BOLETIM DE PROJETOS

2022.1

INSPERDATA

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 2022.1 foram confeccionados 6 trabalhos, abarcando as grandes áreas da Microeconomia, Macroeconomia, Marketing, Direito e Modelagem Preditiva.

Embora com escopos e objetivos largamente distintos, todos os projetos desenvolvidos compartilham um denominador comum: a utilização de ciência dos dados como base para encontrar evidências empíricas acerca dos temas estudados. Nesse sentido, todos os resultados aqui expostos são extraídos das pesquisas feitas pelos grupos. Agradecemos a todos os professores e parceiros de mercado que contribuíram durante o semestre; agradecemos especialmente também aos nossos orientadores Pedro Picchetti, Victor Hugo Alexandrino, Ademar Concon-Neto, Luciana L. Yeung, Matheus Damasceno e Daniel Ferreira os quais, sem suas contribuições, nenhum dos projetos aqui apresentados seriam possíveis.

Desejamos a todos uma boa leitura!

Contatos:
insperdata@gmail.com
@insperdata
@DataInsper

Sumário

Desigualdade de gênero em departamentos de economia e correlatos no Insper

Integrantes: David Gun, Guilhermo Pastore, João Czarnobay

Orientador: Pedro Picchetti

Resumo

O estudo realizado visou analisar a existência de desigualdades de gênero presentes em departamentos de Economia e correlatos dentro do Insper, examinando a magnitude dela, suas fontes e possíveis explicações. Para isso dispusemos de uma base fornecida pelo instituto que continha dados anonimizados de alunos dos cursos de Economia, Administração, Direito e das três Engenharias (Computação, Mecatrônica, Mecânica). Dentre esses dados destacaram-se para o estudo o CR (coeficiente de desempenho do aluno), Ranking e o Período em que o aluno se encontra.

A partir da análise dos dados nota-se o desempenho superior das mulheres em todos os cursos, apresentando mediana e média mais elevadas que seus pares do gênero masculino. Outro dado relevante é a distribuição dos gêneros nos cursos, sempre no intervalo de 20-30% para mulheres e o restante homens, à despeito do curso de Direito, que apresenta uma distribuição em torno de 60% mulheres e 40% homens, sendo o *outlier* dentro da instituição.

1. Introdução e breve revisão bibliográfica

Não é surpresa para ninguém a baixa presença de mulheres em níveis relativos aos homens dentro do Insper. Como informado anteriormente a relação de mulheres dentro das turmas do instituto gira ao redor de 30%. A literatura econômica traça essa baixa razão a fatores como a recente entrada das mulheres no mercado de trabalho, estigmas sociais e desinteresse geral desse grupo por matérias ligadas ao STEM *(Science, technology, engineering and mathematics)*. Este último fator, por sua vez, apresenta literatura mais relacionada à psicologia, trazendo como explicações a maior inclinação das estudantes do gênero feminino por profissões que envolvem o trabalho diretamente com o trato de pessoas, como enfermagem, medicina, psicologia, etc. Por outro lado, notam que os homens são mais inclinados ao estudo de objetos, tendendo a profissões como engenharia, física, química, etc.

Sendo assim, já era esperado a presença de mais homens do que mulheres nos cursos mais ligados ao STEM (engenharias), enquanto que aqueles mais ligados diretamente às Ciências Humanas, como Direito, deveriam apresentar relação inversa. A partir disso, pretendeu-se analisar o desempenho das mulheres para verificar se essa maior inclinação dos homens por disciplinas mais ligadas ao ramo das exatas faria com que esse grupo apresentasse notas

mais elevadas que suas colegas do gênero feminino, que apresentariam menor "aptidão".

Ademais, foi aventada a possibilidade do fato de que a existência de estigmas sociais perante as mulheres e seu ingresso no Ensino Superior faria com que grande parte delas sentisse incapaz de perseguir estudos para STEM. Portanto, ocorreria *ex-ante* um viés de seleção, realizando uma "triagem" das melhores candidatas antes mesmo da realização do vestibular. Com isso, haveria evidências favoráveis para, na média, as candidatas mulheres serem melhores que os candidatos homens. Sendo assim, a hipótese de que a menor inclinação de mulheres por matérias de STEM deveria prejudicar seu desempenho não deveria se mostrar válida, o que foi possível ser concluído através de nossa análise exploratória.

2. Dados

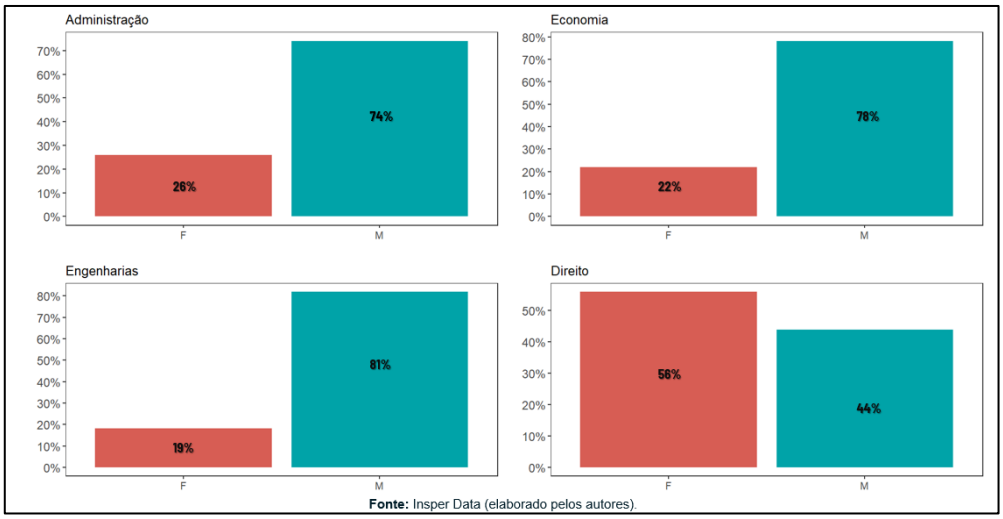
Para o trabalho foi utilizada uma base providenciada pelo Insper. A base contém dados anonimizados dos alunos do Insper até a data de 13/04/2022, com mais de 28 mil observações e 10 variáveis. As variáveis são: Sexo, Curso, Ano, Período, CR, Ranking, Status, Descrição e Data de Divulgação. Vale ressaltar que as variáveis utilizadas para o estudo foram Sexo, Curso, Período, Ranking, CR e Data divulgação. Foram retiradas da análise o Ano, devido ao fato de que apresentava somente o ano de 2021, sendo este o de divulgação dos dados, assim como Status, pois não identificava exatamente a real situação de matrícula do aluno no semestre em que se encontrava. Por fim, foi excluída a Descrição por ser redundante e a Data divulgação.

Para todas as análises do projeto utilizou-se o *software* R. Para a construção de grande parte das análises descritivas, assim como organização da base optou-se pelo uso da biblioteca *Tidyverse*. Para a decomposição do *gap* entre os gêneros entre fatores exógenos e/ou não observáveis com a parcela que é explicada pela diferença entre grupos utilizou-se a biblioteca *Oaxaca*, possibilitando a aplicação do método *Oaxaca-Blinder*.

3. Análise descritiva

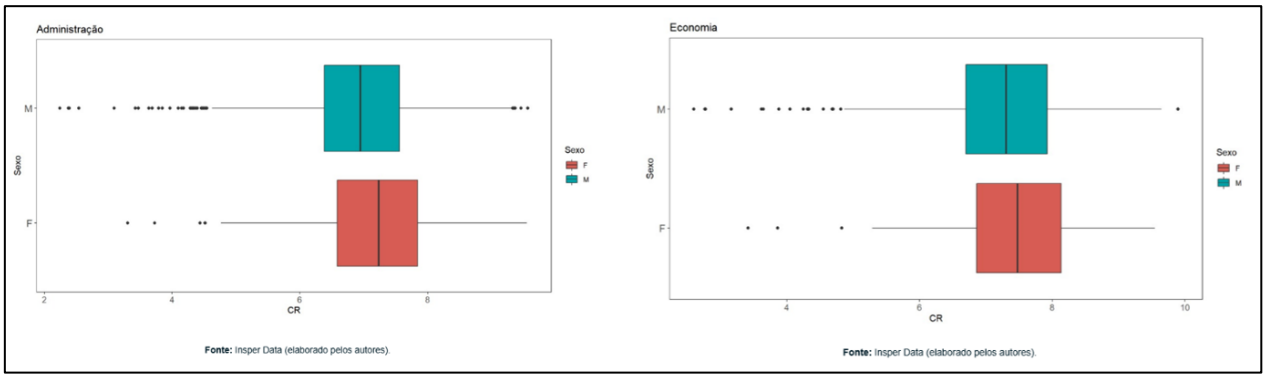
Foi realizada uma análise descritiva para o melhor entendimento das disposições de alunos dos dois gêneros em cada curso, como é possível ver na Figura 1 abaixo:

Figura 1 – Disposição dos gêneros nos cursos dentro do Insper



É possível notar que nos cursos de Economia e Administração a diferença entre as distribuições é muito semelhante, possuindo uma relação um pouco maior de mulheres para o primeiro em relação ao último. Adicionalmente, é interessante notar como ocorre uma inversão nessa relação para o curso de Direito, com quase 60% da turma sendo composta por mulheres. Por outro lado, dentre os cursos analisados, no agregado, os cursos de Engenharia apresentam a menor relação de alunas para alunos.

Figura 2 – *Boxplot* do CR para Administração e Economia



Analisando interturmas, é possível notar como a mediana das mulheres é mais elevada tanto para o curso de Administração como Economia, assim como apresenta uma menor variância entre os respetivos CR's. Por outro lado, os homens em ambos os cursos são os que apresentam mais *outliers*, tanto para observações mais altas de CR, assim como observações mais baixas do coeficiente.

Figura 3 – Medidas resumo do CR para Administração e Economia

- Administração

Sexo	Média	Mediana	Desvio	Mínimo	Máximo
Mulher	7,1932	7,2350	0,9691	3,3028	9,5557
Homem	6,9146	6,9470	1,0473	2,2403	9,5719

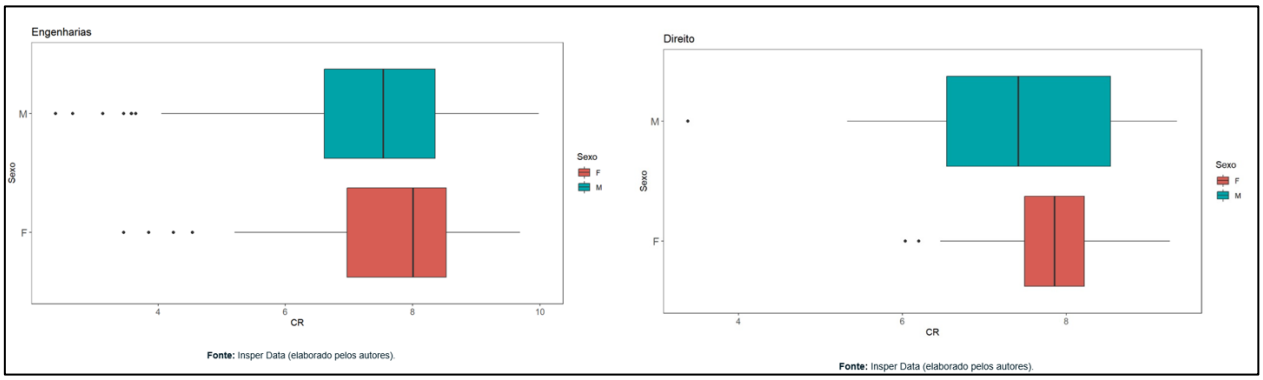
- Ciências Econômicas

Sexo	Média	Mediana	Desvio	Mínimo	Máximo
Mulher	7,4130	7,4733	0,9609	3,4148	9,5441
Homem	7,2814	7,2987	1,0647	2,5979	9,8949

Fonte: Insper Data (elaborado pelos autores).

Isso é bem notável na Figura 2 acima, em que é possível notar que a mediana para ambos os cursos é maior para as mulheres, assim como a média. Contudo, por outro lado, nota-se como o desvio padrão é menor para as mulheres em ambos os cursos, o que indica um grupo mais homogêneo, o que corrobora a hipótese do viés de seleção.

Figura 4 – *Boxplot* do CR para Engenharias e Direito



Fonte: Insper Data (elaborado pelos autores).

Para as Engenharias o padrão se mantém, como é possível ver na Figura 5 acima. A mediana das mulheres é mais elevada que dos homens, enquanto que os *outliers* em níveis de CR mais baixos são observados com maior frequência no grupo dos homens. Por outro lado, para o curso de Direito é notável a grande variância no conjunto de notas dos alunos e a grande assimetria entre os outliers de homens *versus* mulheres, visto que aqueles dos homens são bem abaixo das mulheres.

Figura 5 – Medidas resumo do CR para Engenharias e Direito

- Engenharias

Sexo	Média	Mediana	Desvio	Mínimo	Máximo
Mulher	7,7121	8,0020	1,2699	3,4586	9,6833
Homem	7,4098	7,5349	1,3279	2,3917	9,9812

- Direito

Sexo	Média	Mediana	Desvio	Mínimo	Máximo
Mulher	7,7773	7,8557	0,7930	6,0341	9,2624
Homem	7,3304	7,4115	1,4395	3,3844	9,3501

Fonte: Insper Data (elaborado pelos autores).

Novamente, isso é bem notável na Figura 6 acima, em que é possível notar que a mediana para ambos os cursos é maior para as mulheres, assim como a média. Assim como para os outros cursos, o desvio padrão das mulheres é menor para ambos. Contudo, é muito interessante notar que o desvio padrão para o curso de Direito apresenta uma grande disparidade entre os gêneros, com um desvio notável em relação aos outros cursos. Ademais, o mínimo de CR entre os gêneros é muito dispare, com as mulheres apresentando um CR mínimo quase o dobro daquele dos homens.

Figura 6 – Distribuição dos 10% melhores CRs

Curso	Homens	Mulheres
Administração	73,40%	26,60%
Ciência Econômicas	77,76%	22,24%
Engenharias	81,24%	18,76%
Direito	41,67%	58,33%

Fonte: Insper Data (elaborado pelos autores).

Por fim, para entender melhor como é a distribuição dentre os melhores CR's, optou-se por analisar os 10% melhores de cada curso. É notável a presença de grande parcela de homens dentre os maiores CR's, sempre na faixa de 70-80%. Todavia, como já deve ser sabido a este ponto do boletim, Direito se destaca. Quase 60% do primeiro decil de maiores CR's está concentrado no grupo das mulheres.

4. Análise econométrica

Para mapear os principais resultados de forma clara, foi desenvolvido uma regressão linear múltipla com as principais variáveis da base de dados. O modelo é dado por:

$$log(CR_i) = \beta_0 + \beta_1 * período_i + \beta_2 * sexo_i + \sum_{k=3}^{7} \beta_k * d_{curso_i} + \varepsilon_i$$

A variável período contém o período cursado pelo(a) aluno(a) no segundo semestre de 2021, a variável sexo é uma dummy que designa o sexo do(a) aluno(a) em questão (1 para mulher e 0 para homem), e por fim, a variável d_curso é uma dummy que contém o curso que aquele(a) aluno(a) está fazendo na graduação. Estimando o modelo de regressão obtivemos os seguintes resultados:

Figura 7 – Tabela com os resultados do modelo de regressão estimado

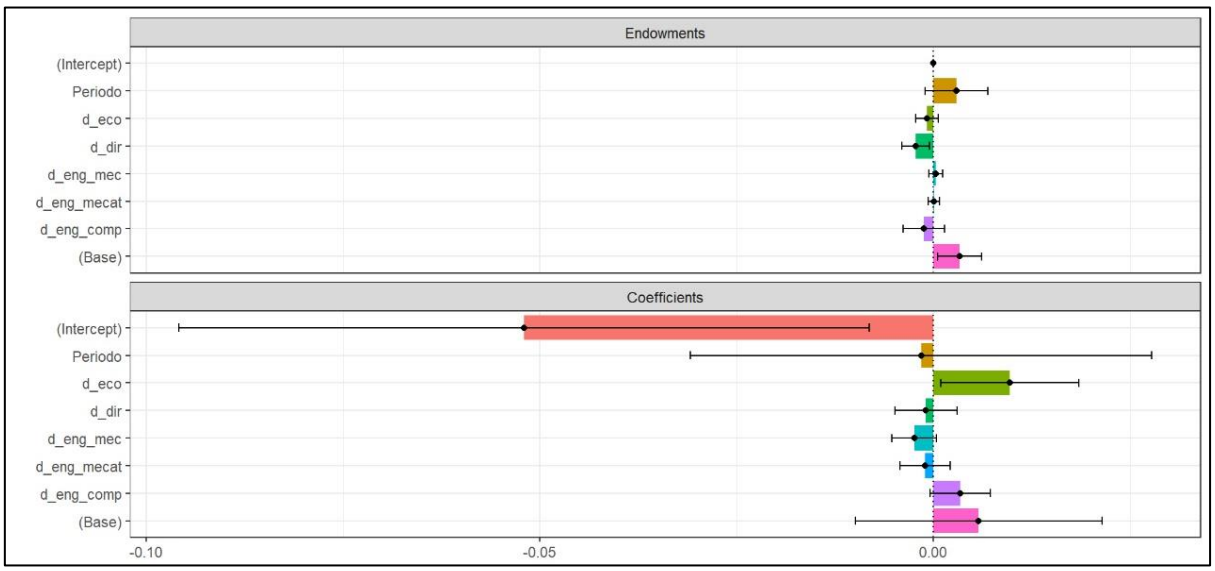
Table 1:

	Dependent variable:
	log(CR)
d_gen1	0.038***
	(0.007)
Periodo	0.017***
	(0.001)
factor(Curso)CIÊNCIAS ECONÔMICAS	0.045***
	(0.007)
factor(Curso)DIREITO	0.111***
	(0.024)
factor(Curso)ENGENHARIA DE COMPUTAÇÃO	0.063***
	(0.011)
factor(Curso)ENGENHARIA MECÂNICA	0.033**
	(0.015)
factor(Curso)ENGENHARIA MECATRÔNICA	0.036***
	(0.014)
Constant	1.846***
	(0.008)
Observations	2,819
R^2	0.091
Adjusted R^2	0.089
Residual Std. Error	0.166 (df = 2811)
F Statistic	40.209*** (df = 7; 2811)
Note:	*p<0.1; **p<0.05; ***p<0.01

Com o modelo devidamente estimado foi possível observar que todos os regressores apresentaram significância estatística, o que é um resultado preliminar importante ao passo que sinalizar que as variáveis escolhidas são relevantes para explicar o logaritmo natural do CR dos alunos. Além disso, é interessante observar para a estimativa de 0,038 para o parâmetro β_2, a qual na prática significa que o valor esperado do CR das mulheres é 3,8% em comparação aos homens, o que converge com os resultados observados na análise descritiva.

