Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Carlos Roberto da Conceição

Modelo para predição de preço de passagens aéreas domésticas

Curitiba
2023

Carlos Roberto da Conceição

Modelo para predição de preço de passagens aéreas domésticas

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Walmes Marques Zeviani

Curitiba
2023

Modelo para predição de preço de passagens aéreas domésticas

Model for predicting domestic airfare prices

Carlos Roberto da Conceição

Os custos com viagens corporativas têm aumentado consideravelmente nos últimos anos, devido ao aumento dos preços das passagens aéreas. As empresas estão cada vez mais empenhadas em reduzir os gastos com viagens, sem comprometer suas operações, pois viajar é necessário, mas economizar é fundamental. Um dos questionamentos mais frequentes é: com que antecedência devemos comprar os bilhetes aéreos para obter os melhores preços? Essa é a questão que buscamos responder!

A definição do preço da passagem aérea não leva em consideração apenas a antecedência da compra. Envolve diversos fatores mercadológicos, custos das empresas aéreas, competição e outros elementos, o que torna a predição do valor da passagem aérea uma tarefa complexa. No entanto, assim como em outros setores, aqueles que compram com maior antecedência têm uma chance maior de obter preços mais baixos em comparação àqueles que deixam a aquisição para datas mais próximas ao evento. .
Palavras-chave: Viagens Corporativas,Antecedência de compra

The costs of corporate travel have been increasing considerably in recent years due to rising airfare prices. Companies are increasingly focusing their efforts on trying to minimize travel expenses without compromising their operations. It's not possible to simply avoid traveling. Traveling is necessary, saving is essential. One of the most recurring questions is: How far in advance should we purchase our air tickets to get the best prices? That's the question to be answered!

The determination of airfare prices takes into account not only the timing of the purchase but also other market factors, airline costs, competitive factors and others that make predicting the value of the airfare not so simple. However, just like in other businesses, those who purchase well in advance have a greater chance of getting lower prices than those who leave the purchase of the product/service closer to the event.
Keywords:

1. Introdução

Os custos com viagens corporativas podem representar uma grande parcela dos custos totais das organizações, e se não controladas de maneira assertiva, podem trazer prejuízos e perdas financeiras no curto e longo prazo.

Anualmente grande parte das organizações fazem o seu planejamento orçamentário para o próximo exercício, que inclui, na maioria das vezes, o orçamento para gastos com viagens. Esse orçamento muitas vezes acaba sendo feito com base em incertezas a respeito dos valores das passagens que serão praticadas pelas cias aéreas no próximo ano. Tentar encontrar um modelo que seja capaz de apresentar previsões dos preços das passagens em um nível satisfatório, seria uma ferramenta de extrema importância para auxiliar nessa definição do orçamento com gastos em viagens.

O presente trabalho tem como objetivo principal aplicar técnicas de machine learning para o desenvolvimento de um modelo de predição do preço de passagens aéreas domésticas, utilizando como base para o treinamento e teste do modelo o consumo anual de passagens aéreas.

Para a criação do modelo será utilizado a linguagem de programação Python, e os pacotes da biblioteca scikit-learn.

Ao longo deste artigo abordarei aspectos a respeito dos critérios de precificação das cias aéreas, uma análise exploratória da base de dados utilizada e por fim o modelo adotado para a previsão dos preços das passagens aéreas e considerações finais a respeito do resultado obtido.

2. Precificação das Passagens Aéreas

"Os modelos de preços das companhias aéreas são complexos, com passagens mantidas a preços baixos o suficiente para que os passageiros os paguem e para

que ao mesmo tempo garantam o lucro das companhias aéreas."(Aitken,2022)

A essa capacidade das companhias aéreas de gerenciar e maximizar a receita obtida por assento disponível em seus voos dá-se o nome de *"manegement yeld*, que em resumo é uma estratégia utilizada pelas companhias aéreas para maximizar a receita por assento disponível, levando em consideração fatores como demanda, concorrência e custos operacionais.

