Instituto CEUB de Pesquisa e Desenvolvimento – ICPD

COMPARANDO MÉTODOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DA DEMANDA DE VIAGENS DE BICICLETAS DA BIXI MONTREAL E ANÁLISE DO EFEITO DA PANDEMIA DE COVID-19 NA DEMANDA DE 2020

Thaís Ferreira Lopes[1]

RESUMO

A previsão de demanda é uma atividade estratégica para uma organização planejar e dimensionar os recursos necessários para a produção de bens e serviços de forma a atender sua demanda e não ter prejuízos. Visto que estima o futuro, ela pode ser impactada com alterações no ambiente externo, como foi o caso da pandemia de COVID-19. Nesse sentido, esse trabalho tem como objetivo analisar como a demanda da BIXI Montreal foi impactada pela pandemia de COVID-19, comparando os valores reais de 2020 com valores projetados utilizando modelos de aprendizado de máquina e dados históricos. Para isso, quatro modelos usando os algoritmos *LinearRegression, DecisionTreeRegressor, RandomForestRegressor e XGBRegressor* de pacotes do *Python* foram elaborados e seus desempenhos avaliados considerando as métricas MAE, MSE, RMSE e Score. A partir disso, o melhor modelo foi escolhido, sendo utilizado para prever a demanda de 2020 e comparar a demanda real com a contrafactual. De forma geral, o modelo de *Random Forest* apresentou o melhor desempenho e foram realizadas três previsões de demanda diferentes. Uma delas apresentou um resultado de demanda muito superestimado. As outras duas apresentaram resultados mais realistas, demonstrando possibilidades da demanda em um cenário mais conservador e outro mais otimista. Por fim, concluiu-se que a demanda de quantidade de viagens da BIXI Montreal sofreu um impacto negativo entre 32% e 125% por causa da pandemia.

Palavras-chave: Previsão de demanda. Aprendizado de máquina. Regressão. COVID-19.

1 Trabalho apresentado ao Centro Universitário de Brasília (UniCEUB/ICPD) como pré-requisito para obtenção de Certificado de Conclusão de Curso de Pós-graduação Lato Sensu em Ciência de Dados e Machine Learning, sob orientação do Prof. Me. Ivandro da Silva Ribeiro. Banca realizada em 4 de novembro de 2021, composta pelos professores-avaliadores Prof. Dr. Gilson Ciarallo e Prof. Me. William Malvezzi.

1 INTRODUÇÃO

A pandemia de COVID-19 impactou diretamente o funcionamento de diversas atividades em todo o mundo. Muitos países introduziram medidas rigorosas para conter a propagação do vírus, como o isolamento social, a quarentena e o *lockdown*. Além disso, foram estabelecidas determinações de espaçamento físico entre pessoas, uso de máscaras faciais, fechamento de empresas e escolas, priorização do trabalho remoto, entre outras que limitaram a circulação da população, o que impactou bastante o setor de transporte. Com isso, observou-se um aumento da procura por modos de transporte ativo, especialmente a pé e de bicicleta, por proporcionarem distanciamento social e reduzirem a aglomeração em ambientes fechados (NIKITAS et al., 2021).

Dentro desse segmento de transporte ativo, encontram-se as empresas de compartilhamento de bicicletas. Observa-se que a demanda desse serviço apresenta um crescimento temporário quando ocorre a interrupção do transporte público devido a fatores internos, como greves e manutenções, e a fatores externos, como desastres naturais e surtos de doenças transmissíveis, como o caso da COVID-19. Porém, estudos sobre a compreensão do efeito do surto de doenças transmissíveis no compartilhamento de bicicletas são relativamente raros (HEYDARI; KONSTANTINOUDIS; BEHSOODI, 2021).

Nesse sentido, pretende-se analisar a BIXI Montreal, que é uma organização sem fins lucrativos criada em 2014 pela cidade de Montreal para gerenciar seu sistema de compartilhamento de bicicletas. Segundo um comunicado da empresa publicado em 14 de novembro de 2019, naquele ano, o negócio estabeleceu um novo recorde de 5,8 milhões de viagens e atingiu o recorde histórico de 320.000 usuários individuais. Esses números representaram um aumento de 80% no uso e 309% nas vendas em relação aos últimos cinco anos (2014-2018). Com base nisso e na pandemia recente, eram esperadas mudanças importantes na demanda da organização no ano de 2020.

Dessa forma, este trabalho tem como objetivo geral analisar como a demanda da BIXI Montreal foi impactada pela pandemia de COVID-19, comparando

os valores reais de 2020 com valores projetados utilizando modelos de aprendizado de máquina e dados históricos. Para isso, o trabalho foi dividido em etapas, definidas pelos seguintes objetivos específicos: a) identificar e tratar a base de dados utilizada na análise; b) elaborar os modelos de previsão de demanda utilizando algoritmos de aprendizado de máquina; c) comparar os desempenhos obtidos nos modelos; d) projetar a demanda da BIXI Montreal para o ano de 2020; e, por fim, e) comparar a demanda projetada com os dados reais, analisando as diferenças de quantidade de viagens.

O presente trabalho foi então estruturado nas seguintes seções: na seção dois apresenta-se a revisão bibliográfica sobre a previsão de demanda, o aprendizado de máquina, os algoritmos e as métricas; na seção três, a metodologia utilizada; na seção quatro, o desenvolvimento do trabalho e a discussão dos resultados obtidos; e, na quinta seção, as considerações finais e sugestões de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

2.1 Previsão de demanda

A previsão de demanda é um método para a determinação de dados futuros, que podem ser baseados em modelos estatísticos, econométricos ou até mesmo modelos subjetivos apoiados em uma metodologia de trabalho clara e previamente definida (MARTINS; LAUGENI, 2015). Ela tem como objetivo dimensionar os recursos necessários para a produção de bens e serviços de forma a atender a demanda de uma organização. Dessa forma, serve como ponto de partida para o planejamento estratégico de muitas atividades de gestão, como produção, vendas e finanças.

A elaboração de uma previsão de demanda pode ser feita utilizando métodos quantitativos, qualitativos ou a combinação de ambos. Os métodos quantitativos são baseados na análise de séries temporais, isto é, conjuntos de dados que denotam a variação de demanda em um certo período de tempo. Por outro lado, os métodos qualitativos baseiam-se em opiniões e estudos realizados por especialistas (PELLEGRINI; FOGLIATTO, 2001).

Neste trabalho, utilizou-se séries temporais de registros de viagens de bicicletas para prever a demanda, o que caracteriza um método quantitativo. Nos próximos tópicos são apresentados os métodos utilizados para manipular esses dados e prever os futuros.

2.2 Aprendizado de Máquina (*Machine Learning*)

O aprendizado de máquina é uma subárea da Inteligência Artificial focada na capacidade de resolução de problemas complexos ou de grande volume de dados. De forma geral, “aplicações baseadas em aprendizado de máquina utilizam heurísticas que buscam por modelos capazes de representar o conhecimento presente em um conjunto de dados” (FACELI et al., 2021, p. 2). Esses conjuntos de dados, normalmente, são estruturados como uma tabela, em que as linhas representam objetos e as colunas, atributos.

> Os atributos podem ser divididos em atributos preditivos, cujos valores descrevem características dos objetos, que formam um vetor de entrada, e atributo alvo, cujo valor rotula o objeto, com uma classe ou valor numérico. Essas denominações têm por origem o frequente uso dos valores dos atributos preditivos de um objeto para predizer o valor de seu atributo alvo. Nem todos os conjuntos de dados possuem atributo alvo. Quando possuem, são chamados de conjuntos de dados rotulados (FACELI et al., 2021, p. 2).

Com base nisso, o aprendizado de máquina utiliza algoritmos ou modelos que são treinados para criar regras ou parâmetros que relacionam os dados de entrada (atributos preditivos) com os dados de saída (atributo alvo), permitindo a realização de tarefas como classificação, previsão e agrupamento de dados. E o aprendizado refere-se à capacidade do algoritmo/modelo melhorar repetidamente o seu desempenho a partir da experiência absorvida após avaliar diferentes conjuntos de dados (LENZ et al., 2020).