Além da regressão múltipla, também foi realizada aplicação da decomposição de Oaxaca-Blinder, método econométrico o qual decompõe gaps entre grupos em uma parcela que é explicada pela diferença entre grupos e uma parcela que é explicada por fatores exógenos e/ou não observáveis. Realizando a decomposição, podemos observar os seguintes resultados:

Figura 8 – Resultado da decomposição de Oaxaca-Blinder



Fonte: Insper Data (elaborado pelos autores)

A partir do gráfico é possível, observar que quase todos os gaps nas variáveis apresentam intervalo de confiança com o número zero contido, ou seja, não apresentam relevância estatística, porém, esse comportamento não foi observado na variável dummy que designa o curso de economia, ou seja, podemos observar que na graduação de economia há um gap de desempenho em que as mulheres vão sistematicamente melhores que os homens, resultado o qual mais uma vez converge com os insights estabelecidos tanto na análise descritiva quanto na regressão.

5. Conclusões e limitações

Durante toda a análise, ficou claro que grande parte das limitações apresentadas pelo estudo se concentram no conteúdo da base de dados, como o fato de a base utilizada ser *cross sectional*, o que dificulta a extração de efeitos causais e inviabiliza uma análise temporal dos dados e o fato de existir poucas variáveis, além disso, dado que a base de dados contém poucas variáveis, acreditamos que o modelo econométrico possa ter problemas de endogeneidade. Mas a limitação mais relevante pode ser o fato da provável existência de um viés de seleção dentro do processo seletivo de entrada no Insper, uma vez que a literatura nos traz evidências que cursos de economia, administração e engenharia tendem a receber mulheres com desempenhos muito mais elevados, o que pode levar a um viés de seleção na amostra coletada.

Apesar das limitações, pode-se concluir com as observações que as estudantes mulheres tem um desempenho mais elevado que os estudantes homens no Insper em todos os cursos oferecidos pela instituição, evidenciados pela análise descritiva dos CRs e pela análise econométrica, apesar de a grande maioria dos cursos ter uma expressiva maioria de homens, coma exceção do curso de direito, que possui um maior número de estudantes mulheres, que também apresentaram um desempenho muito mais expressivo que os estudantes homens.

6. Referências bibliográficas

Rocha et Al. Economistas, *Brazilian Women in Economics*. Disponível em: https://forumseguranca.org.br/wp-content/uploads/2017/03/Relatorio-final_CAF.pdf

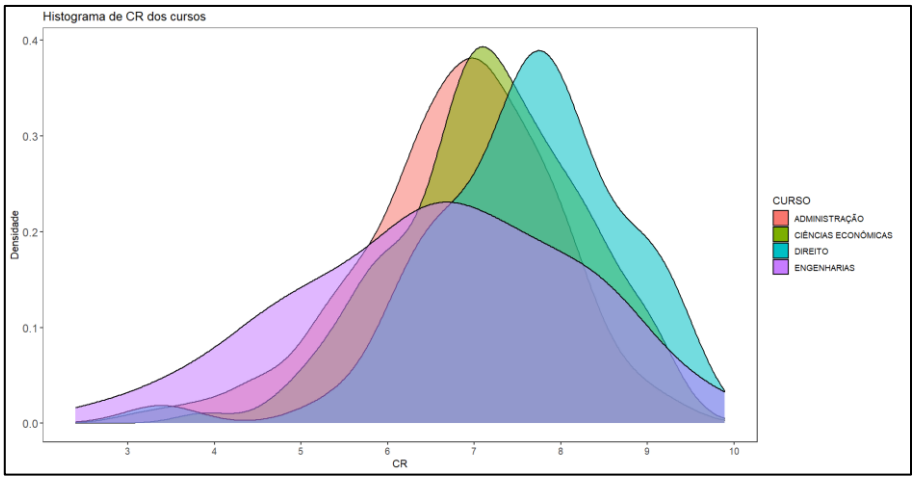
An Application to the Gender Gap of Earnings. Disponível em: https://www.econometrics-with-r.org/3-6-aattggoe.html

B. Jann (2008). The Blinder–Oaxaca decomposition for linear regression models. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/1536867X0800800401

M. Hlavac (2022). oaxaca: Blinder-Oaxaca Decomposition in R. Disponível em: https://cran.r-project.org/web/packages/oaxaca/vignettes/oaxaca.pdf

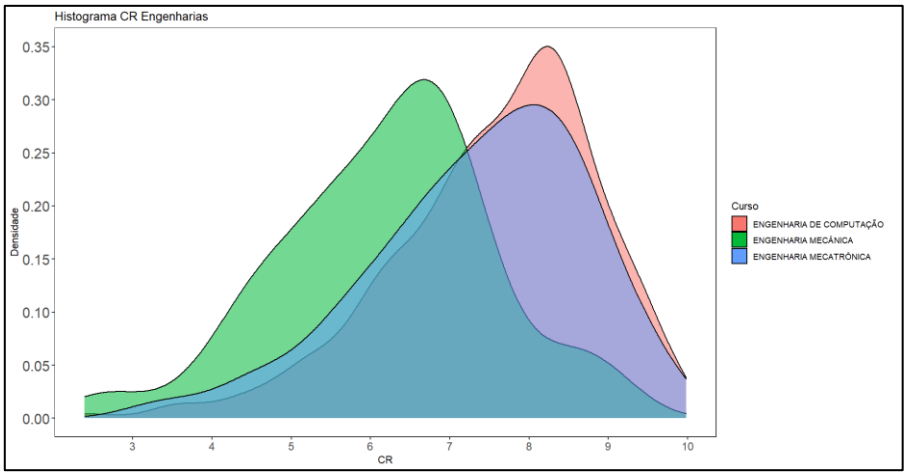
7. Anexos

Anexo 1 – Histograma de CR dos cursos



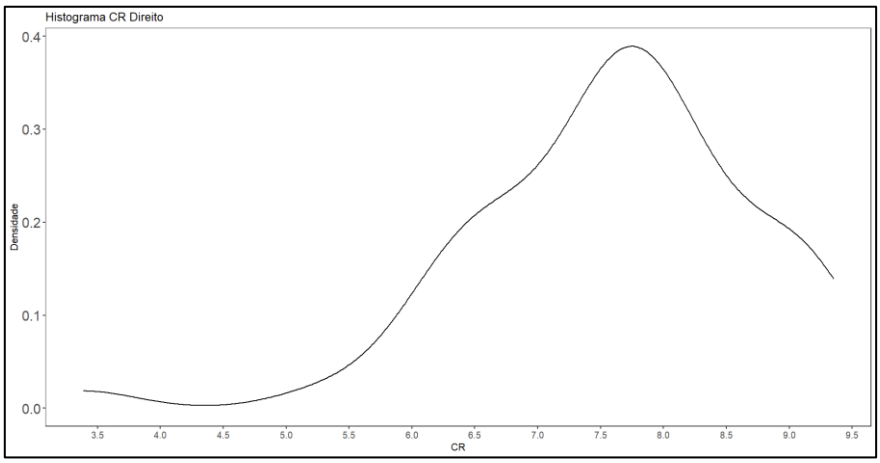
Fonte: Insper Data (elaborado pelos autores)

Anexo 2 – Histograma de CR engenharias



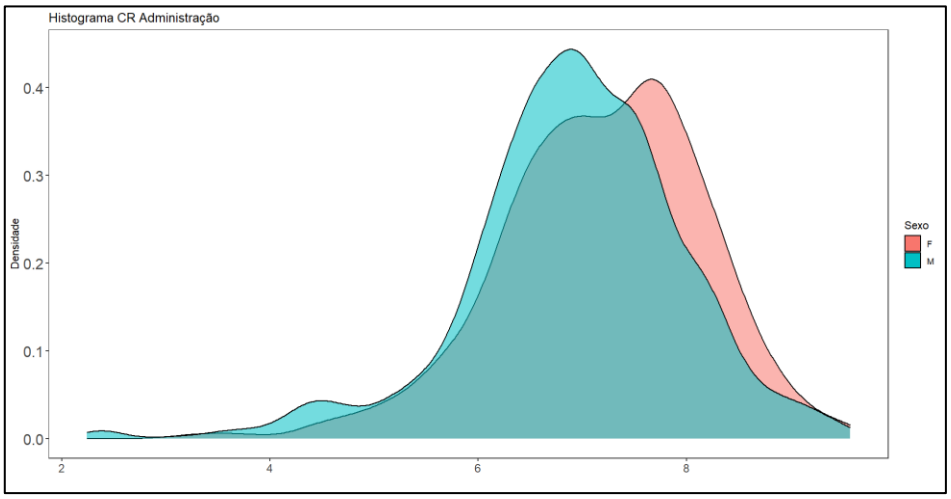
Fonte: Insper Data (elaborado pelos autores)

Anexo 3 – Histograma de CR direito



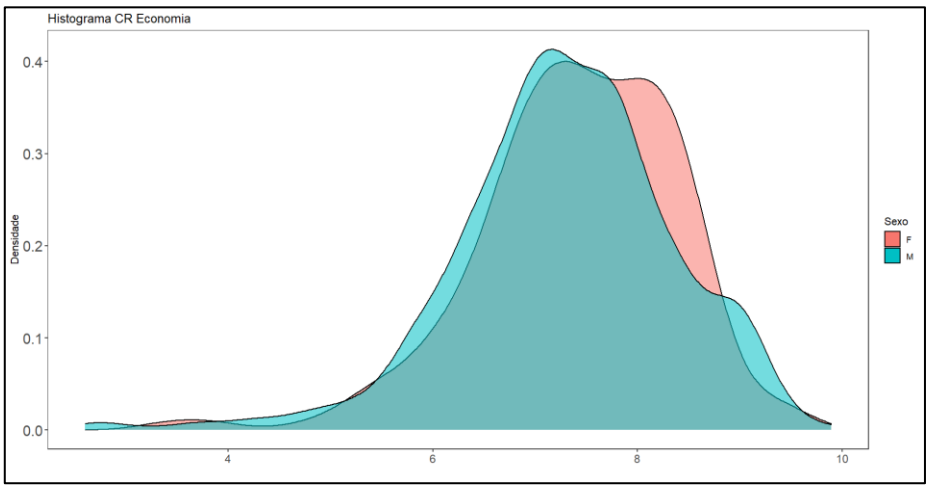
Fonte: Insper Data (elaborado pelos autores)

Anexo 4 – Histograma de CR administração



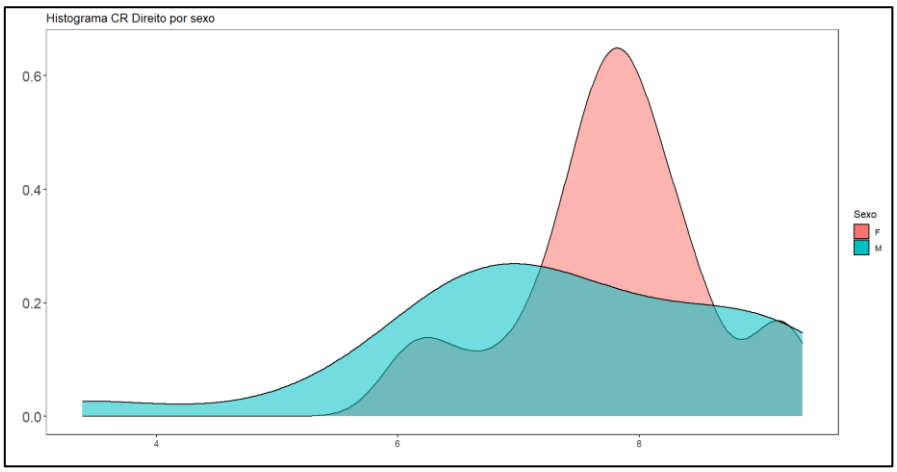
Fonte: Insper Data (elaborado pelos autores)

Anexo 5 – Histograma de CR economia



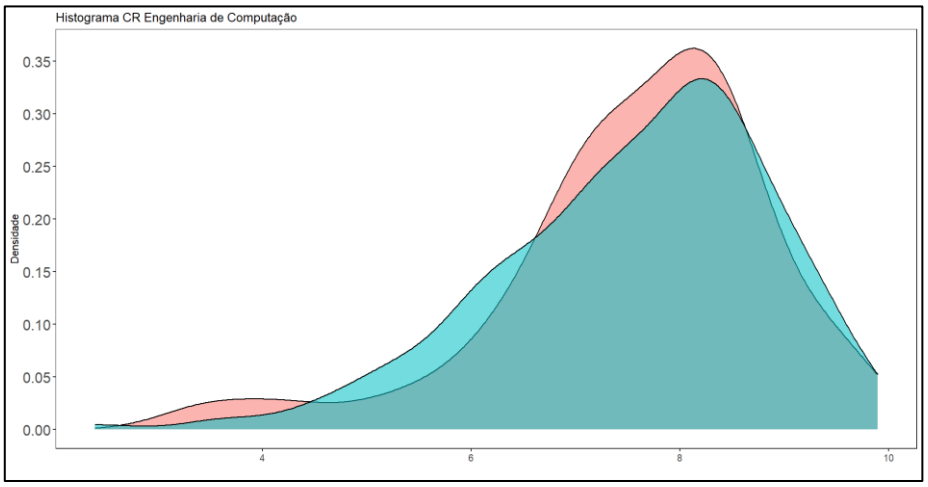
Fonte: Insper Data (elaborado pelos autores)

Anexo 6 – Histograma de CR direito por sexo



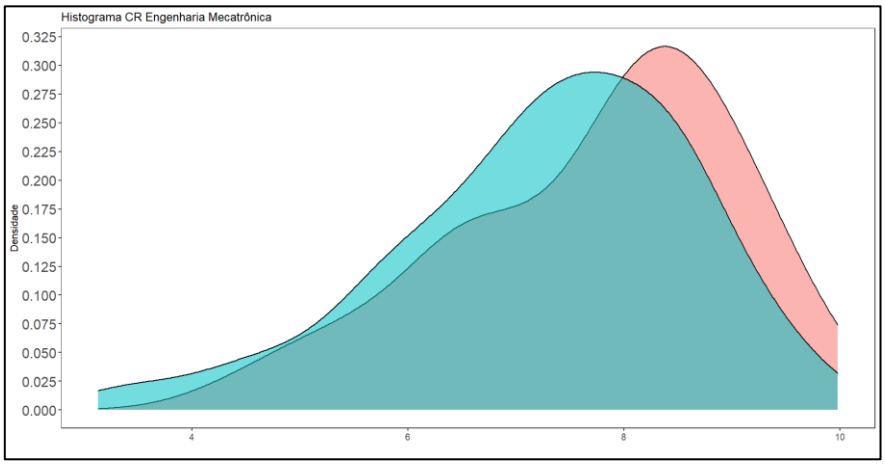
Fonte: Insper Data (elaborado pelos autores)

Anexo 7 – Histograma de CR engenharia de computação



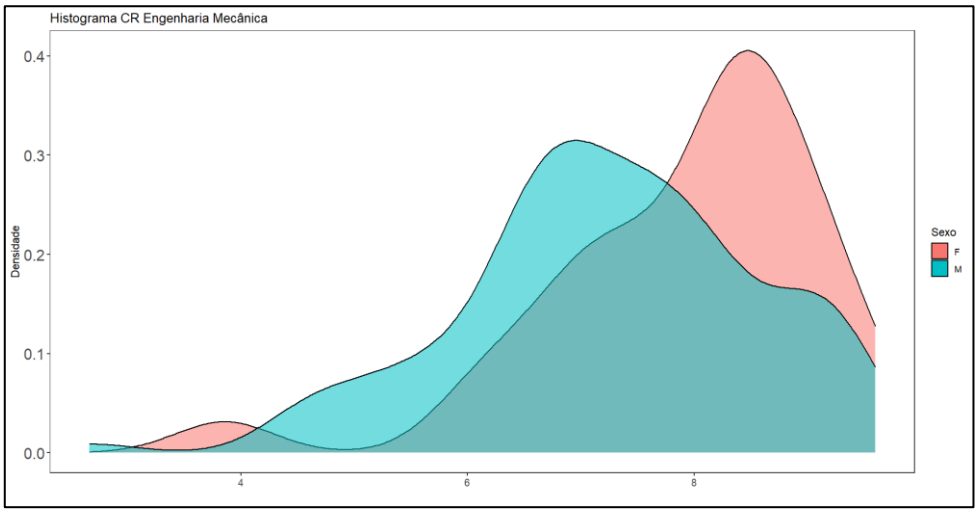
Fonte: Insper Data (elaborado pelos autores)