As próximas figuras apresentam um exemplo de precificação das passagens aéreas de acordo com os assentos disponíveis. Os valores das passagens aqui utilizados tem finalidade exclusiva de exemplificação do modelo de *"manegement yeld*, os preços das passagens aéreas são valores simulados.

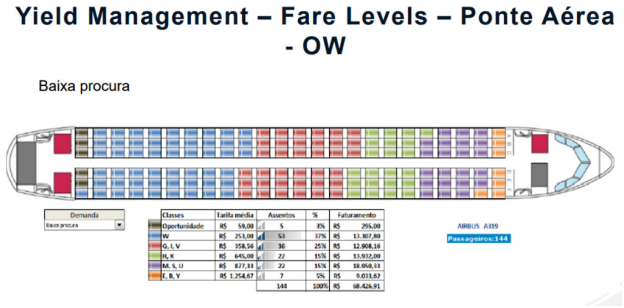

Figura 1: Exemplo de um Modelo de precificação de assentos utilizando o management yeld para demonstrar como as cias aéreas definem os valores de assentos a venda em seus aviões para um voo com baixa procura.

3. Análise exploratória dos dados

Nesta seção é apresentada uma visão geral dos dados e principais indicadores obtidos através da análise exploratório do conjunto de dados utilizado neste estudo.

3.1. Conjunto de dados: Uma visão geral

O conjunto de dados deste estudo é composto pelo histórico de compras reais de uma empresa, referente passagens aéreas domésticas - trechos dentro do Brasil - compradas durante o ano de 2022.

Considerando que a antecedência de compra é um item importante na análise da precificação das passagens aéreas, e que diferentemente das viagens de lazer, onde podemos programar a viagem com uma maior antecedência, nas viagens corporativas o tempo de programação das viagens muitas vezes ocorre com um tempo de antecedência menor que 30 dias, de acordo com os dados observados no gráfico abaixo, as compras com antecedência menor que 30 dias representam 91,47 por cento do conjunto de dados.

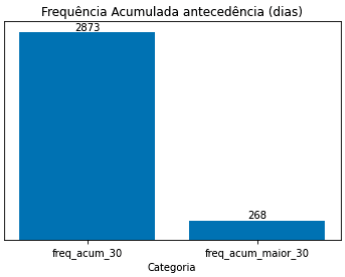

Figura 2: Gráfico de frequência acumulada de compra por antecedência.

Dessa forma os dados para construção do modelo foram centralizados nas transações realizadas com antecedência entre 1 e 30 dias.

