12º Workshop em Engenharia e Tecnologia Espaciais
6, 7, 13 e 14 de novembro de 2021

Predição de dados de telemetria de satélite com o uso de métodos *ensemble*

Ivan Márcio Barbosa [1], Maurício Gonçalves Vieira Ferreira[1], Milton de Freitas Chagas Júnior[1]

[1]Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

Aluno de Doutorado do curso de Engenharia e Gerenciamento de Sistemas Espaciais-CSE.

[1]Coordenação de Rastreio, Controle e Recepção de Satélites - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

[1]Serviço de Relações Institucionais - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

ivan.barbosa@inpe.br

Resumo. *Esse trabalho de pesquisa utiliza o aprendizado de máquina para predição de valores da telemetria TM072 - Battery temperature 1 do satélite de coleta de dados operado pelo INPE. A metodologia utilizada foi a análise bibliográfica sobre ciência de dados, estatística, matemática, aprendizado de máquina etc., a leitura dos dados de telemetria, a análise exploratória desses dados, a criação, o treinamento e a validação do modelo de aprendizado de máquina com o uso do método ensemble do tipo averaging (bagging) e do algoritmo ranfom forest. O modelo final obteve bons resultados com coeficiente de correlação R^2 98.16%, erro médio absoluto (MAE) de 0.167565 e raiz do erro quadrático médio (RMSE) de 0.311292.*

Palavras-chave: Machine learning; Métodos *ensemble;Bagging*

1. Introdução

O volume de dados de telemetria gerados diariamente é muito grande e tende a crescer proporcionalmente ao tempo de vida útil da missão espacial e do número de instrumentos a bordo dela.

Como o volume e a variedade dos dados crescem exponencialmente, é necessária a utilização da Ciência de Dados que, segundo (BOSCHETTI; MASSARON, 2016), é um domínio de conhecimento relativamente novo que requer a integração bem-sucedida da álgebra linear, modelagem estatística, visualização, linguagem de programação, análise de gráficos, aprendizado de máquina, inteligência de negócios, armazenamento e a recuperação de dados.

Considerando que os dados de telemetria do satélite de coleta de dados do INPE possuem alta dimensionalidade com milhões de amostras e 136 variáveis e que isso pode produzir resultados redundantes, provocar *overfitting* no modelo de aprendizado de máquina, exigir excessivo tempo de processamento e complexa análise dos dados etc., faz se necessária a redução da dimensionalidade desses dados.

A redução de dimensionalidade é a operação de eliminar alguns atributos ou variáveis do conjunto de dados de entrada e criar um conjunto restrito de recursos que contém todas as informações necessárias para prever a variável de destino de maneira mais eficaz e confiável. A redução do número de variáveis geralmente também reduz a variabilidade do resultado e a complexidade do processo de aprendizado (bem como o tempo necessário) (BOSCHETTI; MASSARON, 2016).

De acordo com (GOLLAPUDI, 2016), *ensemble methods* (do Inglês, métodos de conjunto) abrangem vários modelos que são construídos independentemente e os resultados desses modelos são combinados e são responsáveis pelas previsões. É fundamental identificar quais modelos independentes devem ser combinados ou incluídos, como os resultados precisam ser combinados e de que maneira alcançar o resultado desejado. Esta é uma classe de técnicas muito poderosa e amplamente adotada.

O método *ensemble* é uma técnica que combina as previsões de vários algoritmos de aprendizado de máquina para fazer previsões mais precisas do que qualquer modelo individual (BROWNLEE, 2020).

O objetivo dos métodos *ensemble* é combinar as previsões de vários estimadores com um dado algoritmo de aprendizado a fim de melhorar a generalização e a robustez de um único estimador (SCIKIT-LEARN DEVELOPERS, 2021).

Os métodos ensemble podem ser divididos basicamente em três tipos: *Averaging*, *Boosting* e *Stacking*.

No método *ensemble* do tipo *averaging* (do Inglês, cálculo da média), o princípio orientador é construir vários estimadores de forma independente e, em seguida, calcular a média de suas previsões. Em média, o estimador combinado é geralmente melhor do que o estimador base porque sua variância é reduzida (SCIKIT-LEARN DEVELOPERS, 2021).

Os métodos *ensemble* do tipo *averaging* podem ser subdivididos em: *pasting*, *bagging*, *subspaces* e *patches*.