Esse aprendizado é dividido em quatro tipos, sendo eles:

a) **Supervisionado:** tem como objetivo aprender a mapear a relação de uma entrada para determinada saída, cujos rótulos corretos são conhecidos e fornecidos. Assim, é um modelo de aprendizado construído para fazer previsões como resposta à entrada de novos dados. Ele utiliza algoritmos de classificação (quando o valor do rótulo é discreto) e técnicas de

regressão (quando o valor do rótulo é contínuo) para desenvolver esses modelos preditivos (SHOBHA; RANGASWAMY, 2018);

b) **Não supervisionado:** ao invés de predizer um valor, como no supervisionado, os algoritmos não supervisionados extraem padrões dos atributos preditivos de um conjunto de dados. A partir disso, pode-se realizar o agrupamento, a associação ou a sumarização dos dados, sendo que (FACELI et al., 2021):

   - Agrupamento é dividir os dados em grupos de acordo com sua semelhança;
   - Sumarização é buscar uma descrição simples e compacta para um conjunto de dados; e
   - Associação é procurar padrões frequentes de associações entre os atributos de um conjunto de dados.

c) **Semissupervisionado:** é a combinação do supervisionado e do não supervisionado, sendo usado quando há poucos dados rotulados para determinada aplicação. Seu objetivo é classificar alguns dos dados não rotulados usando o conjunto de dados rotulados (SHOBHA; RANGASWAMY, 2018); e

d) **Por reforço:** envolve a interação de um agente autônomo com o ambiente em que ele está inserido. De forma geral, o comportamento do agente é recompensado de acordo com as ações que ele realiza no ambiente, assim ele vai aprendendo a escolher as ações ideais para atingir seus objetivos (SHOBHA; RANGASWAMY, 2018).

Neste trabalho foram utilizados algoritmos de aprendizagem supervisionada de regressão para a predição de valores contínuos, que são a quantidade de viagens que a BIXI Montreal irá atender em uma determinada semana, prevendo a demanda semanal de aluguéis de bicicletas.

2.3 Métodos de regressão

Os métodos de regressão são utilizados quando o objetivo é prever variáveis numéricas e contínuas. Nesta seção serão apresentados os algoritmos utilizados no desenvolvimento desse trabalho, sendo eles: Regressão Linear, Árvore de Decisão, Florestas Aleatórias e XGBoost.

2.3.1 Regressão Linear (*Linear Regression*)

Regressão Linear ou *Linear Regression* é um dos métodos mais simples de aprendizado de máquina supervisionado. Ela é uma metodologia estatística que utiliza da relação entre duas ou mais variáveis quantitativas a fim de que a variável de resposta possa ser prevista a partir das anteriores (KUTNER et al., 2004). Ela pode ser simples, quando utiliza apenas um parâmetro para realizar a predição, ou múltipla, quando utiliza dois ou mais parâmetros.

A regressão linear múltipla utiliza a lógica de verificar como um conjunto de variáveis pode explicar outra variável (REHBEIN, 2019). A equação a seguir apresenta sua definição:

$$Y_1 = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + e_n$$ (Equação 1), onde:

- Y_1 é o valor predito;
- α é o valor constante do modelo;
- os x_n apresentam os valores das variáveis analisadas;
- os β_n são os coeficientes do modelo;
- e_n apresenta os resíduos do modelo.

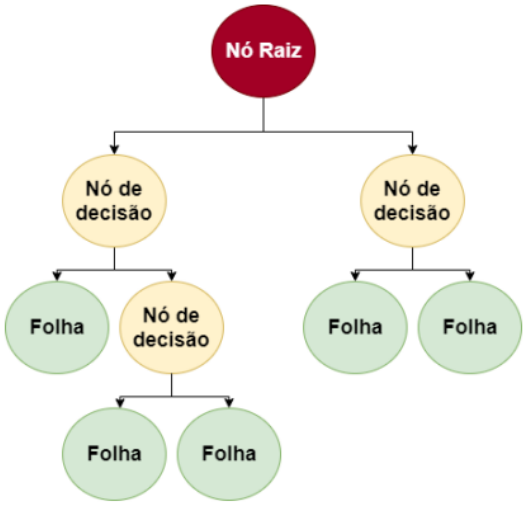
Este método é um dos mais aplicados quando os dados de saída são valores contínuos, mas é bastante sensível a outliers, o que afeta o seu desempenho (MATOS, 2021).

2.3.2 Árvore de Decisão (*Decision Tree*)

Árvore de Decisão ou *Decision Tree* é um algoritmo de aprendizado supervisionado usado para solucionar problemas de classificação e regressão. Seu conceito básico é dividir uma decisão complexa em várias decisões mais simples, que

conduzem a uma solução que é mais fácil de interpretar. Assim, ele é baseado em um esquema de decisão multiestágio ou hierárquico ou em uma estrutura de árvore. Essa estrutura é composta por um conjunto de nós de tomada de decisão binária. Os nós podem ser: raiz (que contém todos os dados), internos ou de decisão (divisões) e terminais (folhas). O processamento desse algoritmo é *top-down*, isto é, inicia no nó raiz e move-se para baixo na árvore até alcançar o nó terminal (XU et al., 2005). A Figura 1 a seguir representa a estrutura de uma árvore de decisão.

Figura 1 - Estrutura de uma árvore de decisão



Fonte: DELGADO FILHO (2020, p. 32).

Para definir cada nó, o algoritmo separa os dados utilizando todas divisões binárias possíveis e seleciona o atributo que realiza a melhor divisão. Para escolher esse atributo, utiliza-se os conceitos de entropia e ganho. A entropia refere-se ao grau de pureza de um conjunto. Este conceito vem da Teoria da Informação, que define a medida de "falta de informação", especificamente o número de bits necessários, em média, para representar a informação em falta, usando codificação ótima (GFBIOINFO). Dado um conjunto S, com instâncias pertencentes à classe i, com probabilidade p_i, o cálculo de entropia é representado pela seguinte equação:

$$Entropia\ (S) = \sum p_i\ log_2\ p_i \text{ (Equação 2).}$$

O ganho refere-se à redução na entropia, logo o Ganho (S, A) significa a redução esperada na entropia de S, ordenando pelo atributo A. Assim, para escolher o melhor atributo é usado o ganho, ou seja, em cada iteração do algoritmo

é escolhido o atributo que apresente o maior ganho. O cálculo do ganho é representado pela seguinte equação:

$$Ganho\ (S, A) = Entropia\ (S) - \sum_{v\ \in\ valores\ (A)} \frac{|S_v|}{|S|}\ Entropia(S_v)$$ (Equação 3).

As árvores de decisão também podem ser aplicadas a problemas de regressão. Nesse caso, a construção do algoritmo também é baseada no particionamento binário repetitivo. Inicialmente, todas as amostras de treinamento são usadas para determinar a estrutura da árvore. O algoritmo separa os dados utilizando todas divisões binárias possíveis e seleciona aquela que divide os dados em duas partes e minimiza a soma dos desvios quadrados da média das partes separadas. Esse processo se repete em cada nó de decisão até chegar nos nós terminais (XU et al., 2005).

Ressalta-se que, pela árvore de decisão ser construída a partir de amostras de treinamento, ela pode apresentar *overfitting,* isto é, o modelo pode sofrer um ajuste excessivo, o que prejudica sua precisão ao utilizar novos dados, reduzindo sua capacidade de generalização. Para evitar isso é importante realizar o processo de poda com um conjunto de dados de validação, que ao ser processado pelo modelo faz com que ele sofra ajustes e melhore sua precisão (XU et al., 2005).

2.3.3 Floresta Aleatória (*Random Forest*)

Floresta Aleatória ou *Random Forest* é um conjunto de diferentes árvores de decisão que também pode ser usado para solucionar problemas de classificação e regressão, porém a criação das árvores que compõem esse algoritmo difere da árvore de decisão comum por utilizar aleatoriedade na seleção de atributos. Dessa forma, as árvores são treinadas com atributos e conjunto de dados distintos, o que garante que cada árvore crie um modelo diferente. De forma geral, as principais etapas do algoritmo são:

a) Criar um conjunto de n amostras utilizando *bootstraping*, isto é, são selecionadas amostras aleatórias com reposição (a mesma linha pode ser escolhida mais de uma vez) da base de dados inicial.

b) Construir uma árvore por amostra. Primeiro são selecionadas, aleatoriamente, n variáveis dentre as variáveis explicativas existentes na base de dados inicial. Com esse subconjunto selecionado é feita a verificação do atributo que melhor divide os dados. Esse atributo é utilizado para iniciar a árvore. Em seguida, mais dois atributos são selecionados entre os restantes e o processo se repete até que o nó terminal atinja um número mínimo de observações pré-definido. Ressalta-se que os subconjuntos de cada nó são diferentes em uma mesma árvore (FREITAS, 2018).

