4intelligence

Scalable
Decision
Science

PROVA DE ECONOMIA

Questão 1

A 4intelligence busca construir dados alternativos a fim de agregar informações não capturadas por indicadores tradicionais às nossas modelagens e análises. Nesse exercício, você será desafiado a desenvolver uma metodologia para um indicador alternativo baseado em avaliações online de estabelecimentos comerciais. Nosso objetivo é avaliar a sua capacidade analítica e a sua criatividade para propor soluções.

Suponha que você tenha acesso a uma base de dados com avaliações de estabelecimentos comerciais feitas no Google. Cada linha dessa base de dados é uma avaliação individual i sobre o estabelecimento j que opera no setor k feita no dia t^0. Uma avaliação consiste em uma nota de 0 a 5 (score) e pode ou não ter um comentário associado (review). Para avaliações com comentários é possível identificar o assunto a que o comentário se refere. A tabela abaixo ilustra a estrutura dos dados.

t^0	j	k	**score**	**review**	**categoria**
06/04/2021	777777	Farmácia	5.0	"Variedade e atendentes muito educados"	"serviço", "ambiente"
06/04/2021	333333	Hotelaria	4.0	"Higiene impecável nos quartos e área comuns"	"serviço"
06/04/2021	333333	Hotelaria	3.0	"Mal localizado"	"localização"
05/04/2021	555555	Farmácia	2.0	"Nunca tem lugar para parar o carro!"	"estacionamento"
05/04/2021	777777	Farmácia	4.0		
05/04/2021	777777	Farmácia	1.0	"Farmácia mais cara da cidade, não recomendo"	"preço"
06/04/2021	777777	Farmácia	2.0	"Progama de cash back é uma enganação"	"preço"

Esse exercício irá conduzi-lo na criação de uma metodologia para um indicador de reputação do varejo com aberturas setoriais. Algumas propriedades desejáveis para esse indicador são: i) avaliações recentes são mais relevantes. Com passar do tempo, uma dada avaliação deve perder importância; ii) deve levar em conta a heterogeneidade dos estabelecimentos, ou seja, o tamanho relativo do estabelecimento indica o quão representativo é o estabelecimento em uma agregação; iii) deve levar em consideração a heterogeneidade entre os setores; iv) frequência diária; e v) facilidade de agregação por setor, por região, etc...

De posse das informações acima, faça o que se pede:

1. Como primeiro passo, estruture uma fórmula para atribuir uma reputação a um estabelecimento j do setor k em um dia arbitrário t a partir de todas as avaliações individuais

que o estabelecimento recebeu até então. Por enquanto, não se preocupe com a coluna de reviews.

2. De posse da reputação de cada estabelecimento, estruture um modo de termos uma agregação para um determinado setor κ em um dia qualquer t. Atente-se para as dificuldades de agregação, é preciso atribuir maior ou menor relevância para algum estabelecimento dentro de um mesmo setor?

3. Suponha agora que queiramos agregar todos os setores. Construa um método para um indicador agregado de reputação em um dia qualquer t. Atente-se para as heterogeneidades setoriais.

4. Como último passo, note que para além da nota em si, as avaliações levam em conta informações qualitativas com os reviews. Como você incorporaria essas observações na sua metodologia?

Questão 2

Assim como os demais países, a economia brasileira foi fortemente afetada pela crise sanitária do coronavírus. Contudo, a magnitude e a direção dos impactos foram bastante distintas a depender dos setores e das regiões.

Utilize as aberturas de três pesquisas do IBGE - PIM, PMC e PMS - para discorrer a respeito dos impactos observados nos setores. Explique as diferenças das dinâmicas em 2020 e apresente a sua visão para o que você imagina que deva acontecer em 2021.

Diretrizes: i) para essa questão, você pode (e deve!) utilizar dados a fim de embasar os argumentos. Visite o Sidra, baixe os dados, visualize e compartilhe na sua resposta; ii) cuidado com "análises de elevador", busque refletir o porquê das dinâmicas recentes; iii) fique à vontade para fazer uso de outras informações, como: PIB e suas aberturas, Pnad, Caged, IPCA, dados de política fiscal, balanço de pagamentos e os instrumentos formais de comunicação da política monetária por parte do Banco Central; iv) cópia e plágio implicam desclassificação imediata; v) síntese e objetividade são apreciadas. Tenha em mente o máximo de 5 páginas com gráficos e tabelas.

