Identificação e Medição de Fatores que Influenciam o Custo de Viagens Corporativas

João Marcelo Hemeritas Heusi
jmarcelo.heusi@gmail.com
Universidade do Vale do Itajaí
Itajaí, Santa Catarina, BRA

Wemerson Delcio Parreira
parreira@univali.br
Universidade do Vale do Itajaí
Itajaí, Santa Catarina, BRA

Anita Maria da Rocha Fernandes
Universidade do Vale do Itajaí
Itajaí, Santa Catarina, BRA

Rudimar Luis Scaranto Dazzi
Universidade do Vale do Itajaí
Itajaí, Santa Catarina, BRA

ABSTRACT

Great gains can be obtained in the efficiency of processes that have a large number of variables that influence their results through the use of machine learning techniques. This work introduces the use of machine learning techniques to obtain a better understanding of the factors that most influence the cost of corporate travel, as well as the construction of a behavioral suggestion model for decreasing this value. The optimization of this process can result in direct savings to companies that performed corporate travel. The final results of these implementations reached the following accuracy: prediction of air tickets with an R^2 value of 0.8 with a multiple linear regression model and an R^2 value of 0.74 with an artificial neural network model.

KEYWORDS

machine learning, corporate travel, suggestion model

1 INTRODUÇÃO

As viagens corporativas são viagens realizadas por funcionários de uma empresa em nome dessa com o propósito de atingir algum objetivo comercial pertinente ao negócio da empresa. Entre os principias objetivos comerciais estão o atendimento ao processo operacional, estreitamento de parcerias, reuniões, missões corporativas e elevação ou consolidação da carteira de clientes [1].

No cenário corporativo, a compra de viagens normalmente ocorre seguindo um processo estabelecido pela própria empresa ou por empresas terceiras, como agências de viagem. Esses processos padronizam as solicitações, cotações, aprovações e emissão delas. É possível utilizar sistemas e modelos de análise de comportamento de usuários para as mais variadas áreas de negócio e empresas. No contexto da compra de viagens corporativas, estes sistemas podem identificar os principais fatores que possam gerar economia na hora de realizar a compra destas viagens, por exemplo. A busca por soluções que utilizam aprendizado de máquina é incentivada pelo crescente número de dados e pela rápida adoção do mercado à serviços baseados em nuvem [2].

As viagens corporativas representam uma parcela significativa dos gastos totais das grandes corporações [3][4]. Por esse motivo, existe uma grande oportunidade pouco explorada para uma otimização no processo de compra das viagens corporativas, visando gerar economia para todas as empresas que possuem dentro de seus processos essas viagens. A identificação de pequenos fatores que influenciam no custo destas viagens tem potencial de gerar centenas de milhares de reais de economia para as grandes empresas, tornando-as mais lucrativas e competitivas.

Assim, para este trabalho, são utilizados dados referentes a compras de passagens aéreas. O motivo desta delimitação se dá pelo fato deste produto ter a maior representatividade dentro dos custos de viagens corporativas, aproximadamente 65% do montante total gasto [5].

Neste trabalho, é proposta a avaliação dos fatores presentes no processo de compra de viagens corporativas, através de diferentes métodos, ferramentas estatísticas e da implementação de modelos de aprendizado de máquina para descobrir se existem padrões que podem ser identificados e que influenciam os custos de viagens corporativas. A partir desta análise então, é proposto construir um modelo, também utilizando métodos de aprendizado de máquina, que otimiza a compra de viagens corporativas.

Uma das áreas dentro do estudo de Inteligência artificial é o aprendizado de máquinas. O objetivo desta área é o desenvolvimento e promover as técnicas computacionais sobre o aprendizado e também a construção de sistemas que possam adquirir conhecimento de forma automática. Para adquirir tal conhecimento, o aprendizado de máquinas funciona a partir de algoritmos de tomada de decisão baseados em experiencias acumuladas através de soluções bem sucedidas e mal sucedidas anteriores. Existem diversos métodos dentro aprendizado de máquina que podem ser utilizados para reconhecimento de padrões, categorização, previsão de valores e diversas outras tarefas [6].

As redes neurais artificiais (RNAs), são modelos matemáticos criados partindo do funcionamento do cérebro humano [7]. Tal como o cérebro humano, a RNA se baseia nas experiências para a construção de aprendizado. As RNAs, através de um banco de dados e com uma análise computacional é capaz de identificar padrões, realizar classificações, dentre outras funções [7, 8].

A regressão é um dos conjuntos de métodos mais comuns e simples de aprendizado de máquina supervisionado. É usualmente utilizado em casos em que a variável de saída do sistema são é quantitativa discreta ou quantitativa contínua. Funciona de forma similar ao modelo de Regressão Linear (RL) múltipla.Este método opera através da construção de uma função linear que minimiza o erro quadrático R^2 e que consequentemente se adéqua da melhor forma a todos os pontos do conjunto de dados utilizado para o treinamento [9].

Este trabalho está organizado como se segue, na Seção 2 apresenta uma revisão da literatura com os trabalhos correlatos. A Seção 3 apresenta uma visão geral das técnicas usadas neste trabalho. Os resultados e a discussão é apresentada na Seção 4. O trabalho é finalizado na Seção 5.

