DOI: 10.48011/asba.v2i1.1721

Uma metodologia orientada a dados sociodemográficos para predição de preços do Uber X

Jefferson Silva * **Luciana Lima** ** **Ivanovitch Silva** *

* *Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, RN (e-mail: jefferson023@ufrn.edu.br, ivan@imd.ufrn.br)*
** *Departamento de Demografia e Ciências Atuariais, Universidade Federal do Rio Grande do Norte, RN (luciana.lima@ccet.ufrn.br)*

Abstract: The urbanization process and the technological development encourages the search for new ways of mensure the urban quality of life. Researches in this field with Uber as data source suggest that Uber waiting time is related to socioeconomic characteristics of the cities. With the aim of verify if Uber price is related to socioeconomic characteristics of pickup location as well, this work study the city of Natal in Rio Grande do Norte and that represent a place where there's a pent-up demand for good quality public transportation. To complete this goal, was collected price data from Uber X service for this city during all 2018 year, besides socioeconomic data aggregated at Human Development Unities (HDUs) level provided by Atlas do Desenvolvimento Humano no Brasil. As methodology was used machine learning techniques to create data oriented regression models. Regression analysis over these models revealed that socioeconomic characteristics of Natal is related to Uber X price data.

Resumo: O avançado processo de urbanização e desenvolvimento tecnológico possibilita a busca por novas formas de avaliar a qualidade de vida no meio urbano. Pesquisas na área e que utilizam a Uber como fonte de dados apontam que o tempo de espera do serviço pode se relacionar com características socioeconômicas das cidades. A fim de se testar a hipótese de que a precificação da Uber relaciona-se às características socioeconômicas dos lugares de embarque do serviço, este trabalho realiza um estudo para a cidade de Natal no Rio Grande do Norte e que representa uma área do Nordeste em que há uma demanda reprimida por transporte público de qualidade. Para atender esse objetivo, foram coletados dados de preços dos serviços do Uber X para essa localidade durante todo o ano de 2018, além de dados socioeconômicos agregados a nível de Unidades de Desenvolvimento Humano (UDH) fornecidos pelo Atlas do Desenvolvimento Humano no Brasil. Como metodologia, empregou-se técnicas de aprendizagem de máquina para a criação de modelos de regressão orientados à dados. Análises de regressão sobre esses modelos revelaram que as características socioeconômicas da cidade de Natal se relacionam com os dados de preço do Uber X.

Keywords: Uber; Data Science; Predictive Analytics; Socioeconomic Indicators; Machine Learning.

Palavras-chaves: Uber; Ciência de Dados; Análise Preditiva; Indicadores Socioeconômicos; Aprendizagem de Máquina.

1. INTRODUÇÃO

O crescimento da população pode induzir um aumento na quantidade de viagens no meio urbano e faz com que as cidades nem sempre consigam se ajustar adequadamente a essa demanda (Aljoufie et al., 2011). Uma tendência emergente para gerenciar e minimizar o impacto dos problemas ocasionados pelo crescimento populacional é a utilização das Tecnologias da Informação e Comunicação (TIC). Esse conceito é conhecido como cidades inteligentes (Ismagilova et al., 2019).

Não existe uma única definição para o que é uma cidade inteligente (Capdevila and Zarlenga, 2015). Entretanto, pode se argumentar que uma cidade é considerada "inteligente" quando o Governo e as organizações investem em capital humano e social, infraestrutura e tecnologias disruptivas com o objetivo de promover um crescimento econômico sustentável e melhorar a qualidade de vida em geral (Obedait et al., 2019). Neste sentido, o Uber pode ser considerado um exemplo de serviço para uma cidade inteligente (Mariano et al., 2019).

Mais que um aplicativo que comercializa serviços de transporte, Wang and Mu (2018) tentaram encontrar relação entre características socioeconômicas de bairros com a acessibilidade dos serviços da Uber. O Estudo utilizou dados do Uber X, UberBlack e dados socioeconômicos da cidade de Atlanta (EUA) para criar modelos espaciais autoregressivos e verificar se fatores socioeconômicos

agregados por bairros influenciavam na acessibilidade do serviço. Por fim, foi concluído que as características socioeconômicas de bairros não tinham uma relação significativa com a acessibilidade dos serviços da Uber, apesar de outras características, como densidade populacional, estarem relacionadas.

Bezerra et al. (2019) também analisou a relação entre o tempo de espera da Uber e características socioeconômicas. O estudo realizou análises sobre dados de tempo de espera do serviço para a cidade de Natal no Rio Grande do Norte, obtendo indicações que essa variável é fortemente relacionado com indicadores de qualidade de vida dos bairros de embarque/desembarque. Além disso, também foi sugerido que os dados da Uber poderiam ser utilizados como indicadores de habitabilidade para cidades e bairros, podendo ainda ser utilizados para o planejamento urbano.

Entretanto, os estudos mencionados, assim como a maioria, pouco tem explorado a dimensão do preço na relação entre o serviço ofertado pela Uber e características socioeconômicas dos locais de embarque e/ou desembarque. Portanto, este trabalho tem como objetivo analisar esse relacionamento a partir dos locais de embarque do serviço da Uber. Para cumprir esse objetivo, foram coletados dados de preço do Uber X e dados sociodemográficos para a cidade de Natal. Esses dados foram utilizados para a construção de modelos preditivos usando técnicas de aprendizado de máquina que, no fim, foram submetidos a análises de regressão para obter os resultados.

