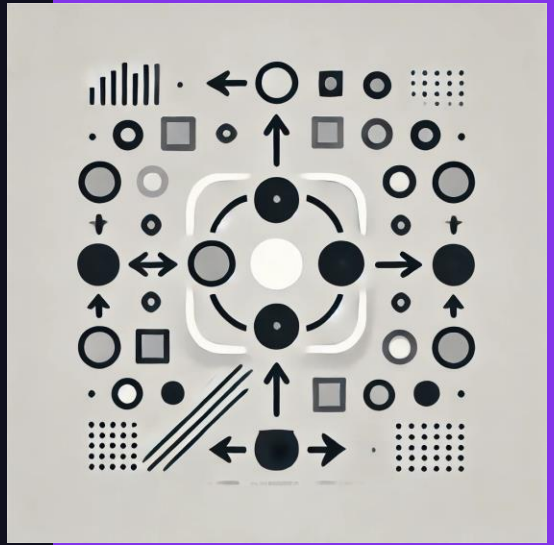


O que são métodos de ensemble

Além disso, os métodos de ensemble exploram a diversidade entre os modelos base para melhorar a precisão das previsões.

A diversidade entre os modelos é essencial porque, se todos os modelos base cometem os mesmos erros, o ensemble não terá um desempenho significativamente melhor do que os modelos individuais.

No entanto, se os modelos cometem erros diferentes, o ensemble pode reduzir a probabilidade de erros globais, resultando em uma melhor performance.



Casos de uso

Previsão Financeira e Decisões de Investimento

No setor financeiro, os métodos de ensemble são amplamente utilizados para melhorar a precisão das previsões e informar decisões de investimento.

Ao combinar as previsões de vários modelos de deep learning treinados em diferentes conjuntos de dados, os métodos de ensemble podem capturar padrões e tendências de mercado complexos de forma mais eficaz.

Isso resulta em previsões mais confiáveis para preços de ações, movimentos de mercado e avaliações de risco.

Casos de uso

Detecção de Anomalias e Fraudes

Os métodos de ensemble se destacam em tarefas de detecção de anomalias e fraudes, combinando as saídas de múltiplos modelos para diferenciar padrões normais de comportamentos incomuns ou suspeitos.

Essa abordagem ajuda a identificar potenciais fraudes, ataques cibernéticos ou falhas de sistema em domínios como bancos, segurança cibernética e processos industriais.

Casos de uso

Diagnóstico Médico e Descoberta de Medicamentos

Na área da saúde, os métodos de conjunto são utilizados para diagnóstico de doenças e descoberta de medicamentos.

Ao combinar previsões de múltiplos modelos de aprendizado de máquina, os métodos de conjunto podem melhorar a precisão da detecção de doenças e ajudar a identificar candidatos a medicamentos potenciais de forma mais eficaz.

Isso é particularmente importante em cenários médicos complexos onde modelos individuais podem ter limitações.

Casos de uso

Sistemas de Recomendação

Os métodos de ensemble são empregados em sistemas de recomendação para fornecer recomendações mais precisas e personalizadas aos usuários.

Ao combinar as saídas de múltiplos modelos de aprendizado profundo, os métodos de ensemble podem capturar diversas preferências dos usuários, lidar com problemas de cold start e melhorar a qualidade das recomendações em várias indústrias, como comércio eletrônico e entretenimento.

Casos de uso

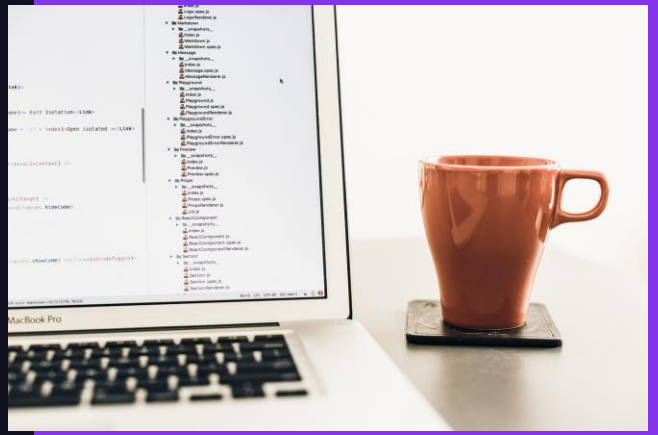
Tarefas de Processamento de Linguagem Natural (NLP)

No campo do NLP, os métodos de conjunto são usados para aprimorar o desempenho em tarefas como análise de sentimentos, classificação de texto e geração de linguagem.

Ao combinar as previsões de múltiplos modelos de aprendizado profundo treinados em diferentes arquiteturas ou conjuntos de dados de NLP, os métodos de conjunto podem alcançar maior precisão e robustez na compreensão e geração de texto semelhante ao humano.

Desafios

Seleção e Ponderação dos Modelos



Selecionar a combinação certa de modelos a serem incluídos no ensemble, determinar a ponderação ideal das previsões de cada modelo e gerenciar os recursos computacionais necessários para treinar e avaliar múltiplos modelos simultaneamente.

Além disso, o ensemble learning nem sempre melhora o desempenho se os modelos individuais forem muito semelhantes ou se os dados de treinamento tiverem um alto grau de ruído.

A diversidade dos modelos - em termos de algoritmos, processamento de recursos e perspectivas de dados - é vital para cobrir um espectro mais amplo de padrões de dados. A ponderação ideal da contribuição de cada modelo, geralmente com base em métricas de desempenho, é crucial para aproveitar seu poder preditivo coletivo.

Portanto, consideração e experimentação cuidadosas são necessárias para alcançar os resultados desejados com o ensemble learning.

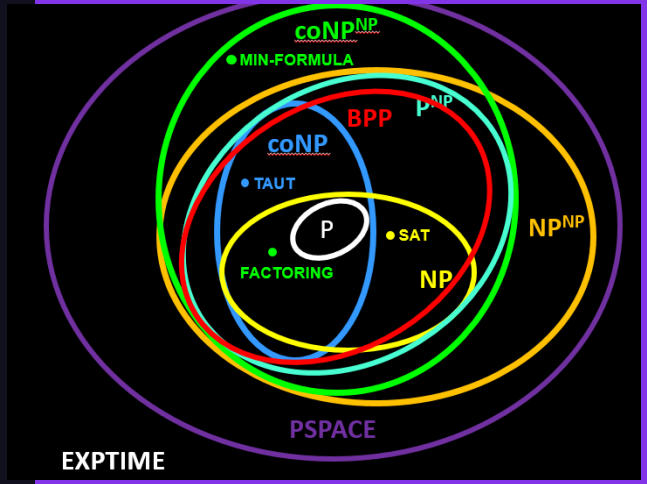
Desafios

Complexidade Computacional

O ensemble learning, envolvendo múltiplos algoritmos e conjuntos de recursos, requer mais recursos computacionais do que modelos individuais.

