**ANÁLISE DE PREÇOS DE RELÓGIOS DE LUXO: PREDIÇÃO E TENDÊNCIAS COM REGRESSÃO LINEAR**

Matheus Mitsuo Yamafuku Benatti

Acadêmico do Curso de Engenharia de Software do Centro Universitário Cesumar – UNICESUMAR, Curitiba – PR.

**RESUMO:** O mercado de relógios de luxo, impulsionado por marcas icônicas e consumidores exigentes, transcende a usabilidade de apenas medição de tempo, transformando essas peças em símbolos de status e investimentos valiosos. A precificação nesse mercado complexo é influenciada por fatores como escassez, reputação da marca e exclusividade percebida. Para analisar e prever os preços desses relógios, a regressão linear surge como uma ferramenta estatística fundamental. Ela permite entender a relação entre o preço e variáveis como marca, material e ano de fabricação. Apesar de sua simplicidade e interpretabilidade, a regressão linear possui limitações, especialmente em modelos complexos com múltiplas variáveis. A não linearidade e a presença de outliers podem comprometer a precisão das previsões. Nesses casos, técnicas avançadas ou aprendizado de máquina podem ser necessários. A análise de preços de relógios de luxo com regressão linear, portanto, busca prever tendências e entender os fatores que influenciam a precificação, reconhecendo as limitações do modelo e buscando aprimoramentos para previsões mais precisas.

**PALAVRAS-CHAVE:** Relógios; Regressão Linear; Mercado de luxo.

## INTRODUÇÃO

O mercado de relógios de luxo é um dos segmentos mais sofisticados da economia global, impulsionado por marcas famosas e consumidores que veem nessas peças muito mais do que simples instrumentos para medir o tempo. Esses relógios são símbolos de status, prestígio e tradição, além de serem considerados ativos financeiros valiosos. De acordo com Köstlmeier e Röder (2023), a precificação desses produtos é influenciada por fatores como escassez, reputação da marca e percepção de exclusividade ao longo do tempo.

Considerando a complexidade envolvida nesse processo de precificação, métodos estatísticos e técnicas de aprendizado de máquina têm sido amplamente utilizados para prever os preços dos relógios de luxo. Entre esses métodos, destaca-se a utilização da regressão linear como ferramenta para identificar a relação entre variáveis como marca, material de fabricação e ano de produção com o preço final atribuído ao relógio. Segundo Maulud e Abdulazeez (2020), além de permitir a previsão de preços, essa abordagem pode, em alguns casos, ajudar a identificar relações entre variáveis independentes e dependentes.

A regressão linear é amplamente reconhecida como uma das técnicas estatísticas mais utilizadas para análise de relações entre variáveis. Essa popularidade se deve à simplicidade de aplicação e à facilidade de interpretação dos resultados obtidos. Em cenários menos complexos, essa abordagem é capaz de produzir estimativas precisas e confiáveis, sendo eficaz tanto em análises preditivas quanto em estudos exploratórios. No entanto, à medida que o número de variáveis envolvidas aumenta e os modelos se tornam mais sofisticados, tornam-se mais evidentes algumas limitações associadas à técnica. Problemas como a presença de relações não lineares entre variáveis ou a ocorrência de outliers podem comprometer significativamente a acurácia dos resultados. Nessas situações, observa-se a necessidade de adoção de técnicas mais avançadas, inclusive métodos baseados em aprendizado de máquina, que possibilitam maior capacidade de generalização e identificação de padrões mais complexos nos dados analisados (Hope, 2020).

## 1 CARACTERÍSTICAS DO MERCADO DE RELÓGIOS DE LUXO

O mercado de relógios de luxo é caracterizado por uma combinação única de tradição, inovação e status, caracterizado pela valorização simbólica, exclusividade e tradição das marcas. Como descrito por Kapferer e Bastien (2012), o segmento de luxo opera sob uma ideia de escassez controlada e construção de desejo, gerando sensação de unicidade do produto e com um valor que transcende apenas materiais e mecanismos.

Nesse contexto, marcas como Rolex, Audemars Piguet e Patek Philippe não estão baseados unicamente na excelência técnica, mas também na relação com elementos culturais, sociais e patrimoniais. Recentemente, observa-se uma mudança no consumidor, que deixa de ver relógios como um bem estético, e passa a ver esses objetos como um meio de investimento. O estudo da Deloitte (2023) revela que, os clientes de um dos países que mais começou a comprar relógios, a Índia, quando pensam em relógio como investimento, são motivados principalmente como uma proteção contra a inflação. Porém, mesmo com os clientes sendo conscientes dos preços, como a economia continua a progredir, as compras consideram principalmente a marca do relógio, seguido do design e só enfim o aspecto custo-benefício.