Anexo 8 – Histograma de CR engenharia mecatrônica



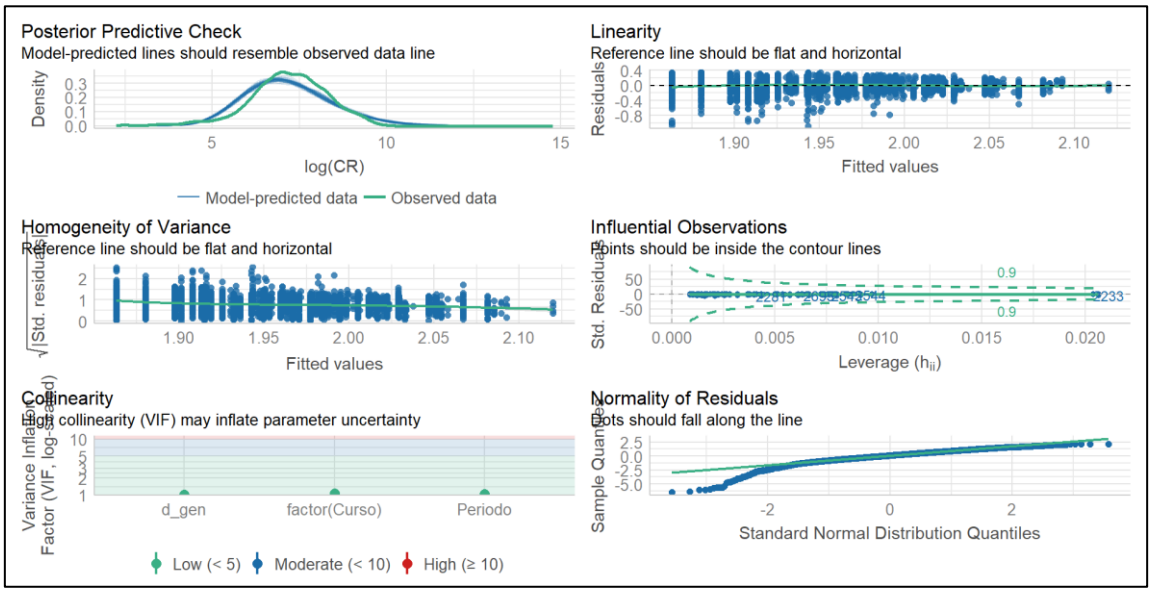
Fonte: Insper Data (elaborado pelos autores)

Anexo 9 – Histograma de CR engenharia mecânica



Fonte: Insper Data (elaborado pelos autores)

Anexo 10 – Posterior Predictive Check, Linearity, Homogeneity of Variance, Influential Observations, Collinearity, Normality of Residuals



Fonte: Insper Data (elaborado pelos autores)

Juros, inflação e atividade: uma abordagem SVAR

Integrantes: Tiago Polloni e Francisco Lopes

Orientador: Victor Hugo C. Alexandrino da Silva

1. Introdução

A macroeconomia contemporânea lida com questões de diversas áreas da economia. Em especial, a análise da inflação, atividade econômica e taxa de juros ocupa um lugar central nessa análise. Dado a conjuntura político-econômica atual, esse tema possui um lugar ainda mais relevante para o futuro da economia nacional e internacional. Em termos de previsão, o objetivo da análise feita aqui é modelar a relação entre as variáveis macroeconômicas supracitadas de forma a prever o que determinado choque, como o ocorrido com o preço do petróleo no início de 2022, pode gerar no comportamento futuro das variáveis relacionadas. De forma geral, é necessário propor um modelo autorregressivo que considere as relações da teoria macroeconômica clássica entre SELIC (taxa de juros), IPCA (*índice de preços no consumidor*) e IBC-BR (*índice de atividade do Banco Central*).

2. Metodologia e literatura

A aplicação metodológica utilizada para a previsão macroeconômica foi inspirada de estudos atuais sobre a modelagem macroeconômica, em especial de métodos formulados por cientistas econômicos que estudam essas relações e suas consequência, além de relatórios sobra a conjuntura atual feita por bancos e instituições internacionais.

Em suma, literaturas como as de Killian (2009-2012) foram muito importantes não só para entender o funcionamento e as relações das variáveis de forma autorregressiva, mas também como elas se interferem entre si (vide Imagem 1 do Anexo). A conclusão geral foi que a teoria comprova esses dois comportamentos da série, de forma que utilizaremos o método SVAR como modelo, de forma a modelar tanto os comportamentos autorregressivos como os interdependentes, isso é possível graças ao modelo em forma de matriz e as relações vetorizadas. Explicaremos de forma resumida como o modelo funciona na próxima sessão.

3. Dados

Para construir a análise desejada, foi preciso construir uma base de dados com os três indicadores mencionados acima; SELIC (taxa de juros), IPCA (*índice de preços no consumidor*) e IBC-BR (*índice de atividade do Banco Central*). A SELIC foi retirada do site do Banco Central, com uma periodicidade mensal, onde tira-se a primeira diferença de forma a encontrar a variação; O

IPCA foi retirado do site do IBGE, também disponível no IPEA-DATA, com periodicidade mensal; O IBC-BR também foi retidão do site do Banco Central com periodicidade mensal, onde foi tirada a primeira diferença do logarítimo, de forma a adequar os dados a análise. Todos os dados estão dentro do período de análise de janeiro de 2004 a dezembro de 2019. É importante ressaltar que os dados de 2020 foram retirados, visto que havia indícios de quebra estrutural, ou seja, a série mudava estruturalmente de comportamento então podia ser incluída na análise.

4. Metodologia SVAR

Dessa forma, o modelo foi construído em cima do VAR, um Vetor de Autoregressão, feito em séries temporais lineares multivariadas em que as variáveis endógenas no sistema são funções dos valores defasados de todas as variáveis endógenas. Isso permite uma alternativa simples e flexível ao sistema de equações estrutural tradicional. Um VAR poderia modelar dados macroeconômicos de forma informativa, sem impor restrições ou relações muito fortes.

Imagem 2 - A matriz de relações SVAR

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_{21} & 1 & 0 \\ \alpha_{31} & \alpha_{32} & 1 \end{bmatrix}$$

$$\Downarrow$$

$$\Phi_i = \begin{bmatrix} 1 & 0 & 0 \\ \phi_{i,21} & 1 & 0 \\ \phi_{i,31} & \phi_{i,32} & 1 \end{bmatrix}$$

Fonte: Insper Data (elaborado pelos autores)

O SVAR é a variação de um modelo VAR, que é um método de estimar múltiplas variáveis em um sistema. Um modelo de série temporal linear multivariado onde as variáveis endógenas no sistema são funções dos valores defasados de todas as variáveis endógenas. Dito de forma simples, o VAR é essencialmente uma generalização do modelo autoregressivo univariado. Comumente, notamos um VAR como um VAR(p) onde p denota o número de defasagens autorregressivas no sistema. Abaixo, vemos um exemplo simples de uma série autorregressiva, onde a variável de interesse y depende de seus próprios valores passados, ponderados pela sua relevância $\boldsymbol{\phi}$ e um fator aleatório $\boldsymbol{\varepsilon_{1,t}}$.

$$\boldsymbol{y_{1,t} = \alpha_{10} - \alpha_{12} y_{2,t} + \phi_{11} y_{1,(t-1)} + \phi_{12} y_{2,(t-1)} + \varepsilon_{1,t}}$$
$$\boldsymbol{y_{2,t} = \alpha_{20} - \alpha_{21} y_{1,t} + \phi_{21} y_{1,(t-1)} + \phi_{22} y_{2,(t-1)} + \varepsilon_{2,t}}$$

Essa restrição sugere que um choque contemporâneo da SELIC (R) afete tanto o IPCA (π) e o IBC (A) no mesmo período. Assim, choques do IPCA (π) afetam somente a o IBC (A) contemporaneamente, mas não a SELIC (R).

Imagem 3 - O modelo SVAR completo

$$\begin{bmatrix} \Delta R_t \\ \pi_t \\ \Delta \ln A_t \end{bmatrix} = \begin{bmatrix} \phi_{01} \\ \phi_{02} \\ \phi_{03} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ \phi_{1,21} & 1 & 0 \\ \phi_{1,31} & \phi_{1,32} & 1 \end{bmatrix} \begin{bmatrix} \Delta R_{t-1} \\ \pi_{t-1} \\ \Delta \ln A_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} 1 & 0 & 0 \\ \phi_{p,21} & 1 & 0 \\ \phi_{p,31} & \phi_{p,32} & 1 \end{bmatrix} \begin{bmatrix} \Delta R_{t-p} \\ \pi_{t-p} \\ \Delta \ln A_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}$$

$$y_t = \begin{bmatrix} \Delta R_t \\ \pi_t \\ \Delta \ln A_t \end{bmatrix}, \Phi_0 = \begin{bmatrix} \phi_{01} \\ \phi_{02} \\ \phi_{03} \end{bmatrix}, \Phi_i = \begin{bmatrix} 1 & 0 & 0 \\ \phi_{i,21} & 1 & 0 \\ \phi_{i,31} & \phi_{i,32} & 1 \end{bmatrix}, \epsilon_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}$$

$$i = \{1, \dots p\}$$

Fonte: Insper Data (elaborado pelos autores)

Vamos primeiro estimar um VAR padrão que reflita uma resposta econômica chave. De acordo com a teoria macroeconômica acredita-se que choques na taxa de juros nominal deveriam representar choques de política monetária. Um choque na variável de política afeta todas as outras variáveis simultaneamente. A variável é afetada por todas as outras dentro do período e é a última da ordem. Por fim, o banco central observa apenas variáveis não políticas com defasagem. Essa relação vai ser representada na forma matricial do SVAR de acordo com as respectivas elasticidades das interações de todas as variáveis. Nas Imagens 4 e 5 do Anexo isso fica mais explícito, com as elasticidades na matriz intermediaria.

4. Análise Descritiva

Para fins da análise descritiva, o que é essencial observar é o comportamento das três séries em questão, IBC, SELIC e IPCA (vide Imagens 6 e 7 no Anexo). As três séries apresentam comportamentos históricos que mostrarão como uma reage a uma mudança no comportamento da outra, mesmo que esse efeito aconteça em um tempo posterior da análise, o que será considerado graças ao modelo autorregressivo em *lags* posteriores. Além disso, as séries parecem se comportar de forma estacionária, com média e variância constante, algo imprescindível para a análise econométrica.

5. Robustez

Para que o modelo possa ser adequadamente interpretado, sem erros de precisão ou viés, deve-se garantir 4 principais "diagnósticos" em que o modelo deve ser "aprovado". Eles sendo:

1. Estacionariedade
2. Resíduos não correlacionados

3. Heterocedasticidade
4. Estabilidade

1. Estacionariedade

Em termos mais simples, uma série histórica estacionária deve-se manter em torno de uma mesma média e não apresentar uma tendência, como visto no exemplo da Imagem 8 no Anexo. Para testar se a série obtida é estacionária, utilizamos a função pp.test() que efetua um teste para estacionariedade chamado de teste de Phillips-Perron:

H_0 : série não estacionária

H_A : série estacionária

Para as três séries (IPCA, IBC e SELIC) tivemos um valor-p $< \alpha = 5\%$, rejeitando a hipótese nula e concluindo que as três séries são estacionárias.

2. & 3. Resíduos não correlacionados e Heterocedasticidade

Correlação entre resíduos e heterocedasticidade são duas características que, caso presentes no modelo, muito provavelmente irão enviesar os intervalos de confiança e, portanto, a estimação do modelo.

Para realizar o teste de não correlação dos resíduos utilizou-se a função serial.test(), que demonstrou que a modelagem não possui correlação entre os resíduos.

Para testar heterocedasticidade utilizou-se a função arch.test(), que efetua o teste ARCH multi variado em que a hipótese nula é de que o modelo não apresenta heterocedasticidade, que foi o resultado obtido para a modelagem.

4. Estabilidade

Uma série estável não apresenta quebras estruturais ao seu decorrer. Pode-se classificar uma quebra estrutural como um pico muito fora da tendencia geral da série que retorna a essa tendência depois de um tempo.

Para identificar se as três séries são estáveis, utilizamos a função stability(), em que se obteve a Imagem 9 no Anexo. Uma quebra estrutural seria representada pelo cruzamento da série histórica e a linha vermelha em cada uma. Como nenhuma das séries cruzam as linhas vermelhas, pode-se concluir que as três séries são estáveis e não possuem quebras estruturais.

6. Regressão e resultados

A partir da programação no R e a modelagem descrita, foram construídos gráficos impulso-resposta. Gráficos de impulso-resposta modelam, dentro da estrutura proposta do SVAR, por exemplo, um comportamento esperado de uma variável dado um choque de outra variável.

No eixo X mede-se o horizonte de tempo (em meses, no caso) depois em que ocorreu o choque. As linhas pontilhadas em vermelho representam o intervalo de confiança de 95%.

A Imagem 10 observada no Anexo mostra o comportamento do IPCA dado um choque na variação da SELIC. Como esperado pela teoria macroeconômica, aumentos na SELIC devem provocar uma queda no nível de preços. Pode-se perceber que, no caso do Brasil, essa resposta do IPCA deve demorar mais de um ano para retornar ao seu nível original.

Visualizando agora a Imagem 11 do Anexo, tem-se o comportamento esperado da variação da SELIC como resposta de um choque em si mesmo. Suponhamos um choque exógeno positivo na variação da SELIC, no Brasil, espera-se que essa variação retorne ao seu nível inicial no curtíssimo prazo de um mês nas estabilize nesse nível somente em um prazo de aproximadamente um ano.

Já na Imagem 12 do Anexo, observa-se o comportamento esperado, de acordo com o modelo SVAR, da variação da SELIC, dado um choque do IPCA. Aqui, novamente condizendo com a teoria macroeconômica, pode-se perceber a intervenção imediata do Banco Central quando se observa um choque no nível de preços. Ainda dentro de aproximadamente 1 ano, deve-se observar, por parte do Banco Central brasileiro um retorno da taxa de variação da SELIC para seu nível inicial, seguido por uma relativa estabilidade da taxa.

Em seguida, na Imagem 13 do Anexo, é possível verificar o comportamento modelado do IPCA dado um choque em si. Nesse gráfico pode-se perceber a rápida queda esperada, de aproximadamente 5 meses, do nível de preços e após essa queda uma estabilização do IPCA em torno de seu nível inicial. Pode-se interpretar aqui que essa brusca queda do IPCA em um curtíssimo prazo é resultado da intervenção do Banco Central, como visto no gráfico anterior.

Seguindo para o gráfico na Imagem 14 do Anexo, este ilustra o comportamento esperado do IPCA dado um choque positivo na atividade brasileira. Nesse caso, como o choque positivo de atividade implica que o Brasil estará produzindo mais do que ele é capaz, tem-se uma forte instabilidade no nível de preços até aproximadamente 14 a 15 meses depois do choque, em que se espera os efeitos das intervenções do Banco Central brasileiro com a taxa SELIC, que é detalhado no gráfico da Imagem 15 do Anexo.

Como visto no gráfico anterior, um choque positivo da atividade brasileira provocará uma instabilidade no nível de preços. Para isso, pode-se observar do gráfico da Imagem 15 que o Banco Central brasileiro deve intervir

alterando a SELIC, ao tentar controlar o IPCA, até aproximadamente 15 meses após o choque positivo, que é exatamente quando se espera uma estabilização do nível de preços (vide gráfico na Imagem 16 do Anexo).

8. Referências bibliográficas

Barsky, R. B. and Kilian, L. 2002. Do We Really Know that Oil Caused the Great Stagflation? A Monetary Alternative. In B.S. Bernanke and K. Rogoff (ed.) *NBER Macroeconomics Annual 2001*, 16: 137-183. Cambridge: MIT Press.

Barsky, R. B. and Kilian, L. 2004. Oil and the Macroeconomy since the 1970s. *Journal of Economic Perspectives*, 18(4): 115-134.

BEN S. BERNANKE, MARK GERTLER, MARK WATSON. A Structural Model of the World Oil Market: The Role of Investment Dynamics and Capacity Constraints in Explaining the Evolution of the Real Price of Oil. Systematic Monetary Policy and the Effiects of Oil Price Shocks 1997

J.H. Stock, M.W. Watson Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions Macroeconomics 2016. *Harvard University, Cambridge, MA, United States †The Woodrow Wilson School, Princeton University, Princeton, NJ, United States. The National Bureau of Economic Research, Cambridge, MA, United States.