Realizadas todas as limpezas e filtros, o conjunto de dados final apresenta 2868 passagens emitidas e 14 variáveis. Utilizando a função info() do pandas obtemos as informações detalhadas do conjunto de dados, apresentados na figura abaixo:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2868 entries, 0 to 3135
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Emissao                   2868 non-null   datetime64[ns]
 1   Embarque                  2868 non-null   datetime64[ns]
 2   Nome Cia                  2868 non-null   object
 3   Trecho                    2868 non-null   object
 4   Tarifa                    2868 non-null   float64
 5   Rt-Ow                     2868 non-null   object
 6   Classe01                  2868 non-null   object
 7   Retorno                   2868 non-null   datetime64[ns]
 8   Duração                   2868 non-null   int64
 9   Dia da Semana (Embarque)  2868 non-null   int64
 10  Dia da Semana (Compra)    2868 non-null   int64
 11  Antecedência (dias)       2868 non-null   int64
 12  Milhas                    2868 non-null   float64
 13  Voos                      2868 non-null   int64
dtypes: datetime64[ns](3), float64(2), int64(5), object(4)
memory usage: 336.1+ KB
```

Figura 3: Dados obtidos através da função info do pandas.

Aplicando a função describe do pandas, obtemos as estatísticas resumidas das colunas numéricas do conjunto de dados, trazendo informações importantes como a média, desvio padrão, valor mínimo e máximo e os quartis. As informações geradas pela função describe auxiliam na identificação de valores atípicos, a entender a dispersão dos dados e obter uma noção geral das escalas dos valores.

Através dos resultados obtidos podemos verificar que os valores de média e mediana da tarifa estão bem próximos, indicando uma que os valores da base de dados têm uma distribuição relativamente simétrica, sem influência significativa de outliers.

	Tarifa	Duração	Dia da Semana (Embarque)	Dia da Semana (Compra)	Antecedência (dias)	Milhas	Voos
Contagem	2868.000000	2868.000000	2868.000000	2868.000000	2868.000000	2868.000000	2868.000000
Média	1465.495603	1.938285	3.493026	4.119247	10.059275	755.782946	2.116806
Desvio Padrão	876.250747	3.649927	1.823658	1.474728	7.245265	573.354377	1.080291
Valor Mínimo	80.900000	0.000000	1.000000	1.000000	0.000000	98.300000	1.000000
primeiro quartil (25° percentil)	800.552500	0.000000	2.000000	3.000000	4.000000	412.000000	1.000000
segundo quartil (mediana ou 50° percentil)	1271.070000	0.000000	3.000000	4.000000	8.000000	664.000000	2.000000
terceiro quartil (75° percentil)	1909.785000	3.000000	5.000000	5.000000	14.250000	956.000000	2.000000
Valor Máximo	6200.480000	31.000000	7.000000	7.000000	30.000000	5180.000000	10.000000

Figura 4: Estatisticas das variáveis númericas do conjunto de dados.

