Data Science I, Coderhouse

Relatório de Análise de Conjunto de Dados de Seguro

Autora: Olga Abramova

Introdução

No setor de seguros altamente competitivo, avaliar o risco com precisão é fundamental para definir preços de prêmios justos e sustentáveis. As seguradoras dependem de uma combinação de dados demográficos, comportamentais e financeiros para orientar suas estratégias de precificação e garantir a lucratividade, ao mesmo tempo em que mantêm a satisfação dos clientes.

O seguinte conjunto de dados sintético de seguros foi escolhido para realizar uma análise dos dados utilizando ferramentas de visualização e estatísticas, com o objetivo de obter insights sobre os relacionamentos entre diversas variáveis. Os dados foram simulados com base em dados do mundo real, para que pudessem ser analisados sem revelar informações sensíveis.

Objetivo

O objetivo desta análise é investigar os relacionamentos entre várias variáveis presentes no conjunto de dados de seguros sintéticos. Utilizando ferramentas de visualização e técnicas estatísticas, buscamos identificar padrões e obter insights relevantes sobre as características dos segurados, a gravidade dos sinistros e os valores dos prêmios, a fim de entender como esses fatores se inter-relacionam.

Este conjunto de dados será utilizado para determinar se a idade de uma pessoa tem um impacto direto na probabilidade de ocorrência de um acidente, e se isso pode ser comprovado pelos dados fornecidos. Também analisaremos como o valor do prêmio é diretamente influenciado pela frequência dos acidentes e idade. Nosso objetivo é fornecer insights para melhorar a avaliação de risco das seguradoras, permitindo que tomem decisões mais bem informadas.

Este relatório investiga duas hipóteses principais com base em um conjunto de dados sintético modelado a partir de dados reais do setor de seguros:

1. **Hipótese sobre Idade e Frequência de Sinistros**: A idade do segurado possui relevância estatística na predição da frequência de sinistros.
2. **Hipótese sobre Frequência de Sinistros e Valor do Prêmio**: A frequência de sinistros influencia o valor do prêmio cobrado pela seguradora.

Por meio de análises estatísticas e visualização de dados, este relatório tem como objetivo testar essas hipóteses e fornecer insights acionáveis sobre como esses fatores influenciam a precificação de seguros. Os resultados obtidos contribuirão para a tomada de decisões orientadas por dados nos processos de avaliação de risco e modelagem de prêmios.

Contexto Comercial

O conjunto de dados simula informações de apólices de seguro, fornecendo dados sobre os segurados, a frequência e a gravidade dos sinistros, além dos valores dos prêmios pagos. Esse tipo de análise pode ser útil para entender os comportamentos dos segurados, ajustar estratégias de precificação e melhorar o processo de tomada de decisões dentro das seguradoras. A análise dos dados também pode apoiar a definição de políticas mais eficazes de gestão de risco e segmentação de clientes.

Problema Comercial

O problema principal é como entender as variáveis que influenciam a frequência e a gravidade dos sinistros, e como essas variáveis se relacionam com o valor do prêmio. Existe a necessidade de otimizar a precificação dos seguros e identificar fatores que possam prever a gravidade dos acidentes e o comportamento dos segurados. A análise pode ajudar a melhorar a estratégia de precificação, ajustar modelos de risco e identificar possíveis áreas de redução de custos.

Contexto Analítico

1. Resumo sobre o conjunto de dados

Os dados utilizados são simulados com base em dados reais de seguros, o que permite realizar a análise sem comprometer a confidencialidade ou segurança das informações. O conjunto de dados é composto por 10.000 observações e 27 variáveis. Após a verificação inicial dos dados, observou-se que não há valores ausentes, o que facilita a análise sem a necessidade de tratamento de dados faltantes. As variáveis analisadas incluem idade, frequência de sinistros, gravidade dos sinistros e valor do prêmio, entre outras.

O conjunto de dados contém uma grande quantidade de informações sobre os segurados, distribuídas em 27 categorias (ou colunas), incluindo, mas não se limitando a: idade, estado civil, seguro anterior, frequência e severidade de sinistros, valor do prêmio, diversos descontos aplicáveis, visitas ao site, consultas, cotações solicitadas, tempo até a conversão, pontuação de crédito e distribuição dos segurados em diferentes regiões.

No entanto, o conjunto de dados não informa o período durante o qual os dados foram coletados, nem torna evidente quais unidades são utilizadas para os valores numéricos, presumindo-se que a moeda empregada seja o dólar americano (USD $).

2. Resumo do Conjunto de Dados

Tamanho: 10.000 linhas × 27 colunas

Valores Ausentes: Nenhum

Tipo de Dados: Contém informações numéricas e categóricas

Faixa Etária: De 18 a 90 anos

Idade Média: Aproximadamente 40 anos

Análise Exploratória de Dados (EDA)

❖ Análise Descritiva

O conjunto de dados compreende um total de 10.000 observações e 27 variáveis. Cada observação representa um ponto de dados único, e as variáveis incluem tanto variáveis preditoras quanto variáveis de resultado, relevantes para a análise de regressão. Foi realizada uma análise exploratória inicial para avaliar a qualidade dos dados. Essa avaliação confirmou que o conjunto de dados não contém valores ausentes nem erros detectáveis. A ausência de dados faltantes ou errôneos garante a confiabilidade dos modelos estatísticos subsequentes.

Com 10.000 observações, o tamanho da amostra é considerado suficiente para a realização de uma análise de regressão linear. Esse grande tamanho de amostra apoia a robustez do modelo e aumenta a confiabilidade dos parâmetros estimados.

Para facilitar o treinamento e a avaliação do modelo, o conjunto de dados foi dividido em dois subconjuntos:

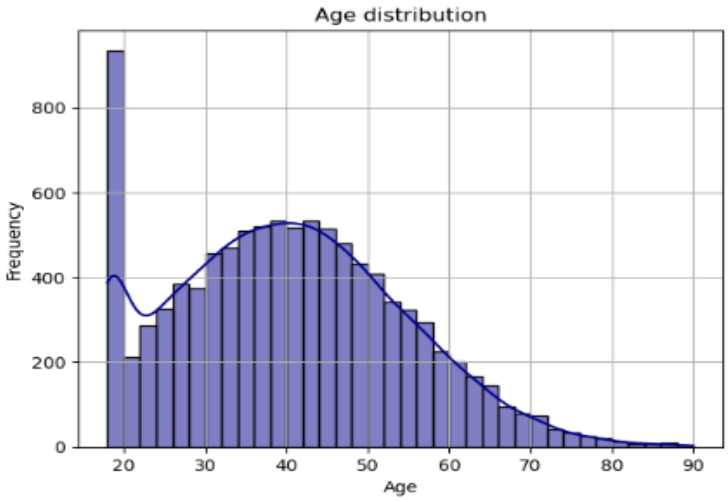
- **Conjunto de Treinamento**: 80% dos dados (8.000 observações) foi alocado para o `conjunto de treinamento, para o desenvolvimento do modelo.
- **Conjunto de Teste**: Os 20% restantes (2.000 observações) foram reservados para a avaliação e validação do modelo.