O método *ensemble* do tipo *bagginng* (*bootstrap agregation*) é uma técnica que realiza a substituição (as amostras podem estar em diferentes conjuntos de dados de treinamento) aleatória de amostras com agregação. Contribui para redução da variância e é muito útil em modelos de aprendizado de máquina baseados em árvores de decisão. O método *bagging* trabalha de modo paralelo e é está disponível na biblioteca *scikit-learn,* tanto para problemas de regressão (*sklearn.ensemble.BaggingRegressor()*), quanto para problemas de classificação (*sklearn.ensemble.BaggingClassifier()*).

Segundo (SCIKIT-LEARN DEVELOPERS, 2021), os métodos *bagging* formam uma classe de algoritmos que constroem várias instâncias de um estimador caixa preta em subconjuntos aleatórios do conjunto de treinamento original e então agregam suas previsões individuais para

formar uma previsão final. Esses métodos são usados como uma forma de reduzir a variância de um estimador base (por exemplo, uma árvore de decisão), introduzindo a randomização em seu procedimento de construção e, em seguida, fazendo um novo conjunto a partir dela.

Como exemplo de métodos *ensemble* do tipo *averaging* podemos citar: *Random Forest*, *Extremely Randomized Trees* etc.

Para predição da variável dependente (*y*) TM072 será utilizado o algoritmo *Random Forest* (do Inglês, florestas aleatórias). O *random forest* é baseado no método ensemble *averaging* do tipo *bagging*, é executado de modo paralelo e é utilizado no aprendizado supervisionado. É um algoritmo de aprendizado de máquina não paramétrico, poderoso e popular, sendo aplicado tanto em problemas de classificação (*sklearn.ensemble.RandomForestClassifier()*), como em problemas de regressão (*sklearn.ensemble.RandomForestRegressor()*).

A abordagem *random forest* é um método *bagging* em que árvores profundas, ajustadas em amostras de *bootstrap*, são combinadas para produzir uma saída com menor variância. No entanto, *random forest* também usa outro truque para tornar as várias árvores ajustadas um pouco menos correlacionadas umas com as outras: ao crescer cada árvore, em vez de apenas amostrar as observações no conjunto de dados para gerar uma amostra de *bootstrap*, também amostramos os atributos e mantemos apenas um subconjunto aleatório deles para construir a árvore (ROCCA, 2019).

De acordo com (SCIKIT-LEARN DEVELOPERS, 2021), o *random forest* é um meta estimador que ajusta um conjunto de árvores de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e o controle do *overfitting*. O tamanho da subamostra é controlado através do parâmetro *max_samples* se *bootstrap* = *True* (padrão), caso contrário, todo o conjunto de dados é usado para construir cada árvore (SCIKIT-LEARN DEVELOPERS, 2021).

Após a criação de modelo de aprendizado de máquina é necessário sabermos a performance e a qualidade do modelo criado e o quanto esse modelo é capaz de prever novos valores ou classificar um conjunto de dados.

De acordo com (DANGETI, 2017), a avaliação de qualquer modelo precisa ser calculada para determinar a qualidade do modelo em relação aos dados reais, para que seu desempenho possa ser aprimorado ajustando os hiper parâmetros e assim por diante. De fato, a precisão de todo o algoritmo de aprendizado de máquina é medida com base em seu tipo de problema. No caso de problemas de classificação, utiliza-se a matriz de confusão e nos problemas de regressão, utiliza-se o erro médio quadrático (MSE) ou os valores ajustados do R^2.

Na regressão, uma variável contínua é prevista e as métricas de erro são obtidas comparando as previsões dos modelos com os valores reais das variáveis de destino e calculando o erro médio (GOLLAPUDI, 2016).

Nesse trabalho de pesquisa serão utilizadas as seguintes métricas:

a) *Mean Absolute Error* – MAE (do Inglês, erro absoluto médio)

A média obtida entre os valores originais e os valores previstos é chamada de erro absoluto médio. Ele também mede a magnitude média do erro, ou seja, o quão distante as previsões estão

dos valores reais. Além disso, o MAE não nos fornece nenhuma direção de erro, isto é, se estamos com overfitting ou undefitting nos dados (SHASHWAT TIWARI, 2019).

Segundo (GOLLAPUDI, 2016), o MAE é mais intuitivo e menos sensível a valores discrepantes.

b) Root Mean Squared Error (do Inglês. raiz do erro quadrático médio)

A raiz do erro quadrático médio é a métrica mais popular utilizada nos problemas de regressão. O RMSE é definido pelo desvio padrão dos erros de previsão. Esses erros de previsão são algumas vezes chamados de residuais. Os resíduos são basicamente a medida da distância dos pontos de dados da linha de regressão (SHASHWAT TIWARI, 2019).

