2022.2

BOLETIM DE PROJETOS

INSPER DATA 2022.2

LAURA

Mensagem inicial

Caro leitor,

Com muita satisfação, o Insper Data convida todos a conhecerem os projetos desenvolvidos durante o semestre.

O Insper Data é uma organização estudantil focada em pesquisa e ciência dos dados. Nosso propósito é garantir o desenvolvimento de forte capacidade analítica e de execução em uma entidade que preza pela excelência e contínuo aprendizado, utilizando métodos estatísticos para resolver problemas reais.

Caminhamos nessa direção através da realização de projetos semestrais, nos quais os grupos escolhem tanto tema quanto orientador, que pode ser alguém com expertise acadêmica ou corporativa – a depender do escopo estudado.

No semestre de 2022.2 foram confeccionados 5 trabalhos, abarcando as grandes áreas da Microeconomia, Macroeconomia, Marketing, e Modelagem Preditiva.

Sumário

A Pandemia e o Efeito da Renda: O impacto no desempenho do ENEM na cidade de São Paulo

Integrantes: Gabriel Villaça Mencacci e Mel Bordin Beloni

Orientador: Pedro Picchetti

Resumo

O presente trabalho busca entender como a pandemia impactou o desempenho escolar de alunos de diferentes classes sociais na cidade de São Paulo. Para isso, foi feita uma análise de Diferenças em Diferenças sobre o desempenho observado no Exame Nacional do Ensino Médio (ENEM) e no qual observou-se que, com a pandemia, houve uma redução média de 10.5 pontos na diferença das notas entre pessoas que advém de famílias com maior renda e pessoas que advém de famílias com menor renda.

Introdução e Revisão Bibliográfica

O presente trabalho utilizou como base teoria o trabalho de Curi, Menezes-Filho, Faria e Educação, T.P (2009), os quais em seu artigo analisam o papel da escola no desempenho dos alunos do ensino médio das melhores escolas particulares e públicas do Estado de São Paulo através do ENEM. Isso de modo a entender se o valor da mensalidade escolar está relacionado com o desempenho médio da escola. Como resultado os autores encontraram que um aumento de 10% no valor da mensalidade escolar aumentava a nota do aluno em 1,1% mesmo com o controle do background familiar.

Outro artigo sobre o qual o presente trabalho se baseou foi o estudo de Lichand e Doria (2022) que buscava entender o impacto que o regresso às aulas presenciais teve em atenuar os retrocessos na educação obtidos na pandemia no estado de São Paulo. Os

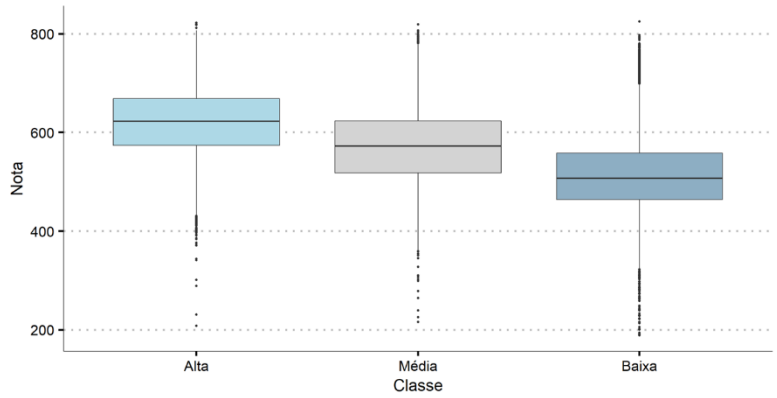
autores encontraram que apesar dos alunos apresentarem um atraso de 55% quando comparado ao que eles teriam aprendido em circunstâncias normais, os mesmos aprenderam cerca de 38 a 45% mais rápido em 2021 do que em um ano típico.

Tendo esses resultados em consideração, o presente trabalho busca entender como a pandemia impactou o desempenho escolar dos alunos da cidade de São Paulo, mas considerando que esse impacto foi de diferente magnitude para as distintas classes sociais.

Análise descritiva

Com a base de Dados do Instituto Nacional de Estudos e Pesquisa Anísio Teixeira foram feitas as análises descritivas. Desse modo, a partir dos dados foi visto que conforme mais renda, em média, espera-se um melhor desempenho no ENEM, como é possível observar na Imagem 01 abaixo.

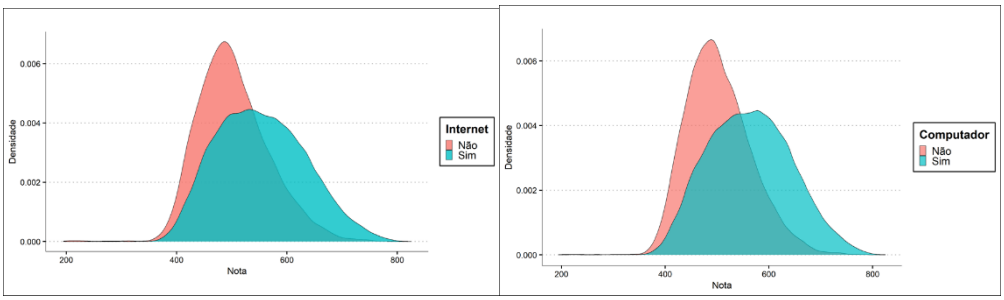
Imagem 01 – Box Plot das notas conforme a renda dos indivíduos



Fonte: Elaboração própria

Uma alternativa para medir a renda dos participantes é partindo de recursos que justificam as diferenças nos desempenhos dos grupos sociais, assim foi comparado o acesso à internet e computadores. Destaca-se que embora haja pouca representatividade no vestibular na cidade de São Paulo, em média, os alunos que não possuem esses recursos do ensino à distância tendem a ter um desempenho pior no vestibular do ENEM.

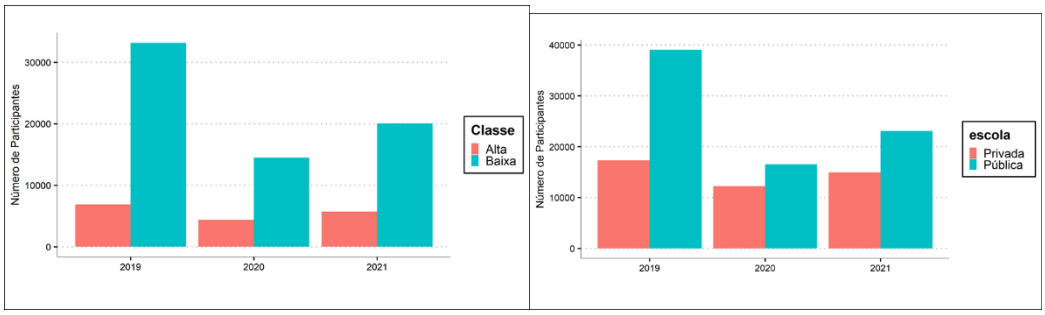
Imagem 02 – Histograma dos indivíduos de acordo com o acesso à internet e computador.



Fonte: Elaboração própria

Além disso, foi estudado o perfil dos estudantes que realizam o exame ao longo dos três anos, para verificar se houve reduções ou alterações no perfil dos alunos. Após a análise, é possível observar na Imagem 03 que houve uma drástica redução na participação de estudantes de escolas públicas, enquanto a participação de alunos de escolas privadas permaneceu aproximadamente constante. Esse resultado está sujeito a interpretação de que houve desistência por parte de alunos de menor renda devido a falta de recursos durante a pandemia.

Imagem 03 – Gráfico de barras ao longo dos anos estudados categorizado por classe social e tipo de escola ao qual o indivíduo frequentou.