Essa divisão garante que o desempenho do modelo possa ser avaliado de forma objetiva com dados inéditos, mitigando o sobreajuste (overfitting) e aumentando a aplicabilidade do modelo.

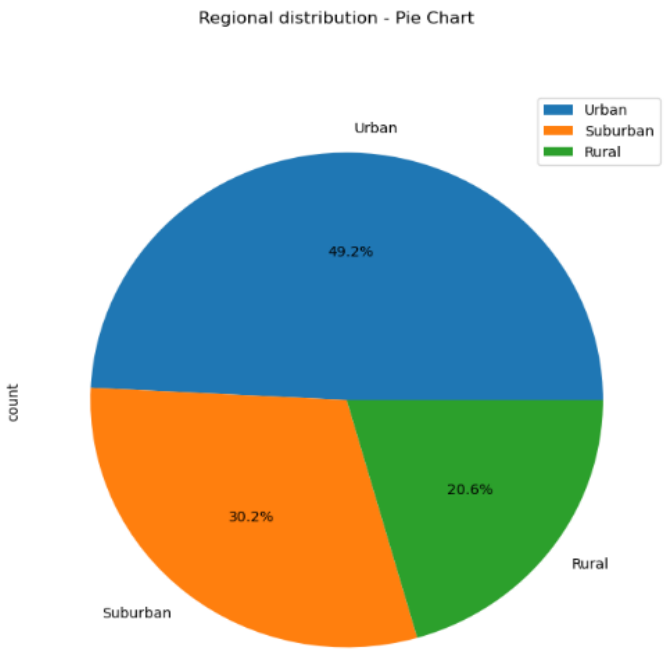
O conjunto de dados foi verificado quanto à qualidade e está adequadamente preparado para a análise de regressão linear. Os subconjuntos de treinamento e teste foram definidos de maneira apropriada, garantindo uma estrutura robusta para o desenvolvimento e avaliação subsequentes.

❖ Observações Iniciais com Resumos Gráficos:

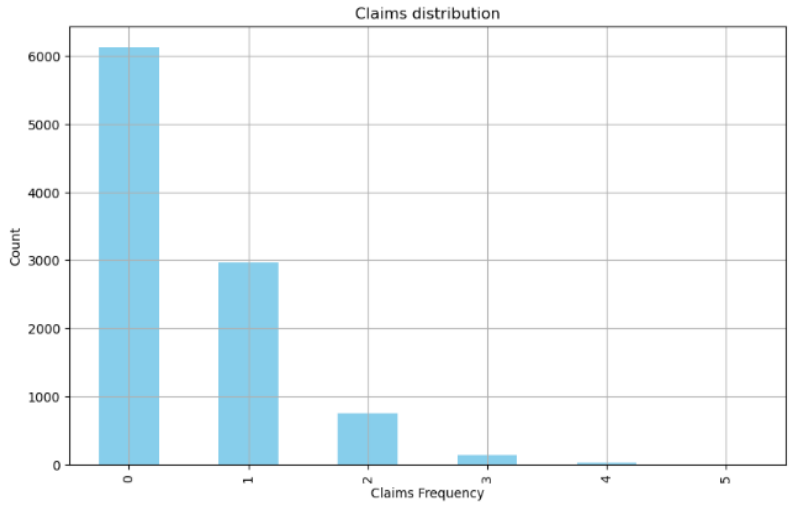
A análise exploratória dos dados revelou diversas informações interessantes. Primeiramente, observamos que a idade dos segurados varia entre 18 e 90 anos, com a média em torno de 40 anos, e uma distribuição assimétrica à direita. A maior parte dos segurados tem 18 anos, com um pico de 822 segurados nessa faixa etária.



Os segurados estão distribuídos conforme o esperado, com 49,2% provenientes de áreas urbanas, 30,2% de áreas suburbanas e 20,6% de áreas rurais.



A maioria dos segurados não registrou sinistros, enquanto apenas um número reduzido tem mais de 3 sinistros registrados.



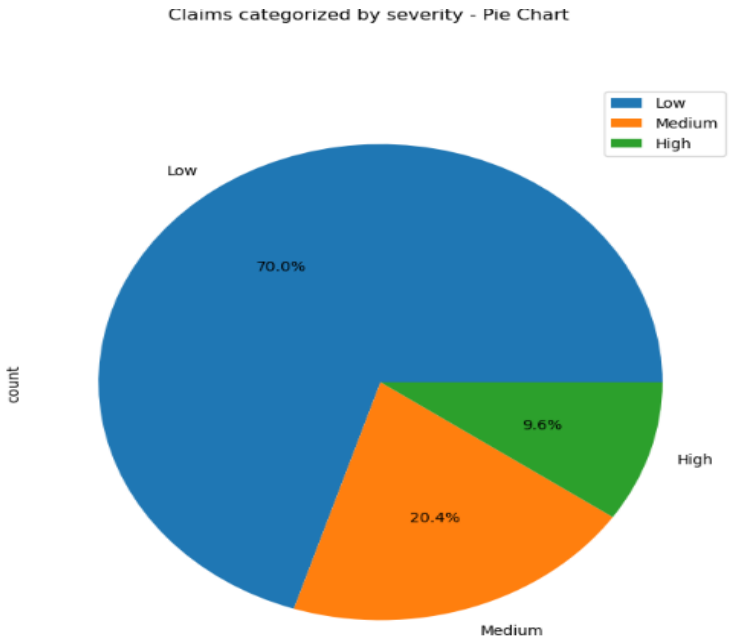
Em relação à relação entre a frequência de sinistros e a idade, observamos que os segurados mais velhos apresentam maior frequência de sinistros do que os mais jovens, contrariando a expectativa inicial. Esse comportamento sugere que, ao contrário da crença comum, os segurados mais velhos têm maior propensão a registrar sinistros.

