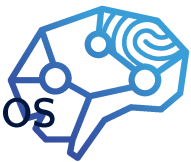
Regressão em Ciência de Dados



Universidade Estadual Paulista, Júlio de Mesquita Filho - UNESP

clayton.pereira@unesp.br

O Poder de Prever: Introdução à Análise Preditiva

Referências

Referências e Sites para consulta

- Scikit-learn
- Harvard CS109

O que é Análise Preditiva?

O que vamos estudar...

- Entender o conceito de análise preditiva
- Introduzir regressão como ferramenta de previsão
- Discutir aplicações reais da análise preditiva
- Refletir sobre limites éticos do uso de dados

O que é Análise Preditiva?

O que vamos estudar...

- Entender o conceito de análise preditiva
- Introduzir regressão como ferramenta de previsão
- Discutir aplicações reais da análise preditiva
- Refletir sobre limites éticos do uso de dados

Segundo Eric Siegel, *"A análise preditiva é o uso de modelos matemáticos para prever o comportamento futuro com base em dados passados."*

O que é Análise Preditiva?

Definição

Análise preditiva é o uso de modelos matemáticos para prever comportamentos futuros com base em dados históricos.

O que é Análise Preditiva?

Definição

Análise preditiva é o uso de modelos matemáticos para prever comportamentos futuros com base em dados históricos.

- Baseada em estatística, aprendizado de máquina e histórico de dados

O que é Análise Preditiva?

Definição

Análise preditiva é o uso de modelos matemáticos para prever comportamentos futuros com base em dados históricos.

- Baseada em estatística, aprendizado de máquina e histórico de dados
- Previsões probabilísticas, não determinísticas

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...
   - Regressão logística aplicada ao comportamento de compra

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...
   - Regressão logística aplicada ao comportamento de compra

2. **Hospitais e readmissão de pacientes**

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...
   - Regressão logística aplicada ao comportamento de compra

2. **Hospitais e readmissão de pacientes**
   - Predição de risco de retorno com base no prontuário

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...
   - Regressão logística aplicada ao comportamento de compra

2. **Hospitais e readmissão de pacientes**
   - Predição de risco de retorno com base no prontuário

3. **Prevenção de fraude bancária**
   - Regressão para identificar padrões de fraudes

Exemplos Reais (Livro de Eric Siegel)

1. **Target e a gravidez prevista**
   - Padrões de compra sugerem gravidez antes mesmo da família saber
     - **ex:** loção sem perfume, suplementos...
   - Regressão logística aplicada ao comportamento de compra

2. **Hospitais e readmissão de pacientes**
   - Predição de risco de retorno com base no prontuário

3. **Prevenção de fraude bancária**
   - Regressão para identificar padrões de fraudes
     - Prever transações fraudulentas com base em padrões anteriores.

Ligando com Regressão

Problema	Técnica	Exemplo
Prever valores contínuos	Regressão Linear	Preço de produtos
Prever eventos binários	Regressão Logística	Diagnóstico médico
Evitar overfitting	Ridge/Lasso	Dados com muitas variáveis

Regressão Linear: Conceito

Conceito

- A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável dependente (Y) e uma variável independente (X). Ela busca ajustar uma reta que melhor representa os dados.

Regressão Linear: Conceito

Conceito

- A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável dependente (Y) e uma variável independente (X). Ela busca ajustar uma reta que melhor representa os dados.
- Seu objetivo é predizer um determinado valor **Y**, baseado em um valor **X**.

Regressão Linear: Conceito

Conceito

- A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável dependente (Y) e uma variável independente (X). Ela busca ajustar uma reta que melhor representa os dados.
- Seu objetivo é predizer um determinado valor **Y**, baseado em um valor **X**.
- Ela busca entender como uma determinada variável se relaciona com a outra.

Regressão Linear: Fórmula

Fórmula Geral

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y = Variável dependente (o que queremos prever)
- X = Variável independente (a explicativa)
- β_0 = Intercepto (valor de Y quando X = 0)
- β_1 = Coeficiente angular, inclinação da reta que representa o quanto **Y** varia com **X**
- ϵ = Erro aleatório

Regressão Logística: Conceito e Fórmula

Conceito

A regressão logística é usada para prever variáveis categóricas binárias, ou seja, quando a saída Y só pode assumir dois valores (por exemplo, 0 ou 1).

Fórmula Geral

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- $P(Y = 1|X)$ = Probabilidade de sucesso (classe 1)
- β_0 = Intercepto
- β_1 = Coeficiente da variável X
- A saída é uma probabilidade entre 0 e 1

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?
- Como evitar viés e discriminação?

