Técnicas de Machine Learning III

Interpretação e proposição dos resultados

Contexto de análise desse estudo de caso

Estudo sobre STARTUP onde investidores desejam realizarem um estudo com base numa amostra de startups americanas para verificar a relação entre o lucro apresentado por essas empresas e os tipos de despesas e investimentos que elas realizam. Ou seja analisar quais são os fatores que explicam alguns fatores determinantes da lucratividade de startups. O investidor deseja investigar direcionadores, do lucro de startups. Modelo preditivo para auxiliar no investimento de companhias.

Y: Lucro (Variável Dependente)

X: P&D; ADM; MKT; Estado(variáveis explicativas)

Estatística F: Avaliação Geral do Modelo

Se o pValor < Nível Significância 🡪 Rejeito H 🡪 Rejeito Hipótese Nula 🡪(Aceito o modelo)

Se o Pvalor > Nível Significância 🡪Não Rejeito H0 🡪Não Rejeito Hipótese Nula 🡪 (Rejeito o Modelo)

Nível de Significância Normalmente usado 5%

p-Valor de cada estatística T segue o mesmo critério que o p-valor da estatística F

O alpha, mesmo que ele não seja significante, ele não deve ser removido do modelo.

A remoção de variáveis vale apenas para as variáveis X.

O alpha (Intercepto) permanece mesmo não sendo estatisticamente significante

Exclusão é para variáveis x, para os β, o α, não

R²: Coeficiente de explicação, mostra o percentual, variando entre 0 e 1, da variabilidade de Y que é explicado pelo conjunto das suas variáveis X

R² Perto de 1 = Modelo Bom

R² Perto de 0 = Modelo Ruim

Quando os β Forem estatisticamente 0, fazemos a remoção deles pelo procedimento de STEPWISE.

Por que usar STEPWISE ?

Porque quando você retira uma variável do seu modelo, pode ser que outra variável que não era estatisticamente significativa passe a ser, pois todas as correlações estão trabalhando para a geração desse modelo.

A interpretação dos modelos de regressão linear múltipla ocorre sempre na condição seteris paribus, ou seja, tudo mais contante. Porque reconhecemos que há uma correlação entre as variáveis explicativas.

Com esse processo um modelo de regressão múltiplo, pode se tornar um modelo de regressão simples.

Coeficiente de β > significa influência positiva

Cuidado com interpretações casuais, quase sempre não é interessante.

Para comparação de modelo usar R² ajustado.

Fitted values: Substituir cada observação da amostra na fórmula do modelo obtido

Fim Bloco 1

Bloco 2

Banco de dados tipo cros-sections: cada linha é independente

Banco de dados em serie temporal: Cada linha representa um momento no tempo

Para modelos estimados por mínimos quadrados ordinários, seja a regressão linear simples ou a regressão linear múltipla, a variável dependente (variável y) tem que ser métrica, ou seja uma variável quantitativa.

Existem outros modelos específicos, que lidam com variáveis Y categóricas, são modelos diferente dos estimados por mínimos quadrados ordinários, podendo ser chamados de modelos de classificação.

Variáveis altamente correlacionadas, acarretam em coeficientes não significativos. Isso não impede que o modelo seja estimado, porque mesmo que uma variável seja altamente correlacionada com outra variável x, se uma delas não passar no critério, ela será removida pelo procedimento de STEPWISE, porque o conteúdo informacional dessa variável já está capturado pela outra variável, altamente correlacionada. A informação não é perdida

* Resíduos pequenos geram um R² alto

Logotipo

Descrição gerada automaticamente

Quanto mais próximo de 0 o resíduo mais próximo de 1 o R²

Y: Preço anunciado

X Características das casas

Get\_Dummies drop-first = true ja descarta um dos parâmetros na dumização, para ficar no intercepto, descartando o primeiro em ordem alfabética

Texto\_formula é uma função genérica para criação de uma formula de regressão com todas as variáveis disponíveis no data Frame.

Muito Cuidado com as interpretações de sinais positivos e negativos quando temos regressões múltiplas, lembrar sempre da clausula ceteris paribus (“Tudo mais constante”) o que significa que aquele coeficiente é valido mantendo todas as outras condições (variáveis) iguais. Alguns efeitos podem ser inesperados por causa do conjunto, do modelo completo, principalmente quando temos correlações muito elevadas.

Fim Bloco 2

Bloco 3

Cuidado com Variáveis que já vem em forma Dummizada

Y: Avaliação do Funcionário

X: Educ; Sexo; Recruta; trein; Idade; Nota; anos\_emp; bateu\_metas; Premiado;

Variáveis de interesse Principal: Tipo de recrutamento, treinamentos

Tese: Será que a forma como o funcionário está sendo recrutado impacta na avaliação? e se com treinamentos essa avaliação é melhorada.

Quando houver missing value antes de iniciar qualquer análise aplicar o tratamento de excluir as linhas faltantes. Função usada para isso: dropna(inplace = true) o implace = true serve para indicar que o novo objeto substituirá o original.

Dummizar apenas as variáveis categóricas que já não estejam Dummizadas

Como no caso especifico eu tenho duas variáveis de interesse principal, pode se estabelecer duas regressões lineares **SIMPLES**, uma para cada variável para ver como elas isoladamente influenciam sobre a variável target(y), Já que no contexto de outras variáveis em conjunto podemos ter uma mudança de interpretação.

Modelo Final

Avaliação = 4,67+0,77trem; -0,07 idade; +0,03 setor; +0,08 metas; + 0,57 premiado; + 0,12 pos; -0,05 masc; -0,42 sourcing

Fim