Este trabalho se subdivide em mais três seções. Na primeira é descrito todo o processo metodológico realizado até a construção dos modelos preditivos para análise de regressão. A segunda descreve todos os resultados encontrados através das análises de regressão nos modelos de regressão linear e não linear. Por fim, a última discute os resultados encontrados, as limitações do trabalho e aponta trabalhos futuros.

2. METODOLOGIA

Nesta seção é descrito, conforme pode ser visto na Figura 1, as etapas realizadas e a metodologia empregada para coleta dos dados de preço da Uber e dados socioeconômicos. Além disso, também é descrito o processo de análise de dados, limpeza e tratamento dos dados e, por fim, o processo de construção dos modelos preditivos para a análise de regressão.

Figura 1. Diagrama de Metodologia.

2.1 Dados de Preço do Uber X

Os dados de preço do Uber X foram obtidos a partir da Interface de Programação de Aplicações (API) de simulação de preço, disponibilizada no site da empresa [1]. Essa API tem como objetivo fornecer uma estimativa de preço mínimo e máximo a partir de duas localidades fornecidas: origem e destino.

A cidade escolhida para coleta dos dados foi a cidade de Natal, localizada na Região Nordeste do Brasil e capital do estado do Rio Grande do Norte. O transporte público informal é o meio de mobilidade urbana predominante na região Norte e Nordeste (IBGE, 2017). Além disso, a quantidade de novos veículos em circulação no estado do Rio Grande do Norte mais que dobrou em apenas uma década (DENATRAN, 2018). Tudo isso aponta para a existência de uma demanda reprimida por serviços de transporte na Região Nordeste e que abre espaço para meios alternativos que são mais acessíveis a mobilidade da população, o que justifica a escolha de Natal como objeto do presente estudo.

Os dados do Uber X foram coletados por uma aplicação em uma infraestrutura dedicada para essa tarefa durante todo o período de coleta. Essa aplicação era responsável por a cada 10 minutos, em paralelo, simular o preço de viagens do Uber X para cada um dos 36 bairros da cidade de Natal. Conforme descrito na Figura 2, para cada bairro selecionado, 10 coordenadas foram escolhidas com base nos seguintes critérios:

- um conjunto fixo de cinco coordenadas de interesse, como hospitais, shoppings, escolas, instituições públicas e etc.
- um conjunto fixo de quatro coordenadas aleatórias que não estejam localizadas sobre nenhum corpo d'água.
- o centroide geográfico do bairro.

O conjunto de dados resultante totalizou cerca de 56 milhões de linhas, coletados no período entre janeiro de 2018 e dezembro do mesmo ano. Em razão do poder computacional que seria necessário para processar esses dados e visando uma melhor otimização do processo de criação do modelo, optou-se por utilizar apenas 10% dos dados de cada bairro. O resultado final foi um conjunto com aproximadamente 5,5 milhões de linhas e, como pode ser visto na Tabela 1, oito colunas no total.

2.2 Dados Socioeconômicos

Como os dados do Uber X não dispõem de informações sociodemográficas dos usuários do serviço, um meio alternativo de obter essas informações é utilizar as características do local de embarque/desembarque das chamadas do serviço. Neste estudo, foram utilizadas somente as informações referentes ao local de embarque. Para isso, utilizou-se dados das Unidades de Desenvolvimento Humano (UDH) fornecidos pelo Atlas do Desenvolvimento Humano no Brasil. Essas informações se baseiam no Censo Demográfico brasileiro (PNUD, IPEA e FJP, 2017).

O Censo Demográfico é realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e representa a principal fonte de referência para as condições de vida da população brasileira em todos os municípios do país (IBGE, 2010).

[1] Disponível em: <https://developer.uber.com/> Acesso em: 12 junho, 2020.

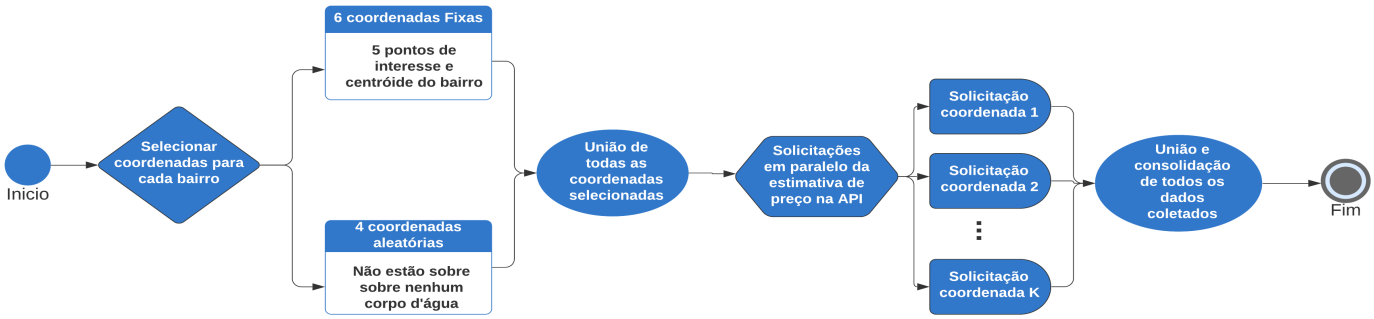

Figura 2. Diagrama do processo de coleta de dados da Uber.