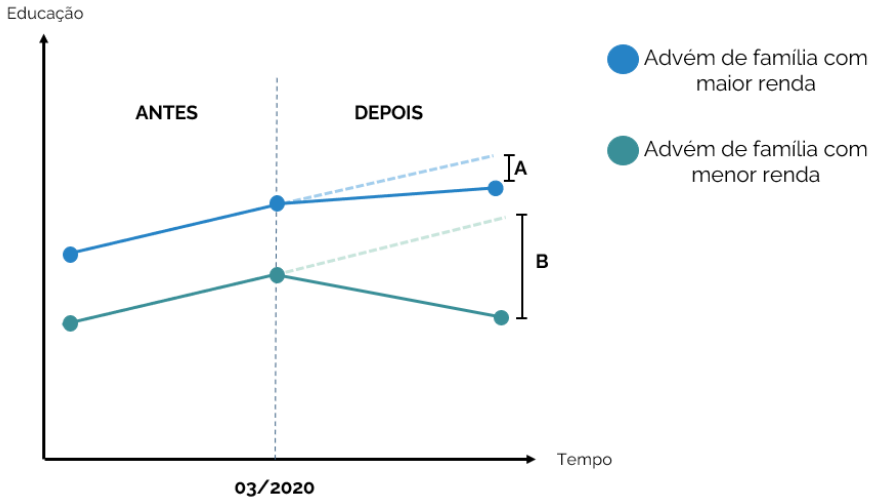


Fonte: Elaboração própria

Metodologia

Para a análise empírica foi utilizada como metodologia uma variação do método de Diferenças em Diferenças (DID), isso dado a natureza do trabalho no qual não há um grupo de controle e sim dois grupos tratados. Os grupos de estudo foram os indivíduos que realizaram o ENEM e que advém de famílias com maior renda e indivíduos que realizaram o ENEM mas que advém de famílias com menor renda. E, tendo como marco temporal de tratamento março de 2020; o início da pandemia. O problema é retratado na imagem 04.

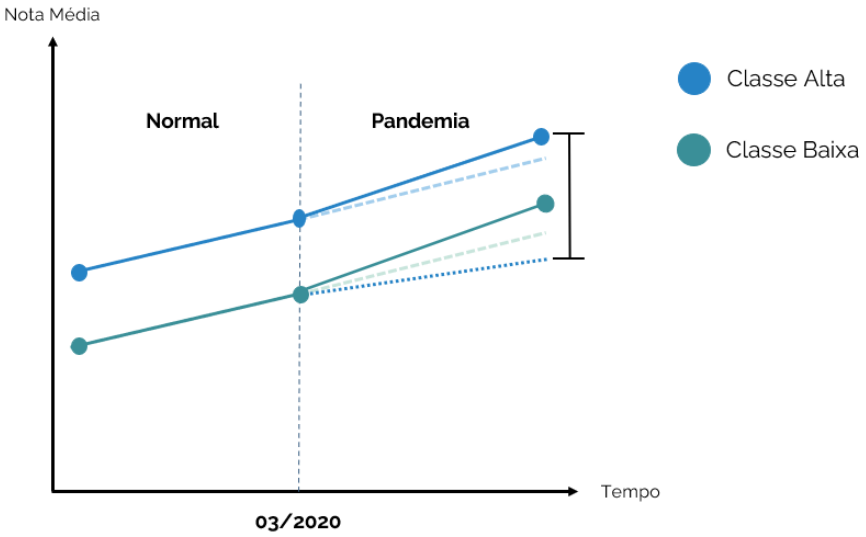
Imagem 04 – Gráfico do método de Diferenças em Diferenças



Fonte: Elaboração própria

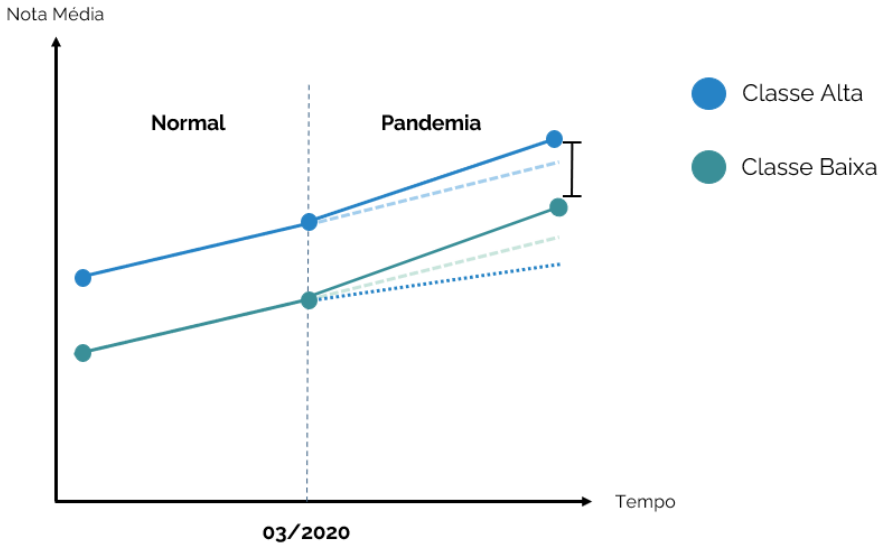
Sendo assim, fez-se uma subtração dupla, a primeira referente à diferença das médias do grupo de indivíduos de classe alta em relação ao que deveria ser a trajetória dos indivíduos de classe baixa caso esses dispusessem dos mesmos recursos que os indivíduos do outro grupo (acesso a mesma nível de educação e utensílios) como é possível observar na Imagem 05, e a segunda referente à diferença entre os dois grupos após o tratamento (Imagem 06)

Imagem 05 – Primeira Diferença



Fonte: Elaboração própria

Imagem 06 – Segunda Diferença



Fonte: Elaboração própria

Além disso, para a efetividade do método fez-se necessário à validação das hipóteses de tendências paralelas e de efeitos homogêneos das "doses" de tratamento.

Conclusões e Limitações

Com a realização do Diferenças em Diferenças foi observado que a pandemia apresentou um efeito médio negativo de 10.5 na nota do ENEM dos indivíduos de São Paulo. Isso é, com a pandemia, houve uma redução média de 10.5 pontos na diferença

das notas entre pessoas que advém de famílias com maior renda e pessoas que advém de famílias com menor renda.

Entretanto, para dar maior robustez a análise, foi estimada uma regressão de modo a testar a relevância estatística do resultado encontrado. Como é possível observar na Imagem 07, variáveis como classe social, escola, sexo, raça, ter acesso a computador e *internet* apresentaram relevância estatística a um nível de confiança de 95%.

Imagem 07 – Tabela dos resultados da estimação da regressão aplicada

	Valor Estimado	Desvio Padrão	Valor - p
Rico	59,55	1,02	< 0,1%
Escola	35,53	0,90	< 0,1%
Sexo	-15,15	0,60	< 0,1%
Computador	16,66	0,72	< 0,1%
Raça	-13,39	0,64	< 0,1%
Pandemia	4,64	0,59	< 0,1%
Internet	4,90	1,24	< 0,1%

Fonte: Elaboração própria

Portanto, com o modelo de regressão, as análises descritivas e o contexto social observado atualmente, foi possível concluir que a pandemia desestimulou os estudantes de menor renda a participar do ENEM, de maneira que apenas os alunos que já teriam um desempenho satisfatório fizessem a prova. Em outras palavras, dadas as condições sociais que os diferentes grupos se encontram, houve uma pré-seleção dos participantes do ENEM, visto que jovens de menor renda que usualmente iriam pior, nem tentaram devido a qualidade da educação entre 2019 e 2021, enquanto os grupos de maior renda com

melhor educação continuaram com o mesmo grupo representante no vestibular. Por isso, as diferenças das notas das classes sociais diminuíram nos últimos anos.

Referências Bibliográficas

Curi, A. Z., Menezes-Filho, N. A., Faria, E. M., & Educação, T. P. (2009). A relação entre mensalidade escolar e proficiência no ENEM. XXXVII Encontro Nacional de Economia. Foz do Iguaçu.

Lichand, G., & Alberto Doria, C. (2022). The Lasting Impacts of Remote Learning in the Absence of Remedial Policies: Evidence from Brazil. The Lasting Impacts of Remote Learning in the Absence of Remedial Policies: Evidence from Brazil (September 3, 2022).