Buscando compreender ainda mais a fundo quais variáveis presentes no conjunto de dados são importantes para a predição do valor da tarifa, foi utilizada a função df.corr() do Pandas, conforme gráfico abaixo:

A função retorna uma matriz de correlação utilizada na geração do gráfico de mapa de calor, que utiliza as cores para representar os valores da matriz, onde as cores mais claras ou intensas indicam maior correlação positiva, enquanto as cores mais escuras indicam maior correlação negativa.

A utilização do gráfico de calor fornece uma visualização clara dos padrões de correlação entre as variáveis, auxiliando na identificação das variáveis que estão mais fortemente relacionadas entre si.

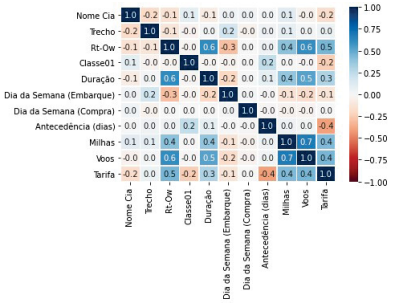

Figura 5: Grafico de correlação das variáveis utilizadas no modelo de predicão.

Através do gráfico acima podemos visualizar que o preço da passagem aéreas possui uma forte correlação negativa com a variável antecedência (dias), demonstrando que há uma redução no preço médio da passagem quando a compra é realizada com uma antecedência maior. Também foi possível notar que o valor da tarifa também tem uma relação negativa com o dia da semana do embarque, indiciando que em determinados dias da semana as passagens apresentam valores médios menores. Outros itens como duração do voo, Milhas voadas e quantidade voos apresentam uma correlação positiva com o valor da passagem aérea, indicando um aumento nos preços das passagens a medida que estas variáveis também aumentam.

O próximo gráfico apresenta a relação do preço da passagem e antecedência, onde podemos visualizar que há uma redução da tarifa média a medida em que há um aumento da antecedência de compra:

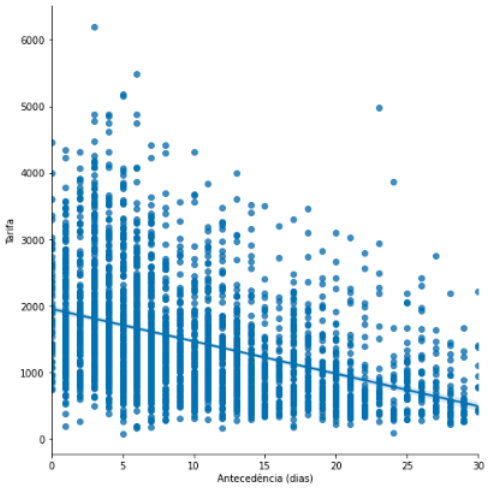

Figura 6: Relação entre o valor da tarifa e antecedência de compra.

Verifica-se através linha de regressão linear deste gráfico a tendência de queda a medida em que a antecedência aumenta, ratificando a variável antecedência como um ponto importante para entender a variação dos valores das passagens.

Conforme verificado no mapa de calor apresentando anteriormente, o dia da semana de embarque também apresenta variação na tarifa média. O gráfico a seguir apresenta a variação da tarifa média de acordo com o dia do embarque:

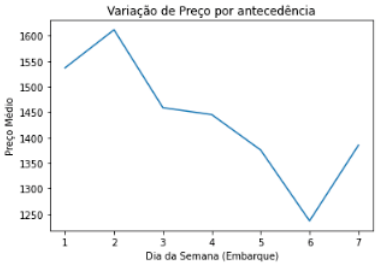

Figura 7: Tarifa Média por dia de embarque, utilizando escala de 1 a 7, onde 1 representa segunda-feira e 7 representa o domingo.

Observa-se que as tarifas apresentam valores médios maiores nos primeiros dias da semana e apresentam queda à medida que se aproxima do final de semana.

4. Criando o Modelo de Machine Learning

Durante a análise do conjunto de dados, e observando todos os aspectos envolvidos na precificação da

passagem aérea, principalmente a existência de fatores internos utilizados pelas empresas aéreas no momento de definir o valor final da passagem, os quais são de difícil mensuração, chegamos à etapa da geração do modelo com 10 variáveis. Essas variáveis são:

- Nome da Companhia Aérea;
- Trecho Voado
- Tipo de Viagem (one way / round trip)
- Classe do Voo
- Duração da viagem
- Dia da semana do Embarque
- Dia da semana de compra
- Antecedência de comoras (em dias)
- Milhas voadas
- Quantidade de voos

O primeiro passo foi transformar todas as variáveis categóricas do conjunto de dados em variáveis numéricas, utilizando a função Lebel Encoder do python. Essa transformação é necessária pois muitos algoritmos de machine learning são formulados em termos de operações matemáticas, como multiplicação de matrizes, e cálculos de distâncias, que são mais facilmente aplicáveis a variáveis numéricas.

Nesta etapa é importante olhar com mais cuidado a variável Preço da passagem, cuja predição é o objetivo deste estudo. O histograma abaixo apresenta a distribuição dos preços das passagens aéreas do conjunto de dados:

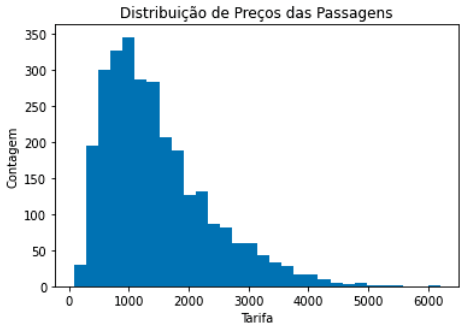

Figura 8: Histograma dos preços das passagens aéreas do conjunto de dados.

O objetivo através deste histograma é fornecer uma visão geral da distribuição dos preços das passagens. Para tentar identificar se os preços estavam concentrados em uma faixa estreita ou se estavam mais amplamente distribuídos.

Apesar da existência de passagens aéreas nas faixas acima de 3 mil reais, é possível perceber uma concentração maior nas faixas entre valores de 1.000 a 2.000 reais. O que é explicado pela característica das passagens que são referentes a compras de trechos dentro do Brasil.

O boxplot a seguir apresenta de forma mais visual essa concentração dos preços das passagens e a distribuição das demais variáveis numéricas, demonstrando não haver grandes outliers que ainda necessitem ser tratados nesta etapa do processo:

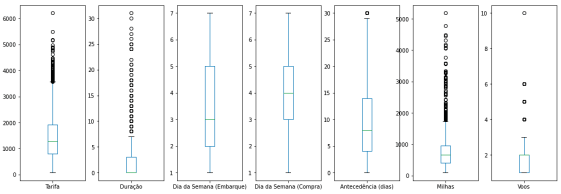

Figura 9: Boxplot variáveis númericas.

Analisados os gráficos, foi realizada a divisão do conjunto de dados em subconjuntos para treinamento e teste, utilizando a função train test split da biblioteca do scikit learn no python.

O tamanho do subconjunto de teste foi definido em 30%, um conjunto de treino de 70% e utilizada uma ramdom state de 7. Essa ramdom state, também chamada de semente aleatória garante que mesmo se a a função for executada várias vezes, se obterá a mesma divisão dos dados em treinamento e teste, desde que o conjunto de dados de entrada seja o mesmo.

Após essa divisão o subconjunto de treino apresenta 2011 observaçõese o subconjunto de teste 862 observações. O modelo de algorítmo escolhido para o modelo supervisionado de Árvore de Decisão. De acordo com Aurélien Géron (2021,p.137) as árvores de decisão são algoritmos versáteis de aprendizado de máquina que podem executar tarefas de classificação, regressão e até mesmo tarefas multioutputs.

Abaixo a representação do código para implementação da árvore de decisão:

```
tree = GridSearchCV(DecisionTreeRegressor(),param_grid, cv = 10)
tree.fit(X_train,y_train)
```

Figura 10: Implementação da árvore de decisão no python.

5. Trabalhos Relacionados

Apesar de não ser um tema amplamente estudado, durante a etapa de pesquisa bibliográfica, foram análisados alguns artigos, que embora tenham uma outra abordagem, têm semelhança no tema preço da passagem.

O tragalho intitulado Sistema de Apoio à Decisão na compra de Passagens Aéreas [Calegario, 2018], apresenta uma abordagem para a criação de um modelo de predição extraindo os dados das passagens aéreas de sistemas web para a extração das informações, baseados em passagens promocionais. Para a criação do modelo preditivo o autor utiliza os métodos de Random Forest e Regressão Linear.

O trabalho intitulado Aplicação de modelos não lineares na precificação de passagens aéreas [Melo, 2018], utiliza modelos não lineares para predição das passagens aéreas utilizando como base de estudos os trechos de Londrina para Curitiba e de Londrina para Congonhas, São Paulo. A autora explora os conceitos dos modelos de regressão linear e não linear, apresentando uma comparação entre os modelos utilizados e ao final apresenta suas conclusões a respeito do modelo escolhido.

6. Resultados

Para a criação do modelo foi utilizado a função GridSearchCV para o modelo supervisionado de Árvore de decisão de regressão, para encontra a melhor combinação de hiperparâmetros para o modelo. Os hiperparâmetros são parametros que não são aprendidos pelo modelo durante o processo de treinamento.

Estes hiperparâmetros controlam aspectos da estrutura da árvore, como a profundidade máxima, o critério de divisão, e outros aspectos.

A métrica de avaliação do modelo foi o erro quadrático médio (MSE), que compara as previsões do modelo com os valores reais dos dados. Ele apresenta uma média dos erros quadrados entre as previsões e os valores reais, fazendo isso para todas as observações. Quanto menor o MSE, melhor o desempenho do modelo, pois neste caso os valores preditos estão mais próximos dos valores reais.

Além do MSE foram utilizadas como métricas para verificação do ajuste do modelo o erro percentual absoluto (MAPE) e o $R\hat{2}$.

O MAPE mede a precisão média das previsões , em percentuais, indicando se o modelo está bem ajustado aos dados. Ele calcula a diferença percentual entre cada um dos valores reais e valores preditos. Através do valor absoluto dessa diferença calcula a média desses erros percentuais para todas as observações. Uma vez que o MAPE é expresso em valores percentuais, a sua interpretação fica mais intuítiva. Valores menores de MAPE indicam que os valores da previsão estão mais próximos dos valores reais, em média. O $R\hat{2}$ fornece uma medida de proporção de variabilidade total dos dados que é explicada pelo modelo. Ele varia entre 0 e 1 e quanto mais proximo do $R\hat{2}$ estiver de 1, melhor ele explica a variabilidade dos dados, apresentando um melhor ajuste do modelo.

Abaixo os resultados de cada uma destas métricas, obtidas para o modelo:

- Erro Quadrático Médio: 372.126;
- Erro Percentual Médio Absoluto: 24.257;
- R^2: 0.813

O próximo gráfico apresenta a comparação entre os valores preditos da variável resposta, Preço da Passagem, no conjunto de dados.

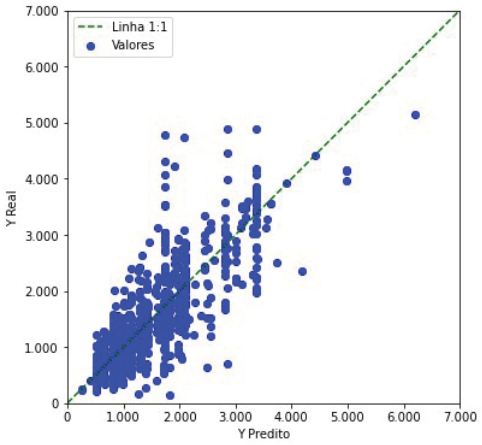

Figura 11: Valores observados contra o valores preditos da variável resposta no conjunto de teste.

7. Conclusões

Se considerarmos a complexidade dos fatores que influenciam a precificação das passagens aéreas, o modelo apresenta um resultado satisfatório ao apresentar um R^2 relativamente alto, de 0,813, indicando que cerca de 81,3% da variabilidade nos preços das passagens aéreas é explicada pelo moodelo.

Outras varárieis como sazonalidade, questões económicas, como câmbio, inflação entre outros são fatores que que influenciam diretamente os preços. Variáveis como estes exemploes, de díficil mensuração, podem explicar os resultados um pouco elevados do MAE e MAPE. Estes resultados indicam que há espaço para refinamento e melhoria do modelo, caso seja possível identificar fatores relevantes, além dos já aplicados ao modelo, além da utilização de bases de dados mais robustas para o treinamento do modelo.

Porém ainda sim o resultado é satisfatório se análisarmos um cenário onde não há modelos de predição de passagens aéreas e o modelo deste estudo pode ser de grande auxilio para o planejamento de viagens, desde que os valores do erros apresentados sejam aceitáveis para a condição utilizada.

Agradecimentos

O autor agradece ao seu Orientador Walmes Marques Zeviani e a todos os professores do curso DSBD - UFPR.

Referências

[1] Géron, Aurélien, *Mãos à obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow* , (Starlin Alta Editora e Consultoria Eireli, Rio de Janeiro, 2021), 2ª ed.

[2] Python Software Foundation. Python Programming Language. Versão 3.9.2. Disponível em: *https://www.python.org/*.

[3] Matplotlib Development Team. Matplotlib: A 2D plotting library. Versão 3.4.2. Disponível em: *https://matplotlib.org/*.

[4] Pandas Development Team. Pandas: Powerful data analysis toolkit. Versão 1.3.0. Disponível em: *https://pandas.pydata.org/*.

[5] SCIKIT-LEARN. Scikit-learn: Machine learning in Python. Versão 0.24.1. Disponível em: <https://scikit-learn.org/>.

[6] Calegario, R.C. (2018). Sistema de Apoio à Decisão na Compra de Passagens Aéreas. Universidade Federal de Pernambuco.

[7] Melo, J. (2018). Aplicação de Modelos não Lineares na Precificação de Passagens Aéreas. Universidade Tecnológica Federal do Paraná, Cornélio Procópio.