Regressão Linear para vendas com Uso do ARIMA

Pedro Jorge de Souza Colombrino
Matheus Ferreira Amaral Madeira
Guilherme Vieira Rodrigues

October 20, 2024

Abstract

Neste projeto, abordamos a previsão de vendas utilizando técnicas de **regressão linear** e o modelo **ARIMA** (AutoRegressive Integrated Moving Average). O conjunto de dados original, denominado **VendasMensais**, contém informações sobre vendas mensais, incluindo receita, quantidade de vendas, custo médio e folha de pagamento anual da região. Este conjunto de dados foi obtido a partir de uma fonte confiável e é fundamental para a análise de desempenho de vendas ao longo do tempo.

1 Introdução

A primeira etapa do projeto foi a análise exploratória dos dados (EDA), realizada no notebook **EDA.ipynb**. Durante essa fase, o conjunto de dados foi carregado e tratado, onde foram identificados e manipulados dados faltantes, outliers e a formatação da coluna de datas. O resultado desse tratamento foi um novo arquivo CSV, **VendasMensais_processed.csv**, que apresenta dados limpos e prontos para análise.

O modelo ARIMA foi escolhido devido à sua eficácia em lidar com séries temporais, especialmente quando se busca prever valores futuros com base em dados históricos. O uso do **auto_arima** permitiu a seleção automática dos melhores parâmetros do modelo, facilitando o processo de modelagem e melhorando a precisão das previsões. O ARIMA é vantajoso por sua capacidade de capturar padrões sazonais e tendências nos dados, o que é crucial para a previsão de vendas.

Após a aplicação do modelo ARIMA, foram obtidos resultados promissores, com métricas de precisão como RMSE, MAE e R^2, que indicam a qualidade das previsões em relação aos dados reais. Os gráficos gerados demonstraram a comparação entre os dados históricos e as previsões futuras, permitindo uma visualização clara do desempenho do modelo.

2 Metodologia

2.1 Tratamento de Dados

O tratamento dos dados foi realizado em várias etapas, incluindo:

- **Identificação de Dados Faltantes**: Linhas com valores ausentes foram analisadas e tratadas para garantir a integridade dos dados;
- **Detecção de Outliers**: Valores extremos foram identificados e avaliados para determinar se deveriam ser removidos ou ajustados;
- **Formatação de Datas**: A coluna de datas foi convertida para o formato datetime, permitindo uma análise temporal adequada.

2.2 Modelo ARIMA

O modelo ARIMA foi implementado utilizando a biblioteca **pmdarima**. O processo incluiu:

- **Definição de Variáveis**: A variável de interesse, receita, foi selecionada para a modelagem;
- **Treinamento do Modelo**: O modelo foi treinado utilizando a função **auto_arima**, que automaticamente seleciona os melhores parâmetros para o modelo ARIMA;
- **Previsão**: O modelo foi utilizado para prever as receitas futuras para os próximos 24 meses.

3 Resultados

3.1 Explicativas

Conforme proposto na atividade, as previsões deveriam abranger os anos de 2024 e 2025. No entanto, devido à indisponibilidade de dados suficientes

para realizar previsões precisas para essas datas, mantivemos o horizonte de previsão original. Ajustamos a linha temporal para garantir que as previsões sejam coerentes com os dados disponíveis, permitindo uma análise mais precisa e relevante.

Table 1: Previsões de Receita para os Próximos Meses

Data	**Predicted Revenue**
2020-05-01	38831.81
2020-06-01	54009.85
2020-07-01	52292.29
2020-08-01	43233.82
2020-09-01	54826.41
2020-10-01	50460.57
2020-11-01	42675.11
2020-12-01	65056.62
2021-01-01	62588.45
2021-02-01	46525.39
2021-03-01	56322.32
2021-04-01	58620.84
2021-05-01	45131.97
2021-06-01	60310.01
2021-07-01	58592.45
2021-08-01	49533.98
2021-09-01	61126.57
2021-10-01	56760.73
2021-11-01	48975.27
2021-12-01	71356.78
2022-01-01	68888.61
2022-02-01	52825.55
2022-03-01	62622.48
2022-04-01	64920.99