Determinação de preços de peças de roupa e acessórios usados: um estudo sobre predição

Integrantes: Carolina Bromfman, Laura Casarin

Orientador: Giuliana Isabella

Resumo

Através de métodos preditivos o trabalho pretende prever quais são os preços que devem ser submetidos em um site de venda e aluguel de peças de roupas e acessórios usados, de modo a reduzir o estoque para certas peças e tentar aumentar a margem de lucro para outras. O resultado foi um modelo preditivo que gera um intervalo de preços que melhor se encaixa no objetivo do projeto.

Observação inicial

Esse semestre, o grupo de Marketing realizou um projeto em conjunto com a uma plataforma que conecta pessoas que desejam vender ou alugar suas próprias peças de roupa e acessórios com outras pessoas que estariam dispostas a alugar ou comprar essas peças em questão. Dessa forma, esse trabalho deve manter algumas informações em sigilo, inclusive o nome da empresa. Por isso, idealmente, não será compartilhado informações (previamente estabelecida em contrato) que possam ser sensíveis para a empresa, a qual contribuiu com dados e informações para esse trabalho.

Introdução

Na tentativa de conectar as pessoas que querem vender ou alugar as peças com aquelas que estão dispostas a comprar ou fazer o aluguel delas, a empresa precisa fechar contratos com as clientes que estão se “desfazendo” de suas peças. Além de outros assuntos contratuais e legais, é importante que esse contrato contenha qual é o intervalo de preços que a plataforma pode atribuir àquela peça ou acessório. Por exemplo, o contrato terá claro quais serão os preços que a empresa e a cliente acordaram.

Atualmente, normalmente, os preços que são acordados entre a empresa e a cliente são pautados em uma conversa simples e informal. Isso significa que não há nenhum racional para a formulação desse intervalo de preço constado no contrato, o que pode acabar maximizando o estoque da empresa ou diminuindo as suas margens de lucro.

Dessa forma, o objetivo desse trabalho é formular um modelo que faça a predição de um intervalo de preços para cada peça de forma a minimizar o seu tempo de venda (é importante dizer que a empresa demandou uma pesquisa para as peças que serão vendidas, e não alugadas).

Ainda, esse intervalo de confiança pode possibilitar a empresa a ajustar os preços às suas estratégias de marketing, de maneira mais racional.

Revisão Bibliográfica

Como o objetivo do trabalho é, em pelo menos um primeiro momento, criar um modelo a fim de montar um intervalo de confiança de preços para os contratos entre a *Empresa* e as donas de cada peça a ser vendida, é preciso entender sobre predição. Para isso, é preciso compreender quais são as principais variáveis que são determinantes para os preços das peças. Segundo o paper "A study on the attributes involved in the hedonic pricing of Brazilian clothing" (Mandotti, Bergmann, 2010), há dois tipos de atributos que devem ser levados em consideração: os atributos intrínsecos, aqueles que descrevem a peça, e os atributos extrínsecos, aqueles externos a peça como marca, tamanho da loja etc.

Ainda, além das características internas ou externas da peça em si, é preciso que uma predição de preços seja coerente com as possíveis estratégias de marketing aplicadas pela plataforma Empresa (Dolgui, Proth, 2010).

Agora que foi compreendido quais variáveis devem ser consideradas para a predição, e considerando que a predição deve permitir a empresa devem permitir a aplicação de estratégias de marketing foi utilizado a metodologia utilizada para tal.

Segundo o *paper* de Jing Lei, Max G'Sell uma maneira de fazer isso é utilizar a mesma tecnologia do Random Forest em forma de intervalo de confiança. Essa solução conseguirá utilizar as variáveis relevantes para atribuir um preço a peça e, ao mesmo tempo, dar flexibilidade maior a empresa a atribuir preços dependendo da estratégia de marketing sugerida pela equipe de marketing da *Empresa*.

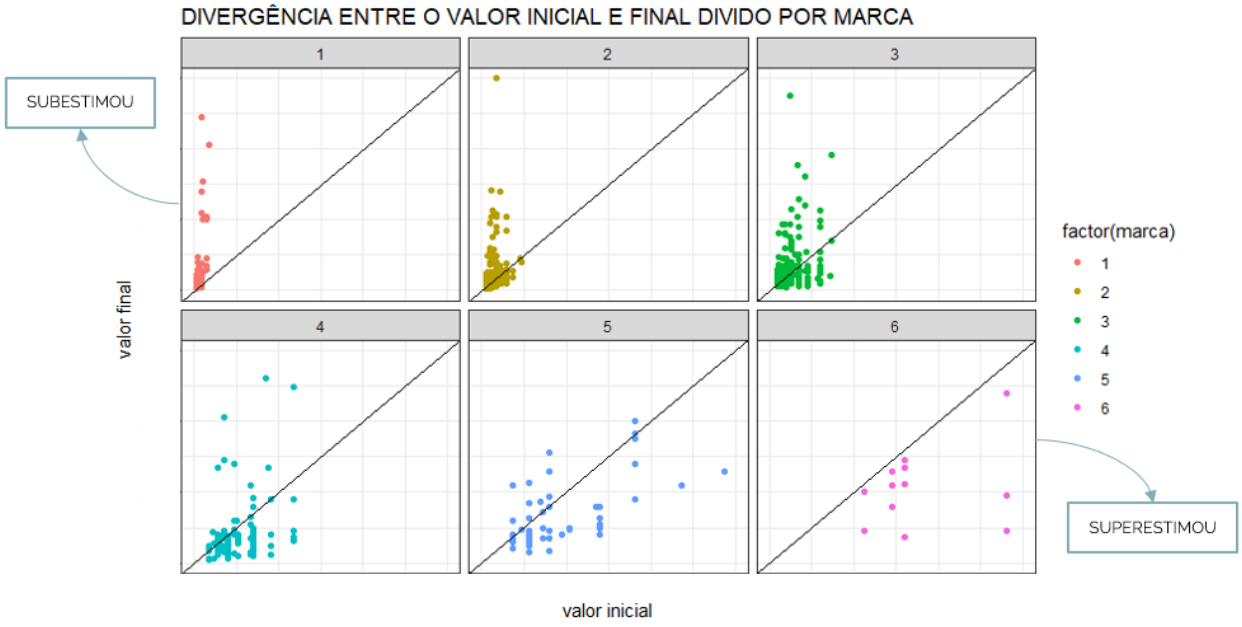
Dados

Como afirmado na seção de introdução em observações iniciais, algumas informações são sensíveis e é preferível que sejam mantidas em sigilo. Isso acontece com os dados fornecidos pela *Empresa*. Apesar disso, pode-se mencionar de forma genérica

como são os dados: a unidade amostral é: peças em forma de “vestido casual” que foi vendido e as variáveis são tópicos que descrevem a peça em questão e o preço de venda daquela peça, que será o alvo a ser predito por esse trabalho.

Análise descritiva

Imagem 01 – Gráficos que comparam preços



Fonte: Elaboração própria

Nota-se nesse resultado descritivo que há divergências entre o valor inicial, aquele que foi colocado para ser vendido em um primeiro momento, e o valor final, aquele que a peça foi vendida. Esses *gaps* sugerem que realmente pode haver um processo de manutenção de altos estoques, quando há a superestimação dos preços, e um processo de redução das margens de lucro por peça quando há subestimação. Visto isso, a próxima seção terá o objetivo de criar um modelo que minimize esses *gaps*, a fim de solucionar tais problemas.

Metodologia

A fim de obter um modelo que consiga prever o preço justo de dada peça, dadas suas características, optou-se por utilizar o Random Forest, na linguagem *python*. Esse modelo foi escolhido por duas principais razões: sua simplicidade e facilidade de utilizar; e ele é ótimo quando se tem muitas colunas e um número amostral não tão alto (como é

o caso). Além disso, ele é um dos modelos de *machine learning* mais completos. O Random Forest funciona, resumidamente, criando muitas árvores de decisões a partir das '*features'* que a base de dados possui. Cada árvore de decisão chega em um preço ideal.