Distribuição da Severidade dos Sinistros:

Biaxo Impacto Sinistros: 70%

Impacto Médio Sinistros: 20.4%

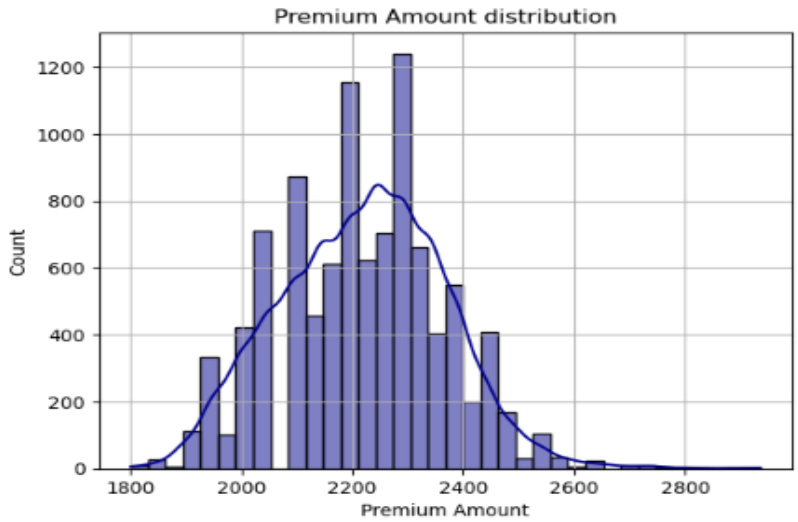
Sinistros de Alta Gravidade: 9.6%



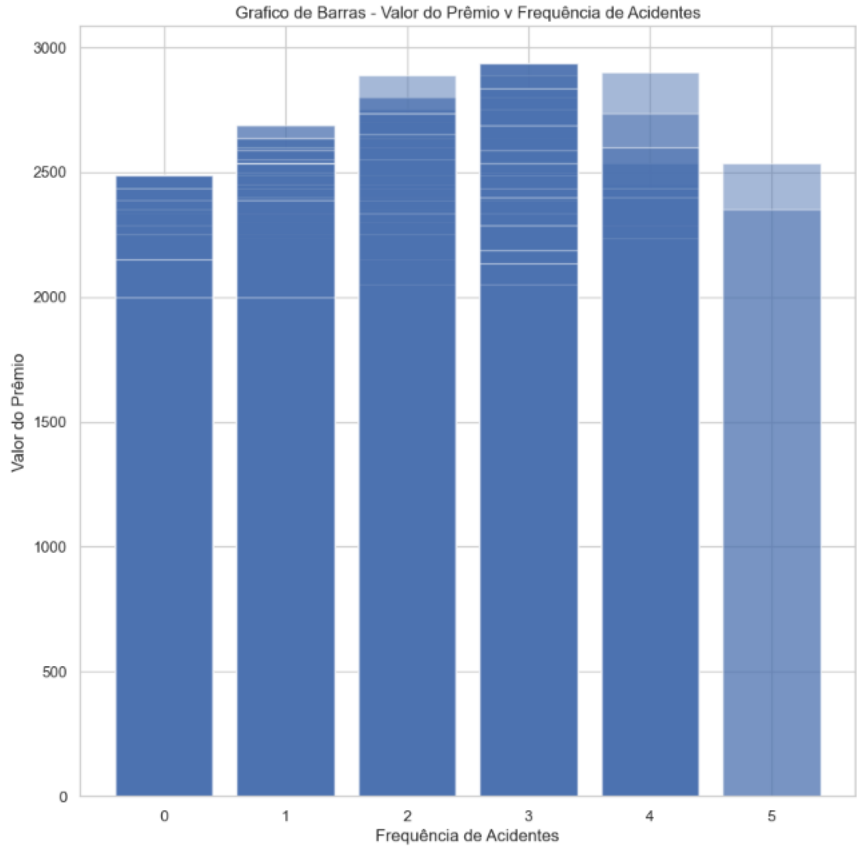
Os gráficos de setores mostram a distribuição da gravidade dos sinistros: 70% dos sinistros são de baixo impacto, 20,4% são de impacto médio, e 9,6% são de alta gravidade. A análise da relação entre a gravidade dos sinistros e a idade revelou que, à medida que a gravidade aumenta, a idade média dos segurados também tende a aumentar, embora uma análise mais aprofundada seja necessária para entender melhor essa relação.

Análise do Valor do Prêmio:

Por fim, a distribuição do valor do prêmio segue uma forma quase normal, com a maioria dos valores situados entre $2.100 e $2.400. O histograma mostra que há uma assimetria à direita, indicando a presença de outliers de valor alto.

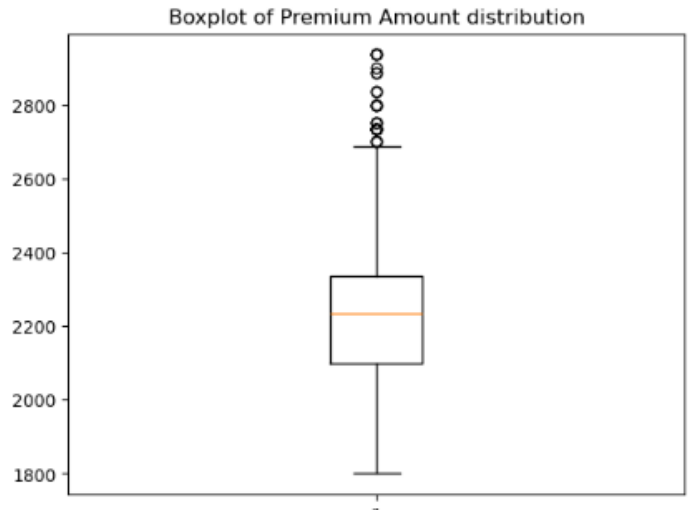


Quando analisamos a relação entre o valor do prêmio e a frequência de sinistros, verificamos que, de fato, o prêmio tende a aumentar conforme o número de sinistros registrados por um segurado aumenta.



