

INSTITUTO TECNOLÓGICO DE AERONÁUTICA



João Vitor Fontenele Lustosa

**UMA ABORDAGEM DA PRECIFICAÇÃO DE
PASSAGENS AÉREAS NO BRASIL UTILIZANDO
ALGORITMOS DE APRENDIZADO DE MÁQUINA**

Trabalho de Graduação
2021

Curso de Engenharia Civil-Aeronáutica

João Vitor Fontenele Lustosa

**UMA ABORDAGEM DA PRECIFICAÇÃO DE
PASSAGENS AÉREAS NO BRASIL UTILIZANDO
ALGORITMOS DE APRENDIZADO DE MÁQUINA**

Orientador

Prof. Dr. Alessandro Vinícius Marques de Oliveira (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

**SÃO JOSÉ DOS CAMPOS
INSTITUTO TECNOLÓGICO DE AERONÁUTICA**

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Fontenele Lustosa, João Vitor

Uma abordagem da precificação de passagens aéreas no Brasil utilizando algoritmos de aprendizado de máquina / João Vitor Fontenele Lustosa.

São José dos Campos, 2021.

66f.

Trabalho de Graduação – Curso de Engenharia Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2021. Orientador: Prof. Dr. Alessandro Vinícius Marques de Oliveira.

1. Aprendizado de máquina. 2. Mercado aéreo. 3. Previsão de preço. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

FONTENELE LUSTOSA, João Vitor. **Uma abordagem da precificação de passagens aéreas no Brasil utilizando algoritmos de aprendizado de máquina**. 2021. 66f.

Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

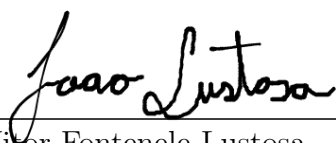
CESSÃO DE DIREITOS

NOME DO AUTOR: João Vitor Fontenele Lustosa

TÍTULO DO TRABALHO: Uma abordagem da precificação de passagens aéreas no Brasil utilizando algoritmos de aprendizado de máquina.

TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2021

É concedida ao Instituto Tecnológico de Aeronáutica permissão para reproduzir cópias deste trabalho de graduação e para emprestar ou vender cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte deste trabalho de graduação pode ser reproduzida sem a autorização do autor.

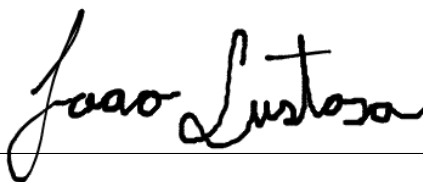


João Vitor Fontenele Lustosa

Av. João Batista Soares de Queiroz Jr, 800
12240-000 – São José dos Campos-SP

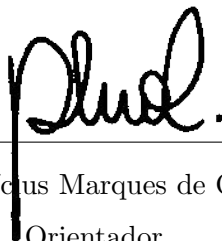
UMA ABORDAGEM DA PRECIFICAÇÃO DE PASSAGENS AÉREAS NO BRASIL UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação



João Vitor Fontenele Lustosa

Autor



Alessandro Vinícius Marques de Oliveira (ITA)

Orientador



Prof. Dr. João Claudio Bassan de Moraes
Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 12 de novembro de 2021.

Dedico este trabalho a minha mãe, Vanesca Fontenele, por ser a mulher mais guerreira que conheço.

Agradecimentos

Inicialmente agradeço ao meu pai Fábio e à minha mãe Vanesca por sempre terem acreditado em mim, me ajudando a superar todas as dificuldades.

Agradeço ao meu coordenador da turma ITA, Danúbio, por ser um exemplo de líder, que proporcionou o melhor ambiente possível para que a nossa turma pudesse estudar com empenho e alcançar o sonho da aprovação. Suas ações levaram a um resultado inesquecível para o Colégio Christus.

Agradeço aos irmãos que fiz durante o curso e que em hipótese alguma me permitiram ficar para trás nessa trajetória. Obrigado por fazer desses anos de ITA os melhores da minha vida até agora. Sem vocês eu não estaria escrevendo esses agradecimentos.

Agradeço especialmente à minha querida namorada Julie, por ser a melhor companheira possível nessa batalha. Sem o seu apoio nada disso seria possível!

Agradeço ao meu professor orientador Alessandro, por todo o conhecimento transmitido, pelo seu carisma, e por provar que com companheirismo e atenção, alunos e professores podem atingir grandes resultados.

Ao meu querido irmão Dudu, quando um dia estiver lendo isso, saiba que eu te amo muito!

*“Choose the ones you love.
Choose your future.
Choose life.”*

— MARK RENTON

Resumo

O preço das passagens aéreas são determinados por inúmeros fatores, como tempo de voo, concentração de mercado, atividade econômica, preço de combustível, entre outros. Cada companhia aérea utiliza algoritmos específicos de gerenciamento de receita, que leva em consideração fatores que são uma incógnita para os consumidores. Este trabalho aborda a precificação do mercado aéreo nacional, buscando prever o preço de passagens aéreas e encontrar variáveis determinantes, utilizando uma base de dados compilada pelo laboratório de transporte aéreo do ITA (LABTAR) contendo variáveis macroeconômicas e dados de voos em trechos nacionais no período entre 2010 a 2018. No desenvolvimento deste trabalho, foi possível treinar seis algoritmos de aprendizado de máquina, utilizando variáveis pré-selecionadas da base de dados. Entre os modelos, o modelo Random Forest obteve a maior precisão preditiva acima de 85%. Em números absolutos, isto representa um erro médio de R\$ 50,00 na determinação de preços que flutuam em torno de uma média de R\$ 470,00. Além disso, esse modelo indicou que os fatores de maior importância preditiva são: tempo agendado de voo, média de quilos de bagagem transportada por passageiro e tamanho da aeronave.

Palavras-chave: Mercado Aéreo, Previsão de Preço, Passagens Aéreas, Aprendizado de Máquina

Abstract

The price of airline tickets is determined by numerous factors, such as flight time, market concentration, GDP, fuel price, among others. Each airline uses specific revenue management algorithms, which take into account factors that are unknown to consumers. This work addresses the pricing of the national air market, seeking to predict the price of air tickets and find determining variables, using a database compiled by the ITA air transport laboratory (LABTAR) containing macroeconomic variables and flight data on national air sections between the period from 2010 to 2018. In the development of this work, it was possible to train six different machine learning algorithms using selected features from the database. Among the models, the Random Forest model had the highest predictive accuracy above 85%, in absolute numbers, this is an average error of R\$ 50.00 in the determination of prices that fluctuate around an average price of R\$ 470.00. In addition, this model indicated that the factors of greatest predictive importance are: scheduled flight time, average kilograms of baggage per passenger and aircraft size.

Keywords: Airfare Prediction, Airline Ticket Price, Machine Learning

Lista de Figuras

FIGURA 1.1 – Fluxograma de interesses da precificação dinâmica. Adaptado de (FIIG T.; SMITH., 2016).	18
FIGURA 2.1 – Gráficos de variação temporal do preço de passagens aéreas. Fonte: (WILLIAMS, 2017)	24
FIGURA 2.2 – Relação entre antecedência e preço para diferentes linhas aéreas nacionais no trecho FOR-GRU. Fonte: (CHAGAS, 2018)	25
FIGURA 2.3 – Ajuste de curva de preço no espaço de tempo anterior ao voo. Fonte: (JANSSEN, 2014)	26
FIGURA 2.4 – Importâncias das variáveis definidas pelo modelo Random Forest. Fonte: (WANG <i>et al.</i> , 2019)	29
FIGURA 3.1 – Fluxograma de execução do trabalho. Fonte: autoria própria, 2021.	30
FIGURA 3.2 – Mapeamento de municípios (à esquerda) e aeroportos brasileiros (à direita) contidos na base de dados. Fonte: autoria própria, 2021. . .	32
FIGURA 3.3 – Malha aérea nacional contida na base. Fonte: autoria própria, 2021.	33
FIGURA 3.4 – Média móvel do preço de passagens deflacionados no período 2010 - 2018. Fonte: autoria própria, 2021.	34
FIGURA 3.5 – Evolução da tarifa média doméstica anual no Brasil e variações percentuais (2002-2017) Fonte: ABEAR, 2018.	34
FIGURA 3.6 – Média móvel do preço de algumas rotas da base no período 2010 - 2018. Fonte: autoria própria, 2021.	35
FIGURA 3.7 – Histograma da distribuição de preços. Fonte: autoria própria, 2021.	35
FIGURA 3.8 – <i>Boxplot</i> da distribuição de preços. Fonte: autoria própria, 2021. . .	36
FIGURA 3.9 – Malha de rotas que contém o aeroporto de Guarulhos - SP (GRU). Fonte: autoria própria, 2021.	37

FIGURA 3.10 – Rotas com preços médios entre R\$150 e R\$240. Fonte: autoria própria, 2021.	38
FIGURA 3.11 – Rotas com preços médios entre R\$620 e R\$774. Fonte: autoria própria, 2021.	38
FIGURA 3.12 – Esquema da partição da base e aplicação individual da função de tratamento de dados. Fonte: autoria própria, 2021.	39
FIGURA 3.13 – Visualização das observações faltantes na base. Fonte: autoria própria, 2021.	40
FIGURA 3.14 – Curva de distribuição normal. Fonte: autoria própria, 2021.	41
FIGURA 3.15 – <i>Trade-off</i> entre viés e variância. Fonte: http://scott.fortmann-roe.com/docs/BiasVariance	
FIGURA 3.16 – Correlação entre PIB e demanda pelo transporte aéreo. Fonte: ABEAR, 2018.	45
FIGURA 3.17 – Matriz de correlação entre as variáveis. Fonte: autoria própria, 2021.	46
FIGURA 3.18 – Condição de contorno para a regressão Lasso. Fonte: http://cursos.leg.ufpr.br/ML4all	
FIGURA 3.19 – Hiperplano de ajuste SVR. Fonte: (SETHI, 2020)	49
FIGURA 3.20 – Função de minimização de erro SVR. Fonte: (SETHI, 2020)	49
FIGURA 3.21 – Esquema de repartição e ajuste do modelo Random Forest.	50
FIGURA 3.22 – Processo iterativo do gradiente descendente.	51
FIGURA 4.1 – Peso de importância preditiva das variáveis no modelo RF.	58
FIGURA 4.2 – Fluxo decisório do algoritmo Random Forest.	59

Lista de Tabelas

TABELA 2.1 – Compilado dos resultados obtidos por Tziridis. Fonte: (TZIRIDIS <i>et al.</i> , 2017)	27
TABELA 2.2 – Descrição das variáveis utilizadas por Wang. Fonte: (WANG <i>et al.</i> , 2019)	28
TABELA 2.3 – Compilado dos resultados obtidos por Wang. Fonte: (WANG <i>et al.</i> , 2019)	29
TABELA 3.1 – Variáveis da base de dados STATAER. Fonte: autoria própria, 2021.	31
TABELA 3.2 – Features selecionadas manualmente.	43
TABELA 4.1 – Compilado dos resultados obtidos com todas as variáveis inclusas. .	53
TABELA 4.2 – Variáveis selecionadas nomeadas por $F\#N$	54
TABELA 4.3 – Resultados obtidos excluindo as features F2 e F3.	54
TABELA 4.4 – Resultados obtidos excluindo a feature F5.	55
TABELA 4.5 – Resultados obtidos excluindo a feature F6.	55
TABELA 4.6 – Resultados obtidos excluindo a feature F7.	55
TABELA 4.7 – Resultados obtidos excluindo a feature F8.	56
TABELA 4.8 – Resultados obtidos excluindo a feature F9.	56
TABELA 4.9 – Resultados obtidos excluindo a feature F10.	56
TABELA 4.10 – Resultados obtidos excluindo a feature F11.	57
TABELA 4.11 – Resultados obtidos excluindo a feature F12.	57
TABELA 4.12 – Resultados obtidos excluindo a feature F13.	57
TABELA 4.13 – Importâncias das features (%) para nove diferentes modelos usando o algoritmo Random Forest.	60
TABELA 4.14 – Resultados obtidos após a otimização de hiperparâmetro.	61

Lista de Abreviaturas e Siglas

RM	Revenue management
LCC	Low cost carrier
CPI	Consumer Price Index
HHI	Índice Herfindahl-Hirschman
ANAC	Agência Nacional de Aviação Civil
ABEAR	Associação Brasileira das Empresas Aéreas
LABTAR	Laboratório de transporte aéreo
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boost
RF	Random Forest

Sumário

1	INTRODUÇÃO	17
1.1	O problema	18
1.2	Objetivo geral	18
1.3	Objetivos específicos	19
1.4	Justificativas	19
1.5	Delimitação do tema	19
1.6	Organização do trabalho	20
1.6.1	Capítulo 2: Revisão de literatura	20
1.6.2	Capítulo 3: Metodologia	20
1.6.3	Capítulo 4: Análise de resultados	20
1.6.4	Capítulo 5: Conclusão	21
2	REVISÃO DE LITERATURA	22
2.1	Gerenciamento de receita	22
2.1.1	Overbooking	22
2.1.2	Personalização	23
2.1.3	Precificação dinâmica	23
2.2	Resultados de previsão de preço da literatura	25
2.3	Modelos de aprendizado de máquina	26
3	METODOLOGIA	30
3.1	Diagrama de execução e boas práticas de <i>Data Science</i>	30
3.2	Base de dados	31
3.3	Análise exploratória dos dados	32

3.4	Pré-processamento de dados	39
3.4.1	Remoção de atributos <i>missing</i>	39
3.4.2	Remoção de <i>outliers</i>	40
3.4.3	<i>Label-encoding</i> para variáveis categóricas	41
3.4.4	Normalizando o <i>dataframe</i>	41
3.5	Seleção de features	42
3.5.1	Seleção manual de features	42
3.5.2	Algoritmos de seleção de features	46
3.6	Algoritmos de aprendizado de máquina	47
3.6.1	Regressão Linear	47
3.6.2	Regressão Lasso	47
3.6.3	Regressão Elastic Net	48
3.6.4	Regressor Support Vector Machine (SVM)	48
3.6.5	Random Forest	50
3.6.6	XGBoost	50
4	ANÁLISE DE RESULTADOS	52
4.1	Métricas de avaliação	52
4.1.1	RMSE	52
4.1.2	MAE	52
4.1.3	MAPE	52
4.1.4	R_{adj}^2	53
4.2	Resultados	53
4.2.1	Variando features de input	54
4.2.2	Random Forest	58
4.2.3	Ajuste de hiperparâmetro	60
5	CONCLUSÃO	62
5.1	Limitações do trabalho	63
5.2	Sugestões para trabalhos futuros	63

REFERÊNCIAS	64
-----------------------	----

1 Introdução

A indústria de transporte aéreo é conhecida por sua precificação dinâmica intertemporal de passagens aéreas. Desde meados dos anos 80, o *yield management* tem sido uma estratégia de grande importância para garantir a receita no mercado aéreo e a saúde do caixa das companhias (WOHLFARTH, 2011). O papel do *yield management* é análogo ao papel do controle de estoque em empresas de manufatura, ou seja, busca-se “vender os assentos certos para o clientes certos no momento certo” (SMITH J. LEIMKUHLE; SAMUELS, 1992).

Tendo em vista a complexidade dos processos decisórios para definição de preço, as companhias aéreas utilizam diversas técnicas computacionais envolvendo análise de dados para tentar atingir previsões de demanda para maximizar o lucro nos trechos (ABDELLA, 2020). Isso resulta naturalmente na flutuação do preço para o mesmo serviço, isto é, tarifas diferentes para cada categoria de assento em um mesmo voo.

