Curso:Pós-Graduação em Ciência de Dados e Inteligência Artificial
Turma:PG PGCD 2025/1 1
Unidade Curricular: Projeto Aplicado - 461344
Professor: Willian Daniel de Matos
Alunos: Adriano Maluf Teixeira, Cristina Aurich, Eliane Nunes da Silva, Julia de Souza Jorge

Relatório Técnico: Projeto Aplicado
Tema: Projeção de Faturamento

Este projeto foi desenvolvido no âmbito da Pós-Graduação em Ciência de Dados e Inteligência Artificial, com o propósito de aplicar técnicas de análise de dados e métodos estatísticos para apoiar a tomada de decisão dos gestores dos negócios. Foram analisados dados históricos de faturamento da instituição com o objetivo de avaliar o comportamento dos dados e, a partir disso:

1) construir um modelo de distribuição mensal de faturamento;
2) elaborar um modelo capaz de prever os valores de faturamento, tendo como base dados históricos.

1. Entendimento do Negócio

A Federação das Indústrias do Estado de Santa Catarina FIESC, é a principal entidade de representação da indústria catarinense e reúne diversas indústrias por meio de sindicatos, além de articular várias frentes para promover o desenvolvimento industrial no Estado representando as indústrias de Santa Catarina.

O Sistema FIESC inclui outras entidades importantes: Serviço Social da Indústria (SESI/SC) com Educação Básica, Saúde, Serviços de Alimentação e Farmácia, Serviço Nacional de Aprendizagem Industrial (SENAI/SC) com Educação Profissional, Ensino Superior, Consultorias e Inovação, Instituto Euvaldo Lodi (IEL/SC) com Desenvolvimento de Talentos e Lideranças e Centro das Indústrias do Estado de Santa Catarina (CIESC) com Associativismo e Serviços Empresariais. As entidades atuam de maneira articulada com a Confederação Nacional da Indústria (CNI) e com os 142 sindicatos industriais filiados à federação para promover um ambiente melhor para os negócios e a indústria.

As entidades do SESI e do SENAI possuem uma organização no Estado em 16 regionais administrativas e operacionais, trazendo capilaridade no atendimento das demandas da indústria e fortalecendo a atuação local, e estão divididas entre: Alto Uruguai Catarinense, Alto Vale do Itajaí, Centro-Norte, Centro-Oeste, Extremo Oeste, Foz do Rio Itajaí, Litoral Sul, Norte-Nordeste, Oeste, Planalto Norte, Serra Catarinense, Sudeste, Sul, Vale do Itajaí, Vale do Itajaí Mirim, Vale do Itapocu.

2. Entendimento dos Dados

Foram obtidos dados históricos de faturamento dos serviços ofertados pelo SESI, SENAI e IEL, referentes ao período de janeiro de 2022 a setembro de 2025, a partir do banco de dados da FIESC. Com o objetivo de garantir a conformidade com a Lei Geral de Proteção de Dados (LGPD – Lei nº 13.709/20) e evitar o uso indevido de informações sensíveis ou sua exposição em sistemas públicos de gerenciamento de dados, foram selecionados apenas os dados estritamente necessários ao estudo, com base em informações publicadas.

O dataset contém informações financeiras e operacionais com as variáveis: Grupo das regionais GEREX (conforme gestão), Regional VP (conforme localização geográfica), Regional, Valor financeiro do faturamento, Filial da regional que foi realizado o faturamento, categoria de cliente (físico ou jurídico), Nível de CR, Grupo de negócio (educação, saúde, tecnologia), Classificação do grupo de negócio: Educação (aprendizagem industrial, técnico, etc) e data do faturamento.

Foi realizada a verificação da qualidade dos dados, identificando a inexistência de valores nulos ou duplicados para tratamento caso fosse necessário, análise de distribuição e consistência das variáveis e conversão de tipos e padronização de unidades com a formatação de moeda e data, além da criação de novas colunas contendo mês e ano, a partir da coluna “data” para aprimorar a análise com a construção de gráficos.

A partir dos dados apresentados do dataset, foram criados gráficos, tabelas e mapas para apoiar na compreensão do comportamento do faturamento e suas variáveis: periodicidade, regional, negócios, etc.

