UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

ALINE DIAS NUNES

Modelos preditivos com aplicação no varejo

Campinas
2022

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

ALINE DIAS NUNES

Modelos preditivos com aplicação no varejo

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. João Batista Florindo.

Campinas
2022

1 Resumo

No setor de varejo, previsão de vendas é uma informação crítica. Não prever as vendas pode acarretar por exemplo em orçamento desbalanceado, incidência de ruptura e excesso de produtos, menos controle sob estoque.

Entretanto, prever não é uma tarefa simples. Dependendo da complexidade do problema de previsão há necessidade de modelos cada vez mais robustos, além da necessidade de mais dados para serem treinados, consequentemente, exigindo maior custo operacional.

Neste trabalho, propõe-se a construção de um modelo de previsão de vendas no varejo do setor de cosméticos que melhore a acurácia de predição.

Para isso, foram analisados modelos de regressão múltipla, de *forecasting* estatístico como SARIMA e SARIMAX, além de modelos considerados estado-da-arte como *Prophet*, projeto *open source* desenvolvido pelo Facebook, e *DeepAR*, algoritmo de previsão do projeto *Amazon SageMaker*, com o intuito de obter um modelo que melhore a acurácia de predição.

Em relação a resultados, o modelo de regressão múltipla apresentou melhor desempenho, o que pode ser explicado pela seleção cuidadosa das *features* mais relevantes.

Conteúdo

2 Introdução

No varejo, antes de expor o negócio a riscos é importante planejamento. Para conseguir traçar um bom plano, primeiramente é necessario ter acesso a dados confiáveis e organizados, assim é possível tomar decisões de forma mais segura e que auxiliam na criação de metas.

Neste trabalho, cujo objetivo é a construção de uma modelo de previsão do total vendido no varejo do setor de cosméticos com menor erro possível, foi feita análise exploratória dos dados com intuito de identificar sazonalidades e alterações no comportamento do total vendido.

Após o entendimento dos dados, foram incorporados modelos de regressão múltipla, de *forecasting* como SARIMA e SARIMAX, além de testada a performance de modelos considerados estado-da-arte como *Prophet*, projeto *open source* desenvolvido pelo Facebook e *DeepAR*, algoritmo de previsão *Amazon SageMaker*.

3 Metodologia

Em busca de identificar padrões e mostrar o que pode acontecer através da analise do histórico dos totais vendidos na loja de cosméticos, foram aplicados conceitos estatísticos para análises preditivas como: regressão múltipla, séries temporais, *framework Prophet* e também o algoritmo *DeepAR*, em busca de identificar a aplicação com a melhor predição. Em seguida, são apresentados o conceito dos modelos, *framework* e algoritmo utilizado.

3.1 Regressão Múltipla

O modelo de Regressão Múltipla é uma das técnicas estatísticas mais simples para se obterem previsões. Leva em consideração variáveis externas e independentes que possam interferir no valor predito.

Ao adicionar variáveis externas no modelo, tem-se uma tendência e/ou sazonalidade determinística, ou seja, o comportamento das variáveis se mantém ao longo de todo período em estudo. Nesse caso, a ordem das observações é irrelevante.

Esse tipo de modelagem é bem importante por permitir que se estabeleça uma relação de causalidade entre as variáveis características, ou seja, uma equação que represente o quanto muda uma determinada resposta quando variamos alguma característica preditora.

O modelo matemático que estabelece a relação funcional entre as variáveis é definido como:

$$y = \beta_0 + \beta_i x_i + \varepsilon, \qquad i = 1, 2, 3, \cdots, k$$

Temo no caso y como sendo o valor predito, variável dependente, e β_0, ..., β_k são parâmetros a serem estimados; já X_1, X_2, ..., X_k são variáveis externas, independentes e ε é um erro aleatório referente a variabilidade em y que não podem ser explicadas pelas variáveis x. Utiliza-se o método de mínimos quadrados para estimação dos parâmetros.

3.2 Séries Temporais

Os métodos de séries temporais são técnicas estatísticas para se obterem previsões levando-se em consideração o tempo, ou seja, as observações vizinhas são dependentes. O interesse é analisar e modelar essa dependência. Portanto, uma série temporal é uma sequência de números coletados em intervalos regulares durante um período de tempo.

A maioria dos métodos de previsão de séries temporais se baseia na suposição de que as observações passadas contêm todas as informações sobre o padrão de comportamento da série temporal. Esse padrão é recorrente ao longo do tempo. O propósito dos métodos de previsão consiste em distinguir o padrão de qualquer ruído que possa estar contido nas observações e então usar esse padrão para prever os valores futuros da série temporal.

A série temporal pode ser decomposta em:

Tendência: Padrão que se destaca quando há uma linha de crescimento ou queda das vendas ao longo do tempo.

Sazonalidade: Padrões que se destacam quando as vendas são influenciadas por fatores sazonais como dias da semana, semana do mês, feriados, etc.

Uma série temporal é considerada estacionária quando não é observado nenhum padrão de tendência, seja crescente ou decrescente, ao longo da série, ou seja, é considerada estacionária se satisfaz os seguintes critérios.

- Esperança de X constante para todo período t.
- Variância de X constante para todo período t.
- Covariância (Y_t, Y_{t+h}) é função somente de h.

3.3 *Prophet*

O Prophet é um *framework open source* desenvolvido pela equipe *Core Data Science* do Facebook. É uma das ferramentas utilizadas para realizar previsões de séries temporais.

A ferramenta utiliza o modelo de séries temporais decomposto com três componentes principais: tendência (g), sazonalidade (s) e feriados (h), combinados na seguinte equação:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Com a ferramenta é possível a modelagem levando-se em consideração sazonalidades anuais, mensais e semanais, assim como feriados nacionais.

3.4 *DeepAR*

O algoritmo de previsão *Amazon SageMaker DeepAR* é um método de aprendizado supervisionado para prever séries temporais escalares (unidimensionais) usando redes neurais recorrentes (RNN). Usa-se a função de verossimilhança binomial para que o modelo seja capaz de realizar predições probabilísticas.

3.5 Aplicação

Em busca de acompanhar tendências de uma loja de cosméticos, existe a necessidade de se fazerem análises preditivas para traçar estratégias para o setor. Para o estudo, foi considerada uma base de dados de uma loja de cosméticos fictícia.

3.6 Banco de Dados

A base de dados é referente a dados do período de 01/01/2018 até 31/03/2021, totalizando assim 1186 dias de observação de uma loja de cosméticos, sendo que em 16 dias a loja não funcionou. A esses dias foi atribuido o valor 0.01 às vendas.