Os fatores que influenciam o preço de um relógio de luxo são múltiplos e muitas vezes relacionados. Entre os mais relevantes estão a edição limitada, que aumenta a escassez e o valor, e colaborações especiais, com celebridades, artistas ou casas de moda. O estado de conservação é outro fator crucial, principalmente no mercado de revenda, onde pequenos detalhes como riscos no mostrador ou a substituição de peças originais podem impactar diretamente no preço final. Relógios com todos os componentes originais, incluindo caixa, pulseira, manual e certificado de compra tendem a ser mais valorizados, pois indicam preservação do patrimônio e aumentam a confiança do comprador.

## 2 COLETA E TRATAMENTO DE DADOS

A base de dados utilizada contém mais de 280 mil registros de relógios de luxo, contendo variáveis como marca, modelo, referência, material da caixa, material da pulseira, ano de produção, tipo de movimento (quartz, automático), condição, tamanho da caixa, sexo-alvo e preço.

O tratamento dos dados consistiu nas seguintes etapas: remoção de valores ausentes ou inválidos na variável alvo (preço), transformação do preço de string para valor numérico, aplicação de logaritmo natural para estabilizar variância e reduzir a influência de *outliers*, e criação de novas variáveis, como *'watch\_age'* (idade do relógio), que considera a coluna ano de produção, removendo qualquer texto e deixando apenas o valor numérico e *'has\_gold'* (presença de ouro nos materiais), que leva em conta se o material da caixa ou da pulseira possuem ouro.

Variáveis categóricas de alta cardinalidade, como a coluna model, foram tratadas com estratégias específicas para evitar a criação de um número excessivo de variáveis dummy e reduzir o risco de *overfitting*. Nesse caso, optou-se por manter apenas os 30 modelos mais frequentes, agrupando os demais na categoria “Other”. Essa técnica ajuda a reduzir a dimensionalidade do conjunto de dados, o que melhora o desempenho do modelo e a interpretabilidade, além de prevenir que modelos pouco representativos influenciem indevidamente a predição de preços.

Nas variáveis categóricas restantes (como marca, movimento, condição, material da caixa), foi aplicada a técnica de codificação alvo (target encoding). Esse método consiste em substituir as categorias por uma estatística calculada a partir da variável dependente, neste caso, o preço do relógio. O valor mais comum utilizado é a média do target para cada categoria.

A codificação alvo (target encoding) foi empregada para lidar com variáveis categóricas de alta cardinalidade, como, por exemplo, a marca ou o modelo dos relógios. Diferentemente de abordagens como one-hot encoding, que podem inflar excessivamente a dimensionalidade do conjunto de dados ao criar colunas binárias para cada categoria única, o target encoding converte cada categoria em um valor numérico contínuo, baseado na média da variável dependente (neste caso, o preço) para aquela categoria. Por exemplo, se a marca "Rolex" apresenta, historicamente, preços médios mais elevados, ela será representada por um valor numérico proporcionalmente maior do que marcas como "Seiko" ou "Casio", que possuem preços médios mais baixos.

Essa técnica se mostra extremamente útil quando existe uma relação estatística entre a variável categórica e a variável de saída, como é comum em bases onde marcas ou modelos carregam valor de mercado. No entanto, para evitar que categorias com poucas ocorrências influenciem excessivamente o modelo, é aplicada uma suavização (smoothing). Essa suavização atua como um balanceamento entre a média da categoria específica e a média global de todo o conjunto de dados. Assim, categorias com muitas amostras se aproximam mais da média real daquela categoria, enquanto categorias raras são puxadas para mais perto da média geral. Isso reduz a variância introduzida por categorias pouco representadas, promovendo uma codificação mais robusta e generalizável.

Na prática, essa suavização pode ser feita com fórmulas que ponderam a média da categoria com a média global, controlando o peso de cada uma com base no número de registros da categoria. Dessa forma, evita-se que o modelo aprenda padrões hipotéticos, garantindo maior estabilidade preditiva e evitando overfitting em categorias raras. Essa abordagem, além de estatisticamente consistente, também contribui para manter o desempenho computacional do pipeline ao reduzir a dimensionalidade em comparação com técnicas tradicionais de codificação categórica. Essa técnica combina a informação local da categoria com o contexto geral do dataset, melhorando a generalização do modelo, especialmente em casos com dados desbalanceados.