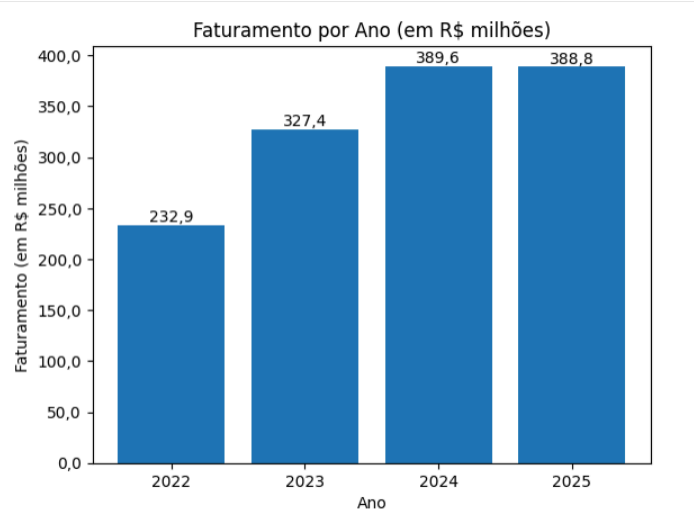


Figura 01 - Faturamento por ano

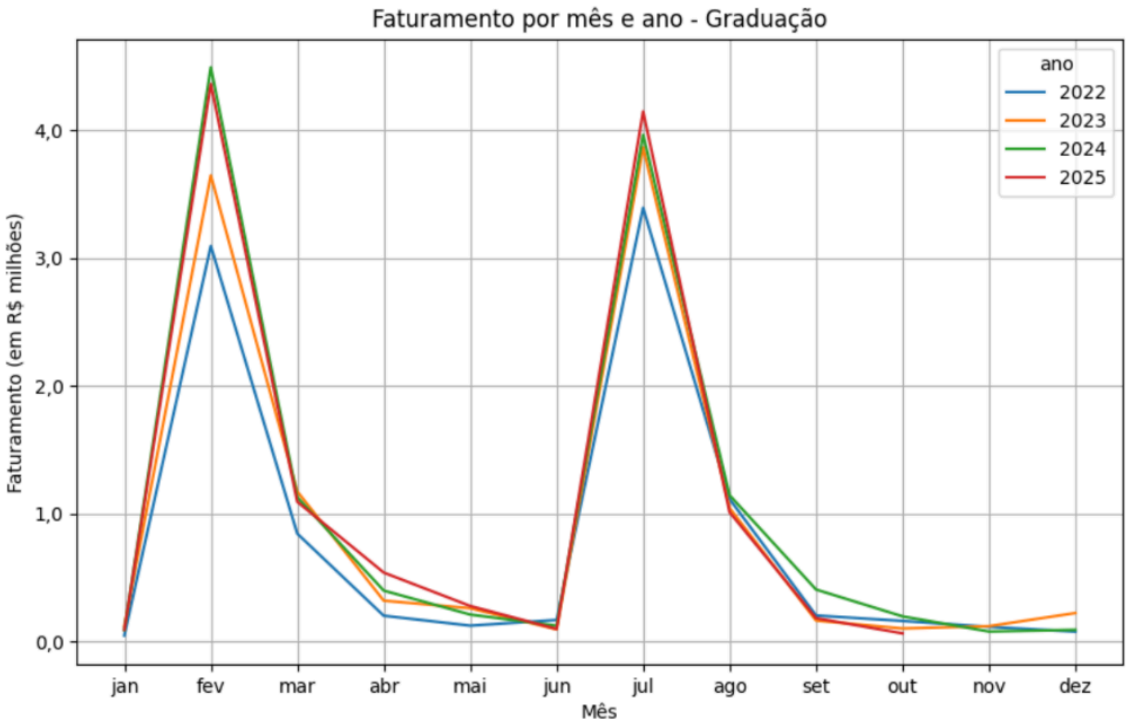


Figura 02 - Faturamento por mês e ano (Graduação)

Analisando os gráficos citados acima, observa-se uma tendência de crescimento anual no faturamento de serviços da FIESC entre os anos de 2022 e 2024, passando de R$ 232,9 milhões em 2022 para R$ 389,6 milhões em 2024. Além disso, observa-se uma considerável variação entre as regionais, evidenciando diferenças geográficas no desempenho financeiro.

Também é possível evidenciar que o faturamento da educação é expressivo e que o negócio “Graduação” possui uma sazonalidade conhecida, pois os meses com maior faturamento (fev e jul) estão relacionados ao período de matrículas.