Dessa forma, o resultado é a média entre valores que todas as árvores chegaram. Vale ressaltar que a quantidade de árvores é escolhida e varia de modelo para modelo. O Random Forest é dividido entre base de treino e teste; a base de treino serve para ele aprender as relações possíveis entre as '*features'* e o preço, e a de teste para testar a eficiência do modelo.

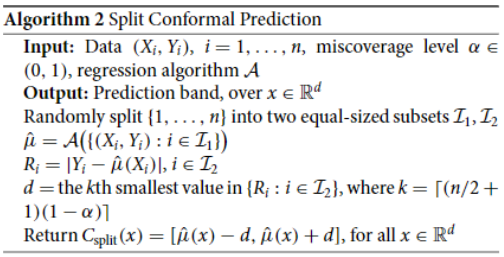
Foi utilizado o *Random Forest Regressor*, dentro da biblioteca *sklearn*, em *python*. Para isso, foi necessário transformar todas as *features* presentes na base em dummies, ou seja, em variáveis binárias. Após isso, utilizou-se a função *train_test_split* para separar aleatoriamente a base em treino e teste. Nesse trabalho, utilizou-se 20% como porcentagem para base de teste. Para que o intervalo de predição dos preços possa ser feito, é necessário utilizar o *out of bag score* quando rodar o modelo.

No caso desse projeto, percebeu-se uma maior eficiência quando as peças eram separadas entre serem de marcas de luxo e não luxo. Isso porque quando o modelo foi rodado com todas as peças conjuntamente, acredita-se que ocorreu um problema de *overfitting*.

Para solucionar esse problema e procurar gerar um resultado que possibilitasse maior adaptabilidade do preço predito às estratégias de marketing da empresa chegou-se à um modelo de predição com um intervalo de confiança sendo o resultado.

Para o intervalo de predição, utilizou-se um modelo do artigo *Distribution-Free Predictive Inference for Regression* (LEI et al., 2018), porém levemente modificado. Isto porque o *Random Forest* já tem o *out of bag score*, o que facilita a montagem do intervalo. Primeiramente, escolheu-se 90% como o nível de confiança. Então foi calculado o intervalo conforme descreve o *paper* (imagem 2).

Imagem 02 – Metodologia do intervalo de preços



Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$

Output: Prediction band, over $x \in \mathbb{R}^d$

Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$

$\hat{\mu} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$

$R_i = |Y_i - \hat{\mu}(X_i)|, i \in \mathcal{I}_2$

$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$

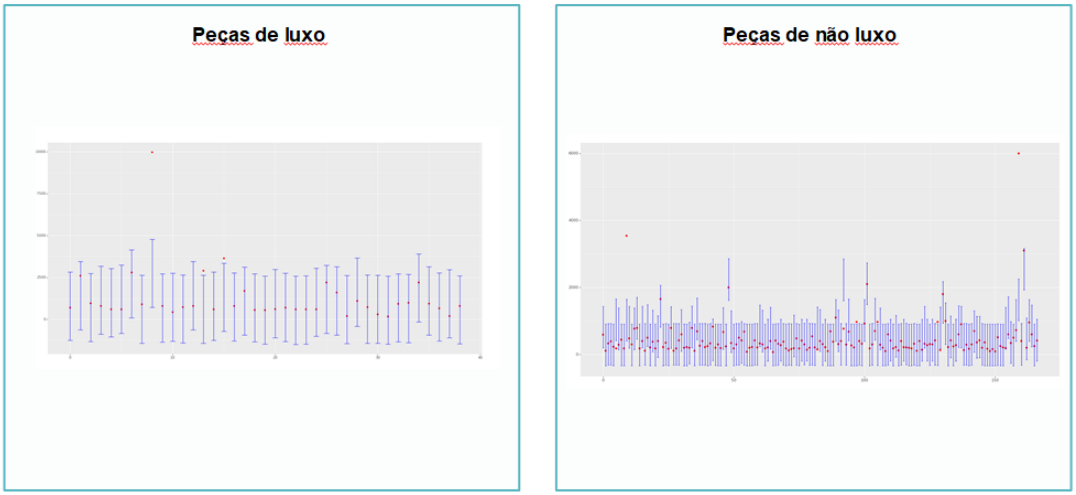
Return $C_{\text{split}}(x) = [\hat{\mu}(x) - d, \hat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

Fonte: Elaboração (LEI et al., 2018)

Conclusões e Resultados

O resultado do projeto foi um intervalo de predição, com 90% de confiança, para o preço ideal da peça. Ressalta-se que foram separadas peças de luxo e de não luxo, para melhor encaixe do modelo.

Imagem 03 – Intervalo preditivo dos preços a serem vendidos



Fonte: Elaboração própria

O resultado foi mais satisfatório para peças de não luxo. Isso pode ser explicado por um número amostral maior. Além disso, peças de luxo tendem a ter mais especificidades, essas que não necessariamente estão explícitas na base de dados.

O trabalho apresenta algumas limitações. Entre elas o baixo número amostral da base de dados e a falta de dados relevantes que não foram fornecidos pela empresa. Dados como informações sobre os consumidores seriam extremamente relevantes para a análise feita.

Referências bibliográficas

BRITTO, Elaine Mandotti de Oliveira et al. **A study on the attributes involved in the hedonic pricing of Brazilian clothing.** The International Review of Retail, Distribution and Consumer Research, v. 29, n. Ja 2019, p. 46-62, 2019Tradução. Disponível em: https://www.tandfonline.com/doi/pdf/10.1080/09593969.2018.1556178?needAccess=true. Acesso em: 21 jan. 2023.

DOLGUI, Alexandre. PROTH, Jean-Marie. **Pricing strategies and models.** Annual Reviews in Control, 2010, 34 (1), pp.101-110. 10.1016/j.arcontrol.2010.02.005. emse-00673983

MARQUES, Paulo C. Marques. **Confidence intervals for the random forest generalization error.** Insper Institute of Education and Research, São Paulo, 2022

LEI, Jing. G'SELL, Max. RINALDO, Alessandro. TIBSHIRANI, Ryan J. WASSERMAN, Larry. **Distribution-Free Predictive Inference for Regression.** Journal of the American Statistical Association, 2018. 113:523, 1094-1111, DOI: 10.1080/01621459.2017.1307116

Aplicação de redes neurais recorrentes para previsão de seres temporais em produtos web

Integrantes: André Corrêa, Thiago Hampl, Rafael Albuquerque, Felipe Catapano
Orientador: Neto Concon

Resumo

O projeto tem como objetivo principal combinar técnicas de Ciência de Dados à análise de movimentações de usuários em produtos Web. Para isso, o trabalho foi feito em conjunto com a empresa Northern, que disponibilizou uma base de dados inicialmente desestruturada, porém já com categorias de ações possíveis no site. Existiu uma demanda do orientador para utilizar como modelo uma Rede Neural Recorrente para prever a tendência das movimentações, principalmente pela sua boa capacidade de entender padrões em situações em que existe uma ordenação temporal. Para trabalhar com esse modelo, foi necessário inferir variáveis numéricas a partir da base abstraindo-a em grafos e, com isso, utilizou-se métricas de rede para calcular pesos diários para cada movimentação de usuário. Por fim, a Rede Neural foi construída e treinada, superando os resultados de comparação e sem caráter de aleatoriedade. O resultado final foi de 0.01959 de erro médio absoluto.

Introdução

Nesse semestre, a equipe de Modelagem Preditiva trabalhou junto à Northern Ventures, uma empresa de negócios digitais, que nos apresentou seu site de viagens, o *Roteirofy*. Esse site consiste em auxiliar seus usuários na busca de um roteiro ideal para suas viagens de modo rápido e personalizado. Nosso orientador e sócio dessa empresa, Neto Concon, nos procurou para que pudéssemos otimizar seu site a partir da base de dados que contêm as interações dos usuários com o site com diversas informações nesse quesito.

A proposta inicial foi integrar a manutenção do *Frontend* da aplicação com *Machine Learning*, a partir de uma rede neural. O objetivo principal seria ajudar a prever o comportamento dos usuários da aplicação da empresa, assim conseguindo antecipar possíveis aumentos ou diminuições do uso de funções contidas no site. As decisões

tomadas a partir daí seriam por parte dos desenvolvedores do site, ou seja, a própria Northern Venture.