Segundo (GORAKALA; USUELLI, 2015), o RMSE é o desvio padrão da diferença entre a classificação real e a classificação prevista.

c) R2

R2 é uma medida estatística de quão perto o ponto de dados é ajustado à linha de regressão. R2 é definido pela variância explicada dividida pela variância total que é explicada pelo modelo linear (SHASHWAT TIWARI, 2019).

O valor de R2 sempre fica entre 0% e 100%, portanto, 0% indica que não há variabilidade dos dados de resposta em torno de sua média e 100% indica como o modelo explica toda a variabilidade dos dados de resposta em torno de sua média. Isso significa claramente que um modelo com valor R2 mais alto é perfeito para o seu modelo (SHASHWAT TIWARI, 2019).

De acordo com (DANGETI, 2017), em alguns casos extremos, o R2 também pode ter um valor menor que zero, o que significa que os valores previstos do modelo têm desempenho pior do que apenas tomar a média simples como uma previsão para todas as observações.

Nesse trabalho de pesquisa utilizaremos o aprendizado de máquina supervisionado, o método ensemble bagging e o algoritmo random forest para predição de valores da variável dependente (y) TM072 - Battery temperature 1.

2. Metodologia

A metodologia utilizada nesse trabalho de pesquisa foi a leitura e análise bibliográfica sobre ciência de dados, estatística, matemática e aprendizado de máquina. Também foram realizadas leituras de artigos científicos sobre aprendizado de máquina e inteligência artificial aplicados à área espacial, estudos sobre as bibliotecas *scikit-learn*, *matplotlib*, *pandas*, *numpy*, *plotly* e sobre métodos ensemble *averaging* (*bagging*), *boosting* e *stacking*.

Após a análise bibliográfica, foi realizada a aquisição do conjunto de dados de telemetria do satélite de coleta de dados, a análise exploratória dos dados (limpeza, transformação, visualização, imputação etc.) com a utilização das bibliotecas *pandas, numpy, plotly* e *matplotlib, a* padronização dos atributos com o *sklearn.preprocessing.StandardScaler* e a seleção de atributos com o método *sklearn.feature_selection.SelectKBest*.

Após a fase da análise exploratória dos dados, foi criado o modelo base de aprendizado de máquina com o uso do método *ensemble random forest*. As melhorias nesse modelo foram realizadas através dos hiper parâmetros (*criterion = 'mae', max_depth = 8, max_features =*

'sqrt', n_estimators = 500, n_jobs = -1, random_state = 42) que foram obtidos através do *sklearn.model_selection.GridSearchCV*. Com isso foi possível obter um modelo de aprendizado de máquina com R^2 melhor que o do modelo base. As métricas de qualidade e desempenho utilizadas durante a fase de validação do modelo preditivo com o método *random forest* foram: *MAE, RMSE* e R^2.

3. Resultados e Discussão

A telemetria TM072 que tem como descrição "*Battery temperature 1*" e que pertence ao equipamento PCU do subsistema PSS e possui valores aceitáveis entre -3°C a 26°C e acurácia de +/- 0.3°C será utilizada como variável dependente (y) e terá seus valores previstos pelo modelo de aprendizado de máquina com o uso do método ensemble *random forest*. A distribuição dos dados da variável dependente (y) TM072 é mostrada na Figura 1.

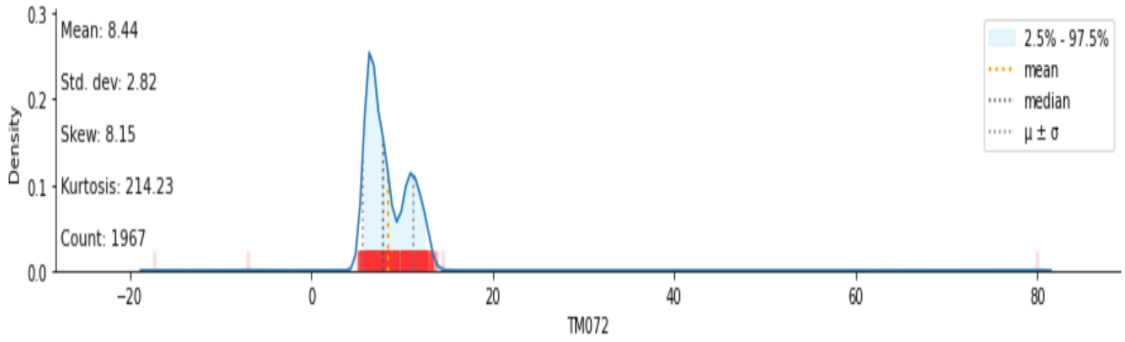


Figura 1. TM072 - *Battery temperature 1*.