Cada dia observado é composto por 14 atributos sendo eles: data, dia, mês, ano, dia da semana, número da semana do mês, feriados, dias não úteis, pagamento, vale, vendas e mês do ano.

Na Figura 1 é possível observar as cinco primeiras observações da base de dados.

Data	dia	mês	ano	Dia_da_semana	Num_Semana_mes	feriados	DiasNaoUteis	Dia_Úteis	pagamento	Vale	Vendas	mês_ano
2018-01-01	1	1	2018	Segunda	1	1	1	0	0	0	0.01	2018-01
2018-01-02	2	1	2018	Terça	1	0	0	1	0	0	296611.15	2018-01
2018-01-03	3	1	2018	Quarta	1	0	0	2	0	0	235446.26	2018-01
2018-01-04	4	1	2018	Quinta	1	0	0	3	0	0	233983.82	2018-01
2018-01-05	5	1	2018	Sexta	1	0	0	4	1	0	331047.31	2018-01

Figura 1: Cinco primeiras observações da Base de dados

Sendo que:

- **Num_Semana_mes:**
  1 = primeira semana do mês, vai do dia 01 até o dia 07.
  2 = segunda semana, vai do dia 08 até dia 14.
  3 = terceira semana, vai do dia 15 até 21.
  4 = quarta semana, vai do dia 22 até 28.
  5 = quinta semana, dia 29 até o último dia do mês.

- **feriados:** 1 quando é feriado no dia, 0 caso contrário.
- **DiasNaoUteis:** 1 quando é domingo e feriado, 0 caso contrário.
- **DiasÚteis:** Numeração dos dias úteis do mês e 0 quando não é um dia util.
- **pagamento:** 1 quando corresponde ao quinto dia util do mês, 0 caso contrário.
- **Vale:** 1 quando corresponde ao pagamento do dia 20 do mês, 0 caso contrário.
- **Vendas:** Valor vendido em reais.

Para complementar a base de dados, foi feita a adição de informações públicas como: PIB, taxa selic, desemprego e Índice geral de preços (IGP-M).

3.7 Análise Exploratória

Com intuito de saber o comportamento das vendas na loja de cosméticos de janeiro/2018 até março/2021, foi feito o gráfico da série para uma análise dos valores vendidos por mês, ano e dia da semana.

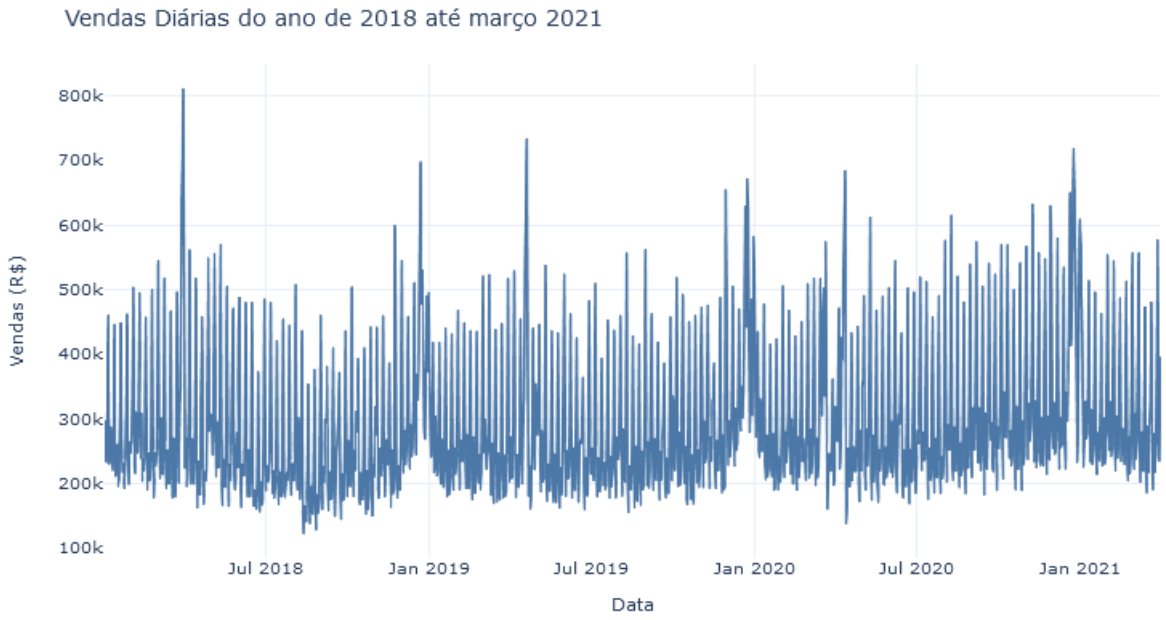


Figura 2: Vendas do ano 2018 até março 2021 por dia

Na Figura 2, aparentemente não existe tendência de alta ou de baixa, a média é constante durante o tempo, um indício de estacionaridade.

Para checar a hipótese foi realizado o teste de Dickey-Fuller, onde:

$$H_0 = \text{Série não é estacionária}$$

$$H_1 = \text{Série é estacionária}$$

Para o período em estudo foi obtido p-valor de aproximadamente zero ou seja, a nível de significância de 5% a série é estacionária.