Dados

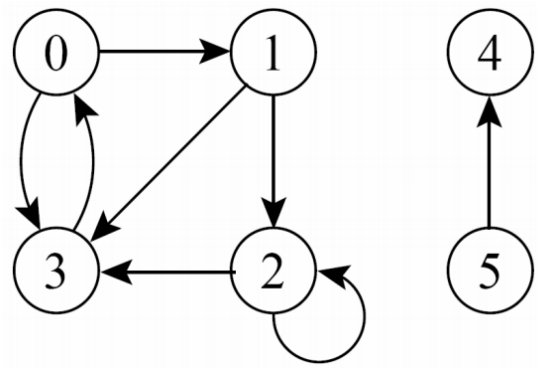
A Northern já possui uma equipe de *Analytics* que serve exatamente para estudar comportamentos de usuários anteriores e, a partir dos dados passados, tenta prever o futuro das *Features* do site. Dessa forma, este projeto não teve a necessidade de trabalhar na extração da base de dados, já que a empresa tem um sistema de gerar observações dos usuários. A empresa, portanto, disponibilizou um histórico de uso de *Features* do *Roteirofy*.

As Features são traduções de qualquer clique do usuário no site para um nome. Alguns exemplos são: *Itinerary Show*, *Itinerary Index, Add Schedule from List*, entre outras. Para o projeto, portanto, seria necessário calcular tendências de crescimento de cada uma dessas classes de interações e, para a empresa, isso seria traduzido em elementos do *Frontend* da aplicação.

A base inicial era baseada em interações completas de cada usuário, ou seja, cada linha continha os caminhos de um usuário do momento que ele entra no site até que ele pare de interagir. Dessa forma, era uma base de dados totalmente com foco nos usuários e não nas *Features*, algo que deveria ser buscado pela equipe. O desafio foi estruturar inputs e outputs numéricos para uma Rede Neural Recorrente, demanda da empresa, a partir de uma base em que cada linha continha uma lista de *Features* que foram acessadas em um momento por um usuário.

A solução encontrada pelo grupo, em conformidade com o coordenador do projeto, foi mapear as interações de usuário em um formato de Grafo direcionado, em que cada vértice é uma *Feature* e cada aresta representa que houve movimentação de uma *Feature* à outra. No entanto, como esse é um problema em que a ordem temporal é importante, foi mapeado um Grafo por dia. Dessa forma, seria possível utilizar métricas de Rede para as movimentações dos usuários e, com isso, calcular valores de input e output, por dia e por *Feature*. Além disso, analisar todos os caminhos em um dia generaliza as interações do site e, sendo assim, os valores não seriam separados por usuário como na base inicial.

Imagem 01 – Visualização de um Grafo Direcionado exemplo



Fonte: Elaboração própria

O valor quantitativo escolhido para ser utilizado de variável foi, simplesmente, a contagem de quantas vezes uma determinada aresta aconteceu em um dia, normalizada pela quantidade total de movimentações no dia. Com isso, a nova base de dados construída a partir daquela desestruturada disponibilizada pela empresa possuía informações sobre o peso, de 0 a 1, de cada movimentação em um dia. Como essas arestas representam um movimento de *Feature* para *Feature*, a nova base de dados não contabilizou as importâncias da Feature do site em si, mas sim o peso de cada caminho possível. Por exemplo, *Itinerary Index-Itinerary Show* representaria a aresta da primeira para a segunda *Feature* e tem um peso para cada dia daqueles disponibilizados para o projeto.

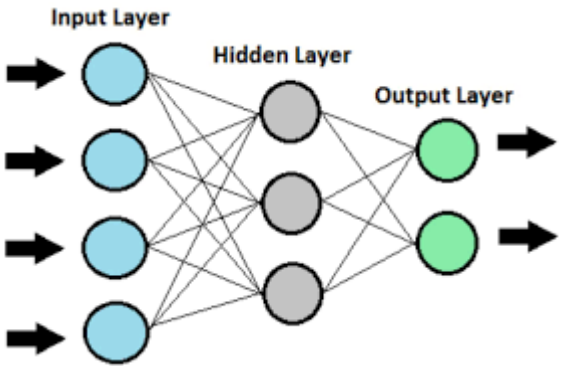
Com os pesos de todas as possibilidades de *features* mapeados por dia, a base se tornou algo maleável e, muito mais importante, treinável em uma Rede Neural. A ideia seria utilizar os pesos dos dias anteriores para calcular os pesos do próximo dia para cada movimentação. Intuitivamente, portanto, o *target* desta modelagem não é apenas um valor, mas sim um peso para cada uma das movimentações.

Modelagem

O modelo escolhido para a previsão dos pesos diários foi uma rede neural recorrente. A estrutura de rede neural foi escolhida pelo orientador visto que modelos mais simples não seriam capazes de abstrair relações entre os pesos diários e poderiam causar "*underfitting*" no *dataset*. A recorrência interna da rede foi determinada de modo a possibilitar a previsão de séries temporais.

Funcionamento básico de uma rede neural clássica:

Imagem 02 – Representação visual da estrutura de uma Rede Neural clássica

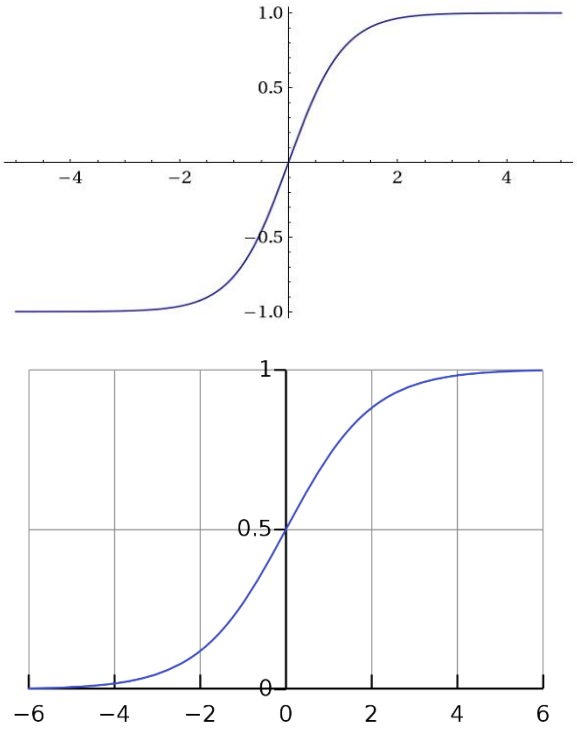


Fonte: Elaboração própria

De forma muito básica, o funcionamento de uma rede neural clássica consiste da transformação de inputs por neurônios e pesos. Cada neurônio representa uma função de ativação (na imagem 1 os neurônios são representados como círculos) e essas funções de ativação são comumente sigmóides ou tangentes hiperbólicas. As linhas representam as conexões entre os vários neurônios e cada uma possui um peso próprio, isto é, a saída de um determinado neurônio vai ser multiplicada por uma constante definida previamente ao entrar como input no neurônio a seguir. Essas constantes são definidas durante o treinamento da rede neural de modo a maximizar a performance da rede neural no *dataset* de validação.

Em suma, todas as entradas em um neurônio são multiplicadas pelos seus respectivos pesos somados e esse resultado é a entrada da função de ativação desse determinado neurônio. A saída dessa função de ativação é então usada como entrada para um neurônio na *layer* seguinte. Esse processo se repete depois para todos os neurônios interconectados até a *layer* de saída, na qual a saída dos neurônios é a própria predição da rede para o conjunto de dados que foram inicialmente utilizados como entrada.

