UNIVERSIDADE DE MARÍLIA

PÓS GRADUAÇÃO - ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

TRABALHO DE CONCLUSÃO DE DISCIPLINA FUNDAMENTOS DE CIÊNCIA DE DADOS E INTELIGÊNCIA DE NEGÓCIOS

Larissa Freitas e Luiz Fernando

Marília - SP

2025

Introdução

A compra de uma aeronave exige uma base de conhecimentos e recursos e boa e prática, mas que é bem melhor acompanhada de uma boa consultoria, pensando nisso a empresa de consultoria para compradores de aeronaves se destaca trazendo mais confiabilidade e uma visão ampla de que será um bom negócio para prever quais os custos estimados da compra de uma aeronave.

A empresa de consultoria para compradores de aeronaves ajuda em diversos fatores, como informações técnicas de modelos de aeronaves, tendências de preços, variações, condições, trazendo mais transparência na hora da compra da aeronave.

Diante do cenário em questão a empresa pode ter uma metodologia CRISP-DM e assim ser cíclica, podendo ser aprimorada, adicionado outros casos de uso, mais dados de aeronaves, com a possibilidade de sua reutilização, já que toda a estrutura já feita para consultas obteve resultados que são possíveis para ajudar com a consultoria para uma compra mais segura e sem dúvidas, podendo ser testados ou experimentados os existentes ou novos modelos preditivos.

O dataset utilizado foi airplane_price, que segue em: <*https://www.kaggle.com/datasets/asinow/airplane-price-dataset*>, que há dados relevantes como preços, modelo e ano de fabriação, são algumas das informações apresentadas no dataset que podem oferecer suporte para a empresa de consultoria aconselhar os compradores, auxiliando na tomadas de decisão antes de efetuar uma compra.

Objetivo

Mas como a empresa de consultoria poderia ajudar o comprador a decidir quanto vai sair uma compra de aeronave de acordo com seus requisitos?

A empresa na verdade não só poderia ajudar o comprador a determinar quanto vai gastar, mas futuramente também, auxiliar na escolha da aeronave na hora da compra, mas estar seguro de qual pode ser a melhor opção considerando diversos fatores, não apenas o preço, mas a capacidade de alcance, além de outros parâmetros definidos para uma decisão final, que possibilita a melhor avaliação de qual seria um preço adequado ao seu perfil, qual seria a qualidade de uma aeronave de acordo com valor, entre outras métricas.

Justificativa

A falta da decisão adequada na aquisição de uma aeronave pode levar a decisão inadequada, prejuízos financeiros e insatisfação, pois sem uma consultoria baseada nos dados o comprador pode adquirir algo incompatível com suas necessidades ou acabar pagando por algo muito acima de suas necessidades.

Não sendo ignorados fatores essenciais como comparações, análise, entre outras questões para estudar qual seria a melhor opção, como fatores importantes por exemplo, manutenção, alcance, gastos futuros, ajuda a refinar e compreender de forma contínua a tomada de decisão.

Metodologia

Para ter garantia na fluidez do processo de consultoria usando técnicas de previsões de modelos foi escolhida a metodologia CRISP-DM, que auxilia no processo de trabalho com os dados e eficiência e assim facilita a parte técnica e de negócios.

Ao todo o framework é elaborado em seis etapas, sendo que as três primeiras tem foco em coleta de dados, organização e análise voltados para a necessidade dos negócios, enquanto que as três últimas etapas são voltadas a criação do modelo, ou seja, modelagem, avaliação e implementação, considerando as variáveis mais adequadas para os negócios.

Com o CRISP-DM a empresa segue as seis etapas, sendo para melhor compreensão de como foram executadas, explicativa a seguir:

1. Compreensão do negócio: O problema definido é que a falta de uma visão ampla sobre a decisão que está se tomando ao comprar uma aeronave é o gasto exacerbado por algo que talvez não atinja os objetivos do comprador que pode gerar frustrações imediatas ou futuras. Por isso é necessário medidas para que a execução de uma compra não gere futuras frustrações.
2. Coleta de dados: Levou-se em conta os dados coletados do dataset e conteúdos destes dados, como verificação de dados não nulos, faltantes, ou outliers e se determinados dados podem ser prejudiciais ou não na predição final do modelo.
3. Preparação dos dados: Consideração por analisar se é necessário a limpeza dos dados ou analisar se é necessário transformações, e após isso estudar as variáveis para que com o andamento da análise seja possível observar

quais eram as melhores *features* para realizar predições de acordo com a necessidade do que o cliente precisa, além de entender a seleção das melhores features.