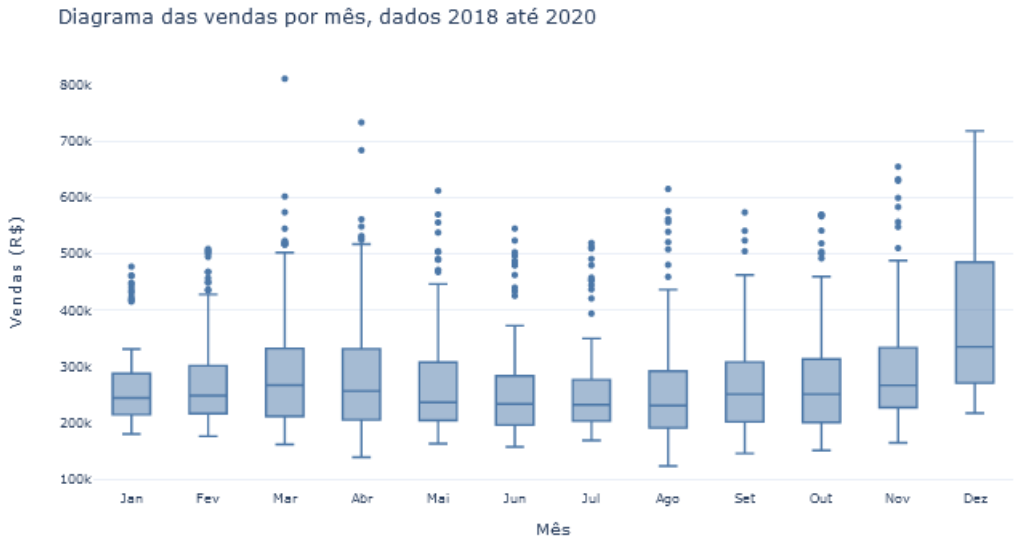


Figura 3: Diagrama vendas por mes, 2018 ate 2020

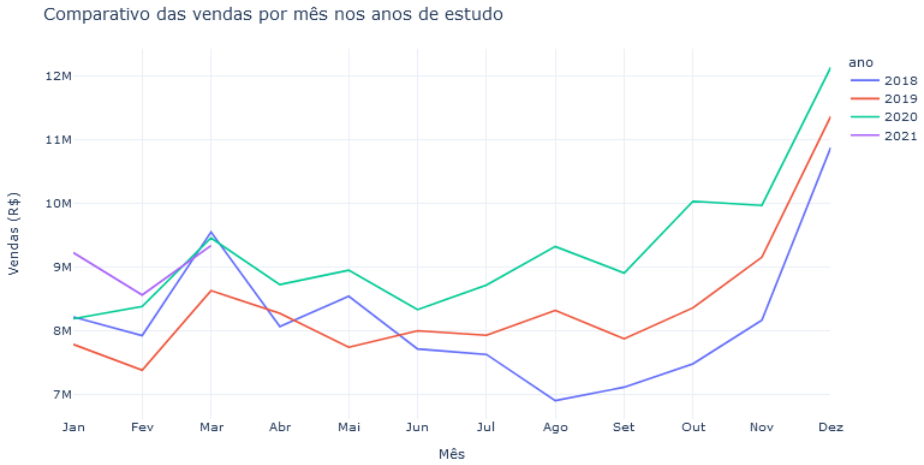


Figura 4: Comparativo das vendas por mês nos anos de estudo

Analisando as figuras 3 e 4 é possível perceber que dezembro foi o mês que vendeu mais em reais. Além disso, há uma flutuação do total vendido entre os meses. Fazendo um comparativo entre os anos têm-se que, o ano 2018 apresentou comportamento diferenciado dos demais.

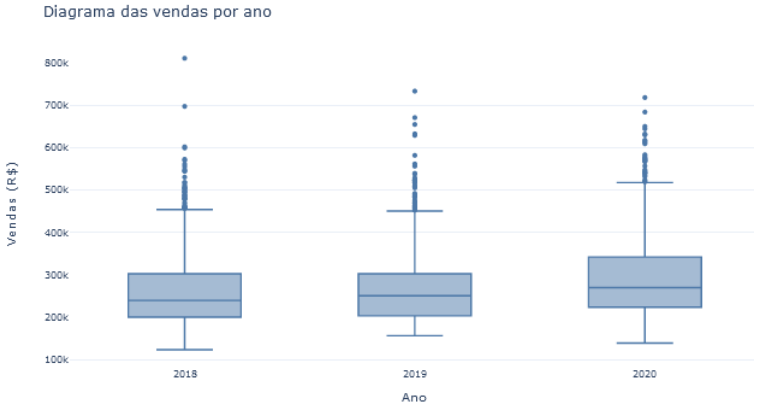


Figura 5: Diagrama das vendas por ano

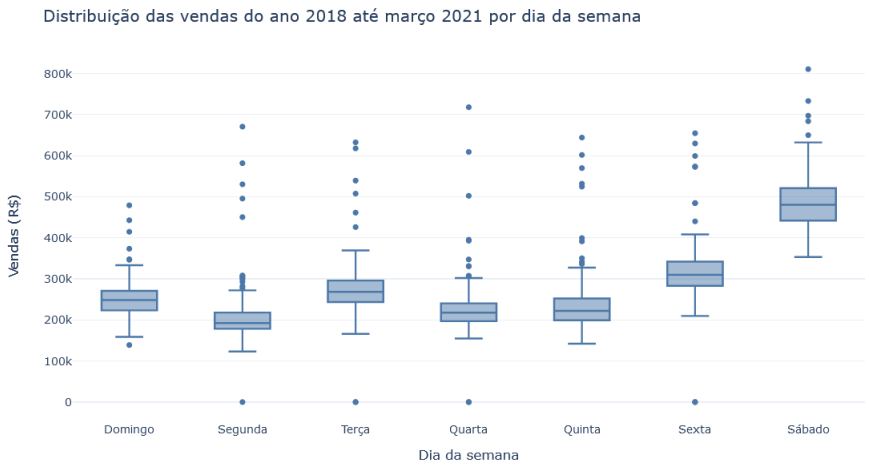


Figura 6: Distribuição das vendas do ano 2018 até março 2021 por dia da semana

Na Figura 6 observa-se que as sextas e sábados são os dias da semana com maiores totais vendidos, enquanto o começo da semana possui os menores.

Com intuito de checar padrões de repetição ao longo do tempo, foi feita a função de autocorrelação

(FAC), marcando a influência do passado no presente. Sua representação é um correlograma e a função de autocorrelação Parcial (FACP) também mede a influência do passado no presente, mas é medida de forma diferente que a FAC. Apresenta-se em seguida a FAC e FACP do período em estudo.

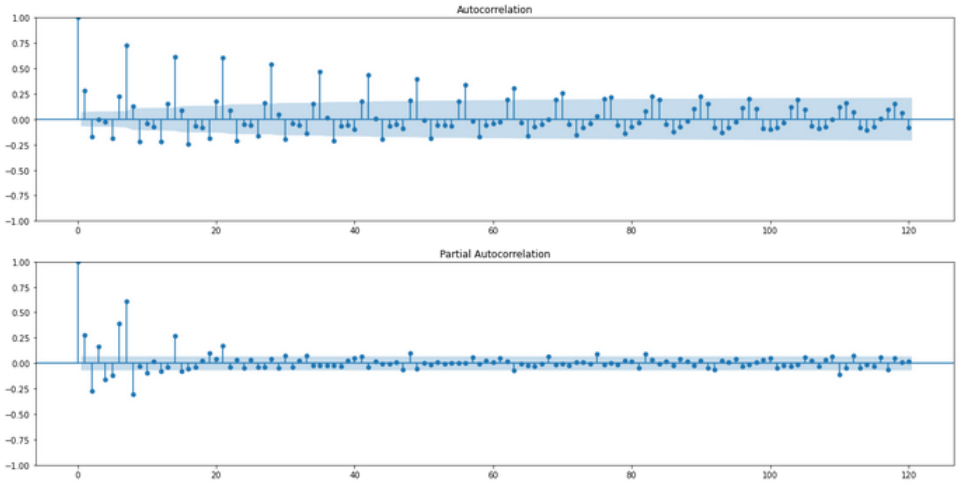


Figura 7: FAC e FACP da série

Na primeira imagem da Figura 7, é possível perceber um comportamento que se repete de sete em sete, mostrando que há sazonalidade semanal.