Questão 3 (Programação)

Para tentar construir algumas análises acerca do comportamento do mercado de trabalho ao longo da pandemia, você recebeu uma base com informações provindas da PNAD Contínua Trimestral para o 4º trimestre de 2019 e também para o 4º trimestre de 2020. A base contém as variáveis descritas na tabela abaixo.

Ano	Ano de referência
Trimestre	Trimestre de referência
UPA	Unidade Primária de Amostragem (UPA)
V1008	Número de seleção do domicílio
V1014	Painel
UF	Unidade da Federação
V2008	Dia de nascimento
V20081	Mês de nascimento
V20082	Ano de nascimento
V2007	Sexo
V2009	Idade
V2010	Cor ou raça
VD3004	Nível de instrução mais elevado alcançado (pessoas de 5 ou mais de idade) padronizado para o Ensino fundamental - SISTEMA DE 9 ANOS
VD4001	Condição em relação à força de trabalho na semana de referência para pessoas de 14 anos ou mais de idade. 1 = Pessoas na força de trabalho
VD4002	Condição de ocupação na semana de referência para pessoas de 14 anos ou mais de idade. 1 = Pessoas ocupadas
VD4020	Rendimento mensal efetivo
VD4019	Rendimento mensal habitual
VD4031	Horas habitualmente trabalhadas na semana de referência
VD4035	Horas efetivamente trabalhadas na semana de referência
V4005	Na semana de referência, tinha algum trabalho remunerado do qual estava temporariamente afastado? Sim = 1

Considere que para identificar um único domicílio na PNAD Contínua, utilizamos três variáveis: “UPA”, “V1008” e “V1014”. Para identificar os indivíduos, utilizamos o domicílio, a data de nascimento e o sexo.

Antes de começar a explorar os dados, devem ser excluídos da base:

- Pessoas com dia de nascimento igual a 99.
- Gêmeos (os gêmeos são identificados da mesma forma que os indivíduos, discriminando adicionalmente o período. Ou seja, observações que tenham mesmo identificador de domicíllio, data de nascimento e sexo em um único período são considerados gêmeos e devem ser excluídos da amostra).

De posse da base limpa, faça o que se pede:

1. No 4º trimestre de 2019, quantas pessoas estavam ocupadas? Analise o perfil desses indivíduos. Proponha visualizações para sua análise. Faça a mesma coisa para 2020 e compare.
2. Considere agora os três grupos abaixo. Como se comportou a renda média (tanto efetiva quanto habitual), em cada um dos casos, quando comparamos o 4º trimestre de 2020

com 4º trimestre de 2019?

Grupos:

- Ocupados.
- Ocupados com redução de jornada (considere como regra de bolso que a pessoa está ocupada com jornada reduzida se as horas efetivamente trabalhadas são ao menos 25% menores que as horas habitualmente trabalhadas).
- Ocupados, mas temporariamente afastados de seu trabalho.

Observações gerais para prova

- Estamos interessados em avaliar não só o conhecimento prévio, mas também a capacidade de aprender do candidato.
- O não cumprimento de alguma tarefa não será desclassificatório.
- O teste deve ser feito individualmente, sem auxílio de terceiros.
- Para a parte escrita, envie-nos um pdf.
- Para a parte de programação, o candidato deve nos enviar de volta um código limpo e reproduzível por meio de um link do github na linguagem que quiser. Prefira notebooks: Rmarkdown para R e Jupyter para Python. Devemos ser capazes de reproduzir qualquer resultado que você queira nos mostrar.
- A qualidade visual dos gráficos gerados, bem como o formato geral da apresentação dos códigos, também serão levados em consideração.
- Os arquivos de input enviados por nós não podem conter nenhuma alteração. Todas as modificações têm que ser feitas em código. Assim, não serão aceitos scripts que operem sobre uma base diferente da versão enviada ao candidato.
- O código tem que ser auto-contido, ou seja, ao ser executado deve chegar aos resultados a que se propôs.
- Fique à vontade para exercer a sua criatividade! Ir além e nos surpreender será considerado um diferencial!