4. Modelagem: Criação e teste de modelos para que seja possível a geração de insights e assim poder compreender qual o melhor modelo para o problema adequado.
5. Avaliação: Da performance e resultados dos modelos experimentados pode ser feita a avaliação dos resultados validando a previsão para garantir confiança e conclusões com segurança.
6. Implementação e monitoramento: Implementação no caso de uso, com exemplo de 3 clientes, utilizando o modelo de melhor performance.

Governança e gestão de dados

A fonte dos dados vieram do dataset *airplane_price,* onde foi verificado se havia colunas com valores nulos (*NaN* ou *None*).

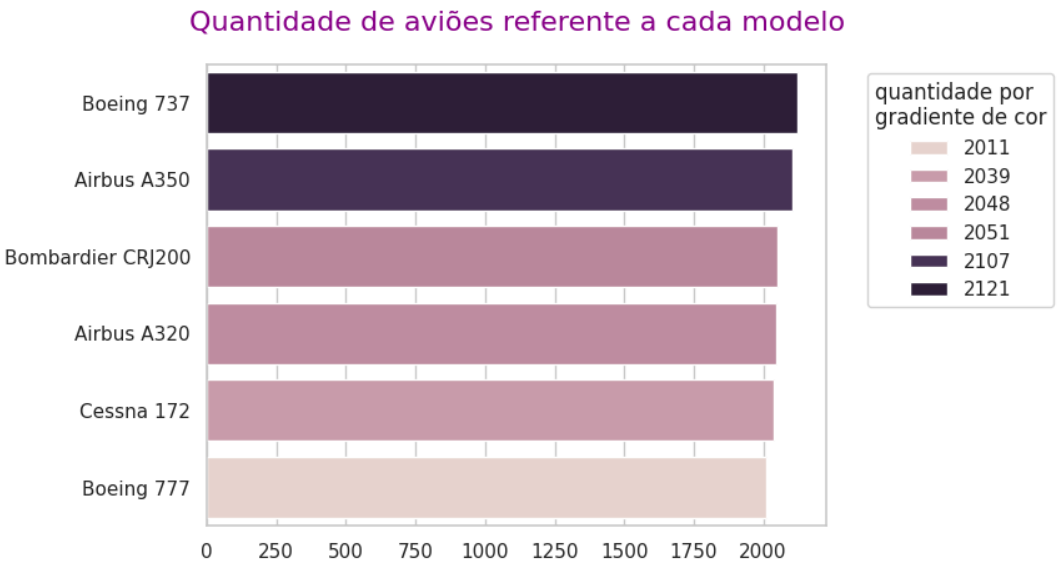
Foi verificado também a qualidade dos dados, onde para melhor compreensão da equipe técnica e de negócios foi traduzido o nome das colunas para maior entendimento, já sendo a língua da nacionalidade da equipe, para melhor compreensão na documentação.

Além da remoção da coluna de idade do avião, já praticamente seria duplicada e sem uso, por existir uma coluna com o ano de fabricação das aeronaves.

Os *outliers* encontrados são de suma importância mantê-los, os outliers devem ser compreendidos e devem ser de conhecimento da equipe o que e quais são eles, considerando os outliers apenas os valores das aeronaves mais caras, tomando como base as maiores aeronaves e realmente mais caras, comparada a aviões de pequeno porte, a empresa precisa ter consideração com essas métricas num todo, considerando que assim podem ter mais precisão na estimativa da precificação de determinado estilo de aeronave.