Ao realizar a mesma análise anterior no entanto com a série diferenciada, foram obtidas as seguintes FAC e FACP.

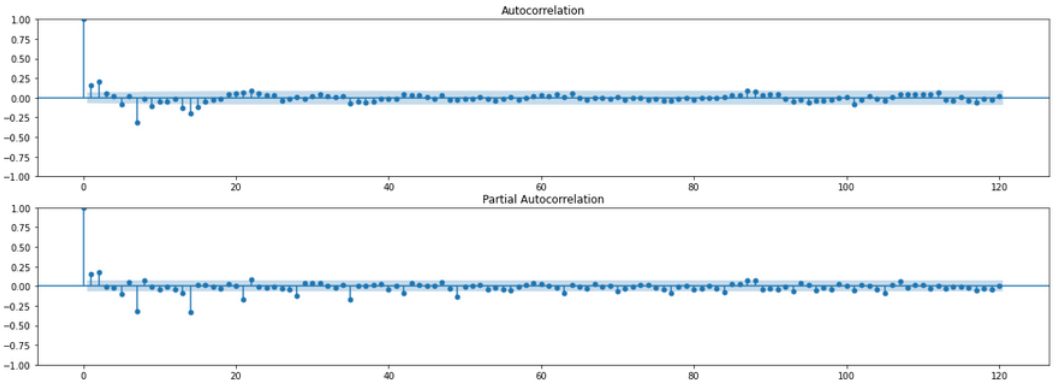


Figura 8: FAC e FACP da série diferenciada

Para identificar a presença de *outliers* foi utilizado *z-score*, fornecendo uma ideia do quanto um determinado ponto está afastado da média dos dados e quantos desvios padrões abaixo ou acima da média dos dados estão. É dado por:

$$z = \frac{x - \mu}{\sigma}$$

Quanto mais longe o *z-score* de uma observação está de zero mais incomum é. Foi utilizado valor de corte 3. As datas cujas vendas estão mais afastadas da média, são:

- 2018-03-31, 2019-04-20, 2020-04-11, 2021-04-03 (sábado de aleluia)
- 2020-11-07
- 2019-11-29, 2020-11-27 ($textitblackfriday$)

- 2018-12-22, 2019-12-21, 2019-12-23, 2019-12-24, 2020-12-19, 2020-12-23, 2020-12-24 (semana do natal)

Ou seja, os dias que apresentaram vendas discrepantes tratam-se de períodos sazonais. Com isso, foram criadas as variáveis sábado de aleluia e *black friday* para contemplar as sazonalidades.

Desde 2020 o mundo enfrenta a pandemia do coronavírus(COVID-19) causada pelo vírus SARS-COV-2. Desde o registro do primeiro caso de COVID-19 no dia 26/02/2020 muitos setores econômicos foram afetados por diversos motivos, dentre eles a necessidade de fechar as portas para o isolamento.

Com a pandemia o PIB, taxa selic, desemprego e Índice geral de preços (IGP-M) podem ter sofrido impactos diferenciados em relação aos periodos antes da pandemia, influenciando assim no total vendido na loja de cosméticos. Com isso, surgiu interesse em checar a correlação do total vendido com as informações públicas.

Na Figura 9 é possível observar-se o correlograma das variáveis públicas com total vendido.

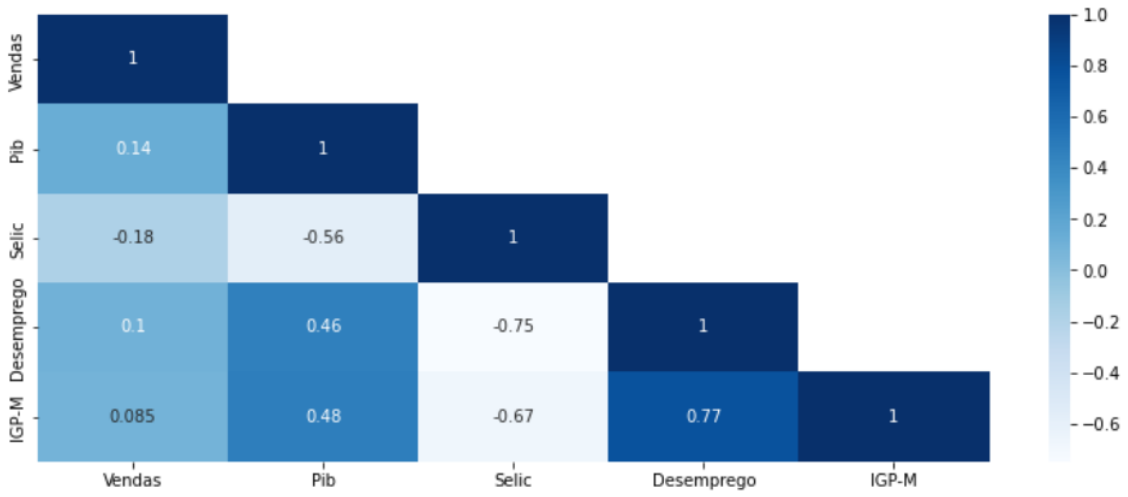


Figura 9: Correlograma variáveis contínuas

As variáveis PIB, taxa selic, desemprego e IGP-M apresentaram correlação bem baixa com as vendas.

3.8 Modelagem

3.8.1 Regressão Múltipla

O modelo de regressão múltipla, examina relação entre variáveis independentes e uma variável dependente e tem capacidade de determinar a influência relativa de uma ou mais variáveis preditoras no valor do critério.

Na modelagem foram consideradas apenas informações a partir de 2019, dado que no ano de 2018 o total vendido teve um comportamento diferenciado em relação aos demais anos.

Os dados foram separados em treino e teste, sendo a base de treino composta por dados de 2019 e 2020 e a base de teste por dados de janeiro até março de 2021.