2 TRABALHOS RELACIONADOS

Durante a etapa de pesquisa bibliográfica foram analisados artigos que apresentam soluções técnicas relacionados à solução de problemas de mercado. Apesar de terem sido empregados em áreas diferentes de estudo deste trabalho, são utilizados de forma análoga. Os trabalhos relacionados são:

(1) O trabalho intitulado *On neural network modeling to maximize the power output of PEMFCs* apresenta uma abordagem similiar a deste trabalho, propondo que a operação ótima de células de combustível devem ser utilizadas para que sejam atingidas suas máximas eficiências, para isso foi implementada uma RNA, além disso, foram investigados os efeitos dos parâmetros importantes e a sua iteração com a variável de saída (a ser otimizada) que se desejava otimizar, registraram um ganho de 23, 6% na potência de saída das células de PEMFCs [10].
(2) O trabalho intitulado "Proposta de um novo modelo de regressão linear para a previsão e controle do indicador de eficiência energética de uma empresa ferroviária" emprega o uso das técnicas estatísticas que para compreender quais são os fatores que mais interferem na variável de saída, em um contexto de eficiência energética de uma empresa ferroviária, foi obtido um modelo matemático que que permite avaliar o impacto na variável de eficiência energética [11] .
(3) O trabalho intitulado "Estimation and prediction of construction cost index using neural networks, time series, and regression" utiliza diferentes técnicas de aprendizado de máquina para se obter um modelo matemático adequado ao comportamento do índice de custo de obras e também para conseguir realizar previsões quanto ao índice de custo de obras dos anos seguintes, foram empregas RL e RNAs e o método de série temporal auto-regressiva, a média absoluta de erros obtida dos métodos de aprendizado de máquina foi de 8, 3 para as previsões realizadas pela RNA e de 17, 5 para as previsões realizadas pelo método de RL [12].

3 MATERIAIS E MÉTODOS

O trabalho tem como foco a identificação dos fatores presentes no *dataset* de maior correlação com o custo de viagens corporativas, como também, a construção de um modelo sugestivo de compras de viagens visando a obtenção do menor custo. Os fatores do *dataset* são categorizados de acordo com seu nível de correlação e se a mesma é positiva ou negativa. Para a realização da identificação dos fatores de maior correlação, foi empregada a RL múltipla. Já para a construção do modelo sugestivo de compras, foram empregadas duas técnicas de aprendizado de máquina, e seus resultados foram comparados quanto à sua capacidade de prever o valor de viagens corporativas.

A partir da análise destes resultados de previsão, foi selecionada a técnica com maior acurácia para a construção do modelo sugestivo, com o objetivo de alcançar o menor custo para as viagens corporativas compradas. Para o desenvolvimento deste trabalho foram utilizadas as implementações das bibliotecas `TersonFlow` e `SciKitLearn` para a linguagem de programação Python. As justificativas para a escolha destas tecnologias é abordada a seguir. A Figura 1 apresenta a sequência de passos seguida para a implementação do projeto.

3.1 Implementação das análises estatísticas

Para a implementação do algoritmo responsável por gerar as análises estatísticas, foram utilizadas as bibliotecas `statistics` e `scipy`. Da primeira biblioteca, Statistics, foram utilizados os métodos `mean`, `median` e `stdev` para gerar análises de média, mediana e desvio padrão. Da biblioteca `Scipy` foram utilizados os métodos pearsonr e levene, para gerar os testes de correlação de Pearson e o teste de igual variância de Levene. Para cada tipo de variável de entrada do *dataset*, um conjunto diferente de análises estatísticas é realizado.

Para as variáveis booleanas, são geradas as analises de valor médio para os valor Verdadeiro da variável e também para o valor Falso da variável. Também é realizado o teste de igual variância de Levene para verificar se há diferença estatística entre os dois tipos de valor possíveis para a variável. Desta análise são retornados os valores calculados, como também o valor p do teste de Levene.

Para as variáveis numéricas, é realizado o teste de correlação de Pearson. Desta análise são retornados o valor de r de correlação, como também o valor p do teste de Pearson.

Já para as variáveis categóricas, também é realizado o teste de igual variância de Levene para todos os valores possíveis. Além disto, são retornadas as informações de média, mediana e desvio padrão dos valores categóricos com maior e menor média de valores.

3.2 Implementação da RL:

O algoritmo de RL foi implementado a partir da biblioteca Scikit-learn para Python. O módulo da biblioteca utilizado foi o de modelos lineares (sklearn.linear_model). O modelo de regressão utilizado foi o implementado na classe LinearRegression, que implementar o algoritmo de RL utilizando o método dos mínimos quadrados.

Da classe `LinearRegression`, foi utilizado o método `fit`, que recebe dois parâmetros, os dados de entrada do *dataset* de treinamento e os dados de saída do mesmo *dataset*. Este método realiza o ajuste dos pesos do coeficiente do modelo de forma que seja minimizada a soma residual dos quadrados entre os quadrados observados no *dataset* e os previstos pela aproximação linear.