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?
- Como evitar viés e discriminação?
  - **ex:** raça, gênero)

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?
- Como evitar viés e discriminação?
  - **ex:** raça, gênero)
- Como evitar viés nos dados?

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?
- Como evitar viés e discriminação?
  - **ex:** raça, gênero)
- Como evitar viés nos dados?
  - **ex:** não ser tendencioso

Reflexões Éticas

- Se você pode prever que uma pessoa fará algo, deve agir com base nisso?
  - O que fazer se um sistema prever que alguém vai cometer um crime?
- Como evitar viés e discriminação?
  - **ex:** raça, gênero)
- Como evitar viés nos dados?
  - **ex:** não ser tendencioso
- Quem é responsável por uma predição errada?

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)
  - **ex:** você usa drogas? você já traiu seu parceiro?

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)
  - **ex:** você usa drogas? você já traiu seu parceiro?
- Sobcobertura

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)
  - **ex:** você usa drogas? você já traiu seu parceiro?
- Sobcobertura
  - A pessoa não tem acesso para responder e responde com um amigo, por exemplo

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)
  - **ex:** você usa drogas? você já traiu seu parceiro?
- Sobcobertura
  - A pessoa não tem acesso para responder e responde com um amigo, por exemplo
- Amostragem de respostas voluntárias

Analisando alguns casos...

Caso 1: **Um determinado humorista resolve fazer uma pesquisa para analisar seu engajamento com seus seguidores, para isso, cria uma pesquisa online.**

A pesquisa aponta que **92%** dos **300** que responderam a pesquisa, são **"loucos"** por ele.

- Qual é a fonte de viés mais preocupante:
  - Viés da resposta (Pessoas não serão sinceras em suas respostas)
  - **ex:** você usa drogas? você já traiu seu parceiro?
- Sobcobertura
  - A pessoa não tem acesso para responder e responde com um amigo, por exemplo
- Amostragem de respostas voluntárias
  - Resultado superestimado e tendencioso.

Analisando alguns casos...

Caso 2: **Uma empresa quer prever a probabilidade de cancelamento de um serviço.**

Buscamos encontrar:

1. Identificar variáveis relevantes
2. Propor modelo de regressão
3. Discutir implicácoes éticas

Analisando alguns casos...

Caso 2: **Uma empresa quer prever a probabilidade de cancelamento de um serviço.**

Buscamos encontrar:

- **Identificar variáveis relevantes:**
  - Tempo de uso do serviço
  - Núumero de reclamações
  - Atrasos no pagamento
  - Engajamento (uso mensal, login, etc.)
  - Tipo de plano ou contrato

Analisando alguns casos...

Caso 2: **Uma empresa quer prever a probabilidade de cancelamento de um serviço.**

Buscamos encontrar:

- **Propor modelo de regressão:**
  - Regressão Logística para prever cancelamento (sim/não)

Analisando alguns casos...

Caso 2: **Uma empresa quer prever a probabilidade de cancelamento de um serviço.**

Buscamos encontrar:

- **Atividade sugerida:**
  - Identificar variáveis adicionais
  - Criar um modelo preditivo
  - Simular uma tomada de decisão com base na predição
  - Discutir implicácoes éticas

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Conjunto de Dados Simulado:

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Conjunto de Dados Simulado:

- Cliente A: 24 meses, sem atraso, plano premium, usa frequentemente

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Conjunto de Dados Simulado:

- Cliente A: 24 meses, sem atraso, plano premium, usa frequentemente
- Cliente B: 5 meses, 2 atrasos, plano básico, pouco engajado

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Conjunto de Dados Simulado:

- Cliente A: 24 meses, sem atraso, plano premium, usa frequentemente
- Cliente B: 5 meses, 2 atrasos, plano básico, pouco engajado
- Cliente C: 10 meses, 1 reclamação, plano intermediário, uso médio

Atividade Sugerida - Simulação em Sala

Cenário: Você é cientista de dados em uma empresa de serviços por assinatura.

Desafio: Com base nos dados apresentados no *dataset*, responda:

- Qual cliente tem maior risco de cancelamento?
- Quais variáveis mais influenciam esse risco?
- O que você recomendaria para reter esse cliente?

Conjunto de Dados Simulado:

- Cliente A: 24 meses, sem atraso, plano premium, usa frequentemente
- Cliente B: 5 meses, 2 atrasos, plano básico, pouco engajado
- Cliente C: 10 meses, 1 reclamação, plano intermediário, uso médio

Tarefa: Defina qual cliente você abordaria e por quê.

Dúvidas?