Foram testados vários modelos, entre eles os quatro modelos citados a seguir:

Modelo	R^2 treino	R^2 teste	MAE teste	RMSE teste
Modelo 1	0,755	0,73	42407	50752
Modelo 2	0,753	0,72	43502	51752
Modelo 3	0,695	0,88	23165	33787
Modelo 4	0,706	0,90	22727	31128

A variável preditora é a variável de vendas e as variáveis explicativas seguem descritas a seguir nos seus respectivos modelos.

Modelo 1: composto por: feriados, pagamento, vale, semana do mês, dia da semana, *outlier black friday*, *outlier* semana santa, PIB, taxa selic, desemprego e IGP-M

Modelo 2: Foram removidas as variáveis que não foram significativas do modelo 1, ficando assim apenas: semana do mês, dia da semana, *outlier black friday*, *outlier* semana santa, PIB, taxa selic, desemprego.

Modelo 3: Foram desconsideradas todas as variáveis públicas, ficando assim apenas: feriados, pagamento, vale, semana do mês, dia da semana, *outlier black friday*, *outlier* semana santa.

Modelo 4: semana do mês, dia da semana, *outlier black friday*, *outlier* semana santa e desemprego.

O modelo com melhor desempenho foi o Modelo 4, por apresentar os menores valores de MAE e RMSE, maior valor de R^2 e melhor ajuste. Nas figura abaixo é possível visualizar a predição da base de teste.

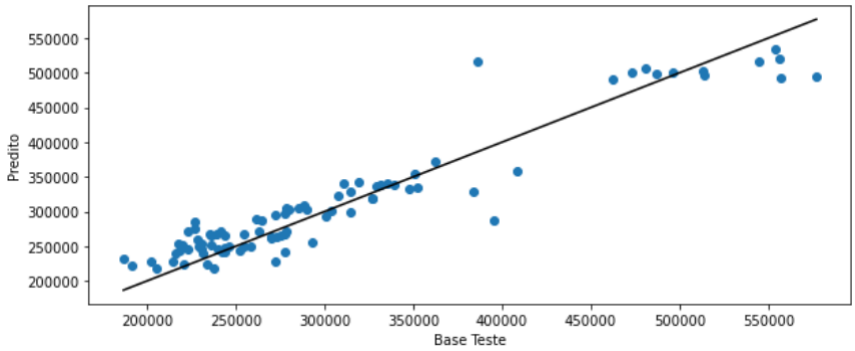


Figura 10

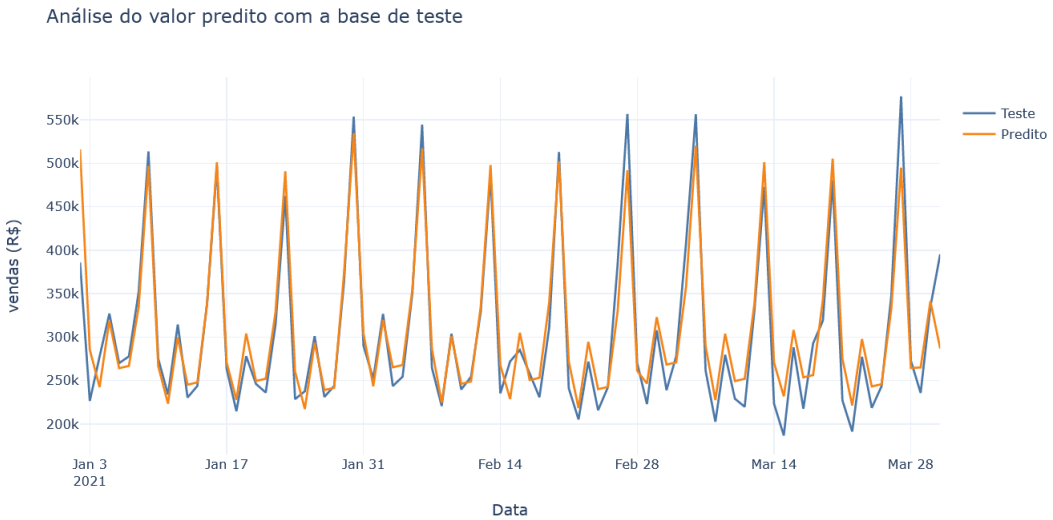
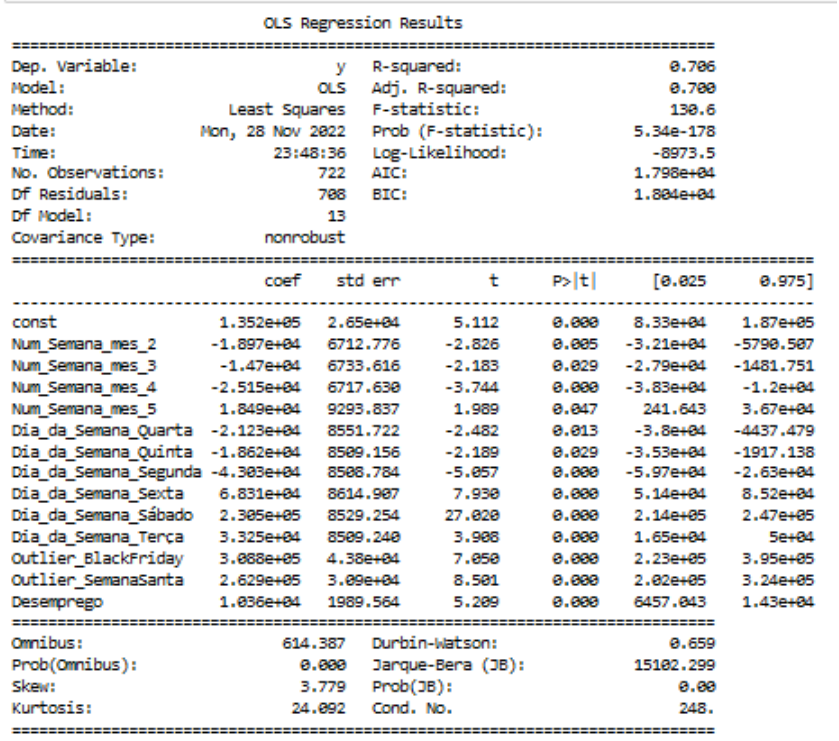