Embora o processamento paralelo ofereça uma solução, orquestrar um ensemble de modelos em vários processadores pode introduzir complexidade tanto na implementação quanto na manutenção.

Além disso, mais computação nem sempre leva a um melhor desempenho, especialmente se o ensemble não for configurado corretamente ou se os modelos amplificarem os erros uns dos outros em conjuntos de dados ruidosos.



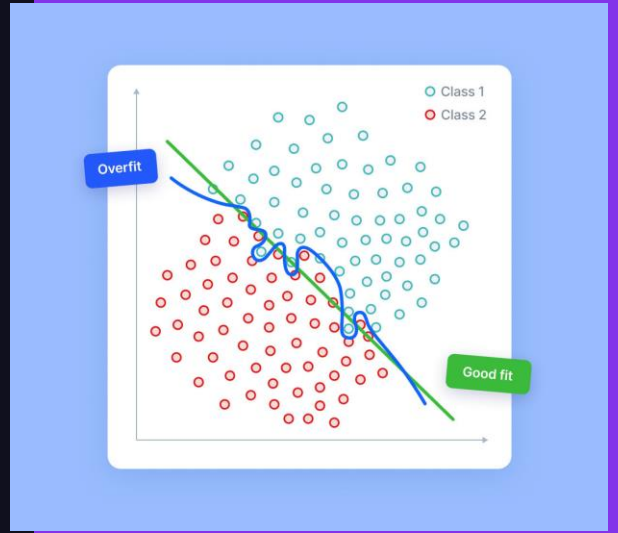
Desafios

Diversidade e Overfitting

O ensemble learning requer modelos diversos para evitar viés e melhorar a precisão. Ao incorporar diferentes algoritmos, conjuntos de recursos e dados de treinamento, o ensemble learning captura uma gama mais ampla de padrões.

No entanto, se os modelos individuais forem muito diferentes ou se os dados de treinamento tiverem ruído excessivo, o ensemble pode sofrer de overfitting, generalizando mal para novos dados.

Portanto, é crucial encontrar um equilíbrio entre a diversidade dos modelos e a qualidade dos dados de treinamento para obter os melhores resultados com o ensemble learning.



Visão geral dos principais métodos de ensemble

Bagging

Cria múltiplas versões de um modelo de aprendizado, cada uma treinada em um subconjunto aleatório dos dados de treino. As previsões desses modelos são então combinadas para produzir a previsão final.

Exemplo: Random Forest

Boosting

Constrói um conjunto de modelos sequencialmente, onde cada novo modelo tenta corrigir os erros cometidos pelos modelos anteriores.

Exemplos: Catboost, XGBoost, LightGBM, Adaboost

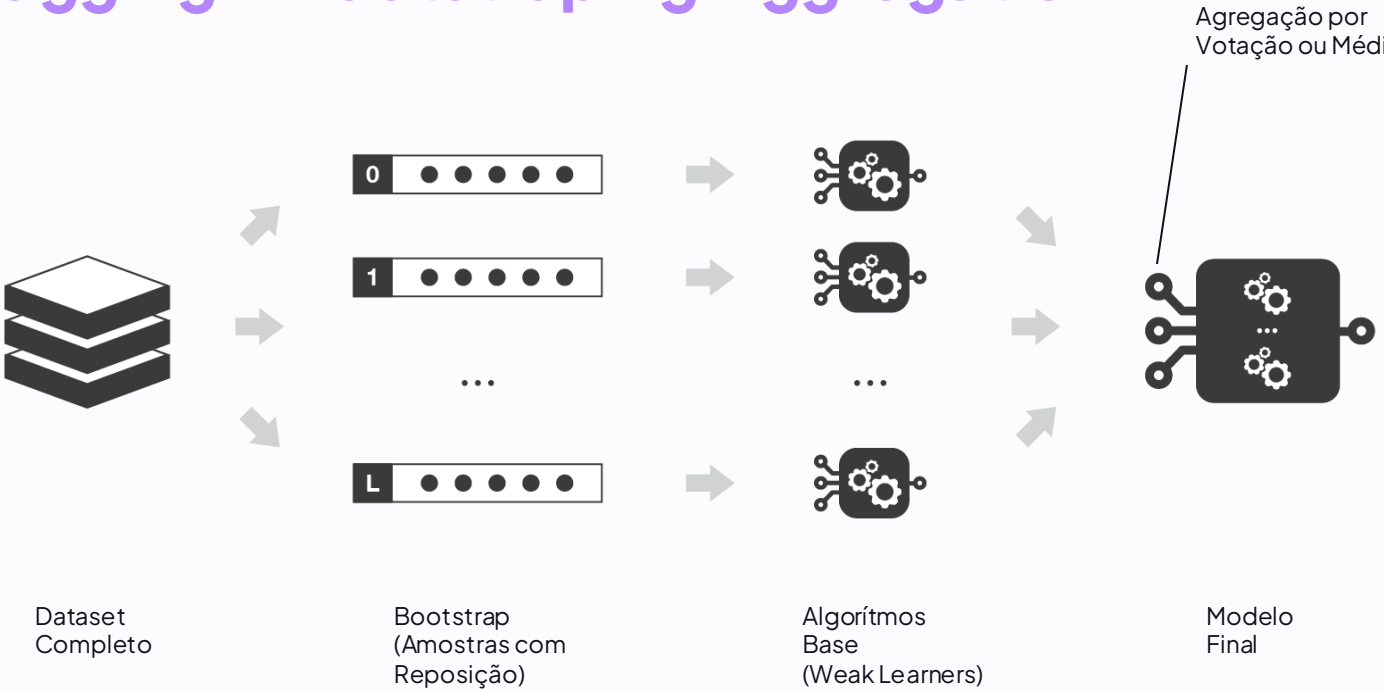
Stacking

Diferentes modelos são treinados e suas previsões são usadas como entradas para um modelo de nível superior (meta-modelo), que aprende a melhor maneira de combinar essas previsões para fazer a previsão final.