Durante o pré-processamento das variáveis numéricas do conjunto de dados, foi adotado o uso do *RobustScaler*, uma técnica de normalização especialmente adequada para dados que contêm *outliers* ou distribuições assimétricas. Diferentemente de escalonadores como o *StandardScaler*, que centraliza os dados na média e os redimensiona com base no desvio padrão, o *RobustScaler* baseia-se em medidas estatísticas menos sensíveis a valores extremos, sendo elas a mediana e o intervalo interquartil (IQR).

Mais precisamente, o *RobustScaler* transforma cada valor subtraindo a mediana da variável e dividindo pelo intervalo interquartil, ou seja, a diferença entre o terceiro quartil, Q3, e o primeiro quartil, Q1. Essa operação resulta em dados escalonados em torno de zero, mas com uma distribuição mais resistente à distorção causada por valores incomuns, o que é extremamente quando há grande dispersão nos preços.

Esse escalonamento robusto é benéfico para modelos lineares e baseados em distância, como regressões lineares, pois garante que variáveis com diferentes escalas não dominem a função de custo ou distorçam o espaço de similaridade. Ao manter a influência dos *outliers* controlada, o modelo consegue aprender padrões mais representativos da tendência geral dos dados, sem ser muito influenciado por exceções raras que, muitas vezes, são ruídos ou casos muito específicos, como edições extremamente raras e valiosas de relógios.

Assim, o uso do *RobustScaler* contribui para a estabilidade do treinamento, a generalização do modelo e uma maior confiabilidade nas métricas de desempenho obtidas, promovendo um aprendizado mais robusto em um cenário caracterizado por alta variabilidade de preços e características técnicas diversas.

## 3 APLICAÇÃO DA REGRESSÃO LINEAR

A regressão linear foi aplicada neste estudo por se tratar de um modelo estatístico interpretável, eficiente e adequado para estabelecer relações entre variáveis independentes e uma variável dependente contínua (o preço). O uso da regressão linear se deu por conta de que o preço pode ser expresso como uma combinação linear ponderada das variáveis explicativas, como marca, material, ano de produção e tamanho. Além disso, ela permite analisar a importância de cada variável por meio dos coeficientes estimados, tornando-se uma ferramenta importante para identificar fatores que mais influenciam o valor dos relógios.

A implementação foi realizada em ambiente Python, utilizando bibliotecas modernas como Pandas, Scikit-learn, SHAP, TPOT e PyCaret, que encapsula e automatiza muitas das etapas do pipeline de modelagem preditiva. Inicialmente, os dados foram pré-processados com imputação de valores ausentes via mediana, escalonamento robusto das variáveis numéricas (*RobustScaler*), e codificação de variáveis categóricas por meio de *target encoding*.

A etapa de treinamento do modelo ocorreu dentro da classe PyCaretAdapter, desenvolvida para integrar-se ao sistema e centralizar o processo de configuração e avaliação de modelos. O método train\_model() identifica o tipo de tarefa (neste caso, regressão) e executa a função reg\_setup() do PyCaret com os seguintes parâmetros relevantes:

* normalize=True: normalização das variáveis numéricas.
* polynomial\_features=True: geração de termos polinomiais de segunda ordem.
* remove\_multicollinearity=True: remoção de colinearidade entre atributos com limiar acima de 0.95.
* use\_gpu=True: uso de aceleração por GPU.

A variável dependente definida foi o preço (price), e as variáveis independentes foram todas as demais colunas após o pré-processamento. O reg\_compare() selecionou automaticamente o melhor modelo com base em métricas de regressão (R², RMSE, MAE), e foi possível gerar gráficos e salvar métricas com funções integradas.

Os coeficientes obtidos fornecem uma interpretação direta do impacto de cada variável no preço estimado. Por exemplo, um coeficiente positivo para a variável tamanho indica que relógios maiores tendem a ter preços mais altos, enquanto um coeficiente negativo para material da pulseira sugere que esse material reduz o valor em relação à referência (como metal).