Imagem 03 e 04 – Imagens de funções sigmóide e tangente hiperbólica



Fonte: Elaboração própria

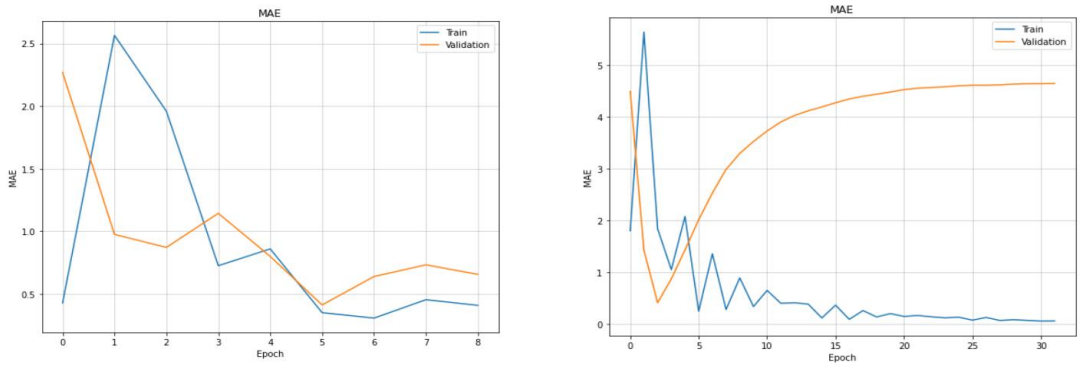
Em uma rede neural recorrente, entretanto, uma das *layers* internas tradicionais é substituída por uma "*layer*" recorrente, isto é, uma série de neurônios que não só são capazes de passar resultados para frente, mas, também, são capazes de guardar resultados anteriores, permitindo uma espécie de memória ajustável. Memória essa que torna possível a previsão de séries temporais, visto que, em séries temporais, intervalos de tempo seguintes são influenciados por intervalos de tempo anteriores.

Validação

O *dataset* fornecido possui 602 colunas, ou seja, 602 *features* para serem previstos diariamente, para apenas 111 dias de dados. Salta aos olhos o imenso problema que a dimensionalidade do banco de dados representa para qualquer modelo preditivo. Inicialmente alguns modelos com muitos *layers* e sem regularização foram treinados, para checar se mesmo com a dimensionalidade problemática do *dataset* era possível atingir acurácias razoáveis. Contudo, os modelos treinados apresentavam acurácias erráticas.

Os modelos foram treinados com os primeiros 70 dias e os últimos 41 dias foram usados para teste.

Imagem 05 e 06 – Distribuição do erro nas épocas de treinamento

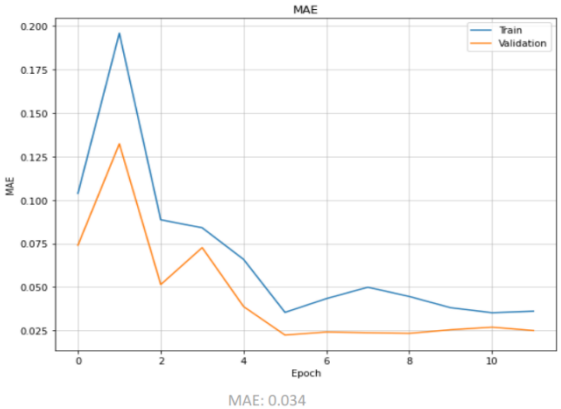


Fonte: Elaboração própria

As imagens acima ilustram o histórico de dois modelos treinados para o mesmo conjunto de dados e com os mesmos hiperparâmetros. A imagem 6 claramente ilustra um overfit do modelo. (MAE - *mean absolute error*).

De forma a reduzir a dimensionalidade do conjunto de dados, foram selecionadas as 5 *features* com maior atividade (maior número de valores não nulos) e foi possível observar resultados muito mais estáveis.

Imagem 07 – Distribuição do erro nas épocas de treinamento



Fonte: Elaboração própria

De forma a simular o funcionamento esperado para o modelo, o método escolhido para validação foi o "*backtracking*". Foi decidido que o modelo seria treinado diariamente utilizando os últimos 20 dias para a previsão do dia seguinte, sendo que a previsão do dia seguinte foi sempre comparada ao valor verdadeiro desse dia para avaliação da

performance do modelo. Deste modo, iterando por todo o *dataset* a partir do vigésimo dia o modelo foi avaliado em seu erro médio absoluto.

Imagem 08 – Overview do que é o resultado

Fonte: Elaboração própria

O resultado obtido para a performance deste modelo final foi um MAE médio de aproximadamente 0.01959. Nota-se uma melhora em relação ao erro da rede treinada com os primeiros 70 dias, ainda que o modelo esteja sendo avaliado em todo o conjunto de dados.

Conclusão

A partir da utilização e ajuste de uma rede neural recorrente, como sugerido pelo orientador, foi possível criar um modelo com alta acurácia, simulando condições reais de operação nas plataformas da empresa. Considerando isso, provou-se não apenas a coerência do *dataset* fornecido, gerado pelas ferramentas da Northern, mas também a sua compatibilidade com a estratégia de tabulação e previsão diária com "*backtracking*".

Assim, ao desconsiderar possíveis limitações no recorte de *features* estabelecido por conta da dimensionalidade do *dataset*, é seguro dizer que o objetivo central do trabalho foi atingido.

Referências Bibliográficas

***As referências bibliográficas desse projeto foram citadas de maneira diferente, já que ele não se utilizou diretamente das referências. Por isso, adotou-se a forma mais detalhada de se referir à elas.**

O livro "Hands On Machine Learning with Scikit-Learn, Keras & Tensorflow", de Aurélien Géron consegue passar por temas básicos da programação de aprendizagem de máquina até conceitos mais avançados, como Redes Neural Recorrentes, e por isso foi utilizado no projeto para consulta em todas as etapas do projeto, tanto em análises descritivas iniciais como para a modelagem e a validação em si.

Além desse livro, a publicação "Analyzing Social Networks" de BORGATTI, S. P.; EVERETT, M. G.; JOHNSON, J. C, foi utilizado para o melhor entendimento dos conceitos de métricas de rede utilizados no projeto.

Uso do estimador Arellano-Bond para determinar causalidade entre produtividade e *churn* (*turnover*) de firmas

Integrantes: Fábio Gerevini Canton, João Victor Czarnobay

Orientador: Mariana Orsini

Resumo

O trabalho busca compreender como a mudança na produtividade por trabalhador de firmas listadas na bolsa de valores impacta o turnover de seus funcionários no período estudado. Para o estudo foi utilizado o modelo para dados em painel Arellano-Bond, o qual não produziu evidências suficientes para corroborar nossa hipótese, sendo necessário dados a firma, e não setorizados, para conclusões mais claras.

Introdução

Neste semestre, o grupo de finanças buscou fazer uma análise que buscasse utilizar conceitos das finanças corporativas, e problemas da economia que tenham forte relevância para o mercado. Nosso objetivo foi estabelecer uma análise que busque observar uma relação entre o emprego, a produtividade e o *turnover* nas firmas.

Com isso, o trabalho buscou analisar, por meio de métodos econométricos e estatísticos, a relação entre o *turnover* e a produtividade por trabalhador em diferentes firmas de diferentes setores, e qual impacto que a produtividade e outros fatores relacionados, como o salário, tamanho das firmas, composição da força de trabalho, trazem no *turnover* das empresas.

Nossa hipótese inicial estipulava que a produtividade por trabalhador traz um impacto negativo no turnover, ou seja, o *turnover* tende a ser maior em setores com baixa produtividade por trabalhador, em empresas mais novas.

Base de Dados

O projeto contou com o uso de três bases de dados principais, que continham informações da força de trabalho, dos resultados financeiros de um número selecionado de empresas, e de dados cadastrais dos funcionários e das empresas.

A base utilizada pela equipe para estudar o comportamento das variáveis estudadas é a RAIS (Relação Anual de Informações Sociais), que tem como objetivo suprir as necessidades de controle das atividades trabalhistas do país e prover dados para estudos sobre a força de trabalho. O principal objetivo com o uso da base era a retirada de dados sobre o turnover das empresas listadas na bolsa de valores, dados esses como, número de funcionários, data de contratação, data de desligamento e motivo de desligamento, para que se pudesse estabelecer uma taxa sobre o turnover de funcionários de cada empresa.