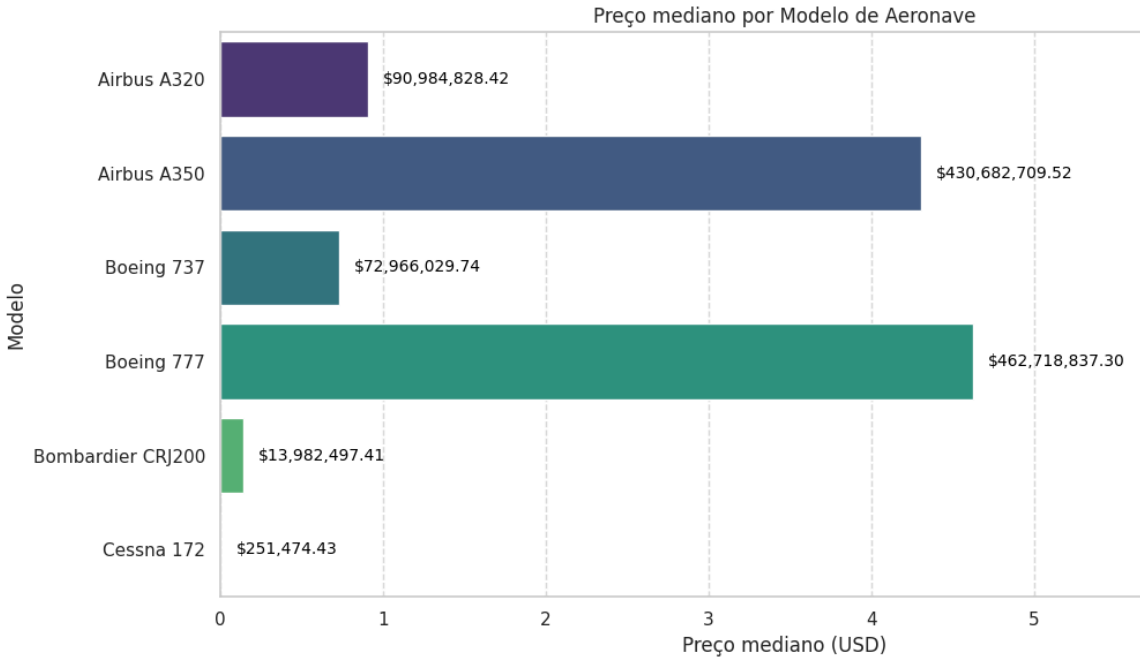
Análise exploratória

Foi analisado quantas vezes cada modelo de avião apareceu no dataset, sendo coletados os números absolutos, em seguida, após a normalização, para facilitar a visualização e entendimento dos stakeholders.



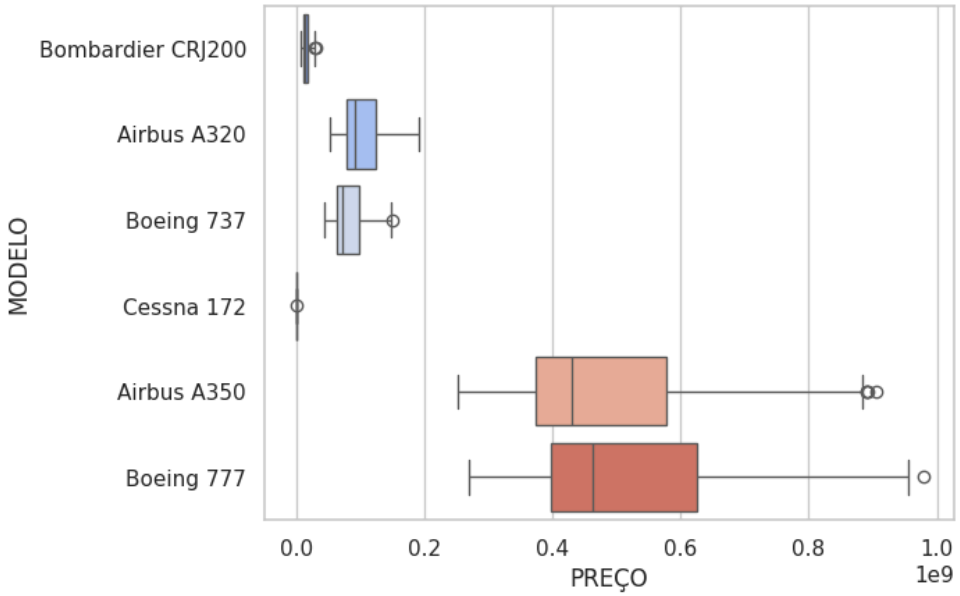
Foi verificado também a quantidade de cada tipo de motor, sendo, 2039 motores *Piston* e 10338 motores *Turbofan*, então a empresa precisou entender a diferença, quais modelos usam quais motores, onde foi descoberto que apenas *Cessna 172* usaria *Piston*, considerando que é único que não é considerado comercial, além de usar apenas 1 motor, enquanto os outros modelos tem a quantidade de 2 motores cada. Mais a frente após testar os modelos com *Cessna 172*, para variados perfis de compradores, seria interessante sua remoção para avaliar como o modelo se sairia, para o perfil de compradores de aviões comerciais e assim ver como o modelo se comportaria.