Voting

Diferentes modelos de aprendizado são treinados em todo o conjunto de dados, e suas previsões são combinadas por meio de votação (para classificação) ou média (para regressão)

Bagging – Bootstraping Aggregation



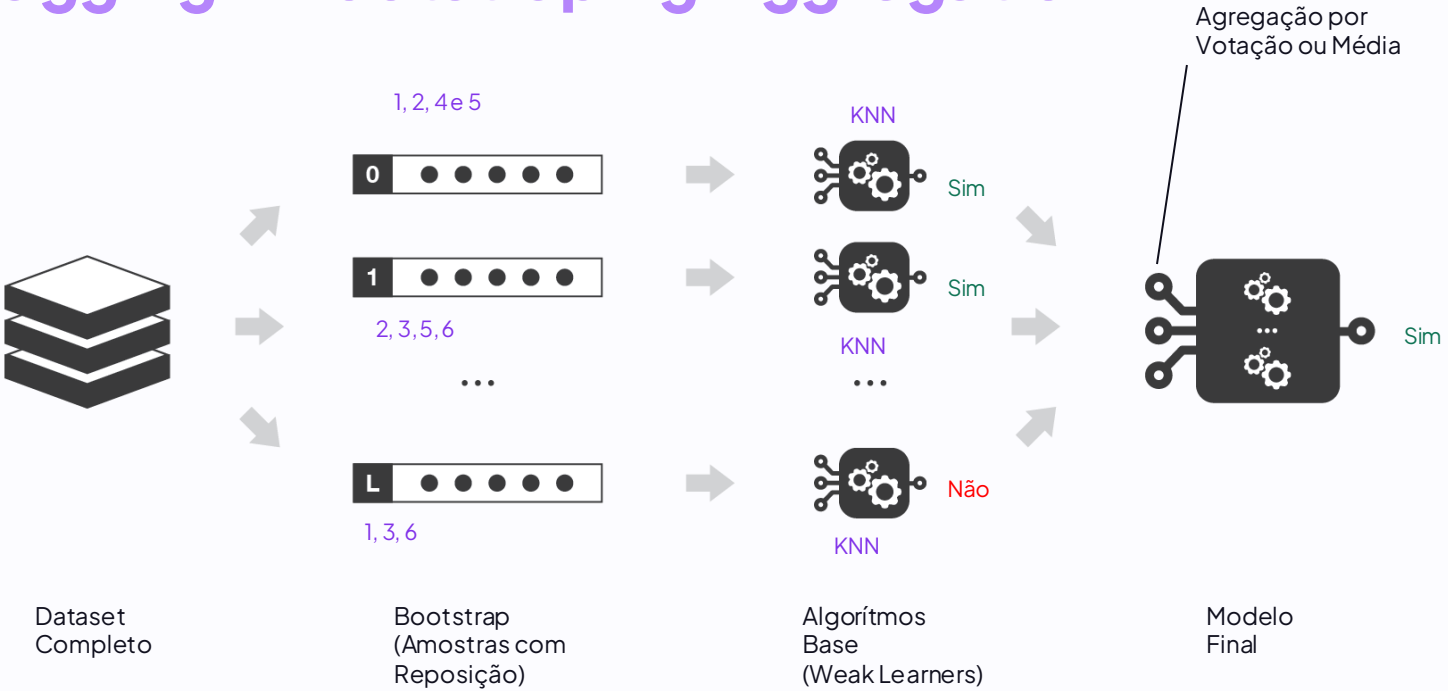
Dataset
Completo

Bootstrap
(Amostras com
Reposição)

Algorítmos
Base
(Weak Learners)

Modelo
Final

Bagging – Bootstraping Aggregation



Bagging

Vantagens no Uso

Redução da variância e do overfitting
Facilidade de paralelização
Aumento da robustez e precisão do modelo

Desvantagens no Uso

Maior custo computacional devido ao treinamento de múltiplos modelos
Pode não ser tão eficaz se os modelos base tiverem alta correlação

Bagging

Variações do Bagging

Pasting
Você extrai amostras do dataset, considerando um subconjunto de registros, sem reposição

Bagging
Você extrai amostras do dataset, considerando um subconjunto de registros, com reposição

Random Subspaces
Você extrai amostras do dataset, considerando um subconjunto de features

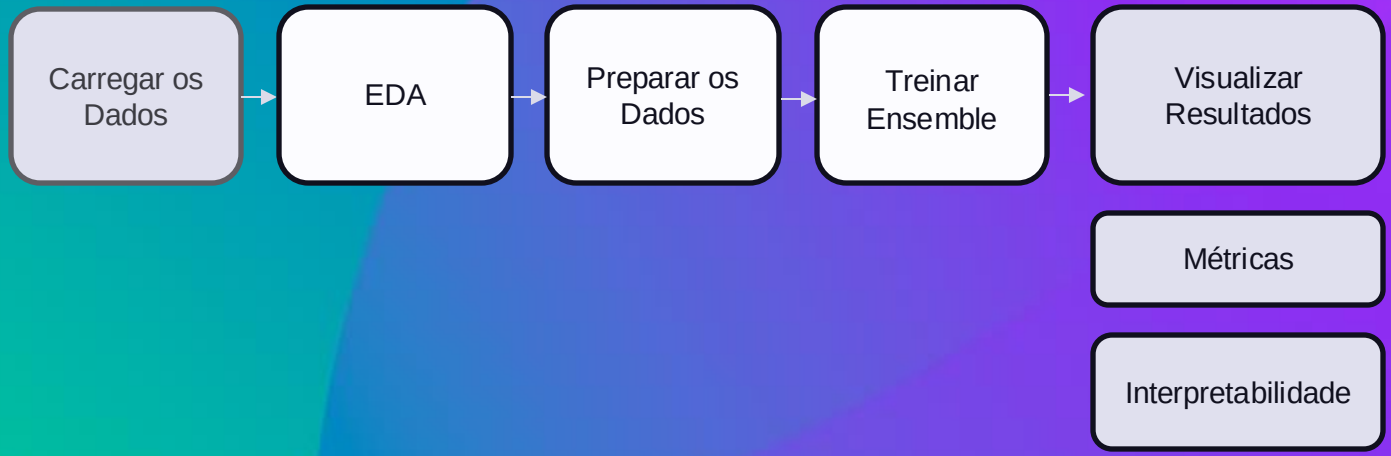
Random Patches
Você extrai amostras do dataset, considerando um subconjunto de registros e de features

Projeto – Bagging Classificação

Uma EdTech possui uma plataforma de vendas para oferta de seus produtos de educação e com o objetivo de priorizar melhor suas ações comerciais, quer desenvolver uma estratégia para ampliar seu fator de conversão de leads em vendas. Atualmente as informações referentes a este processo de vendas encontram-se em uma ferramenta de CRM, da qual podem ser extraídos alguns insights.