Foi realizada uma análise exploratória dos dados por meio de um boxplot com o objetivo de examinar a distribuição da variável *prêmio*. A visualização obtida indicou que a distribuição apresenta uma leve assimetria à direita, atribuída principalmente à presença de diversos outliers com valores elevados. Esses outliers correspondem a prêmios excepcionalmente altos, que podem justificar uma investigação mais aprofundada para verificar se refletem adequadamente os fatores de risco subjacentes.

Embora a identificação desses outliers seja relevante, uma análise detalhada de sua validade e das avaliações de risco associadas extrapola o escopo do presente projeto. Ainda assim, sua ocorrência deve ser registrada como um possível ponto de análise em estudos futuros.



❖ Resumos Numéricos:

- **Resumo de 5 números (5-number summaries):**

Foi realizada uma análise estatística descritiva abrangente utilizando a função **summary()** para obter os resumos de cinco números de cada variável. Essa análise incluiu o valor **mínimo**, o **primeiro quartil (25%)**, **a mediana (50%)**, o **terceiro quartil (75%)** e o valor **máximo**, além da **média** e do **desvio padrão**.

Os principais resultados obtidos a partir da análise estatística incluem:

- **Idade dos Segurados**: As idades dos segurados variam de 18 a 90 anos, com uma idade média de aproximadamente 40 anos.
- **Valores dos Prêmios**: Os valores dos prêmios registrados variam de US$ 1.800 a US$ 2.936, com uma média de US$ 2.219,57.

Essas estatísticas descritivas fornecem insights valiosos sobre a distribuição e as tendências centrais das variáveis-chave presentes no conjunto de dados.

❖ Análise de Componentes Principais (PCA - Principal Component Analysis)

A **análise de componentes principais (PCA)** foi realizada para reduzir a dimensionalidade dos dados e identificar os principais fatores que explicam a variabilidade presente no conjunto original.

Variância Explicada pelos Componentes:

O Primeiro Componente Principal explica aproximadamente 11,75% da variância total dos dados.

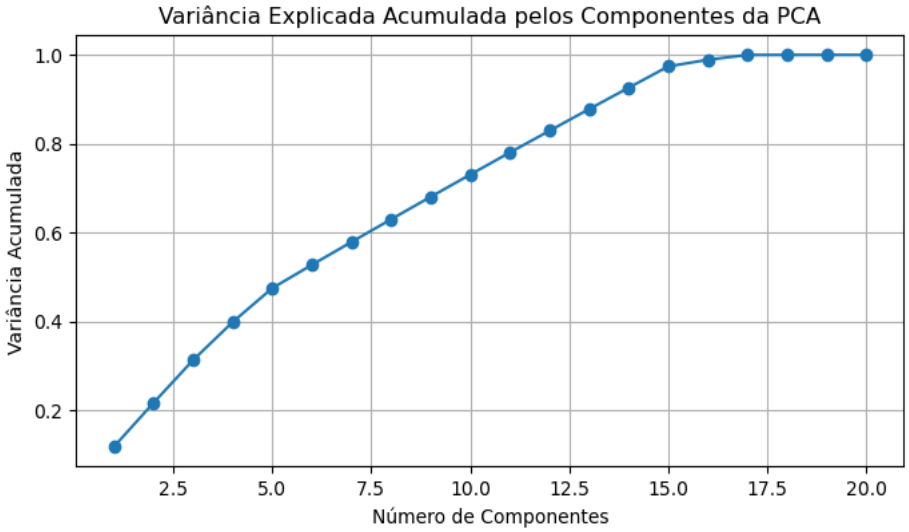
O Segundo Componente Principal explica aproximadamente 9,88% da variância total.

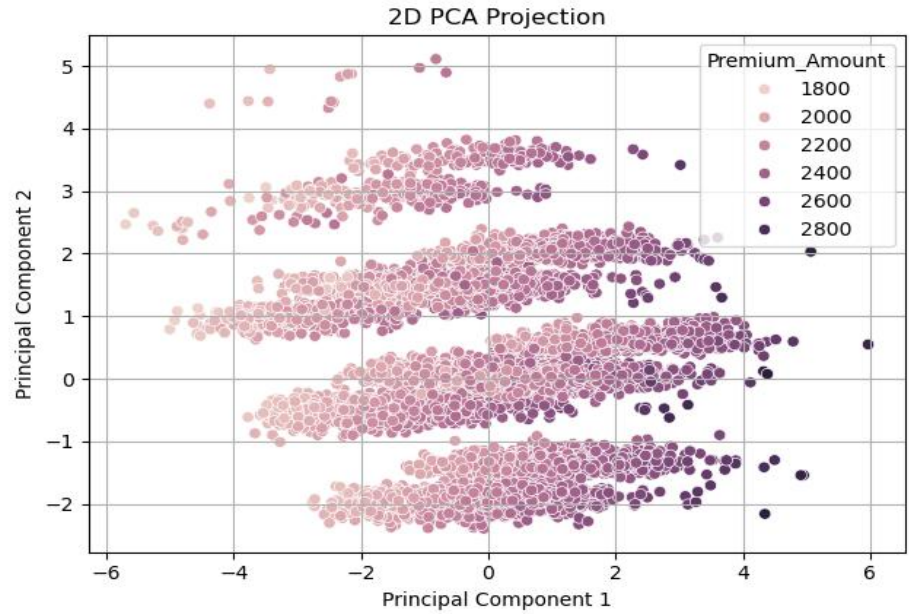
Combinados, os dois primeiros componentes preservam cerca de 21,63% da variância original dos dados.

Interpretação dos Resultados:

A preservação de apenas 21,63% da variância indica que, ao reduzir os dados para duas dimensões, cerca de 78,37% da informação original é perdida. Portanto, a projeção bidimensional não é suficiente para representar adequadamente a estrutura complexa dos dados originais.

Apesar da correta aplicação do PCA, a utilização dos dois primeiros componentes não proporciona uma explicação satisfatória da variabilidade total. Recomenda-se a inclusão de um número maior de componentes principais para assegurar uma maior retenção da informação e, consequentemente, uma melhor representatividade do modelo dimensional.





<u>A matrix de *loadings de PCA:*</u>

A **matriz de loadings de PCA** revela a contribuição de cada variável original para os componentes principais extraídos.

Valores absolutos mais altos indicam maior influência de uma variável sobre um componente específico.