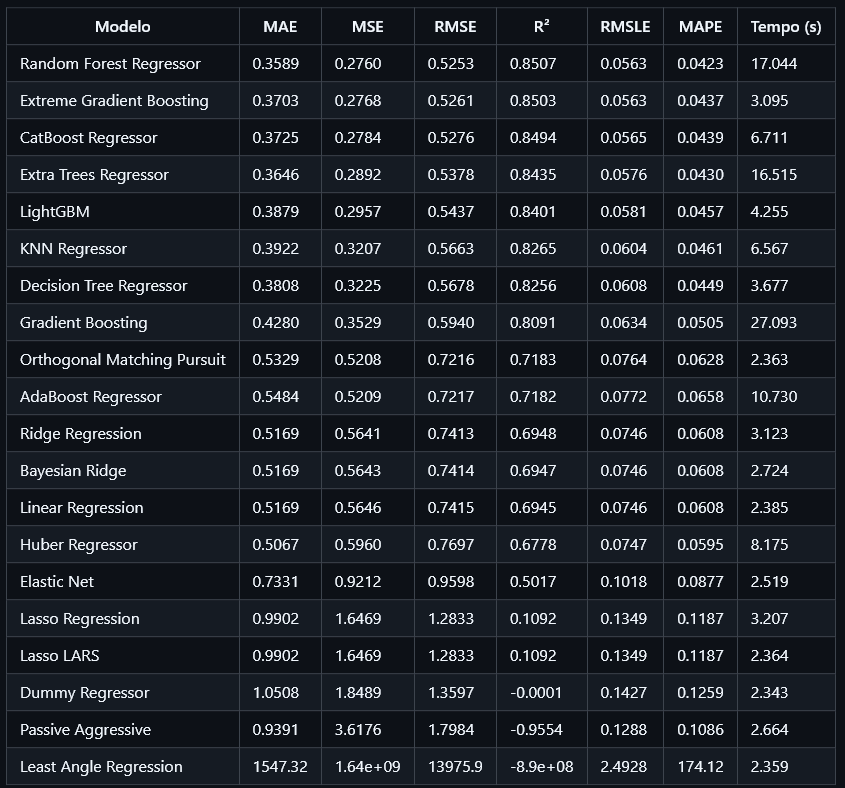
Durante a análise dos coeficientes e seus valores-p (*p-values*), observou-se que algumas variáveis apresentaram forte associação estatística com o preço, send altamente significativas. Entre elas destacam-se:

* Marca (*brand*), com destaque para Rolex, Patek Philippe e Audemars Piguet.
* Ano de produção, que se correlaciona positivamente com o preço em modelos mais novos.
* Material da caixa (*case\_material*), onde metais nobres como ouro e platina impactam fortemente no aumento do preço.
* Condição (*condition*), sendo relógios “novo sem uso” ou “com etiqueta” mais valiosos.
* Movimento (*movement*), com movimentos automáticos geralmente mais valorizados.

Por outro lado, algumas variáveis como sexo (*gender*) ou material da pulseira em categorias menos comuns mostraram coeficientes baixos e *p-values* altos, indicando fraca ou nenhuma influência significativa sobre o preço final. Ainda assim, mesmo essas variáveis podem contribuir marginalmente ao desempenho geral do modelo.

Os modelos geraram o seguinte resultado:

**Figura 1** – Desempenho dos modelos de regressão



Fonte: o autor (2025)

## 4 AVALIAÇÃO DO MODELO

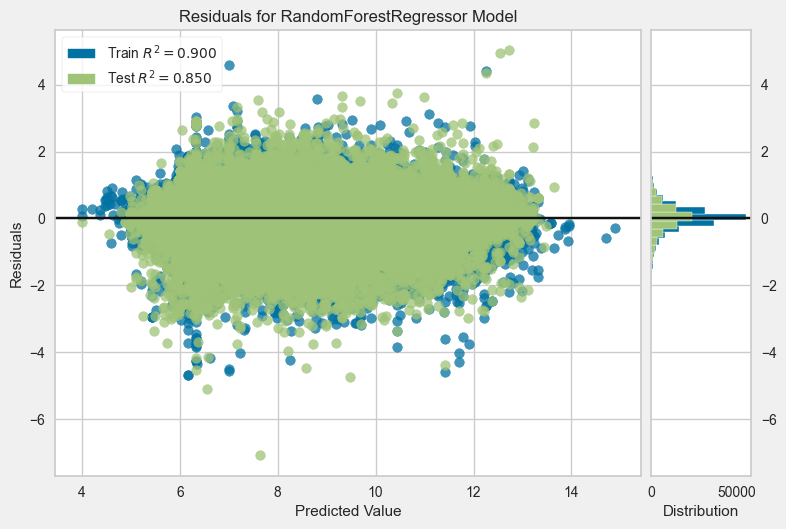
A avaliação do modelo analisa sua capacidade de generalização e precisão preditiva. Para isso, foram utilizadas diferentes métricas que capturam aspectos variados do erro e da qualidade do ajuste. Com base na figura 1, as Métricas de Avaliação consideradas foram as seguintes:

* MAE (Erro Médio Absoluto): Mede o erro médio entre os valores previstos e os reais, ignorando o sinal. É uma métrica simples e interpretável. Um MAE de 0.35, por exemplo, indica que, em média, o modelo erra 0.35 unidades do log-preço.
* MSE (Erro Quadrático Médio): Eleva ao quadrado os erros antes de calcular a média, penalizando com mais força grandes discrepâncias. Embora seja sensível a outliers, é útil para cenários onde erros grandes precisam ser fortemente penalizados.
* RMSE (Raiz do Erro Quadrático Médio): É a raiz quadrada do MSE. Por manter a unidade da variável-alvo (log-preço), é mais fácil de interpretar. Quanto menor, melhor.
* R² (Coeficiente de Determinação): Representa a proporção da variância da variável de saída explicada pelo modelo. Um R² de 0.85 indica que 85% da variabilidade nos preços dos relógios é explicada pelas variáveis independentes.
* RMSLE (Raiz do Erro Logarítmico Médio Quadrático): Mede o erro entre logaritmos dos valores reais e previstos. É útil em problemas de regressão onde se espera que os erros relativos importem mais do que os absolutos, como no caso de preços.
* MAPE (Erro Percentual Absoluto Médio): Expressa o erro médio em termos percentuais, facilitando a interpretação de quanto o modelo erra, em média, em relação ao valor verdadeiro.
* Tempo (s): Tempo necessário para treinar o modelo. Importante para avaliar a viabilidade de implementação em sistemas em tempo real ou com restrições computacionais.

Entre todos os modelos avaliados, o Random Forest Regressor se destacou com o melhor desempenho global. Ele obteve o menor MAE, MSE, RMSE e o maior R², além de um excelente RMSLE e MAPE. Seu único ponto negativo é o tempo de treinamento, que, embora não seja proibitivo, é mais alto que alternativas como o XGBoost (3.1s) ou LightGBM (4.3s).

Modelos lineares simples como Linear Regression, Ridge e Lasso apresentaram desempenho inferior (R² em torno de 0.69 ou menor), indicando que a relação entre as variáveis independentes e o preço não é completamente linear. Foram aplicadas técnicas como normalização robusta, remoção de multicolinearidade, e engenharia de atributos polinomiais, mas mesmo assim, os modelos lineares não conseguiram superar os métodos baseados em árvores.

**Figura 2** – Gráfico Residual do Random Forest Regressor



Fonte: o autor (2025)

A figura 2 representa o gráfico de resíduos do modelo Random Forest Regressor. Esse tipo de gráfico avalia a qualidade do ajuste de modelos de regressão, pois ele mostra a diferença entre os valores reais e os valores previstos pelo modelo ao longo da distribuição das previsões. No eixo horizontal estão os valores previstos (já transformados pelo log do preço), enquanto o eixo vertical mostra os resíduos, ou seja, os erros cometidos pelo modelo. Cada ponto azul corresponde a uma observação do conjunto de treino e cada ponto verde corresponde a uma do conjunto de teste.