Desta forma, para apoiar no desenvolvimento desta estratégia, iremos trabalhar num algoritmo de classificação que possa prever se um lead será ou não convertido em venda. E dado o volume de dados e as features disponíveis, adotaremos o método Bagging de ensemble, usando algoritmos supervisionados de classificação.

Estrutura do Projeto – Bagging Classificação



Code Time . . .

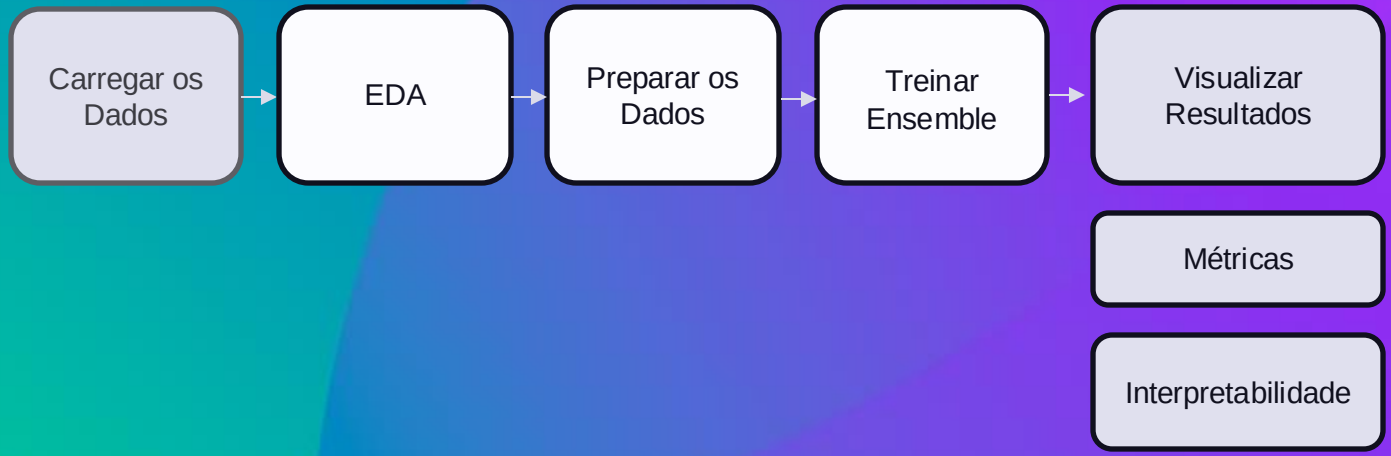
rocketseat.com.br

Projeto – Bagging Regressão

Uma HealthTech realizou uma pesquisa com mais de 1.300 pessoas sobre gastos realizados com saúde nos últimos 2 anos para criar uma oferta mais assertiva de planos de saúde para pequenas e médias empresas. Dentre os componentes para formatar esta oferta, é importante conseguir estimar os gastos de saúde dos funcionários de uma possível empresa cliente, com base em algumas características destes funcionários.

Desta forma, para apoiar no desenvolvimento desta oferta, iremos trabalhar num algoritmo de regressão que possa estimar os gastos de saúde de funcionários. Considerando o volume de dados e as features disponíveis, adotaremos o método Bagging de ensemble, usando algoritmos supervisionados de regressão.

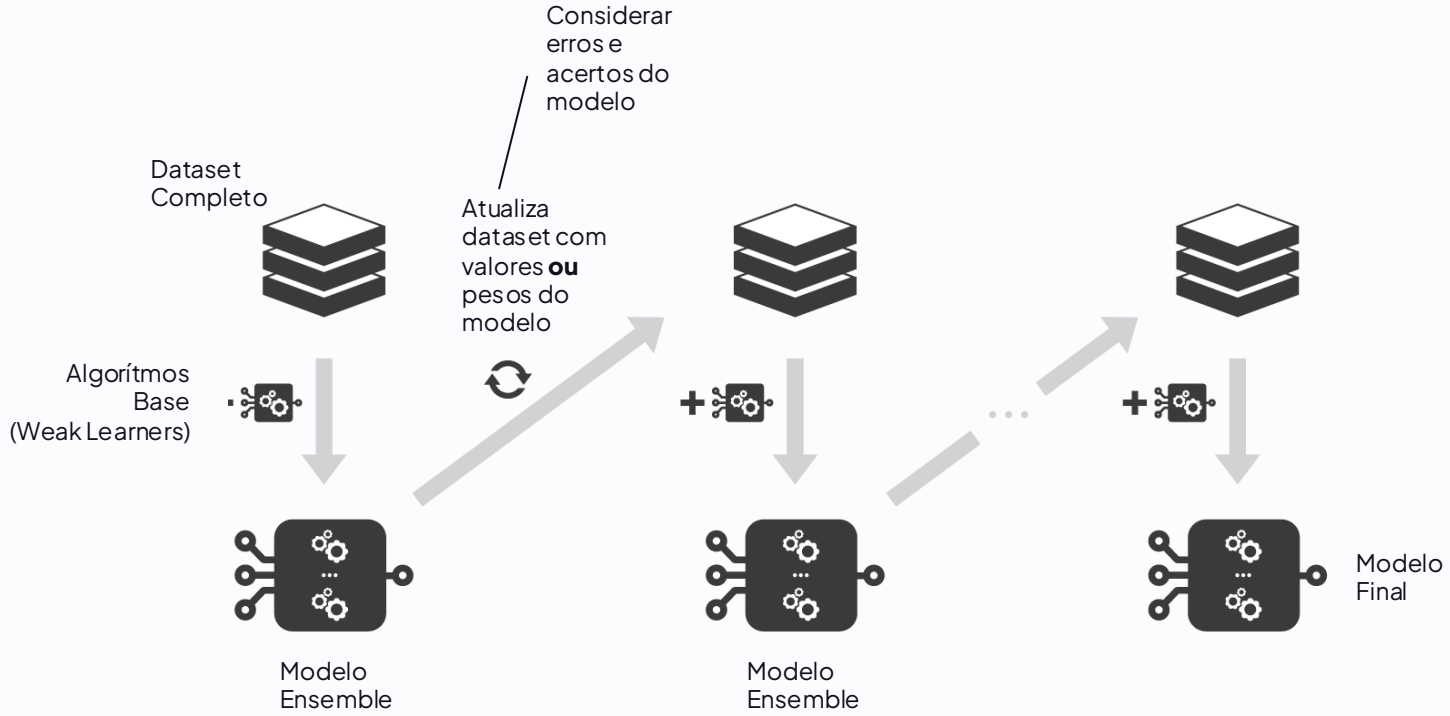
Estrutura do Projeto – Bagging Regressão



