**Conceitos Estatísticos para IA – Turma 20IA**

**Trabalho 3**

Integrantes Grupo 5:

RM 341455 Orlando Gomes Mariano

RM 340399 Gabriel Roger

RM 341293 Felipe Neri

RM 341608 Carlos Kis

**Habilidades desenvolvidas: Análise exploratória dos dados. Análise de correlação e associação. Modelo preditivo usando a regressão linear simples**

1. Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa (arquivo: Vendas\_2016a2018.csv). Use o modelo de regressão linear simples:

|  |  |  |  |
| --- | --- | --- | --- |
| Data | ano | Vendas | Budget\_Advertising |
| jan/16 | 2016 | 1160081 | 72800 |
| fev/16 | 2016 | 1622540 | 123392 |
| mar/16 | 2016 | 1597260 | 135761 |
| abr/16 | 2016 | 1640675 | 148064 |
| mai/16 | 2016 | 1511270 | 159746 |
| jun/16 | 2016 | 1634073 | 183353 |
| jul/16 | 2016 | 1856971 | 190722 |
| ago/16 | 2016 | 1585566 | 197802 |
| set/16 | 2016 | 2041672 | 248891 |
| out/16 | 2016 | 1933557 | 256353 |
| nov/16 | 2016 | 2076910 | 298805 |
| dez/16 | 2016 | 1740202 | 268925 |
| jan/17 | 2017 | 1657258 | 104000 |
| fev/17 | 2017 | 2317914 | 176274 |
| mar/17 | 2017 | 2281800 | 193944 |
| abr/17 | 2017 | 2343821 | 211520 |
| mai/17 | 2017 | 2158957 | 228209 |
| jun/17 | 2017 | 2334390 | 261933 |
| jul/17 | 2017 | 2652815 | 272460 |
| ago/17 | 2017 | 2265094 | 282574 |
| set/17 | 2017 | 2916674 | 355559 |
| out/17 | 2017 | 2762225 | 366219 |
| nov/17 | 2017 | 2967014 | 426865 |
| dez/17 | 2017 | 2486002 | 384179 |
| jan/18 | 2018 | 2071573 | 130000 |
| fev/18 | 2018 | 2897392 | 220343 |
| mar/18 | 2018 | 2852250 | 242431 |
| abr/18 | 2018 | 2929776 | 264401 |
| mai/18 | 2018 | 2698696 | 285262 |
| jun/18 | 2018 | 2917988 | 327417 |
| jul/18 | 2018 | 3316019 | 340575 |
| ago/18 | 2018 | 2831367 | 353218 |
| set/18 | 2018 | 3645842 | 444449 |
| out/18 | 2018 | 3452781 | 457774 |
| nov/18 | 2018 | 3708768 | 533582 |
| dez/18 | 2018 | 3107503 | 480224 |
| jan/19 | 2019 |  | 91000 |
| fev/19 | 2019 |  | 154240 |
| mar/19 | 2019 |  | 169702 |
| abr/19 | 2019 |  | 185081 |
| mai/19 | 2019 |  | 199683 |
| jun/19 | 2019 |  | 229192 |
| jul/19 | 2019 |  | 238403 |
| ago/19 | 2019 |  | 247253 |
| set/19 | 2019 |  | 311114 |
| out/19 | 2019 |  | 320442 |
| nov/19 | 2019 |  | 373507 |
| dez/19 | 2019 |  | 336157 |

1. Definição do problema:

***Encontrar o valor de budget advertising para o próximo ano, e entender se o valor do budget advertising supre a quantidade de vendas.***

1. Objetivo do estudo:

***Prever o faturamento do próximo ano.***

1. Classifique os tipos de variáveis:

|  |  |  |
| --- | --- | --- |
| Variável | Tipo da variável | Regra de data mining |
| Data | ID |  |
| Ano | ***Qualitativa Ordinal*** |  |
| Vendas | ***Quantitativa Discreta*** | ***Target*** |
| Budget Advertising | ***Quantitativa Contínua*** | ***Preditora*** |

1. Qual a variável resposta do modelo? Qual a variável preditora?

***Variável Resposta = Vendas***

***Variável Preditora = Budget Advertising***

1. Apresente o gráfico de dispersão entre Vendas e Budget. Interprete o resultado.

***É possível verificar no gráfico abaixo, a correlação das variáveis vendas e budget, sugerindo por exemplo que quando há aumento de budget consequentemente temos também um aumento nas vendas.***