A interpretação detalhada dos cinco primeiros componentes principais:

Componente Principal 1 (PC1) – Custo do Prêmio versus Descontos/Conversão

Variáveis com maior contribuição:

Premium_Amount (0.542)

Policy_Adjustment (0.295)

Claims_Adjustment (0.184)

Premium_Adjustment_Credit (0.282)

Total_Discounts (-0.353)

Conversion_Status (-0.269)

Interpretação:

Este componente está fortemente associado ao valor do prêmio. Valores mais elevados do prêmio, bem como ajustes relacionados a políticas e sinistros, aumentam esse componente. Por outro lado, altos descontos e maiores taxas de conversão reduzem seu valor. Isso sugere um eixo de troca entre maior custo do prêmio e incentivos como descontos ou facilidade de conversão.

Componente Principal 2 (PC2) – Descontos e Conversão

Variáveis com maior contribuição:

Total_Discounts (0.493)

Multi_Policy_Discount (0.325)

Safe_Driver_Discount (0.303)

Bundling_Discount (0.207)

Conversion_Status (-0.478)

Time_to_Conversion (0.479)

Interpretação:

Reflete um eixo centrado em estratégias promocionais. Clientes com mais descontos e maior tempo até conversão tendem a ter valores mais altos nesse componente, enquanto aqueles com maior probabilidade de conversão contribuem negativamente. Isso sugere que o excesso de incentivos pode estar associado a uma menor eficiência no processo de conversão.

Componente Principal 3 (PC3) – Risco Financeiro e Ajustes de Crédito

Variáveis com maior contribuição:

Premium_Adjustment_Credit (0.450)

Credit_Score (-0.450)

Premium_Amount (0.224)

Total_Discounts (0.270)

Interpretação:

Esse componente representa um eixo de risco financeiro. Altos ajustes de prêmio baseados em crédito combinados com baixas pontuações de crédito elevam o componente, indicando maior risco. Isso aponta para perfis de clientes que exigem compensações de preço devido ao seu maior risco de inadimplência.

Componente Principal 4 (PC4) – Segmentação por Faixa Etária

Variáveis com maior contribuição:

Age (0.691)

Is_Senior (0.683)

Interpretação:

Este componente separa nitidamente os segurados com base na idade. Com forte peso das variáveis relacionadas à idade e status de idoso, é possível distinguir claramente entre clientes mais jovens e mais velhos, com pouca influência de outras variáveis.

Componente Principal 5 (PC5) – Ação da Política versus Solvência de Crédito

Variáveis com maior contribuição:

Policy_Adjustment (0.436)

Claims_Adjustment (0.305)

Premium_Amount (0.368)

Premium_Adjustment_Credit (-0.415)

Credit_Score (0.433)

Interpretação:

Este componente aborda o comportamento das políticas de seguro e a confiabilidade financeira dos clientes. Ele sugere que os clientes com mais ajustes em políticas e sinistros, mas com menores ajustes de prêmio e maior pontuação de crédito, formam um grupo distinto — capturando a tensão entre necessidade de ação corretiva e estabilidade financeira.

Conclusão:

A análise de loadings proporcionou insights valiosos sobre os principais eixos latentes que estruturam os dados.

Regressão Linear com PCA:

A modelagem de Regressão Linear com PCA foi testada para avaliar o desempenho preditivo do modelo de regressão linear aplicado à variável alvo Premium_Amount, a partir de dados transformados por Análise de Componentes Principais (PCA).

Métricas de Avaliação:

Coeficiente de Determinação (R^2):

Valor obtido: 0,989

O R^2 representa a proporção da variância da variável dependente que é explicada pelo modelo. Neste caso, o valor de 0,989 indica que 98,9% da variabilidade nos valores de Premium_Amount é explicada pelas variáveis preditoras incluídas no modelo. Trata-se de um valor extremamente elevado, sugerindo um excelente ajuste do modelo aos dados, com forte capacidade explicativa.

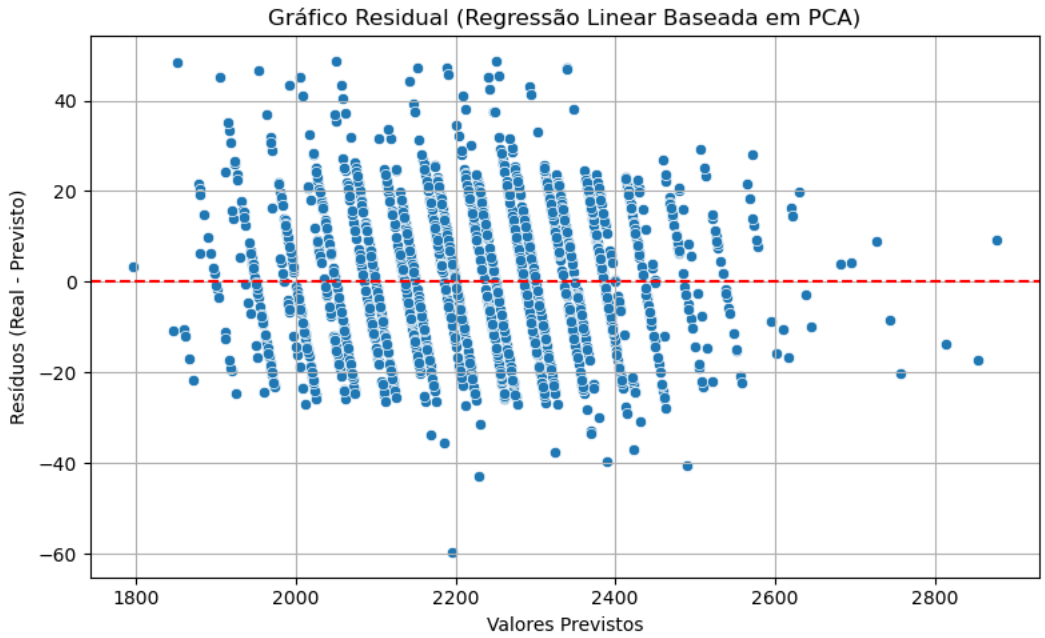
Raiz do Erro Quadrático Médio (RMSE):

Valor obtido: 15,697

O RMSE mede o erro médio da previsão do modelo em unidades da variável dependente. Um valor de 15,697 indica que, em média, o modelo apresenta uma diferença de aproximadamente 15,7 unidades em relação aos valores reais de Premium_Amount. Quanto menor o RMSE, melhor o desempenho do modelo. O resultado obtido é considerado satisfatório, reforçando a boa precisão preditiva do modelo.

