João Pedro Libonati

11720472

Respostas da prova de Predictive Analytics

**1.** A diferença entre regressão linear e regressão logística está no objetivo dessas regressões. Ambas as técnicas usam variáveis numéricas na equação. Porém, a diferença está no resultado gerado pela equação. A regressão linear é usada para obter um valor númerico enquanto que a regressão logística é usada para obter um valor binário, podento até resultar em um valor categórico. Portanto, usa-se o modelo de regressão logística quando o objetivo é obter uma variável binária.

Um exemplo de regressão linear é achar a temperatura de uma placa de vídeo a partir do modelo da placa, tamanho da placa, número de ventuinhas da placa, tamanho do gabinete, potência da placa, etc.

Um exemplo de regressão logística é fazer uma predição se o computador irá desligar a partir da temperatura média da cpu, temperatura da placa de vídeo e o consumo de energia do computador.

**2.**

**a)** Como a variável independente é um número (uma porcentagem), é adequado usar a regressão linear, já que a regressão linear nos devolve uma variável numérica contínua. Poderia ser usado também a técnica de regressão logística para obter como resultado uma classificação.

**b)** Sim. Há causalidade, porque a pontuação foi feita com base nestas duas variáveis, ou seja, esta variável deve ser uma representação do efeito, porque se não for, a variável não faz sentido de ter sido criada.

**c)** Sim. Como podemos ver no "Basic Summary", o R-squared é de 0.8, mostrando que o modelo é bom para predizer a pontuação. E os p-values de ambas variáveis estão mostrando que o modelo tem baixíssima probabilidade de estar errado. Com isso, o modelo aparenta ser bem eficaz.

**3.**

**a)** Não, já que a variável de resposta é do tipo binário, não deve ser usada a técnica de regressão linear, mas sim, a regressão logística.

**b)** Considerando as variáveis após o stepwise, essas variáveis fazem sentido de serem aplicadas. Já que elas descrevem como os clientes atuam e mais especificamente, padrões de inadimplência. Mesmo tendo variáveis de 2005, faz sentido usar essas variáveis para predizer a situação de 2008, já que há muitas variáveis independentes no modelo.

**c)** Mesmo com um p-value baixo para muitas variáveis no modelo com stepwise, não possui um bom grau de predição. Porque o R-squared está muito baixo (0.12), ou seja, essas variáveis provavelmente explicam pouco a variável dependente.

**X.** Não foi feita.

Questões abaixo para referência:

Disciplina: **Predictive Analytics**

Professor: **André Insardi – andre.insardi@espm.br**

Período: **Eletiva** Aplicação: **2020\_02**

**Nome Aluno: João Pedro Libonati**

|  |
| --- |
| **PROVA PREDICTIVE ANALYTICS** |

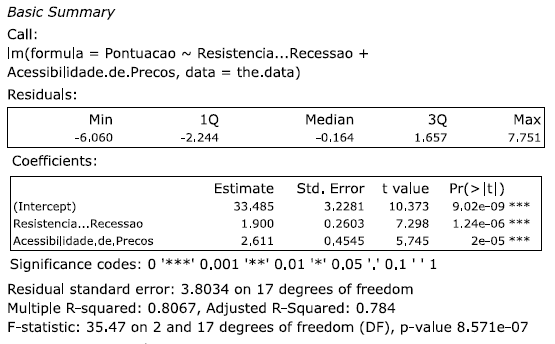
**[2,0] 1.** Explique com suas palavras a diferença entre Regressão Linear e Regressão Logística, cite exemplos para ilustrar sua explicação.

**[4,0] 2.** A revista *SmartMoney* avaliou 65 regiões metropolitanas para determinar onde os preços das casas eram mais altos. Uma cidade ideal obteria uma pontuação 100 se todos os fatores medidos fossem os mais favoráveis possíveis. Regiões com pontuações a partir de 60 são consideradas de primeira linha em termos de valorização de preço, e as regiões com uma pontuação abaixo de 50 podem sofrer dete­rioração de preços das moradias. Dois dos fatores avaliados foram a resistência da região à recessão econômica e sua acessibilidade e preços. Ambos os fatores foram classificados usando-se uma esca­la que varia de 0 (pontuação baixa) a 10 (pontuação alta). Os dados obtidos de uma amostra de 20 cidades avaliadas pela *SmartMoney* são apresentados a seguir (*SmartMoney,* fevereiro de 2015).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Resistência | Acessibilidade |  |
| Região Metropolitana | à Recessão | de Preços | Pontuação |
| Tucson | 10 | 7 | 70,7 |
| Fort Worth | 10 | 7 | 68,5 |
| San Antonio | 6 | 8 | 65,5 |
| Richmond | 8 | 6 | 63,6 |
| Indianapolis | 4 | 8 | 62,5 |
| Filadélfia | 0 | 10 | 61,9 |
| Atlanta | 2 | 6 | 60,7 |
| Phoenix | 4 | 5 | 60,3 |
| Cincinnati | 2 | 7 | 57,0 |
| Miami | 6 | 5 | 56,5 |
| Hartford | 0 | 7 | 56,2 |
| Birmingham | 0 | 8 | 55,7 |
| San Diego | 8 | 2 | 54,6 |
| Raleigh | 2 | 7 | 50,9 |
| Oklahoma City | 1 | 6 | 49,6 |
| Orange County | 4 | 2 | 49,1 |
| Denver | 4 | 4 | 48,6 |
| Los Angeles | 0 | 7 | 45,7 |
| Detroit | 0 | 5 | 44,3 |
| Nova Orleans | 0 | 5 | 41,2 |

Com o objetivo de criar um modelo para prever a pontuação da região, foi desenvolvido em Alteryx uma regressão linear para os dados apresentados.