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

1. Qual a correlação entre Vendas e Budget? Interprete o resultado.

***Temos uma correção FORTE, atingindo o número 0.8516129, ou seja, um valor mais próximo de 1.***

1. Apresente a saída da regressão linear simples. Interprete os resultados.

***Intercepto = 1060550***

***Var. preditora = 4,964***

1. Qual a acurácia do modelo ajustado?

***~73% (0,7252)***

1. Apresente a análise de resíduos. Interprete os resultados.

*Residuals*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Min*** | ***1Q*** | ***Median*** | ***3Q*** | ***Max*** |
| *-655330* | *-256271* | *-30444* | *234875* | *743028* |

***Devido ao uso de uma variável apenas (Budget\_Advertising), temos uma diferença considerável entre o erro mínimo (R$ -655.330,00) e o máximo (R$ 743.028,00)***

1. Qual a previsão de Vendas para 2019? Utilize o arquivo Budget\_2019.csv.

***R$ 26.902.666,888 – aplicando o modelo ajustado de RL.***

1. Use os dados *Bike\_Sharing.csv* para construir as análises descritivas, correlação, associação e modelo preditivo para previsão do número de bikes alugadas por mês.

Descrição:

Os sistemas de compartilhamento de bicicletas são uma nova geração de aluguel de bicicletas tradicional, onde todo o processo de associação, locação e devolução tornou-se automático. Através destes sistemas, o usuário pode facilmente alugar uma bicicleta a partir de uma determinada posição e retornar em outra posição. Atualmente, existem cerca de 500 programas de compartilhamento de bicicletas em todo o mundo, compostos por mais de 500 mil bicicletas. Hoje, existe um grande interesse nesses sistemas devido ao seu importante papel no trânsito, questões ambientais e de saúde.

Fonte de dados: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

Atividades:

1. Classifique o tipo de variável

|  |  |  |  |
| --- | --- | --- | --- |
| **Variável** | **Descrição** | **Tipo de Variável** | **Regra de data mining** |
| **instant** | record index | ID |  |
| **dteday** | date | ID |  |
| **season** | season (~~1:springer, 2:summer, 3:fall, 4:winter~~)  1:winter, 2:spring, 3:summer, 4:fall | ***Qualitativa Ordinal*** | ***Preditora*** |
| **yr** | year (0: 2011, 1:2012) | ***Qualitativa Ordinal*** | ***Preditora*** |
| **mnth** | month ( 1 to 12) | ***Qualitativa Ordinal*** | ***Preditora*** |
| **hr** | hour (0 to 23) |  | ***Preditora*** |
| **holiday** | weather day is holiday or not | ***Qualitativa Ordinal*** | ***Preditora*** |
| **weekday** | day of the week | ***Qualitativa Ordinal*** | ***Preditora*** |
| **workingday** | if day is neither weekend nor holiday is 1, otherwise is 0. | ***Qualitativa Ordinal*** | ***Preditora*** |
| **weathersit** | 1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | ***Qualitativa Ordinal*** | ***Preditora*** |
| **temp** | Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) | ***Quantitativa Continua*** | ***Preditora*** |
| **atemp** | Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale) | ***Quantitativa Continua*** | ***Preditora*** |
| **hum** | Normalized humidity. The values are divided to 100 (max) | ***Quantitativa Continua*** | ***Preditora*** |
| **windspeed** | Normalized wind speed. The values are divided to 67 (max) | ***Quantitativa Continua*** | ***Preditora*** |
| **casual** | count of casual users | ***Quantitativa Discreta*** | ***Target*** |
| **registered** | count of registered users | ***Quantitativa Discreta*** | ***Target*** |
| **cnt** | count of total rental bikes including both casual and registered | ***Quantitativa Discreta*** | ***Target*** |

1. Definição do problema:

***Indisponibilidade de bicicletas para aluguel.***

1. Objetivo do estudo:

***Prever a quantidade de bicicletas para locação. Variável CNT***

Obs: escolha uma variável resposta: cnt ou casual ou registered.