Conclusão

Com base nas métricas avaliadas, conclui-se que o modelo de regressão linear apresenta excelente capacidade de explicação (R^2 = 0,989) e erro médio relativamente baixo (RMSE = 15,697). Esses resultados indicam que o modelo está bem ajustado aos dados e pode ser considerado confiável para realizar previsões de Premium_Amount no contexto analisado.



<u>O modelo usando as características originais (sem PCA):</u>

A modelagem de Regressão Linear sem PCA foi testada para avaliar o desempenho preditivo do modelo de regressão linear aplicado à variável alvo Premium_Amount, sem Análise de Componentes Principais (PCA).

<u>Métricas de Avaliação:</u>

Coeficiente de Determinação (R^2):

Valor obtido: 1.000

O valor de R^2 igual a 1.000 indica que 100% da variância da variável alvo (Premium_Amount) é explicada pelo modelo. Essa pontuação representa um ajuste perfeito, no qual as previsões realizadas pelo modelo coincidem exatamente com os valores reais.

Raiz do Erro Quadrático Médio (RMSE):

Valor obtido: 0.000

O RMSE quantifica o erro médio entre os valores previstos e os valores reais. O valor zero indica que não há diferença entre as previsões e os valores observados, caracterizando ausência total de erro.

Análise Crítica:

Embora as métricas indiquem desempenho ideal, esses resultados sugerem fortemente a ocorrência de sobreajuste (overfitting). O sobreajuste ocorre quando o modelo aprende em excesso os dados de treinamento, inclusive seus ruídos e flutuações específicas, o que pode comprometer a capacidade de generalização para novos dados.

Modelos sobreajustados tendem a apresentar resultados irreais e excessivamente otimistas nos conjuntos utilizados para treino/teste, mas podem ter desempenho significativamente inferior em dados externos ou reais não vistos anteriormente.

Conclusão:

Os resultados obtidos – R^2 = 1.000 e RMSE = 0.000 – apontam para um modelo com ajuste perfeito, mas que, muito provavelmente, não generaliza adequadamente fora do conjunto analisado.

Comparar Ambos os Modelos:

Após comparar o desempenho de dois modelos de regressão linear, sendo um construído com a aplicação de Análise de Componentes Principais (PCA) e outro utilizando todas as variáveis originais sem redução de dimensionalidade, os seguintes resultados foram definidos:

Modelo com PCA:

R^2 (Coeficiente de Determinação): 0,989

O modelo explica 98,9% da variância da variável alvo (Premium_Amount). Trata-se de um resultado excelente, demonstrando que o modelo é capaz de realizar previsões com elevado grau de precisão, mesmo após a redução do número de variáveis.

RMSE (Raiz do Erro Quadrático Médio): 15,697

O erro médio entre os valores previstos e reais é de aproximadamente 15,7 unidades. Essa métrica indica a magnitude média do erro cometido pelo modelo nas previsões.

Modelo sem PCA:

R^2 (Coeficiente de Determinação): 1,000

O modelo explica 100% da variância da variável alvo, sugerindo um ajuste perfeito aos dados utilizados. Embora esse resultado possa parecer ideal, levanta indícios de sobreajuste (overfitting), indicando que o modelo pode ter memorizado os dados em vez de aprender padrões generalizáveis.

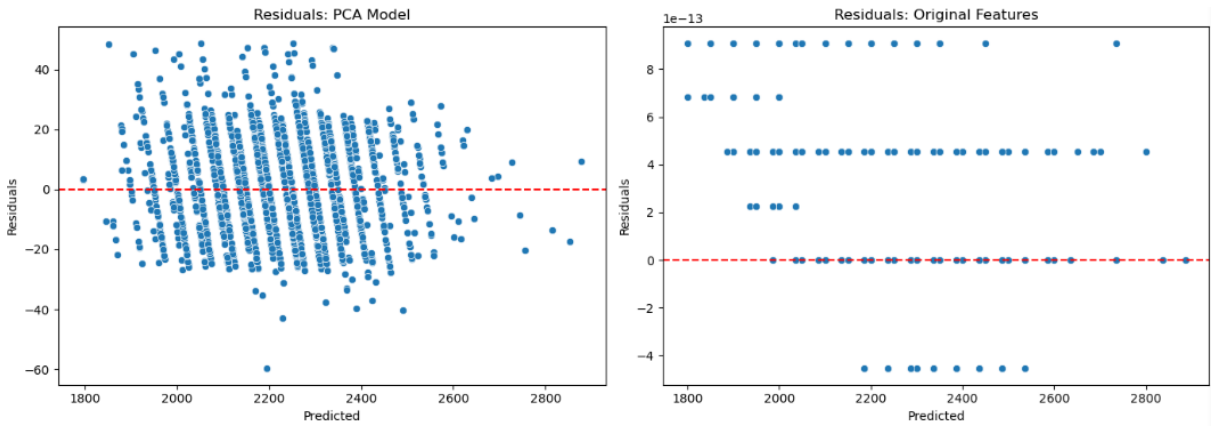
RMSE (Raiz do Erro Quadrático Médio): 0,000

O erro médio é exatamente zero, o que reforça a hipótese de sobreajuste, pois o modelo não apresenta nenhuma discrepância entre os valores previstos e reais nos dados fornecidos.

O modelo com PCA apresenta desempenho robusto, conciliando alta capacidade explicativa com menor risco de sobreajuste, o que favorece sua capacidade de generalização para novos dados. Já o modelo sem PCA, embora apresente métricas perfeitas, possui fortes indícios de sobreajuste, o que compromete sua confiabilidade para aplicações em contextos reais com dados não vistos.

Conclusão

Ambos os modelos demonstram desempenho elevado. No entanto, recomenda-se cautela com o modelo sem PCA devido aos sinais claros de sobreajuste. O modelo com PCA se mostra uma alternativa mais confiável e estável, especialmente em contextos que exigem capacidade de generalização e robustez analítica.



❖ Resumo da Análise de Correlação:

Foi realizada a análise da matriz de correlação de Pearson entre todas as variáveis numéricas do conjunto de dados. Essa matriz permite identificar a força e direção das relações lineares entre pares de variáveis.

Interpretação da Matriz:

Cada célula da matriz apresenta um coeficiente de correlação que varia entre:

+1.000: correlação linear positiva perfeita (ambas as variáveis aumentam juntas);

0.000: ausência de correlação linear;

–1.000: correlação linear negativa perfeita (quando uma variável aumenta, a outra diminui).