Tabela 1. Amostra da base de dados de preço do Uber X.

coordenada origem	coordenada destino	bairro origem	bairro destino	udh origem	udh destino	data e hora	preço médio Uber X
(-35.23712, -5.810864)	(-35.20719, -5.850292)	Bom Pastor	Capim Macio	1240810200046	1240810200019	2018-01-26 15:47:16	17.0
(-35.235188, -5.809281)	(-35.20729, -5.816953)	Bom Pastor	Lagoa Nova	1240810200046	1240810200020	2018-01-25 06:44:13	10.5
(-35.243916, -5.809263)	(-35.20719, -5.850292)	Bom Pastor	Capim Macio	1240810200046	1240810200019	2018-01-25 16:09:17	18.5
(-35.245975, -5.8088417)	(-35.19499, -5.810603)	Bom Pastor	Tirol	1240810200046	1240810200001	2018-01-31 04:37:56	15.0
(-35.243862, -5.812955)	(-35.185234, -5.795018)	Bom Pastor	Mãe Luiza	1240810200046	1240810200003	2018-01-28 06:46:21	20.5

O Censo Demográfico fornece os dados socioeconômicos agregados em um conjunto de áreas contíguas denominado setor censitário. Enquanto as UDHs foram construídas buscando agregar os dados de forma que gerem áreas mais homogêneas e que captem melhor condições socioeconômicas (PNUD, IPEA e FJP, 2017). Contudo, tanto o Censo Demográfico quanto as UDHs possuem uma periodicidade decenal, o que implica o uso de dados para a última década e que pode representar uma limitação quanto à atualidade de suas informações.

As variáveis socioeconômicas selecionadas, como pode ser visto na Tabela 2, se dividiram em três categorias: características de condições de vida, características de atividade econômica e proxy de demanda do serviço do Uber X. As variáveis que expressam as condições de vida da população foram o Índice de Gini e o IDHM (Índice de Desenvolvimento Humano Municipal). A variável que expressa as características de atividade econômica foi a taxa de pessoas que exerciam atividades econômicas entre 25 e 29 anos. Por fim, as variáveis que podem representar um proxy de demanda do serviço do Uber X foram a densidade demográfica e o percentual de pessoas em domicílios vulneráveis à pobreza que gastam mais de uma hora até o trabalho.

Tabela 2. Variáveis sociodemográficas.

Variável	Definição
Índice de Gini	Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Assume o valor 0 quando não há desigualdade e tende a 1 à medida que a desigualdade aumenta.
IDHM	É uma média geométrica das dimensões renda, educação e longevidade, com pesos iguais.
Taxa de atividade econômica	Percentual de pessoas entre 25 e 29 anos de idade que eram economicamente ativas, ou seja, estavam empregadas durante o período de coleta do Censo Demográfico.
Densidade demográfica	Razão entre a população residente total de uma UDH pela sua área em km^2.
Deslocamento de pessoas vulneráveis	Percentual de pessoas que vivem em domicílios vulneráveis à pobreza (com renda per capita inferior a 1/2 salário mínimo de agosto de 2010) e que gastam mais de uma hora em deslocamento até o local de trabalho.

Fonte: Adaptado do dicionário dos indicadores do Atlas do Desenvolvimento Humano no Brasil.

2.3 Análise Exploratória dos Dados

A Análise Exploratória dos Dados (AED) é uma abordagem para análise de dados em que se emprega uma variedade de técnicas gráficas com o objetivo de maximizar o discernimento, estrutura, extração de variáveis mais importantes e outros fatores do conjunto de dados (Natrella, 2010). No conjunto de dados de preço do Uber X, a exploração das características do conjunto de dados acontece por meio de gráficos de suas variáveis quantitativas e mapas coropléticos para analisar relações espaciais.

As UDHs do município de Natal aparentam ser um fator decisivo na determinação do preço do Uber X. Conforme indicado na Figura 3, as 36 UDHs apresentaram uma média de preço diferenciada a partir da localização de origem. As UDHs mais afastadas do centro do município e, principalmente, aquelas que estão localizadas nas zonas mais ao Norte apresentaram uma média de preço mais elevada. Enquanto as UDHs mais próximas do centro do município mostraram médias mais baixas. Isso pode ocorrer por diversos motivos, inclusive o deslocamento de trabalhadores da Zona Norte para as Zonas Centro/Sul devido essas áreas centralizarem as principais atividades econômicas do município.

O horário de uso do serviço também se mostra como um outro influenciador do preço. Conforme visto na Figura 4, existem quatro períodos principais de tempo durante o dia onde ocorrem picos no preço médio do Uber X, atingindo o maior valor entre 17 horas e 18 horas. Contudo, como visto na Figura 5, a diferença entre o preço normal e durante um horário de pico não é bastante significativa.

2.4 Limpeza e Tratamento dos Dados

Uma etapa importante que deve ser realizada antes da criação de modelos preditivos é a limpeza e tratamento

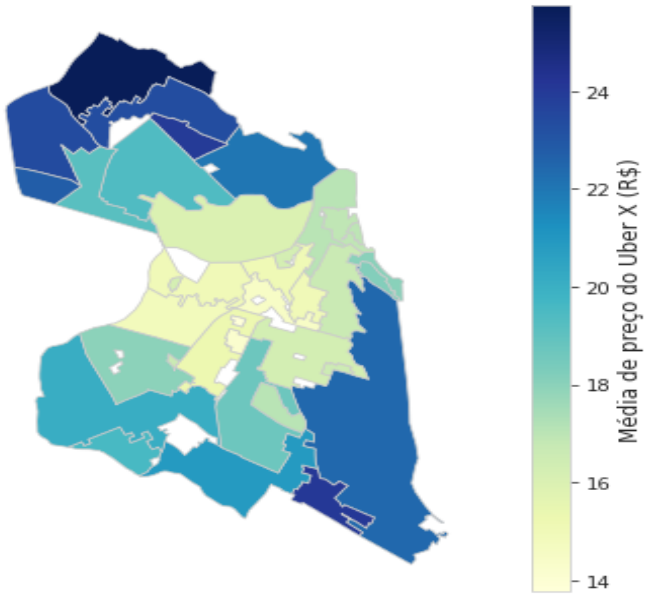

Figura 3. Mapa coroplético indicando a média de preço do Uber X por UDHs.