3. Criação de função para distribuição mensal do faturamento

Foi desenvolvida uma função em python com o objetivo de projetar o faturamento mensal, com base na distribuição real de faturamento dos anos de 2023 e 2024. Para isso, ele aplica uma lógica de ponderação que calcula o peso percentual de cada mês dentro do total faturado no período selecionado, distribuindo a meta anual informada proporcionalmente entre os meses. A aplicação permite filtrar os dados por regional, grupo de negócio e negócio, ajustando automaticamente os resultados conforme os recortes escolhidos. Além disso, o código gera uma tabela detalhada com os valores históricos consolidados, seus respectivos pesos e a meta mensal projetada.

A interface possibilita que o usuário selecione filtros e visualize os resultados de forma dinâmica. O gráfico é gerado a partir da soma consolidada das metas mensais respeitando os filtros selecionados e apresenta uma barra única para cada mês, facilitando a compreensão da projeção ao longo do ano. Com essa combinação de filtros, cálculos automáticos, exibição tabular e visualização gráfica, o código funciona como uma ferramenta analítica que apoia o planejamento financeiro e o acompanhamento de metas, permitindo análises rápidas, intuitivas e comparáveis entre diferentes unidades e segmentos da organização. Foram criados filtros interativos e tabela de faturamento mensal projetada:

Regional: Sudeste
Grupo: SUPERIOR
Negócio: Graduação
Meta anual: 2200000
Atualizar

Filtro – Regional: Sudeste | Grupo: SUPERIOR | Negócio: Graduação | Meta anual: R$ 2.200.00

	ds_grupo_negocio	ds_negocio	gr_regional	mes	soma_2_anos	peso_%	meta_mensal
0	SUPERIOR	Graduação	Sudeste	1	R$ 824,67	0.022540	R$ 495,87
1	SUPERIOR	Graduação	Sudeste	2	R$ 1.335.011,50	36.488024	R$ 802.736,52
2	SUPERIOR	Graduação	Sudeste	3	R$ 329.214,71	8.997971	R$ 197.955,35
3	SUPERIOR	Graduação	Sudeste	4	R$ 86.722,24	2.370259	R$ 52.145,70
4	SUPERIOR	Graduação	Sudeste	5	R$ 66.329,92	1.812904	R$ 39.883,89
5	SUPERIOR	Graduação	Sudeste	6	R$ 858,90	0.023475	R$ 516,45
6	SUPERIOR	Graduação	Sudeste	7	R$ 1.010.122,25	27.608275	R$ 607.382,05
7	SUPERIOR	Graduação	Sudeste	8	R$ 698.936,84	19.103074	R$ 420.267,64
8	SUPERIOR	Graduação	Sudeste	9	R$ 84.175,55	2.300654	R$ 50.614,39
9	SUPERIOR	Graduação	Sudeste	10	R$ 32.714,75	0.894147	R$ 19.671,23
10	SUPERIOR	Graduação	Sudeste	11	R$ 3.284,19	0.089762	R$ 1.974,77
11	SUPERIOR	Graduação	Sudeste	12	R$ 10.570,73	0.288915	R$ 6.356,13