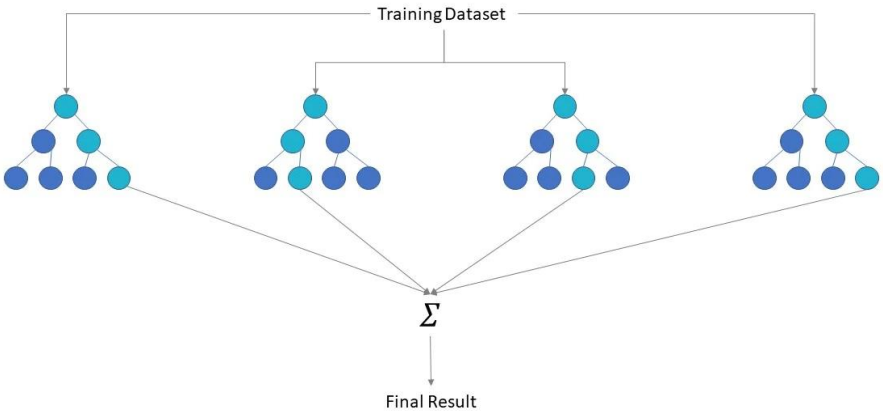
c) Para realizar uma classificação, as classes determinadas por cada árvore da floresta são computadas como um voto, a classe definida refere-se à opção com mais votos. “No caso de uma floresta de regressão, as previsões de cada árvore são uma média da variável de interesse, e a previsão da floresta é uma média das previsões dadas pelas árvores que a compõem” (FREITAS, 2018), conforme a seguinte equação:

$$\hat{y}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_b^*(x)$$ (Equação 4), onde:

- B é a quantidade de amostras *bootstraping*; e
- $\hat{y}_b^*$ é a resposta predita pela b-ésima árvore *random forest*.

A Figura 2 a seguir representa um exemplo de uma floresta aleatória.

Figura 2 - Exemplo de uma floresta aleatória



Fonte: IBM, 2020.

Por causa da aleatoriedade das amostras de dados e das variáveis consideradas na tomada de decisão nos nós, a Floresta Aleatória reduz problemas de *overfitting*. Além disso, a combinação de múltiplas árvores de decisão obtém uma previsão mais precisa e estável.

2.3.4 XGBoost

XGBoost é a abreviação do pacote *eXtreme Gradient Boosting*. Pode-se dizer que ele é um algoritmo de floresta aleatória aprimorado pelo método *boosting*. Esse método inicia a construção da floresta a partir de uma primeira árvore construída com alto grau de sobreajuste aos dados de treinamento. Em seguida, novas árvores são construídas de forma sequenciada, a partir de um aprimoramento da árvore anterior, buscando reduzir o grau de sobreajuste a cada nova rodada de ajuste (CHEN; GUESTRIN, 2016).

2.4 Métricas de avaliação do desempenho dos algoritmos de regressão

A avaliação de um algoritmo de aprendizado de máquina supervisionado é normalmente feita por meio da análise do desempenho dele em predizer rótulos de dados novos, que não foram apresentados durante o treinamento (MONARD; BARANAUSKAS, 2003). Nos tópicos abaixo serão apresentadas as métricas que foram utilizadas nesse trabalho para avaliar o desempenho dos modelos.

2.4.1 Erro Médio Absoluto (MAE)

O Erro Médio Absoluto ou *Mean Absolute Error* (MAE) representa a média da diferença absoluta entre os valores reais e os previstos. A fórmula de cálculo é:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \vee \hat{y}_i - y_i \vee$$ (Equação 5), onde:

- $\hat{y}_i$ é o valor previsto pelo método;
- y_i é o valor real;
- *n* é o tamanho da amostra;

O MAE apresenta o valor mínimo igual a zero e não possui valor máximo, porém quanto menor seu valor, melhor o modelo. Ele representa uma métrica estável para modelos que devem prever muitos dados ou dados sazonais, nas quais prever a

tendência e sazonalidade dos números é mais importante do que os valores absolutos de cada dia (AZANK, 2020).

2.4.2 Erro quadrático médio (MSE)

O Erro Quadrático Médio ou *Mean Squared Error* (MSE) é o MAE elevado ao quadrado. A fórmula de cálculo é:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \text{ (Equação 6)}$$

Da mesma forma que o MAE, o valor mínimo do MSE é zero e ele não possui valor máximo, mas quanto menor, melhor o modelo. Ele é comumente usado para verificar a acurácia de modelos e em comparação ao MAE, por elevar o erro ao quadrado, ele dá um maior peso aos maiores erros, sendo uma boa métrica para problemas nos quais grandes erros não são tolerados. Ressalta-se, porém, que sua interpretação não é direta, visto que a unidade do MSE é u^2 enquanto a predição de valores é u. (AZANK, 2020).

2.4.3 Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio ou *Root Mean Squared Error* (RMSE) é a raiz quadrada do MSE. Assim, ele melhora a interpretabilidade do MSE, pois retorna à unidade de medida do modelo. A fórmula de cálculo é:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}} \text{ (Equação 7)}$$

Seu valor mínimo também é zero e ele não possui valor máximo, mas quanto menor, melhor o modelo.

2.4.4 Score

O Score refere-se a função *.score()* da biblioteca *Scikit-learn* do *Python*. Para executá-la é preciso passar dois principais parâmetros: os dados de entrada de treino ou teste (X) e os dados de saída (y), conforme apresentado a seguir:

modelo.score(X_treino, y_treino) ou modelo.score(X_teste, y_teste)

O objetivo dessa função é verificar o valor que seu modelo previu e comparar com o valor esperado. O resultado varia entre 0 e 1, em que quanto mais próximo de 1, melhor o modelo consegue explicar os valores observados. Ressalta-se, porém, que é importante observar as diferenças do *score* entre o conjunto de treino e teste, pois quando o modelo apresenta um *score* alto no conjunto de treino, e baixo no de teste, existe um problema de *overfitting.*

2.5 Trabalhos relacionados

Ao pesquisar trabalhos relacionados ao tema abordado nesse estudo, foram encontrados alguns artigos sobre previsão de demanda de sistemas de compartilhamento de bicicleta, assim como artigos avaliando o efeito da pandemia nesse serviço. A respeito do primeiro grupo, Sathishkumar, Park e Cho (2020) utilizaram uma técnica de mineração de dados para prever a demanda de aluguel de bicicletas por hora, visando o reabastecimento necessário das estações ao longo do dia. Para isso, eles compararam o desempenho dos algoritmos *Linear regression, Support Vector Machine, Boosted Trees, Extreme Gradient Boosting Tree e Gradient Boosting Machine*, sendo esse o que apresentou melhor resultado.

Wang et al. (2021) desenvolveram um modelo de regressão com coeficientes que variam espacialmente para investigar como o uso da terra, a infraestrutura sociodemográfica e de transporte afetam a demanda de compartilhamento de bicicletas em diferentes estações. Para isso, definiram uma regressão específica para cada estação e usaram uma estrutura de gráfico para encorajar as estações próximas a ter um coeficiente similar. O modelo proposto apresentou capacidade de previsão superior em comparação a outros modelos (*Random forest, SVM regression, KNN, Nearest neighbors average, Linear regression, Regression kriging, Geographically weighted regression (GWR)* e *Graph regularization with circle buffer*). Esse estudo também utilizou os dados da BIXI Montreal.

Sobre o segundo grupo, Heydari, Konstantinoudis, Behsoodi (2021) estudaram o efeito da pandemia de COVID-19 no sistema de compartilhamento de bicicletas de Londres (*Santander Cycles*) durante o período de março a dezembro de 2020. Para isso, utilizaram *Bayesian hierarchical models with a second-order random*

walk specification para explicar a correlação temporal entre os dados. Por fim, compararam o número observado de aluguéis de bicicletas e de tempo de duração com seus respectivos contrafactuais, o que teria sido se a pandemia não tivesse acontecido, para estimar o impacto causado pela pandemia. Outros estudos relacionados ao efeito de surtos de doenças transmissíveis e da pandemia de COVID-19 em sistemas de compartilhamento de bicicleta são citados por esses autores.

De forma geral, o diferencial desse trabalho em relação a esses estudos correlatos é que foram comparados alguns algoritmos diferentes para prever a demanda e realizadas previsões em diferentes cenários, o que possibilitou identificar com uma margem o impacto da pandemia de COVID-19.

2.6 Caracterização da organização

A BIXI Montreal é uma organização sem fins lucrativos criada em 2014 pela cidade de Montreal, no Canadá, para gerenciar seu sistema de compartilhamento de bicicletas. O serviço da BIXI funciona por temporada, que, normalmente, corresponde ao período de 15 de abril a 15 de novembro, durante a primavera, verão e outono na cidade. Nos outros meses as estações de bicicleta são retiradas por causa do inverno.