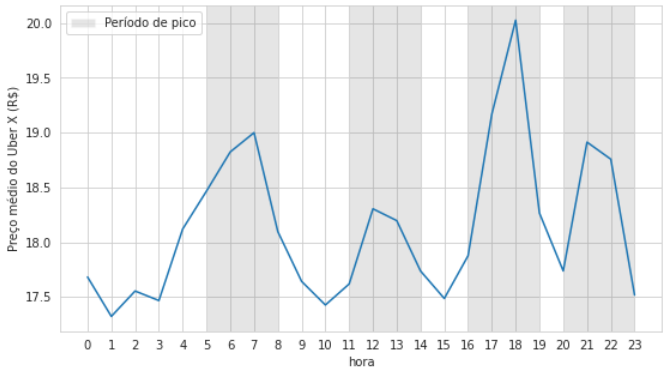

Figura 4. Média de preço do Uber X por hora do dia.

dos dados. A importância desta etapa é justificada por dados sem processamento não serem uniformes e previsíveis (Haughton et al., 2003). Além disso, pesquisas realizadas na área mostraram que existem custos econômicos milionários associados ao uso de dados em estado bruto (Jesmeen et al., 2018).

O primeiro passo realizado foi em relação aos dados faltantes. Além da qualidade de conjuntos de dados incompletos ser questionável, é difícil para muitas técnicas de aprendizado de máquina conseguirem processar, com êxito, dados faltantes (Tsai and Chang, 2016). Quando existe esse tipo de problema, esses dados devem ser removidos por completo ou substituídos por outros gerados a partir dos dados pré-existentes, esse processo é chamado imputação. Entretanto, a remoção ou imputação dos dados pode causar enviesamento dos resultados caso seja utilizado sem o conhecimento do problema que gerou os dados faltantes (van Ginkel et al., 2019; Ramli et al., 2013).

Para os dados do Uber X coletados, os dados faltantes apresentam origens distintas e que necessitam de um estudo aprofundado para descobrir sua causa. Por este motivo e pela quantidade total de dados faltantes repre-

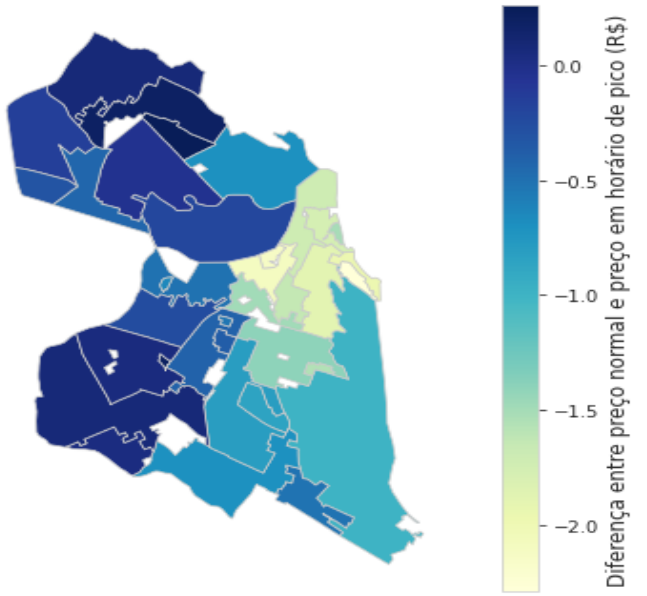

Figura 5. Mapa coroplético indicando a diferença entre o preço normal do Uber X e o preço durante horário de pico, entre 16 horas e 19 horas.

sentarem apenas uma pequena fração de todo o conjunto de dados, como pode ser visto na Figura 6, optou-se por remover todos os dados que faltavam. Outro ponto importante no tratamento de dados é garantir que a presença de *outliers* no conjunto de dados não afete os resultados finais, gerando dados inválidos e que não retratam a realidade.

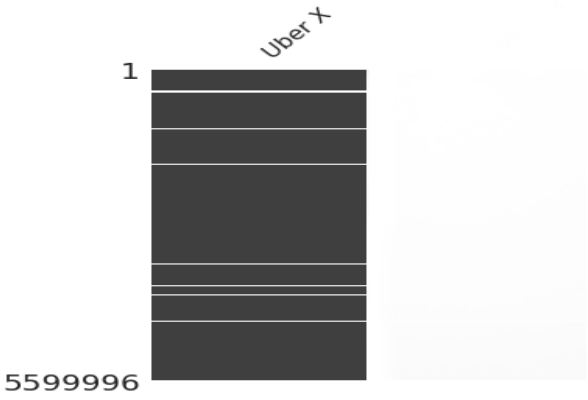

Figura 6. Matriz indicando a dispersão de dados faltantes.