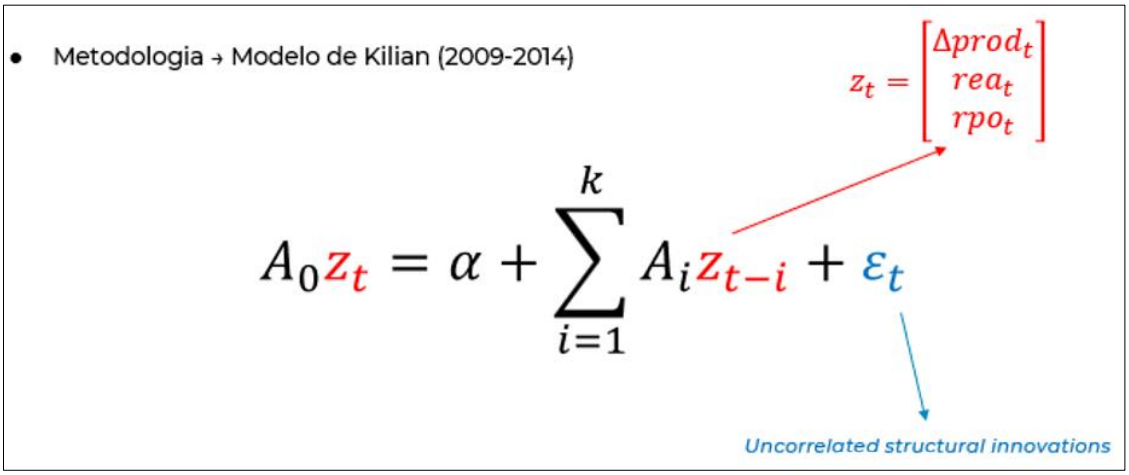
V.A. Ramey. Macroeconomic Shocks and Their Propagation 2016. University of California, San Diego, CA, United States. NBER, Cambridge, MA, United States

Takuji Fueki, Hiroka Higashi, Naoto Higashio, Jouchi Nakajima, Shinsuke Ohyama and Yoichiro Tamanyu. BIS Working Papers No 725 Identifying oil price shocks and their consequences: the role of expectations in the crude oil market

D. E. Reinhard Ellwanger. A Structural Model of the Global Oil Market. International Economic Analysis Department; Bank of Canada

9. Anexo

Imagem 1 – Exemplo da função autorregressiva de Killian(2009)



Fonte: Insper Data (elaborado pelos autores)

Imagem 4 – Modelo SVAR exemplo para a variação do preço do petróleo

$$\begin{bmatrix} \epsilon_t^{\text{production}} \\ \epsilon_t^{\text{consumption}} \\ \epsilon_t^{\text{real price of oil}} \\ \epsilon_t^{\text{GDP}} \end{bmatrix} = \begin{bmatrix} 1, & \eta_S \cdot \gamma_D, & \eta_S, & \eta_S \cdot (\eta_I \cdot \gamma_D + \gamma_G) \\ \eta_D \cdot \gamma_S, & 1, & \eta_D, & \eta_I \\ \gamma_S, & \gamma_D, & 1, & \eta_I \cdot \gamma_D + \gamma_G \\ \eta_G, & 0, & 0, & 1 \end{bmatrix} \cdot \begin{bmatrix} \text{supply shock}_t, \\ \text{oil market specific demand shock}_t \\ \text{storage demand shock}_t \\ \text{economic growth shock}_t \end{bmatrix}$$

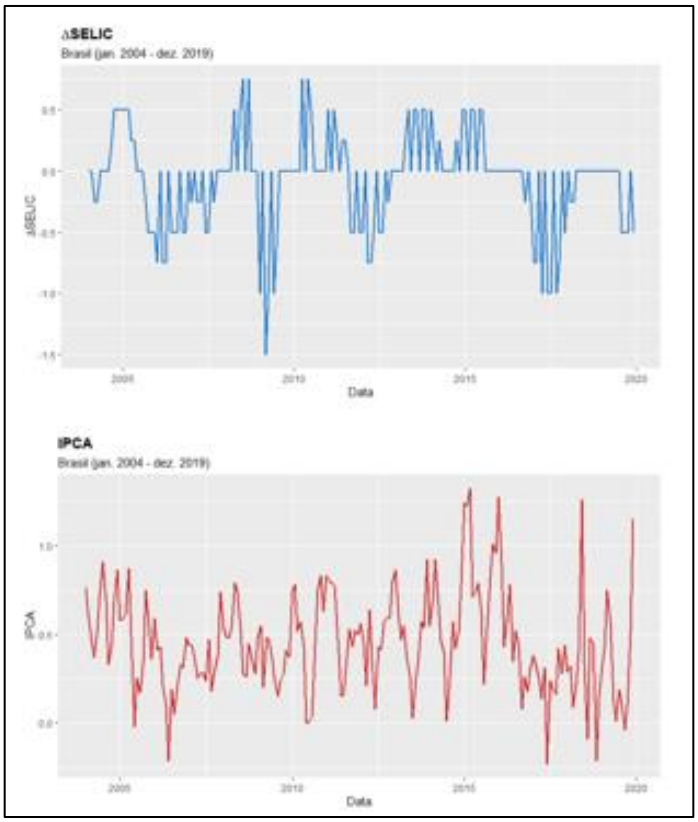
Fonte: Bank of japan 2-1-1 Nihonbashi-Hongokucho, Chuo-ku, Tokyo 103-0021, Japan

Imagem 5 – Disposição das variáveis

$$\begin{bmatrix} \Delta \text{SELIC} \\ \text{IPCA} \\ \Delta \ln(\text{IBC}) \end{bmatrix} := \begin{bmatrix} \Delta R \\ \pi \\ \Delta \ln A \end{bmatrix}$$

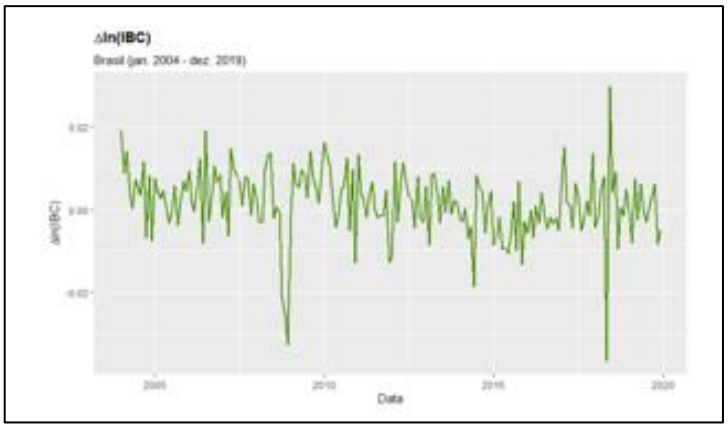
Fonte: Insper Data (elaborado pelos autores)

Imagem 6 – Série SELIC(delta) e IPCA



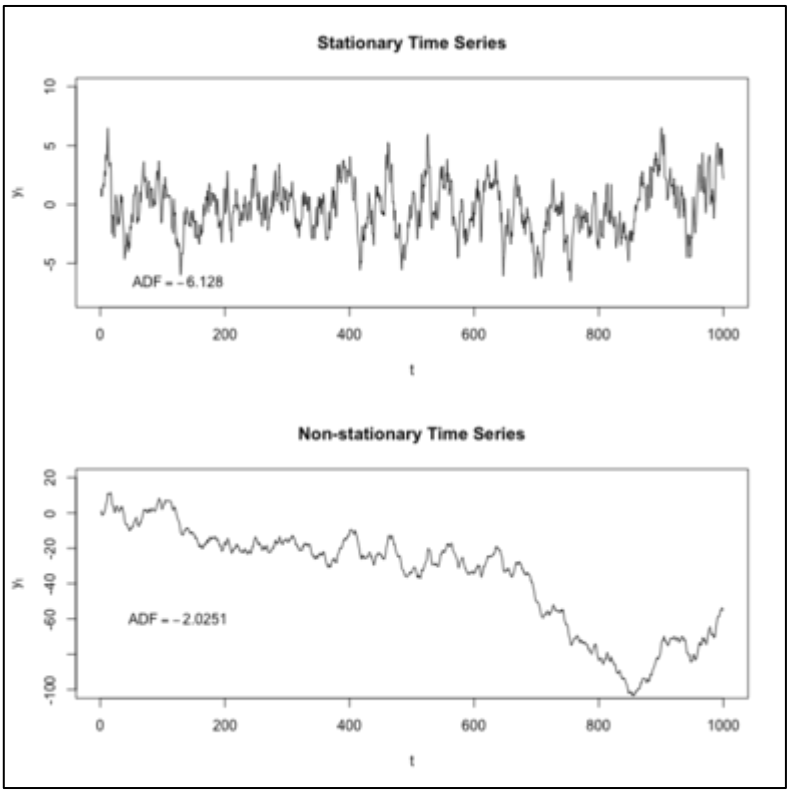
Fonte: Insper Data (elaborado pelos autores)

Imagem 7 – Série da primeira diferença do logaritmo do IBC



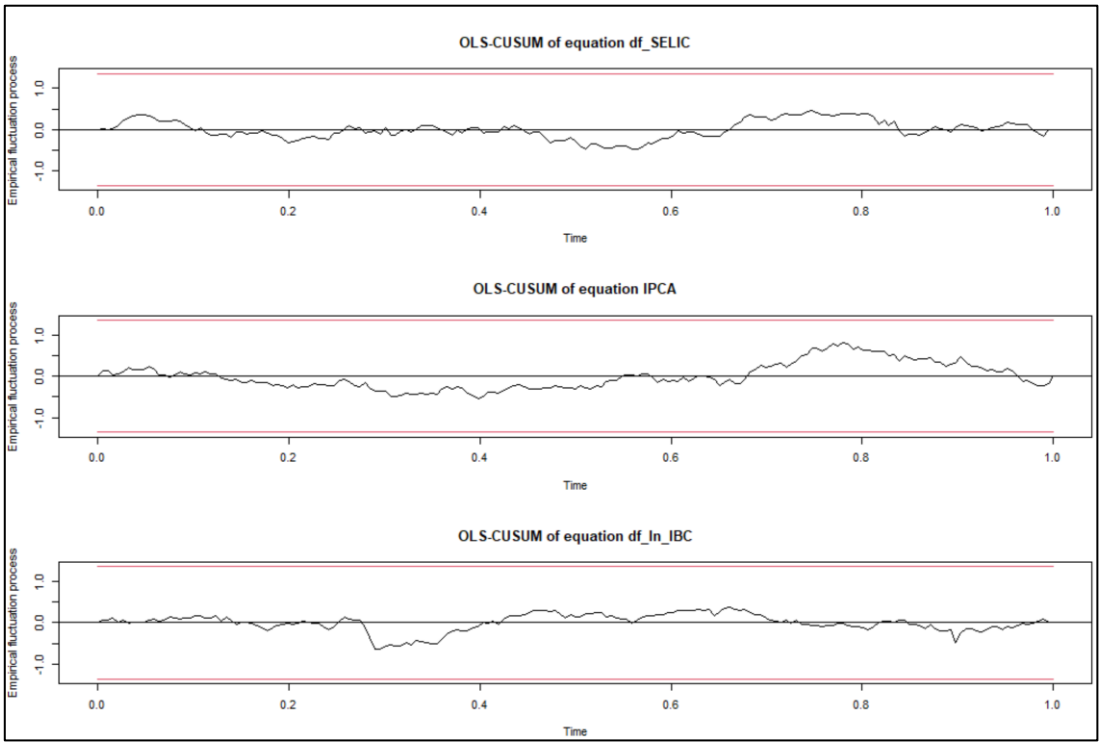
Fonte: Insper Data (elaborado pelos autores)

Imagem 8 – Exemplo de uma série histórica estacionária



Fonte: O'Reilly

Imagem 9 – Retorno da função *stability*()

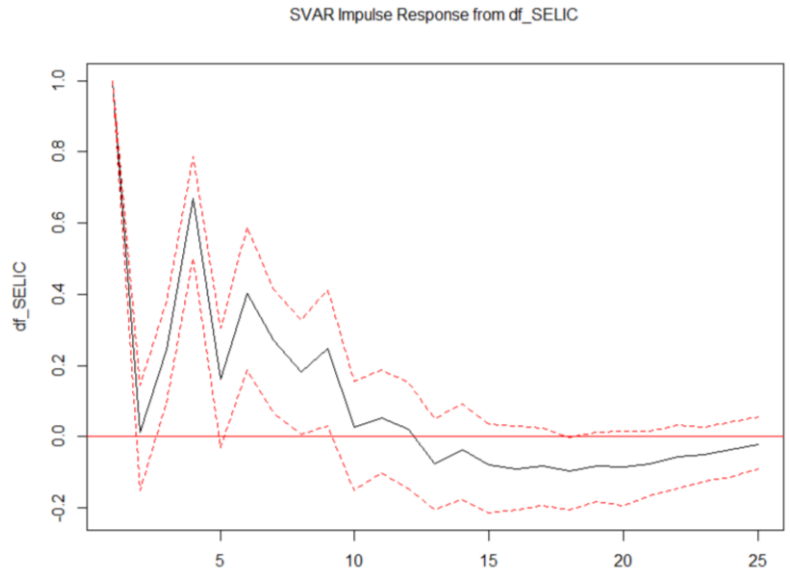


Fonte: Insper Data (elaborado pelos autores)

Imagem 10 – Comportamento do IPCA dado um choque na variação da SELIC

Fonte: Insper Data (elaborado pelos autores)

Imagem 11 – Comportamento esperado da variação da SELIC como resposta de um choque em si mesmo

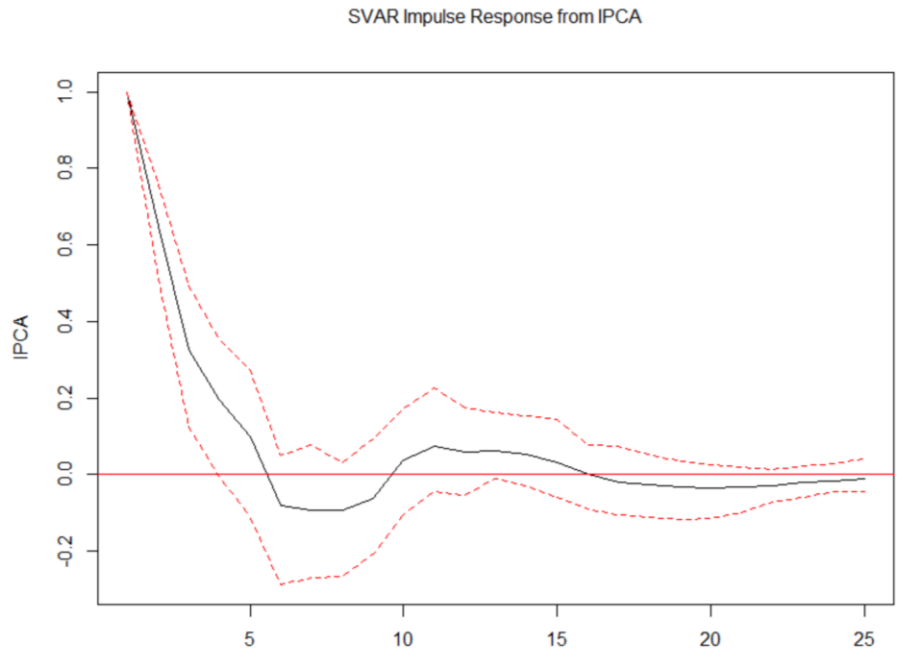


Fonte: Insper Data (elaborado pelos autores)

Imagem 12 – Comportamento esperado, de acordo com o modelo SVAR, da variação da SELIC, dado um choque do IPCA

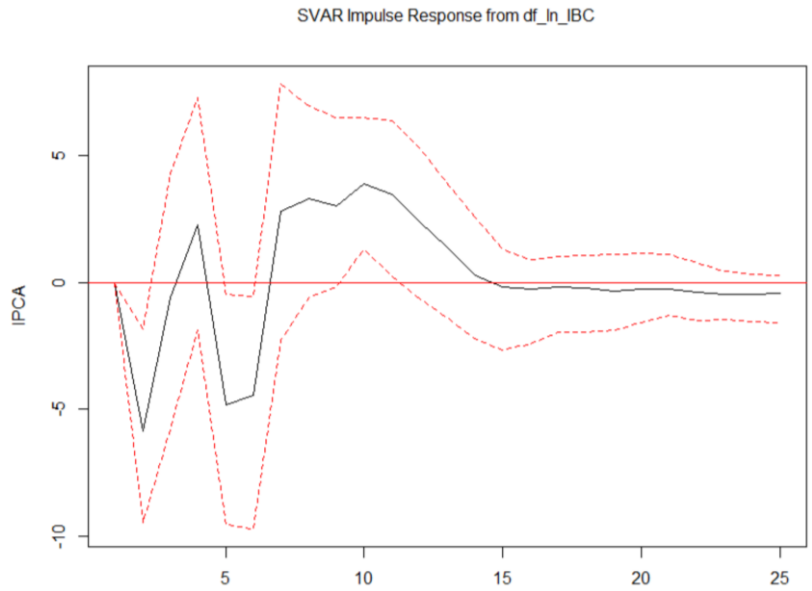
Fonte: Insper Data (elaborado pelos autores)

Imagem 13 – Comportamento modelado do IPCA dado um choque em si



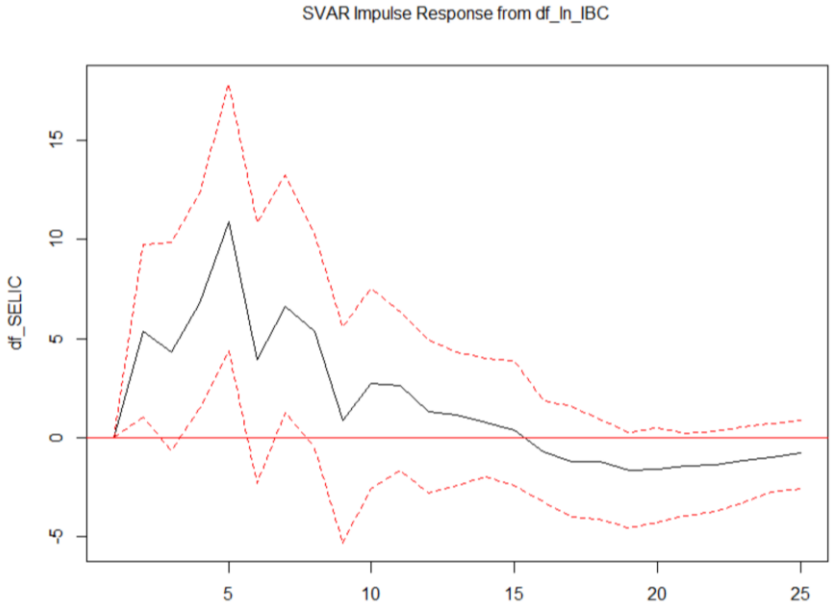
Fonte: Insper Data (elaborado pelos autores)

Imagem 14 – Comportamento esperado do IPCA dado um choque positivo na atividade brasileira



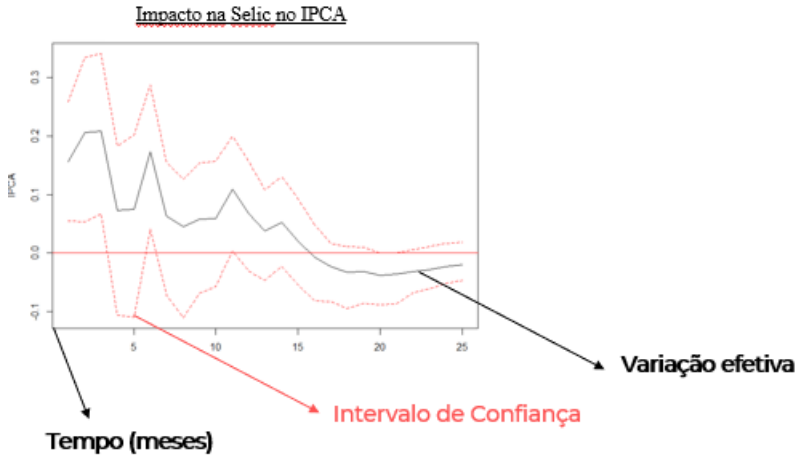
Fonte: Insper Data (elaborado pelos autores)

Imagem 15 - Os efeitos das intervenções do Banco Central brasileiro com a taxa SELIC



Fonte: Insper Data (elaborado pelos autores)

Imagem 16 – Impacto da Selic no IPCA



Fonte: Insper Data (elaborado pelos autores)

Análise da relação entre preços e volume de avaliação em sites de comparação de preços

Integrantes: Carolina Bromfman, Livia Reis e Michel Wachslicht

Orientador: Ademar Concon-Neto

1. Objetivo e motivação

Tendo como ponto de partida o escopo de Marketing, esse projeto foi desenvolvido de forma a auxiliar empresas que possuam um ambiente de avaliação na web de seus produtos/serviço, a compreender melhor os consumidores. Assumindo como premissa que avaliações são um referencial da percepção do consumidor sobre um certo produto e que tal percepção esteja atrelada ao cumprimento de expectativas, tem-se que um preço elevado pode gerar uma maior percepção de valor que, caso não seja suprida, pode gerar insatisfação e mais avaliações negativas. Tal reflexão inicial é apenas uma hipótese do impacto do preço, desse modo, o objetivo do projeto é auxiliar empresas a entenderem de que forma a precificação pode influenciar as avaliações de seus produtos. Por outro lado, no que tange os consumidores, investiga-se possíveis vieses causados pelos preços dos produtos presentes nas avaliações em sites de review.