Em contraste, na perspectiva do passageiro, essas flutuações podem ser uma fonte de preocupação intensa na decisão de compra, o que pode levar passageiros mais sensíveis ao preço a hesitar ou serem impulsivos no momento de compra (WOHLFARTH, 2011).

Assim, a precificação dinâmica se classifica como um fator de extrema relevância tanto para a companhia aérea, quanto para os passageiros. Fiig et al. (2016) comparou a performance de precificação dinâmica em relação a preços estáticos e descobriu ganho de receita com a utilização do algoritmo dinâmico.

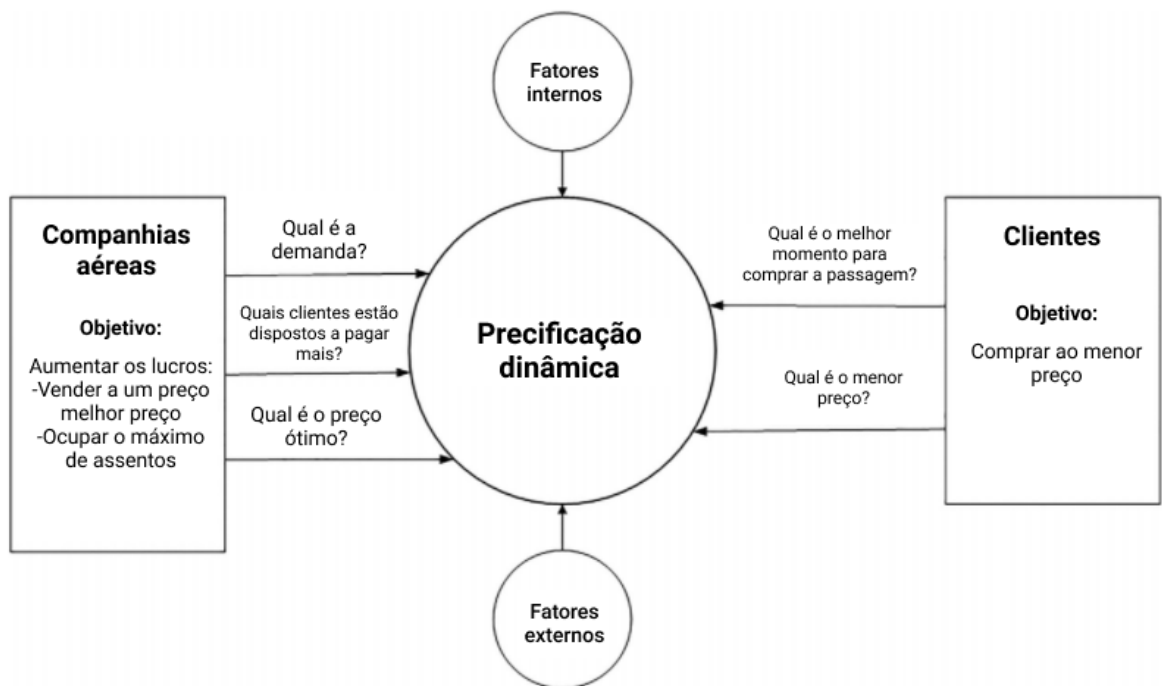


FIGURA 1.1 – Fluxograma de interesses da precificação dinâmica.
Adaptado de (FIIG T.; SMITH., 2016).

Portanto, modelar uma previsão para as tarifas aéreas pode trazer uma margem de compra para o consumidor e um posicionamento estratégico de mercado para as linhas aéreas, no que se trata de obter uma estimativa do seu faturamento de acordo com cada uma de suas rotas.

1.1 O problema

Utilizando dados históricos de rotas aéreas nacionais, será possível construir um bom modelo de aprendizado de máquina para prever preços médios de tarifas aéreas em um determinado trecho? Quais atributos mais impactam esses preços? Como os consumidores e as companhias aéreas podem se beneficiar de tais previsões?

1.2 Objetivo geral

O presente trabalho busca explorar uma base extensa de dados chamada STATAER que foi compilada pelo Núcleo de Transporte Aéreo do ITA, contendo voos nacionais para treinar e validar um modelo de aprendizado de máquina que preveja o preço médio de passagens nos trechos.

1.3 Objetivos específicos

O presente trabalho busca cumprir os seguintes objetivos específicos:

- Explorar os dados de rotas aéreas nacionais
- Atribuir importância de variáveis e seu impacto no preço
- Treinar e validar diferentes algoritmos para previsão do preço médio de uma dada rota
- Comparar o desempenho de cada algoritmo

1.4 Justificativas

Este trabalho pretende abordar ambas perspectivas sobre a precificação de passagens aéreas, tanto das companhias aéreas quanto do consumidor. Há espaço para explorar hipóteses e modelos de previsão dentro do contexto do mercado aéreo, visto que falta precisão e generalidade nos modelos existentes (ABDELLA, 2020). Gillen (2009) argumentou sobre a dificuldade de analisar a precificação do mercado aéreo:

“A falta de literatura examinando essas características de preços pode ser explicada pela dificuldade em obter esses dados em uma grande escala. Esses dados devem ser coletados longitudinalmente (de diferentes fontes confiáveis), e os sites podem bloquear a coleta intensiva de dados, como foi vivido por McAfee e Te Velde (MCAFEE, 2006).”

Com os dados obtidos pelo laboratório de transporte aéreo, temos uma certa vantagem no desenvolvimento deste trabalho. Dessa maneira, é válido uma análise intensiva sobre as variáveis dos dados que gerem uma discussão detalhada sobre seus impactos na precificação. Assim, essa pesquisa pode contribuir para um melhor discernimento da precificação do mercado aéreo nacional.

1.5 Delimitação do tema

O mercado aéreo muda constantemente e a sua precificação é extremamente complexa para ser avaliada. Podemos comentar, por exemplo, sobre a pandemia de COVID-19 em 2020 que impactou fortemente o setor. Sabemos que esse impacto não foi só momentâneo e sim transformou a perspectiva de muitos passageiros em relação às suas necessidades

quanto ao deslocamento aéreo. Dessa forma, a qualidade de previsão obtidas neste trabalho, para dados entre 2010 e 2018, pode não ser obtida com dados de anos futuros, requerendo modelos e variáveis adaptadas. É válido comentar as limitações da base de dados utilizada no trabalho (STATAER), que contém apenas preços médios por trechos aéreos, captados entre 2010 e 2018. Dessa forma, não há como analisar a distribuição de intervalo de preços dentro de um mesmo voo. Além disso, sentiu-se falta de algumas variáveis que poderiam ser úteis na análise, como: horário de partida, linha aérea efetuando a rota, dias para a partida, load factor, entre outras.

1.6 Organização do trabalho

O presente trabalho está organizado conforme a seguinte sequência dos capítulos.

- Revisão de literatura;
- Metodologia;
- Análise de resultados;
- Conclusão;

1.6.1 Capítulo 2: Revisão de literatura

Neste capítulo contextualizamos a precificação do mercado aéreo no que se trata do gerenciamento de receitas e sob a perspectiva da precificação dinâmica. Além disso, passaremos sobre trabalhos da literatura que estudaram algoritmos preditivos de passagens aéreas e vamos entender os resultados por eles obtidos.

1.6.2 Capítulo 3: Metodologia

Este capítulo consiste em discretizar toda a metodologia de análise exploratória de dados, assim como as melhores práticas adotadas de *Data Science* para tratamento de dados, treino e validação de algoritmos de *Machine Learning*. Além disso, exploramos o funcionamento de todos os modelos utilizados.

1.6.3 Capítulo 4: Análise de resultados

Os resultados preditivos obtidos serão ilustrados nesses capítulos, juntamente com uma discussão a respeito dos algoritmos desenvolvidos e suas respectivas performances de ajuste.

1.6.4 Capítulo 5: Conclusão

Nesse último capítulo, enfatizamos os principais *take-aways* dessa pesquisa, limitações dos resultados e sugestões para pesquisas futuras na mesma linha.

2 Revisão de literatura

2.1 Gerenciamento de receita

O objetivo original do gerenciamento de receita (*Revenue Management*) - como ele evoluiu no transporte aéreo de passageiros na década de 1970, após a desregulamentação do mercado de aviação dos EUA - era a tarefa de vender de forma otimizada um estoque fixo e perecível dentro de um determinado horizonte de tempo. Mais precisamente, tendo produtos definidos com base em um conjunto de serviços que usam diferenciação de preços (por exemplo, tarifas padrão e econômicas no caso de passagens aéreas), a ideia era controlar dinamicamente a disponibilidade desses produtos ao longo do tempo para maximizar em geral receitas obtidas (ou lucros) (KLEIN SEBASTIAN KOCH, 2020).

A maioria das pesquisas envolvendo gerenciamento de receita foca no processo de controle de vendas a um nível operacional. Sob a perspectiva da demanda, há duas classes de modelos. Na primeira, assume-se segmentos de consumidores com diferentes alternativas de compra, chamado de ‘*choice-based RM*’ ou gerenciamento de receita baseado nas opções. No segundo modelo, os consumidores só consideram a possibilidade de compra de um único produto, caso não esteja disponível, estes não irão realizar a compra. Nos dias de hoje, o modelo de produtos independentes está ultrapassado, visto que há uma influência mútua na decisão de compra dos passageiros (KLEIN SEBASTIAN KOCH, 2020).

2.1.1 Overbooking

Overbooking descreve a prática de vender mais produtos do que a capacidade fisicamente disponível, a fim de se proteger contra os casos em que os clientes cancelam antes da prestação do serviço ou não comparecer no momento da prestação do serviço (no-shows).

Por exemplo, com base em dados da Continental Airlines, Gorin (2006) estimou que 4% dos clientes não aparecem na partida e que 15% a 18% da receita total das melhorias obtidas pelo gerenciamento de receita resultam de overbooking.

Portanto, empresas aéreas devem ser capazes de prever padrões de chegadas no sistema

de reservas, para que possam ter controle na definição de critérios para a flutuação do preço das passagens com foco no controle de estoque de assentos (OLIVEIRA, 2008).

2.1.2 Personalização

Adaptar uma oferta a um cliente com base em observados ou as informações derivadas sobre esse indivíduo são chamadas de ofertas customizadas. Personalização de oferta é atualmente um dos tópicos mais discutidos no gerenciamento de receitas na indústria aérea. Isso se deve principalmente ao recente e contínuo mudanças na tecnologia de distribuição das companhias aéreas, oferecendo novas possibilidades no que diz respeito a tornar muito mais flexível e ofertas personalizadas (WESTERMANN, 2013). Primeiras tentativas foram realizadas para propor modelos que dinamicamente decidem sobre disponibilidade com base nas características específicas do cliente, por ex. a viagem final ou fazer ofertas de tarifas (WITTMAN; BELOBABA, 2018). Os dados do cliente podem, por exemplo, ser obtido durante o processo de reserva e potencialmente vinculado com dados sobre o cliente específico obtidos por meio da companhia aérea, programas de fidelidade e também contém o histórico específico de reserva etc. Portanto, uma forte integração com análises e mineração de dados técnicas é necessário para que as companhias tenham um bom modelo de personalização.

O comportamento de compra dos passageiros é diferente do que a maioria dos modelos de gerenciamento de receita da literatura estipula. A maioria dos passageiros é sensível ao preço e não é perfeitamente segmentada, ou seja, irá procurar pagar a menor tarifa possível no trecho desejado (BOYD; KALLESEN, 2004). Boyd articulou:

“É importante perceber que a precificação da demanda não está relacionada como uma companhia faz negócios, mas de como os passageiros comportam-se; o aumento das operadoras de baixa tarifa e o crescimento das reservas pela Internet refletem os esforços para acertar a precificação de acordo com a análise comportamental dos clientes.”

2.1.3 Precificação dinâmica

A precificação dinâmica é uma função obtida após todo o processamento de variáveis incluídas no modelo de gerenciamento de receita. Essa técnica é essencial quando o produto fornecido possui duas características. Primeiro, o produto expira em um determinado momento, ou seja, possui um prazo de validade. Em segundo, a capacidade é fixada com bastante antecedência e pode ser aumentada apenas a um custo marginal relativamente alto.

Várias teorias foram desenvolvidas para explicar a dispersão dos preços sendo amplamente aceito o fenômeno de preço de equilíbrio competitivo. A observação de múltiplas passagens aéreas para a mesma rota levou a uma variedade de pesquisas que examinam a relação entre a concentração de mercado e a dispersão de preço, muitas vezes usando dados de bilhetes transacionados da Pesquisa de Origem e Destino do Departamento de Transporte dos EUA banco de dados (DB1B) (GERARDI; SHAPIRO, 2007).

Williams (2017) analisou a oscilação das tarifas aéreas. A figura 2.1 mostra a frequência e magnitude das tarifas ao longo do tempo, tendo como referência o dia do voo.

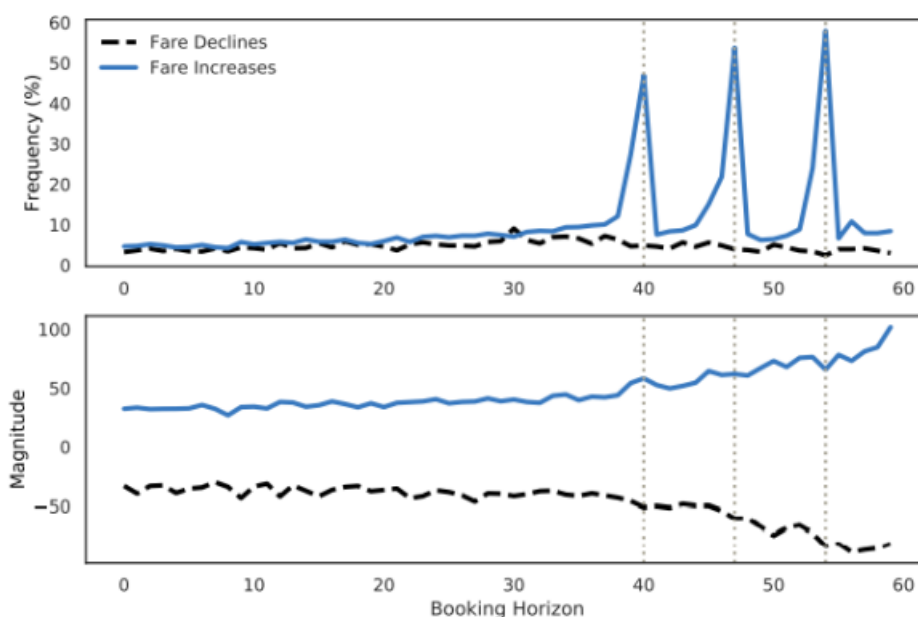


FIGURA 2.1 – Gráficos de variação temporal do preço de passagens aéreas.
Fonte: (WILLIAMS, 2017)

Antes da partida, o gráfico inferior indica a magnitude dessas mudanças de tarifa (ou seja, um gráfico das primeiras diferenças, condicionado à direção da mudança de tarifa). Por exemplo, no gráfico superior, quarenta dias antes da partida, cerca de 5 por cento das tarifas aumentam e 5 por cento das tarifas diminuem. Os 90% restantes das tarifas são mantidos constantes. Analisando o gráfico inferior, a magnitude do aumento e diminuição da tarifa quarenta dias depois é de aproximadamente \$50. O gráfico superior confirma que os aumentos e quedas nas tarifas ocorrem ao longo do tempo. Os picos observados no gráfico superior indicam o momento em que promoções com desconto na passagem expiram.

Chagas (2018) analisou a variação do preço de passagens aéreas no trecho nacional GRU-FOR, obtendo um intervalo de probabilidade de preço de acordo com a antecedência da data do voo.