A Figura 2 mostra o diagrama dos passos executados pelo algoritmo da RL. A seguir é detalhado o funcionamento de cada um dos blocos apresentados na 2:

(1) O software faz a requisição para o banco de dados `MongoDB` buscar os documentos do *dataset* armazenados na nuvem e os recupera.
(2) O *dataset* passa pelo algoritmo de tratamento de dados.
(3) O *dataset* é dividido em dois conjuntos separados, um para a realização do treino e o outro para testes. A proporção utilizada para este trabalho foi de 90% para a realização do treinamento e 10% para os testes.

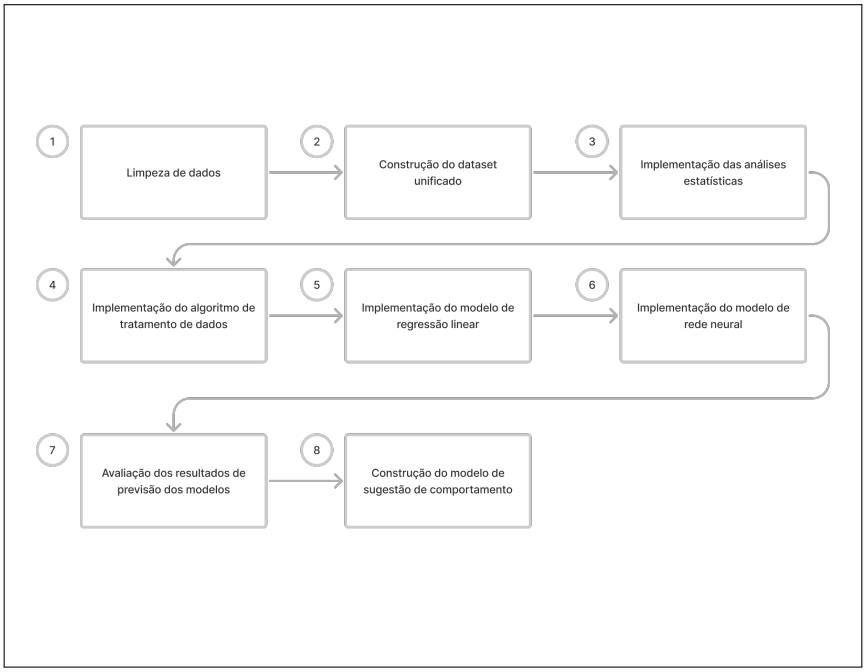

Figura 1: Fluxo do plano de implementação.

(4) É executada o método fit recebendo como parâmetros os dados de entrada e saída do *dataset* de treinamento.
(5) São realizadas as previsões utilizando o modelo gerado pelo método fit, as previsões são realizadas utilizando o *dataset* de testes.

3.3 Implementação das RNAs

O algoritmo implementado para a construção do modelo de rede neural utilizado utiliza as bibliotecas Keras. Foram utilizados dois módulos da biblioteca para a implementação do modelo de rede, Models (`keras.models`) e Layers (`keras.layers`).

O modelo de RNA implementado foi um modelo sequencial de rede neural profunda, utilizando como base a classe `Sequential` do módulo `Models` da biblioteca Keras. Esse modelo tem suas camadas organizadas de forma linear e sequencial. O modelo implementado possui 7 camadas ao todo. Todas as camadas do modelo são camadas densa com a implementação da classe Dense do módulo `Layers` da biblioteca Keras.

A primeira camada é a camada de entrada e possui 15 neurônios, um para cada uma das variáveis de entrada. A rede neural ainda conta com outras 5 camadas ocultas com 15 neurônios cada. Por fim, a camada de saída possui apenas um neurônio. Todas as camadas utilizam a função de ativação ReLU.

A Figura 3 mostra o diagrama dos passos executados pelo algoritmo da RL.

A seguir são detalhados os funcionamentos de cada um dos blocos apresentados na Figura 3:

(1) O software faz a requisição para o banco de dados MongoDB buscar os documentos do *dataset* armazenados na nuvem e os recupera.

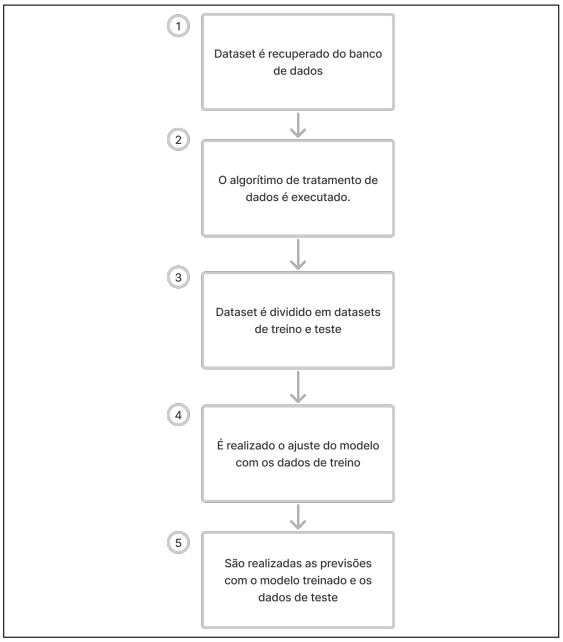

Figura 2: Diagrama em blocos dos passos executados pelo algoritmo de RL

(2) O *dataset* passa pelo algoritmo de tratamento de dados.
(3) O *dataset* é dividido em três conjuntos de dados, um utilizado para o treinamento do modelo de rede neural, o segundo é utilizado para a validação do modelo de rede