2. Revisão da literatura

Segundo a análise econométrica de dados textuais apresentada no artigo *"Deriving the Pricing Power of Product Features by Mining Consumer Reviews"* (ARCHAK N.; GHOSE A.; IPEIROTIS P.G.; 2011), conclui-se que o efeito exercido pelo preço e pelo volume de avaliações sobre as vendas é contrário. Isto é, mais especificamente, através do uso de *Text Mining*, por meio de NLP, somada a metodologia Crowdsourcing, uma maneira semi-automatizada de usar a inteligência humana, ao invés de totalmente automatizada; os autores concluem, no artigo, que enquanto o preço tem relação negativa sobre as vendas, o volume tem relação positiva sobre elas.

No que diz respeito à influência social que permeia o ato da compra, conforme "*Arousal, valence, and volume: how the influence of online review characteristics differs with respect to utilitarian and hedonic products*" (REN, J.; NICKERSON, J.V.; 2019), há evidências de que o alto volume de avaliações aumenta a visibilidade do produto e pode o tornar mais socialmente desejável. Indo além, pode-se utilizar a teoria de Bens de Veblen para entender o fenômeno de enfoque aqui, a qual diz respeito aos produtos que possuem a procura proporcional ao seu preço elevado, uma vez que o valor percebido nesses produtos está relacionado ao status que eles proporcionam.

Tendo em vista que o conceito de Bens de Veblen se aplica para bens de luxo, optou-se por usar hotéis como objeto para as análises. Isso porque

assumiu-se o pressuposto de que bens de luxo estão relacionados com características como requinte e qualidade, as quais estão presentes na hotelaria (BELLAICHE; MEI-POCHTLER; HANISCH, 2010).

3. Construção da base de dados

Através da biblioteca Selenium, presente na linguagem de programação Python, realizou-se um *Web Scraping* do site Google Hotels, do qual coletou-se os dados de 379 hotéis do estado do Rio de Janeiro. Tal site tem como vantagem, além da ampla gama de hotéis e do fato de possuir uma página para cada hotel, o que auxilia na coleta das informações. Em contrapartida, o site de *reviews* contempla imóveis do Airbnb, os quais foram excluídos da coleta, por não possuírem avaliações escritas.

Os dados estáticos como nome dos hotéis, número de avaliações, quantidade de estrelas e preço foram recolhidos com base nas classes de cada item, presentes na linguagem de marcação utilizada na construção de sites (HTML). Para elementos não estáticos, como botões de troca de página e mudança de aba, utilizou-se o XPath. Além disso, também fez-se o uso da biblioteca unidecode para remover acentuação e outros símbolos utilizados pela língua brasileira, e presentes nas avaliações coletadas, os quais não são compreendidos pelo Python. Em suma, foram coletados os seguintes dados: nome do hotel, preço, quantidade de avaliações, avaliação média, avaliações escritas por consumidores e as respectivas notas dadas por estes.

4. Análises descritivas

Enxergar os dados que foram coletados é de extrema importância para o desenvolvimento do projeto, tendo-se obtido os dados de interesse via *Web Scraping*, como descrito previamente. Isso porque uma análise descritiva ajuda a se ter uma visão geral dos dados quantitativos coletados, sendo esses o preço, volume de avaliações e avaliação média do hotel (1-5 estrelas).

Primeiramente, é necessário ressaltar que foram retirados dois *outliers* dessa análise, o Copacabana Palace e o Fasano. Os seus preços divergiam muito dos demais, custando mais de R$1700 reais, fato que impacta desde a dispersão dos dados até a linearidade dos mesmos. Ainda, ressalta-se que nenhuma conclusão é formulada a partir das análises descritivas a seguir.

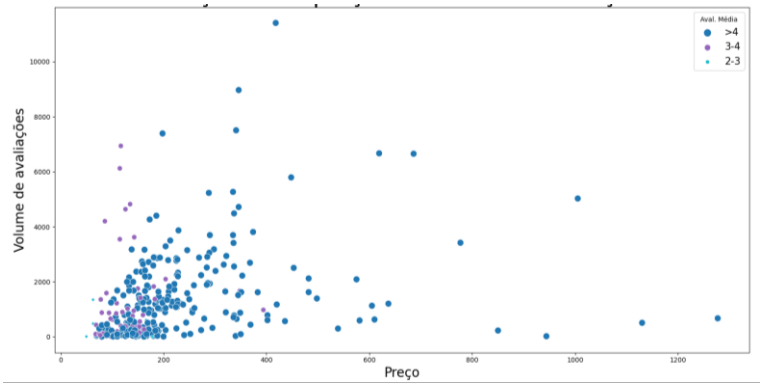
Conforme mostra a Imagem 1 dos anexos, tanto o preço, quanto o volume possuem uma certa dispersão dos dados, justificado por alguns hotéis em que o preço é mais elevado que o restante. Tal fato indica baixa homogeneidade da amostra e, possivelmente, uma gama de características e padrões de comportamento diferentes para/com certos hotéis. A Imagem 2, histograma do preço, mostra essa grande dispersão dos preços.

Outra variável de extrema importância para a pesquisa é o volume de avaliações. Como a tabela (Imagem 1) demonstra, a variação é muito grande,

o que pode dificultar a análise, uma vez que quando hotéis recebem poucas avaliações, a sua nota pode ficar viesada. Contudo, mesmo com a amostra pequena, apenas 10% dos hotéis possuem menos de 100 avaliações.

Para tentar, descritivamente, enxergar uma relação entre o preço e a quantidade de avaliações de cada hotel, foi feito o gráfico a seguir. A partir dele não se pode notar uma tendência, porém, percebe-se que o aumento da dispersão de pontos nos preços mais elevados, sendo isso um indício da hipótese de heterogeneidade de comportamento citada anteriormente.

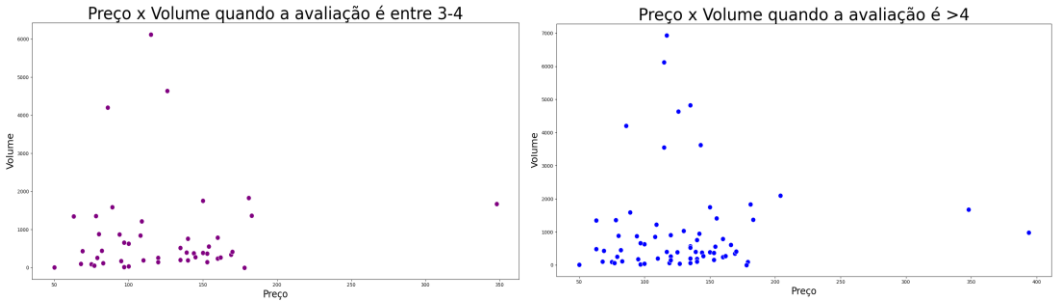
Imagem 4 - Gráfico de dispersão entre preço e volume de avaliações



Fonte: Autoria Própria

Além do mais, o gráfico está colorido para a avaliação média, pois essa pesquisa também tenta enxergar se outros fatores como a avaliação do hotel podem impactar na relação preço e volume. Os dois gráficos a seguir mostram a relação do preço com o volume, porém agora separado por avaliação média. O primeiro quando a avaliação está entre 3 e 4, e o segundo, quando esta é maior do que 4. Há somente um hotel que possui avaliação menor do que 3. É possível, então, perceber que a avaliação média que o hotel recebeu não tem impacto aparente na relação do preço e volume de avaliações.

Imagem 5 - Gráfico de dispersão



Fonte: Autoria Própria

Ademais, a fim de entender qual seria o modelo mais adequado para a regressão linear que seria feita em sequência, testou-se as quatro formas de modelagem - *lin-lin*, *lin-log*, *log-lin* e *log-log* - e dentre elas a que mais se aproximou de um modelo linear foi a *log-log* (Imagem 6). Vale ressaltar que uma hipótese para que as extremidades tenham assumido um formato mais exponencial é a concentração de hotéis com valores medianos ou *outliers* de preços elevados.

Para investigar mais a fundo a hipótese, limitou-se o logaritmo dos preços (eixo x do gráfico) para valores até 10^2,6, desse modo, foi possível se aproximar ainda mais de um modelo linear, o que pode indicar que a hipótese anteposta é razoável (Imagem 7). Vale ressaltar que o modelo log-log captura a elasticidade da variável resposta (volume), com a variável independente (preço) e, dado a horizontalidade da linha, pode indicar uma relação elástica, ou seja, o número de avaliações varia em resposta às mudanças de preço.

5. Metodologia

5.1. Análise das avaliações

A fim de construir um modelo que abrangesse a percepção do consumidor, optou-se por realizar uma análise de sentimentos das avaliações escritas coletadas, por meio de técnicas de NLP, as quais permitem que uma máquina entenda a linguagem humana. Primeiramente, de modo a entender quais eram as principais palavras citadas nas avaliações escritas, utilizou-se o método *Wordcloud*.

Imagem 7 - Nuvem de *Tags*

Fonte: Autoria Própria

Baseado nas palavras mais recorrentes nas avaliações, foram determinados os temas para classificação de tais. Por meio do método de classificação de texto *Zero Shot*, baseado em NLI, linguagem natural que determina a relação lógica entre dois textos, definiu-se a probabilidade de cada rótulo - comida, localização, preço, limpeza,

infraestrutura, cômodo e atendimento - ser o assunto central tratado em cada avaliação. Nesse sentido, foi possível compreender qual é o principal fator presente na observação do consumidor sobre determinado hotel.

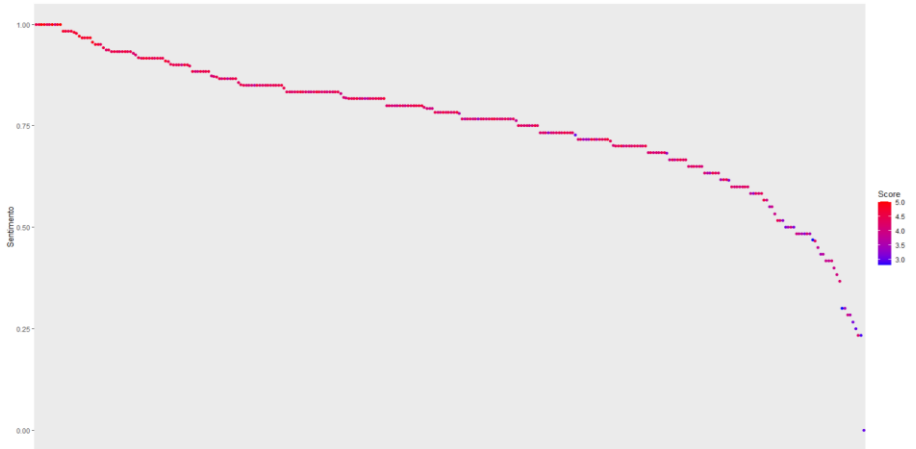
Apoiando-se na mesma lógica, através dos rótulos "positivo" e "negativo", também foi possível compreender se a percepção do consumidor, em sua avaliação, foi favorável ao hotel. Vale ressaltar que o método *Zero Shot* implica no uso de modelos pré prontos utilizando a biblioteca *Transformers*.

Depois de rodar o modelo, foi possível listar a importância das classes em cada avaliação. Assim, pode-se identificar quais classes apareceram mais vezes como mais importantes, como foi feito na Imagem 8 dos anexos. A partir dele, foi identificado que o atendimento é a classe com maior importância nas avaliações; enquanto a comida é a de menor importância. Vale ressaltar que tal resultado não era o esperado pelo grupo, o qual imaginava que preço seria a variável mais relevante, porém uma hipótese para isso não ter ocorrido é que o preço funciona como uma régua para a expectativa do cliente sobre o atendimento do hotel. Desse modo, quando o preço é alto mas o serviço não é de qualidade, o consumidor irá pontuar esse segundo aspecto em sua avaliação, ao invés do preço elevado.

Analogamente, pode-se utilizar o mesmo raciocínio para entender se as avaliações são majoritariamente positivas ou negativas. O resultado está presente na Imagem 9, em que se identificou que a grande maioria das avaliações são entendidas como positivas.

Buscando entender se as avaliações escritas representam a nota com a qual o hotel é atribuído, o seguinte gráfico foi criado. Nele é visto que, apesar de tender à realidade, alguns hotéis possuem alta taxa de comentários positivos, mesmo que possua nota baixa.

Imagem 10 - Porcentagem de avaliações positivas por hotel pelo nota do hotel



Fonte: Autoria Própria

Além disso, foi treinado um modelo de Random Forest para para que fosse estudado a importância das variáveis no impacto do volume de avaliações. Apesar de criar um modelo de previsão não ser um foco, é um bom método para se entender quais variáveis são mais relevantes. Segundo a Imagem 11, é interessante notar que a variável Nome é a mais importante, e isso se dá, possivelmente, pelo fato de cada hotel ter muitas características específicas que não estão sendo abrangidas pelas outras variáveis.

5.2. Regressão

Enfim, com o propósito de responder nossa questão inicial, optou-se pela regressão linear múltipla, uma vez que a variável resposta, volume de avaliações, é métrica. Dessa forma, a fim de cumprir com a análise inicial proposta, a principal variável independente é o preço em reais, sendo assim, busca-se interpretar o coeficiente da mesma (β_1).

Além das variáveis já citadas, de modo tanto a contemplar a percepção do consumidor, quanto a evitar a exclusão de variável relevante, adicionou-se as seguintes variáveis e interações:

- Número de estrelas;
- Avaliação majoritariamente positiva ou negativa? (*dummy*: assume 1 quando for, majoritariamente, positiva e 0 negativa);
- Comida é fator mais relevante nas avaliações? (*dummy*);
- Localização é fator mais relevante nas avaliações? (*dummy*);
- Preço é fator mais relevante nas avaliações? (*dummy*);
- Atendimento é fator mais relevante nas avaliações? (*dummy*);
- Limpeza é fator mais relevante nas avaliações? (*dummy*);
- Interação entre Preço e Relevância do preço;
- Interação entre Número de estrelas e Avaliação é positiva ou negativa.

Segue em sequência a construção da equação:

$$\ln(Volume) = \beta_0 + \beta_1 \ln(Preço) + \beta_2 Estrela + \beta_3 Percep + \beta_4 R_{Comida} + \beta_5 R_{Loc} + \beta_6 R_{Preço} + \beta_7 R_{Atend} + \beta_8 R_{Limpeza} + \beta_9 R_{Comodo} + \beta_{10} Preço * R_{Preço} + \beta_{11} Estrela * Percep$$

Vale ressaltar que as *dummys* foram criadas a partir da premissa que, para ser o fator mais relevante, ou seja, o que assumirá o número 1, ele precisa assumir a maior probabilidade na avaliação, dado que a

somatória das probabilidades de todos os fatores para cada avaliação deve dar 100%. Os demais fatores que não assumirem a posição de maior relevância recebem o valor 0. Nessa lógica, a variável "Perceb" recebe o valor 1 quando a probabilidade da avaliação ser positiva é maior que 50%.

6. Limitações

O estudo possui limitações tanto no momento da coleta de dados, quanto na modelagem. Durante esse primeiro momento, a principal limitação a ser constatada tem relação com a data em que os preços foram coletados e as avaliações foram feitas. Foi necessário assumir como premissa preços estáticos, o que não ocorre na prática, porém tentou-se compensar tal incoerência com a variável de importância do preço na avaliação. Ainda no período de coleta das avaliações, enfrentou-se um problema relacionado à opção de "ler mais", desse modo, não foi possível coletar todas as avaliações por completo. Porém, a parte coletada foi suficiente para a execução das análises utilizando NLP. Também é possível citar o fato dos hotéis coletados serem apenas da região do Rio de Janeiro.

Em termos da modelagem, é relevante citar o fato do *Zero Shot* não ser completamente otimizado para a língua portuguesa, também havendo restrição de informações características de cada hotel.

7. Resultados e conclusão

Após a realização da regressão, a qual pode ser observada na Imagem 12 dos anexos, constatou-se que todas as variáveis independentes, com exceção das interações, têm efeito positivo sobre o volume de avaliações. É relevante citar dois principais principais coeficientes, β_1 e β_3, ou seja, os coeficientes do preço e da percepção do consumidor sobre o hotel com base em sua avaliação.

O primeiro indica que uma variação de 1% no preço resultará em aumento de 1% no volume de avaliações, fato que é condizente com o pressuposto de elasticidade do modelo de regressão log-log. Já o segundo coeficiente indica que, quando a percepção passa de negativa (valor 0 da variável) para positiva (valor 1 da variável), a quantidade de avaliações aumenta em 200%. Tal efeito pode indicar que consumidores sentem-se mais motivados a avaliar quando sua experiência em certo hotel foi positiva.