Na estatística, *outliers* ou pontos fora da curva são observações que apresenta um grande afastamento numérico das outras observações. No contexto da base de dados analisada, *outliers* aparecem por meio de preços das viagens simuladas do Uber X com custos extremamente altos, como mostra a Figura 7. Para detectar *outliers* é necessário utilizar algumas técnicas estatísticas. As três técnicas mais utilizadas são: *standard deviation* (SD), *median absolute deviation* (MAD) e *interquartile range* (IQR) (Yang and Rahardja, 2019).

Entre os métodos de detecção de *outliers* apresentados, o escolhido foi o IQR. Essa técnica utiliza quartis ao contrário de métodos que utilizam média ou o desvio padrão e isso a torna menos afetada por *outliers* (Kobalşi

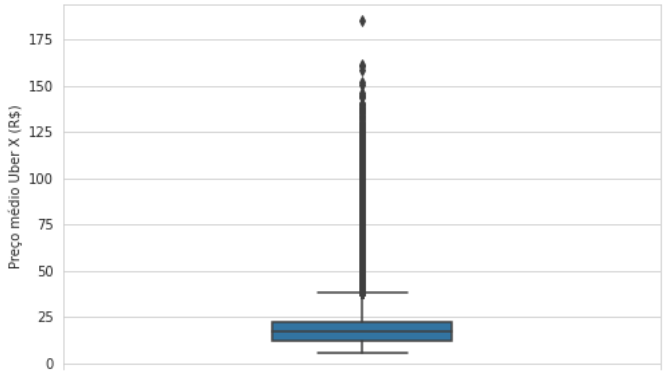

Figura 7. Gráfico de caixa indicando outliers detectados pelo IQR.

and Ünsal, 2019). O IQR funciona com base em dois limitantes *Tmin* e *Tmax* calculados de acordo com as Equações 1, 2 e 3.

$$IQR = Q3 - Q1 \quad (1)$$

$$Tmin = Q1 - 1.5 * IQR \quad (2)$$

$$Tmax = Q3 - 1.5 * IQR \quad (3)$$

Onde *Q1* e *Q3* correspondem, respectivamente, 25% e 75% no conjunto de dados ordenado. Todos os valores que não estão entre *Tmin* e *Tmax* são considerados *outliers* e, posteriormente, esses dados são removidos.

O último passo realizado no tratamento dos dados foi reescalar variáveis numéricas para que não interfiram no modelo preditivo. No conjunto de dados utilizado, a única variável que necessitou dessa transformação foi o código das UDHs de origem e destino já que esse número é composto por 15 dígitos. Entre todas as técnicas de transformação e codificação de variáveis, a codificação ordinal foi escolhida. Essa codificação funciona tanto para dados categóricos quanto dados numéricos e, de modo geral, designa um número inteiro para cada valor ou categoria presente no conjunto de dados. A desvantagem dessa técnica é que ela implica que existe uma ordem entre todas as categorias ou valores da variável (Von Eye and Clogg, 1996, cited in Potdar et al., 2017). Entretanto, conforme foi mostrado na etapa de Análise Exploratória dos Dados, existe de fato uma ordem de preço que varia entre UDHs e que faria esta técnica ideal para esse caso.

2.5 Criação de Modelos Preditivos

Para analisar o relacionamento entre as variáveis socioeconômicas e o preço do Uber X, foi utilizada a técnica estatística de análise de regressão. A análise de regressão acessa as formas específicas de relacionamento entre variáveis, prevendo ou estimando o valor de uma variável de acordo com o valor de outra (Jain et al., 2016). No presente estudo, as variáveis socioeconômicas seriam algumas das variáveis utilizadas para estimar o valor de preço dos serviços do Uber X. Para atingir esse objetivo, foram criados modelos de regressão linear e não linear baseado em Árvores de Decisão utilizando aprendizagem de máquina.

Os modelos foram criados utilizando o algoritmo de *gradient boosting*, fornecido pela biblioteca XGBoost [2] e que está disponível para diversas linguagens de programação. Esse algoritmo foi escolhido devido ao tamanho da base de dados utilizada e que impossibilitou que a maioria dos outros algoritmos conseguissem treinar o modelo de regressão em um período de tempo viável. Para treinar os modelos de regressão, foi utilizado 80% da base de dados como dados de treinamento e os 20% restantes foram utilizados como dados de teste para avaliar o desempenho do modelo.

Por fim, os *hiperparâmetros* utilizados para os modelos de regressão foram selecionados com base nos melhores resultados obtidos no conjunto de dados de treinamento. Para o modelo linear, os *hiperparâmetros max_depth*, *booster*, *eta* e *n_estimators* assumiram, respectivamente, os valores como segue: 7, *'gblinear'*, 0.5 e 80. Já para o modelo não linear, os *hiperparâmetros max_depth*, *booster*, *reg_lambda* e *n_estimators* assumiram, respectivamente, os seguintes valores: 7, *'gbtree'*, 0.6 e 80. Para fins de reprodutibilidade, ambos os modelos utilizaram o *hiperparâmetro random_state* com valor igual a 42.