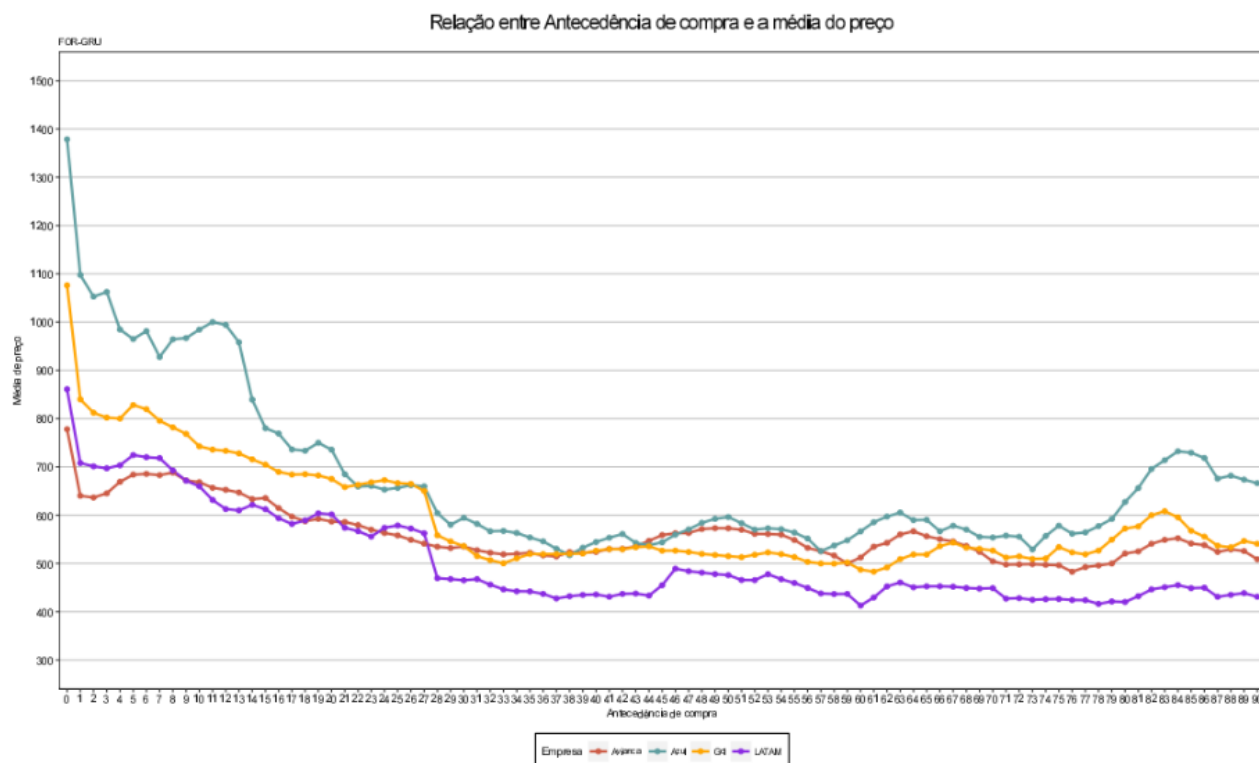


FIGURA 2.2 – Relação entre antecedência e preço para diferentes linhas aéreas nacionais no trecho FOR-GRU.

Fonte: (CHAGAS, 2018)

Percebemos no gráfico da Figura 2.2 um padrão de variação intertemporal das tarifas para as semelhantes às diferentes linhas aéreas. Esse estudo fornece uma boa estimativa para que passageiros adquiram passagens aéreas em um momento de preço ótimo.

2.2 Resultados de previsão de preço da literatura

De acordo com estudos da literatura, Groves e Gini (2013) aplicaram o modelo de regressão parcial de mínimos quadrados (PLS) para otimizar a compra de passagens aéreas, com 75,3% de precisão. Papadakis (2012) previu se o preço do tíquete vai cair no futuro, lidando com o problema como um tarefa de classificação usando Ripple Down Rule (RIDOR) (74,5% de acerto), Regressão Logística (69,9% de acerto) e Support Vector Machine (SVM) linear (69,4% de acerto). Ren e Yang (2015), estudaram o desempenho de regressão linear (77,06% de acerto.), Nave Bayes (73,06% de acerto), Softmax Regression (76,84% de acerto) e SVM (80,6% de acerto).

2.3 Modelos de aprendizado de máquina

Janssen (2014) usou um modelo de regressão linear com quantis para modelar o comportamento de preço dado o número de dias de antecedência ao voo. Janssen obteve uma curva de ajuste ilustrada na figura abaixo:

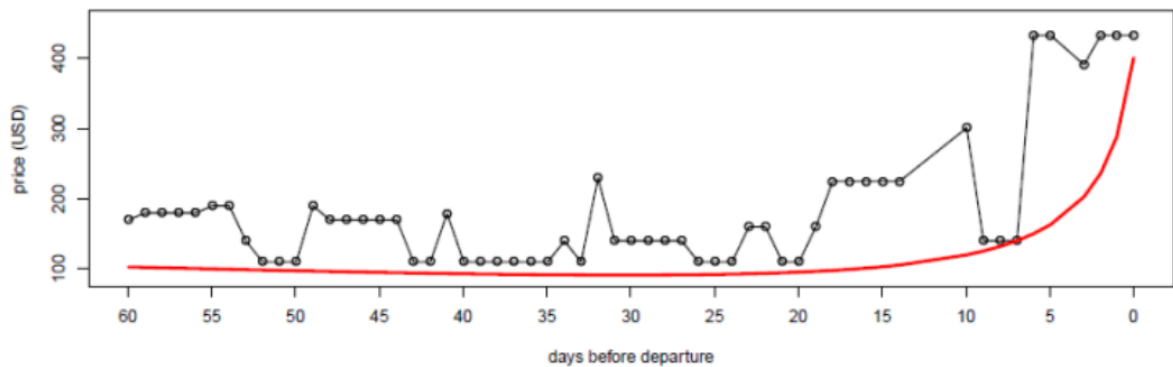


FIGURA 2.3 – Ajuste de curva de preço no espaço de tempo anterior ao voo.

Fonte: (JANSSEN, 2014)

O resultado mostra pouca variação de preço até 10 dias de antecedência do voo, quando o preço escala de forma aproximadamente exponencial.

Tziridis (2017) estudou a previsão de preço de passagens aéreas, analisando 8 modelos de aprendizado de máquina diferentes. Os modelos utilizados foram:

- Multilayer Perceptron (MLP)
- Generalized Regression Neural Network
- Extreme Learning Machine (ELM)
- Random Forest Regression Tree
- Regression Tree
- Bagging Regression Tree
- Regression SVM (Polynomial and linear)
- Linear Regression (LR)

Nesse estudo, as variáveis utilizadas, foram:

- F1: Horário de decolagem

- F2: Horário de pouso
- F3: Quantidade de bagagem livre (0,1 ou 2)
- F4: Dias de antecedência
- F5: Número de conexões
- F6: Feriado (sim ou não)
- F7: Voo noturno (sim ou não)
- F8: Dia da semana

O resultado obtido por K. Tziridis com todas as features descritas acima, está representado na tabela:

TABELA 2.1 – Compilado dos resultados obtidos por Tziridis.
Fonte: (TZIRIDIS *et al.*, 2017)

Modelo	Precisão (%)	Tempo de execução (s)
Multilayer Perceptron	80.28	20.88
Generalized Regression Neural Network	66.83	0.13
Extreme Learning Machine	66.68	0.05
Random Forest Regression Tree	85.91	5.50
Regression Tree	84.13	0.04
Bagging Regression Tree	87.42	17.05
Regression SVM (Polynomial)	77.00	1.23
Regression SVM (Linear)	49.40	0.34
Linear Regression	57.25	0.10

Wang (2019) também analisou e modelou algoritmos de previsão para a precificação de passagens, utilizando as seguintes variáveis:

TABELA 2.2 – Descrição das variáveis utilizadas por Wang.
Fonte: (WANG *et al.*, 2019)

Nome da variável	Descrição
Classe de assento	Indicador de classe executiva ou econômica
Distância	Distância entre os aeroportos de origem e destino
Volume de passageiros	Número total de passageiros entre o aeroporto de origem e de destino
Load Factor	Taxa de ocupação do voo
Presença de LCC	Indica a presença de LCC operando no mercado
Fator de competição	Índice Herfindahl-Hirschman de concentração de mercado
Preço do petróleo	Média trimestral do preço de petróleo bruto
CPI	Média trimestral do índice de preços ao consumidor
Trimestre	Período do ano por trimestre (1, 2, 3 e 4)

Os modelos de aprendizado de máquina utilizados foram:

- M1: Linear Regression
- M2: SVM Regressor
- M3: Multi Layer Perceptron
- M4: XGBoost
- M5: Random Forest

Analisando as variáveis, Wang conseguiu atribuir importâncias para cada uma com o modelo Random Forest.

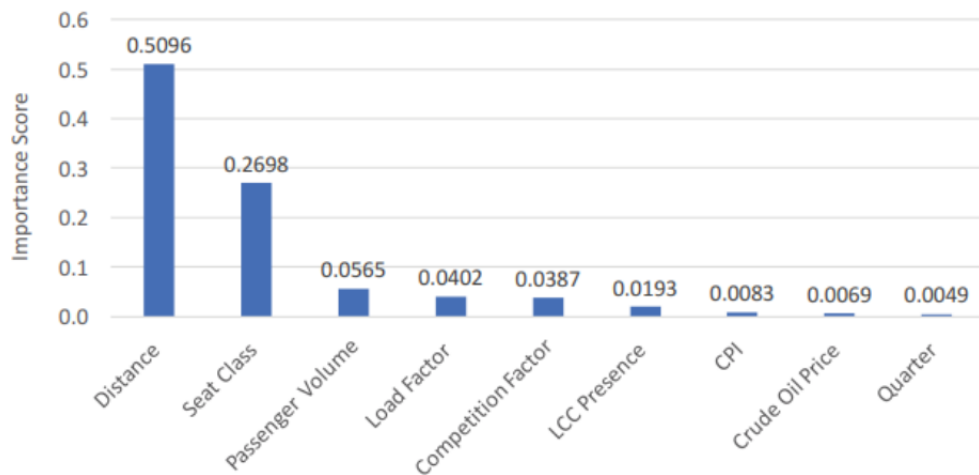


FIGURA 2.4 – Importâncias das variáveis definidas pelo modelo Random Forest.
Fonte: (WANG *et al.*, 2019)

Treinando e validando os modelos descritos, foi obtido os seguintes resultados:

TABELA 2.3 – Compilado dos resultados obtidos por Wang.
Fonte: (WANG *et al.*, 2019)

Modelo	RMSE	R_{adj}^2
LR	112.039	0.599
SVM	66.83	0.615
MLP	94.569	0.715
XGBoost	90.419	0.739
RF	70.575	0.804

Assim, utilizando o modelo Random Forest, obteve o melhor desempenho na previsão de preço, com R_{adj}^2 de 0,804 e RMSE de 70,575.

3 Metodologia

3.1 Diagrama de execução e boas práticas de *Data Science*

No presente trabalho, buscou-se utilizar as melhores práticas de Machine Learning desde a exploração inicial dos dados até a análise de resultados dos modelos aferidos. A metodologia utilizada está sequenciada no fluxograma abaixo.



FIGURA 3.1 – Fluxograma de execução do trabalho.

Fonte: autoria própria, 2021.

A linguagem de programação *Python* foi utilizada para o desenvolvimento dos modelos. Essa linguagem foi concebida no ano 1989 por Guido van Rossum no Centro de Matemática e Tecnologia da Informação (CWI, Centrum Wiskunde e Informatica), na Holanda. Dentro do ecossistema de bibliotecas *Python*, foi utilizada principalmente a biblioteca Pandas, famosa por seus recursos de análise de dados. O ambiente de desenvolvimento utilizado foi o *Google Colab*, uma plataforma poderosa para rodar scripts na nuvem, no formato similar ao do *Jupyter Notebook*.

3.2 Base de dados

A base de dados utilizada é do laboratório de transporte aéreo do ITA (LABTAR) e tem como abreviatura STATAER. A base bruta contém 234 colunas com mais de 65.000 observações de rotas aéreas nacionais entre os anos de 2010 e 2018. A base contém dados de rotas com mais de 100 pax/mês e com no mínimo 6 meses de observações macroeconômicas, assim como dados específicos da rota, variáveis de perfil socioeconômico, características das companhias aéreas e dos voos. Na tabela abaixo, ilustram-se algumas das variáveis da base.

TABELA 3.1 – Variáveis da base de dados STATAER.

Fonte: autoria própria, 2021.

Variável	Descrição
p	Preço médio das passagens aéreas na rota. Deflacionado IPCA jan 2019.
cp	Identificador da rota
speed	Velocidade média do avião na rota
pop _o	População na cidade de origem
pop _d	População na cidade de destino
seatschart	Total de assentos em voos fretados na rota
quarter	Identificador de trimestre
cargokg	Total de quilos de carga transportados na rota, englobando todos os voos regulares
bagppax	Média de quilos de bagagem transportados por passageiro nos voos da rota
asize	Tamanho do avião
schfltime	Tempo médio planejado de voo (horas)
turnmea	Tempo de turnaround
usd	Taxa de câmbio R/ <i>US</i> -comercial - compra - média
tcrewsiz	Tripulação mínima requerida pela matrícula da aeronave
conn	Proporção de pax em conexão na rota, tanto doméstico quanto internacional.
avstl _o	Média das etapas de voo do aeroporto de origem em km.
fja1	Preço de combustível

3.3 Análise exploratória dos dados

A análise exploratória tem como foco entender o comportamento dos preços ao longo do tempo e também como ele se comporta para cada rota. Para o início da exploração, primeiro temos que entender quais aeroportos estão contidos na base e quais são as rotas da malha aérea nacional. Para efeito de visualização tridimensional, foi utilizado o software gratuito *kepler.gl* (<https://kepler.gl/>).

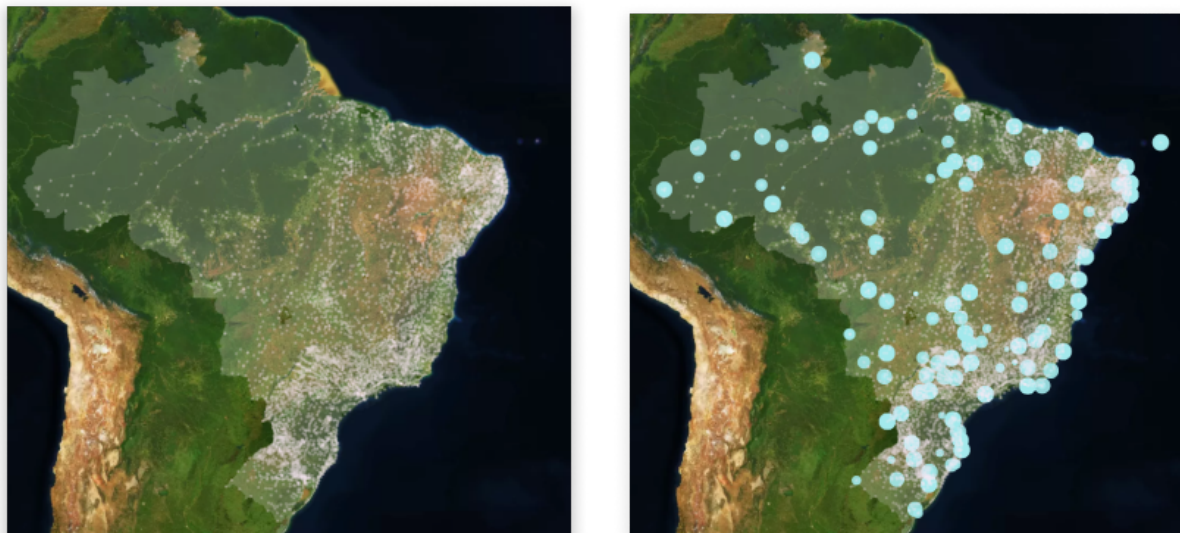


FIGURA 3.2 – Mapeamento de municípios (à esquerda) e aeroportos brasileiros (à direita) contidos na base de dados.