Sendo assim, devido aos resultados encontrados, é recomendável que empresas do ramo de hotelaria não deixem de lado a variação no volume de avaliações dada uma mudança de preço. Isso porque, como citado na revisão da literatura, quanto maior o número de avaliações, maior será a visibilidade do seu serviço e, assim, mais desejável ele será. Ademais, é interessante que essas empresas também incentivem a avaliação do estabelecimento, visto que

a motivação para escrever um comentário é duplicada quando a experiência é positiva.

Em termos da teoria, uma constatação interessante é que mesmo que o preço tenha, diretamente, uma influência negativa sobre as vendas, uma vez que ele aumenta o volume de avaliação, ele tem um efeito indiretamente positivo sobre as vendas.

8. Referências bibliográficas

ARCHAK N.; GHOSE A.; IPEIROTIS P.G. (2011) **Deriving The Pricing Power Of Product Features By Mining Consumer Reviews.** Management Sci.

REN, J.; NICKERSON, J.V. (2019) **Arousal, Valence, And Volume: How The Influence Of Online Review Characteristics Differs With Respect To Utilitarian And Hedonic Products.** European Journal of Information.

BELLAICHE, J. M.; MEI-POCHTLER, A; HANISCH, D. (2010). **The New World Of Luxury: Cough Between Growing Momentum And Lasting Change.** The Boston Consulting Group.

"The Goods That Become More Desirable The More Expensive They Get". Jake Courage. Disponível em: https://blog.42courses.com/home/2018/7/10/the-goods-that-become-more-desirable-the-more-expensive-they-get

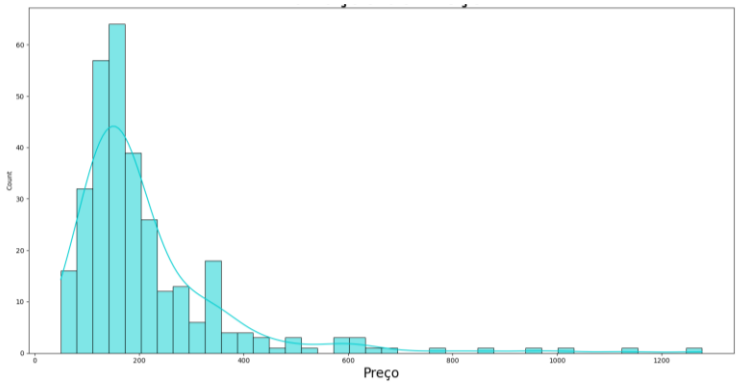
9. Anexo

Imagem 1 - Tabela sumário

	Preço	Volume	Aval. média
Média	214	1335	4,2
Mediana	168	868	2,3
Max	1277	11406	5
Min	50	2	2,8
Q1 (25%)	130	243	4,1
Q3 (75%)	236	1828	4,5
Desvio Padrão	157,6	1587,8	0,39

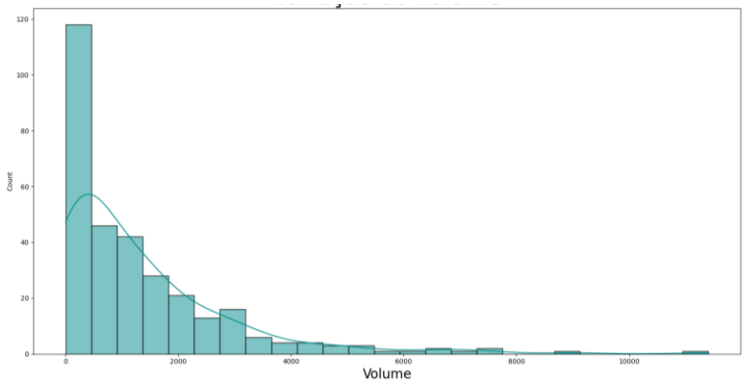
Fonte: Autoria Própria

Imagem 2 - Histograma de preço



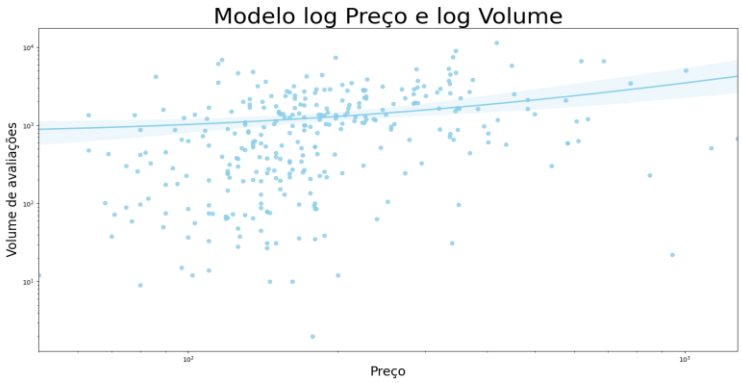
Fonte: Autoria Própria

Imagem 3 - Histograma do volume



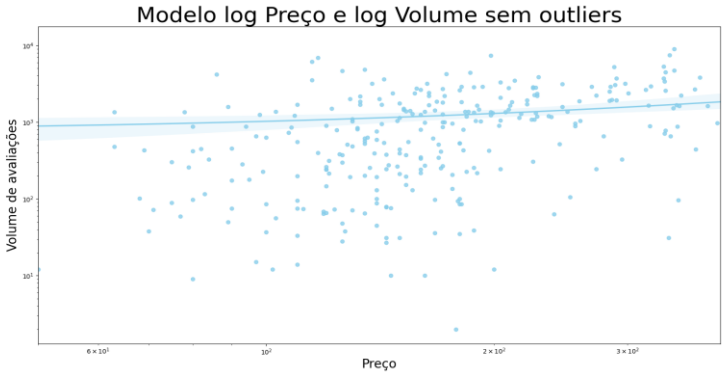
Fonte: Autoria Própria

Imagem 6 - Modelo *log-log*



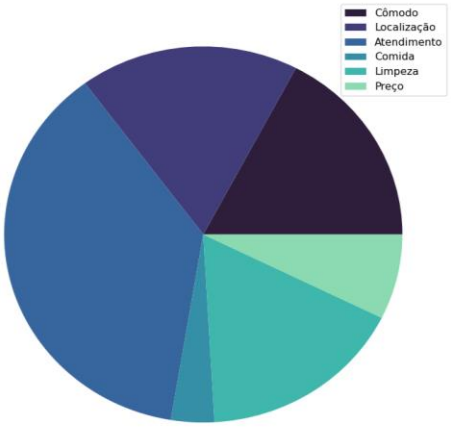
Fonte: Autoria Própria

Imagem 7 - Modelo *log-log* com um corte dos valores maiores



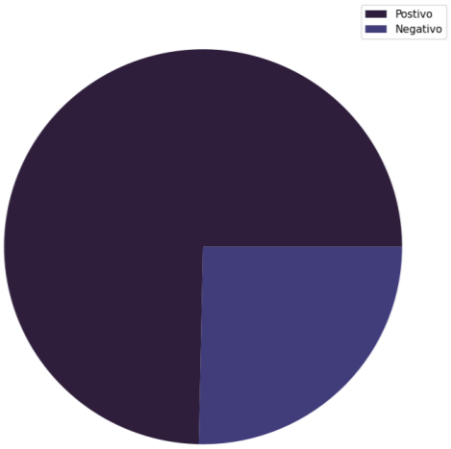
Fonte: Autoria Própria

Imagem 8 - Frequência em que as classes apareceram como mais importantes



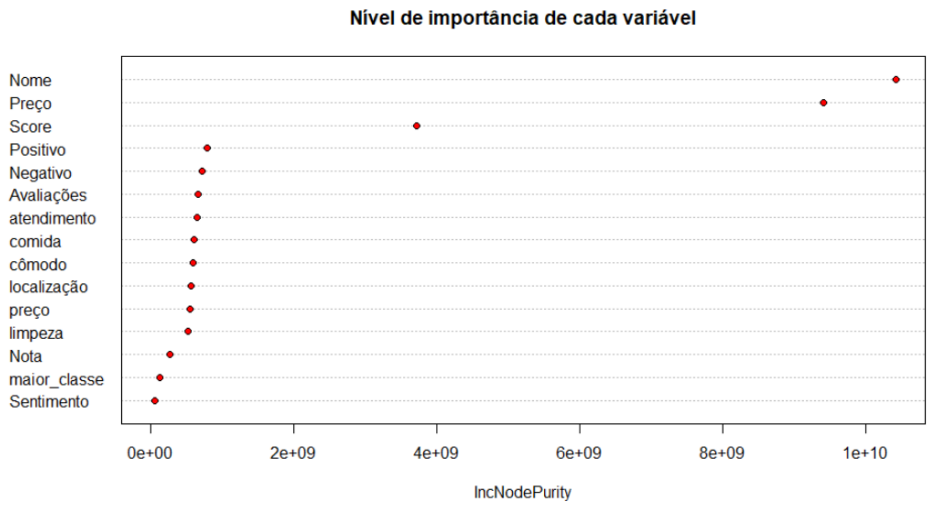
Fonte: Autoria Própria

Imagem 9 - Frequência da polaridade das avaliações



Fonte: Autoria Própria

Imagem 11 - Nível de importância de cada variável



Fonte: Autoria Própria

Imagem 12 - Saída da regressão em R

```
Residuals:
     Min       1Q   Median       3Q      Max
-6.0125  -0.8032   0.1956   0.8264   2.8548

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.8489269  0.1743874   4.868 1.14e-06 ***
Preço_log       1.0118025  0.0189605  53.364  < 2e-16 ***
Score           0.1570384  0.0428490   3.665 0.000248 ***
Percep          2.0390421  0.2106131   9.681  < 2e-16 ***
R_Comida        0.3156335  0.0548566   5.754 8.87e-09 ***
R_Loc           0.1755152  0.0279597   6.277 3.52e-10 ***
R_Preço         0.2285336  0.0578705   3.949 7.88e-05 ***
R_Atend         0.2755093  0.0277585   9.925  < 2e-16 ***
R_Limpeza       0.1426529  0.0317767   4.489 7.19e-06 ***
R_Comodo               NA         NA      NA       NA
R_Preço:Preço -0.0006304  0.0002034  -3.099 0.001946 **
Score:Percep  -0.5861745  0.0503267 -11.647  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.154 on 17477 degrees of freedom
  (60 observations deleted due to missingness)
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1923
F-statistic: 417.2 on 10 and 17477 DF,  p-value: < 2.2e-16
```

Coleta de dados de decisões do Superior Tribunal de Justiça

Integrantes: Laura Casarin, Victor Alves, Yasmin Bocatto

Orientadora: Luciana L. Yeung

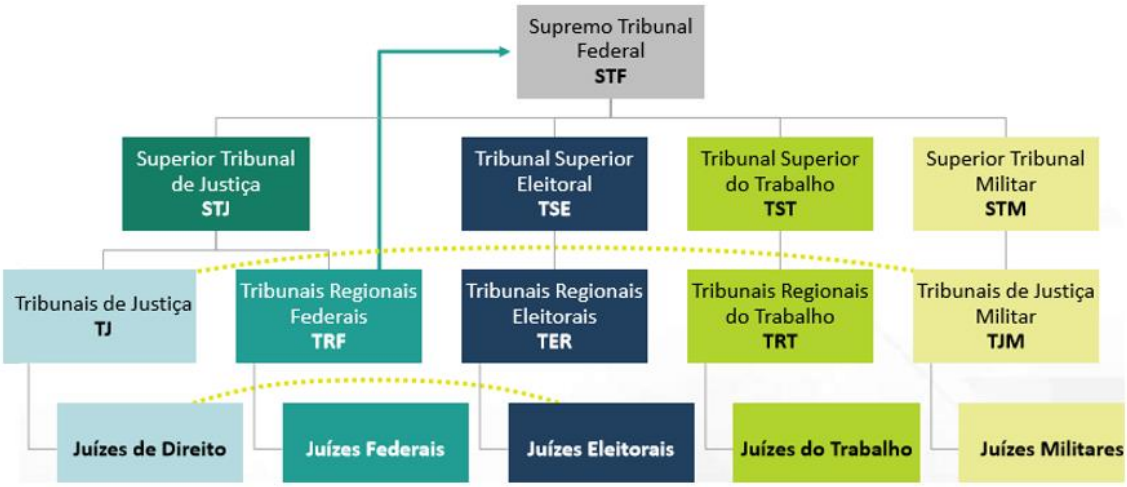
Resumo

As novas técnicas de inteligência artificial permitem a extração mais rápida de informações de milhares de documentos, algo que pode incentivar estudos e análises que antes não eram possíveis manualmente. Desse modo, como principal objetivo deste projeto, pretende-se a aprendizagem de coleta automática de dados de decisões jurídicas de tribunais brasileiros para eventual análise. Assim, para a sua realização foram necessárias duas etapas: uma coleta manual cujo aprendizado teve foco em uma familiarização com termos jurídicos e uma coleta automatizada a partir de ferramentas como *web scraping* e *Regular Expressions* (*Regex*).

1. Fonte da pesquisa

Este projeto selecionou acórdãos do Supremo Tribunal de Justiça (STJ) como objeto de estudo. Estes são casos decididos monocraticamente por um magistrado que, na maioria das vezes, partem das instâncias inferiores e sobem para o STJ através de recursos. Esse processo acontece da seguinte forma: caso uma das partes não esteja de acordo com a decisão dos Juízes de direito da primeira instância, ela recorre à segunda instância. Ao recorrer, o caso passa a ser julgado pelo Tribunal de Justiça (TJ). Se, novamente, uma das partes estiver insatisfeita com a decisão dos desembargadores, o processo finalmente chega ao STJ e será julgado por um ministro – no caso de acórdãos. Então, no Supremo Tribunal de Justiça os acórdãos estão em sua terceira instância, logo, são julgados de maneira definitiva sem a possibilidade de novos recursos. Resultando, assim em menores riscos a respeito de uma futura mudança da decisão ao considerar uma futura análise.

Figura 1 - Sistema Judiciário brasileiro



Fonte: Jusbrasil

2. Primeira etapa

a. Tipo dos acórdãos coletados

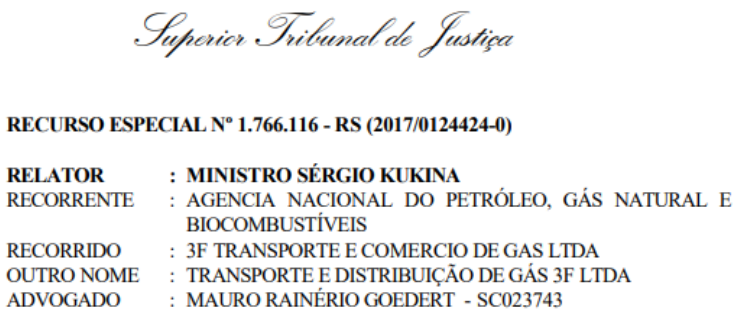
Para fins deste projeto, supondo que uma pesquisa será feita a respeito do tema, foram coletados dados de acórdãos da última instância em que uma das partes, ou seja, recorrente ou recorrida, é uma agência reguladora. Sendo a definição formal do termo:

> "Agência reguladora é pessoa jurídica de direito público interno, geralmente constituída sob a forma de autarquia especial ou outro ente da administração indireta, cuja finalidade é regular e/ou fiscalizar a atividade de determinado setor da economia de um país." (Fonte: Jusbrasil)

b. Coleta manual

Deste modo, escolhido o tema, a fase inicial de coleta manual se inicia com uma leitura extensa de todos os acórdãos, ao mesmo tempo que a identificação do acórdão, o estado, o relator, as partes, a data e a decisão eram digitadas em uma planilha do Excel.

Figura 2 - Acórdão exemplo de agência reguladora recorrente



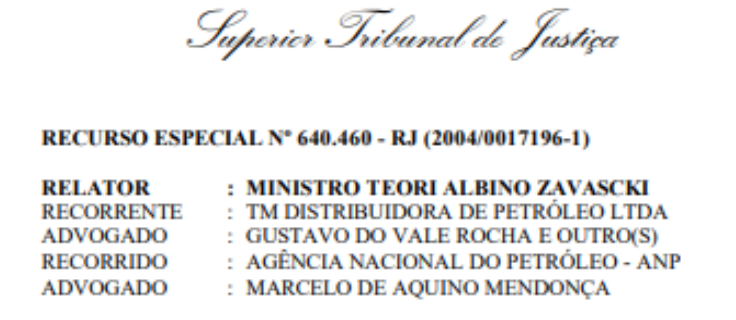
Superior Tribunal de Justiça

RECURSO ESPECIAL Nº 1.766.116 - RS (2017/0124424-0)

RELATOR	**: MINISTRO SÉRGIO KUKINA**
RECORRENTE	: AGENCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS
RECORRIDO	: 3F TRANSPORTE E COMERCIO DE GAS LTDA
OUTRO NOME	: TRANSPORTE E DISTRIBUIÇÃO DE GÁS 3F LTDA
ADVOGADO	: MAURO RAINÉRIO GOEDERT - SC023743

Fonte: site do STJ

Figura 3 – Acórdão exemplo de agência reguladora recorrida



Superior Tribunal de Justiça

RECURSO ESPECIAL Nº 640.460 - RJ (2004/0017196-1)

RELATOR	**: MINISTRO TEORI ALBINO ZAVASCKI**
RECORRENTE	: TM DISTRIBUIDORA DE PETRÓLEO LTDA
ADVOGADO	: GUSTAVO DO VALE ROCHA E OUTRO(S)
RECORRIDO	: AGÊNCIA NACIONAL DO PETRÓLEO - ANP
ADVOGADO	: MARCELO DE AQUINO MENDONÇA

Fonte: site do STJ

3. Segunda etapa

A construção da base de dados por ferramentas se deu por meio de outras duas etapas: a coleta e o tratamento.