Figura 11

Os resultados da modelagem usando mínimos quadrados aparecem em seguida.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.706
Model:	OLS	Adj. R-squared:	0.700
Method:	Least Squares	F-statistic:	130.6
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	5.34e-178
Time:	23:48:36	Log-Likelihood:	-8973.5
No. Observations:	722	AIC:	1.798e+04
Df Residuals:	708	BIC:	1.804e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.352e+05	2.65e+04	5.112	0.000	8.33e+04	1.87e+05
Num_Semana_mes_2	-1.897e+04	6712.776	-2.826	0.005	-3.21e+04	-5790.507
Num_Semana_mes_3	-1.47e+04	6733.616	-2.183	0.029	-2.79e+04	-1481.751
Num_Semana_mes_4	-2.515e+04	6717.630	-3.744	0.000	-3.83e+04	-1.2e+04
Num_Semana_mes_5	1.849e+04	9293.837	1.989	0.047	241.643	3.67e+04
Dia_da_Semana_Quarta	-2.123e+04	8551.722	-2.482	0.013	-3.8e+04	-4437.479
Dia_da_Semana_Quinta	-1.862e+04	8509.156	-2.189	0.029	-3.53e+04	-1917.138
Dia_da_Semana_Segunda	-4.303e+04	8508.784	-5.057	0.000	-5.97e+04	-2.63e+04
Dia_da_Semana_Sexta	6.831e+04	8614.907	7.930	0.000	5.14e+04	8.52e+04
Dia_da_Semana_Sábado	2.305e+05	8529.254	27.020	0.000	2.14e+05	2.47e+05
Dia_da_Semana_Terça	3.325e+04	8509.240	3.908	0.000	1.65e+04	5e+04
Outlier_BlackFriday	3.088e+05	4.38e+04	7.050	0.000	2.23e+05	3.95e+05
Outlier_SemanaSanta	2.629e+05	3.09e+04	8.501	0.000	2.02e+05	3.24e+05
Desemprego	1.036e+04	1989.564	5.209	0.000	6457.043	1.43e+04

Omnibus:	614.387	Durbin-Watson:	0.659
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15102.299
Skew:	3.779	Prob(JB):	0.00
Kurtosis:	24.092	Cond. No.	248.

Figura 12

O Modelo 4 considerando semana do mês, dia da semana, outlier *black friday*, *outlier* semana santa e desemprego como atributos(X_i) para predizer o total vendido(Y) resultou em $R^2 = 0{,}706$ e $R^2_{ajust} = 0{,}70$, ou seja, o modelo consegue explicar 70% da varibilidade das vendas.

Ao nível de significância de 5% todas as variáveis são significativas.

Validação Cruzada

Para avaliar a capacidade de generalização do Modelo 4, foi utilizado o método de validação cruzada *K-Fold*.

O método consiste em dividir o modelo K-vezes, onde o modelo usa $K-1$ partes para treinar e a parte restante para validar.

Para testar a capacidade de generalização do modelo anterior foi considerado $K = 10$.

Foi obtido que o verdadeiro *score* que o modelo está generalizando é K-Fold = 0.72, com média= 0.73 e desvio padrão= 0.12.

3.8.2 Séries Temporais

Na modelagem foram considerados dados apartir de 2019, sendo 2/3 da base para treinamento do modelo e 1/3 para teste.

Na análise exploratória não foi observada nenhuma presença de tendência de alta nem de baixa, mas foi identificada presença de sazonalidade semanal.

Foram considerados modelos de séries temporais que levam em consideração sazonalidade sendo eles os modelos SARIMA e SARIMAX.

Na aplicação SARIMAX foram consideradas as variáveis exógenas: feriado, pagamento, quinto dia útil, pagamento no dia 20 (vale), dia da *black friday* e sábado de aleluia.

Para mensurar a qualidade dos modelos foi utilizada a função *auto-arima*, que retorna o melhor modelo segundo o Critério de Informação de Akaike (AIC) ou o Critério Bayesiano de Schwarz(BIC).

Para escolha do melhor modelo foram ulilizadas as métricas R^2, MAE e RMSE.

Na tabela seguinte têm-se os melhores modelos segundo o critério AIC e BIC.

Modelo	R_2 treino	R_2 teste	MAE teste	RMSE teste
SARIMA(3,0,0)(0,1,2)$_7$	0,43	0,558	43823	80146
SARIMA(2,0,0)(0,1,2)$_7$	0,43	0,558	43759	80118
SARIMA(0,0,3)(0,1,1)$_{14}$	0,33	0,554	45173	80499

Modelo 1: Melhor modelo pelo critério AIC, sazonalidade igual a 7

Modelo 2: Melhor modelo pelo critério BIC, sazonalidade igual a 7

Modelo 3: Melhor modelo pelo critério BIC, sazonalidade igual a 14

A seguir é apresentada a análise do modelo SARIMA, que obteve a melhor predição dentre os modelos de séries temporais testados, SARIMA(3,0,0)(0,1,2)$_7$.

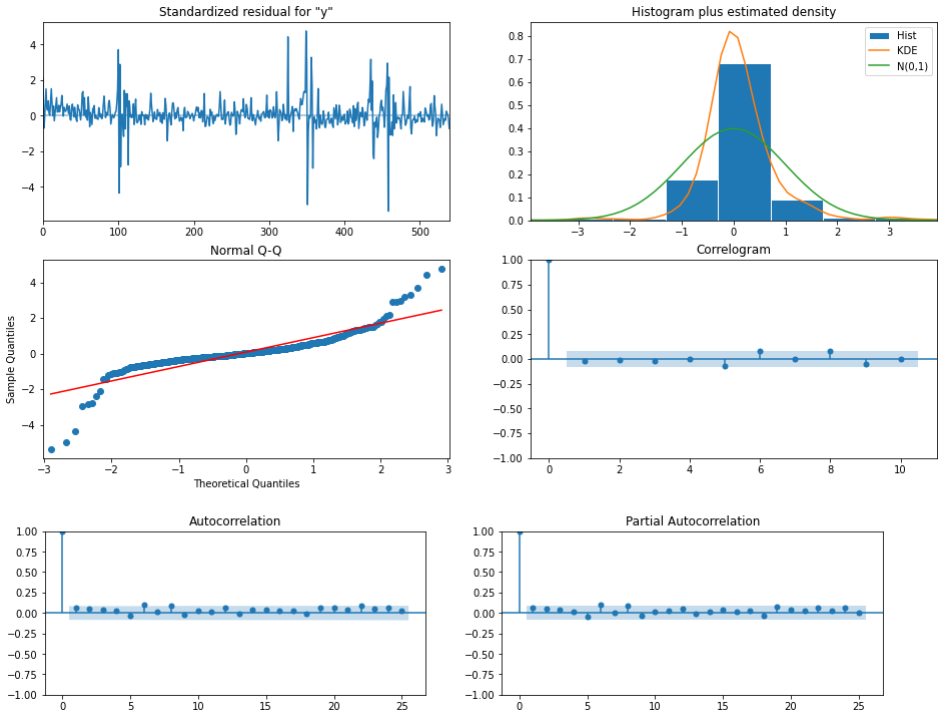


Figura 13: Avaliação dos resíduos base de treinamento modelo SARIMA(3,0,0)(0,1,2)$_7$

Com intuito de checar a normalidade e a independência dos resíduos, formam feitos os testes de Shapiro-Wilk e Ljung-Box.