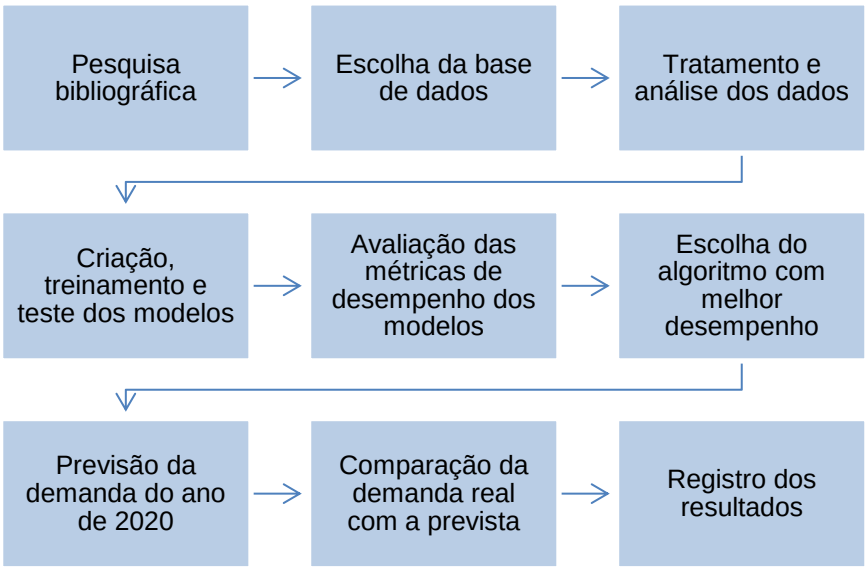
Esse serviço preenche um nicho importante no sistema de transporte público de Montreal, conectando os usuários, que podem ser membros ou usuários ocasionais (como turistas), entre as lacunas nos sistemas de ônibus e metrô e fornecendo uma alternativa saudável e ecológica para os deslocamentos de curta duração.

3 METODOLOGIA DO TRABALHO

Esse trabalho é classificado, de acordo com Fontelles et al. (2009), como: aplicado, pois tem como objetivo a aplicação dos conhecimentos obtidos por meio da pesquisa para resolver o problema de previsão de demanda de uma empresa real; observacional: visto que foram realizadas manipulações e análises nos dados disponibilizados pela empresa, mas não houve nem um tipo de intervenção no objeto da pesquisa; quantitativo, pois utilizou dados numéricos para atingir conclusões mensuráveis e estatísticas; transversal e retrospectivo, visto que os dados utilizados

consideram um intervalo de tempo passado. Assim, a elaboração do trabalho foi dividida nas etapas apresentadas na Figura 3 a seguir.

Figura 3 - Etapas da metodologia do trabalho



Fonte: Autora.

Primeira etapa: procedeu-se com uma pesquisa bibliográfica sobre os algoritmos e as métricas utilizadas, além do levantamento de trabalhos relacionados ao tema analisado, a fim de adquirir embasamento teórico e técnico. Para isso, utilizaram-se livros, artigos científicos, trabalhos de conclusão de graduação e dissertações de mestrado.

Segunda etapa: consistiu em buscar uma base de dados que possibilitasse a análise proposta no trabalho. Assim, foram utilizados os dados de histórico de viagens dos anos 2019 e 2020, disponibilizados no *Open Data* do site da empresa BIXI Montreal (https://bixi.com/en/open-data).

Terceira etapa: realizou-se o tratamento dos dados utilizando as bibliotecas da linguagem de programação *Python* para possibilitar uma análise preliminar dos dados. Nessa etapa, os registros unitários das viagens foram agregados a fim de se obter a quantidade de viagens por semana. Além disso, foram simuladas diferentes bases para prever a demanda de 2020. Por fim, foram criados alguns gráficos para visualização dos dados e os dados foram divididos para treino e teste dos modelos.

Quarta etapa: consistiu na criação, treinamento e teste dos modelos com cada um dos algoritmos (LinearRegression, DecisionTreeRegressor, RandomForestRegressor e XGBRegressor).

Quinta etapa: avaliou-se o desempenho dos modelos por meio dos resultados das métricas (MAE, MSE, RMSE e Score). Dessa forma, foi definido o algoritmo que apresentou melhor desempenho.

Sexta etapa: foram feitas previsões da demanda do ano de 2020, utilizando o melhor algoritmo. Em seguida, compararam-se os resultados preditos com os reais, para analisar a diferença na quantidade de viagens e o impacto que a pandemia causou na demanda da empresa.

4 DESENVOLVIMENTO DO PROJETO

Os tópicos apresentados a seguir trazem mais detalhes de cada etapa descrita na metodologia e os resultados obtidos em cada uma delas.

4.1 Tratamento dos dados

Os dados disponibilizados pela BIXI Montreal consistem nos registros unitários de viagens de bicicleta realizadas em todas as estações em Montreal, no Canadá. Ao todo foram utilizadas 10 planilhas csv, sendo 7 referentes aos meses de abril a outubro de 2019, 1 com o acumulado dos meses de abril a novembro de 2020 e 2 com as informações das estações no ano de 2019 e de 2020, respectivamente. As bases com os registros de viagens estavam organizadas em colunas conforme apresentado no Quadro 1 a seguir.

Quadro 1 - Variáveis das bases de dados

Nome da coluna	Descrição
start_date	dia e hora do início da viagem
start_station_code	código da estação de início da viagem
end_date	dia e hora do fim da viagem
end_station_code	código da estação de fim da viagem
duration_sec	duração da viagem em segundos

Nome da coluna	Descrição
is_member	se o passageiro é ou não é membro

Fonte: Autora.

A partir disso, para obter a demanda de viagens por semana, os dados foram acumulados por semana, trajeto (isto é, viagens que começaram e finalizaram nas mesmas estações) e se foram realizadas por membros ou não. Para cada combinação “semana,inicio,fim,membro” foi contabilizada a frequência, definindo assim a quantidade de viagens realizadas. Após isso, as linhas duplicadas foram excluídas e as colunas ano e mês foram adicionadas. As Figuras 4 e 5 apresentam os *datasets* de registros de viagens de 2019 e 2020 após o tratamento, respectivamente.

Figura 4 - *Dataset* de registros de viagens de 2019

	start_station_code	end_station_code	is_member	ano	mes	qtd_viagens	n_semana
0	6001	6132	1	2019	4	1	0
1	6411	6411	1	2019	4	1	0
2	6097	6036	1	2019	4	1	0
3	6310	6345	1	2019	4	1	0
4	7029	6250	0	2019	4	1	0
...	...	...	...	...	...	...	...
2302921	6212	6733	1	2019	10	1	29
2302922	6129	7095	0	2019	10	1	29
2302923	6202	6090	1	2019	10	1	29
2302924	6901	6063	1	2019	10	1	29
2302925	7149	6338	1	2019	10	1	29

2302926 rows × 7 columns

Fonte: Autora.

Figura 5 - *Dataset* de registros de viagens de 2020

	start_station_code	end_station_code	is_member	ano	mes	qtd_viagens	n_semana
0	6212	6250	1	2020	4	2	0
1	6369	6072	1	2020	4	3	0
2	6207	7038	1	2020	4	3	0
3	6202	6212	1	2020	4	3	0
4	6159	7067	1	2020	4	1	0
...	...	...	...	...	...	...	...
1704429	7073	6209	1	2020	11	1	31
1704430	6154	7035	1	2020	11	1	31
1704431	6108	6017	1	2020	11	1	31
1704432	6104	6119	1	2020	11	1	31
1704433	6148	6370	1	2020	11	1	31

1704434 rows × 7 columns

Fonte: Autora.

Em seguida, para simular os dados de entrada que o gestor teria para fazer a previsão da demanda de 2020, foram criadas duas bases:

a) Uma com todas as combinações possíveis de semana, estação de início, estação de fim e membros e não membros. Para isso, utilizou-se a lista com os códigos de todas as estações que estavam em funcionamento em 2020, disponível na planilha Station_2020; e
b) Uma com a junção dos registros de 2019 e de 2020, incluindo novos trajetos devido a novas estações criadas.

Por fim, essas bases foram combinadas com a de registro de viagens reais de 2020, resultando na coluna de quantidade de viagens por semana de cada trajeto. Nos trajetos que não tinham registro em 2020, os valores nulos foram substituídos por zero. As Figuras 6 e 7 a seguir apresentam os *datasets* de cada uma dessas bases simuladas.

Figura 6 - *Dataset* da base simulada com todas as combinações possíveis de 2020

	n_semana	start_station_code	end_station_code	is_member	ano	qtd_viagens	mes
0	0	6001	6001	0	2020	0.0	4
1	0	6001	6001	1	2020	2.0	4
2	0	6001	6002	0	2020	0.0	4
3	0	6001	6002	1	2020	3.0	4
4	0	6001	6003	0	2020	0.0	4
...	...	...	...	...	...	...	...
26296379	31	8022	8036	1	2020	0.0	11
26296380	31	8022	8069	0	2020	0.0	11
26296381	31	8022	8069	1	2020	0.0	11
26296382	31	8022	8022	0	2020	0.0	11
26296383	31	8022	8022	1	2020	0.0	11

26296384 rows × 7 columns

Fonte: Autora.