Como medida do *turnover*, foi utilizada a razão entre o número total de desligamentos (funcionários que se demitiram ou foram demitidos) e o número total que compõe a força de trabalho das firmas.

A base de dados fornecida pelo Valor Pro foi utilizada para retirar a receita bruta de cada firma, que depois seria utilizada como *proxy* de produtividade para o estudo. A base secundária gerada a partir da junção da Raiz e receita bruta das firmas do Valor Pro, com o cruzamento realizado a partir do cruzamento dos CNPJs das duas bases, foi de extrema importância para a realização do projeto. A partir dela foi possível estimar a regressão para aferir causalidade entre as mesmas.

A variável principal utilizada para a estimação do modelo foi a razão entre a receita bruta e o número de funcionários das empresas, que foi usada como proxy para a produtividade por trabalhador das firmas trabalhadas.

Outra base de dados utilizada para a realização do projeto foi o CAGED (Cadastro Geral de Empregados e Desempregados), que tem como objetivo o registro de admissões e demissões via CLT. A base foi usada principalmente para comparação de resultados destoantes e utilização de valores faltas na base secundária gerada.

Modelo *Arellano Bond*

Arellano bond é um modelo dinâmico para dados em painel. Diferentemente de dados em painel estáticos, são incluídas defasagens da variável dependente como regressor,

contudo, incluir defasagens da variável dependente como regressor implica na violação da exogeneidade estrita, uma vez que as variáveis defasadas são prováveis de estarem correlacionadas com efeitos aleatórios e o erro.

No método Arellano Bond, a primeira diferença da equação da regressão é tirada para se eliminar efeitos individuais, assim, defasagens mais profundas da variável dependente são utilizadas como instrumentos para se tomar diferenças na variável dependente.

Nas técnicas tradicionais de dados em painel, adicionar defasagens mais profundas da variável dependente reduz o número de observações disponíveis. Por exemplo, se as observações estiverem disponíveis em T períodos de tempo, então, após a primeira diferenciação, apenas T-1 lags são utilizáveis. Então, se os K lags da variável dependente forem usados como instrumentos, apenas as observações T-K-1 são utilizáveis na regressão. Isso cria uma compensação: adicionar mais defasagens fornece mais instrumentos, mas reduz o tamanho da amostra. O método de Arellano-Bond contorna esse problema.

Para se melhorar a precisão do método Arellano-Bond, é usado o sistema GMM (*General Method of Movements).* Quando se tem uma alta variância do termo de efeito individual em observações individuais, ou quando a equação da regressão chegar perto de ser um passeio aleatório, o modelo Arellano-bond pode performar de maneira fraca para amostras finitas. Isso ocorre devido pois as variáveis dependentes defasadas são instrumentos fracos nessas circunstâncias.

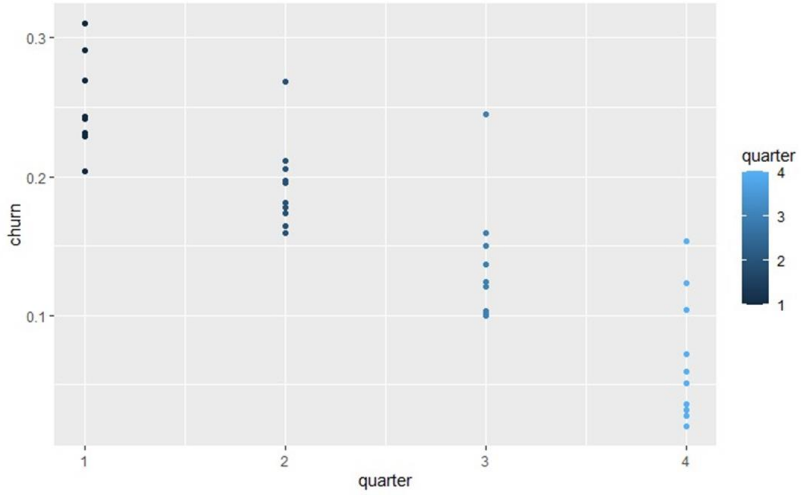
Blundell e Bond (1998) derivaram uma condição sob a qual é possível usar um conjunto adicional de condições de momento. Essas condições de momento adicionais podem ser usadas para melhorar o desempenho de pequenas amostras do estimador de Arellano-Bond.

Para a realização do modelo, foi utilizado a produtividade por trabalhador como variável dependente e o *turnover* das empresas calculado pelo autores como variável explicativa, e vice-versa.

Análise primária

Na análise descritiva, pode-se notar a existência de uma sazonalidade decrescente no turnover das empresas, sendo o turnover no primeiro trimestre o maior deles, decrescendo nos trimestres seguintes, até o quarto trimestre, que registrou seu menor valor nos anos de análise.

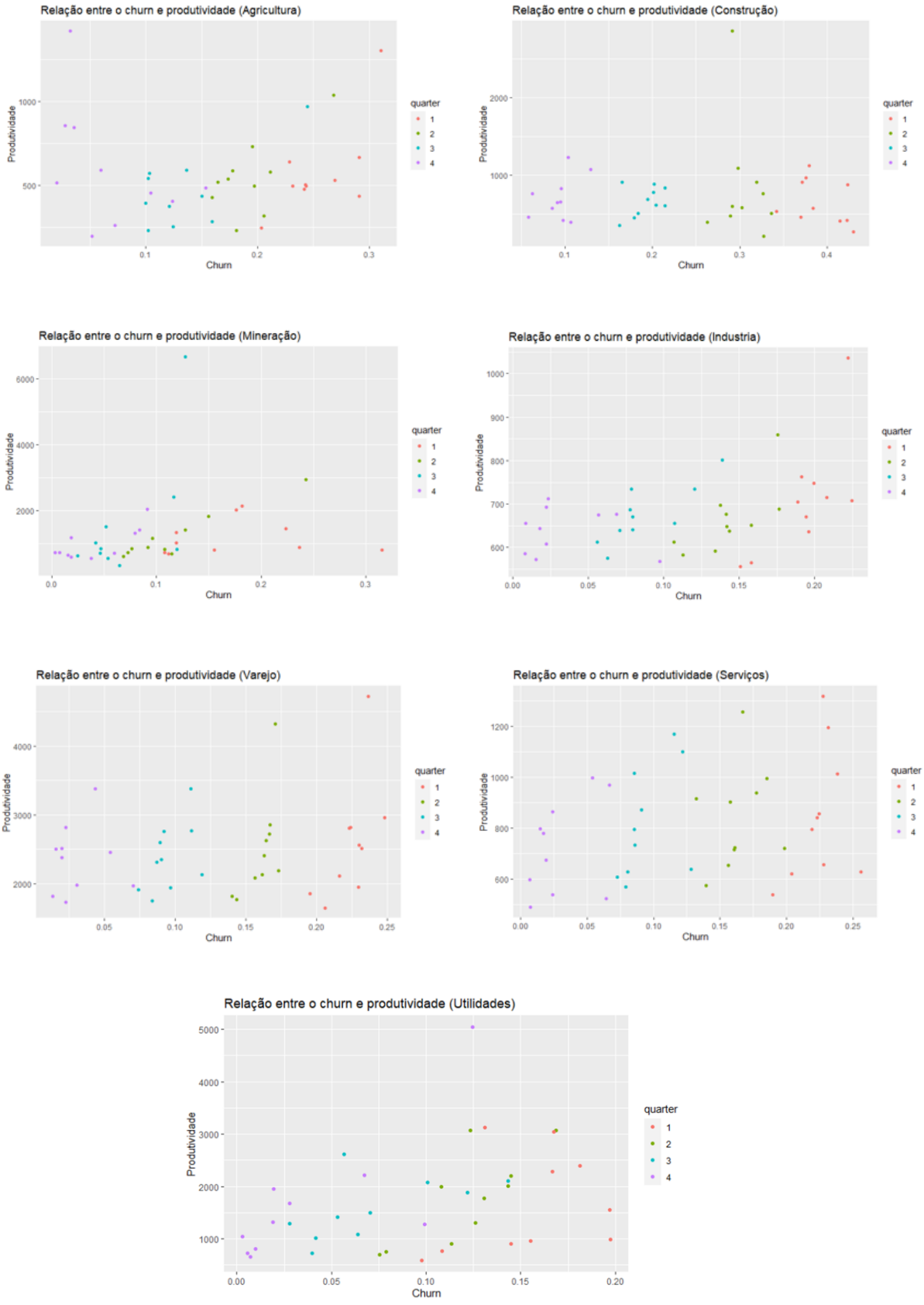
Imagem 01 – Turnover (churn) por trimestre



Fonte: Elaboração própria

Além da sazonalidade do turnover, as análises descritivas não evidenciaram nenhuma relação linear clara entre o turnover nas empresas e com o seus níveis de produtividade em todos os setores analisados. A sazonalidade observada anteriormente também se viu na análise descritiva setorial, ademais pode-se observar a maior concentração dos níveis de produtividade em alguns setores, como nos setores de mineração e construção.