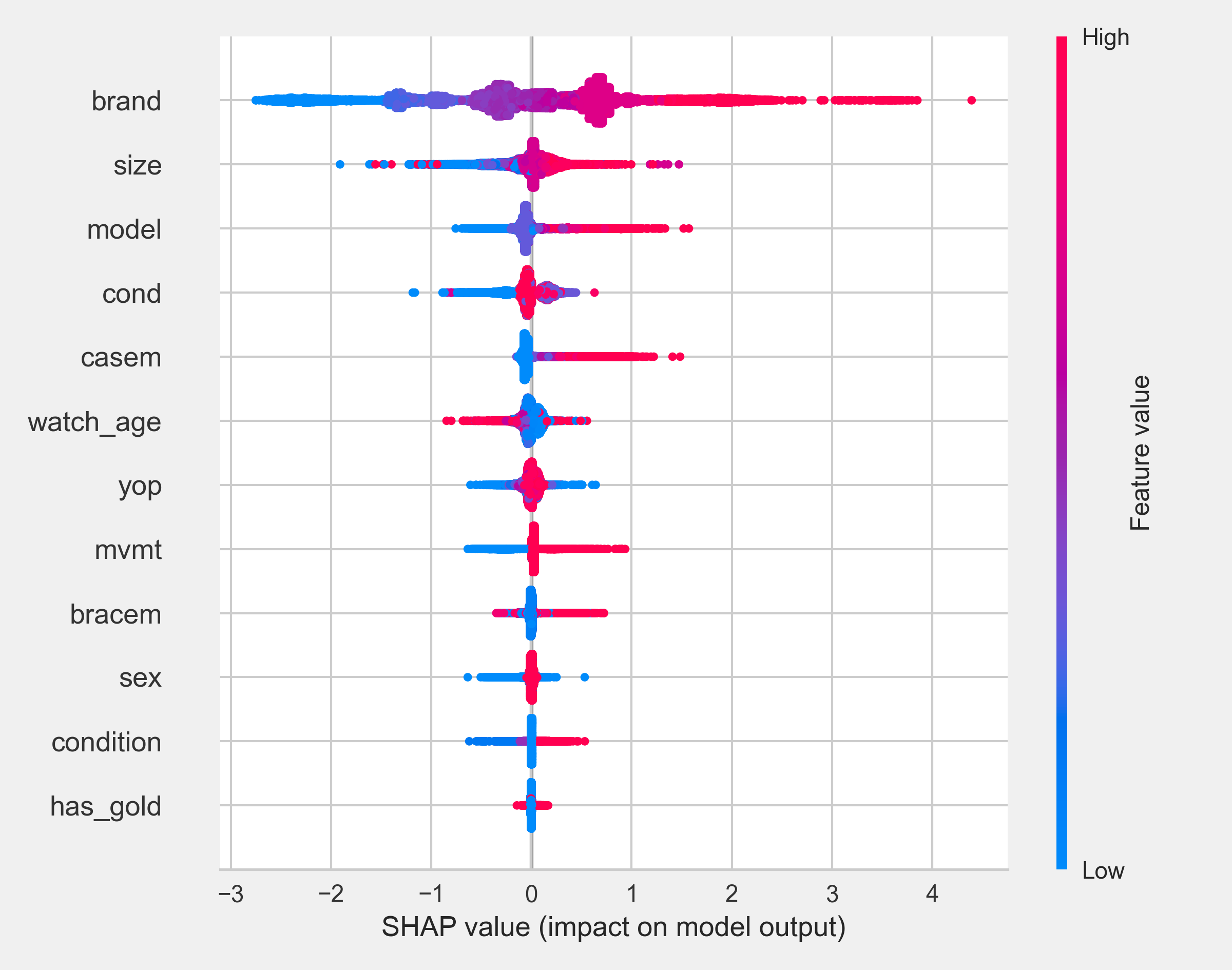
Observando a distribuição dos pontos no gráfico, nota-se que os resíduos estão concentrados ao redor de zero, o que é um indicativo positivo: significa que o modelo não está sistematicamente superestimando nem subestimando os preços dos relógios. Além disso, a dispersão dos resíduos é relativamente uniforme ao longo da faixa de valores previstos, além de que a variância dos erros é praticamente constante. Isso sugere que o modelo é consistente em sua capacidade de previsão para diferentes faixas de preço.

Outro ponto relevante é a comparação entre os desempenhos nos conjuntos de treino e de teste. O coeficiente de determinação (R²) no treino foi de 0.900, enquanto no teste foi de 0.850. Esses valores indicam que o modelo conseguiu explicar 90% da variabilidade dos preços no conjunto de treinamento e 85% no conjunto de teste, o que demonstra uma generalização bastante sólida. A diferença entre os dois valores é pequena, o que reforça a ideia de que o modelo não sofreu *overfitting*, ou seja, ele não aprendeu apenas os padrões específicos do conjunto de treino, mas foi capaz de aplicar o modelo em dados não usados antes.

À direita do gráfico, há um histograma com a distribuição dos resíduos. A forma dessa distribuição é aproximadamente simétrica e centrada em zero tanto para os dados de treino quanto de teste. Isso indica que a maioria dos erros cometidos pelo modelo é pequena e que os erros extremos são raros, o que está em linha com os resultados quantitativos apresentados anteriormente.

Em resumo, a análise visual do gráfico de resíduos complementa as métricas numéricas ao confirmar que o Random Forest Regressor apresenta um ajuste robusto e balanceado. Os resíduos distribuídos de maneira simétrica, a proximidade dos R² entre treino e teste e a ausência de padrões evidentes nos erros apontam para um modelo confiável, que pode ser utilizado com segurança para prever preços de relógios de luxo com base nas variáveis doo modelo.

**Figura 3** – Impacto de cada variável no modelo



Fonte: o autor (2025)

A figura 3 mostra as variáveis do modelo e seus impactos para o resultado final. No eixo y, estão listadas as variáveis (ou *features*) do modelo, ordenadas de cima para baixo pela sua importância média na predição. Ou seja, quanto mais alta a variável está na lista, maior foi seu impacto no resultado final do modelo ao longo de todas as amostras.

No eixo x, temos os valores SHAP, que representam o impacto da variável no resultado da predição. Valores positivos indicam que aquela característica está empurrando a previsão para cima (aumentando o preço do relógio), enquanto valores negativos indicam que a variável está puxando a previsão para baixo (reduzindo o preço).