Principais Relações Identificadas:

Correlações Fortes (|r| > 0.5)

Safe_Driver_Discount & Total_Discounts: 0.5868

- Relação positiva forte: motoristas considerados seguros recebem mais descontos.

Multi_Policy_Discount & Total_Discounts: 0.6768

- Descontos por múltiplas apólices estão fortemente associados ao total de descontos aplicados.

Policy_Adjustment & Premium_Amount: 0.6634

- Ajustes de apólice impactam diretamente o valor do prêmio.

Claims_Adjustment & Premium_Amount: 0.4391

- Correlação positiva moderada: maior número de ajustes por sinistros tende a aumentar o valor do prêmio.

Premium_Adjustment_Credit & Credit_Score: –0.7878

- Correlação negativa forte: pontuações de crédito mais altas estão associadas a menores ajustes no prêmio.

Time_to_Conversion & Conversion_Status: –0.9978

- Correlação negativa quase perfeita. Pode indicar sobreposição conceitual ou problemas de codificação (por exemplo, se Conversion_Status = 1 representa conversão imediata).

Correlações Fracas ou Nulas (|r| < 0.1)

Age & Premium_Amount: –0.0295

- Relação praticamente inexistente entre idade e valor do prêmio.

Website_Visits & Conversion_Status: 0.0253

- Visitas ao site têm baixa capacidade preditiva de conversão.

Inquiries & Quotes_Requested: 0.0034

- Quase nenhuma correlação; pode indicar falhas ou ruído no processo de coleta de dados.

Conclusão

A matriz de correlação oferece uma visão estratégica sobre quais variáveis influenciam significativamente os resultados principais do negócio, como o valor do prêmio e o comportamento de conversão. A análise permite também identificar possíveis inconsistências no conjunto de dados, o que é fundamental para a construção de modelos preditivos mais robustos.

❖ Escopo da Análise

Este relatório tem como foco a análise de duas relações fundamentais presentes no conjunto de dados:

- A relação entre a idade dos segurados e a frequência de sinistros;
- Impacto da frequência de sinistros sobre o valor do prêmio de seguro.

O objetivo dessas análises é oferecer insights sobre a possível influência da idade na ocorrência de sinistros, bem como entender de que forma o comportamento histórico de sinistros afeta a precificação dos seguros.

❖ Construção do Modelo

Um modelo de regressão linear foi selecionado para examinar quantitativamente as relações entre as variáveis escolhidas, fornecendo uma estrutura estatisticamente robusta para a análise.

Para garantir a aplicabilidade do modelo e minimizar o risco de sobreajuste (overfitting), o conjunto de dados foi particionado em dois subconjuntos:

80% dos dados (8.000 observações) foram utilizados para o treinamento do modelo;

Os 20% restantes (2.000 observações) foram reservados para teste e validação do desempenho preditivo.

Essa abordagem permite uma avaliação confiável da eficácia do modelo em dados não vistos anteriormente. to quantitatively examine the relationships between the chosen variables, providing a statistically robust framework for analysis.

Hipótese 1 (Ho): A **Idade** do segurado possui relevância estatística na predição da **Frequência de Sinistros**

Ao ajustar o modelo com as variáveis Idade e Frequência de Sinistros para a **Hipótese 1 (H_0): impacto da Idade na Frequência de Sinistros**, o modelo apresentou os seguintes resultados principais:

- **R-quadrado:**
  Valor de R^2 foi 0.000, indicando que o modelo não explica praticamente nenhuma da variabilidade observada na frequência de sinistros. Portanto, a variável Idade não é um preditor relevante para a frequência de sinistros neste conjunto de dados.
- **Idade**: Indica uma leve diminuição da frequência de sinistros com o aumento da idade. No entanto, o p-valor elevado (0.570) mostra que essa relação não é estatisticamente significativa.
- O modelo como um todo não é estatisticamente significativo, reforçando que a variável **Idade** não contribui para a explicação da frequência de sinistros.

Conclusão

A análise de regressão linear não identificou uma relação estatisticamente significativa entre a **Idade** dos segurados e a **Frequência de Sinistros**. Os resultados indicam que a variável **Idade**, isoladamente, não é um bom preditor da **Frequência de Sinistros**, e que modelos alternativos podem ser mais adequados para essa análise.

- **<u>Hipótese 2 (Ho):</u>** A **Frequência de Sinistros** influencia o **Valor do Prêmio** cobrado pela seguradora

Ao ajustar o modelo com as variáveis Frequência de Sinistros e Valor do Prêmio para a **Hipótese 2 (H_0): impacto da Frequência de Sinistros na Valor do Prêmio**, o modelo apresentou os seguintes resultados principais:

- Intercepto (const): 2.182,93 - Valor base do prêmio, quando não há ocorrência de sinistros.
- Claims Frequency: 73,70 - A cada sinistro adicional, o prêmio aumenta, em média, R$ 73,70.

Os resultados confirmam a hipótese de que o valor do prêmio aumenta proporcionalmente à frequência de sinistros.

- R-quadrado: 0,126 - Isso indica que 12,6% da variabilidade no valor do prêmio pode ser explicada pela frequência de sinistros. Apesar de não ser elevado, esse valor é aceitável para dados financeiros, onde múltiplos fatores influenciam o resultado.

Conclusão

O modelo de regressão OLS revelou uma relação estatisticamente significativa entre a frequência de sinistros e o valor do prêmio de seguro. A cada aumento na frequência de sinistros, observa-se um acréscimo correspondente no prêmio. Embora o R^2 seja moderado, o modelo é robusto e útil para compreender o impacto direto da frequência de sinistros sobre a precificação do seguro.

❖ Modelo de regressão linear múltipla que inclui tanto Frequência de Sinistros quanto Idade como preditores de Valor do Prêmio.

Objetivo da Análise

A presente análise teve como objetivo avaliar a influência conjunta da frequência de sinistros e da idade do segurado sobre o valor do prêmio de seguro, por meio de um modelo de regressão linear múltipla.

