2º EDSBD
2º Encontro de Data Science & Big Data · UFPR
7 de Julho de 2023 Campus Jardim Botânico · UFPR · Curitiba/PR dsbd.leg.ufpr.br

Explorando as dificuldades da aplicação de dados do mundo real na previsão de volume de chuva com dados de radares meteorológicos

André Vitor Kuduavski[1], Luiz E. S. Oliveira[2]

[1]Aluno do programa de Especialização em Data Science & Big Data, *andre.kuduavski@ufpr.br*;
[2]Professor do Departamento de Estatística - DEST/UFPR, *luiz.oliveira@ufpr.br*.

Resumo

Este trabalho descreve o processo de análise de regressões utilizando dados reais e aborda os desafios e soluções encontrados ao lidar com esse tipo de informação. O estudo utiliza um conjunto de dados coletados por radares meteorológicos no estado do Paraná, com o objetivo de prever o volume de chuva. Foram realizadas análises exploratórias, tratamento de dados ausentes, identificação e remoção de outliers, além do balanceamento. Em seguida, foram aplicados seis modelos de regressão para estimar o volume de chuva. Os resultados foram avaliados utilizando a métrica RMSE (Root Mean Squared Error).

Palavras chaves: *Dados reais, regressões, imputação de dados, chuva.*

Introdução

O objetivo principal desse trabalho é descrever o caminho e os desafios ao utilizar dados reais para realização de análises de regressões. Mostrar possíveis soluções para problemas específicos relacionados a natureza dos dados e no final medir os resultados após as tratativas implementadas. Como *case* desse estudo foram utilizados dados de radares meteorológicos distribuídos pelo estado do Paraná fornecidos pelo SIMEPAR (Sistema de Tecnologia e Monitoramento Ambiental do Paraná).

Figura 1: Resumo dos dados

ID	EST	TIME	AZIMUTH	RANGE	UH	UV	DBZH	DBZV	KDP	ZDR	RHOHV	X	Y	Z	LAT	LON	ALT	TP_EST	DISTANCIA
0	Loanda	09/01/2018 12:30	10	220250	null	null	null	null	null	-0.72	0.4	38227.34	216798.04	4775.95	-22.92	-53.15	4895.95	2.5	219
1	orto_Formos	09/01/2018 12:30	15	73750	-8.03	null	-9.92	null	-0.5	2.41	0.54	19085.25	71227.13	963.36	-24.23	-53.33	1083.35	0	74
2	Paranavai	09/01/2018 12:30	29	226750	6	5.17	4.74	3.38	1.84	1.33	0.68	109874.72	198219.24	5003.38	-23.09	-52.44	5123.38	0.2	230
3	ampo_Moura	09/01/2018 12:30	52	144750	null	null	null	null	null	5.56	0.47	114032.22	89091.74	2496.08	-24.07	-52.4	2616.08	0.35	152
4	Ubirata	09/01/2018 12:30	57	66750	null	-8.08	null	null	null	-4.25	0.49	55974.14	36350.03	844.52	-24.55	-52.97	964.51	0	71

Figura 2: Amostra dos dados

Material e Métodos

► Imputação de dados

Durante a exploração dos dados foi possível identificar que cerca de 87% dos registros possuíam ao menos uma coluna com dados faltantes. Para poder seguir com o objetivo do trabalho foi necessário analisar estratégias para tratar esse problema, e assim foram utilizadas técnicas de **imputação de dados** que resultaram em um aumento de 70% nos dados elegíveis para as próximas etapas.

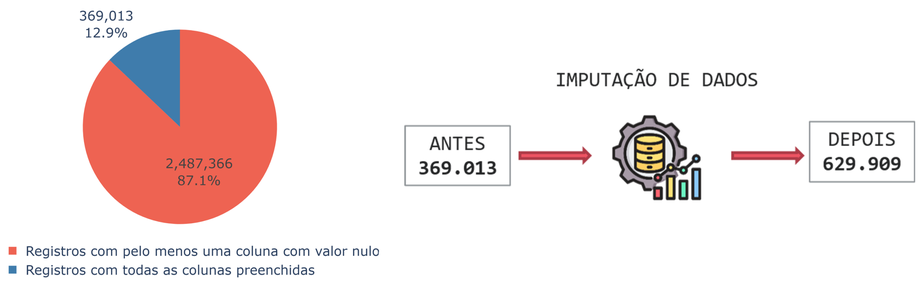


Figura 3: Resultado imputação de dados

► Outliers e Desbalanceamento

Outros tratamentos aplicados na base de dados foi em relação a *outliers* e desbalanceamento. Para os outliers foi aplicado uma filtragem na consulta aos dados e para o desbalanceamento foi implementado um algoritimo de oversampling.