Imagem 02 a 08 – Gráficos de correlação entre Turnover (churn) e produtividade por setor econômico



Fonte: Elaboração própria

Resultados e conclusões

Ao aplicar o modelo econométrico, a equipe realizou inicialmente uma regressão linear simples utilizando a produtividade por trabalhador como variável dependente e o *turnover* das empresas como variável explicativa, e os resultados indicam uma rejeição da hipótese inicial estipulada pelo grupo, ao observar que a produtividade não impacta o resultado do turnover das empresas.

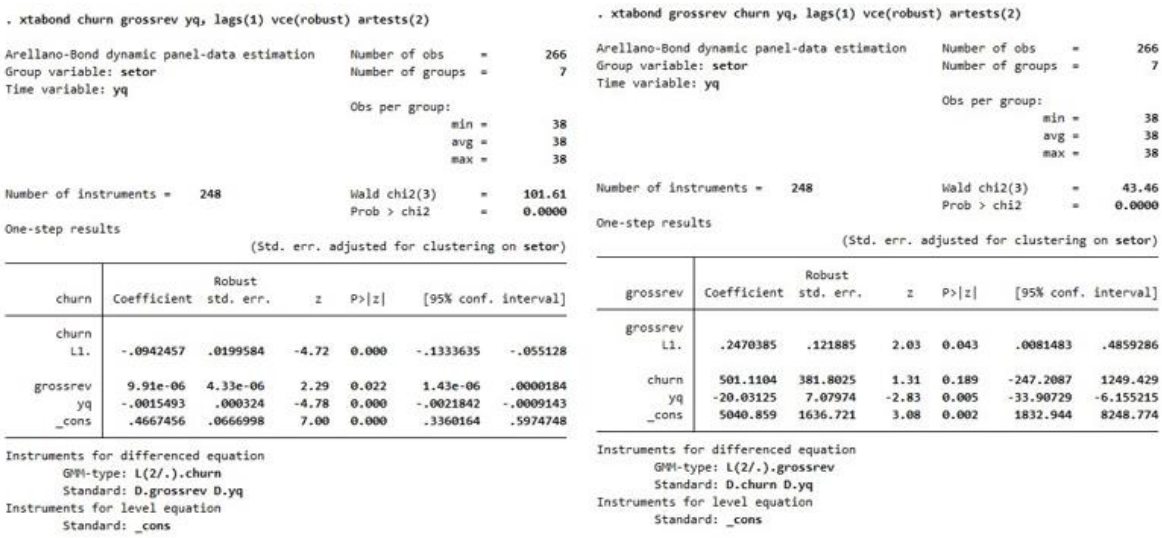
Imagem 09 – Resultado da primeira regressão

```
. reghdfe churn grossrev setor, noabsorb
(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =        280
Absorbing 1 HDFE group                            F(   2,    277) =      18.85
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.1198
                                                  Adj R-squared   =     0.1134
                                                  Within R-sq.    =     0.1198
                                                  Root MSE        =     0.0863

------------------------------------------------------------------------------
       churn | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    grossrev |   .0000128   6.42e-06     2.00   0.047     1.93e-07    .0000255
       setor |  -.0174159   .0028543    -6.10   0.000    -.0230347   -.0117971
       _cons |   .1941048   .0118078    16.44   0.000     .1708604    .2173492
------------------------------------------------------------------------------
```

Fonte: Elaboração própria

Com a aplicação do modelo *Aureliano Bond*, o grupo optou por realizar duas regressões, uma utilizando a produtividade por trabalhador como variável dependente e o *turnover* das empresas como variável explicativa, e a outra fazendo a análise inversa, utilizando o *turnover* das empresas como variável dependente e o produtividade por trabalhador como variável explicativa. Os resultados indicam uma conclusão semelhante à regressão anterior, não indicando uma conclusão da hipótese inicial, mas sim uma rejeição da mesma.

Imagem 10 – Resultado final das regressões



```
. xtabond churn grossrev yq, lags(1) vce(robust) artests(2)

Arellano-Bond dynamic panel-data estimation     Number of obs     =       266
Group variable: setor                           Number of groups  =         7
Time variable: yq
                                                Obs per group:
                                                              min =        38
                                                              avg =        38
                                                              max =        38

Number of instruments =    248                  Wald chi2(3)      =    101.61
                                                Prob > chi2       =    0.0000
One-step results
                                        (Std. err. adjusted for clustering on setor)
------------------------------------------------------------------------------
             |               Robust
       churn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       churn |
         L1. |  -.0942457   .0199584    -4.72   0.000    -.1333635   -.055128
             |
    grossrev |   9.91e-06   4.33e-06     2.29   0.022     1.43e-06    .0000184
          yq |  -.0015493    .000324    -4.78   0.000    -.0021842   -.0009143
       _cons |   .4667456   .0666998     7.00   0.000     .3360164    .5974748
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).churn
        Standard: D.grossrev D.yq
Instruments for level equation
        Standard: _cons
```

```
. xtabond grossrev churn yq, lags(1) vce(robust) artests(2)

Arellano-Bond dynamic panel-data estimation     Number of obs     =       266
Group variable: setor                           Number of groups  =         7
Time variable: yq
                                                Obs per group:
                                                              min =        38
                                                              avg =        38
                                                              max =        38

Number of instruments =    248                  Wald chi2(3)      =     43.46
                                                Prob > chi2       =    0.0000
One-step results
                                        (Std. err. adjusted for clustering on setor)
------------------------------------------------------------------------------
             |               Robust
   grossrev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   grossrev |
         L1. |   .2470385   .121885     2.03   0.043     .0081483    .4859286
             |
       churn |   501.1104   381.8025     1.31   0.189    -247.2087    1249.429
          yq |  -20.03125   7.07974    -2.83   0.005    -33.90729   -6.155215
       _cons |   5040.859   1636.721     3.08   0.002     1832.944    8248.774
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).grossrev
        Standard: D.churn D.yq
Instruments for level equation
        Standard: _cons
```

Fonte: Elaboração própria

Pelos resultados apresentados, não foi observado um resultado que confirmasse a nossa hipótese inicial. Na análise descritiva, não foi observada uma relação clara entre a produtividade e o *turnover*, que apresenta uma sazonalidade entre os trimestres que os afetam de forma decrescente.

Nossos modelos indicam que não existe uma relação entre a produtividade e o turnover das firmas, e portanto não foi apresentada uma conclusão clara, sendo necessárias mais observações a nível firma, sendo esta a principal limitação de nossa análise, o baixo número de observações gerado pela análise setorial.

Referências Bibliográficas

BASE DOS DADOS. **CAGED**. Disponível em: https://basedosdados.org/dataset/br-me-caged?bdm_table=microdados_antigos. Acesso em: 15 jan. 2023.

DIEPPE, Alistair. **Global Productivity**: Trends, Drivers and Policies. 1. ed. [S.l.: s.n.], 2021. p. 1-464.

ORSINI, Mariana. The Cleansing and Sullying Effects of Recessions in Heterogeneous Sectors. ***Instituto de Ensino e Pesquisa (Insper)**, Insper, v. 1, n. 1, p. 2, set./2022.