Com a análise bivariada pode ser observado as estatísticas dos preços, em tabela e em forma visual para melhor compreensão técnica, assim com base no uso principal da mediana de preços de acordo com os modelos, mostrando os valores de forma mais equilibrada.

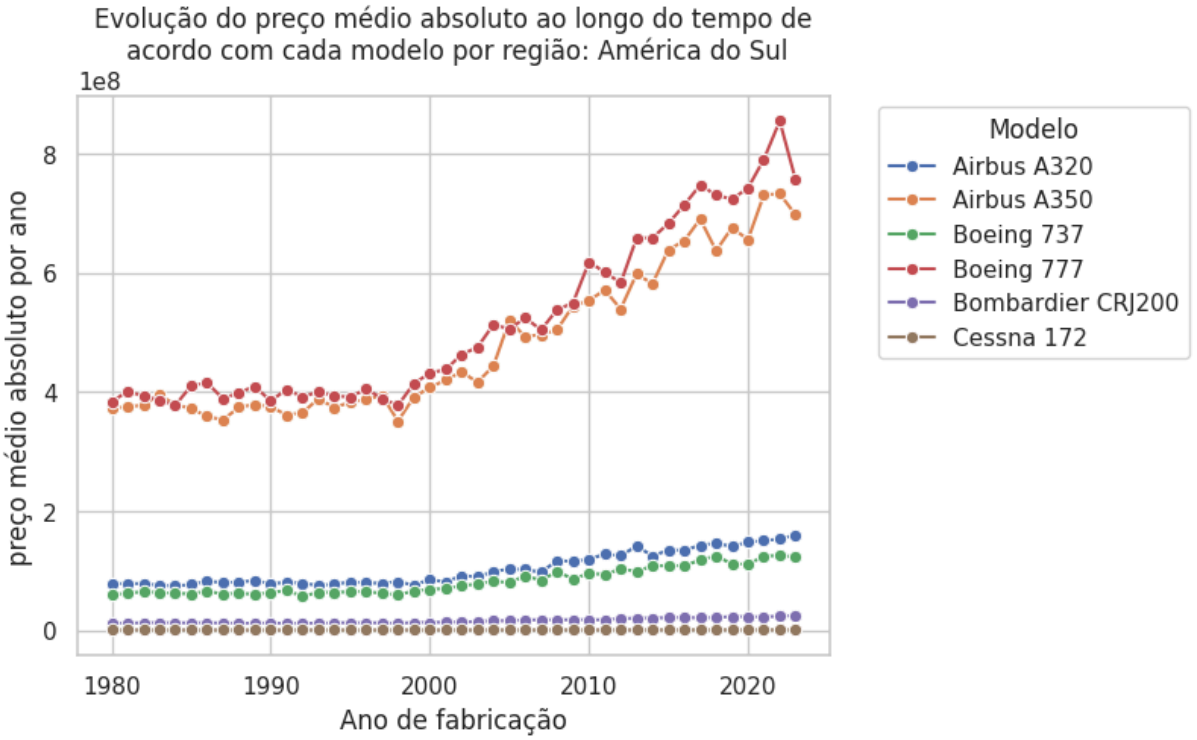


Foi importante analisar também se poderia haver variação dos preços de acordo com modelo e ano de fabricação das aeronaves e dependendo do perfil do comprador e qual sua finalidade, como maior alcance, capacidade, ou custo de manutenção por hora, assim pode-se entender melhor por meio das métricas qual a melhor opção para compradores que procuram aeronaves e que buscam custo menor de manutenção, ou qual o consumo de combustível por litro/hora de cada modelo.

Além disso foram avaliados os *boxplots* para a demonstração de *outliers* de acordo com o modelo e preços das aeronaves, demonstrando que o modelo *Airbus 320* é o mais equilibrado no quesito de demonstrar mais segurança, pois não obteve nenhum *outlier*, isso pode trazer ainda mais segurança na previsão indicando que os preços do *Airbus 320* tendem a ser mais consistentes no mercado, isso não significa que todos os compradores deveriam fazer aquisição de um *Airbus 320*, apenas demonstra que foi o preço que houve mais equilíbrio.