Cada ponto do gráfico representa uma observação do conjunto de dados (ou seja, um relógio específico), e a cor indica o valor da variável naquela observação: azul para valores baixos e rosa/vermelho para valores altos.

brand (marca): É a variável mais impactante no modelo. A concentração de pontos vermelhos à direita mostra que marcas com valores altos (associadas a marcas de luxo como Rolex ou Patek Philippe, após target encoding) têm forte influência positiva no preço previsto. Já pontos azuis (marcas menos valiosas) puxam o preço para baixo.

* size (tamanho da caixa): Tem impacto relevante. Tamanhos maiores (vermelho) tendem a aumentar o preço, mas há certa dispersão — nem sempre relógios grandes são caros. Tamanhos pequenos (azul) geralmente reduzem a previsão.
* model (modelo): Também importante, embora mais disperso. Modelos com maior valor codificado aumentam o preço, mas o efeito é mais espalhado, o que pode indicar variabilidade dentro de modelos de uma mesma marca.
* cond (condição do relógio): Condições melhores (valores altos, em rosa) geralmente aumentam o preço, enquanto piores condições reduzem. Porém, esse impacto não é tão intenso quanto o de marca ou tamanho.
* casem (material da caixa): Pode ter influência positiva quando o material é mais nobre (ex: ouro), mas o efeito geral é moderado.
* watch\_age (idade do relógio): Quanto mais velho (valores altos, vermelho), mais ele tende a reduzir o preço, conforme visto pelos pontos vermelhos indo para a esquerda.
* yop (ano de produção): Correlação inversa com a idade. Valores altos (mais novos) aumentam o preço.
* mvmt (tipo de movimento): Pode impactar positivamente (ex: movimento automático de alta qualidade), mas não é o mais determinante.
* bracem (material da pulseira): Influência pequena, mas ainda observável. Materiais nobres como ouro ou titânio tendem a elevar o valor.
* sex (sexo para o qual o relógio foi projetado): Tem impacto baixo. Em geral, relógios masculinos (normalmente maiores) tendem a ter valores ligeiramente mais altos.
* condition e has\_gold: Aparecem com menor importância. A presença de ouro (has\_gold), por exemplo, tem impacto pontual, mas não é dominante.

Esse gráfico é essencial para interpretabilidade do modelo, pois mostra de forma clara quais variáveis influenciam mais o preço previsto e em que direção essa influência atua. No seu caso, ele reforça a relevância da marca e do tamanho como os principais determinantes do valor dos relógios de luxo, e ajuda a validar a lógica do modelo com conhecimento de domínio do mercado de relógios.

## 5 LIMITAÇÕES DO MODELO

Apesar do bom desempenho dos modelos de regressão aplicados, especialmente do Random Forest Regressor, é importante destacar algumas limitações próprias da modelagem e ao próprio conjunto de dados utilizado. Essas limitações podem impactar tanto a acurácia quanto a interpretação dos resultados.

Um dos primeiros desafios enfrentados foi a multicolinearidade entre variáveis. A presença de forte correlação entre atributos, como o material da caixa e o material da pulseira, ou entre marca e modelo, pode confundir algoritmos mais simples, como a regressão linear, pois dificulta a identificação do impacto isolado de cada variável. Para mitigar esse problema, foi aplicada uma etapa de remoção automática de multicolinearidade via PyCaret, mas ainda assim, pequenas correlações persistem.

Outro ponto crítico foram os outliers. Em um mercado tão volátil e influenciado por fatores subjetivos como o de relógios de luxo, é comum encontrar valores extremamente altos ou baixos que fogem à lógica dos dados gerais. Esses outliers podem distorcer o ajuste de modelos como a regressão linear clássica, afetando a inclinação da reta e tornando as previsões menos confiáveis para o restante dos dados. Embora técnicas como o uso do *RobustScaler* tenham ajudado a atenuar esse problema ao reduzir a sensibilidade a valores extremos, ele ainda persiste como uma fonte de ruído.

Além disso, alguns modelos têm dificuldade em capturar relações não lineares complexas entre os atributos e o preço. Por exemplo, o impacto do tamanho da caixa ou do ano de produção sobre o preço pode seguir padrões não monotônicos ou com pontos de inflexão que não são bem representados por uma função linear simples. Por isso, o uso de técnicas como *polynomial features* e algoritmos baseados em árvores (como Random Forest, XGBoost e CatBoost) foi fundamental para melhorar a performance geral.