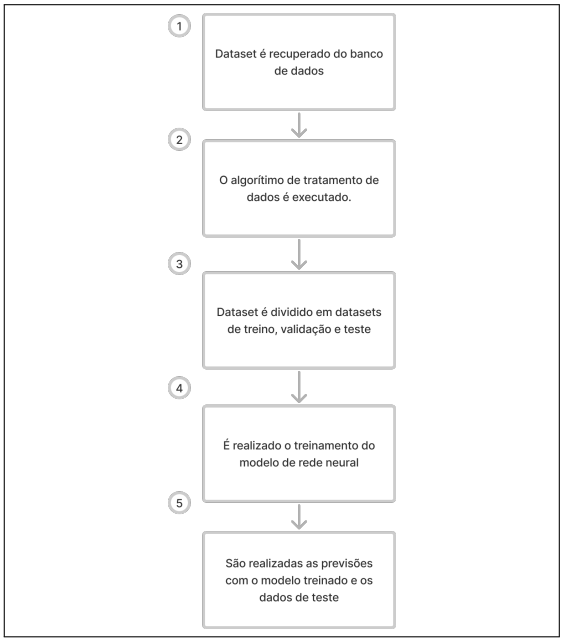

Figura 3: Diagrama em blocos dos passos executados pelo algoritmo da RNA

neural e o terceiro é utilizado para a realização dos testes. A proporção utilizada para este trabalho foi de 80% para a realização do treinamento, 10% para a validação e 10% para os testes.

(4) É criado o modelo de rede neural com as especificações citadas na Seção 3.3. Em seguida é realizado o treinamento da rede neural, utilizando como parâmetros do método fit os conjuntos de dados de treinamento e validação.

(5) São realizadas previsões utilizando o *dataset* de testes como parâmetro.

3.4 Implementação do modelo de sugestão de comportamento

O modelo de sugestão de comportamento foi construído a partir dos resultados obtidos através das implementações dos modelos de aprendizado de máquina. Para a construção deste, foram utilizados os coeficientes de cada uma das variáveis geradas pelo modelo ajustado de RL. Cada coeficiente foi categorizado como tendo um impacto positivo ou negativo sobre o valor das passagens através do sinal do coeficiente. A grandeza deste impacto que as variáveis possuem no valor das passagens também foi determinada pelo módulo de cada um destes coeficientes.

O retorno fornecido pelo modelo de sugestão, portanto, indica para o usuário qual a relação de cada uma das variáveis de entrada com o valor de passagens aéreas. Esta indicação é dada com frases que seguem o padrão: "Aumentando o valor da variável × o custo das passagens aéreas irá diminuir", "Diminuindo o valor da variável *Y* o custo das passagens aéreas irá diminuir", "O preço das passagens aéreas esta no seu menor valor quando a variável qualitativa *Z* possui um destes valores: [A,B,C]" e também "O preço das passagens aéreas está no seu maior valor quando a variável qualitativa W possui um destes valores: [D,E,F]".

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos no trabalho. Os resultados são provenientes das análises estatísticas, da RL, da RNA e do modelo de sugestão.

4.1 Resultados das análises estatísticas

Os tipos de análises estatísticas realizadas neste trabalho foram: média, mediana, desvio padrão, teste de igual variância de Levene e o teste de correlação de Pearson. A base de dados utilizada para a geração das análises estatísticas é composta por todo o *dataset*, contando com 30696 amostras de dados de vendas de passagens aéreas. Conforme mencionado na Seção 3.1, para cada tipo de variável do *dataset*, foram realizadas analises diferentes. Para as variáveis categóricas, com o objetivo de avaliar se existia diferença no preço da passagem aérea entre os valores que cada variável pode possuir, foi utilizado o teste de igual variância de Levene, como também, foram levantados as médias de valores de preço das passagens aéreas para cada valor.

É possível observar na Tabela 1 que para todas as variáveis o valor de p do teste de igual variância estatística são próximos a 0 ou são 0. Valores baixos de p sugerem que os conjuntos de dados não possuem igual variância. Também é possível analisar que a média dos valores de passagens aéreas variam diferem bastante quando comparamos os valores que produzem a maior média com os valores que produzem a menor média.

Com as variáveis numéricas, foi realizado o teste de correlação de Pearson. O teste foi realizado passando como entrada o conjunto de valores da variável e o conjunto de valores de passagens aéreas. O resultado produzido pelo teste é o valor de R gerado pelo cálculo de correlação e o valor de p da hipótese de teste. Para os testes de correlação de Pearson, o valor de R varia de -1 à 1, em que os valores próximos a estes limites indicam respectivamente uma correlação forte negativa e uma correlação forte positiva, valores próximos a 0 indicam que não existe correlação positiva ou negativa. Valores elevados (maior do que 1) de p retornados pelo teste de correlação de Pearson sugerem que há correlação linear entre as variáveis, valores baixos (próximos a zero) de p sugerem que não há correlação entre as variáveis.