Foi visualizado a evolução de preços médio absoluto ao longo dos anos de cada modelo, além das vendas por região, para entender se as variações de câmbio poderiam alterar muito os preços ao longo dos anos, sendo filtrado por região, com o preço médio absoluto, segue um exemplo com a filtragem da região da América do Sul:



Além dessas análises, foram feitas outras análises julgadas pela empresa como necessárias para tomar decisões de como recorrer com os dados, e

consideração de mantimento de variáveis ou retiradas de variáveis para predições, testes e ajustes para melhoria contínua.

Análises estas que incluíram o preço relativo e absoluto de cada modelo em relação à média de preço de todas as aeronaves, visualização de tendências dos tipos de motores ao longo dos anos, visualização de *lineplots* para os dados categóricos com evolução do preço de acordo com cada modelo. Além de análise com heatmap para ver o grau de correlação entre as variáveis.

Em seguida houve a engenharia e seleção de variáveis, onde foi decidido o que seria utilizado para os modelos com Mix (para criar combinações de colunas) e Binning (para agrupar as features contínuas e categóricas e facilitar a análise) para fazer a filtragem.

Modelo de machine Learning

Foram analisados três modelos, sendo eles *LinearRegression*, *DecisionTreeRegressor* e *RandomForestRegressor*, as variáveis de preço, consumo de combustível por hora e alcance em quilometragem podem ser variáveis numéricas contínuas, enquanto que o ano de fabricação pode ser além de numérico, tanto contínuo como discreto, além de poder ser usado como categórico com o uso de *qcut*; seguindo quantidade de motores que também pode ser discreto, dependendo da análise a empresa pode usar também como categórico, enquanto que variáveis como modelo, tipo do motor ou região de venda são categóricos.

Seguindo com a ideia de que foram escolhidos três modelo, entende-se que o uso do *LinearRegression*, *RandomForestRegressor* ou *DecisionTreeRegressor,* para regressão poderia testar diferentes métricas e encontrar o que realmente a empresa busca para cada caso de uso, levando a consideração estatística de preços, exclusivo para cada cliente que é o que tornará a consultoria eficaz.

Mais eficaz ainda com o uso futuro de novos modelos, mas categóricos, onde o “y” tomaria uma variável categórica, por exemplo com a variável de modelo de aeronave do dataset e o uso de por exemplo DecisionTreeClassifier, para ver qual o melhor modelo de aeronave de acordo com as métricas o cliente poderia fazer aquisição de acordo com seus requisitos.

Até mesmo ser previsto com um modelo de regressão qual o valor previsto de compra deste cliente e em seguida qual o modelo de aeronave previsto com um modelo para dados categórico, criando uma junção ainda melhor de previsões para

tornar a compra ainda mais eficaz e sem erros equivocados, assim o cliente teria em mente qual o melhor modelo de aeronave e qual a média de preço ele gastaria na compra, tornando a consultoria uma experiência ainda mais agradável.

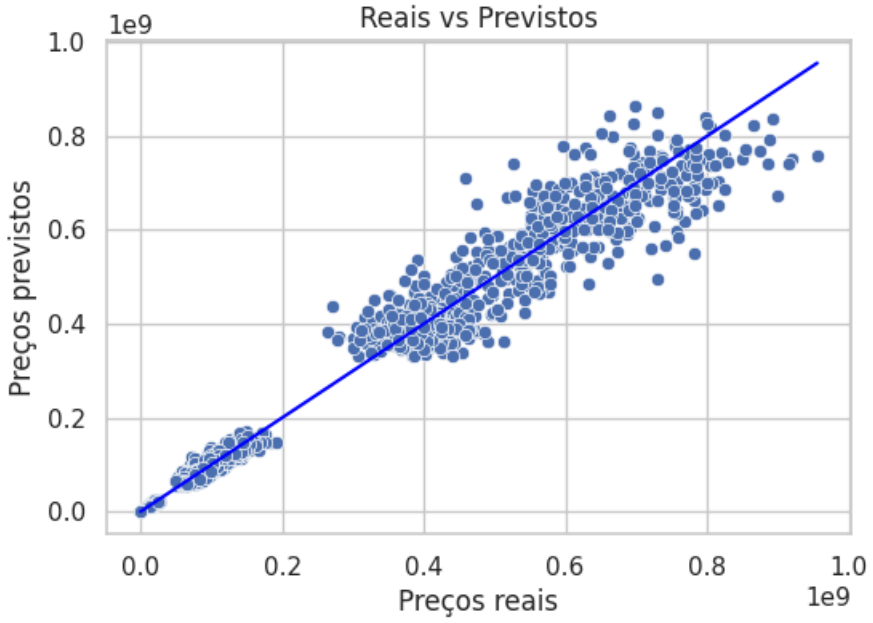
Segue abaixo os resultados obtidos de cada modelo:

⚔️ COMPARAÇÃO ENTRE OS MODELOS	MAE	MSE	R²
Regressão Linear	54146162.19785749	5211601017802536.0	0.9025020764084212
Random Forest	18296753.233630717	1248099633801526.5	0.9766507216658792
Decision Tree Regressor	22406421.397854738	1867159739298715.5	0.9650694293416618

O score dos modelos de regressão retornando o R^2, revelam que todos foram bem, com RandomForest se sobressaindo, isso de acordo com os números, mas a empresa pode avaliar bem mais a seguir estes dados.

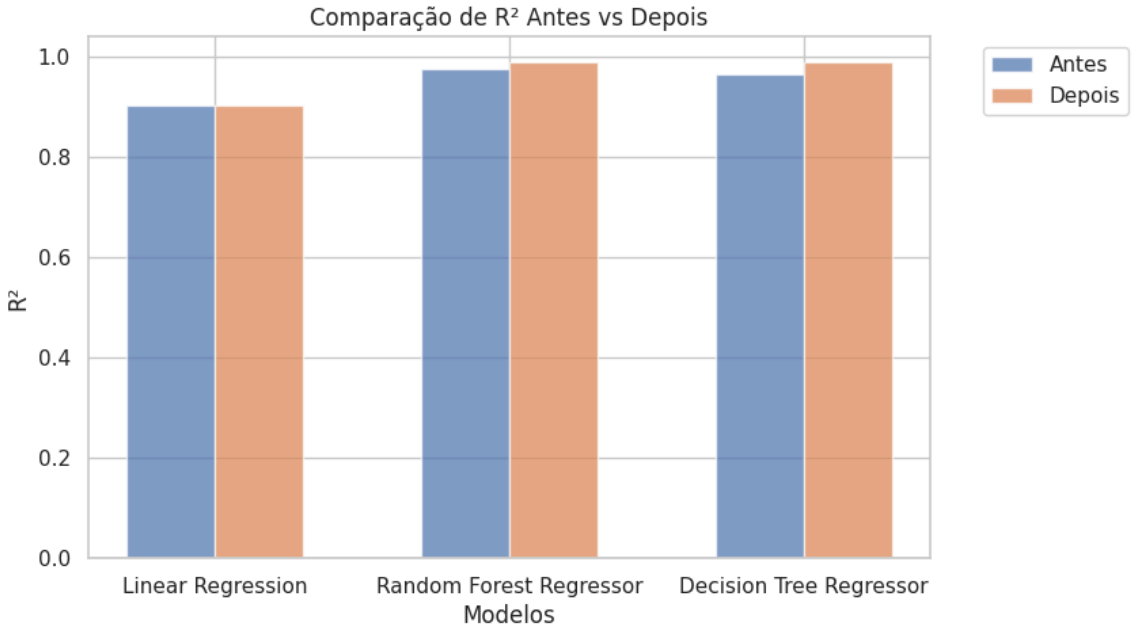
Logo, para simulação foram usados 3 clientes, com requisitos que abordaram diferentes critérios com os dados de ano de fabricação, alcance em km e consumo de combustível, sendo um deles João da Silva que obteve a previsão de gasto de 805.854.300,33 USD (mais detalhes no material da equipe técnica).

Análise das previsões do modelo *RandomForestRegressor* com gráfico de Dispersão de resíduos:



Como por exemplo fazer a limpeza do modelo *Cessna 172* que é a única aeronave utilizando o motor *Piston*, além disso a única aeronave pequena de milhares de dólares, enquanto que as outras aeronaves são de milhões de dólares, causou poucos *outliers*, ao analisar a retirada, promovendo a limpeza do modelo *Cessna 172,* além da remoção da coluna de tipo de motor, já que haviam dois tipos

de motor e *Cessna 172* era o único grupo usando o motor *Piston*, enquanto todos os outros modelos usam o motor *Turbofan*, consta diferença mínima e sua remoção não impactou diretamente nas taxas de erros, não houve diferença considerável nos outliers e houve mínimo aumento em R^2.