Fonte: Próprio autor

Na Figura 1 é possível observar que a variável dependente (*y*) TM072 tem média de 8.44°C, desvio padrão 2.82 e que a maior parte das amostras estão entre 2.5% e 97.5%. A Figura 2 ilustra o gráfico de dispersão onde é possível observar os pontos fora da curva.

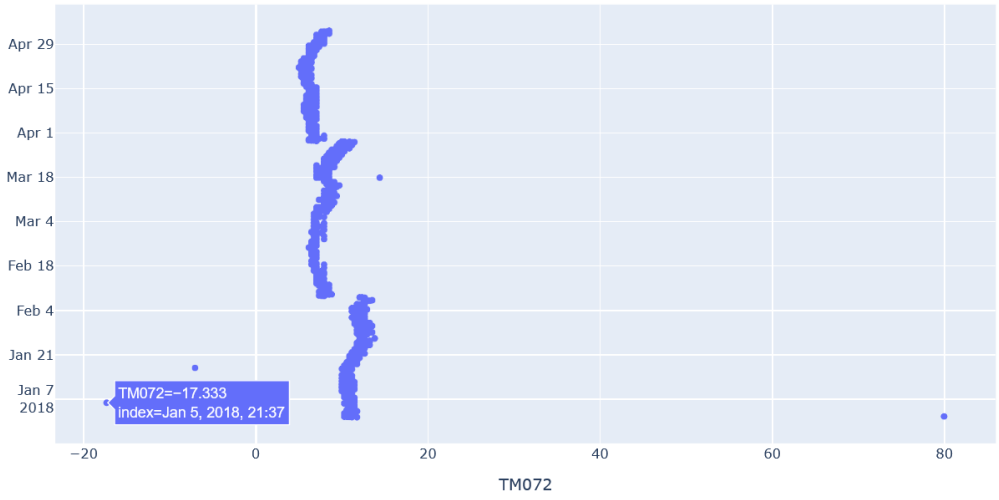


Figura 1. Gráfico de dispersão da variável dependente (*y*) TM072.

Fonte: Próprio autor

Como conjunto de dados de telemetria do satélite de coleta de dados possui alta dimensionalidade, foi necessário realizar a seleção de atributos. A seleção de atributos foi feita com uso do método *Univariate Feature Selection sklearn.feature_selection.SelectKBest (mutual_info_regression e k = 16)*. A Tabela 1 ilustra os 16 atributos que foram selecionados após a execução do *SelectKBest*.

Tabela 1. Variáveis independentes (X).

Fonte: Próprio autor

Variável independente	Descrição
TM027	*Receiver AGC voltage*
TM081	*Battery temperature 2*
TM113	*DCP temperature monitor*
TM114	*MGE temperature monitor*
TM119	*OBC temperature monitor*
TM121	*ENC temperature monitor*
TM122	*PCU temperature monitor*
TM127	*BAT temperature monitor*
TM129	*RDU temperature monitor*
TM130	*SS1 temperature monitor*
TM131	*SS2 temperature monitor*
TM133	*CP sup. temperature monitor*
TM134	*BP temper. monitor (near BAT)*
TM135	*BP temper. monitor (near PCU)*
TM353	*Extra telemetry*
TM354	*Extra telemetry*

Após a padronização dos dados, a seleção dos atributos, a divisão dos dados em dados de treinamento com 1.475 amostras (75% do conjunto de dados) e dados de testes com 492 amostras (25% do conjunto de dados), foi criado o modelo de aprendizado de máquina com o método *ensemble random forest* com 16 atributos e 1.967 amostras em um notebook Dell Latitude com processado Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz e 08GB de memória RAM.

Inicialmente foi criado, treinado e testado um modelo de aprendizado de máquina de linha de base com o método *random forest*. Logo após foi utilizado *sklearn.model_selection.GridSearchCV* para obter melhores parâmetros (*criterion = 'mae', max_depth = 8,max_features = 'sqrt', n_estimators = 500, n_jobs = -1, random_state = 42*) para criação de um outro modelo de aprendizado de máquina com a métrica R^2 maior . A Tabela 2 ilustra os resultados alcançados pelo modelo de aprendizado de máquina *random forest* utilizado como linha de base e o modelo *random forest* final.