1. Apresente as tabelas descritivas e gráficos das variáveis. Interprete os resultados.

-- Comentário da professora em 01:06:00

|  |  |  |
| --- | --- | --- |
|  | fa\_holiday | fr\_holiday |
| 0 | 710 | 0.97127223 |
| 1 | 21 | 0.02872777 |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

***Notamos que a grande demanda se dá em dias normais (não feriados). Isso nos permite deduzir que o uso das bikes podem ser usados aos locais de trabalho, por exemplo, sendo usadas menos para lazer, conforme vemos no gráfico.***

|  |  |  |
| --- | --- | --- |
|  | fa\_weekday | fr\_weekday |
| 0 | 105 | 0.1436389 |
| 1 | 105 | 0.1436389 |
| 2 | 104 | 0.1422709 |
| 3 | 104 | 0.1422709 |
| 4 | 104 | 0.1422709 |
| 5 | 104 | 0.1422709 |
| 6 | 105 | 0.1436389 |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

***Os valores da distribuição estão muito próximos entre os dias da semana.***

|  |  |  |
| --- | --- | --- |
|  | fa\_workingday | fr\_workingday |
| 0 | 231 | 0.3160055 |
| 1 | 500 | 0.6839945 |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

***Maior demanda em dias úteis do que em finais de semana, informação também identificada na tabela Holiday.***

|  |  |  |
| --- | --- | --- |
|  | fa\_weathersit | fr\_weathersit |
| 1 | 463 | 0.63337893 |
| 2 | 247 | 0.33789330 |
| 3 | 21 | 0.02872777 |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

***Verificamos a maior incidência em dias mais “limpos”.***

|  |  |  |
| --- | --- | --- |
|  | fa\_season | fr\_season |
| 1 | 181 | 0.2476060 |
| 2 | 184 | 0.2517100 |
| 3 | 188 | 0.2571819 |
| 4 | 178 | 0.2435021 |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

***Maior volume de uso no outono.***

Gráfico

Descrição gerada automaticamente

***Distribuição quase simétrica, sem presença de outliers.***

1. Quais variáveis foram normalizadas? Por quê? Apresente a fórmula utilizada.

***Essa ação é necessária pois essas variáveis não serão utilizadas com seus valores iniciais e não utilizaremos esses valores para “cálculos” (variáveis não dicotômicas).***

***As variáveis foram colocadas em “bases” semelhantes, para diminuir o efeito das grandezas das variáveis que influenciam no modelo.***

*Variáveis normalizadas:*

* ***Temp, Atemp, Hum, Windspeed***

***Fórmulas utilizadas:***

1. ***Temp****: Valores divididos entre o mínimo e máximo (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 ) – normalizada em graus Celsius.*
2. ***Atemp:*** *Valores divididos entre o mínimo e máximo* (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50) *– normalizada em graus Celsius.*
3. ***Hum:*** *Valores divididos por 100(máximo)*
4. ***Windspeed:*** *Valores divididos por 67 (máximo)*
5. Apresente a análise de associação e correlação de Pearson. Quais variáveis têm correlação com a variável resposta?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | temp | atemp | hum | windspeed | cnt |
| temp | 10.000.000 | 0.9917016 | 0.1269629 | -0.1579441 | 0.6274940 |
| atemp | 0.9917016 | 10.000.000 | 0.1399881 | -0.1836430 | **0.6310657** |
| hum | 0.1269629 | 0.1399881 | 10.000.000 | -0.2484891 | -0.1006586 |
| windspeed | -0.1579441 | -0.1836430 | -0.2484891 | 10.000.000 | -0.2345450 |
| cnt | 0.6274940 | **0.6310657** | -0.1006586 | -0.2345450 | 10.000.000 |

***A variável atemp apresenta a correlação mais forte dentre as variáveis preditoras em relação a resposta.***

1. Criar o script R para o cálculo da variável Zscore da ingestão de sal. Utilize a base de dados “sal\_lab1.xlsx”.
2. Faça o histograma da variável original e Zscore.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

Gráfico, Histograma

Descrição gerada automaticamente

1. Qual a diferença entre as duas variáveis?

O primeiro histograma mostra como está a distribuição dos dados com base na quantidade de ingestão de sal em gramas.

O segundo mostra o Z-score, onde cada dado está posicionado, ou seja, se ficou acima ou abaixo da média. Quanto mais próximo do zero melhor. Poderíamos classificar o z-score de 1 até 2 sendo bom, de 2 até 3 regular e acima de 4 inaceitável. Depende muito da quantidade da amostra, quanto mais participantes participem do ensaio, melhor.

Se o z-score é negativo indica que o ponto está abaixo da média, positivo acima da média e 0 próximo da média, um ponto incomum é se está acima de 3 ou abaixo de -3.

Data de entrega: 10/09/2021

Regina Bernal

26/08/2021