3. RESULTADOS

O modelo de regressão linear apresentou Erro Quadrático Médio (EQM) de 47.48 e Raiz do Erro Quadrático Médio (REQM) de 6.89 no conjunto de testes utilizado para avaliar o desempenho do modelo. Essas duas medidas indicam o erro do modelo em relação valor que era esperado e são calculadas conforme as Equações 4 e 5.

$$EQM = \frac{1}{N}\sum_{i=1}^{N}(E_i - R_i)^2 \quad (4)$$

$$REQM = \sqrt{EQM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - R_i)^2} \quad (5)$$

Onde N é a quantidade de linhas do conjunto de teste, E e R são, respectivamente, o i-ésimo valor estimado e o i-ésimo valor real de preço do conjunto de dados de teste. O REQM também pode ser interpretado como o desvio padrão do erro. De modo geral, isso significa que a diferença entre o preço estimado pelo modelo linear e o preço real do Uber X é, aproximadamente, de R$ 6,89 a mais ou a menos. Este valor representa um indicativo alto para o modelo linear considerando que o preço das corridas da Uber no conjunto de dados final se concentrava entre R$ 6,00 e R$ 38,00

Os resultados do modelo de regressão linear, conforme a Tabela 3, mostraram que todas as variáveis sociodemográficas do local de embarque, com exceção da taxa de atividade econômica, se relacionaram com o preço do Uber X. O grau e sentido do relacionamento é determinado com base nos coeficientes de regressão, valores muito altos ou muito baixos indicam variáveis que foram significativas para determinação do preço. Enquanto coeficientes bem próximos de zero indicam variáveis que não se relacionaram bem e que tiveram pouca influência no preço. Uma exceção foi a variável de densidade demográfica que, devido a sua magnitude, não foi afetada por um valor de coeficiente próximo de zero. Além disso, a densidade demográfica também foi a única variável com um coeficiente negativo, o que mostra que o aumento de um habitante por km^2 fez com que o preço reduzisse por R$ 0,0001.

[2] Disponível em: <https://xgboost.readthedocs.io/en/latest/> Acesso em: 12 junho, 2020.

Tabela 3. Resultado do modelo de regressão linear.

Variável independente	Coeficiente	Erro Padrão
gini	6.1445	0.125
idhm	4.5552	0.085
intercepto	3.6673	0.095
deslocamento pop vulnerável	1.0748	0.004
hora	0.0769	0.000
mes	0.0569	0.001
atividade 25a29 anos	0.0437	0.001
udh destino	0.0370	0.000
dia	0.0301	0.000
udh origem	0.0237	0.001
densidade demográfica	-0.0001	0.000

Ainda sobre o modelo de regressão linear, o coeficiente das variáveis indicativas de condição de vida mostrou que um aumento de 0.1 no índice de Gini e IDHM fez o preço do Uber X aumentar, respectivamente, cerca de R$ 0,61 e R$ 0,45. Como o índice de Gini indica desigualdade e o IDHM indica qualidade de vida, isso mostra que a desigualdade é um fator que influencia mais para o aumento do preço que a qualidade de vida. Por fim, a variável que mais contribuiu para o aumento de preço do Uber X foi a variável que indicava o deslocamento de pessoas vulneráveis. Mesmo com um coeficiente de regressão menor que as variáveis socioeconômicas de condição de vida, essa variável assumiu valores de 0 à 100. Por este motivo, o aumento de 1% no deslocamento de pessoas vulneráveis fez com que o preço do Uber X aumentasse cerca de R$ 1,07.

O modelo de regressão não linear baseado em Árvores de Decisão conseguiu estimar o preço do Uber X com maior precisão comparado ao modelo linear. O modelo não linear obteve EQM de 11.34 e REQM de 3.25 no conjunto de dados de teste, o que representa uma melhoria de cerca de 52% em relação ao outro modelo. Como mostra a Figura 8, as variáveis socioeconômicas mais importantes para essa melhoria foram a taxa de atividade econômica, as variáveis de proxy de demanda do serviço e o índice de Gini. A importância de variáveis é medida com base no percentual de vezes que cada variável contribuiu para estimar o preço com maior precisão. Isso mostrou que as variáveis socioeconômicas também se relacionaram com o preço para o modelo de regressão não linear. Uma característica que, se mantendo constante em ambos os modelos, serviu para enfatizar ainda mais a importância das variáveis sociodemográficas na estimativa de preço.

Por fim, foi observada a forma como algumas das variáveis socioeconômicas influenciaram as estimativas geradas pelo modelo não linear. Isso foi feito através da técnica de *partial dependence plots* que, para todos os dados no conjunto de treinamento, mostra em gráficos a forma como os valores de uma variável altera o valor médio estimado pelo modelo (Goldstein et al., 2013). Nesse sentido, como mostra a Figura 9, aumentar os valores da variável de taxa de atividade econômica para valores maiores que 72% fez com que a estimativa de preço do Uber X aumentasse em média até 1290%. Da mesma forma, como indica a Figura 10, aumentar os valores da variável de deslocamento de população vulnerável para valores maiores que 1% fez com que o preço também aumentasse em média até 566%. Contraditoriamente a esse resultado, aumentar o Índice

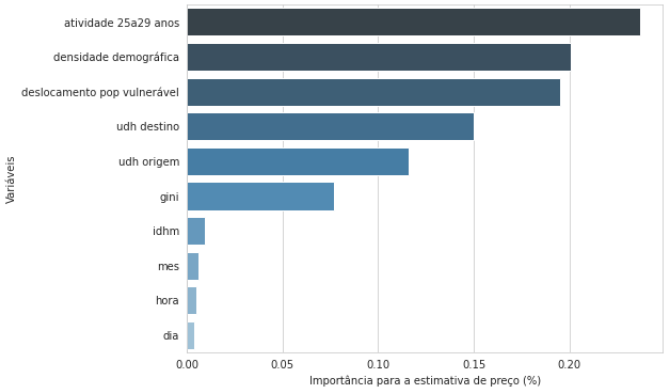

Figura 8. Indicador da importância de variáveis para o modelo de regressão não linear

de Gini para valores maiores que 0.42 e a densidade demográfica para valores maiores que 5000 habitantes por km^2, como mostra a Figura 11 e Figura 12, fez com que o preço estimado fosse reduzido em média até 114% e 195%. Além disso, valores menores que 5000 habitantes por km^2 para a densidade demográfica exibiu um efeito contrário na estimativa de preço, ou seja, de aumento.