Podemos observar que o valor de R das variáveis *lowestPrice* e *journeys* possuem os maiores valores de R, sugerindo que dentre as variáveis númericas, estas possuem o maior impacto no preço das passagens aéreas.

Para as variáveis booleanas, também foi realizado o teste de igual variância de Levene, como também foram extraídas a média dos valores de passagens aéreas quando as variáveis assumem valores verdadeiros e a média dos valores das passagagens aéreas quando as variáveis assumem valores falsos.

Neste caso, pode ser observado que para as variáveis *isInternational* e *policy3* os valores de p do teste de igual variância são baixos, sugerindo que não existe igual variância entre os conjuntos de dados. Diferentemente, os valores de p para as variáveis *policy1* e *policy2* o valor é mais alto, sugerindo que existe igual variância entre os conjuntos de dados. Assim, pode-se esperar que as variáveis

Tabela 1: Análises estatísticas das variáveis categóricas

Variável	Levene p-value	Valor que produz maior média	Valor que produz menor média	Maior média	Menor média
supplier	0	EMIRATES	HAHN AIR	14012, 47	492, 43
departureStation	0	DUS	CUN	18962, 71	59, 2
arrivalStation	-2×10^{218}	GDL	TRC	12654, 66	244, 09

Tabela 2: Análises estatísticas das variáveis númericas

Variável	Pearson R	Pearson p_value
flightDuration	0, 17	$-6, 42 \times 10^{204}$
journeys	0, 25	$-3, 61 \times 10^{69}$
timeToApprove	$-0, 01$	0, 06
lowestPrice	0, 67	0, 72
highestPrice	0, 007	0, 18

isInternational e *policy3* tenham maior influência sobre o preço das passagens aéreas.

Tabela 3: Análises estatísticas das variáveis booleanas

Variável	Levene p-value	Média quando verdadeiro	Média quando negativo
isInternational	0	6451, 73	877, 03
policy1	0, 45	965, 91	951, 66
policy2	0, 26	968, 13	956, 8
policy3	0, 098	817, 69	1058, 37

4.2 Resultados da RL

De acordo com Seção 3.2, para o treinamento do modelo, o *dataset* original foi dividido em 2 grupos distintos, um para treinamento e outro para a realização dos testes. O *dataset* utilizado para apresentar os resultados desta seção foi o *dataset* de testes. A forma como a separação dos dados foi realizada é apresentada na Seção 3.2.

Os resultados foram obtidos a partir da execução do método predict da classe `LinearRegression` à qual o modelo ajustado pertence. O método `predict` recebe como parâmetro um *dataset*. Para realizar as predições foi utilizado os dados do *dataset* de testes como parâmetro de entrada. Ao final da execução, o método *predict* retorna os valores previstos para cada uma das entradas.

Com os resultados gerados pelo método predict, foi calculado o erro absoluto médio (MAE) do modelo e o valor de R^2 do modelo. Estes valores foram gerados utilizando os métodos `mean_absolute_error` do módulo `metrics` da biblioteca Scikit-learn e o método score da classe `LinearRegression` respectivamente. Os resultados obtidos no modelo estão apresentados na Tabela 4. Nesta, é possível conjecturar de que existe uma linearidade nas amostras de compras de passagem do dataset utilizado neste trabalho. Nos objetivos deste trabalho não foram definidos valores alvo por conta dos trabalhos correlatos utilizarem outros *datasets* de outras áreas de mercado. Portanto, a avaliação quanto à qualidade dos resultados obtidos precisa ser realizada em comparação ao outro modelo construído neste trabalho, a rede neural.

Tabela 4: Métricas de desempenho do modelo de RL.

Métrica	Valor
MAE	179, 06
R^2	0, 8

Tabela 5: Dados estatísticos do preço das passagens aéreas

Métrica	Valor
Quantidade	30696
Média	962, 93
Mediana	760, 16
Desvio padrão	1338, 58

Quanto ao valor de erro médio absoluto, o modelo de RL realiza a previsão do custo das passagens aéreas com uma média de erro de R$179,06. Como o desvio padrão do próprio *dataset* e os preços das amostras variam da casa de dezenas de reais a valores próximos a 5 mil reais, como pode ser visto na Tabela 5 e Figura 4 respectivamente. Pode-se considerar que o resultado obtido é satisfatório.

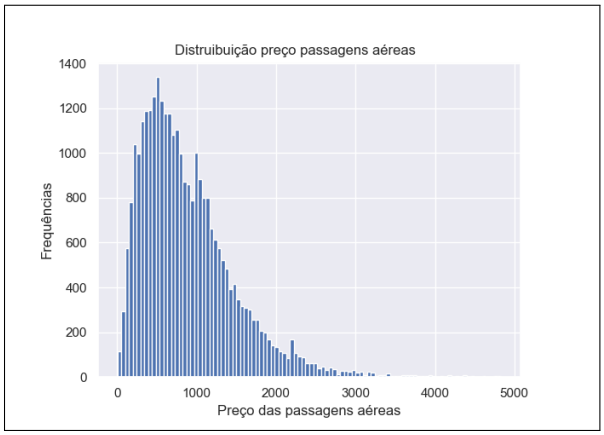

Figura 4: Gráfico de distribuição dos valores de passagens áreas por frequência de aparecimento no *dataset*.