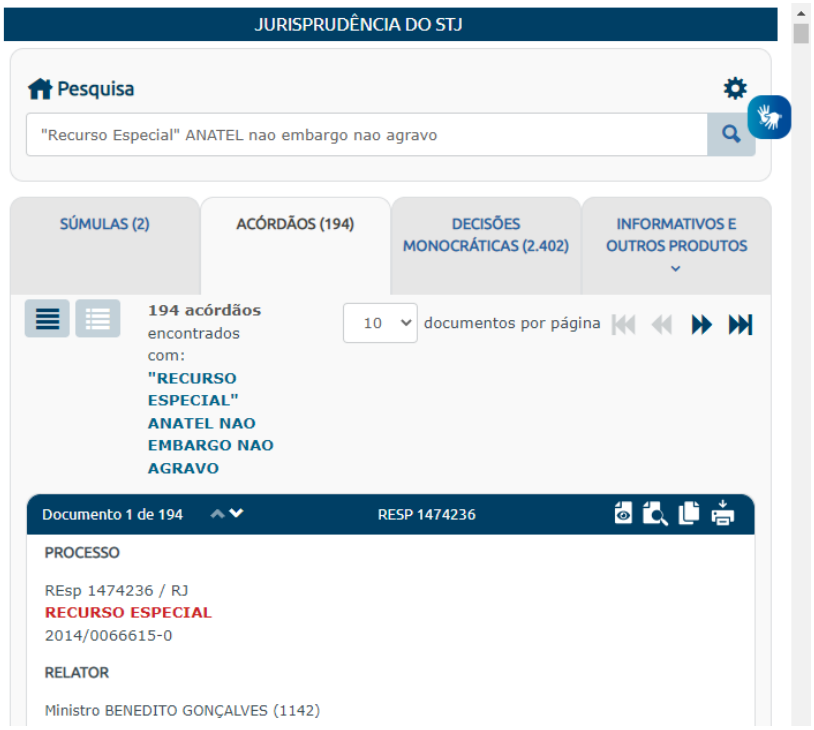
Aplicando ferramentas relacionadas à mineração de dados web (*web scraping*), foi construído um programa automatizado (*bot*). O objetivo deste foi coletar informações de dois sites: STJ – Jurisprudência do STJ e STJ – Consulta Processual. Do primeiro foram coletadas as seguintes informações dos processos: número, estado, data de julgamento, data de publicação, nome do ministro e ementa. Do segundo, foram coletadas as partes dos mesmos processos – recorrente e recorrido.

Inicialmente o programa foi instruído a inserir automaticamente a frase de pesquisa alvo: "'Recurso Especial' (nome da agência reguladora) não embargo não agravo", a qual indica para o site que o usuário deseja obter os recursos especiais que não tenham natureza de embargo ou agravo e que estejam relacionados à uma agência reguladora de interesse. Posteriormente, assim como se fosse feito por um usuário humano, o programa recebe as informações da página de resultado e navega através dela coletando todas as informações desejadas para a base de dados.

Figura 4 – Pesquisa alvo no site do STJ

Fonte: captura de tela elaborada pelos autores do site do STJ

Figura 5 – Resultado da pesquisa alvo no site do STJ



Fonte: captura de tela elaborada pelos autores do site do STJ

Dessa etapa é importante ressaltar duas variáveis importantes para o processo de construção da base dados: a variável ementa possui o corpo textual do processo, o que permitirá a obtenção da variável referente à decisão do

processo e a variável número do processo permite que cada processo seja identificado unicamente, permitindo consultas e atualizações futuras.

Com as informações em mãos, o programa dá início à coleta dos dados presentes no segundo site. Dado que no primeiro foram obtidos os números de identificação de cada processo, o programa é orientado a buscar estes mesmos números novamente. Quando realizado o processo, o programa recebe todas as informações presentes na página e através de um filtro consegue capturar todas as partes envolvidas no acórdão.

Todo esse processo gera duas bases de dados distintas com uma variável identificadora semelhante, possibilitando a união de ambas por meio de um comando chamado *join*. Tendo em mãos a base unida, o programa parte para a segunda etapa do processo, o tratamento dos dados, o qual é realizado utilizando técnicas de manipulação de textos com a ferramenta *Regular Expressions (Regex)*.

A base de dados conta com a ementa dos acórdãos do STJ. Nela, estão informações relevantes, como se o processo foi provido, não provido ou parcialmente provido, que significa, respectivamente, se a decisão do ministro relator do STJ está de acordo, não acordo ou parcialmente de acordo com o juiz da instância anterior. Assim, para coletar esses dados de cada caso foi preciso utilizar o *Regex*, uma vez que foi necessário coletar poucas palavras em um grande texto, no caso, a ementa do acórdão.

Figura 6 – Exemplo de ementa

EMENTA

ADMINISTRATIVO E PROCESSUAL CIVIL. **RECURSO ESPECIAL.** MULTA ADMINISTRATIVA. VALOR. REDUÇÃO JUDICIAL PARA MONTANTE AQUÉM DO MÍNIMO LEGAL. OFENSA AO PODER DE POLÍCIA. INOCORRÊNCIA. RAZOABILIDADE E PROPORCIONALIDADE.
1. "O Poder Judiciário, no exercício de sua competência constitucional (ex vi do art. 5º, XXXV, da CF/88), pode examinar os atos praticados pela Administração Pública, notadamente no que tange à legalidade ou a sua legitimidade, não havendo que se falar em invasão do mérito administrativo quando o magistrado reduz o valor da multa, com fulcro nos princípios da razoabilidade e proporcionalidade." (AgInt no AREsp 1.067.401/SC, Rel. Ministro Gurgel de Faria, Primeira Turma, DJe 9/8/2018).
2. No caso, a empresa autora, ora recorrida, ajuizou ação de procedimento ordinário objetivando, entre outras providências, a redução do valor de multa a ela imposta pela **ANP,** em virtude da constatação de não observância de normas legais na disposição de recipientes de gás.
3. A Corte regional, por sua vez, confirmou a sentença apelada, no que esta reduziu o valor da sanção pecuniária, invocando, para tanto, critérios de razoabilidade e de proporcionalidade, em conformidade com entendimento deste Superior Tribunal de Justiça.
4. Não há falar em ofensa ao poder de polícia da **ANP,** como aventado nas razões recursais, senão que, atento às peculiaridades do caso concreto, o julgador, pela perspectiva da razoabilidade e da proporcionalidade, não vislumbrou compatibilidade entre a infração glosada pela autoridade fiscalizadora e o elevado quantum da multa aplicada.
5. **Recurso especial** não provido.

Fonte: RECURSO ESPECIAL Nº 1.766.116 - RS (2017/0124424-0)

O método conhecido como expressões regulares consiste em buscar padrões de caracteres em *strings* (variáveis de texto). Dessa forma, a fermenta possibilita criar padrões que auxiliam a combinar, localizar e gerenciar um texto, expressão ou palavras.

No caso do trabalho, bastou encontrar um padrão para as palavras que se referem à um provimento do processo, depois à não provimento e provimento parcial. As expressões strings de *REGEX* encontradas são:

Para não provimento, "([Nn][ãa]o |[Dd]es|[Ii]m|[Nn]egar |NÃO |NEGAR)(provido|conhecido|provimento|PROVIDO|PROVIMENTO)(s)?". Já para provimento "(parcial(mente)? provi(do|mento)|provido(s)? em parte)".

Figura 7 – Exemplo de ementa sendo filtrada pelo Regex

Fonte: captura de tela elaborada pelos autores do site https://regex101.com/

Existe uma dificuldade em filtrar apenas provimento dado que este podia ser seguido da palavra "parcial" ou precedido do advérbio "não" e, mesmo utilizando mais filtros, devido ao fato de que são várias ementas e cada ministro possuía um modo de escrever, poderiam ocorrer erros durante a coleta. Uma forma encontrada para resolver essa situação foi fazer uso de uma condicional: primeiro era buscado o não provimento, se não fosse encontrado

seria iniciada uma busca pelo provimento parcial, para, finalmente, caso não fosse nenhum dos dois casos, o acórdão teria sido provido.

Dessa forma, foi possível encontrar os dados requeridos em todas as ementas referentes aos acórdãos do STJ. Unindo então a etapa de coleta e tratamento, tem-se a base completa elaborada por esse trabalho.

4. Conclusão

Com a coleta e o tratamento da base, mais especificamente filtrando informações relevantes da variável ementa, chegou-se a uma base com 680 informações, tendo como variáveis o número do acórdão, estado de origem, parte recorrente ou aquele que recorre sob a decisão da instância anterior, parte recorrida, data de julgamento, nome do ministro relator e grau de provimento.

5. Limitações

As dificuldades encontradas no trabalho se restringem ao uso de *Regex*, uma vez que não foi possível encontrar um padrão para a palavra provimento que não evidenciasse e encontrasse palavras referente aos outros graus de provimento. Ainda, devido a variação de maneira de escrita entre os ministros e entre os casos foi minimizada a possibilidade de coleta de outras informações relevantes da emente.

6. Referências bibliográficas

- "O que é uma sentença ou acórdão válido?", Escola Brasileira de Direito. Disponível em: https://ebradi.jusbrasil.com.br/artigos/422151030/o-que-e-uma-sentenca-ou-acordao-valido
- "A Ciência de Dados na área do Direito: Novos inputs para a gestão empresarial", Alexandre Zavaglia. Disponível em: https://www.thomsonreuters.com.br/pt/juridico/blog/a-ciencia-de-dados-na-area-do-direito-novos-inputs-para-a-gestao-empresarial.html#:~:text=Al%C3%A9m%20da%20organiza%C3%A7%C3%A3o%20dos%20dados,s%C3%A3o%20despadronizados%20e%2C%20assim%2C%20n%C3%A3o
- "Poder Judiciário - A terceira instância - Julgamento de casos polêmicos", Érika Finati. Disponível em: https://educacao.uol.com.br/disciplinas/cidadania/poder-judiciario---a-terceira-instancia-julgamento-de-casos-polemicos.htm
- "E o regex? Como vai?", Antônio Marcos dos Santos da Rosa. Disponível em: https://cwi.com.br/blog/e-o-regex-como-vai/

Modelagem preditiva de *churn*

Integrantes: Felipe Catapano, Julia Brown e Paulo Kim

Orientador: Matheus Damasceno

1. Objetivo e motivação

O projeto foi desenvolvido em parceria com a empresa Mottu, uma startup que permite que seus clientes aluguem motos a partir de diferentes planos de assinatura. Seus principais clientes consistem em entregadores de aplicativos como Rappi, Ifood, etc. O orientador do projeto, Matheus Damasceno, atua na área de estratégia de negócios e dados da empresa. O *churn*, uma métrica referente à quantidade de clientes que cancelaram o serviço de uma empresa em um determinado período de tempo, é sempre um motivo de preocupação, sendo um valor que deve ser acompanhado de perto.

Com isso, o orientador sugeriu que fosse feita a modelagem preditiva do *churn* da empresa, com o intuito de, inicialmente, entender os dados e as variáveis que poderiam afetar essa métrica, para finalmente poder-se construir modelos que poderiam prever se um cliente iria cancelar o serviço da empresa no futuro ou não. Para isso, houve uma etapa de análise exploratória dos dados onde algumas hipóteses foram testadas, e depois partiu-se para a construção dos modelos, onde foi utilizado também um algoritmo de *Feature Selection*.

2. Base de Dados

Os dados disponibilizados pela startup continham diferentes informações sobre seus clientes – dessa forma, eram todas organizadas a partir do ID dos usuários - sendo que as principais bases enviadas foram:

- Histórico de locações: informações sobre a locação das motos, como: data de início e fim, plano do usuário, região, etc.
- Histórico de pagamentos: dados sobre as diferentes parcelas omitidas: ID da parcela, tipo de plano, valor do aluguel, se há alguma multa de trânsito ou manutenção a ser paga, data de vencimento, data de pagamento, etc.
- Histórico de parcelamentos: caso algum cliente leve uma multa de trânsito ou precise pagar por alguma manutenção da moto, a Mottu possui um sistema de parcelamentos o qual permite que seus clientes parcelem esse acréscimo no seu pagamento. Essa base possui informações sobre esses parcelamentos, como: valor do acréscimo no pagamento, quantidade de parcelas, data de início e vencimento, etc.
- Histórico de clientes: dados sobre os clientes, como: idade, data da primeira CNH, data de validade da CNH, local da residência

- Histórico de multas
- Histórico de manutenções

As bases possuíam dados desde o início de 2020, quando a Mottu começou a operar, até aproximadamente março de 2022.

3. Análise exploratória dos dados

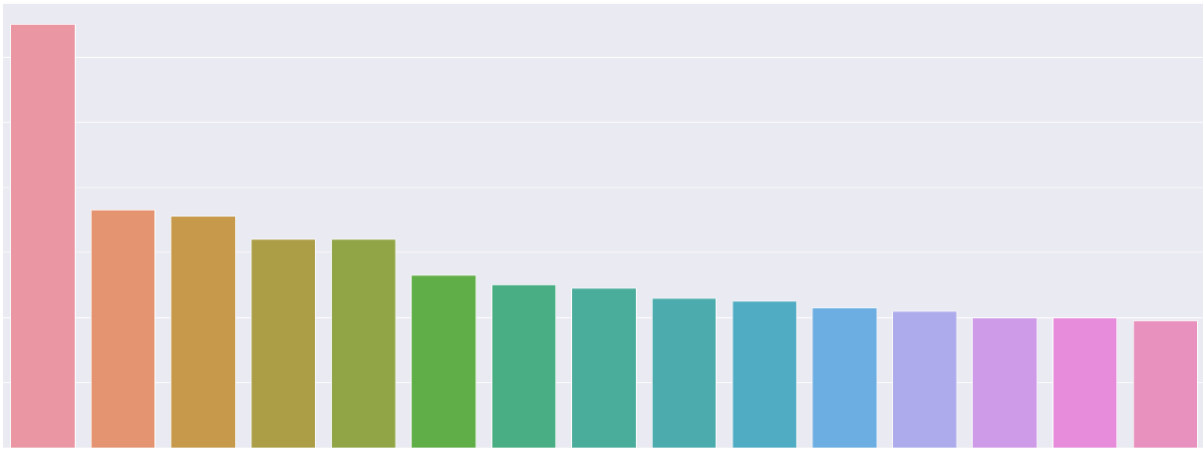
O primeiro passo do projeto consistiu em explorar as bases e construir algumas hipóteses sobre como as diferentes hipóteses impactavam o *churn*. Com isso, foi possível direcionar a análise exploratória dos dados, a qual teve como objetivo, portanto, testar a validez das hipóteses.

É importante ressaltar que essa etapa foi dividida em três vertentes, sendo que cada uma focou em diferentes variáveis:

- Variáveis demográficas dos clientes: localização da residência, idade e data da primeira CNH.
- Variáveis sobre os diferentes planos: *churn* e reviews do aplicativo.
- Variáveis sobre o sistema de parcelamentos: acréscimo no parcelamento.

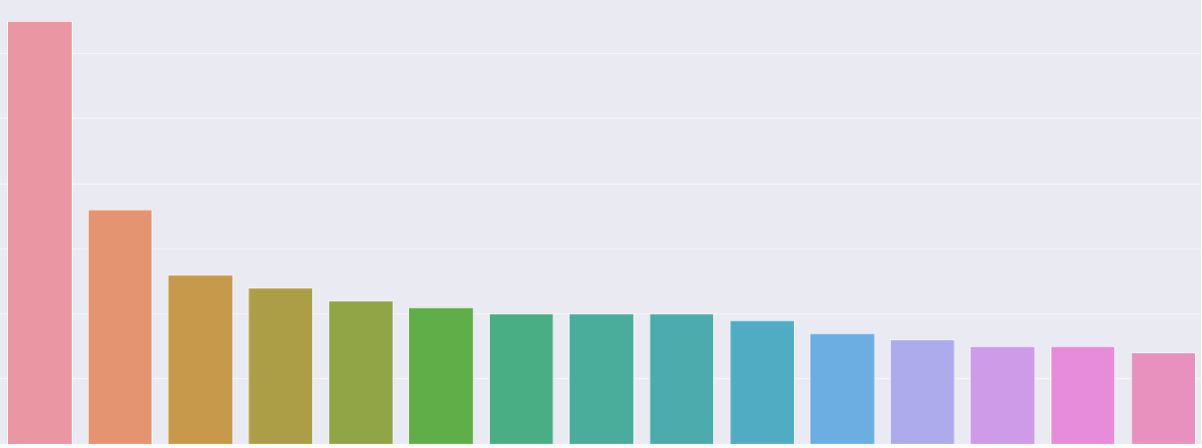
Em relação as variáveis demográficas, a primeira análise tinha como objetivo determinar se havia algum tipo de correlação entre *churn* e localização, isto é, se em alguma região era possível encontrar um número maior de usuários que cancelaram o plano. Para isso, foram feitos diversos gráficos para entender a distribuição dos clientes, incluindo *heatmaps,* como mostram as imagens 1, 2 e 3:

Figura 4 - Distribuição dos clientes em diferentes bairros



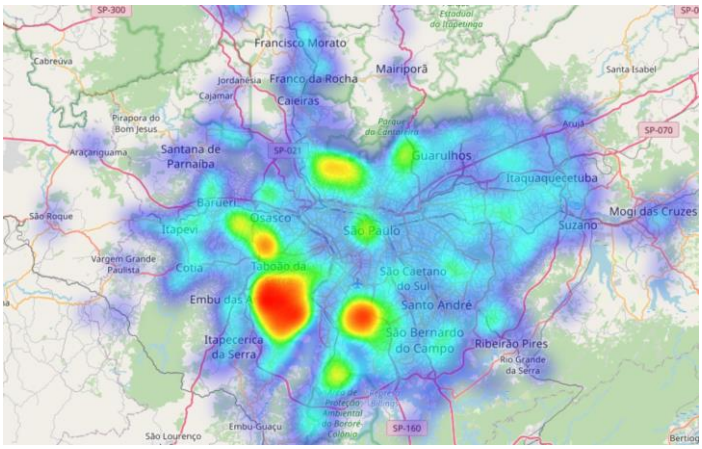
Fonte: elaborado pelos autores

Figura 5 - distribuição dos clientes que cancelaram o plano em diferentes bairros



Fonte: elaborado pelos autores

Figura 6 - *heatmap* da localização dos clientes



Fonte: elaborado pelos autores

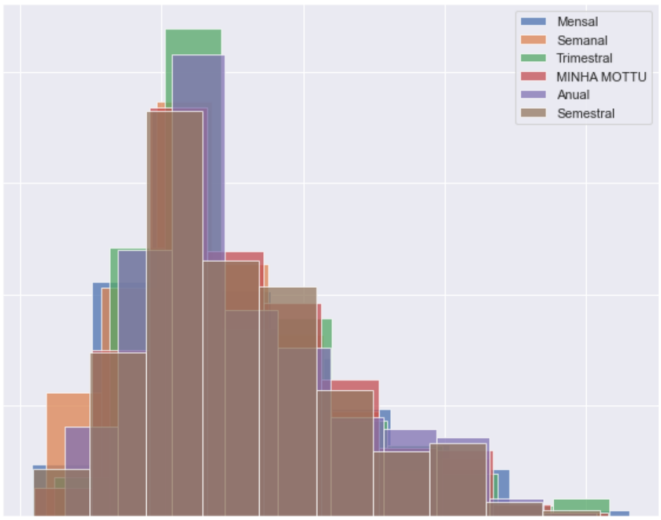
Ademais, também foram feitas análises que buscavam entender se havia alguma correlação entre a idade dos motoristas e a variável *churn* e, por fim, se a data da primeira habilitação impactava o cancelamento. A razão para essa última hipótese vem do fato de que, para motoristas que ainda estão em sua primeira CNH, caso eles recebam uma multa gravíssima, grave ou mais de uma média, eles perdem sua habilitação provisória e precisam recomeçar o processo, o que levaria esses clientes a cancelarem o serviço.