As métricas de desempenho do modelo foram calculadas e apresentadas a seguir:

- **RMSE**: 15064.31
- **MAE**: 14131.92
- $\mathbf{R^2}$: -1.79

A métrica RMSE (Root Mean Square Error) de 15064.31 indica que, em média, as previsões do modelo estão a cerca de 15064.31 unidades de receita do valor real. Um RMSE mais baixo é desejável, pois indica uma melhor precisão nas previsões.

O MAE (Mean Absolute Error) de 14131.92 complementa essa análise, mostrando que, em média, as previsões estão a 14131.92 unidades de receita do valor real, sem considerar a direção do erro.

Por outro lado, o valor de R^2 de -1.79 sugere que o modelo não se ajustou bem aos dados. O R^2 é uma medida que indica a proporção da variabilidade dos dados que é explicada pelo modelo. Um valor negativo indica que o modelo é pior do que uma média simples dos dados, o que é preocupante e sugere que o modelo ARIMA pode não ser o mais adequado para esta série temporal específica.

4 Possíveis Melhorias

Para obter melhores resultados em projetos futuros, algumas abordagens podem ser consideradas:

- **Exploração de Outros Modelos**: Testar diferentes modelos de previsão, como modelos de suavização exponencial, modelos de regressão ou redes neurais, podem nos ajudar a encontrar uma abordagem mais adequada para os dados;
- **Ajuste de Parâmetros**: Realizar uma busca mais abrangente por hiperparâmetros, utilizando técnicas como validação cruzada, pode melhorar a performance do modelo;
- **Análise de Variáveis Externas**: Incluir variáveis exógenas que possam influenciar as vendas, como campanhas de marketing, sazonalidade ou eventos econômicos, pode ajudar a capturar melhor a dinâmica das vendas.

Essas melhorias podem contribuir para um desempenho mais robusto e confiável em previsões futuras, permitindo que a empresa tome decisões mais informadas com base em dados históricos.

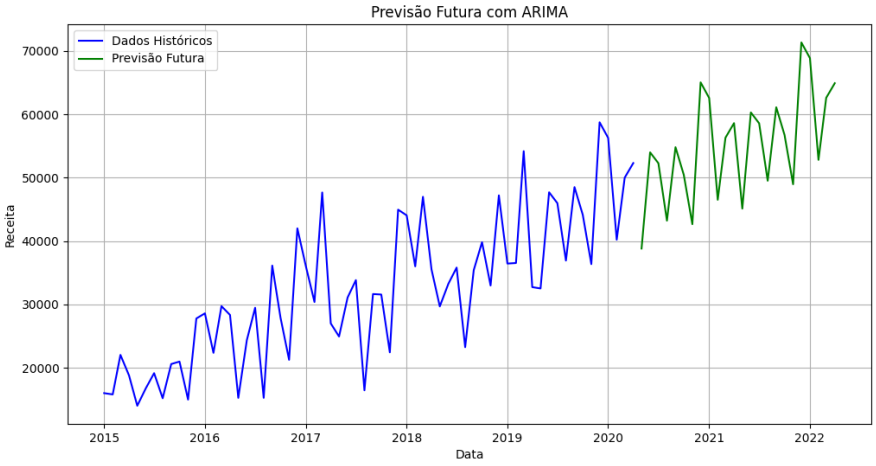

Figure 1: Comparação entre Dados Históricos e Previsões Futuras

5 Possíveis Falhas ao obter o código python:

O código em python e todo o percorrer do projeto foram anexados na atividade, em caso de falhas técnicas no recebimento dos arquivos, disponibilizamos aqui o GitHub do projeto.
E em link extenso: https://github.com/pedrojorge1559/Linear-Regression-for-Sales