Figura 7 - *Dataset* da base simulada com a junção dos registros de 2019 e 2020

	start_station_code_x	end_station_code_x	is_member_x	ano_x	mes_x	n_semana_x	qtd_viagens_y
0	6001.0	6132.0	1.0	2020.0	4.0	0.0	1.0
1	6411.0	6411.0	1.0	2020.0	4.0	0.0	2.0
2	6097.0	6036.0	1.0	2020.0	4.0	0.0	0.0
3	6310.0	6345.0	1.0	2020.0	4.0	0.0	0.0
4	7029.0	6250.0	0.0	2020.0	4.0	0.0	0.0
...	...	...	...	...	...	...	...
3282629	7073.0	6209.0	1.0	2020.0	11.0	31.0	1.0
3282630	6154.0	7035.0	1.0	2020.0	11.0	31.0	1.0
3282631	6108.0	6017.0	1.0	2020.0	11.0	31.0	1.0
3282632	6104.0	6119.0	1.0	2020.0	11.0	31.0	1.0
3282633	6148.0	6370.0	1.0	2020.0	11.0	31.0	1.0

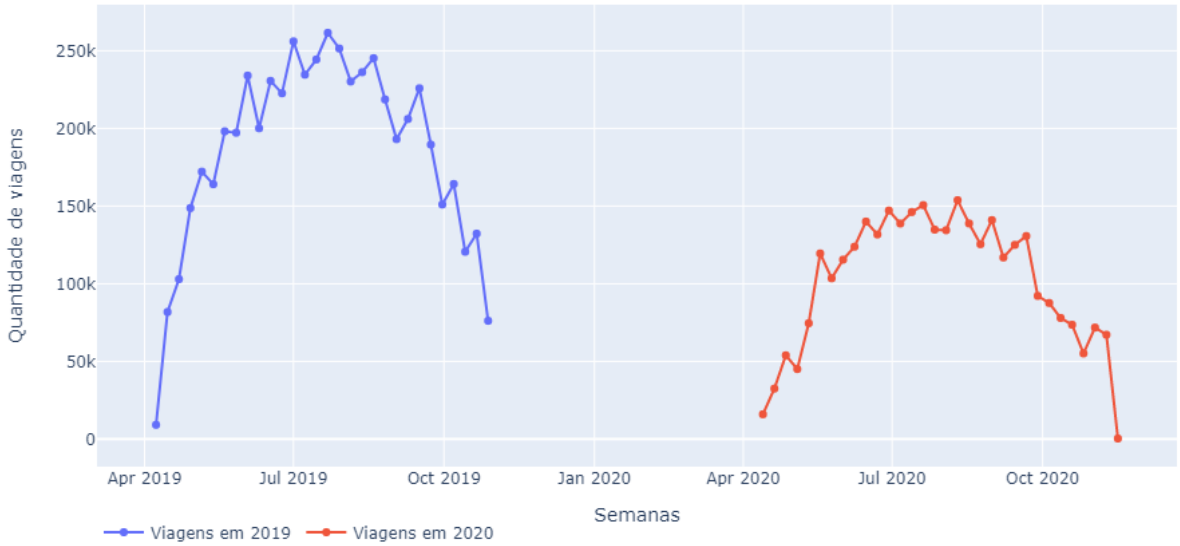
3282634 rows × 7 columns

Fonte: Autora.

4.1.1 Análise dos dados

A fim de entender os dados, foi feita uma análise preliminar com alguns gráficos. O primeiro analisou a quantidade de viagens por semana em 2019 e 2020, conforme apresentado no Gráfico 1 a seguir.

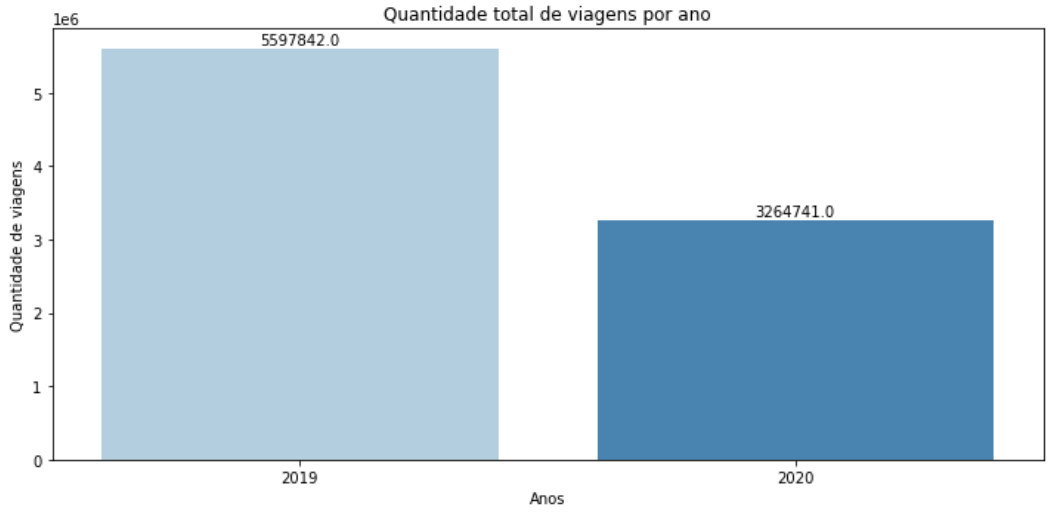
Gráfico 1 - Quantidade total de viagens por semana



Fonte: Autora.

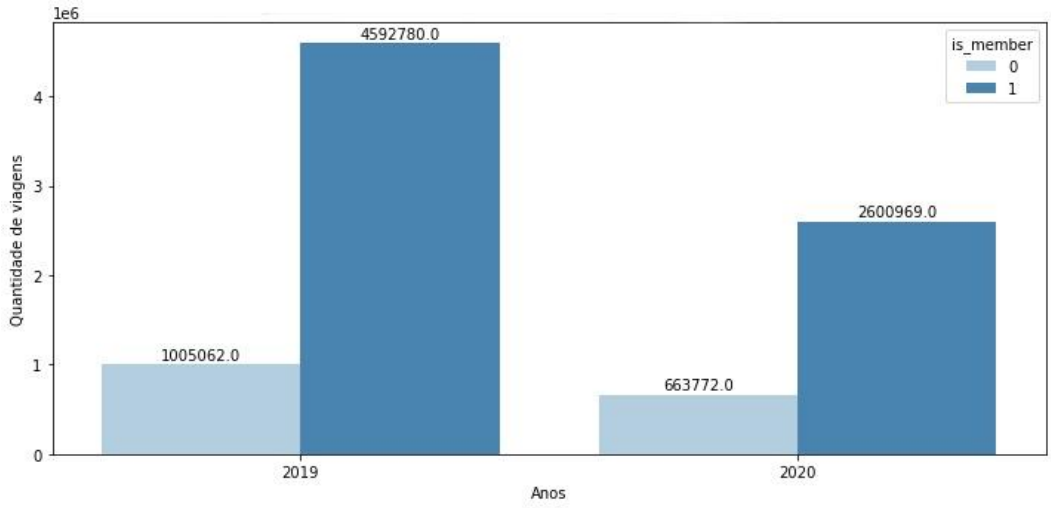
Observa-se no Gráfico 1 que as curvas das demandas apresentam o mesmo formato, o que caracteriza a sazonalidade presente nesse serviço. Além disso, é possível ver a redução da demanda em 2020 causada pela pandemia de COVID-19. Para analisar a proporção dessa redução, os dados foram acumulados por ano. Os Gráficos 2 e 3 a seguir apresentam a quantidade total de viagens por ano e divididas entre os membros e não membros.

Gráfico 2 - Quantidade total de viagens por ano



Fonte: Autora.

Gráfico 3 - Quantidade total de viagens por ano dividida por membros e não membros[2]



Fonte: Autora.

Observa-se no Gráfico 2 que 2020 teve 2.333.101 viagens a menos que 2019. Isso representa uma redução de quase 42% de viagens. Já no Gráfico 3, observa-se que a quantidade de viagens de membros foi maior que a de não membros nos dois anos. Porém, calculando a proporção da quantidade de viagens de cada tipo em relação ao total do ano, em 2020 a quantidade de viagens feitas por não membros representou 20% das viagens, 2% a mais que em 2019.

2 Legenda: 0 é não membro (usuário ocasional) e 1 é membro.

4.1.2 Divisão dos dados de treino e teste

Para o treinamento e teste dos modelos, considerou-se apenas os dados de 2019. Assim, foram definidos o X e o y, sendo que:

a) X corresponde às colunas start_station_code_x, end_station_code_x, is_member_x, ano_x e semana_x; e
b) y corresponde à coluna viagens.