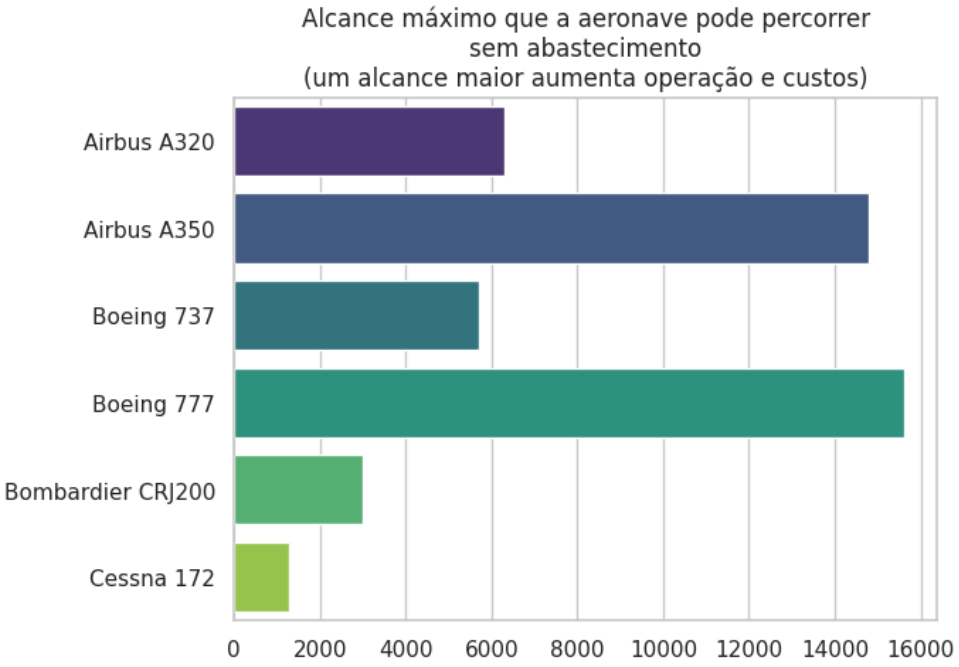
Storytelling e resultados

Análise, avaliações e consultas podem ser melhoradas constantemente, além de automatizar cada vez mais o uso de algumas funções pode tornar o sistema ainda mais cíclico, no momento presente a empresa oferece consultoria para compradores terem uma estimativa de gastos com sua compra de aeronaves de acordo com dados requeridos que resultam em uma previsão estatística com uso de modelos de regressão.

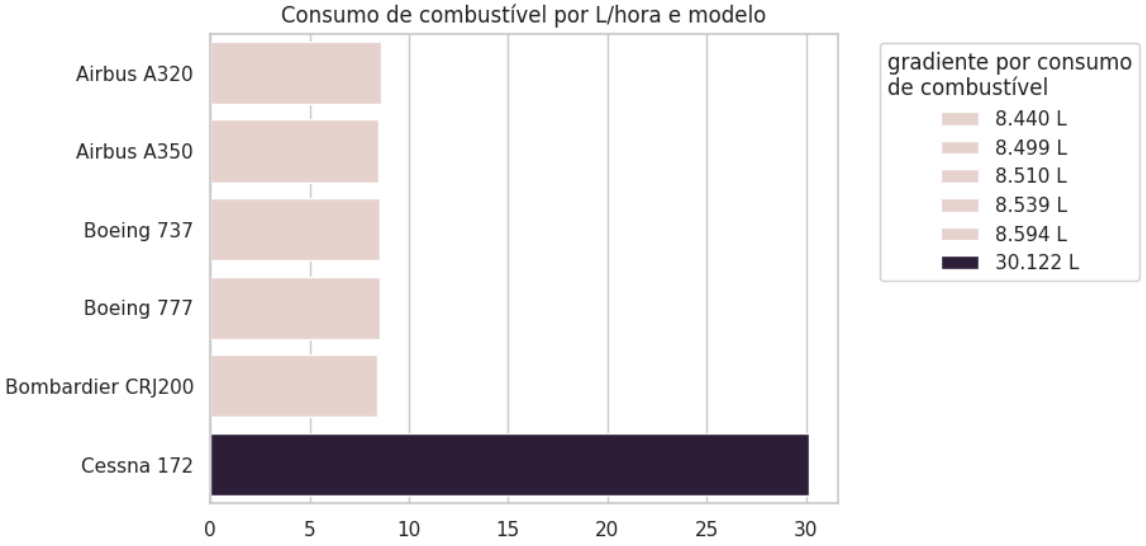
Isso pode ser aprimorado futuramente, como dito anteriormente, trazer ainda mais segurança na compra do cliente com a junção de modelos de regressão e categóricos a consultoria pode atribuir uma experiência ainda mais agradável ao poder prever qual o melhor tipo de aeronave ou outras previsões categóricas.