Code Time . . .

rocketseat.com.br

Boosting


Considerar erros e acertos do modelo
Dataset Completo
Atualiza dataset com valores **ou** pesos do modelo
Algorítmos Base (Weak Learners)
Modelo Ensemble
Modelo Ensemble
Modelo Final

Boosting

Vantagens no Uso

Alta precisão devido à redução do viés.
Eficaz para problemas complexos e conjuntos de dados ruidosos.
Flexibilidade para ajustar parâmetros e melhorar a performance.

Desvantagens no Uso

Maior suscetibilidade ao overfitting se não ajustado corretamente.
Custo computacional mais alto e tempo de treinamento mais longo.

Boosting

Variações do Boosting

Adaptative Boosting (AdaBoost)
No AdaBoost, a ênfase está em ajustar o peso das observações mal classificadas em cada iteração. Cada modelo subsequente é treinado para focar mais nas instâncias que os modelos anteriores falharam em prever corretamente. O modelo final é uma combinação ponderada de vários modelos fracos.

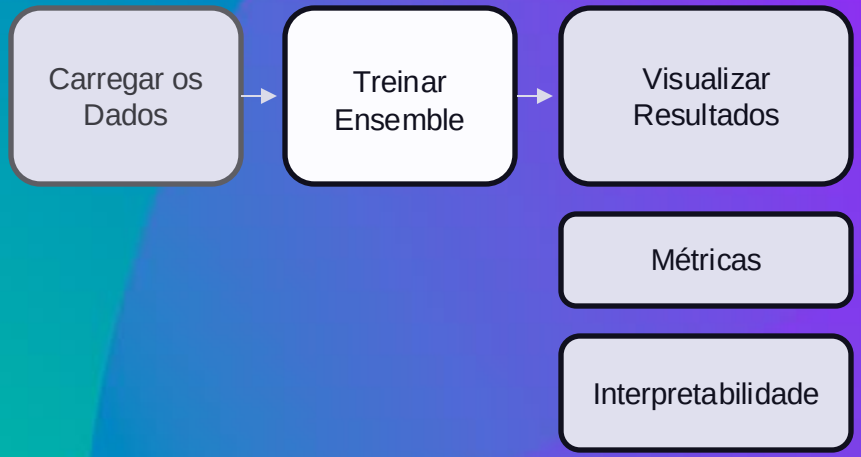
Gradient Boosting
O Gradient Boosting constrói modelos sequenciais onde cada novo modelo tenta corrigir os erros residuais dos modelos anteriores. Ele utiliza a otimização de gradiente para ajustar as previsões, visando minimizar o erro global.

Projeto – Boosting Classificação

Uma EdTech possui uma plataforma de vendas para oferta de seus produtos de educação e com o objetivo de priorizar melhor suas ações comerciais, quer desenvolver uma estratégia para ampliar seu fator de conversão de leads em vendas. Atualmente as informações referentes a este processo de vendas encontram-se em uma ferramenta de CRM, da qual podem ser extraídos alguns insights.

Desta forma, para apoiar no desenvolvimento desta estratégia, iremos trabalhar num algoritmo de classificação que possa prever se um lead será ou não convertido em venda. E dado o volume de dados e as features disponíveis, adotaremos o método Adaptative Boosting de ensemble, usando algoritmos supervisionados de classificação.

Estrutura do Projeto – Boosting Classificação



Code Time . . .

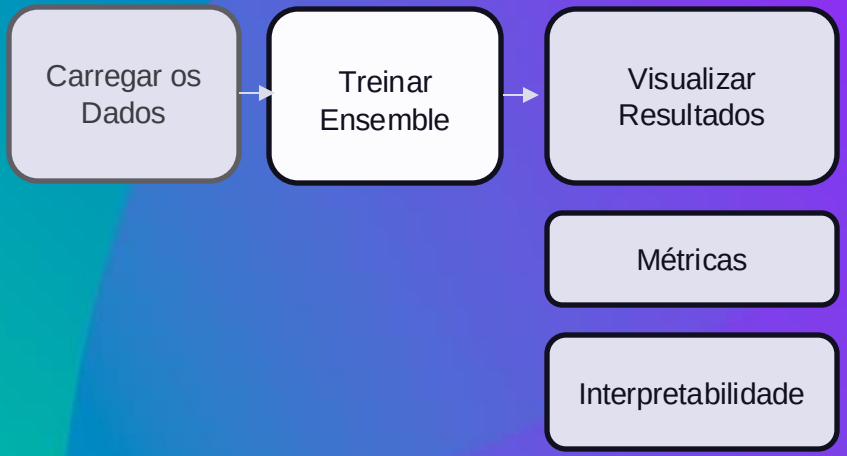
rocketseat.com.br

Projeto – Boosting Regressão

Uma HealthTech realizou uma pesquisa com mais de 1.300 pessoas sobre gastos realizados com saúde nos últimos 2 anos para criar uma oferta mais assertiva de planos de saúde para pequenas e médias empresas. Dentre os componentes para formatar esta oferta, é importante conseguir estimar os gastos de saúde dos funcionários de uma possível empresa cliente, com base em algumas características destes funcionários.

Desta forma, para apoiar no desenvolvimento desta oferta, iremos trabalhar num algoritmo de regressão que possa estimar os gastos de saúde de funcionários. Considerando o volume de dados e as features disponíveis, adotaremos o método Adaptative Boosting de ensemble, usando algoritmos supervisionados de regressão.