Após isso, a divisão dos dados em treino e teste foi feita utilizando o pacote train_test_split da biblioteca Scikit-learn do Python. Já para a previsão dos valores de 2020, foram considerados 3 *datasets* diferentes, sendo definidos 3 Xs:

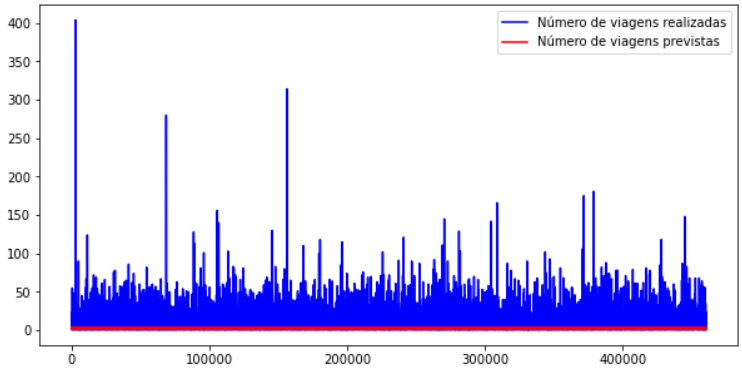
a) X_1: utilizando a base de registros de viagens de 2020;
b) X_2: utilizando a base simulada com todas as combinações possíveis de 2020;
c) X_3: utilizando a base simulada com a junção dos registros de 2019 e 2020.

4.2 Criação, treino e teste dos modelos e análise dos resultados

Para a criação dos modelos utilizou-se alguns pacotes da biblioteca *Scikit-learn* e *xgboost* do *Python.* Ao todo foram criados 4 modelos utilizando algoritmos diferentes, sendo eles: *LinearRegression, DecisionTreeRegressor*, *RandomForestRegressor* e *XGBRegressor*. Foram utilizadas as configurações de parâmetros padrão.

Após o treinamento dos modelos com os dados de treino, foi feita a previsão com os dados de teste e criados alguns gráficos a fim de observar os valores reais e os previstos. Os Gráficos 4, 5, 6 e 7 a seguir apresentam os resultados obtidos em cada modelo, respectivamente.

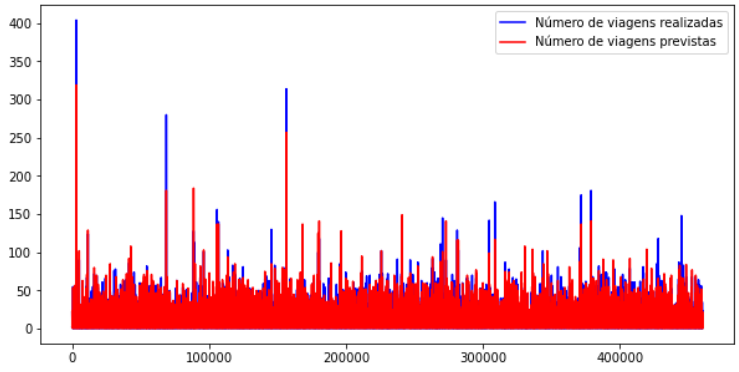
Gráfico 4 - Comparação viagens realizadas e previstas com o modelo Regressão Linear



Fonte: Autora.

O Gráfico 4 acima representa a comparação de viagens realizadas (azul) e de viagens previstas com o modelo Regressão Linear (vermelho). Observa-se que esse modelo não gerou bons resultados, visto que as viagens previstas se concentraram na base, não acompanhando as colunas de viagens realizadas. Isso significa que o modelo previu uma quantidade menor de viagens do que de fato aconteceu, assim o modelo não conseguiu encontrar uma boa relação entre os dados de entrada a fim de prever o dado de saída.

Gráfico 5 - Comparação viagens realizadas e previstas com o modelo Árvore de Decisão

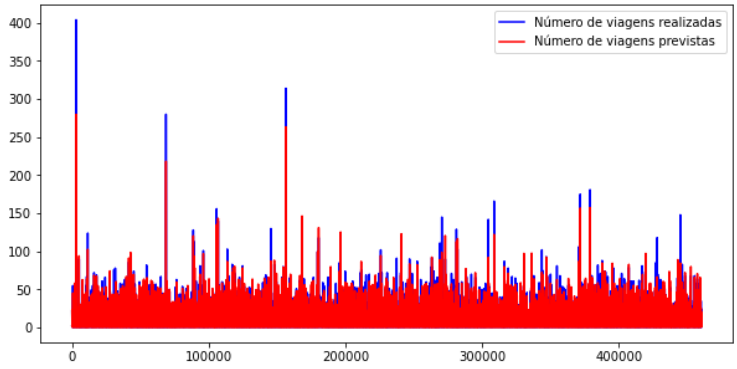


Fonte: Autora.

O Gráfico 5 acima representa a comparação de viagens realizadas (azul) e de viagens previstas com o modelo Árvore de Decisão (vermelho). Diferente do Gráfico 4, observa-se que esse modelo apresentou resultados melhores, pois as colunas de viagens previstas ficaram bem próximas das colunas de viagens realizadas. Assim, esse modelo previu uma quantidade próxima de viagens do que de

fato aconteceu, o que significa que ele encontrou uma relação melhor entre os dados de entrada a fim de prever o dado de saída.

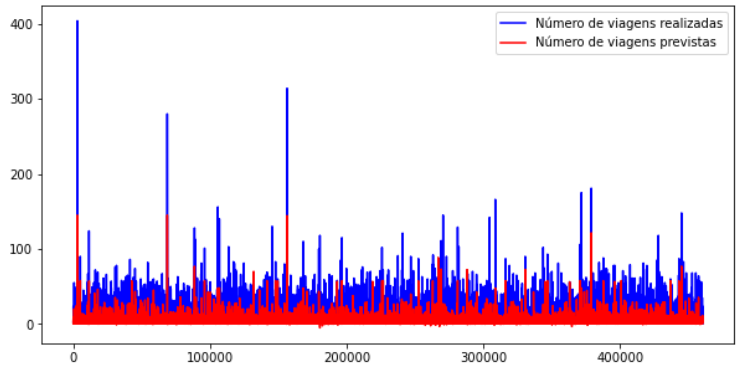
Gráfico 6 - Comparação viagens realizadas e previstas com o modelo *Random Forest*



Fonte: Autora.

O Gráfico 6 acima representa a comparação de viagens realizadas (azul) e de viagens previstas com o modelo *Random Forest* (vermelho). Observa-se que ele apresentou um resultado muito próximo do Gráfico 5, conseguindo prever uma quantidade próxima de viagens do que de fato aconteceu.

Gráfico 7 - Comparação viagens realizadas e previstas com o modelo *XGBoost*



Fonte: Autora.

O Gráfico 7 acima representa a comparação de viagens realizadas (azul) e de viagens previstas com o modelo *XGBoost* (vermelho). Observa-se que ele apresentou um resultado melhor que o Gráfico 4 e pior que os Gráficos 5 e 6. Dessa forma, o modelo conseguiu encontrar uma relação entre os dados de entrada, pois as colunas de viagens previstas seguem o mesmo formato das viagens realizadas, porém a previsão resultou em uma quantidade menor de viagens.

Deste modo, os modelos de Árvore de Decisão (Gráfico 5) e *Random Forest* (Gráfico 6) apresentaram resultados melhores comparados ao de Regressão Linear (Gráfico 4) e *XGBoost* (Gráfico 7). Porém, para identificar o modelo com melhor performance, também foram analisados os resultados das métricas de avaliação do desempenho, apresentados na Tabela 1 a seguir.

Tabela 1 - Métricas de avaliação de desempenho dos modelos

Métricas	Regressão Linear	Árvore de Decisão	*Random Forest*	*XGBoost*
MAE	1,664465	1,237867	1,074162	1,484255
MSE	12,401189	4,936242	3,147269	8,666466
RMSE	3,521532	2,221766	1,774054	2,943886
Score - treino	0,008153	1,000000	0,962666	0,303777
Score - teste	0,007610	0,604983	0,748144	0,306477

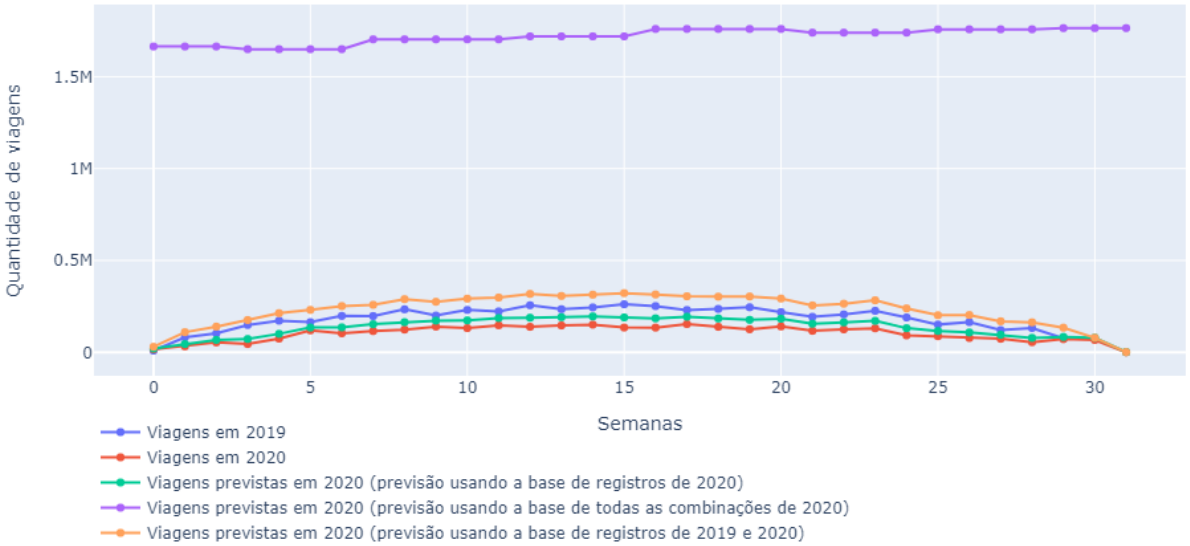
Fonte: Autora.