Fonte: autoria própria, 2021.

Na figura acima, à esquerda, temos todos os municípios Brasileiros pontilhados em branco. Na figura à direita, temos todos os aeroportos da base destacados em um círculo azul-claro, que variam de tamanho conforme a concentração de rotas daquele aeroporto.

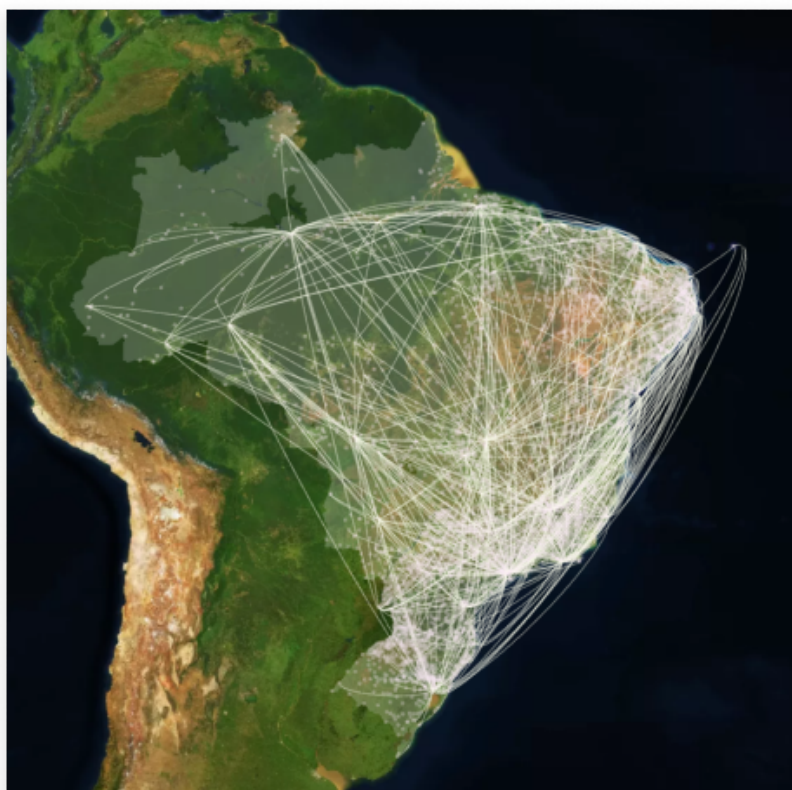


FIGURA 3.3 – Malha aérea nacional contida na base.
Fonte: autoria própria, 2021.

Foi construída a malha aérea da base, que contém um total de 1074 rotas distintas.

No que se trata da análise de preço ao longo do intervalo de tempo da base (2010 - 2018), foi analisada a variação da média móvel de todos os voos nesse período.

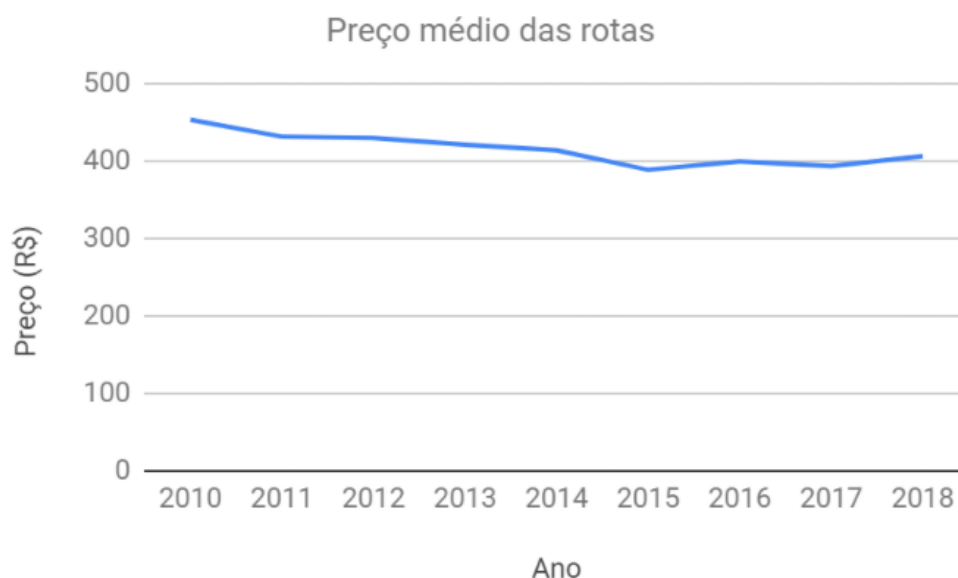


FIGURA 3.4 – Média móvel do preço de passagens deflacionados no período 2010 - 2018.
Fonte: autoria própria, 2021.

Dessa maneira, observou declínio de 10,3% do preço médio ao longo desses anos. Comparando com dados históricos da Associação Brasileira das Empresas Aéreas (2018) na figura abaixo, percebemos que a base de dados utilizada está coerente.

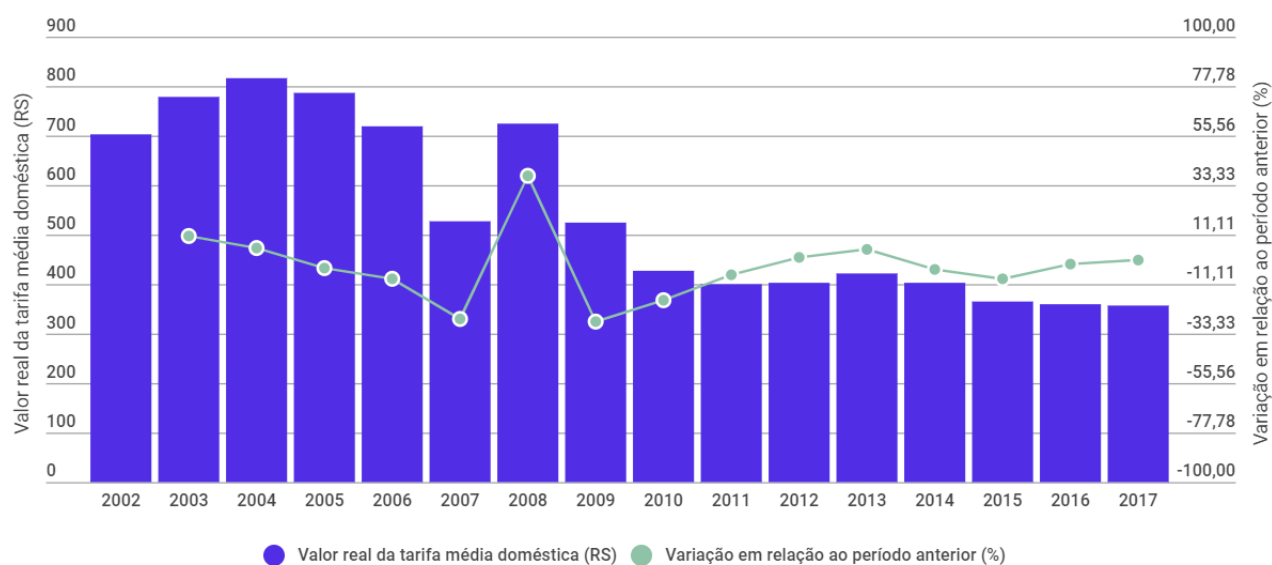


FIGURA 3.5 – Evolução da tarifa média doméstica anual no Brasil e variações percentuais (2002-2017)
Fonte: ABEAR, 2018.

Analisando essa mesma variação, segmentando por rota, observou-se que cada rota teve variações distintas nesse intervalo, demonstrado na figura abaixo com alguns exemplos.

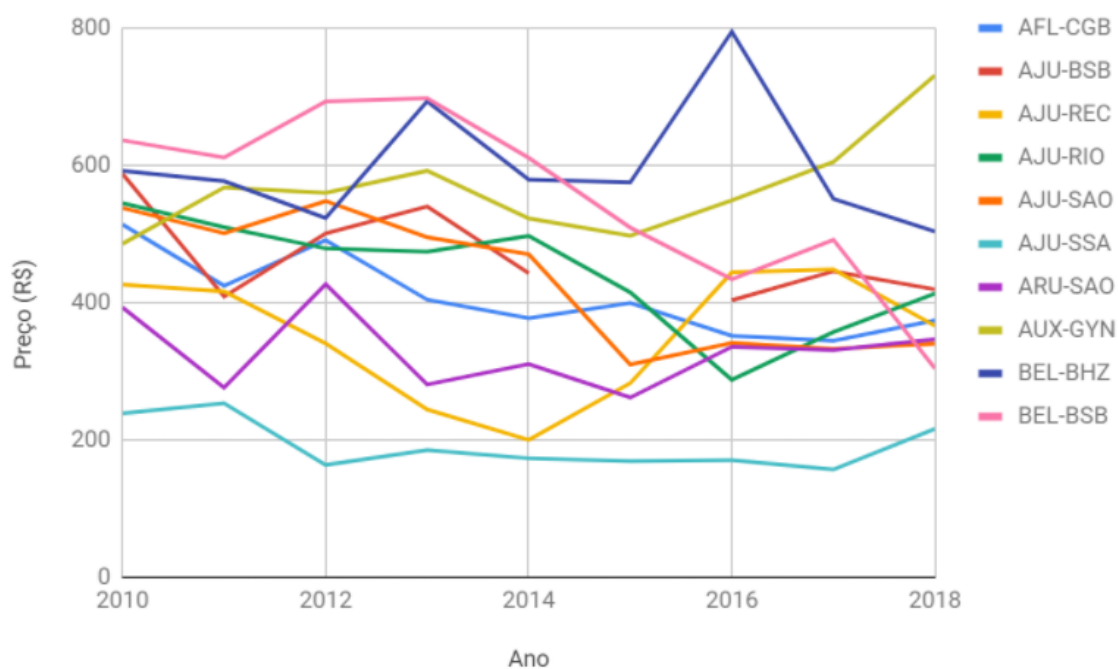


FIGURA 3.6 – Média móvel do preço de algumas rotas da base no período 2010 - 2018.
Fonte: autoria própria, 2021.

De todas as observações de preço, obtemos a distribuição de acordo com o histograma abaixo, indicando que a maior concentração de observações de preço está entre o intervalo de R\$ 250 e R\$ 500.

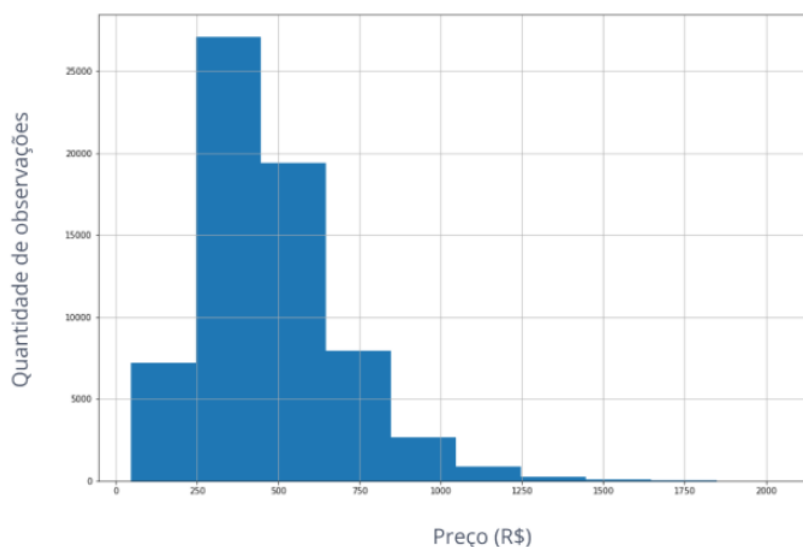


FIGURA 3.7 – Histograma da distribuição de preços.
Fonte: autoria própria, 2021.

Também foi gerado um *boxplot* que nos fornece uma visualização clara dos quartis e

da presença de observações *outliers*.

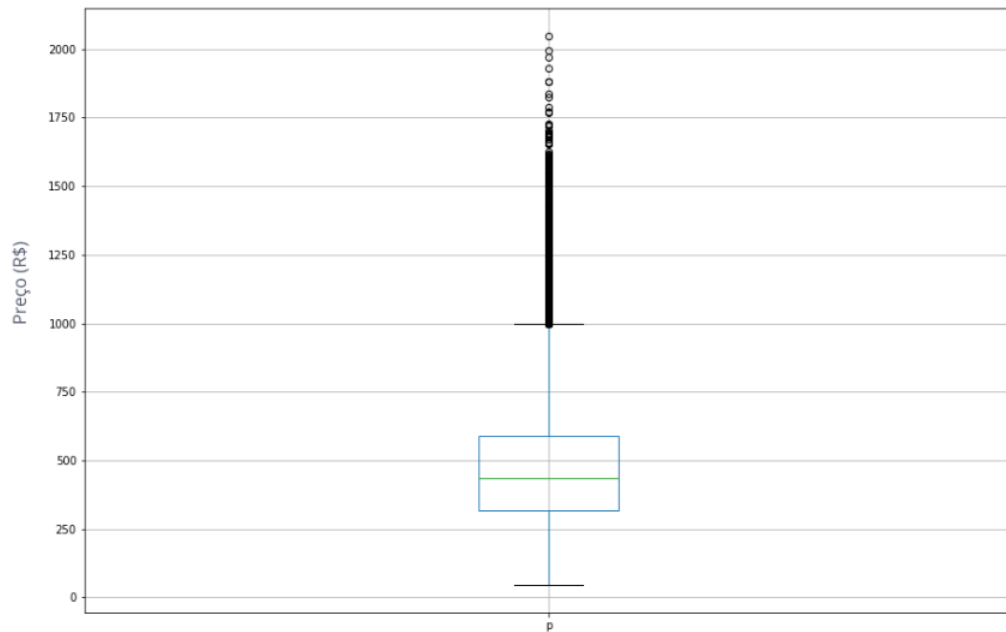


FIGURA 3.8 – *Boxplot* da distribuição de preços.
Fonte: autoria própria, 2021.

Com efeito, foi realizada então a análise de preço médio para cada rota ao longo de todo o período, que pode ser elucidado na figura abaixo, destacando o aeroporto de Guarulhos (GRU).



FIGURA 3.10 – Rotas com preços médios entre R\$150 e R\$240.
Fonte: autoria própria, 2021.



FIGURA 3.11 – Rotas com preços médios entre R\$620 e R\$774.
Fonte: autoria própria, 2021.

3.4 Pré-processamento de dados

Uma etapa que requer extrema cautela para o cientista de dados é a parte de tratamento de dados, pois a qualidade do modelo é altamente dependente de um bom pré-processamento.