Figura 03 - Gráfico de colunas para projeção de faturamento mensal

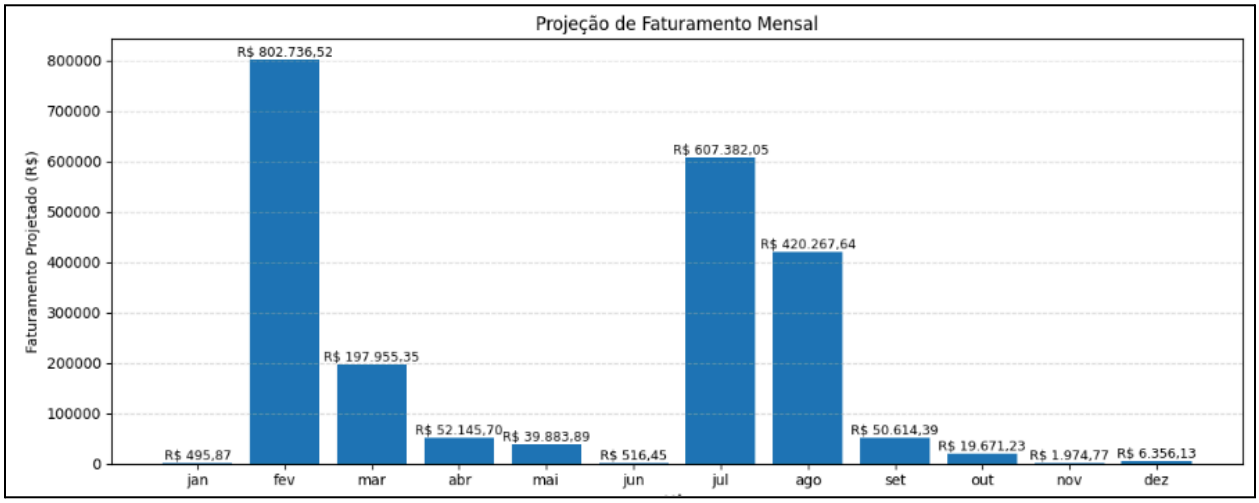


Figura 04 - Projeção de Faturamento Mensal

Observa-se que o gráfico acima condiz com o comportamento de faturamento no negócio “Graduação” evidenciado na figura 05, onde os picos de faturamento estão concentrados no período de matrículas, fevereiro e julho.

3.1 Considerações da Função para distribuição mensal do faturamento

Com a função criada, a operação que hoje é realizada manualmente para distribuição da meta ao longo do ano, através de planilhas eletrônicas, poderá ser feita de maneira mais ágil e confiável. Além disso, exige baixo esforço comparado ao modelo tradicional, pois para gerar uma nova distribuição é

necessário somente a atualização do dataset com os dados dos dois últimos anos e a substituição do período desejado no código da função.

4. Modelo Preditivo com Deep Learning

4.1 Escolha e justificativa da técnica

Considerando os dados históricos de faturamento das entidades do Sistema FIESC apresentados neste material, a proposta do presente projeto visa elaborar um modelo de máquina que seja capaz de prever os valores de faturamento, tendo como base para o aprendizado esses dados históricos. Por conta da complexidade dos dados, optou-se por desenvolver um modelo focado na predição de valores de um único negócio, no caso a **Graduação**, também levando-se em conta os constantes desafios e a forte concorrência que a entidade SENAI enfrenta nesta modalidade. Além disso, como já visto na etapa de entendimento dos dados, esse negócio possui uma característica forte de sazonalidade, que pode permitir um processamento com menos ruídos. Também optou-se, considerando a complexidade da projeção, em não especificar os valores preditos por regional.

O algoritmo escolhido para o desenvolvimento desse modelo foi o **Long Short-Term Memory (LSTM)**. A escolha dessa técnica se fundamenta na natureza sequencial e temporal das informações financeiras. O faturamento de uma empresa é fortemente influenciado por padrões históricos, sazonalidades e tendências que se desenvolvem ao longo do tempo. Modelos tradicionais de regressão não seriam tão eficazes para capturar essas dependências temporais, pois tratam os dados como independentes entre si.

4.2 Preparação dos Dados

A primeira etapa do processo foi criar um novo dataframe (df_filtrado), a partir do dataframe principal (df), contendo a seleção do negócio "**Graduação**".

Considerando a proposta apresentada, foram mantidas apenas as colunas de "data" e "valor", que consistem nas features essenciais para a predição do modelo. Ficou definido, desta forma, o "X" ("data") e o "y" ("valor") do novo dataframe.

```
<class 'pandas.core.frame.DataFrame'>
Index: 3053 entries, 124855 to 130666
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   data    3053 non-null   datetime64[ns]
 1   valor   3053 non-null   float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 71.6 KB
```

Figura 05 - Informações do dataframe filtrado

Como já observado anteriormente, não há valores nulos no dataframe. No entanto, existem datas repetidas e valores faltantes (não há valores de faturamento em todos os dias do ano), o que pode ocasionar ruído e afetar a capacidade do modelo de aprender corretamente as relações entre períodos.

Para garantir a qualidade e consistência das informações, foi necessário aplicar dois procedimentos principais: agrupamento por data e valor e interpolação para valores faltantes.

O agrupamento visa eliminar duplicidades que poderiam distorcer os padrões temporais e comprometer a integridade da série histórica. Já a interpolação para valores faltantes foi aplicada para lidar com lacunas na sequência temporal. A ausência de dados pode prejudicar algoritmos como o LSTM, que dependem da continuidade da série para capturar dependências de longo prazo.