Outro fator que desafia a modelagem é a dinâmica imprevisível do mercado de relógios de luxo, que não pode ser completamente explicada apenas por características físicas e técnicas do produto. Por exemplo, o preço de um modelo específico da Rolex pode disparar de uma semana para outra simplesmente porque uma celebridade foi vista usando o mesmo modelo em um evento de grande visibilidade. De forma semelhante, fenômenos como o estouro gerado em redes sociais, edições limitadas relançadas ou a especulação por parte de investidores podem inflacionar o preço de certos modelos muito além de seu valor estimado historicamente. Nenhum modelo puramente estatístico, baseado apenas em atributos objetivos, consegue capturar esses comportamentos de curto prazo, o que se traduz em erros de previsão mesmo com bons indicadores globais de desempenho.

Portanto, embora o modelo proposto ofereça um bom ponto de partida para estimativas de preços, é essencial reconhecer que ele funciona melhor para capturar padrões gerais e históricos, mas tem limitações importantes frente a comportamentos atípicos e flutuações de mercado movidas por fatores externos e subjetivos.

## 6 CONCLUSÃO

Para prever os preços de relógios de luxo com base em um conjunto de mais de 280 mil registros, abrangendo diversas marcas, modelos, anos, materiais e outras características relevantes, foram utilizados a regressão linear e modelos de aprendizado de máquina. Após um processo de pré-processamento, incluindo o uso de escalonamento robusto, codificação alvo para variáveis categóricas e remoção de multicolinearidade, foram treinados e comparados diversos algoritmos para encontrar o melhor resultado e desempenho.

Os resultados indicaram que os modelos baseados em árvores, especialmente o Random Forest Regressor, apresentaram melhor desempenho, com um R² de 0,85 no conjunto de teste e um erro médio absoluto (MAE) de 0,3589, superando outros modelos lineares. A análise dos resíduos e das métricas de avaliação confirmou a consistência das previsões e a ausência de overfitting relevante.

No entanto, o estudo também evidenciou limitações relacionadas a outliers, multicolinearidade e relações não lineares, além de destacar que fatores como tendências sociais e especulações financeiras, comuns no mercado de luxo, não são capturados por esse tipo de modelagem. Esses elementos, muitas vezes subjetivos e voláteis, exigem abordagens mais flexíveis e adaptativas, capazes de lidar com dados não estruturados e com variáveis latentes.

Como trabalhos futuros, a incorporação de variáveis contextuais externas, como dados de redes sociais, histórico de leilões e menções na mídia, bem como o uso de modelos mais avançados como redes neurais profundas que combinem diferentes fontes de informação são boas alternativas para se obter ainda mais precisão na predição. Além disso, a implementação de técnicas explicativas, pode contribuir para a interpretabilidade dos modelos, favorecendo sua adoção em contextos comerciais e estratégicos.

Em resumo, a análise realizada não apenas fornece uma ferramenta preditiva útil para o mercado de relógios de luxo, como também contribui para o entendimento de quais variáveis realmente impactam o valor desses itens. O uso da ciência de dados aplicada ao mercado de luxo abre caminho para soluções cada vez mais precisas, inteligentes e alinhadas às transformações digitais que moldam o comportamento do consumidor contemporâneo.

## REFERÊNCIAS

DELOITTE. ***The Deloitte Swiss Watch Industry Study 2023***. Deloitte AG, 2023. Disponível em: <https://www.wjinitiative2030.org/wp-content/uploads/2024/03/Deloitte_swiss-watch-industry-study-2023_EN.pdf>. Acesso em: 15 abr. 2025.

HOPE, Thomas M.H. Chapter 4 - Linear regression. In**: *MECHELLI, Andrea; VIEIRA, Sandra (ed.). Machine Learning. London: Academic Press***, 2020. p. 67–81. ISBN 9780128157398. Disponível em: https://doi.org/10.1016/B978-0-12-815739-8.00004-3. Acesso em: 29 mar. 2025.

KAPFERER, Jean-Noël; BASTIEN, Vincent. **The Luxury Strategy: Break the Rules of Marketing to Build Luxury Brands.** 2. ed. London: Kogan Page, 2012.

KÖSTLMEIER, Siegfried; RÖDER, Klaus. ***The Global Market for Luxury Watches and Asset Pricing***. abril 2023. Disponível em: <https://ssrn.com/abstract=4431680>. Acesso em: 29 mar. 2025.

MAULUD, D.; ABDULAZEEZ, A. M. A review on linear regression comprehensive in machine learning. ***Journal of Applied Science and Technology Trends***, [S. l.], v. 1, n. 2, p. 140–147, 2020. DOI: 10.38094/jastt1457. Disponível em: <https://jastt.org/index.php/jasttpath/article/view/57>. Acesso em: 30 mar. 2025.