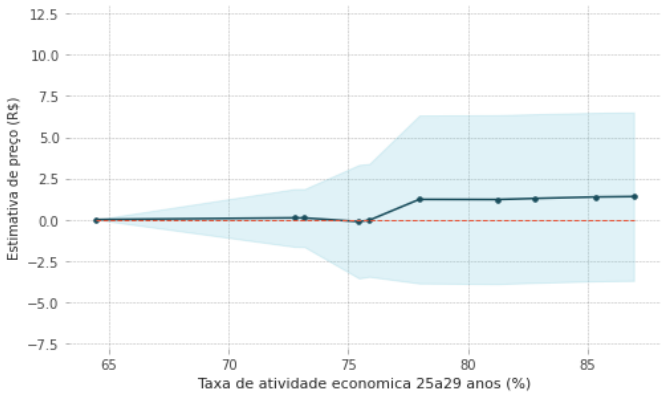

Figura 9. Influência da taxa de atividade econômica na estimativa de preço para o modelo de regressão não linear.

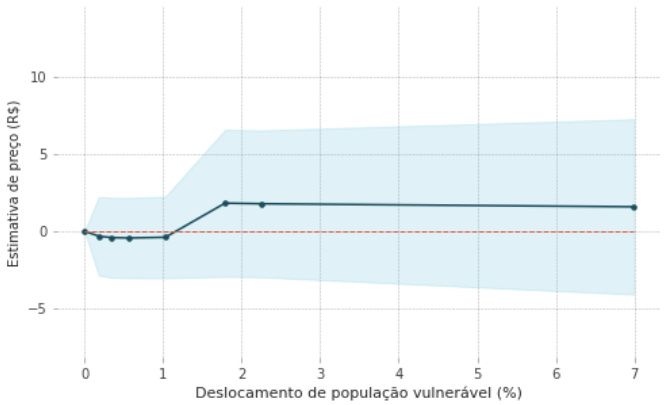

Figura 10. Influência do deslocamento de pessoas vulneráveis a pobreza na estimativa de preço para o modelo de regressão não linear.

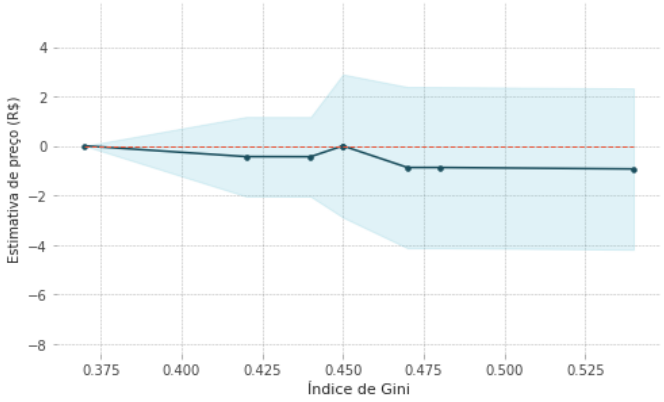

Figura 11. Influência do índice de Gini na estimativa de preço para o modelo de regressão não linear.

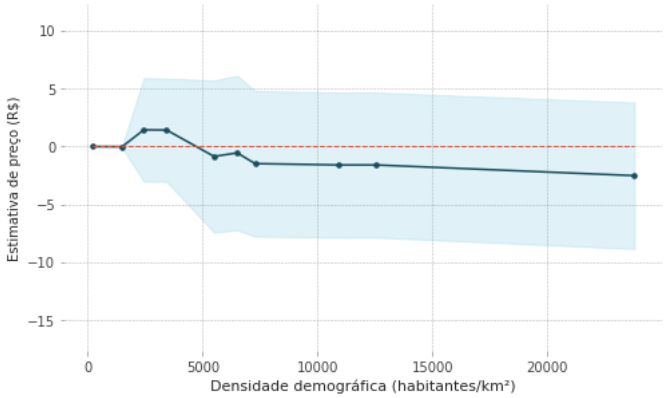

Figura 12. Influência da densidade demográfica na estimativa de preço para o modelo de regressão não linear.

4. CONCLUSÃO

Os resultados mostraram que as variáveis sociodemográficas do local de embarque se relacionaram com o preço do Uber X. Ademais, esse relacionamento se mostrou um relacionamento não linear visto que o desempenho exibido pelo modelo de regressão linear foi inferior ao desempenho do modelo não linear. Com isso, fica evidente que o preço dos serviços do Uber X é influenciado por características socioeconômicas do local de embarque e realça a importância dos dados de preço da Uber como um potencial indicador de características sociodemográficas.

Além disso, a forma como o Índice de Gini influenciou na estimativa de preço mostra a desigualdade como um fator decisivo para o aumento dos preços. Ademais, a desigualdade em excesso também influenciou para que o preço do Uber X reduzisse. Outra característica que exibiu resultados semelhantes ao Índice de Gini foi o deslocamento de pessoas vulneráveis à pobreza. Além de indicar que o aumento na quantidade de pessoas que se encaixam nessas condições socioeconômicas faz com que o preço aumente, também mostra que o preço do Uber X é sensível a questões de mobilidade urbana, abrindo possibilidades para os dados da Uber sejam utilizados como possíveis indicadores de problemas de mobilidade.