A partir deste ponto, foi realizada separação do conjunto de dados em **treino (80%)**, **validação (10%)** e **teste (10%)**.

A próxima etapa consistiu na aplicação do método **RobustScaler** para normalização dos dados. Essa é uma etapa essencial para o treinamento de modelos baseados em redes neurais, como o LSTM, pois esses algoritmos são sensíveis à escala das variáveis.

O método RobustScaler foi escolhido por sua capacidade de reduzir o impacto de outliers na normalização. Diferentemente de técnicas como MinMaxScaler, que podem ser fortemente influenciadas por valores extremos, o RobustScaler utiliza medidas estatísticas robustas — mediana e intervalo interquartílico (IQR) — para ajustar a escala dos dados. Isso garante que a distribuição seja centralizada e redimensionada de forma mais estável, mesmo em séries temporais com picos sazonais ou variações abruptas, como é o caso do dataframe atual.

Após a normalização dos dados, foi necessário preparar a série temporal para o treinamento do modelo LSTM. Para isso, utilizou-se o método **TimeSplitter**, que é uma ferramenta voltada para a criação de janelas temporais (windows size) a partir de séries históricas.

Modelos como LSTM não trabalham diretamente com dados em formato de série contínua; eles precisam de sequências de observações para aprender padrões temporais. O TimeSplitter automatiza esse processo, dividindo os dados em blocos (janelas) que representam períodos consecutivos, permitindo que o modelo capture dependências de curto e longo prazo.

4.3 Elaboração do Modelo de Rede Neural

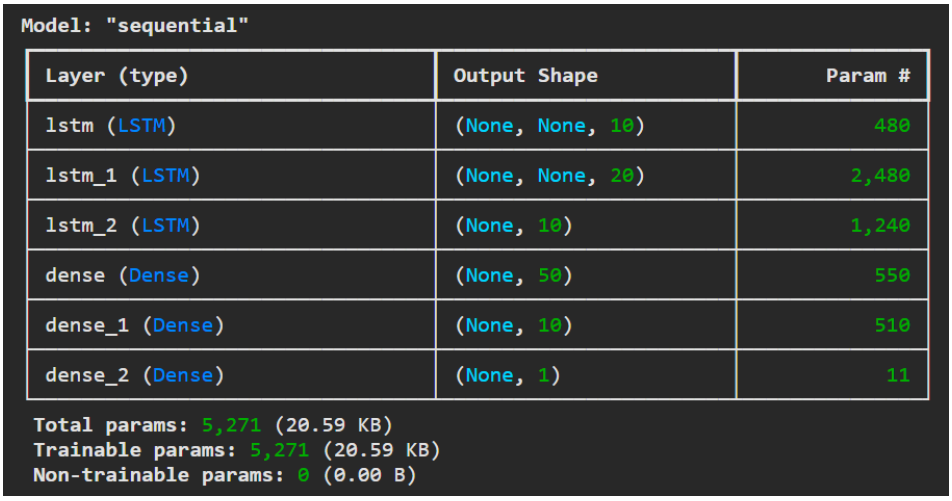
O modelo de rede neural proposto foi desenvolvido com as seguintes sequências de camadas:

- Camadas LSTM: Três camadas recorrentes com diferentes números de unidades (10, 20, 10) e ativação ReLU. As duas primeiras usam return_sequences=True para manter a saída sequencial, permitindo que camadas posteriores capturem dependências temporais mais complexas. A última camada LSTM resume a sequência para uma única saída.

- Camadas densas: Duas camadas totalmente conectadas (50 e 10 neurônios) com ativação ReLU, responsáveis por aprender relações não lineares após a extração de padrões temporais.

- Camada de saída: Uma única unidade com ativação linear, adequada para problemas de regressão, como previsão de faturamento.

Essa arquitetura foi escolhida para equilibrar capacidade de aprendizado temporal (via LSTM) e flexibilidade para ajustes não lineares (via camadas densas), garantindo previsões mais precisas.



Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, None, 10)	480
lstm_1 (LSTM)	(None, None, 20)	2,480
lstm_2 (LSTM)	(None, 10)	1,240
dense (Dense)	(None, 50)	550
dense_1 (Dense)	(None, 10)	510
dense_2 (Dense)	(None, 1)	11

Total params: 5,271 (20.59 KB)
Trainable params: 5,271 (20.59 KB)
Non-trainable params: 0 (0.00 B)

Figura 6 - Esquema da rede neural

Uma etapa fundamental para o processo de otimização e aprendizado da máquina é definir a forma e as métricas que o modelo deve considerar no treinamento. Essa definição é executada pelo método compile e foi composta pelas seguintes métricas:

- **loss = "mse"**: orienta o treinamento para minimizar erros quadráticos, adequado para regressão.
- **optimizer = Adadelta**: garante ajustes adaptativos dos pesos, melhorando estabilidade.
- **metrics = ["mape"]**: fornece uma métrica percentual intuitiva para avaliar previsões de faturamento.

4.4 Treinamento do Modelo

O treinamento do modelo foi ajustado com os seguintes hiperparâmetros:

a) epochs = 900

Define o número de vezes que todo o conjunto de treino será processado pelo modelo. Um valor alto (900) é adequado para redes LSTM, pois elas precisam de várias iterações para aprender padrões temporais complexos como sazonalidade e tendência. Esse número foi escolhido para garantir convergência sem interromper o aprendizado prematuramente.

b) batch_size = 128

Indica quantas amostras serão processadas antes da atualização dos pesos. Um batch maior (128) melhora a eficiência computacional e a estabilidade do gradiente, reduzindo oscilações durante o treinamento. Em séries temporais, isso ajuda a manter consistência no aprendizado.

c) validation_data = (X_valid_scaled, y_valid_scaled)

Permite avaliar o desempenho do modelo em dados não vistos durante o treino, monitorando métricas como MSE e MAPE a cada época. Essa validação é essencial para detectar overfitting e ajustar hiperparâmetros se necessário.

Após o treinamento, foi analisado o histórico das métricas para avaliar se o modelo aprendeu de forma adequada e se não houve problemas como overfitting ou underfitting. O gráfico da curva de aprendizado mostra a evolução da loss (erro) para treino e validação ao longo das épocas.