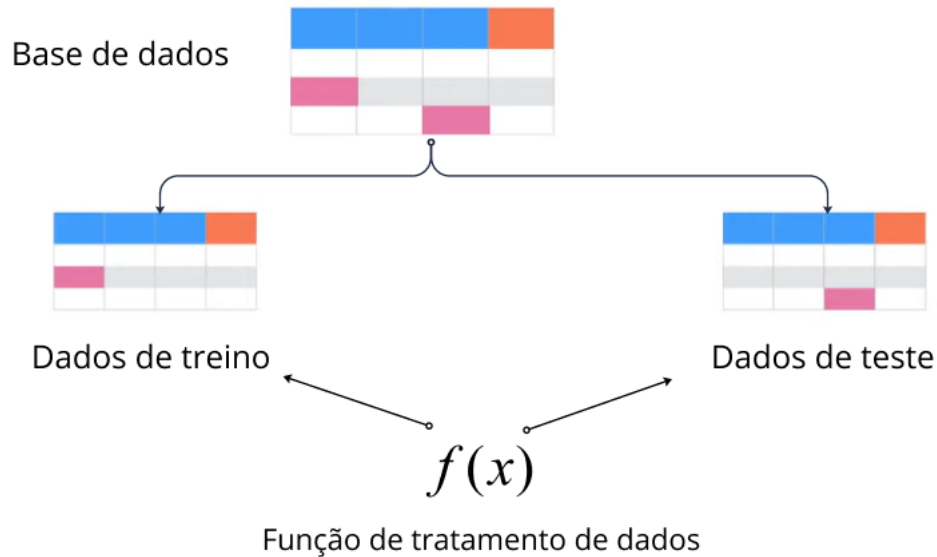


FIGURA 3.12 – Esquema da partição da base e aplicação individual da função de tratamento de dados.

Fonte: autoria própria, 2021.

A função de tratamento de dados contém uma sequência de manipulações de tratamento para deixar as bases de treino e teste adequadas para a análise. Essa função evita o problema de vazamento de dados na análise. O vazamento de dados ocorre quando as informações de fora do conjunto de dados de treinamento são usadas para criar o modelo. Essas informações adicionais podem permitir que o modelo aprenda ou saiba algo que de outra forma não saberia e, por sua vez, invalidar o desempenho estimado do modo que está sendo construído. Os métodos de tratamento incluídos nessa função foram descritos abaixo. Vale ressaltar que a base foi repartida na proporção 80/20 para dados de treino e de teste respectivamente.

3.4.1 Remoção de atributos *missing*

Foi realizada a remoção de linhas com dados faltantes, reduzindo o tamanho da base de 65.571 linhas para 41.796, uma redução de 36,25%.

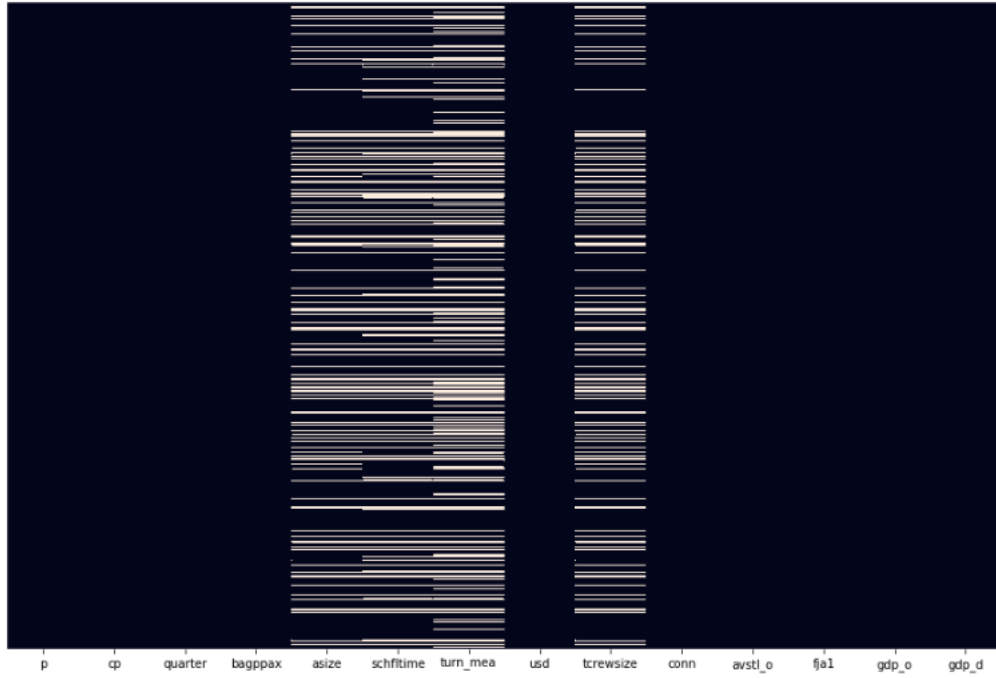


FIGURA 3.13 – Visualização das observações faltantes na base. Fonte: autoria própria, 2021.

3.4.2 Remoção de *outliers*

Um *outlier* é uma observação que parece ser inconsistente com o restante de um conjunto de dados (ESCALANTE, 2005). Foram removidas as observações de preço consideradas outliers no extremo superior do intervalo. Para isso, foi utilizado o indicador estatístico *z-score* da distribuição normal.

$$Z = \frac{X - \mu}{\sigma} \quad (3.1)$$

Portanto, observações com *z-score* maiores que 3 ou menores que -3 são identificadas e removidas. Podemos visualizar esse intervalo na curva de distribuição normal da Figura 3.14.

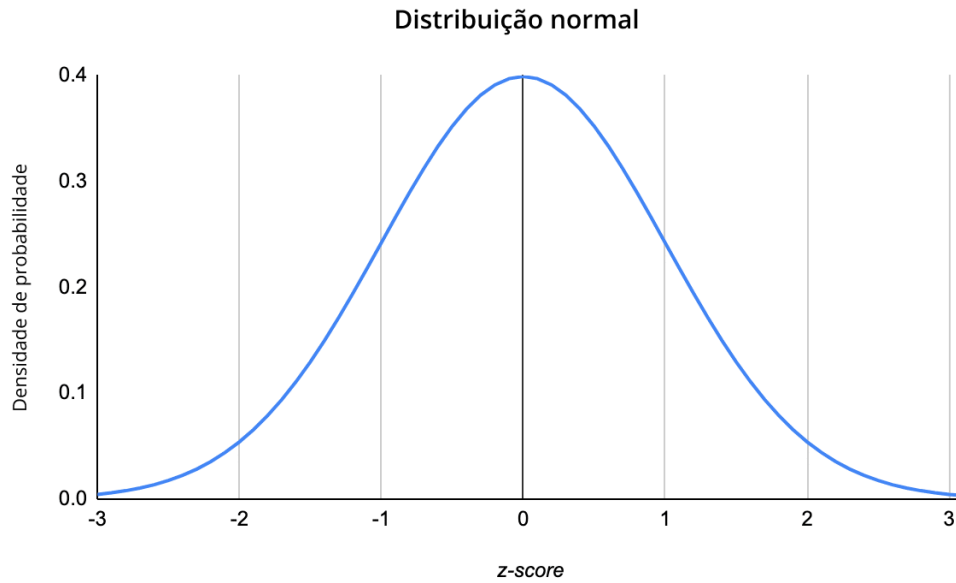


FIGURA 3.14 – Curva de distribuição normal. Fonte: autoria própria, 2021.

A base original continha preços no intervalo entre R\$ 45,96 e R\$ 2047,69. Após a remoção de *outliers*, esse intervalo foi reduzido para preços entre R\$ 45,96 e R\$ 932,69.

3.4.3 *Label-encoding* para variáveis categóricas

Esse método transforma as variáveis categóricas (*dummies*) em uma lista de números sequenciais atrelados a cada observação em formato de texto. A coluna que foi transformada foi a ‘cp’: Identificador da rota aérea (Ex: AAX-BHZ).

3.4.4 Normalizando o *dataframe*

Para um melhor desempenho dos algoritmos de aprendizado de máquina, todas as variáveis foram normalizadas para que o desvio padrão seja 1 e a média seja 0, utilizando a função *Standard Scaler*. As variáveis foram normalizadas para que a diferença entre escalas não interferisse na capacidade decisória dos algoritmos.

$$X_{norm,(i,j)} = \frac{X_{(i,j)} - mean(X_j)}{std(X_j)} \quad (3.2)$$

(FERREIRA *et al.*, 2019) articulou que esse método de normalização é sensível à presença de *outliers* na base e a dados esparsos. Além disso, é melhor aplicável para dados distribuídos em uma curva normal.

3.5 Seleção de features

A seleção de features é indispensável para a elaboração de um bom modelo. Como a base inicial contém mais de 250 colunas com diferentes variáveis, é preciso reduzir a quantidade dessas variáveis para obter uma boa performance com um custo de processamento computacional razoável. Além disso, temos a questão do *trade-off* entre a variância e o viés, que está associada à complexidade do modelo em número de variáveis. Neste trabalho, buscou-se selecionar um número adequado de variáveis que evitasse o *overfitting* e *underfitting*.

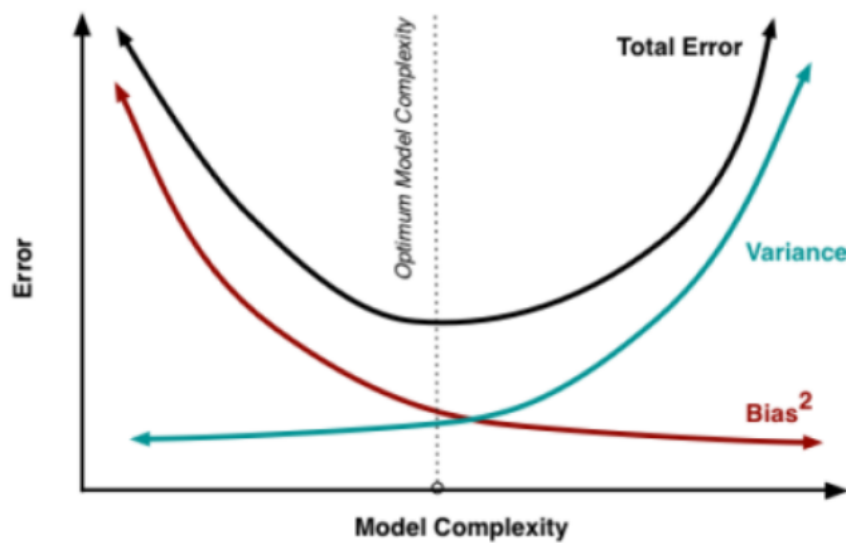


FIGURA 3.15 – *Trade-off* entre viés e variância.

Fonte: <http://scott.fortmann-roe.com/docs/BiasVariance.html>

3.5.1 Seleção manual de features

Foram selecionadas 12 variáveis dentre os dados, com base em resultados obtidos pela literatura no que tange à sua influência na precificação.

TABELA 3.2 – Features selecionadas manualmente.

Variável	Descrição
p	Preço médio das passagens aéreas na rota. Deflacionado IPCA jan 2019.
cp	Identificador da rota
gdp _o	PIB na cidade de origem
gdp _d	PIB na cidade de destino
quarter	Identificador de trimestre
bagppax	Média de quilos de bagagem transportados por passageiro nos voos da rota
asize	Tamanho do avião
schfltime	Tempo médio planejado de voo (horas)
turnmea	Tempo de turnaround
usd	Taxa de câmbio R/ <i>US</i> -comercial - compra - média
tcrewsiz	Tripulação mínima requerida pela matrícula da aeronave
conn	Proporção de pax em conexão na rota, tanto doméstico quanto internacional.
avstl _o	Média das etapas de voo do aeroporto de origem em km.
fja1	Preço de combustível

Devemos nos atentar que, no momento, essas escolhas se tratam de hipóteses com base na literatura para o desenvolvimento do nosso modelo preditivo.

Podemos justificar cada seleção com base nas seguintes argumentações:

- F1: Identificador da rota

Espera-se que cada rota tenha uma variação característica de preço, fazendo-se necessário a inclusão desse indicador para que os modelos entendam esse padrão.

- F2: Identificador de trimestre

Sabemos que o mercado aéreo é precificado conforme as estações do ano, nesse sentido, essa variável será útil para categorizar a observação conforme o período do voo.

- F3: Média de quilos de bagagem por passageiro

Parte da receita das linhas aéreas está associada com a bagagem transportada, portanto, espera-se que esse indicador tenha influência no preço médio final da passagem.

- F4: Tamanho do avião

O tamanho da aeronave está associado a diferentes custos operacionais, principalmente de manutenção. Dessa forma, aeronaves maiores naturalmente vão ser mais custosas, o que pode se refletir no preço ao consumidor.

- F5: Tempo médio de voo

O tempo de voo está atrelado ao consumo de combustível que vai determinar majoritariamente os custos operacionais.

- F6: Turnaround time

O tempo de turnaround é definido como o intervalo necessário para a aeronave desembarcar os passageiros e se preparar para a próxima viagem, às vezes necessitando de limpezas na cabine e nos assentos. Cheng-lung (2000) estimou que 22% dos custos operacionais relacionados a atrasos estão associados diretamente e indiretamente com as defasagens causadas pelo tempo de turnaround. Portanto, incluímos essa variável na nossa análise para avaliar o seu impacto no preço das passagens aéreas.

- F7: Taxa de câmbio R\$/USD

A taxa de câmbio influencia diretamente no preço de diversos componentes da aeronave, majoritariamente importados do exterior e precificados em dólar.

- F8: Tamanho da tripulação da aeronave

O tamanho da tripulação está associado com gastos de transporte de pessoal, alocação da tripulação e manutenção de bordo, refletindo, possivelmente, no preço final ao consumidor.

- F9: Média de etapas de voo

Sriram (2003) constatou que a manutenção de aeronave requer uma série de configurações que está atrelada a combinação de números de horas de voos assim como os ciclos de aterrissagem e decolagem. Portanto, podemos inferir que o número de etapas de voo pode refletir num custo maior para o preço médio das passagens aéreas contendo esses trechos.

- F10: Proporção de passageiros em conexão na rota

Passageiros buscam voos em conexão para pagar mais barato na passagem de modo geral. Porém, sabemos que essa proporção varia para cada rota de acordo com a alocação dinâmica de assentos. Dessa forma, foi incluída essa variável no nosso modelo para entendermos qual o seu impacto na determinação do preço.

- F11: Preço de combustível

McConnachie (2013) classificou o impacto do preço de combustível como o mais significativo nas despesas operacionais das linhas aéreas. Dessa forma, é de grande importância incluir essa variável no modelo.

- F12, F13: PIB da cidade de destino e PIB da cidade de origem

Ambas as variáveis estão associadas com características determinantes da demanda por viagens aéreas nas cidades. O gráfico abaixo, retirado do Panorama da Aviação Brasileira (2018), deixa claro essa correlação.