Como apresentado anteriormente, quanto menor o valor do MAE, MSE e RMSE, mais próximo o modelo está de acertar. Já para o Score, quanto maior esse valor, mais explicativo é o modelo. Com base nisso e comparando as métricas dos modelos, observa-se na Tabela 1 que o algoritmo *Random Forest* apresentou os melhores resultados para todas essas métricas, exceto para o Score – treino, que o modelo Árvore de Decisão ficou com 1. A respeito disso, ressalta-se que quando o modelo apresenta Score alto para os dados de treino e baixos para os dados de teste, isso demonstra que o modelo apresenta problemas de *overfitting*, não conseguindo generalizar. No caso do modelo de Árvore de Decisão, a diferença entre os Scores foi em torno de 0,40. Apesar de não ser um intervalo maior que 0,5, ele demonstra que o modelo está instável a novos dados, sendo necessárias ajustes para que ele demonstre melhor desempenho.

4.3 Previsão da demanda de 2020 e comparação com os dados reais

Para realizar a previsão da demanda de 2020 utilizou-se o modelo do *Random Forest*, visto que ele foi o que apresentou melhor desempenho. Assim, utilizou-se o comando .*predict()* para os 3 Xs apresentados anteriormente. O Gráfico 8 a seguir apresenta a quantidade de viagens obtidas em cada previsão e os valores reais de 2019 e 2020.

Gráfico 8 - Comparação das previsões de demanda para 2020



Fonte: Autora.

Observa-se no Gráfico 8 os seguintes pontos:

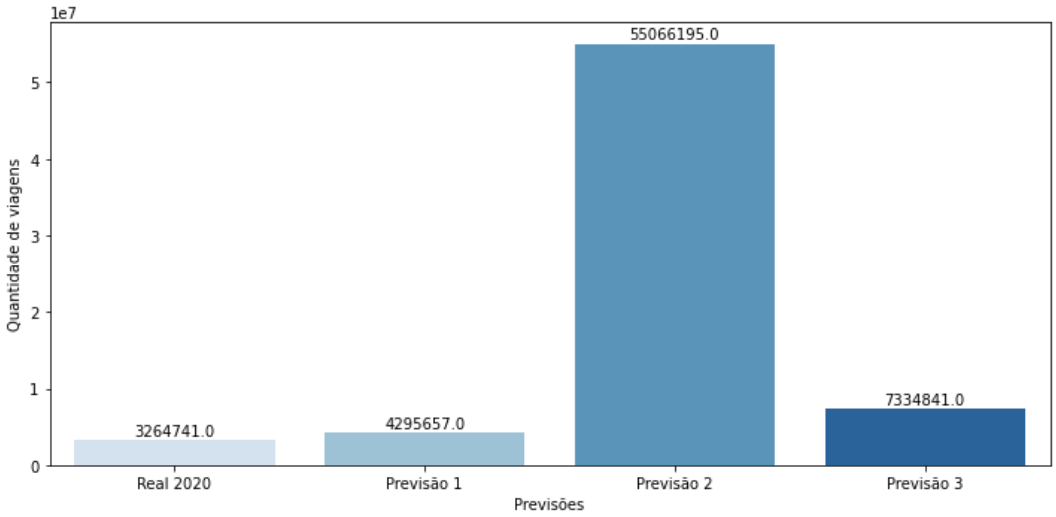
a) a curva da **Previsão 1** (usando a base de registros de viagens de 2020) ficou superior, porém bem próxima da curva de viagens reais de 2020. Provavelmente isso ocorreu por essa previsão utilizar o mesmo *dataset* de registro de 2020, tendo, portanto, a mesma quantidade de linhas. Assim, por causa do padrão dos dados de 2019, o modelo previu uma quantidade maior de viagens do que realmente ocorreu.

b) a curva da **Previsão 2** (usando a base simulada com todas as combinações possíveis de 2020) apresentou um resultado muito superior das outras curvas, assim como um formato diferente. Isso demonstra que o modelo superestimou muito a demanda, provavelmente por causa do tamanho da base utilizada, que contava com 26.296.384 linhas enquanto as outras tinham em média 2.493.534 linhas.

c) a curva da **Previsão 3** (usando a base simulada com a junção dos registros de 2019 e 2020) apresentou o melhor resultado entre as três previsões, ficando acima da curva de viagens reais de 2019, o que era esperado, se não fosse a pandemia, visto que a empresa finalizou 2019 com recordes de viagens, usuários e aumento de vendas.

Com base nisso, calculou-se a quantidade total de viagens reais e previstas por ano para cada uma das simulações. O Gráfico 9 a seguir apresenta os resultados.

Gráfico 9 - Quantidade total de viagens por ano – Real 2020 x Previsões



Fonte: Autora.

De forma geral, o Gráfico 9 confirma os pontos observados anteriormente. Analisando a diferença entre os valores a fim de identificar como a demanda da BIXI foi impactada pela pandemia de COVID-19, temos que:

a) na **Previsão 1**, a demanda real teria sido 32% menor do que o previsto;
b) na **Previsão 2**, a demanda real teria sido 1587% menor do que o previsto; e
c) na **Previsão 3**, a demanda real teria sido 125% menor do que o previsto.

Por fim, infere-se que os resultados da Previsão 2 foram muito altos, sendo pouco realistas. Já os resultados da Previsão 1 e 3 seriam os mais possíveis de ocorrer no ano de 2020, caso a pandemia de COVID-19 não tivesse ocorrido, sendo que a Previsão 1 demonstra um cenário mais conservador e a Previsão 2, um cenário mais otimista. Assim, pode-se dizer que a pandemia causou um impacto negativo entre 32% e 125% nas demandas da organização.

5 CONSIDERAÇÕES FINAIS

5.1 Conclusões

A previsão de demanda é um método para a determinação de dados futuros, sendo uma atividade estratégica para uma organização planejar e dimensionar os recursos necessários para a produção de bens e serviços de forma a atender sua demanda e não ter prejuízos. Visto que estima o futuro, a previsão de demanda pode ser impactada com alterações no ambiente externo, como foi o caso da pandemia de COVID-19 que impactou o funcionamento de diversas atividades. No caso do segmento de transporte, especificamente o de compartilhamento de bicicletas, a demanda costuma apresentar um crescimento temporário quando ocorre a interrupção do transporte público, porém poucos estudos analisaram o efeito de doenças transmissíveis na demanda desse serviço.

Assim, esse trabalho teve como objetivo geral analisar como a demanda da BIXI Montreal foi impactada pela pandemia de COVID-19, comparando os valores reais de 2020 com valores projetados utilizando modelos de aprendizado de máquina e dados históricos. Depreende-se que esse objetivo foi alcançado por meio da elaboração de quatro modelos de algoritmos diferentes e a utilização do modelo com melhor desempenho para fazer previsões da demanda de 2020, considerando dados de entrada diferentes.

Os objetivos específicos que auxiliaram na consecução desse objetivo geral também foram atingidos. Em relação a isso, conclui-se que o modelo utilizando o algoritmo *Random Forest* foi o que apresentou melhor desempenho entre os analisados. Apesar disso, esse modelo ainda pode ser ajustado a fim de apresentar melhores métricas. De forma geral, a capacidade de processamento e memória do computador utilizado foi um fator limitante durante esse trabalho, não sendo possível realizar testes alterando os parâmetros do modelo, nem utilizar uma série temporal maior ou com registros diários, porém isso é recomendado para processos futuros de melhoria do modelo.

Ademais, conclui-se que o modelo criado possibilitou várias previsões da demanda de 2020, o que permitiu uma comparação de diferentes cenários da demanda de viagens de bicicleta, caso a pandemia de COVID-19 não tivesse ocorrido.

A respeito das previsões, a Previsão 2, que utilizou como dados de entrada uma base simulada com todas as combinações possíveis de 2020, apresentou um resultado muito superestimado de demanda. A justificativa para isso é o tamanho maior da base utilizada. Apesar de ser uma forma interessante de simular os dados de entrada, pois utiliza informações que o gestor teria facilmente, seria preciso treinar e testar os modelos já utilizando bases nesse formato para avaliar o desempenho.