Existem algumas limitações que valem a pena serem destacadas para trabalhos futuros. Uma das limitações é em relação ao conjunto de dados de preço do Uber X disponibilizados pela empresa. Esses dados são simulados e, portanto, não representam valores reais pagos pelo serviço. São apenas estimativas geradas pelo algoritmo de precificação da Uber e que podem não levar em consideração características que fariam o preço real aumentar ou diminuir. Uma outra limitação é o fato desse contemplar apenas um dos serviços ofertados pela Uber. Utilizar dados de outros serviços da empresa e até mesmo dados de outros serviços de transporte pode gerar novos ou até melhores resultados. A última das limitações é o uso de dados sociodemográficos para a última década e provenientes do Censo Demográfico 2010. No período de 10 anos entre cada censo, os dados socioeconômicos podem mudar bastante e não refletir de maneira mais destacada a realidade da população. Revisar esse trabalho utilizando fontes de dados sociodemográficos mais recentes, como o do Censo Demográfico 2020, pode melhorar o desempenho dos modelos existentes e gerar novos resultados.

AGRADECIMENTOS

O presente artigo foi realizado com apoio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS

Aljoufie, M., Zuidgeest, M., Brussel, M., and Maarseveen, M. (2011). *Urban growth and transport: Understanding the spatial temporal relationship*, volume 116, 315–328. doi:10.2495/UT110271.

Bezerra, A., Alves, G., Silva, I., Rosati, P., Endo, P., and Lynn, T. (2019). A preliminary exploration of uber data as an indicator of urban liveability. doi:10.1109/CyberSA.2019.8899714.

Capdevila, I. and Zarlenga, M. (2015). Smart city or smart citizens? the barcelona case. *Journal of Strategy and Management*, 8. doi:10.1108/JSMA-03-2015-0030.

DENATRAN (2018). Ministério da infraestrutura. URL `https://infraestrutura.gov.br/component/content/article/115-portal-denatran/8552-estat\%C3\%ADsticas-frota-de-ve\%C3\%ADculos-denatran.html`. Viewed 2020-06-06.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2013). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24. doi:10.1080/10618600.2014.907095.

Haughton, D., Robbert, M., Senne, L., and Gada, V. (2003). Effect of dirty data on analysis results. 64–79.

IBGE (2010). IBGE - Conceitos e métodos - Metadados. URL `https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?edicao=9673&t=conceitos-e-metodos`. Viewed 2020-06-06.

IBGE (2017). Ligações rodoviárias e hidroviárias 2016.

Ismagilova, E., Hughes, L., Dwivedi, Y.K., and Raman, K.R. (2019). Smart cities: Advances in research—an information systems perspective. *International Journal*

of Information Management, 47, 88 – 100. doi:10.1016/j.ijinfomgt.2019.01.004.

Jain, S., Chourse, S., Dubey, S., Jain, S., Kamakoty, J., and Jain, D. (2016). Regression analysis-its formulation and execution in dentistry.

Jesmeen, M., Hossen, J., Sayeed, S., Ho, C., Tawsif, K., Rahman, M.A., and Hossain, M. (2018). A survey on cleaning dirty data using machine learning paradigm for big data analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, 10, 1234–1243. doi:10.11591/ijeecs.v10.i3.pp1234-1243.

Kobalşi, A. and Ünsal, A. (2019). A comparison of the outlier detecting methods: An application on turkish foreign trade data. *Journal of Mathematics and Statistical Science*, 5, 213–234.

Mariano, A., Ramírez-Correa, P., Alfaro, J., Painén-Aravena, G., and Machorro, F. (2019). O papel da aceitação da tecnologia nas cidades inteligentes: Um estudo das percepções dos usuários do uber brasil. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, 17, 571–583.

Natrella, M. (2010). *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH.

Obedait, A.A., Youssef, M., and Ljepava, N. (2019). Citizen-centric approach in delivery of smart government services. In A. Al-Masri and K. Curran (eds.), *Smart Technologies and Innovation for a Sustainable Future*, 73–80. Springer International Publishing, Cham. doi:10.1007/978-3-030-01659-3_10.

PNUD, IPEA e FJP (2017). *Atlas do Desenvolvimento Humano nas Regiões Metropolitanas Brasileiras*. Programa das Nações Unidas para o Desenvolvimento (PNUD), Brasília.

Potdar, K., Pardawala, T., and Pai, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175, 7–9. doi:10.5120/ijca2017915495.

Ramli, M., Yahaya, A.S., Ramli, N., Md Yusof, N.F.F., and Abdullah, M.M.A.B. (2013). Roles of imputation methods for filling the missing values: A review. *Advances in Environmental Biology*, 7, 3861–3869.

Tsai, C.F. and Chang, F.Y. (2016). Combining instance selection for better missing value imputation. *Journal of Systems and Software*, 122, 63 – 71. doi:10.1016/j.jss.2016.08.093.

van Ginkel, J., Linting, M., Rippe, R., and Voort, A. (2019). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 102, 1–12. doi:10.1080/00223891.2018.1530680.

Wang, M. and Mu, L. (2018). Spatial disparities of uber accessibility: An exploratory analysis in atlanta, usa. *Computers, Environment and Urban Systems*, 67, 169–175. doi:10.1016/j.compenvurbsys.2017.09.003.

Yang, J. and Rahardja, S. (2019). Outlier detection: how to threshold outlier scores? 1–6. doi:10.1145/3371425.3371427.