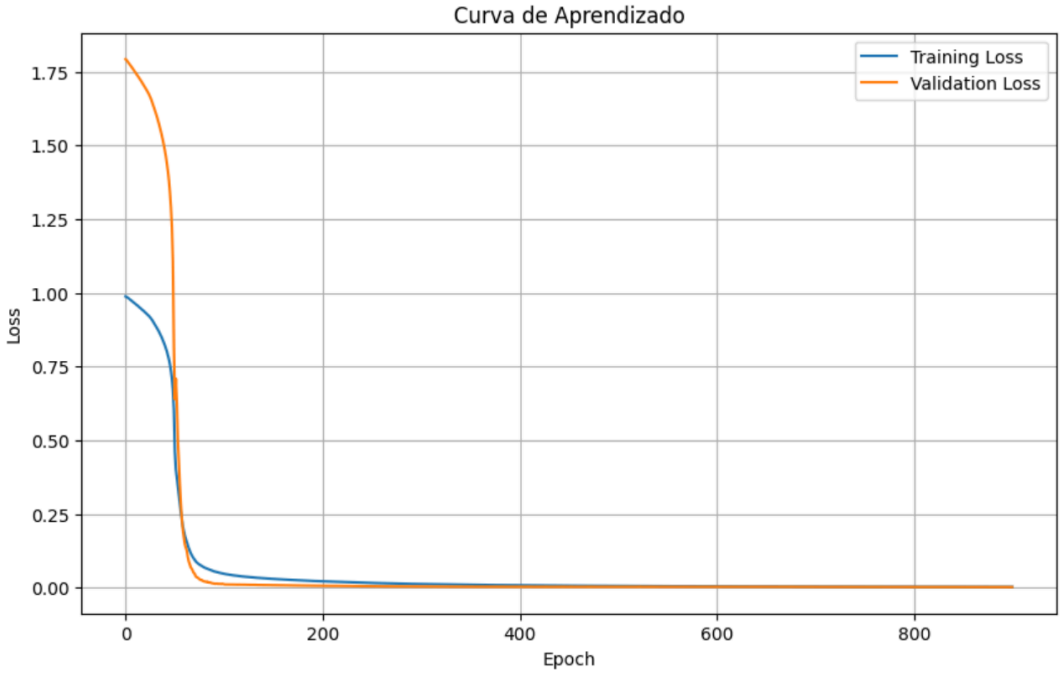


Figura 7 - Curva de Aprendizado do Modelo

Em linhas gerais, tanto a perda de treino quanto a de validação diminuem consistentemente, indicando aprendizado efetivo. Há uma convergência por volta da época 70, com uma queda acentuada na perda de validação, seguida por estabilização próxima à perda de treino. As curvas permanecem próximas até o final, sugerindo que o modelo generaliza bem para dados não vistos e minimiza os riscos de overfitting. Após cerca de 200 épocas, ambas as curvas praticamente atingem zero e permanecem estáveis até o final, indicando que o prolongamento para 900 épocas não trouxe ganhos significativos e que esse número poderia ser reduzido consideravelmente, poupando tempo sem perda de qualidade.

4.5 Avaliação do Modelo

Para avaliar o desempenho do modelo de previsão, foram utilizadas as seguintes métricas, com os seguintes resultados:

- **MAE: 39.473,69**

Indica que, em média, as previsões diferem cerca de R$ 39,5 mil do valor real. Considerando os dados de faturamento da Graduação, esse valor é relativamente baixo, tendo em mente a escala dos dados.

- **MAPE: 9,25%**

Um erro percentual inferior a 10% é considerado muito bom para previsões financeiras, mostrando alta precisão e boa capacidade de generalização.

- **R^2: 0,9979**

O modelo explica 99,79% da variabilidade da série, o que indica excelente ajuste e forte correlação entre valores previstos e reais.

O gráfico a seguir compara os dados de treino, validação e teste.

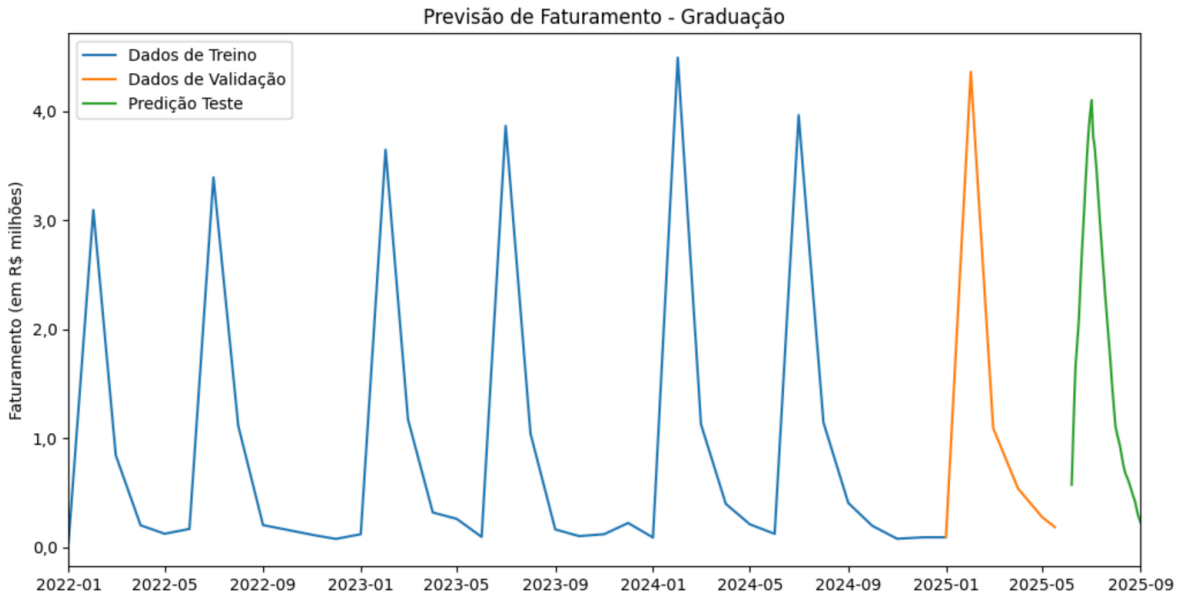


Figura 8 - Gráfico de Avaliação do Modelo (treino, validação e teste)

Esse gráfico mostra que o modelo LSTM reproduziu corretamente a sazonalidade e tendência do faturamento, com previsões coerentes para o período de teste.

Um novo gráfico foi plotado, focando nos dados preditos e comparando com os dados reais do dataframe.