Por outro lado, as Previsões 1 e 3 apresentaram resultados mais realistas, podendo ser utilizadas como um intervalo possível de comportamento da demanda da organização. Visto que a demanda de um serviço é algo que pode ser influenciado pelo ambiente externo, a previsão de vários cenários é uma estratégia interessante de se adotar, pois cria uma margem de segurança para o gestor planejar e tomar decisões. Ressalta-se que a diferença nos resultados dessas previsões deu-se pelos dados de entrada utilizados, assim é importante que o gestor se atente à base que alimentará o modelo e aprimore a qualidade desses dados ao longo do tempo, para atingir resultados melhores.

Por fim, como o objetivo era identificar o impacto da pandemia na demanda da BIXI, comparando o real com os contrafactuais, estima-se que a demanda sofreu um efeito negativo entre 32% e 125% em quantidade de viagens, o que representa uma redução significativa da demanda.

5.2 Sugestões de trabalhos futuros

Com base nos resultados obtidos e oportunidades identificadas ao longo da realização desse trabalho, sugere-se que em trabalhos futuros sobre esse tema a previsão de demanda seja realizada considerando os registros diários de viagens, assim como uma série temporal maior seja utilizada para treinar o modelo. A hipótese é que com isso os modelos consigam prever melhor dados futuros, pois poderá analisar a tendência por um período maior, assim como identificar a sazonalidade no decorrer dos dias da semana, por exemplo.

Outra sugestão é realizar outros testes alterando os parâmetros dos modelos elaborados para analisar o comportamento do desempenho. Outros algoritmos diferentes dos que foram usados, como algoritmos de redes neurais, também podem ser testados.

Por fim, sugere-se que os modelos sejam testados utilizando bases com todas as combinações possíveis de cada ano. Além disso, para os casos de novas estações serem criadas de um ano para o outro, sugere-se que a previsão da demanda considere a tendência de estações próximas, similar a lógica de Wang et al. (2021), visto que provavelmente a demanda está relacionada a proximidade com lugares com muita movimentação de pessoas.

COMPARING MACHINE LEARNING METHODS FOR PREDICTING BIXI MONTREAL BICYCLE TRAVEL DEMAND AND ANALYZING THE EFFECT OF THE COVID-19 PANDEMIC ON 2020 DEMAND

ABSTRACT

Demand forecasting is a strategic activity for an organization to plan and dimension the resources necessary for the production of goods and services in order to meet demand and not have losses. It can be impacted by changes in the external environment, as was the case with the COVID-19 pandemic. In this sense, the objective of this work is to analyze how the demand of BIXI Montreal was impacted by the COVID-19 pandemic, comparing the real values of 2020 with values using Machine Learning models and historical data. For this, four models were developed using Python's LinearRegression, DecisionTreeRegressor, RandomForestRegressor and XGBRegressor algorithms. The performance of each model was evaluated considering the MAE, MSE, RMSE and Score metrics. The best model was chosen and used to forecast the 2020 demand and compare the real demand with a counterfactual. Overall, the Random Forest model showed the best performance and it was used to perform three different demand forecasts. One of the forecasts had a highly overestimated demand result. The other two forecasts showed more realistic results, demonstrating the possibilities of demand in a more conservative and a more optimistic scenario. Finally, it was concluded that BIXI Montrel's travel quantity demand was negatively impacted between 32% and 125% because of the pandemic.

Key words: Demand forecast. Machine learning. Regression. COVID-19.

REFERÊNCIAS

AZANK, Felipe. Como avaliar seu modelo de regressão: As principais métricas para avaliar seus modelos de regressão. Medium, 2020. Disponível em: <https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96>. Acesso em: 18 out. 2021.

BIXI. BIXI shatters records for 2019. Bixi, 2019. Disponível em: <https://bixi.com/en/media>. Acesso em: 30 set. 2021.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794, DOI:https://doi.org/10.1145/2939672.2939785. Acesso em: 21 out. 2021.

DELGADO FILHO, Antonio Jorge Ferreira. **Análise de métodos de regressão para previsão de demanda de curto prazo.** 2020. Dissertação (Mestrado) – Universidade Federal de Pernambuco, Recife, 2020. Disponível em: <https://repositorio.ufpe.br/handle/123456789/38123>. Acesso em: 05 out. 2021.

FACELI, K. et al. **Inteligência Artificial:** uma abordagem de aprendizado de máquina. 2. Ed. Rio de Janeiro: LTC, 2021.

FONTELLES, M. J. et al. **Metodologia da Pesquisa Científica:** Diretrizes para a Elaboração de um Protocolo de Pesquisa. Belém: UNAMA, 2009, p. 8.

FREITAS, Cristiana Caldeira Garcia de. **Demanda por seguro de automóvel no Rio de Janeiro.** 2018. Dissertação (Mestrado) – Fundação Getulio Vargas, Escola de Pós-Graduação em Economia, Rio de Janeiro, 2018. Disponível em: <http://bibliotecadigital.fgv.br/dspace/handle/10438/24096>. Acesso em: 05 out. 2021.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

GFBIOINFO. Árvore de decisão. Disponível em: http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id=199. Acesso em: 10 nov. 2021.

HEYDARI, S.; KONSTANTINOUDIS, G.; BEHSOODI, A. Effect of the COVID-19 pandemic on bike-sharing demand and hire time: Evidence from Santander Cycles in London. arXiv: 2107.11589, 2021. Acesso em: 04 out. 2021.

IBM. Random Forest. IBM Cloud Education, 2020. Disponível em: <https://www.ibm.com/cloud/learn/random-forest>. Acesso em: 21 out. 2021.

KUTNER, M. H. et al. **Applied linear statistical models.** 4th ed. New York: McGraw-Hill/Irwin series Operations and decision sciences, 2004.

LENZ, M. L. et al. **Fundamentos de aprendizagem de máquina.** Porto Alegre: SAGAH, 2020.

MARTINS, P. G; LAUGENI, F. P. **Administração da Produção.** 3 ed. São Paulo: Saraiva, 2015.

MATOS, Gonçalo Ribeiro de. **Machine learning aplicado à gestão de activos físicos industriais.** Dissertação (Mestrado) – Instituto Superior de Engenharia de Lisboa, Lisboa, 2021. Disponível em: <https://repositorio.ipl.pt/handle/10400.21/13524>. Acesso em: 22 out. 2021.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos Sobre Aprendizado de Máquina.** Sistemas Inteligentes Fundamentos e Aplicações. 1 ed. Barueri-SP: Manole Ltda, 2003.

NIKITAS, A. et al. Cycling in the Era of COVID-19: Lessons Learnt and Best Practice Policy Recommendations for a More Bike-Centric Future. **Sustainability**, v. 13, n. 9, p. 4620, 2021, DOI: https://doi.org/10.3390/su13094620. Acesso em: 04 out. 2021.

PELLEGRINI, F. R.; FOGLIATTO, F. S. Passos para implantação de sistemas de previsão de demanda: técnicas e estudo de caso. **Production**, v. 11, n. 1, p. 43–64, 2001, DOI: https://doi.org/10.1590/S0103-65132001000100004. Acesso em: 22 out. 2021.

REHBEIN, Matheus Henrique. **Estudo comparativo de deep learning e regressão linear na predição de mensagens processadas pela plataforma de integração Guaraná.** Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2019.

Disponível em: <https://bibliodigital.unijui.edu.br:8443/xmlui/handle/123456789/6705>. Acesso em: 22 out. 2021.

SATHISHKUMAR, V. E.; PARK, J.; CHO, Y. Using data mining techniques for bike sharing demand prediction in metropolitan city. **Computer Communications**, v. 153, p. 353-366, 2020, DOI: https://doi.org/10.1016/j.comcom.2020.02.007. Acesso em: 22 out. 2021.

SHOBHA, G.; RANGASWAMY, S. Chapter 8 - Machine Learning. In: SHOBHA, G.; RANGASWAMY, S. Handbook of Statistics. Elsevier, v. 38, 2018, p. 197-228, DOI: https://doi.org/10.1016/bs.host.2018.07.004. Acesso em: 04 out. 2021.

WANG, Xudong et al. Modeling bike-sharing demand using a regression model with spatially varying coefficients. **Journal of Transport Geography,** v. 93, p. 103059, 2021, DOI: https://doi.org/10.1016/j.jtrangeo.2021.103059. Acesso em: 22 out. 2021.

XU, M. et al. Decision tree regression for soft classification of remote sensing data. **Remote Sensing of Environment,** v. 97, n. 3, 2005, p. 322-336, DOI: <https://doi.org/10.1016/j.rse.2005.05.008>. Acesso em: 05 out. 2021.