Resultado:

- R^2: 0,127
  R^2 ajustado: 0,127
  Modelo explica aproximadamente 12,7% da variação nos valores dos prêmios. Embora esse percentual seja relativamente modesto, ele é considerado adequado em contextos reais do setor de seguros, nos quais diversos outros fatores — como localização geográfica, tipo de veículo e perfil de risco — também exercem influência significativa.
- Intercepto (const): 2.194,58 - Valor base do prêmio quando idade e frequência de sinistros são zero.
- Claims Frequency: 73,67 - A cada sinistro adicional, o prêmio aumenta em média R$ 73,67.
- Idade: -0,29 - Cada ano adicional de idade reduz o prêmio em cerca de R$ 0,29.

Frequência de sinistros: preditor positivo forte e altamente significativo.

Idade: efeito negativo pequeno, mas estatisticamente significativo.

- Interpretação do Efeito da Idade: embora estatisticamente relevante, o efeito da idade sobre o valor do prêmio é muito limitado do ponto de vista prático. A redução de R$ 0,29 por ano de idade pode refletir uma estrutura de precificação em que os segurados mais velhos já possuam prêmios base mais elevados, conforme observado nos dados simulados.
- Estatística F: 727,5 ($p < 0.00001$) → Modelo estatisticamente significativo no geral.
- Durbin-Watson: 1,97 → Ausência de autocorrelação nos resíduos.
- Distribuição dos resíduos: Assimetria e curtose indicam conformidade com a normalidade.

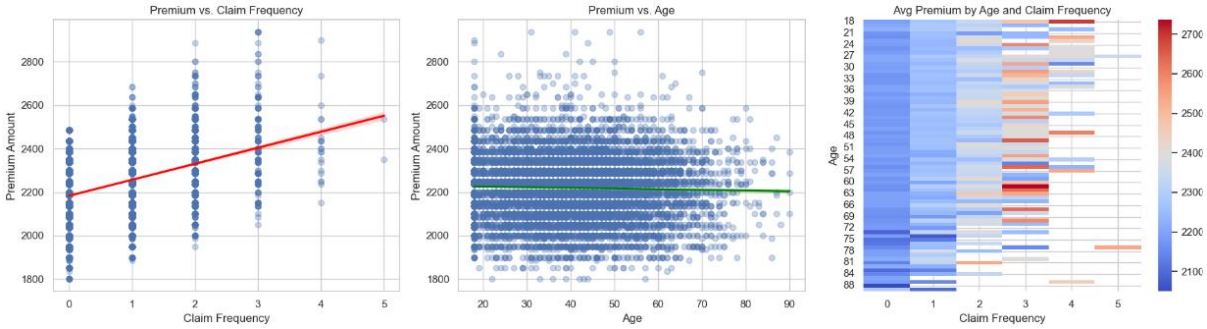
Os diagnósticos confirmam que o modelo atende adequadamente às suposições da regressão linear, aumentando a robustez das conclusões.

Conclusão

A regressão linear múltipla confirma que:

- ✓ A frequência de sinistros tem forte impacto positivo no valor do prêmio;
- ✓ A idade apresenta um efeito negativo estatisticamente significativo, porém sem relevância prática expressiva;
- ✓ O modelo é estatisticamente consistente e representa um avanço em relação a versões anteriores.

Resumo Analítico dos Gráficos:



1. **Gráfico de Dispersão – Prêmio vs. Frequência de Sinistros**

A linha de regressão indica uma tendência positiva, sugerindo que prêmios mais altos estão associados a maior frequência de sinistros.

Há uma dispersão considerável ao redor da linha, indicando a influência de variáveis adicionais além da frequência de sinistros.

2. **Gráfico de Dispersão – Prêmio vs. Idade**

A linha de tendência sugere um possível aumento nos prêmios com a idade, embora de forma não linear.

A alta dispersão dos dados revela que a idade, isoladamente, não explica de forma robusta as variações nos valores dos prêmios.

3. **Heatmap – Prêmio Médio por Idade e Frequência de Sinistros**

O heatmap mostra que prêmios mais elevados ocorrem em segurados mais velhos com alta frequência de sinistros.

Prêmios mais baixos concentram-se em idades mais jovens e menor frequência de sinistros.

O gradiente de cores facilita a identificação visual de padrões, evidenciando zonas de risco mais elevado.

Conclusão Geral

Os gráficos revelam que tanto a idade quanto a frequência de sinistros estão positivamente associadas aos valores dos prêmios, mas nenhuma das variáveis, isoladamente, é suficiente para explicar toda a variabilidade. Isso sugere a necessidade de considerar fatores adicionais na modelagem e precificação de seguros.

Conclusão do Relatório

A análise realizada sobre o conjunto de dados sintético de seguros permitiu investigar de forma detalhada os fatores que influenciam a precificação de prêmios, com foco especial na idade dos segurados e na frequência de sinistros.

Com base nos resultados estatísticos obtidos, podemos concluir o seguinte:

Hipótese 1: A idade do segurado possui relevância estatística na predição da frequência de sinistros.

- Hipótese rejeitada. A análise de regressão mostrou que não há relação estatisticamente significativa entre a idade dos segurados e a frequência de sinistros ($R^2 \approx 0{,}000$; p-valor > 0,05). Portanto, não há evidência suficiente para afirmar que a idade influencia diretamente a frequência com que sinistros ocorrem neste conjunto de dados.

Hipótese 2: A frequência de sinistros influencia o valor do prêmio cobrado pela seguradora.

- Hipótese não rejeitada. Os resultados indicaram uma relação positiva e estatisticamente significativa entre a frequência de sinistros e o valor do prêmio ($R^2 \approx 0{,}126$; $p < 0{,}0001$), o que confirma que, à medida que a frequência de sinistros aumenta, o valor do prêmio também tende a aumentar.

Além disso, ao considerar um modelo de regressão linear múltipla com ambas as variáveis (idade e frequência de sinistros) como preditores do valor do prêmio, verificou-se que:

A frequência de sinistros continua sendo um forte preditor positivo e estatisticamente significativo.

A idade, embora estatisticamente significativa, apresentou um impacto prático mínimo sobre o valor do prêmio.

Por fim, os modelos testados — com e sem PCA — mostraram boa capacidade explicativa, embora o modelo sem PCA tenha apresentado indícios de sobreajuste. Isso reforça a importância do uso criterioso de técnicas de redução de dimensionalidade e validação cruzada para garantir a generalização dos resultados.

Os achados deste relatório reforçam a necessidade de uma abordagem multivariada na avaliação de risco e precificação de seguros, considerando não apenas variáveis individuais, mas suas interações e a complexidade dos dados.