A Figura 5 mostra o diagrama de dispersão relacionando os valores previstos pelo modelo de RL com os valores reais do *dataset* de testes, nos permitindo observar quão distantes os valores previstos

estão do valor real as previsões se encontram. Para adicionar à esta avaliação, na tabela 5, estão algumas informações do preço das passagens aéreas. Na Figura 4 também pode ser observada a distribuição dos valores de passagens aéreas.

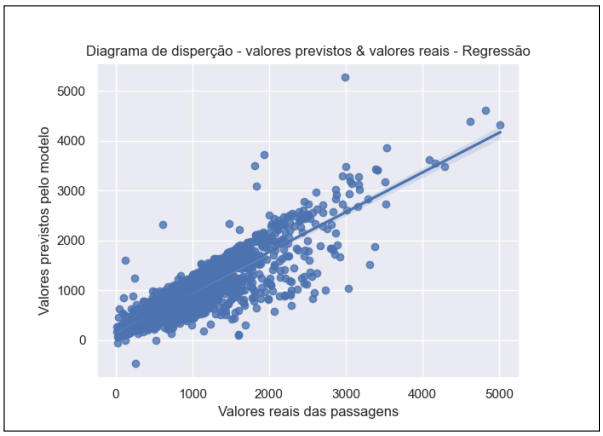

Figura 5: Diagrama de dispersão relacionando os valores previstos pelo modelo de RL com os valores reais do *dataset* de testes

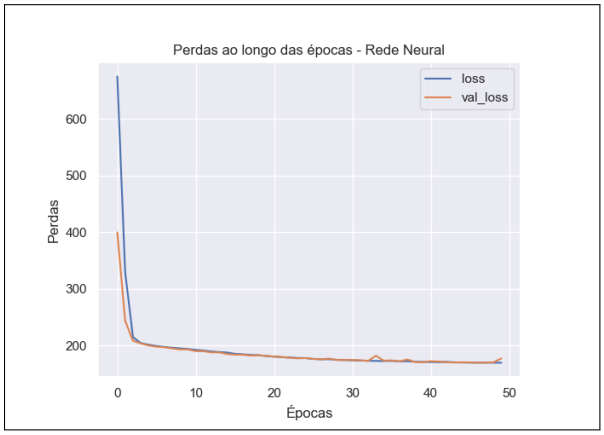

Figura 6: Gráfico do comportamento das taxas de erro ao longo das épocas de treinamento.

Tabela 6: Métricas de desempenho do modelo da rede neural.

Métrica	Valor
MAE	174, 09
R^2	0, 74

4.3 Resultados da RNA

Como apresentado na Seção 3.3, o dataset foi dividido em três grupos, um para treino, um para validação e um para testes. O dataset utilizado para a construção dos resultados apresentados nessa seção é o *dataset* de testes, contendo 10% das amostras do dataset original. A forma como a separação dos dados foi realizada também está apresentada na Seção 3.3.

Foram utilizadas 50 épocas para o treinamento do modelo de rede neural, a quantidade de épocas foi decidida usando como parâmetro a estabilização da taxa de perda de treinamento e validação. A taxa de perda do modelo atingiu a estabilização com um valor próximo à 173. O R^2 alcançado pelo modelo utilizando como entrada os dados de teste foi de 0, 76. Para que não exista *overfitting* a taxa de perda deve se estabilizar e a taxa de perda do treinamento e validação devem ser similares. Tal comportamento pode ser observado na Figura 6 em que é apresentado um gráfico com a variação das taxas de perda de treinamento e validação ao longo das épocas de treinamento

A Tabela 6 apresenta os valores de erro médio absoluto e R^2 obtidos nos testes realizados com o modelo implementado de rede neural. De forma similar à análise realizada sobre os resultados obtidos com o modelo de RL, o objetivo da implementação do modelo de rede neural não era de atingir um percentual de acurácia específico ou de realizar melhorias partindo de modelos já implementados. Portanto, da mesma forma como o resultado obtido com o modelo de RL foi avaliado como positivo para o trabalho, este também foi considerado desta forma.

Uma hipótese levantada durante a etapa de desenvolvimento foi a de que a rede neural poderia obter um resultado superior de forma expressiva frente ao modelo de RL, entendendo que a rede neural poderia conseguir identificar padrões não lineares no modelo. No entanto, como abordado na Seção 4.4 à seguir, ambos os modelos obtiveram métricas de desempenho muito próximas. A não obtenção de um resultado com desempenho superior no modelo de rede neural não foi considerado como um problema, mas sim como uma possível evidência a mais de que o conjunto de variáeis no *dataset* possui uma influência linear sobre o valor das passagens aéreas.

A Figura 7 mostra o diagrama de dispersão relacionando os valores de saída previstos com os valores reais. Os valores previstos foram gerados utilizando como entrada o *dataset* de testes.