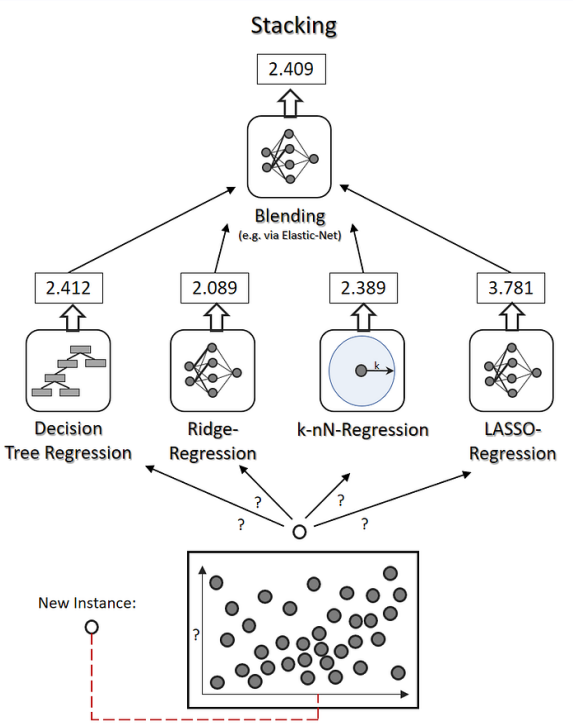
Estrutura do Projeto – Boosting Regressão



Code Time . . .

rocketseat.com.br

Stacking


Stacking
2.409
Blending
(e.g. via Elastic-Net)
2.412
2.089
2.389
3.781
Decision
Tree Regression
Ridge-
Regression
k-nN-Regression
LASSO-
Regression
?
?
?
?
New Instance:
?

Stacking

Vantagens no Uso

Potencial para melhorar significativamente a precisão.
Flexibilidade para combinar diferentes tipos de modelos.
Redução do viés e variância ao usar um meta-modelo.

Desvantagens no Uso

Maior complexidade e custo computacional.
Mais difícil de implementar e ajustar corretamente.
Maior risco de overfitting se não configurado adequadamente.

Stacking

Variações do Stacking

Vanilla
Usa apenas as predições ou classificações finais de cada modelo para treinamento do meta-modelo.

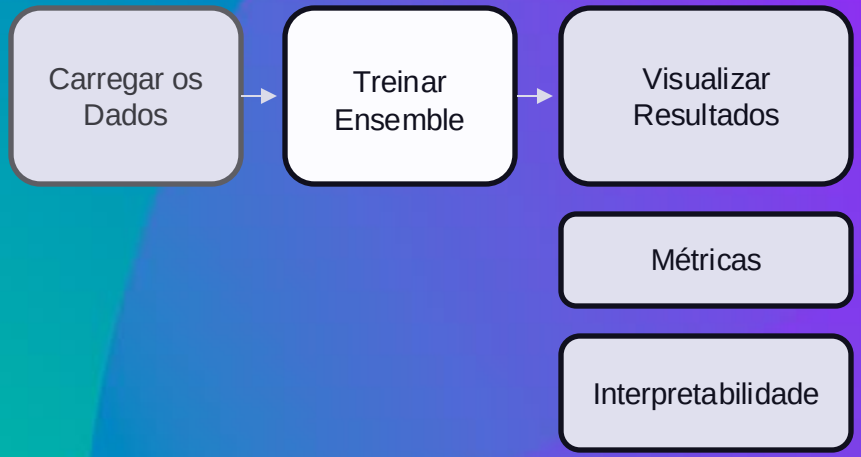
Blending
Além das predições ou classificações finais de cada modelo, usa os dados originais inseridos em cada modelo base, para treinamento do meta-modelo.

Projeto – Stacking Classificação

Uma EdTech possui uma plataforma de vendas para oferta de seus produtos de educação e com o objetivo de priorizar melhor suas ações comerciais, quer desenvolver uma estratégia para ampliar seu fator de conversão de leads em vendas. Atualmente as informações referentes a este processo de vendas encontram-se em uma ferramenta de CRM, da qual podem ser extraídos alguns insights.

Desta forma, para apoiar no desenvolvimento desta estratégia, iremos trabalhar num algoritmo de classificação que possa prever se um lead será ou não convertido em venda. E dado o volume de dados e as features disponíveis, adotaremos o método Stacking de ensemble, usando algoritmos supervisionados de classificação.

Estrutura do Projeto – Stacking Classificação



Code Time . . .

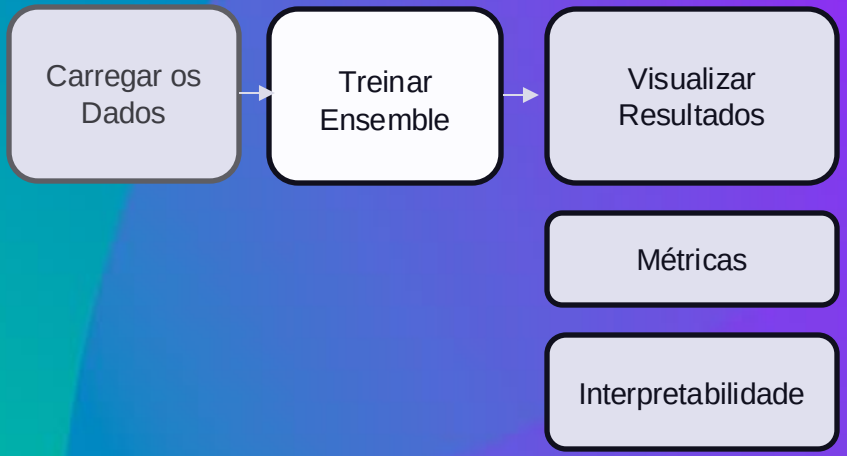
rocketseat.com.br

Projeto – Stacking Regressão

Uma HealthTech realizou uma pesquisa com mais de 1.300 pessoas sobre gastos realizados com saúde nos últimos 2 anos para criar uma oferta mais assertiva de planos de saúde para pequenas e médias empresas. Dentre os componentes para formatar esta oferta, é importante conseguir estimar os gastos de saúde dos funcionários de uma possível empresa cliente, com base em algumas características destes funcionários.

Desta forma, para apoiar no desenvolvimento desta oferta, iremos trabalhar num algoritmo de regressão que possa estimar os gastos de saúde de funcionários. Considerando o volume de dados e as features disponíveis, adotaremos o método Stacking de ensemble, usando algoritmos supervisionados de regressão.