Teste de Shapiro-Wilk

Utilizado para avaliar se uma distribuição é semelhante a uma distribuição normal, tal que:

H_0: Os erros seguem distribuição normal

H_1: Os erros não seguem distribuição normal

Caso o p-valor seja menor que o nível de significância 0.05 há evidências que os erros não seguem distribuição normal, caso contrário temos que os erros se assemelham a uma distribuição normal.

Aplicando o teste para os resíduos do modelo SARIMA$(3,0,0)(0,1,2)_7$ foi obtido p-valor de aproximadamente zero, ou seja, os erros não seguem distribuição normal. A mesma conclusão também pode ser observada nas imagens do *Histogram plus estimated density* e na Normal Q-Q.

Teste Ljung-Box

O teste examina m auto-correlações entre os resíduos. Se as auto-correlações forem muito pequenas, conclui-se que o modelo não exibe falha significativa de ajuste.

H_0: O modelo não exibe falha de ajuste (resíduos independentes)

H_1: O modelo exibe falha de ajuste (resíduos dependentes)

Aplicando o teste para os resíduos do modelo SARIMA$(3,0,0)(0,1,2)_7$ foi obtido p-valor=0.13, ou seja, o modelo não exibe falha de ajuste (independentes).

Nas imagens *Autocorrelation* e *Partial Autocorrelation* aparentemente todos os resíduos encontram-se dentro do limite de Bartlett dando indícios de independência dos resíduos.

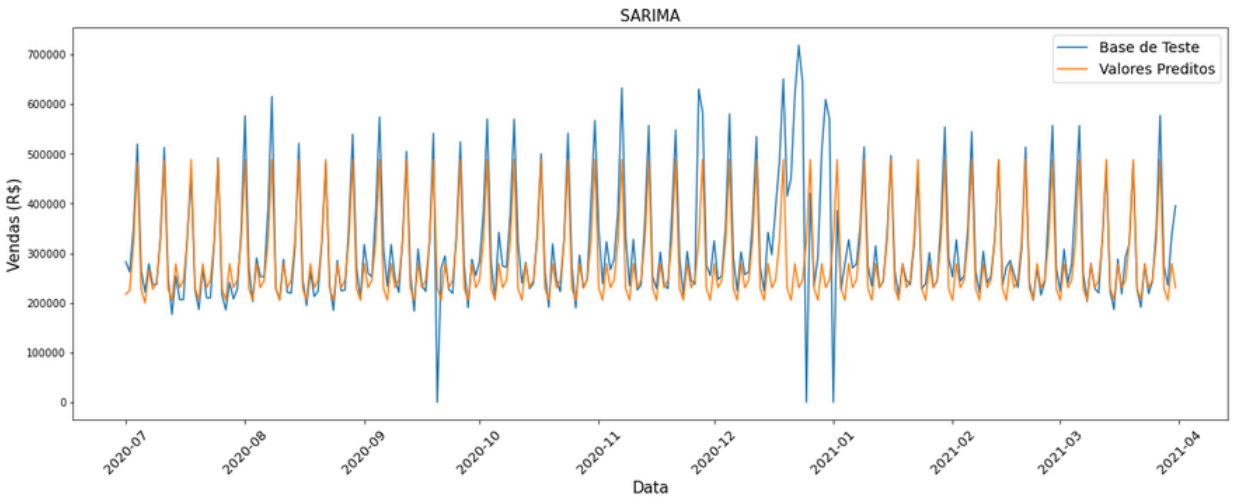


Figura 14

É importante ressaltar que o modelo acima não é capaz de fazer uma boa predição das vendas da loja de cosméticos, pois apresentou acurácia baixa. Na Figura 13 percebe-se que o modelo não foi capaz de predizer bem os picos.

A modelagem pelo SARIMAX resultou que nenhuma das variáveis exógenas são significativas para predizer as vendas da loja de cosméticos, além disso apresentou acurácia bem baixa, motivo pelo qual não foram apresentados os valores.

3.8.3 Prophet

Dado que o *Prophet* modela levando em consideração a sazonalidade anual, mensal e semanal e também os feriados do Brasil e datas específicas (personalizadas), como por exemplo dia de *black friday*, além de

apresentar boa performance com dados diários e ter um tempo de execução rápido, torna-se uma aposta de se ter um bom modelo preditivo para prever o total vendido.

Para a aplicação *Prophet*, foram considerados dados apartir de 2019, sendo 2/3 das observações utilizadas para treinamento e 1/3 para teste.

O modelo que apresentou melhor acurácia leva em consideração a sazonalidade semanal, anual, feriados nacionais do Brasil e tendência ajustada de forma a se tornar menos flexível.

Abaixo segue a acurácia da base de teste.

Modelo	MAE teste	RMSE teste
Prophet	37264	62326

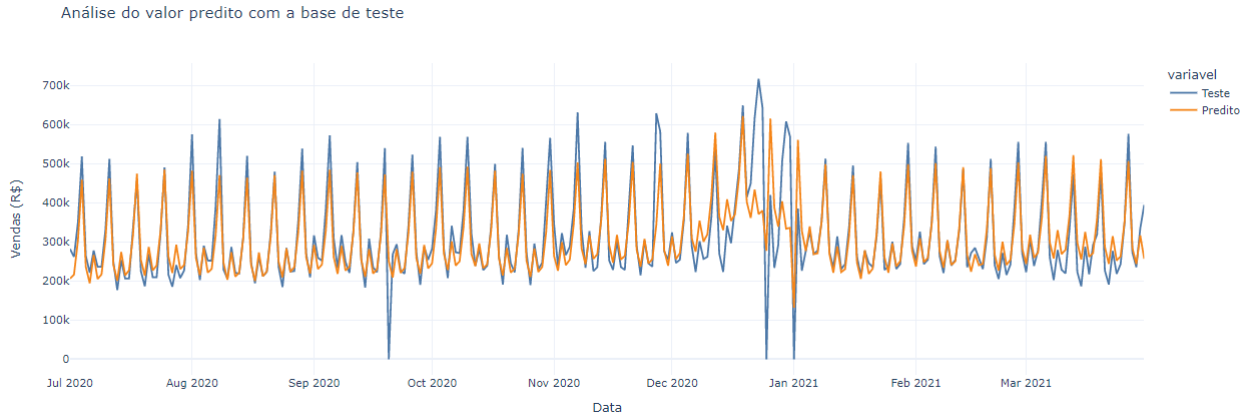


Figura 15

É possível observar que o modelo não conseguiu capturar alguns picos, sendo aconselhável testar a adição de outras informações.