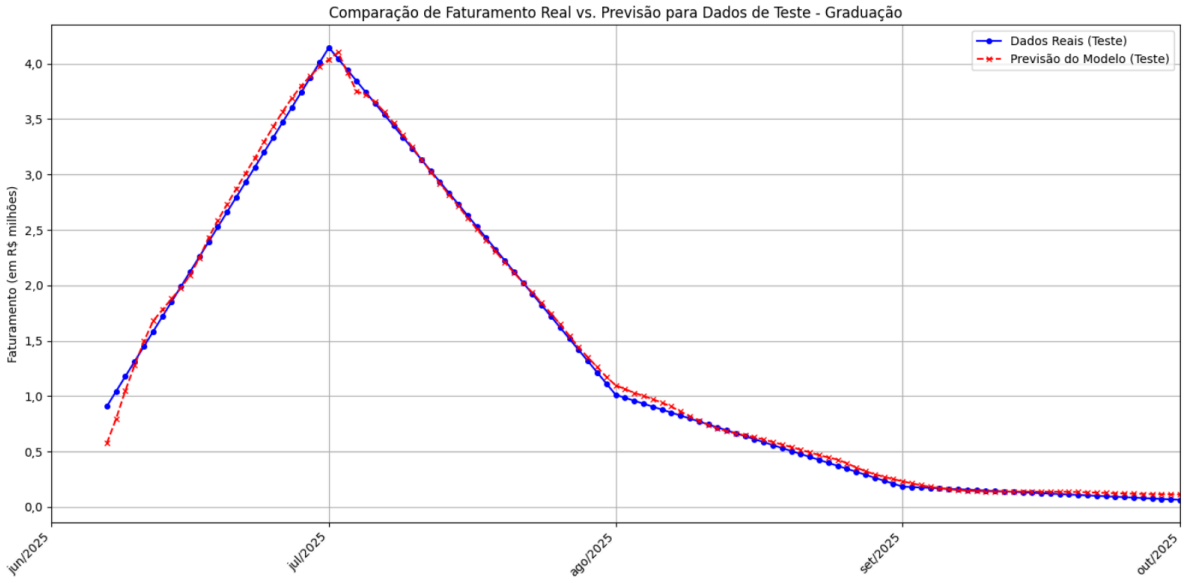


Figura 9 - Gráfico de Avaliação do Modelo (dados reais X previsão)

4.6 Considerações do Modelo Preditivo com Deep Learning

O modelo apresentado neste trabalho conseguiu capturar a sazonalidade da série, reproduzindo os picos anuais e a tendência geral. A previsão segue o comportamento esperado, indicando que o LSTM aprendeu bem os padrões temporais. No entanto, é importante destacar que a série utilizada (valores históricos de faturamento de cursos de Graduação) apresenta uma forte característica de sazonalidade, com picos anuais nos meses de fevereiro e julho, coincidindo com os períodos de matrículas. Como um desdobramento da proposta apresentada, sugere-se o treinamento do modelo para dados de outras linhas de serviços, que não possuam características de sazonalidade forte ou regularidade nas linhas históricas. Para esses casos, há forte probabilidade de ajustes nos hiperparâmetros do modelo.

Também destaca-se que o modelo possui limitações e uma das mais evidentes é que ele depende da qualidade dos dados apresentados. Registros indevidos, dependendo da consulta utilizada para o treinamento, podem interferir na projeção final. Outra limitação é a ausência de variáveis externas na projeção realizada pela máquina, como fatores macroeconômicos, investimentos realizados pela entidade, mudanças de estratégia do negócio que impactam no faturamento, etc.

Por fim, como possíveis ações de melhoria, destaca-se a ampliação dos dados do dataset utilizado, com a inclusão de dados anteriores a 2022. Outra sugestão seria a utilização de outros algoritmos de aprendizado de máquina, como modelos ARIMA e até mesmo Regressão Linear, para aplicação em datasets com forte sazonalidade e poucos ruídos na série temporal. A vantagem, em relação à modelos de rede neural, seria a utilização de uma técnica menos complexa, com menos hiperparâmetros para serem calibrados e que exigem menos capacidade computacional.

5. Implementação / Entrega

5.1 Consolidação dos notebook

O projeto foi desenvolvido no **Google Colab**, com um notebook:

1. **Projeto Aplicado.ipynb**

5.2 Arquivos

- **dados_faturamento.xlsx**
- **geojs-42-mun.json**
- **Municipios_por_regionalVP.xlsx**
- **Relatório Final Projeto Aplicado.pdf**