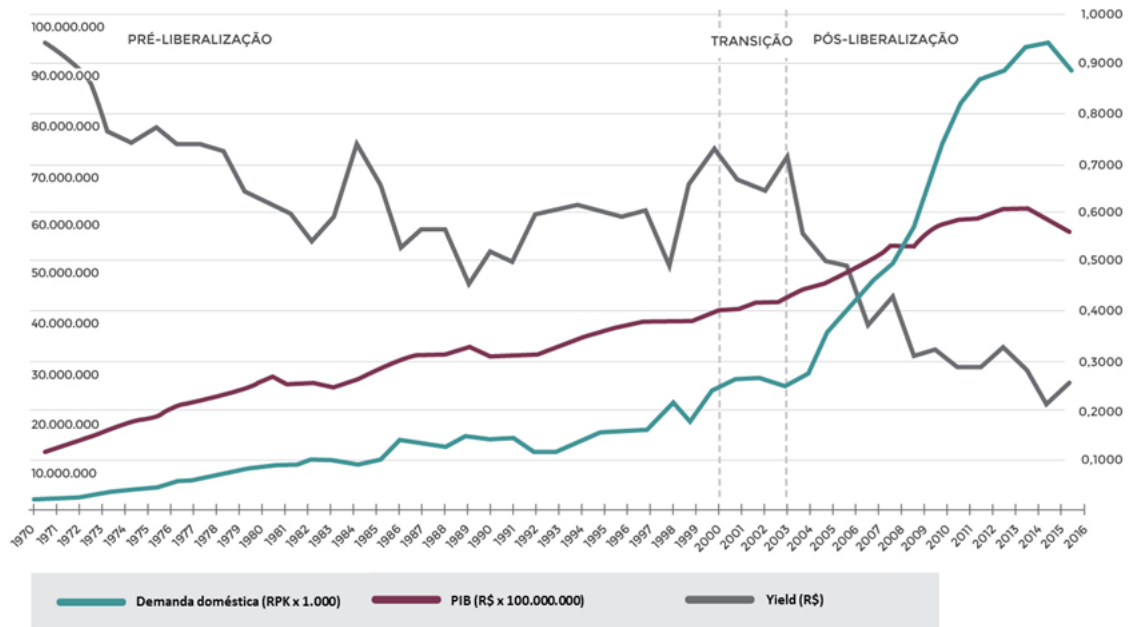


FIGURA 3.16 – Correlação entre PIB e demanda pelo transporte aéreo. Fonte: ABEAR, 2018.

Podemos, portanto, analisar a matriz de correlação entre as variáveis selecionadas.

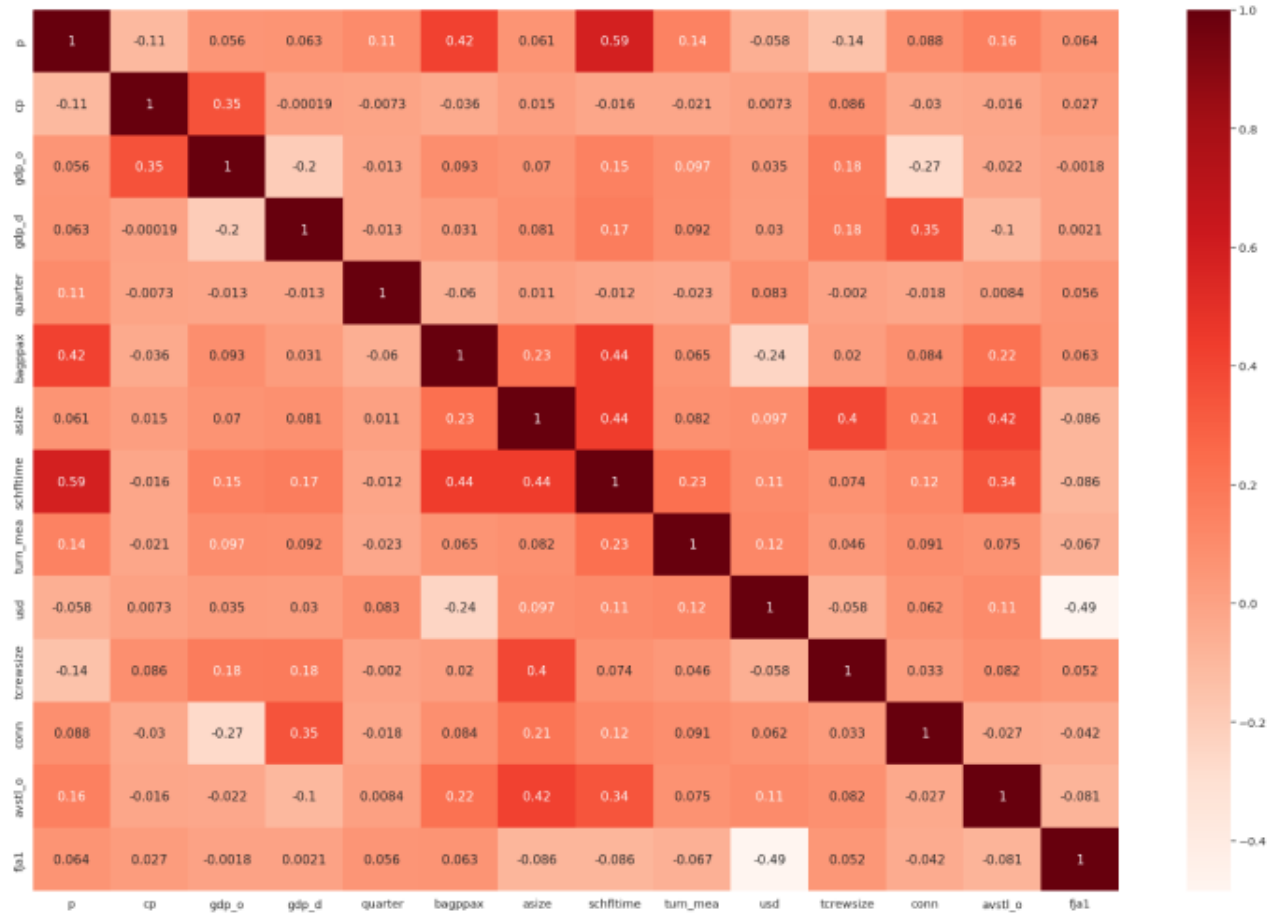


FIGURA 3.17 – Matriz de correlação entre as variáveis.

Fonte: autoria própria, 2021.

Na Figura 3.14 percebe-se uma correlação significativa do preço com o tempo de voo e o peso de bagagem por passageiro.

3.5.2 Algoritmos de seleção de features

3.5.2.1 Select K-Best

Aplicando em toda a extensão de variáveis da base, esse algoritmo retorna features altamente correlacionadas, que não colaboram para melhorar a performance dos modelos preditivos. Um objetivo fundamental da análise de regressão é isolar a relação entre cada variável independente e a variável dependente. Além disso, de acordo com a teoria de informação e aprendizado computacional *Minimum description length* (MDL), a descrição mais curta dos dados fornece um melhor modelo. Por exemplo, esta seleção incluiu as features schfftime e actfftime, que são o tempo de voo agendado e o tempo de voo real presentes na base, respectivamente. Logo, não iremos prosseguir com esse algoritmo de seleção.

3.5.2.2 Sequential forward selection

De forma similar ao Select-K-Best, esse método também define o número de features e vai adicionando a essa lista de seleções conforme percorre a base de dados. O resultado obtido foi que o algoritmo optou pelas variáveis mais correlacionadas, selecionando as mesmas que o Select-K-Best. Portanto, prosseguiremos a análise com a seleção manual de feature, porém, não deixando de levar em consideração as indicações dos algoritmos em nossa análise.

3.6 Algoritmos de aprendizado de máquina

Foram testados alguns algoritmos para a análise deste trabalho. Nem sempre é trivial a seleção dos algoritmos, pois o desempenho está estritamente associado às características da base de dados. Dessa maneira, o objetivo é encontrar um modelo que melhor se adequa ao *dataset* e gera o melhor output preditivo.

3.6.1 Regressão Linear

A regressão linear tenta estimar o valor de uma variável dependente em função das demais variáveis independentes. A equação geral da regressão linear é dada por:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n \quad (3.3)$$

Em que y é a variável dependente, β_0 é o coeficiente linear, β_i com i pertencente $[1, n]$, n inteiro, são coeficientes que são calculados representando cada uma das variáveis.

3.6.2 Regressão Lasso

A regressão LASSO (Least Absolute Shrinkage and Selection Operator) é um método de regressão penalizada desenvolvida por Tibshirani (1996) como uma alternativa à Regressão Ridge. Essa regressão é feita adicionando um termo na função de custo, de forma que minimizar a função de custo automaticamente minimize também os coeficientes. Quando há múltiplas features altamente correlacionadas (ou seja, features que se comportam da mesma maneira) a regularização Lasso seleciona apenas uma dessas features e zera os coeficientes das outras, de forma a minimizar a penalização.

$$\min \left\{ \frac{1}{N} \sum_{i=1}^j (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{i=1}^m |\beta_i| \right\} \quad (3.4)$$

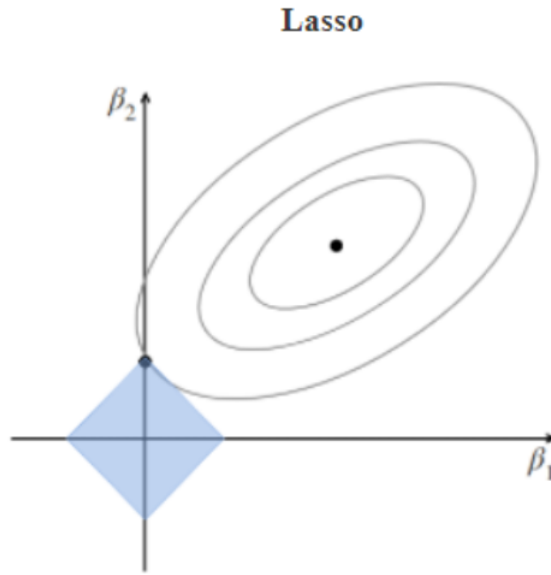


FIGURA 3.18 – Condição de contorno para a regressão Lasso.

Fonte: <http://cursos.leg.ufpr.br/ML4all/apoio/Regularizacao.html>

3.6.3 Regressão Elastic Net

A regressão Elastic Net se trata de combinar os termos de regularização de Lasso e de Ridge (Método análogo ao Lasso, porém com diferente condição de contorno). Assim, obtemos o melhor de cada algoritmo, porém temos que enfrentar o problema de determinar dois hiperparâmetros para obter soluções ótimas.

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^n (y_i - x_i' \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^m \hat{\beta}_j^2 + \alpha \sum_{j=1}^m |\hat{\beta}_j| \right) \quad (3.5)$$

3.6.4 Regressor Support Vector Machine (SVM)

O modelo SVM pode ser usado tanto como classificação como regressão. O modelo utilizado será o de regressão (SVR) que delimita um hiperplano de decisão sobre as observações, criando uma fronteira de regressão entre os dados. O melhor modelo é aquele que se ajusta entre as variáveis minimizando o erro quadrático e obdecendo às condições de contorno (SETHI, 2020).

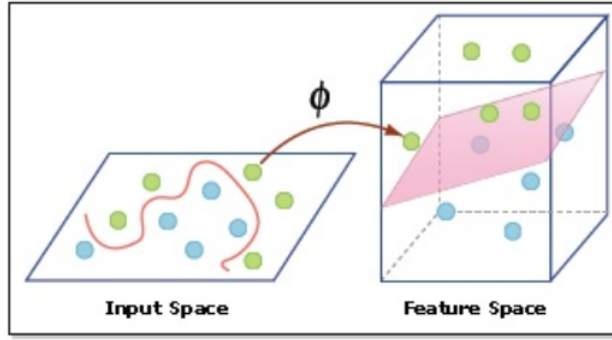


FIGURA 3.19 – Hiperplano de ajuste SVR.
Fonte: (SETHI, 2020)

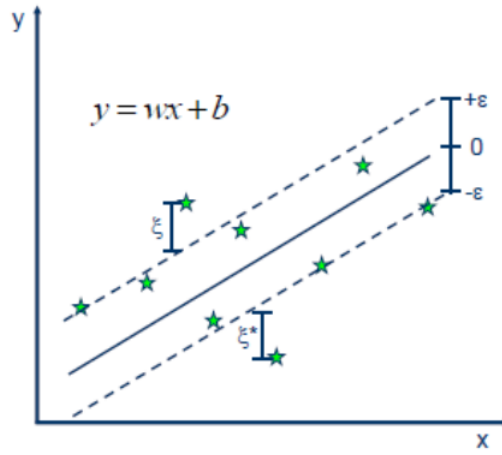


FIGURA 3.20 – Função de minimização de erro SVR.
Fonte: (SETHI, 2020)

Função de minimização de erro:

$$\min \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*) \right\} \quad (3.6)$$

Condições de contorno:

$$y_i - wx_i - b \leq \epsilon + \xi_i \quad (3.7)$$

$$wx_i + b - y_i \leq \epsilon + \xi_i^* \quad (3.8)$$

$$\xi_i, \xi_i^* \geq 0 \quad (3.9)$$

3.6.5 Random Forest

O modelo Random Forest foi publicado pela primeira vez na literatura por L. Breiman em 2001 (BREIMAN, 2001) e consiste em múltiplas árvores de decisão em paralelo. A 'floresta' gerada pelo algoritmo Random Forest é treinada por meio de *Bagging*. *Bagging* é um meta-algoritmo de conjunto que melhora a precisão dos algoritmos de aprendizado de máquina. Os modelos de RF demonstraram ser preditores robustos para tamanhos de amostra pequenos e dados dimensionais elevados.

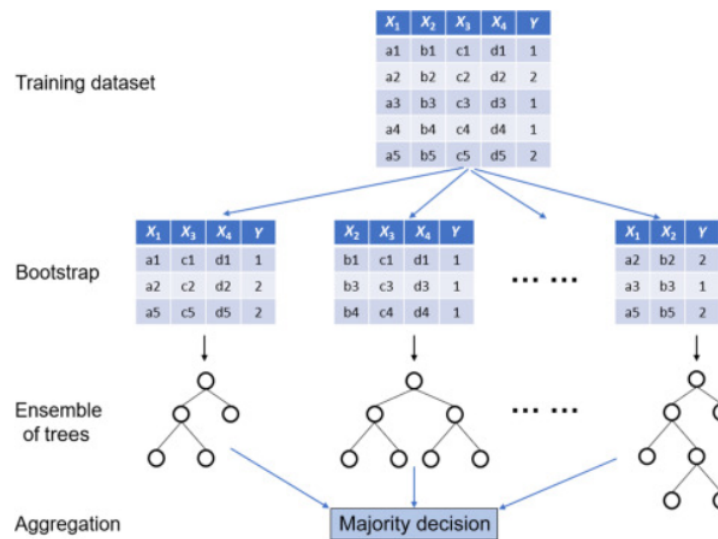


FIGURA 3.21 – Esquema de repartição e ajuste do modelo Random Forest.

3.6.6 XGBoost

O algoritmo XGBoost foi desenvolvido como um projeto de pesquisa na Universidade de Washington. Seus desenvolvedores são Tianqi Chen e Carlos Guestrin e apresentaram o algoritmo por meio de um artigo na Conferência SIGKDD. XGBoost é o termo abreviado para o pacote Extreme Gradient Boosting, um algoritmo de aprendizado de máquina baseado na árvore de decisão, que usa uma estrutura de Gradient Boosting, um caso especial de reforço em que os erros são minimizados pelo algoritmo de gradiente descendente (*gradient descent*).

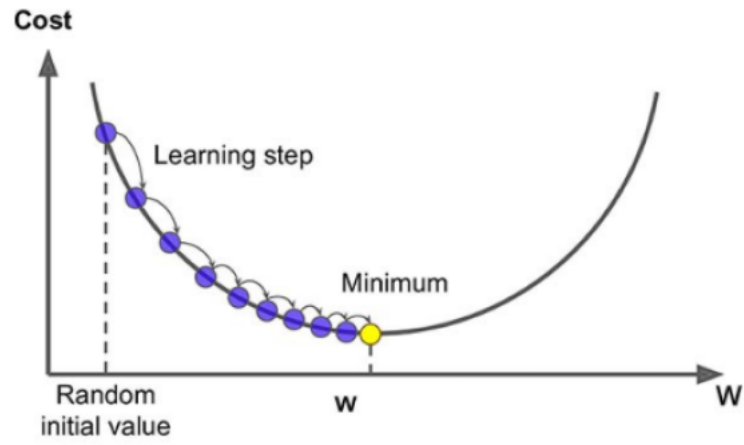


FIGURA 3.22 – Processo iterativo do gradiente descendente.