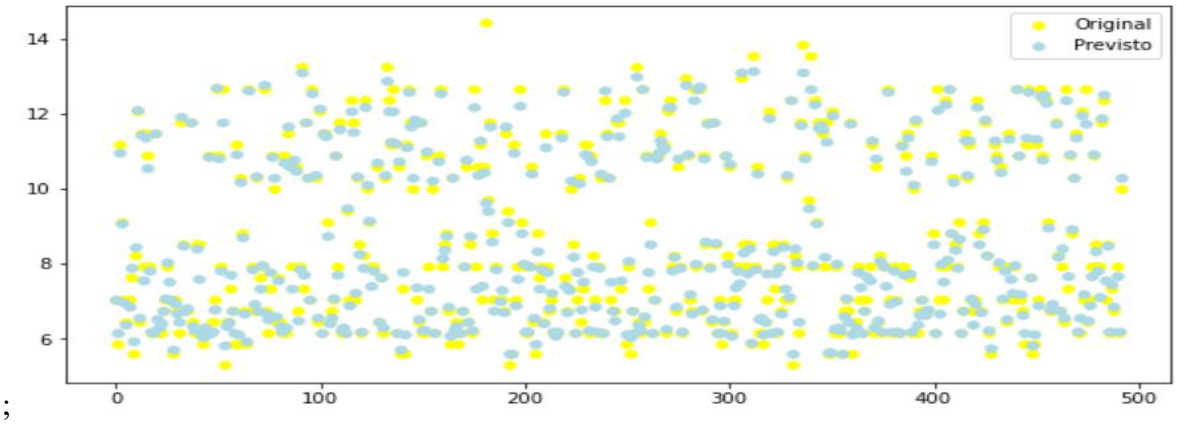
Tabela 2. Métricas de avalição do modelo

Fonte: Próprio autor

Random Forest	R^2	MAE	RMSE
Modelo base	0.969362	0.142217	0.377117
Modelo final	0.981649	0.167565	0.311292

Na Tabela 2 é possível observar que o modelo final obteve melhorias em relação ao modelo de linha de base. A melhoria no coeficiente de correlação R^2 e no RMSE foi em decorrência do uso do *sklearn.model_selection.GridSearchCV*. No entanto, o valor obtido na métrica MAE do modelo final obteve valor um pouco maior que o valor do modelo de aprendizado de máquina utilizado como linha de base.

A Figura 3 ilustra os dados de reais (*y_test*) e os dados de temperatura da bateria 1 previstos (*y_pred*) pelo modelo de aprendizado de máquina com o uso do método ensemble *random forest*.



;

Figura 2. Comparação entre os dados reais e os dados previstos.

Fonte: Próprio autor

Na Figura 2 é possível observar que os valores reais (y_test) da telemetria TM072- *Battery temperature 1 que* foram previstos (y_pred) estão bem próximos um do outro e que há aproximadamente 500 amostras nesse conjunto de dados com valores entre 5°C e 14°C.

4. Conclusão

A utilização do aprendizado de máquina pode trazer muitas contribuições para a operação contínua dos satélites do INPE. Conforme mostrado na Tabela 2, o modelo final de aprendizado de máquina resultou no coeficiente de determinação (R2) com 98.16% e RMSE com 0.311292. Com isso podemos considerar que o modelo de aprendizado de máquina criado com a utilização do métodos ensemble random forest atingiu um bom resultado na predição dos valores da temperatura da Bateria 1 do satélite de coleta de dados.

Referências

BOSCHETTI, A.; MASSARON, L. **Python Data Science Essentials**. Second ed. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016. 361 p. ISBN(978-1-78646-213-8).

BROWNLEE, J. **Bagging and Random Forest Ensemble Algorithms for Machine Learning**. Disponível em: <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>. Acesso em: 18 set. 2021.

DANGETI, P. **Statistics for Machine Learning**. First ed. BIRMINGHAM - MUMBAI: Packt Publishing Ltd., 2017. 424 p. ISBN(978-1-78829-575-8).

GOLLAPUDI, S. **Practical Machine Learning**. BIRMINGHAM - MUMBAI: BIRMINGHAM - MUMBAI, 2016. 261 p. ISBN(978-1-78439-968-9).

GORAKALA, S. K.; USUELLI, M. **Building a Recommendation System with R**. First ed. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2015. 135 p. ISBN(978-1-78355-449-2).

SCIKIT-LEARN DEVELOPERS. **Ensemble methods**. Disponível em: <https://scikit-learn.org/stable/modules/ensemble.html#>. Acesso em: 8 ago. 2019.

SHASHWAT TIWARI. **Complete Guide to Machine Learning Evaluation Metrics**. Disponível em: <https://medium.com/analytics-vidhya/complete-guide-to-machine-learning-evaluation-metrics-615c2864d916>. Acesso em: 25 jun. 2020.