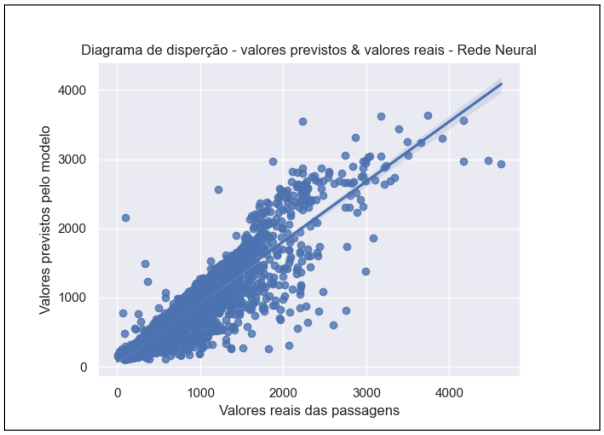

Figura 7: Diagrama de dispersão relacionando os valores previstos pelo modelo da rede neural com os valores reais do *dataset* de testes.

4.4 Comparação dos resultados produzidos pelos diferentes modelos do trabalho

Para a comparação entre os resultados produzidos pelos modelos foram utilizadas algumas métricas de cada um destes modelos – o valor do MAE produzido por cada um deles, o valor do R^2 do modelo, o tempo de execução dos algoritmos destes modelos – que estão apresentados na Tabela 7. Nesta é possível observar que o valor de R^2 e o MAE estão próximos. Existe uma diferença de $0,06$ no valor de R^2 obtido nos modelos, onde a RL teve um melhor desempenho. Já observando o MAE obtido nos modelos, existe uma diferença absoluta de $4,97$, onde a rede neural mostrou melhor acurácia na previsão de valores. Este trabalho avaliou estes resultados próximos como uma demonstração de natureza parcialmente linear das variáveis de entrada com o valor das passagens aéreas dentro do conjunto amostral de compras de passagens aéreas do *dataset* utilizado. Porém, o tempo de execução da RNA foi superior a 3x o tempo da RL, em virtude da natureza mais complexa do método. Apesar da execução dispensar requisitos de tempo real, em virtude na natureza offline da análise, é importante para uma tomada de decisão em virtude do esforço computacional empregado

Tabela 7: Métricas de desempenho dos modelos de RL e rede neural

Modelo	Valor do R^2	MAE	Tempo para execução (segundos)
RL	$0,8$	$179,06$	$7,36$
Rede Neural	$0,74$	$174,09$	$23,65$

Assim, devido ao melhor desempenho do modelo de RL no valor obtido no R^2 e pelo tempo menor de execução, os resultados e coeficientes produzidos por este modelo foram utilizados para o desenvolvimento do modelo de sugestão de comportamento.

4.5 Comparação dos resultados obtidos com os trabalhos correlatos

Para a avaliação comparativa entre os trabalhos correlatos e este trabalho, foram utilizadas as informações: media absoluta de erro em relação ao valor médio do valor, valores de R^2 e R^2 ajustado. A Tabela 8 apresenta o comparativo da previsão de valores de R^2 deste trabalho com os diferentes modelos implementados com o valor de R^2 obtido pelo trabalho realizado por [11]. De forma análoga, a Tabela 9, apresenta um comparativo dos valores de R^2 do modelo elaborado neste trabalho com o trabalho de [12].

Os resultados obtidos neste trabalho foram inferiores numericamente, no entanto, como a natureza dos *datasets* e o conjunto de dados utilizados não são os mesmos, a comparação destes resultados não tem como objetivo averiguar se um modelo performa melhor que o outro, e sim de avaliar como os modelos implementados desempenham com o *dataset* deste trabalho utilizando as mesmas métricas que os trabalhos correlatos utilizaram.

Quando analisamos os resultados obtidos por [12], podemos observar que o resultado obtido pelo modelo de RNA foi superior na métrica média percentual absoluta de erro ao modelo de RL. Este mesmo comportamento pôde ser observado neste trabalho, em

Tabela 8: Métricas de desempenho dos modelos de RL e rede neural

Trabalho	Técnica	Objetivo	Resultado
[11]	RL	Construção do modelo para previsão do valor de E.E	Previsão do valor de E.E. com R^2 de 93,8%
Este trabalho	RL	Prever o valor das passagens aéreas	Previsão do valor das passagens aéreas com R^2 de 80%
Este trabalho	RNA	Prever o valor das passagens aéreas	Previsão do valor das passagens aéreas com R^2 de 74%

que a RNA também obteve uma média percentual absoluta de erro inferior ao modelo de RL implementado.

4.6 Resultados do modelo de sugestão

Conforme apresentado na Seção 3.4, os resultados do modelo de sugestão construído dependem diretamente dos resultados obtidos no modelo de RL. Como mencionado também na Seção 3.4, com o modelo de RL construído, é possível extrair o módulo e sinal de cada um dos coeficientes vinculados à cada uma das variáveis de entrada.

Com esses coeficientes extraídos, eles foram classificados entre tendo uma correlação positiva ou negativa com o valor de passagens aéreas. Os módulos dos valores desses coeficientes também foi normalizado em uma escala de 0 a 100 para demonstrar o grau de impacto de cada um dos coeficientes no valor das passagens aéreas. A Tabela 10 exibe os resultados obtidos e a classificação realizada.