3.8.4 DeepAR

Pela capacidade de se adaptar a uma arquitetura de rede neural recorrente, baseada em LSTM semelhante ao problema de previsão probabilística e pela vantagem de treinar várias centenas ou milhares de séries temporais simultaneamente, oferece potencial escalabilidade de modelo significativa e acaba sendo um modelo mais complexo comparado ao *Prophet*. O fato de ser um modelo mais complexo faz com que o processo de execução como um todo tenha um custo computacional mais elevado, comparado aos modelos citados anteriormente.

Para a codificação em *Python*, foi utilizada a *API Apache MXNet* além de diversas bibliotecas e pacotes do *Python*, destacando-se o uso do pacote *GluonTS*. *GluonTS* é um pacote de Python para modelagem probabilística de séries temporais, focado em modelos de aprendizado profundo.

Para o início da modelagem é necessário primeiramente definir os hiperparâmetros do *DeepAR* são eles:

- **"prediction_length":** O número de etapas de tempo que o modelo é treinado para prever.
- **"contest_length":** Número de períodos de tempo que o modelo levará em consideração para previsão.
- **"epochs":** Número máximo de passagens para examinar os dados de treinamento.
- **"num_cells":** Número de células que serão usadas dentro da rede neural recorrente, valor este que é sempre um inteiro positivo, variando de 30 a 100, a fim de se obter um resultado satisfatório.

- **"num_layers":** Número de camadas ocultas da RNN, é sempre um valor inteiro positivo, com valor padrão 2.

Na modelagem foram definidos os seguintes hiperparâmetros:

- **"prediction_length"**= 273
- **"contest_length"**= 547
- **"epochs"**= 300
- **"num_cells"**= 40
- **"num_layers"**= 2

Foram considerados dados a partir de 2019, sendo 2/3 das observações utilizadas para treinamento e 1/3 para teste.

A modelagem forneceu várias métricas para a base de teste. Em seguida são apresentados os valores de duas delas: MAPE e RMSE.

Modelo	MAPE teste	RMSE teste
DeepAR	334131	81075

Na Figura 15, tem-se a previsão probabilistica pela modelagem DeepAR. Fornece uma estimativa de quão confiável é o modelo e permite que decisões posteriores com base nessas previsões levem em consideração essa incerteza.

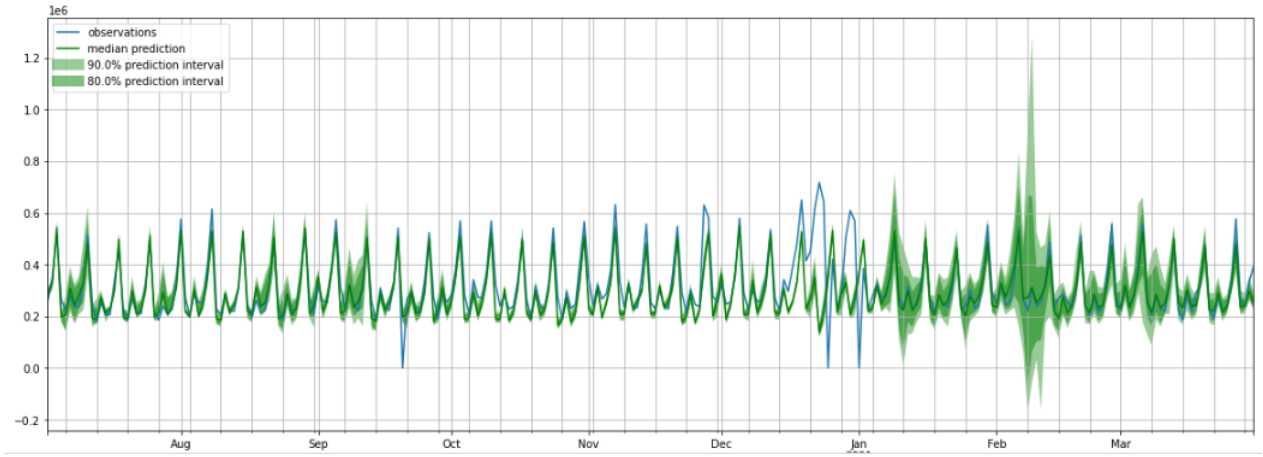


Figura 16

4 Conclusão

Após testar vários modelos conclui-se que o modelo dentre os testados capaz de fazer a melhor predição do total vendido foi o de regressão múltipla, com as variáveis explicativas sendo semana do mês, dia da semana, *outlier black friday*, *outlier* semana santa e desemprego.

Os resultados apresentam uma perspectiva interessante pois o método com melhor desempenho foi justamente o mais simples deles. A explicação para este fenômeno está no tipo de tarefa e de dados acessados por cada algoritmo. O método de regressão múltipla leva todas as variáveis preditivas em consideração e estas se mostraram poderosas para a obtenção de uma boa previsão. Já abordagens como *Prophet*, *DeepAR* e afins

levam em conta apenas o histórico de vendas. Mesmo sendo ferramentas poderosas, os resultados ilustram bem o quanto o tipo de dado que é analisado acaba tendo importância muito maior do que o algoritmo em si.

REFERÊNCIAS:

1. João Florindo, Aprendizado de Máquinas (Notas de aula e Slides).
2. Marketing por dados, acesso em 05/12/2022, https://marketingpordados.com/analise-de-dados/modelos-preditivos-para-que-servem-e-quais-sao-os-diferentes-tipos/
3. acesso em 05/12/2022, https://kekayan.medium.com/forecasting-with-deepar-for-busy-people-ed67f9d9a00d
4. Thiago Ildeu Albuquerque Lira, Modelos Neurais para Regressão em Séries Temporais, 2020
5. acesso em 05/12/2022, https://erico-coutojr.medium.com/modelagem-com-prophet-9b96b81bec0
6. Viniboscoa, acesso em 05/12/2022, https://www.viniboscoa.dev/blog/prophet-prevendo-o-futuro-em-series-temporais
7. Viniboscoa, acesso em 05/12/2022, https://www.viniboscoa.dev/blog/prophet-prevendo-o-futuro-em-series-temporais
8. acesso em 05/12/2022, https://towardsdatascience.com/prophet-in-a-loop-a875516ef2f9