O algoritmo de aprendizado iterativo é descrito pela expressão:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3.10)$$

4 Análise de resultados

Este capítulo sintetiza todos os resultados dos modelos de aprendizado de máquina construídos em nossa análise. Vale ressaltar que os modelos foram importados de bibliotecas *python* chamada *Scikit-Learn*, que são um compilado de códigos *open-source* que contém contribuições de milhares de desenvolvedores pelo mundo.

4.1 Métricas de avaliação

As métricas de avaliação selecionadas neste trabalho são as mais utilizadas para modelos de regressão. Portanto, nossa análise irá se basear nessas métricas para seleção do melhor modelo.

4.1.1 RMSE

Root mean squared error:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2} \quad (4.1)$$

4.1.2 MAE

Mean absolute error:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_{test}^{(i)} - \hat{y}_{test}^{(i)}| \quad (4.2)$$

4.1.3 MAPE

Mean absolute percentage error:

$$MAPE = \frac{100\%}{m} \sum_{i=1}^m \left| \frac{y_{test}^{(i)} - \hat{y}_{test}^{(i)}}{y_{test}^{(i)}} \right| \quad (4.3)$$

4.1.4 R_{adj}^2

O coeficiente de determinação varia entre 0 e 1 e é uma medida de ajuste generalizada para modelos estatísticos.

$$SQ_{exp} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4.4)$$

$$SQ_{tot} = \sum_{i=1}^n (y_i - \hat{y})^2 \quad (4.5)$$

$$R^2 = \frac{SQ_{exp}}{SQ_{tot}} \quad (4.6)$$

$$R_{adj}^2 = 1 - \left[\frac{n-1}{n-k-1} (1 - R^2) \right] \quad (4.7)$$

Em que k é o número de preditores e n é o tamanho da amostra.

4.2 Resultados

Para entendermos o desempenho dos modelos, vamos analisar como estes se comportam em relação as features. Para isso, foram executados os modelos retirando uma das features e analisando o impacto daquela variável nas métricas de desempenho.

TABELA 4.1 – Compilado dos resultados obtidos com todas as variáveis inclusas.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	73.828	0.771	54.526	14.934
2	XGBoost	97.005	0.605	73.897	20.922
3	Linear Regression	111.204	0.481	85.140	24.186
4	Lasso	111.204	0.481	85.141	24.186
5	Elastic Net	111.229	0.481	85.170	24.193
6	SVM	139.296	0.186	107.784	30.339

Portanto, o Random Forest se mostrou o melhor modelo, com precisão de 85.08% que representa um erro médio de R\$ 54,44 no preço das passagens. Sabendo-se que o preço médio da base de dados é de R\$ 473,22, esse erro é aceitável para realizar estimativas nessas rotas.

Em seguida, vamos repetir a execução dos modelos removendo algumas variáveis da base.

4.2.1 Variando features de input

Foram nomeadas todas as features do modelo segunda a tabela 4.2.

TABELA 4.2 – Variáveis selecionadas nomeadas por $F\#N$.

Variável	Descrição
F1	Identificador da rota
F2	PIB na cidade de origem
F3	PIB na cidade de destino
F4	Identificador de trimestre
F5	Média de quilos de bagagem transportados por passageiro nos voos da rota
F6	Tamanho do avião
F7	Tempo médio planejado de voo (horas)
F8	Tempo de turnaround
F9	Taxa de câmbio R/ <i>US</i> -comercial - compra - média
F10	Tripulação mínima requerida pela matrícula da aeronave
F11	Proporção de pax em conexão na rota, tanto doméstico quanto internacional.
F12	Média das etapas de voo do aeroporto de origem em km.
F13	Preço de combustível

Portanto, obtemos os seguintes resultados:

TABELA 4.3 – Resultados obtidos excluindo as features F2 e F3.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	75.645	0.760	55.657	15.203
2	XGBoost	97.286	0.604	74.109	20.915
3	Linear Regression	111.292	0.481	85.259	24.155
4	Lasso	111.292	0.481	85.259	24.155
5	Elastic Net	111.307	0.481	85.271	24.158
6	SVM	139.898	0.180	108.383	30.521

TABELA 4.4 – Resultados obtidos excluindo a feature F5.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	71.765	0.784	52.329	14.169
2	XGBoost	98.106	0.596	75.015	21.261
3	Linear Regression	113.340	0.461	87.159	24.736
4	Lasso	113.340	0.461	87.159	24.736
5	Elastic Net	113.362	0.461	87.189	24.746
6	SVM	139.105	0.188	107.644	30.271

TABELA 4.5 – Resultados obtidos excluindo a feature F6.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	77.753	0.746	57.074	15.635
2	XGBoost	99.670	0.583	75.671	21.436
3	Linear Regression	113.660	0.458	87.295	24.820
4	Lasso	113.660	0.458	87.296	24.820
5	Elastic Net	113.665	0.458	87.303	24.821
6	SVM	139.392	0.184	107.903	30.377

TABELA 4.6 – Resultados obtidos excluindo a feature F7.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	89.762	0.664	67.095	19.107
2	XGBoost	109.569	0.500	84.219	24.276
3	Linear Regression	133.032	0.262	104.450	30.470
4	Lasso	133.033	0.262	104.451	30.471
5	Elastic Net	133.056	0.262	104.469	30.476
6	SVM	148.083	0.086	115.218	32.474

TABELA 4.7 – Resultados obtidos excluindo a feature F8.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	75.622	0.771	55.241	15.041
2	XGBoost	99.807	0.601	76.018	21.378
3	Linear Regression	115.302	0.467	88.260	24.946
4	Lasso	115.302	0.467	88.261	24.946
5	Elastic Net	115.320	0.467	88.282	24.951
6	SVM	141.659	0.196	109.684	30.744

TABELA 4.8 – Resultados obtidos excluindo a feature F9.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	74.940	0.764	55.431	15.180
2	XGBoost	97.247	0.603	74.077	20.926
3	Linear Regression	111.328	0.480	85.188	24.206
4	Lasso	111.329	0.480	85.189	24.206
5	Elastic Net	111.353	0.480	85.213	24.211
6	SVM	139.208	0.187	107.708	30.318

TABELA 4.9 – Resultados obtidos excluindo a feature F10.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	74.924	0.761	55.463	15.307
2	XGBoost	98.352	0.589	75.142	21.380
3	Linear Regression	111.147	0.475	85.064	24.249
4	Lasso	111.147	0.475	85.064	24.249
5	Elastic Net	111.176	0.475	85.101	24.259
6	SVM	134.884	0.227	103.906	28.843

TABELA 4.10 – Resultados obtidos excluindo a feature F11.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	73.140	0.778	53.700	14.604
2	XGBoost	97.762	0.603	74.745	21.151
3	Elastic Net	112.479	0.474	86.296	24.594
4	Lasso	112.481	0.474	86.303	24.597
5	Linear Regression	112.481	0.474	86.304	24.598
6	SVM	139.772	0.188	108.286	30.607

TABELA 4.11 – Resultados obtidos excluindo a feature F12.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	74.005	0.770	54.306	14.854
2	XGBoost	97.373	0.602	74.179	20.977
3	Linear Regression	111.149	0.482	85.045	24.159
4	Lasso	111.149	0.482	85.045	24.159
5	Elastic Net	111.176	0.482	85.081	24.168
6	SVM	138.376	0.197	107.152	29.951

TABELA 4.12 – Resultados obtidos excluindo a feature F13.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	74.666	0.766	54.936	15.016
2	XGBoost	97.228	0.603	73.861	20.899
3	Linear Regression	111.585	0.477	85.349	24.199
4	Lasso	111.585	0.477	85.349	24.199
5	Elastic Net	111.610	0.477	85.382	24.208
6	SVM	139.206	0.187	107.708	30.318

De modo geral, percebemos que os modelos Linear Regression, Lasso e Elastic Net geraram resultados bem próximos. Apesar dos modelos Lasso e Elastic Net serem de regressão penalizada, percebemos que essa penalização não obteve vantagem na nossa amostra de dados.

Por outro lado, o modelo RF se sobressai em todas as circunstâncias, como seria de se esperar de acordo com os trabalhos de (WANG *et al.*, 2019) e (TZIRIDIS *et al.*, 2017) que estavam em condições preditivas semelhantes. A *feature* que mais impactou negativamente o desempenho dos modelos foi a F7 (Tempo agendado de voo). Por outro lado, notamos que ao excluir a *feature* F5 (Média de quilos transportados por passageiros) trouxe o melhor desempenho, com $R_{adj}^2 = 0.784$ e precisão de 85,83%.

4.2.2 Random Forest

O modelo Random Forest possui um algoritmo de análise de importância de feature construído internamente no pacote Scikit-Learn. Esse algoritmo é chamado de Importância de Gini ou Diminuição média na impureza (MDI), que calcula a importância de cada recurso como a soma sobre o número de divisões (em todas as árvores) que incluem o recurso, proporcionalmente ao número de amostras que ele divide.

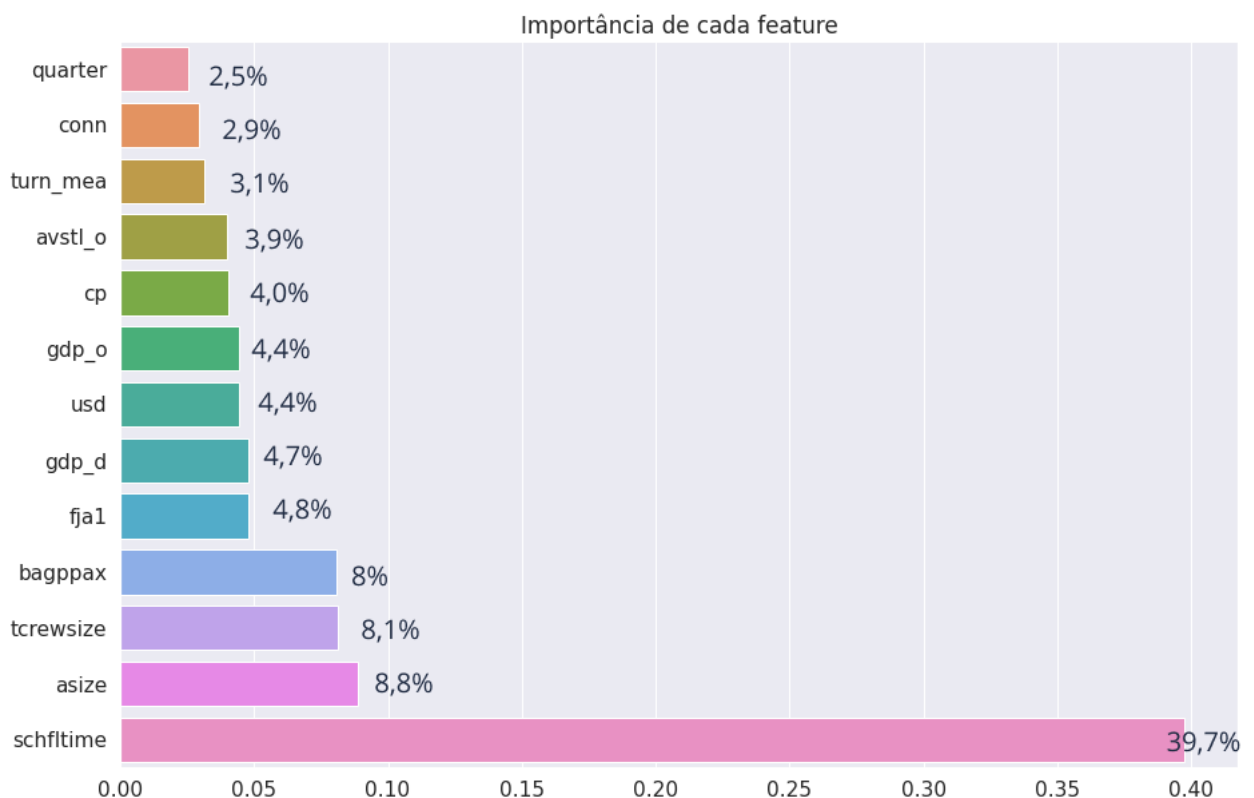


FIGURA 4.1 – Peso de importância preditiva das variáveis no modelo RF.

Percebemos que a variável de tempo de voo definitivamente é a mais significativa na predição do preço (39,7%), seguida pelo tamanho da aeronave (8,8%) e pelo tamanho médio da tripulação (8,1%). Por outro lado, percebemos que o trimestre teve a menor importância preditiva (2,5%), contrariando a nossa hipótese de que o período do ano teria um peso significativo na determinação dos preços.

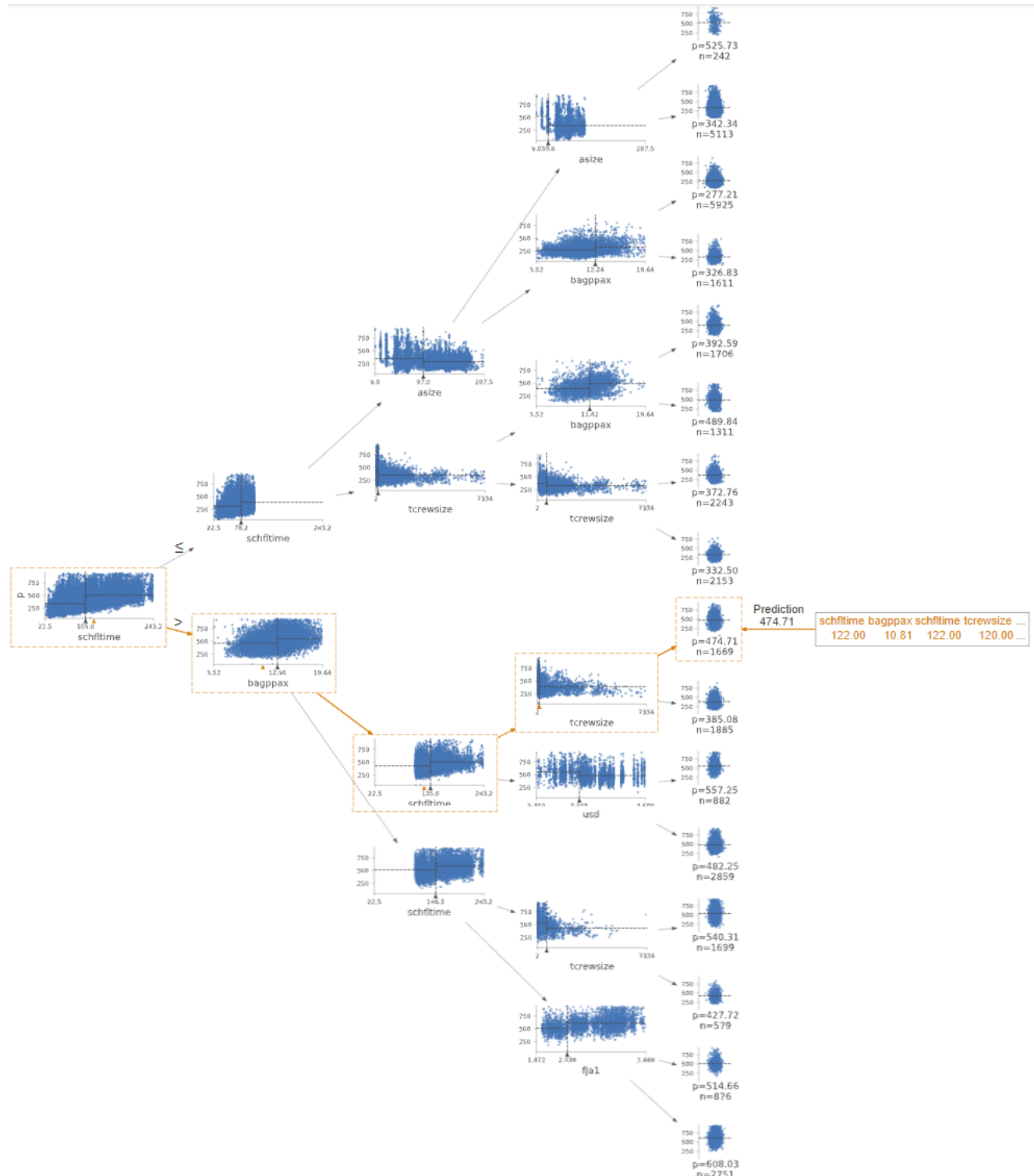


FIGURA 4.2 – Fluxo decisório do algoritmo Random Forest.