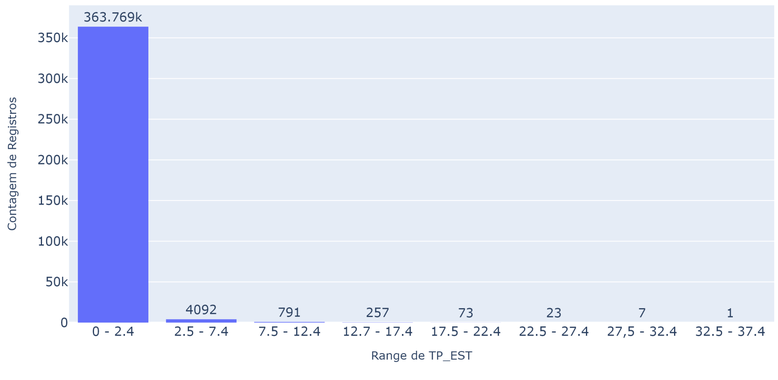


Figura 4: Desbalanceamento nos dados

O algorítimo de *oversampling* tem o objetivo de reduzir o número de observações da classe majoritária e foi desenvolvido em duas etapas, a primeira faz o balanceamento na base completa de uma só vez e a segunda faz balanceamentos separados para cada estação meteorológica.

► Fluxo Final

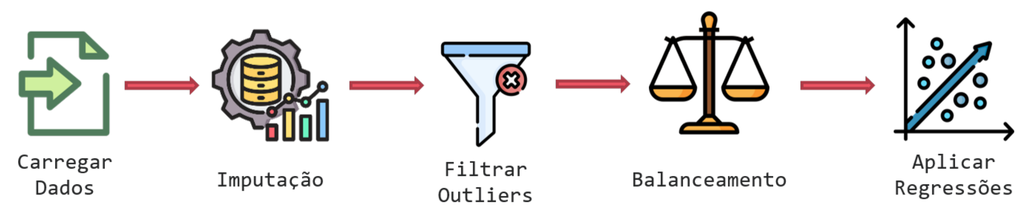
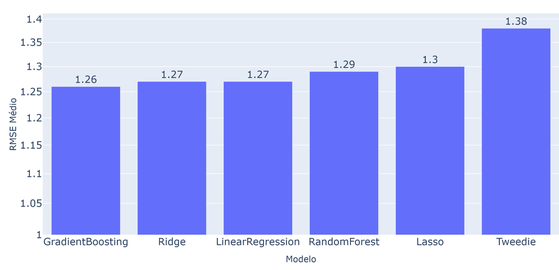


Figura 5:Fluxo final

Resultados e discussões

► Principais resultados

Após a execução do fluxo final foram obtidos dois resultados, um avaliando os modelos em todos os registros e outro aberto por estação de coleta dos dados. Como métrica foi utilizado a média de RMSE de 30 execuções para cada cenário.



estacao	melhorRegressao	rmseMedio
Aguas_do_Vere	Ridge	1.17
Altonia	Ridge	1.1
Assis_Chateaubriand	Ridge	0.93
Baixo_Iguacu	GradientBoosting	1.15
Bela_Vista_Jusante	GradientBoosting	0.5
Boa_Vista_da_Aparecida	Ridge	1.03
Campo_Mourao	Ridge	1.28
Coronel_Domingos_Soares	Ridge	1.13
Derivacao_do_Rio_Jordao	RandomForest	0.49
Foz_do_Iguacu__Itaipu	Ridge	1.12
Guaira	Ridge	0.85
Laranjeiras_do_Sul	Ridge	1.01
Loanda	Ridge	1.22
Palotina	Ridge	0.95
Paranavai	Ridge	1.12
Pato_Branco	RandomForest	0.82
Porto_Formosa	Ridge	1.23
Porto_Santo_Antonio	GradientBoosting	1.43
Reservatorio_Salto_Caxias	Ridge	0.78
Salto_Caxias	Ridge	1.27
Santa_Helena	Ridge	1.13
Segredo	Ridge	1.1
Solais_Novo	Ridge	1.08
Toledo	Ridge	0.95
Ubirata	Ridge	1.23
Umuarama	Ridge	0.71

Figura 6:Resultados

Para a execução com todos os dados o Gradient Boosting teve menor RMSE e na execução por estações foram obtidios um melhor resultado para cada estação. Ao comparar os dois cenários foi possível afirmar que existem modelos que melhor se ajustam aos dados quando realizado execuções separadas para cada estação meteorológica.

Conclusões

Neste trabalho, foram abordados os desafios enfrentados ao utilizar dados reais para análises de regressões, e após a identificação de problemas foram aplicadas técnicas de imputação de dados, filtragem de *outliers* e *undersampling*. Em seguida, foram realizadas regressões utilizando seis modelos diferentes e os resultados foram avaliados utilizando a métrica RMSE. Este estudo contribui para a compreensão dos processos envolvidos na análise de regressões com dados reais e fornece *insights* sobre o desempenho dos modelos utilizados.

Principais Referências

SIMEPAR, Sistema Meteorológico do Paraná, http://www.simepar.br

YNOUE, Rita Yuri; BOIASKI, Nathalie T.; DA SILVA, Gyrlene A. M.; AMBRIZZI, Tércio; REBOITA, Michelle S. Previsão de Clima. Artigo. Licenciatura em Ciências, USP/Univesp. Disponível em: https://midia.atp.usp.br/plc/plc0009/impressos/plc0009$_12.pdf$.

Agradecimentos

Aos professores do curso de DSBD e especialmente ao Prof. Luiz Eduardo S. Oliveira que auxiliou na escolha do tema e prestou apoio durante o desenvolvimento.