Depois de realizado esse processo, são geradas mensagens de resposta para o usuário. Nestas mensagens fica escrito como uma variação em uma variável de entrada impactaria o valor das passagens aéreas. No caso das variáveis categóricas, é construída uma mensagem identificando quais são os melhores valores para elas de forma a diminuir os valores das passagens aéreas. A Figura 8 exibe as mensagens que são retornadas para o usuário.

Além disso, foi possível obter através da implementação dos modelos de rede neural e RL, a determinação de quais fatores possuem o maior impacto no custo das passagens aéreas, como também foi um valor normalizado para este impacto e se o mesmo é positivo ou negativo.

Por fim, foi gerado um modelo de sugestão de comportamento, que entrega mensagens humanizadas para o usuário, que o permite compreender como as alterações e variações em cada uma das variáveis irá impactar o custo das passagens aéreas, como também permite que ele visualize qual o impacto de cada uma delas.

5 CONCLUSÃO

Para o desenvolvimento deste trabalho foram implementados dois algoritmos de aprendizado de máquina diferentes, utilizando a linguagem de programação Python para realizar a previsão de preços de passagens aéreas: RL Múltiplas e RNAs.O dataset utilizado para para a implementação das análises estatísticas e para o treinamento

Tabela 9: Métricas de desempenho dos modelos de RL e rede neural

Trabalho	Técnica implementada	Objetivo	Resultado
[12]	RNA	Prever o valor do CCI	Previsão com média percentual absoluta de erro de 8, 3%
[12]	Regressão Linear	Prever o valor do CCI	Previsão com média percentual absoluta de erro de 17, 5%
Este trabalho	RNA	Prever o valor das passagens aéreas	Previsao com média percentual absoluta de erro de 25, 6%
Este trabalho	Regressão Linear	Prever o valor das passagens aéreas	Previsao com média percentual absoluta de erro de 31, 9%

Tabela 10: Métricas de desempenho dos modelos de RL e rede neural

Variável	Coeficiente normalizado	Correlação
isInternational	100	Negativa
journeys	10,58	Positiva
requestMonth	3,73	Positiva
issueMonth	2,4	Negativa
policy3	1,32	Positiva
policy2	0,9	Negativa
policy1	0,63	Positiva
flightDuration	0,41	Negativa
lowestPrice	0,28	Positiva
daysInAdvance	0,1	Negativa
daysToIssue	0,08	Negativa
timeToApprove	0,01	Negativa
highestPrice	0,00	Negativa

Figura 8: Resultados de saída das sugestões de comportamento geradas pelo modelo.

dos modelos de aprendizado de máquina implementados – com 30.696 amostras de compras de passagens aéreas – foi fornecido por uma empresa do ramo de viagens.

A acurácia dos dois modelos implementados se assemelha muito, sendo que a Rede Neural obteve um valor um pouco inferior de erro médio absoluto quando comparado com a RL. Já a RL obteve um R^2 superior ao da rede neural. Embora estas métricas se assemelhem bastante, há uma diferença um pouco maior quanto ao tempo que cada um dos algoritmos leva para executar.

Analisando o resultado obtido e a aplicação das técnicas utilizado, é possível fornecer recomendações para obterem economias na redução de custos através do modelo de sugestão de comportamento.

Considera-se como trabalhos futuros a exploração de outras arquiteturas de modelos de aprendizado de máquina para o aprimoramento dos resultados.Adicionalmente, considera-se análise de outros produtos de turismo para avaliação da correlação dos fatores de maior impacto no preço de hospedagens, locações de carro e afins.

REFERÊNCIAS

[1] FLY AEOLUS. 2017 business travel statistics. 2017. URL https://flyaeolus.com/blog/2017-business-travel-statistics/.
[2] MRFR. Global machine learning market research report. 2020. URL https://www.marketresearchfuture.com/reports/machine-learning-market-2494.
[3] SALESTRIP. Share of gdp generated by business tourism spending worldwide from 1995 to 2019. 2021. URL https://www.statista.com/statistics/1194725/business-tourism-share-of-gdp/.
[4] SALESTRIP. How much do companies spend on business travel? 2020. URL https://cdn2.hubspot.net/hubfs/5132914/Reports/2020_Business_Travel_Trends_Whitepaper.pdf.
[5] ABRACORP. Pesquisa de vendas da abracorp 2019. 2019. URL https://www.abracorp.org.br/bi-2019.
[6] John D Kelleher, Brian Mac Namee, and Aoife D'arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* MIT press, 2020.
[7] S HAYKIN. *Redes Neurais: Princípios e Prática.* Artmed, 2007. ISBN 9788577800865.
[8] D S ACADEMY. Deep learning book. 2007. URL https://www.deeplearningbook.com.br/.
[9] RUSSELL and P NORVIG. Artificial intelligence: A modern approach. 2009.
[10] Fereshteh Salimi Nanadegani, Ebrahim Nemati Lay, Alfredo Iranzo, and J. Antonio Salvaand Bengt Sunden. On neural network modeling to maximize the power output of pemfcs. 2020.
[11] G ROCHA. Proposta de um novo modelo de regressão linear para previsão e controle do indicador de eficiência energética de uma empresa ferroviária. 2013.
[12] Y Elfahham. Estimation and prediction of construction cost index using neural networks, time series, and regression. 2019.