Não se obteve nenhum resultado significante para as duas primeiras hipóteses feitas a partir da análise descritiva.

Na segunda vertente da análise, o principal objetivo era entender como a variável *churn* se comportava de acordo com os diferentes planos, além de

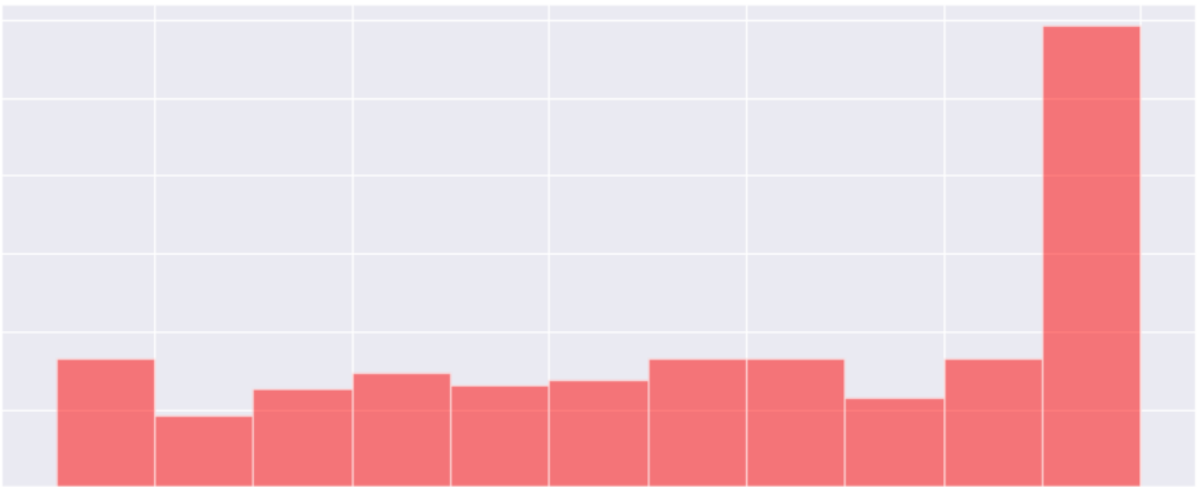
entender algumas outras características dos clientes que saíram. Os gráficos a seguir (imagens 4 e 5) exemplificam algumas análises feitas:

Figura 7 - distância dos clientes que cancelaram o plano até a Mottu



Fonte: elaborado pelos autores

Figura 8 - dispersão da perda de clientes ao longo do ano para o plano Semanal



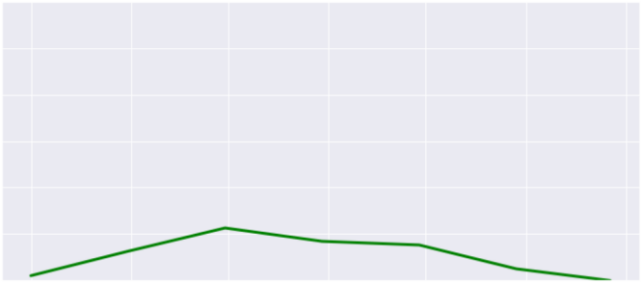
Fonte: elaborado pelos autores

Um resultado importante que pode ser observado foi que, na análise da dispersão dos clientes ao longo do ano, em todos os planos notou-se uma maior perda no final do ano.

A outra análise feita nessa vertente utilizou *web scrapping* para explorar as reviews deixadas nos aplicativos da Mottu, analisando comentários positivos e a permanência dos usuários.

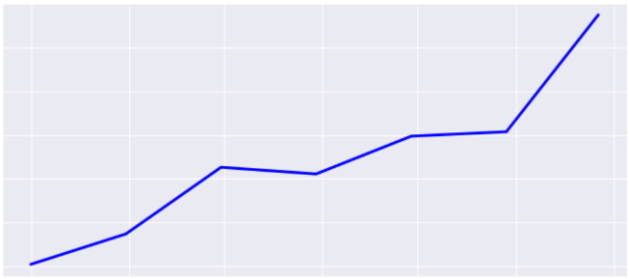
Por fim, a vertente do sistema de parcelamentos tinha como principal objetivo procurar uma correlação entre o *churn* e o aumento do acréscimo no pagamento dos clientes. A hipótese por trás da análise era que, mesmo com a possibilidade de parcelar, pagar multas ou manutenções seria muito difícil para alguns clientes, o que os levava a cancelar os planos. Nesse caso, foi calculado o saldo futuro que os clientes precisavam pagar nos próximos 30 dias para o intervalo de data de 08/2021 (quando o sistema de parcelamentos iniciou) até o final de 02/2022. Depois foi feita uma média mensal (figuras 6 e 7) e essa foi comparada com o *churn rate*.

Figura 9 - média clientes que deram churn



Fonte: elaborado pelos autores

Figura 10 - média clientes que não deram churn



Fonte: elaborado pelos autores

4. Modelagem

Como o objetivo do projeto é prever se o usuário cancelará o plano no futuro próximo ou não, o tipo de modelo que se utiliza aqui é um classificador. Dessa forma, foram escolhidos cinco modelos: Naive Bayes, Regressão Logística, SVM, AdaBoost e Random Forest.

Para preparar a construção dos modelos e montar as bases necessárias, foi adotada uma estrutura de dados centrada nos clientes, deixando as *features* (variáveis) relevantes facilmente acessíveis, incluindo eventos de *churns*,

pagamentos, parcelamentos, manutenções e multas para cada cliente, e ordenando os eventos por tempo. Com isso, os dados foram reorganizados de forma a tornar a ocorrência de *churn,* ou não, a variável alvo. Os seguintes fatores foram incluídos, tendo em mente os resultados das hipóteses testadas na etapa de análise exploratória dos dados:

- De cliente: idade, se CNH é provisória, e se é da capital de São Paulo
- De plano: Qual é, permanência, sazonalidade
- De pagamento: Quantos pagamentos e parcelamentos, gasto com multas e manutenções, pontos na CNH e desconto

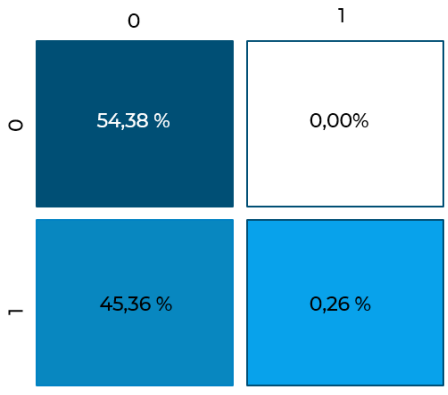
Feito isso, os modelos foram construídos e, após uma avaliação inicial, seus hiperparâmetros foram otimizados a partir do método de *Grid Search* disponibilizado na biblioteca do scikit-learn. Hiperparâmetros se referem, basicamente, aos *settings* de um modelo, os quais podem ser ajustados de acordo com as necessidades do usuário. A sua otimização se' 23leciona um conjunto de parâmetros que gera um melhor resultado de acordo com o objetivo dos criadores do modelo.

Outro método utilizado no processo de modelagem foi o algoritmo *Boruta,* de *Feature Selection*. Esse algoritmo analisa todas as variáveis selecionadas e escolhe as mais importantes para o problema a ser solucionado, definindo, então, quais variáveis devem entrar no modelo.

5. Resultados e conclusões

As imagens 8, 9, 10, 11 e 12 mostram os resultados dos modelos criados, sem a inclusão da ferramenta de *Feature Selection*.

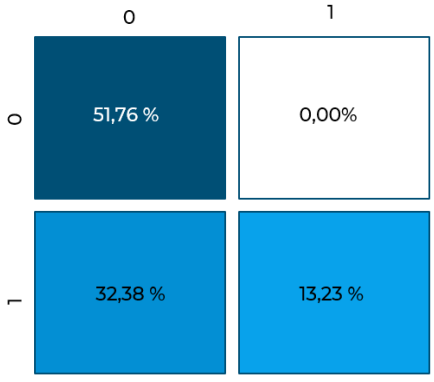
Figura 11 - resultado do AdaBoost



Acc: 55%

Fonte: elaborado pelos autores

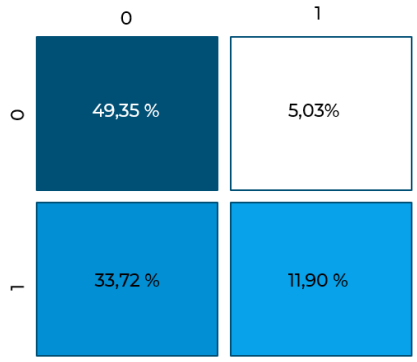
Figura 12 - resultado da Regressão Logística



Acc: 65%

Fonte: elaborado pelos autores

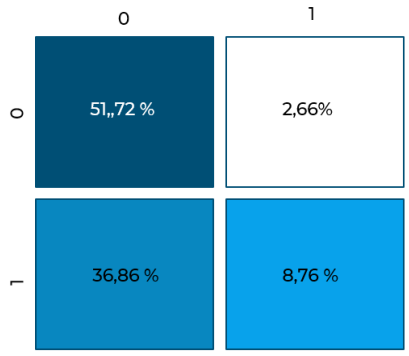
Figura 13 - resultado do Naive Bayes



Acc: 61%

Fonte: elaborado pelos autores

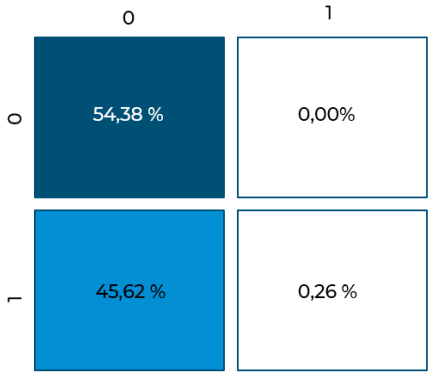
Figura 14 - resultado do Random Forest



Acc: 60%

Fonte: elaborado pelos autores

Figura 15 - resultado do SVM



Acc: 54%

Fonte: elaborado pelos autores

Para compreender os resultados, deve-se ter em mente que os valores 0 e 1 – ou seja, não deu *churn* e deu *churn*, respectivamente - na parte superior da matriz (horizontal) referem-se ao que o modelo adivinhou. A parte esquerda da matriz (vertical), referem-se aos resultados reais.

Ao analisar as matrizes de confusão, nota-se que todos os modelos tem uma tendência a adivinhar 0, ou seja, eles estão "viciados". Isso compromete severamente a efetividade deles, e permite uma conclusão de que os modelos montados não são capazes de prever o fenômeno do *churn* corretamente.

O grupo acredita que a razão para a falta de efetividade dos modelos pode estar relacionada a diferentes problemas como: montagem errônea ou incompleta da base de treino, falta de profundidade no entendimento do problema, outras variáveis que poderiam descrever o fenômeno, falta de entendimento completo da vida real em relação aos dados.

Vale-se mencionar que a inclusão do algoritmo de *Feature Selection* não melhora os resultados. Entretanto, isso não significa que não é possível retirar

conclusão importantes dessa ferramenta. A imagem 13 mostra os resultados do *Boruta*, onde é informado quais variáveis devem ser mantidas no modelo ou não. A partir dela, observa-se que variáveis referentes a CNH, pagamentos, sistema de parcelamentos (multas e manutenções), permanência e data são importantes para o modelo, o que contribui para confirmar as hipóteses iniciais construídas e para entender o que afeta o *churn*.

Figura 16 - resultado do algoritmo *Boruta* de *Feature Selection*

Fonte: elaborado pelos autores

Em suma, apesar dos problemas encontrados com os modelos construídos, ainda foi possível tirar conclusões importantes sobre o problema, entendendo melhor quais *features* impactam o *churn* e como elas o fazem. Esse resultado permite uma conclusão parcial do objetivo inicial do projeto, e é fundamental para a gestão e o planejamento estratégico da empresa, a qual deve ter essas variáveis em mente ao pensar em como reduzir essa métrica.

Projeto realizado em parceria com a GCB - Investimentos

Análise de crédito

Integrantes: Adney Costa, Antoniao Ehler e Eduardo Araujo

Orientador: Daniel Ferreira

1. Objetivo e Motivação

Este projeto foi desenvolvido em parceria com a empresa GCB - Investimentos, holding especializada no mercado financeiro e de capitais, com foco em investimentos, consultoria financeira, securitização, antecipação de recebíveis e tokenização de ativos ilíquidos. No decorrer do projeto foram utilizadas diversas ferramentas de programação, não necessariamente relacionadas ao estudo da modelagem, como construção de API`s em Flask, considerações para um modelo microeconômico de inadimplência e análise descritiva de variáveis da própria empresa.

A motivação principal do grupo foi vivenciar uma experiência profissional e de grande valor curricular, além de preservar e construir o nome do Insper Data entre as empresas para que cada vez mais oportunidades como essa apareçam para os novos membros.

Figura 17 - Logo GCB - Investimentos

2. Contato com a empresa

O contato com a empresa foi realizado por intermédio da atual presidente na época, Yasmin Bocatto, que organizou e esclareceu todos os pontos. A iniciativa para essa parceria foi tomada pela própria empresa que

conhecia o Insper por um projeto realizado com outra organização estudantil, a Blockchain Insper, e que havia feito um processo de tokenização para eles.

Feito o primeiro contato de interesse, foi introduzido o funcionamento do Insper Data por nossa parte, e nos foi apresentado o interesse deles por um grupo de modelagem preditiva. Havia dois grupos de modelagem na época e ambos com trabalhos sendo realizados com empresa, como o outro grupo já estava conversando com a Mottu – Aluguel de motos, nosso grupo foi o escolhido para assumir o projeto.

Após as apresentações e escolhido o grupo, nós conversamos mais detalhadamente sobre a proposta de parceria que eles tinham. Escolhemos trabalhar com um modelo de predição de inadimplência para clientes que adiantavam recebíveis. Feita a escolha, passamos para a parte de conclusão onde formalizamos o acordo por meio de contratos e visita a sede da empresa.

3. Estrutura do projeto

O projeto tinha como objetivo principal prever por métodos de *Machine Learning* clientes inadimplentes, porém não foi bem estruturado pela empresa como seria feito o processo.

Ficamos sob responsabilidade do Daniel Ferreira, coordenador da área de Data Science da Adiante Recebíveis, uma das empresas que compõem a holding GCB. Foi estabelecido que deveríamos comparecer presencialmente na empresa pelo menos duas vezes na semana, o que precisou ser organizado entre os membros do grupo pois era necessário alocar um horário vago na grade horária do Insper simultaneamente para os três. Feito isso, nós decidimos ir as terças e quintas das 9:40 às 13:00, o que impossibilitou a participação nas reuniões semanais da entidade.

4. Rotina na Empresa

Dado que o projeto não foi bem estabelecido e o Daniel não estava muito preparado para a tarefa de liderança, nós não focamos muito no projeto estabelecido previamente que seria a predição de clientes inadimplentes, mas acabamos sendo incluídos na rotina da empresa o que nos possibilitava trabalhar com desafios diários não necessariamente voltados para Data Science.

Focamos principalmente em três frentes que foram trabalhadas no decorrer do projeto e podem ser listadas como:

- Análise descritiva das *features* para o modelo

Nessa etapa nós utilizávamos uma ferramenta comprada pela empresa que era o BigDataCoorp, onde nós conseguíamos consultar informações públicas sobre empresas e pessoas através apenas do CNPJ ou CPF. Nós criamos relatórios embasados em estudos sobre as variáveis que acreditávamos serem interessantes para a construção do modelo. Aprendemos a realizar consultas via API e entender a importância de uma variável para um modelo.

- Análise microeconômica para um modelo de inadimplência

Paralelamente com a análise descritiva das variáveis nós construíamos um modelo com base em fatores microeconômicos relevantes para a predição de inadimplência. Conseguíamos validar as teorias com outros funcionários da empresa.

- Construção de API em Flask para uso interno da empresa

Mais para o final do projeto foi nos proposto o desafio de construir uma API para uso interno onde os funcionários poderiam enviar informações e rodar modelos de predição apenas por meio dessa API. Para a construção dessa ferramenta foi necessário o estudo do *framework* Flask.

Com isso nós finalizamos o nosso projeto sem o entregável de um modelo de predição, mas com mini tarefas diárias que ajudaram a empresa no decorrer do semestre.

5. Considerações Finais

Apesar da falta de um entregável foi uma experiência agregadora profissionalmente, porém acredito que seja necessário elaborar um padrão de apresentação do Data para empresas que desejam realizar projetos em parceria. Nessa apresentação deve ser mostrado o calendário do Data e como nossa proposta de projeto com um entregável se encaixa na proposta da empresa.