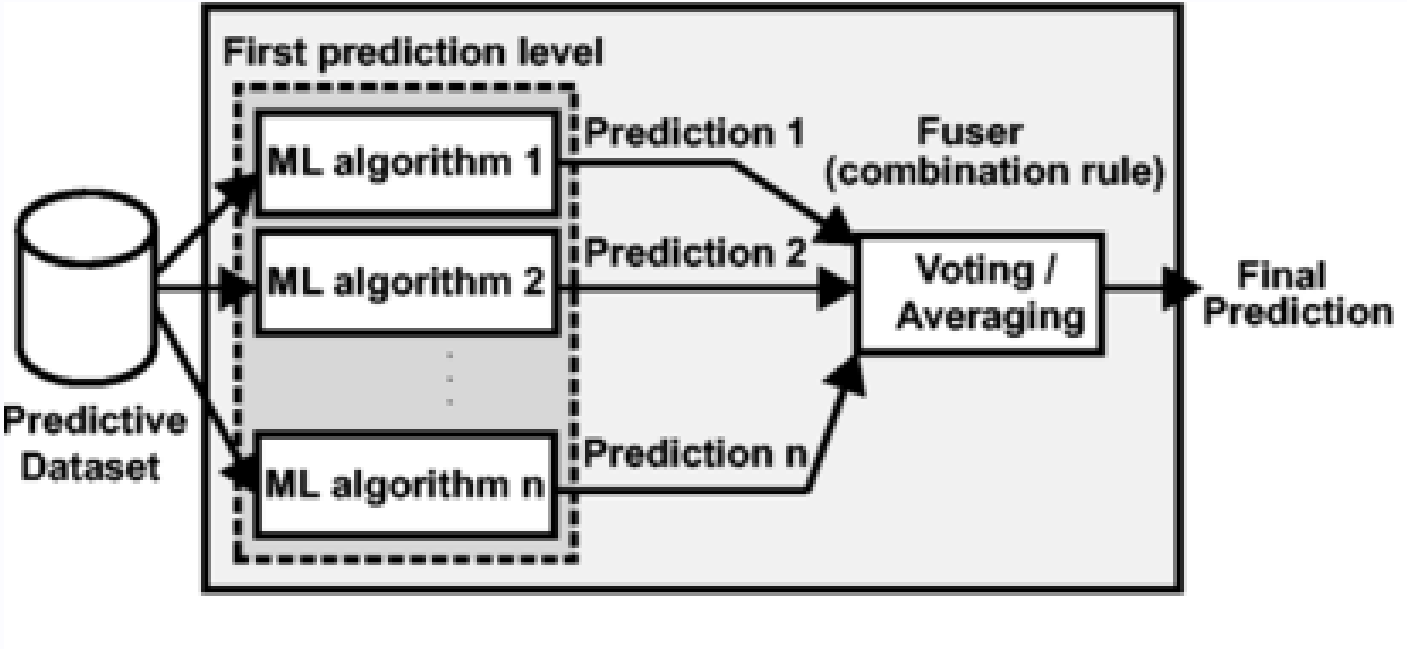
Estrutura do Projeto – Stacking Regressão



Code Time . . .

rocketseat.com.br

Voting


First prediction level
ML algorithm 1
ML algorithm 2
ML algorithm n
Prediction 1
Prediction 2
Prediction n
Fuser
(combination rule)
Voting /
Averaging
Final
Prediction
Predictive
Dataset

Voting

Vantagens no Uso

Simplicidade e facilidade de implementação.
Robustez ao combinar modelos diferentes.
Melhor desempenho geral ao utilizar modelos complementares.

Desvantagens no Uso

Pode não ser tão eficaz se os modelos individuais não forem robustos.
Dependência da qualidade e diversificação dos modelos base.

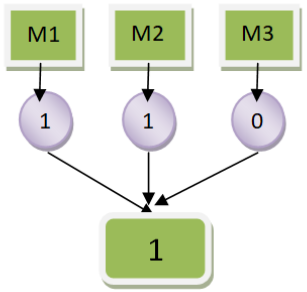
Voting

Variações do Voting (Para problemas de Classificação)

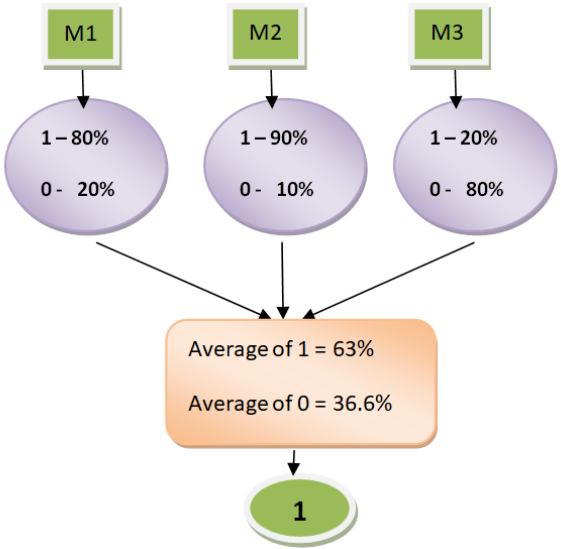
Hard Voting
Na hard voting, a previsão final é feita com base na votação majoritária das previsões dos modelos base. Cada modelo do ensemble faz sua própria previsão, e a classe ou rótulo que recebe a maioria dos votos é a previsão final.

Soft Voting
Na soft voting, a previsão final é feita com base na média ponderada das probabilidades de previsão de cada modelo base. Ou seja, cada modelo prevê uma probabilidade para cada classe, e essas probabilidades são somadas e ponderadas, resultando na previsão da classe com a maior probabilidade combinada.

Voting Classificação



Hard Voting



Soft Voting

Voting

Variações do Voting (Para problemas de Regressão)

Hard Voting
Na hard voting, a previsão final é feita com base na média aritmética das previsões.

Soft Voting
Na soft voting, a previsão final é feita com base na média ponderada dos valores e dos pesos de cada modelo.

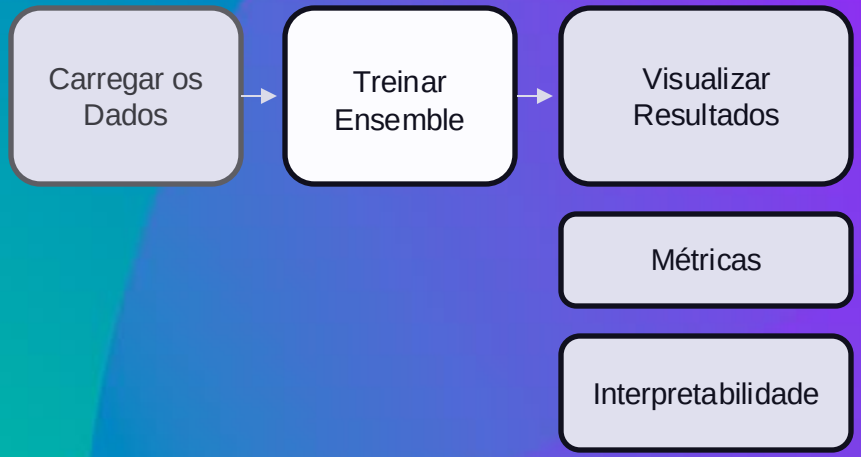
Obs:
No sklearn, não existe um hiperparâmetro pra definir o tipo de voting para regressão, portanto a implementação do soft voting precisa ser construída, pois o padrão é trabalhar com hard voting.