O resultado do modelo está apresentado no *report* a seguir:



1. A técnica de Regressão Linear escolhida para a construção do modelo preditivo de Pontuação é adequada? Explique. [2.0]
2. Existe relação de causa e efeito entre as variável dependente e as variáveis Independentes? Explique. [1.0]
3. Em sua análise o modelo possui um bom grau de predição? Explique. [1.0]

**[4,0] 3.** Com a crise econômica mundial de 2008, os níveis de inadimplência se elevaram nas aplicações de crédito e as instituições financeiras buscaram sistemas mais apurados para a gerência de risco.

Os dados financeiros e sociais do cliente são analisados a partir de modelos preditivos. Essa avaliação tem um impacto significativo nos lucros das empresas, as decisões devem ser criteriosas, pois caso errem pode provocar prejuízos enormes.

O dataset conta com 23 atributos, 30000 registros, trata a probabilidade de inadimplência de clientes de cartão de crédito em Taiwan

O conjunto de dados (base de dados anexa no canvas) empregou uma variável binária (default payment next month – inadimplência [Sim = 1, Não = 0]), como variável de resposta. Seguem mais informações dos atributos:

* X1: Valor do crédito dado (dólar): inclui tanto o crédito ao consumo individual como o crédito (suplementar) da sua família.
* X2: Sexo (1 = masculino, 2 = feminino).
* X3: Educação (1 = pós-graduação, 2 = universidade, 3 = ensino médio, 4 = outros).
* X4: Estado civil (1 = casado, 2 = solteiro, 3 = outros).
* X5: Idade (anos).
* X6 – X11: Histórico de pagamentos anteriores. Rastreamento dos últimos registros de pagamentos mensais (de abril a setembro de 2005) da seguinte forma:
* X6 = o status de pagamento em setembro de 2005
* X7 = o status de pagamento em agosto de 2005

…

* X11 = o estado de pagamento em abril de 2005
* A escala de mensuração para o status de reembolso é:
* -1 = pagamento em dia
* 1 = atraso de pagamento por um mês
* 2 = atraso de pagamento por dois meses

…

* 8 = atraso de pagamento por oito meses
* 9 = atraso de pagamento por nove meses ou mais
* X12-X17: Valor da fatura (em dólar).
* X12 = valor da fatura em setembro de 2005
* X13 = valor da fatura em agosto de 2005

…

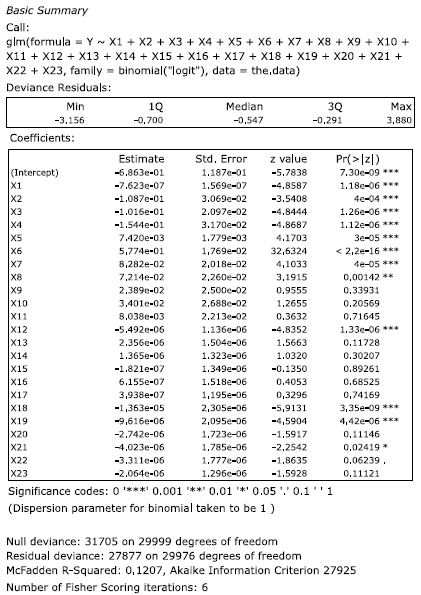
* X17 = valor da fatura em abril de 2005
* X18-X23: Valor do pagamento anterior (em dólar).
* X18 = valor pago em setembro de 2005;
* X19 = valor pago em agosto de 2005;

…

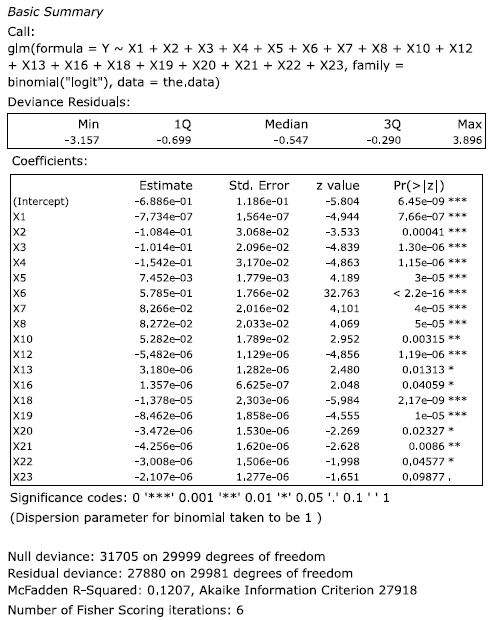
* X23 = valor pago em abril de 2005.

Com o objetivo de criar um modelo para prever a inadimplência no banco de Tawian, foi desenvolvido em Alteryx uma regressão logística para os dados apresentados.

O resultado do modelo está apresentado no *report* a seguir:



Modelo 01 - sem stpewise



Modelo 02 - com stpewise

1. A técnica de Regressão Logística escolhida para a construção do modelo preditivo é adequada? Explique. [2.0]
2. Existe relação de causa e efeito entre as variável dependente e as variáveis Independentes? Explique. [1.0]
3. Em sua análise o modelo possui um bom grau de predição? Explique. [1.0]

**[3,0] X.** Questão bônus – Os dados da questão 02 e questão 03 estão disponíveis no Canvas, será concedido 1,5 ponto adicional para cada workflow desenvolvido em Alteryx. Esta questão NÃO é obrigatória, os pontos conquistados nela serão somados até o limite de 10 pontos da prova.