Podemos observar na Figura 4.2 o fluxo de decisão do algoritmo, estimando o preço médio da passagem aérea no trecho de R\$474,71 com base nos valores das variáveis dependentes.

Para explorar de maneira mais intensiva o modelo Random Forest, analisamos o de-

sempenho adicionando progressivamente as variáveis, explicitadas na tabela 4.13.

TABELA 4.13 – Importâncias das features (%) para nove diferentes modelos usando o algoritmo Random Forest.

	Variável	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
F7	schfltime	54,35	48,82	44,75	43,00	42,48	41,23	40,57	40,14	39,12
F6	asize	25,88	20,06	17,25	15,06	12,44	11,38	10,59	9,87	8,95
F4	bagppax	19,76	15,42	12,89	11,20	10,66	9,66	9,17	8,97	8,22
F10	tcrewsiz		15,68	13,59	11,73	11,28	10,30	9,53	8,99	8,02
F13	fja1			11,50	9,33	8,72	7,77	7,23	6,86	4,89
F3	gdp _d				9,66	8,36	7,19	6,42	5,74	4,78
F8	turnmea					6,02	5,04	4,30	4,09	3,28
F2	gdp _o						7,24	6,37	5,57	4,47
F12	avstl _o							5,77	5,27	4,08
F1	cp								4,44	4,02
F9	usd									4,50
F11	conn									2,90
F4	quarter									2,58
	RMSE	98,07	88,01	80,67	78,18	76,41	73,65	72,41	73,15	75,64
	R^2_{adj}	0,615	0,693	0,742	0,758	0,758	0,775	0,782	0,777	0,771

De modo geral, percebemos um aumento progressivo da precisão do modelo até a inclusão da 9ª variável no modelo (7). Os modelos (8) e (9) não apresentaram incremento no desempenho, que pode ser justificado pela baixa importância das variáveis adicionadas: *avstl_o*, *cp*, *usd*, *conn* e *quarter*.

4.2.3 Ajuste de hiperparâmetro

Para atingir um melhor desempenho do modelo Random Forest, buscou-se melhorar o modelo por ajustes de hiperparâmetros. No aprendizado de máquina, um hiperparâmetro é um parâmetro cujo valor é usado para controlar o processo de aprendizado. Abaixo segue exemplos de hiperparâmetros do modelo RF:

- *n_estimators* = Número de árvores na floresta;
- *max_features* = Máximo número de variáveis considerada para subdividir um nó;

- `max_depth` = Máximo número de camadas em cada árvore de decisão;
- `min_samples_split` = Mínimo número de dados em um nó antes da subdivisão;

Para obtermos o melhor ajuste de hiperparâmetros, foi utilizado um algoritmo chamado de *Random Search Cross Validation*. Esse algoritmo de validação cruzada realiza vários testes com diferentes combinações de hiperparâmetros e encontra a melhor combinação possível, ou seja, aquela que configura máxima precisão de acerto dentre os intervalos especificados.

TABELA 4.14 – Resultados obtidos após a otimização de hiperparâmetro.

	Modelo	RMSE	R_{adj}^2	MAE	MAPE (%)
1	Random Forest	68.946	0.800	50.659	13.875

Portanto, percebemos um aumento de precisão de 1,328%, uma melhoria considerável para nossa precisão, reduzindo o erro absoluto no preço em R\$ 3,86.

5 Conclusão

Este trabalho teve como objetivo estudar e treinar modelos de aprendizado de máquina para realizar previsões assertivas de preço de passagem aéreas em trechos nacionais, para que tenhamos em evidência quais variáveis contribuem para essa flutuação periódica das tarifas. Na literatura do transporte aéreo, trabalhos similares foram realizados para prever o preço de passagens aéreas nos Estados Unidos, porém não foram encontrados trabalhos similares no contexto nacional.

Nessa análise, foi possível entender o grau de dificuldade de conseguir prever com eficácia essa precificação, pois entendemos que as políticas de RM, cada vez mais complexas, torna a proposta desse trabalho ainda mais difícil de ser alcançada. No entanto, possuindo uma base de dados de tamanho significativa com dados relevantes e se baseando na literatura do transporte aéreo e de estudos de tMachine Learning, pudemos atravessar as dificuldades e chegar a um bom resultado final, com um erro de predição adequado quando comparamos com demais resultados obtidos na literatura.

Foi possível explorar a base de dados para entender alguns padrões da precificação da malha aérea nacional. Dentre a nossa seleção de variáveis, o modelo Random Forest indicou que o tamanho da tripulação, a média de bagagem por passageiros (kg), o tamanho da aeronave e o tempo médio de voo agendado tiveram importância de Gini maiores que 8%, configurando as melhores variáveis preditivas. Dentre essas variáveis, o tempo de voo se sobressaiu na determinação da precificação, de tal forma que ao retirar essa variável dos modelos preditivos, tivemos uma redução de precisão de 4,24%. Dada essa correlação, podemos inferir que estas variáveis provavelmente estão incluídas no modelo de gerenciamento de receita das companhias aéreas nacionais.

Entre os modelos de aprendizado de máquina selecionados, o modelo Random Forest se destacou em todas as análises, obtendo uma precisão máxima, após os ajustes de hiperparâmetro, de 86,12% com $R_{adj}^2 = 0.80$. Este resultado significa que há um erro médio de R\$50 na previsão de preço das passagens. Em suma, podemos concluir que o modelo Random Forest conseguiu alcançar um bom aprendizado na base de dados utilizada, obtendo uma precisão adequada frente às limitações do nosso estudo.

5.1 Limitações do trabalho

Na maioria dos trabalhos de aprendizado de máquina e de ciência de dados, a qualidade do resultado obtido depende majoritariamente da qualidade da base de dados. Apesar das funções de tratamentos utilizadas, sabemos que temos como limitação as variáveis disponíveis na base STATAER utilizada, assim como possíveis erros pela extensão das observações da base.

5.2 Sugestões para trabalhos futuros

Para progredir os estudos nessa linha, sugere-se testar novas bases de dados e novas variáveis utilizando o modelo Random Forest, para que se possa descobrir outras correlações em relação ao preço. Nesse sentido, espera-se que os métodos de previsão devam evoluir com o passar do anos, visto que o mercado aéreo e o comportamento dos passageiros se adaptam. Portanto, analisar o desempenho de previsão de preço para diferentes intervalos de tempo pode trazer evidências de como o mercado aéreo se transforma e assim obter padrões mais claros de quais variáveis se mantêm preditivas ou não.

Referências

ABDELLA, J. A. Airline ticket price and demand prediction: A survey. **Journal of King Saud University - Computer and Information Sciences**, v. 32, n. 0, p. 12–17, 2020.

Disponível em:

<<https://www.sciencedirect.com/science/article/pii/S131915781830884X>>. Acesso em: 12 ago. 2021.

BOYD, E.; KALLESEN, R. Practice papers: The science of revenue management when passengers purchase the lowest available fare. **Journal of Revenue and Pricing Management**, v. 3, p. 171–177, 07 2004.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 10 2001.

CHAGAS, L. V. C. das. Modelagem computacional e econômico financeiro da evolução temporal dos preços de passagens aéreas: Um estudo sob a perspectiva das companhias aéreas nacionais. **Biblioteca do ITA**, 2018. Disponível em:

<<http://www.civil.ita.br/graduacao/tgs/resumos/display.php?File=tg2018-11.php>>. Acesso em: 12 ago. 2021.

ESCALANTE, H. J. A comparison of outlier detection algorithms for machine learning. **Programming and Computer Software**, 01 2005.

FERREIRA, P.; LE, D. C.; ZINCIR-HEYWOOD, N. Exploring feature normalization and temporal information for machine learning based insider threat detection. In: **2019 15th International Conference on Network and Service Management (CNSM)**. [S.l.: s.n.], 2019. p. 1–7.

FIIG T., O. G. R. A.; SMITH., B. Dynamic pricing — the next revolution in rm? **Journal of Revenue and Pricing Management**, v. 15, p. 360–379, 2016. Disponível em: <<https://link.springer.com/article/10.1057/rpm.2016.28>>. Acesso em: 15 ago. 2021.

GERARDI, K.; SHAPIRO, A. The effects of competition on price dispersion in the airline industry: A panel analysis. **SSRN Electronic Journal**, 02 2007.

GILLEN, B. M. D. Price volatility in the airline markets. **Transportation Research Part E Logistics and Transportation Review**, v. 45, n. 0, p. 693–709, 2009. Disponível em: <https://www.researchgate.net/publication/227427271_Price_volatility_in_the_airline_markets>. Acesso em: 08 out. 2021.

- GORIN, T. O.; BRUNGER, W.; WHITE, M. No-show forecasting: A blended cost-based, pnr-adjusted approach. **Journal of Revenue and Pricing Management**, v. 5, p. 188–206, 2006.
- GROVES, W.; GINI, M. An agent for optimizing airline ticket purchasing. In: . [S.l.: s.n.], 2013. v. 2, p. 1341–1342.
- JANSSEN, T. A linear quantile mixed regression model for prediction of airline ticket prices. 2014.
- KLEIN SEBASTIAN KOCH, C. S. A. K. R. A review of revenue management: Recent generalizations and advances in industry applications. **European Journal of Operational Research**, v. 283, p. 397–412, 2020. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/S0377221719305272>>. Acesso em: 12 ago. 2021.
- MCAFEE, R. Dynamic pricing in the airline industry. **Handbook on Economics and Information Systems**, v. 1, 01 2006.
- MCCONNACHIE, D.; WOLLERSHEIM, C.; HANSMAN, R. The impact of fuel price on airline fuel efficiency and operations. In: . [S.l.: s.n.], 2013. ISBN 978-1-62410-225-7.
- OLIVEIRA, R. B. F. A. V. M. de. Overbooking, gerenciamento de receitas e previsão de demanda: Estudo empírico das posições em sistemas de reservas de companhias aéreas. **REVISTA DE ADMINISTRAÇÃO CONTEMPORÂNEA**, v. 12, 2008.
- PAPADAKIS, M. Predicting airfare prices. In: . [S.l.: s.n.], 2012.
- REN Y. YANG, S. Y. R. Prediction of airline ticket price, technical report, stanford university. In: . [S.l.: s.n.], 2015.
- SETHI, A. Support vector regression tutorial for machine learning. **Machine Learning**, 2020. Disponível em: <<https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>>. Acesso em: 02 OUT. 2021.
- SMITH J. LEIMKUHLE, R. D. B.; SAMUELS. Yield management american airlines. **2011 10th International Conference on Machine Learning and Applications and Workshops**, v. 22, n. 0, p. 8–31, 1992.
- SRIRAM, C.; HAGHANI, A. An optimization model for aircraft maintenance scheduling and re-assignment. **Transportation Research Part A-policy and Practice**, v. 37, p. 29–48, 2003.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246. Disponível em: <<http://www.jstor.org/stable/2346178>>.
- TZIRIDIS, K.; KALAMPOKAS, T.; PAPAKOSTAS, G. A.; DIAMANTARAS, K. I. Airfare prices prediction using machine learning techniques. In: **2017 25th European Signal Processing Conference (EUSIPCO)**. [S.l.: s.n.], 2017. p. 1036–1039.

WANG, T.; POUYANFAR, S.; TIAN, H.; TAO, Y.; ALONSO, M.; LUIS, S.; CHEN, S.-C. A framework for airfare price prediction: A machine learning approach. In: **2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)**. [S.l.: s.n.], 2019. p. 200–207.

WESTERMANN, D. The potential impact of iata's new distribution capability (ndc) on revenue management and pricing. **Journal of Revenue Pricing Management**, v. 12, 11 2013.

WILLIAMS, K. Dynamic airline pricing and seat availability. **SSRN Electronic Journal**, 01 2017.

WITTMAN, M.; BELOBABA, P. Customized dynamic pricing of airline fare products. **Journal of Revenue and Pricing Management**, v. 17, p. 78–90, 2018.

WOHLFARTH, T. A data-mining approach to travel price forecasting. **2011 10th International Conference on Machine Learning and Applications and Workshops**, 2011. Disponível em: <<https://ieeexplore.ieee.org/document/6146948>>. Acesso em: 20 jun. 2021.

WU, C.-L.; CAVES, R. Aircraft operational costs and turnaround efficiency at airports. **Journal of Air Transport Management**, v. 6, p. 201–208, 10 2000.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO TC	2. DATA 24 de novembro de 2021	3. DOCUMENTO Nº DCTA/ITA/TC-117/2021	4. Nº DE PÁGINAS 66
5. TÍTULO E SUBTÍTULO: Uma abordagem da precificação de passagens aéreas no Brasil utilizando algoritmos de aprendizado de máquina			
6. AUTOR(ES): João Vitor Fontenele Lustosa			
7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES): Instituto Tecnológico de Aeronáutica – ITA			
8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR: Aprendizado de máquina; Mercado aéreo; Previsão de preço			
9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO: Aprendizado de máquina; Mercado aéreo; Previsão de preço			
10. APRESENTAÇÃO:		(X) Nacional () Internacional	
Trabalho de Graduação, ITA, São José dos Campos, 2021. 66 páginas.			
11. RESUMO: O preço das passagens aéreas são determinados por inúmeros fatores, como tempo de voo, concentração de mercado, atividade econômica, preço de combustível, entre outros. Cada companhia aérea utiliza algoritmos específicos de gerenciamento de receita, que leva em consideração fatores que são uma incógnita para os consumidores. Este trabalho aborda a precificação do mercado aéreo nacional, buscando prever o preço de passagens aéreas e encontrar variáveis determinantes, utilizando uma base de dados compilada pelo laboratório de transporte aéreo do ITA (LABTAR) contendo variáveis macroeconômicas e dados de voos em trechos nacionais no período entre 2010 a 2018. No desenvolvimento deste trabalho, foi possível treinar seis algoritmos de aprendizado de máquina, utilizando variáveis pré-selecionadas da base de dados. Entre os modelos, o modelo Random Forest obteve a maior precisão preditiva acima de 85%. Em números absolutos, isto representa um erro médio de R\$ 50,00 na determinação de preços que flutuam em torno de uma média de R\$ 470,00. Além disso, esse modelo indicou que os fatores de maior importância preditiva são: tempo agendado de voo, média de quilos de bagagem transportada por passageiro e tamanho da aeronave. Palavras-chave: Mercado Aéreo, Previsão de Preço, Passagens Aéreas, Aprendizado de Máquina			
12. GRAU DE SIGILO: (X) OSTENSIVO () RESERVADO () SECRETO			