Para não técnicos, no caso de stakeholders, a empresa demonstra questões como características que impactam no preço, como alcance em km, consumo de combustível e ano de fabricação (como demonstrado anteriormente na análise) e preço mediano por modelo de aeronave.

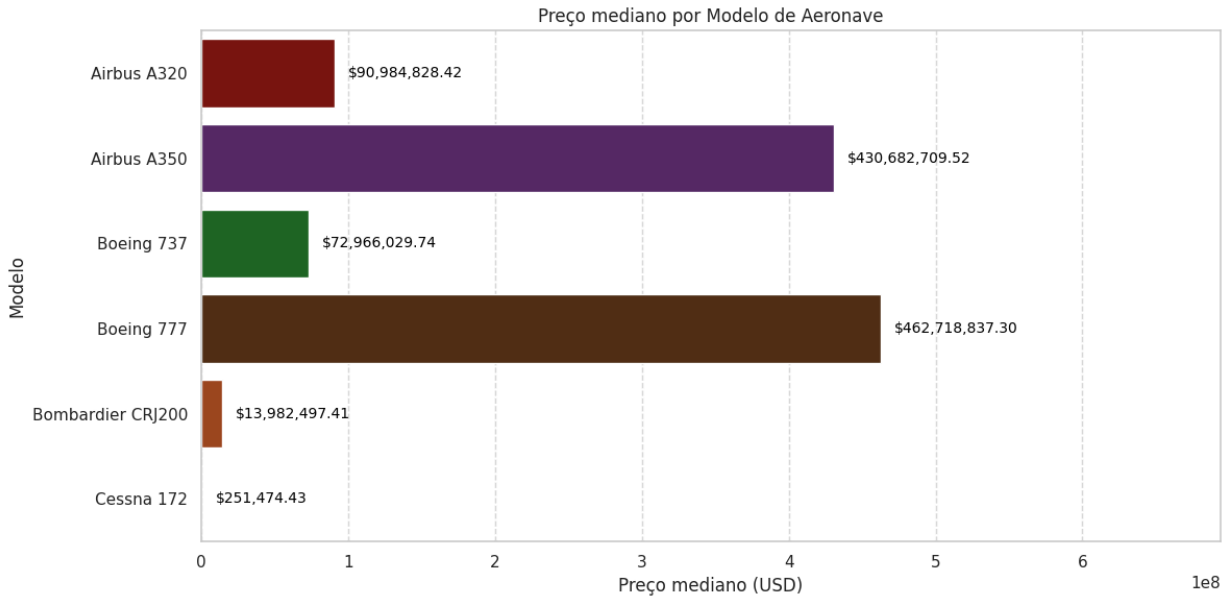
Storytelling do impacto no alcance em quilômetros:



Do consumo de combustível por litro/h:



Do preço mediano por modelo de aeronave:



Conclusão

Após leitura, análise, tratamento, criação e seleção de melhor modelo preditivo estatístico para contribuição de consultoria com clientes compradores de aeronaves destaca-se a importância da aplicabilidade prática seguindo metodologias o que reforça as soluções, além da ideia de que podem ser aprimorados passos técnicos cada vez mais completos para maior desenvoltura do sistema é mostrado o quão importante para os negócios podem ser as previsões de modelos criados a partir de governança de dados e Machine Learning.

Estes aspectos trazem vantagens nos negócios, ajudando nos problemas como, inconsistência em tomada de decisões ou falta de dados necessários para fazer bons negócios, mostra-se a justificativa da importância de que não importa o cenário é muito relevante a análise, tratamento e averiguação de dados e outliers para a confiança no uso dos dados e finalidade do uso de um modelo de Machine Learning que possa agregar para os negócios.