Projeto – Voting Classificação

Uma EdTech possui uma plataforma de vendas para oferta de seus produtos de educação e com o objetivo de priorizar melhor suas ações comerciais, quer desenvolver uma estratégia para ampliar seu fator de conversão de leads em vendas. Atualmente as informações referentes a este processo de vendas encontram-se em uma ferramenta de CRM, da qual podem ser extraídos alguns insights.

Desta forma, para apoiar no desenvolvimento desta estratégia, iremos trabalhar num algoritmo de classificação que possa prever se um lead será ou não convertido em venda. E dado o volume de dados e as features disponíveis, adotaremos o método Voting de ensemble, usando algoritmos supervisionados de classificação.

Estrutura do Projeto – Voting Classificação



Code Time . . .

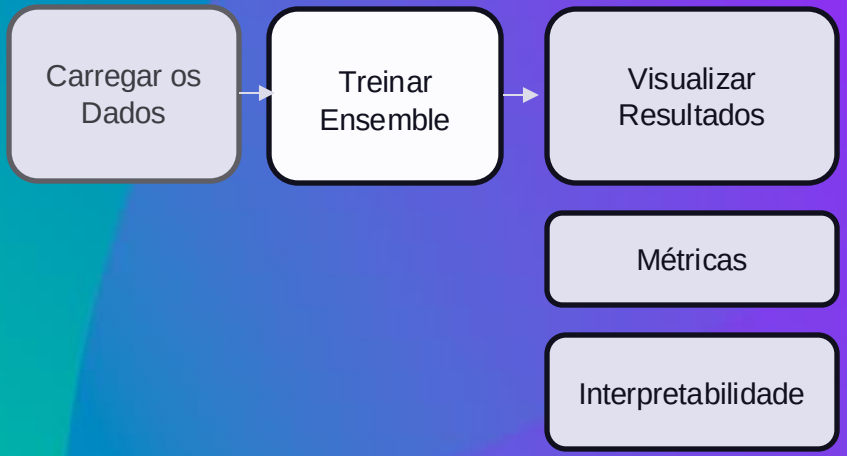
rocketseat.com.br

Projeto – Voting Regressão

Uma HealthTech realizou uma pesquisa com mais de 1.300 pessoas sobre gastos realizados com saúde nos últimos 2 anos para criar uma oferta mais assertiva de planos de saúde para pequenas e médias empresas. Dentre os componentes para formatar esta oferta, é importante conseguir estimar os gastos de saúde dos funcionários de uma possível empresa cliente, com base em algumas características destes funcionários.

Desta forma, para apoiar no desenvolvimento desta oferta, iremos trabalhar num algoritmo de regressão que possa estimar os gastos de saúde de funcionários. Considerando o volume de dados e as features disponíveis, adotaremos o método Voting de ensemble, usando algoritmos supervisionados de regressão.

Estrutura do Projeto – Voting Regressão



Code Time . . .

rocketseat.com.br

Comparativo entre os métodos

Categoria	Melhor para	Quando evitar
Bagging	Reduzir variância, modelos instáveis (como árvore de decisão)	Quando o modelo base é estável
Boosting	Reduzir viés e variância, problemas complexos	Se os dados forem muito ruidosos (risco de overfitting)
Stacking	Combinar diferentes tipos de modelos, aproveitando suas forças e superando suas fraquezas.	Se o ajuste for muito complexo ou o risco de overfitting for alto
Voting	Combinar diferentes modelos sem entrar em detalhes de camadas complexas, mas ainda assim obter uma melhoria de performance.	Quando os modelos são muito semelhantes ou o problema for muito complexo

Soluções vencedoras

Mercari Price Suggestion Challenge

Desafio:
O objetivo desta competição era prever o preço de produtos listados no site Mercari, com base em descrições, categorias e outras características dos itens. Os participantes tinham que lidar com um grande conjunto de dados desbalanceado e variáveis textuais.

Solução:
A equipe vencedora utilizou uma abordagem de ensemble que combinava modelos de Gradient Boosting (como XGBoost) e Random Forest. Eles também implementaram técnicas de stacking, onde as previsões de vários modelos eram usadas como entradas para um modelo final. Além disso, a equipe fez uso extensivo de processamento de linguagem natural (NLP) para extrair características significativas das descrições dos produtos, melhorando assim a performance do modelo.

Soluções vencedoras

Data Science Bowl 2019

Desafio:
Nesta competição, os participantes foram desafiados a criar um modelo que pudesse identificar a presença de células cancerígenas em imagens de microscopia. O desafio envolvia a análise de imagens complexas e a extração de características relevantes.

Solução:
Os vencedores usaram uma combinação de técnicas de ensemble, incluindo Convolutional Neural Networks (CNNs) e modelos clássicos como Random Forest. Eles aplicaram bagging e blending, onde múltiplos modelos CNN foram treinados em diferentes subconjuntos dos dados e suas previsões foram combinadas para melhorar a acurácia. A engenharia de features e o aumento de dados também foram cruciais para o sucesso da solução.

Soluções vencedoras

Google Cloud & NCAA ML Competition 2020

Desafio:
Esta competição focava na previsão dos resultados dos jogos do torneio da NCAA, utilizando dados históricos sobre partidas anteriores, estatísticas dos times e outros fatores relevantes.

Solução:
A equipe vencedora implementou um ensemble que combinava Logistic Regression, Random Forest, e XGBoost. Eles utilizaram técnicas de stacking para combinar as previsões desses modelos, além de realizar uma análise cuidadosa das features, incluindo variáveis derivadas das estatísticas dos jogos. A modelagem cuidadosa e a validação cruzada ajudaram a garantir que o modelo fosse robusto e capaz de generalizar bem para novos dados.