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.