EDA - Collision Detection Challenge

Nico I. G. Ramos GRR20210574

I. Introdução

Foi desafiador realizar uma análise exploratória de dados (EDA) no dataset escolhido para o projeto, de detecção de risco de colisão espacial, devido à sua complexidade e ao grande volume de dados. O conjunto de treino consiste em 162634 registros com 103 características cada. A alta variedade de características, em conjunto com a falta de conhecimento prévio sobre o domínio, dificultou a identificação de padrões, relações e insights relevantes.

O processo de EDA foi mais um processo iterativo de limpeza e redução de dimensionalidade considerando os insights parciais obtidos, do que uma análise direta dos dados. Esse processo foi necessário por conta da alta dimensionalidade do dataset, que dificultava a visualização e compreensão dos dados.

O relatório está organizado da seguinte forma:

- Primeiras linhas do dataset original e do dataset final;
- Descrição da preparação e da limpeza dos dados pré-análise;
- Etapas de redução de dimensionalidade considerando a dispersão e a correlação entre as features;
- Estatísticas descritivas do dataset;
- Insights;
- Divisão do dataset em treino, validação e teste.
- Apêndice com o significado das colunas mencionadas.

A. Visualização das Primeiras Linhas do Dataset

As primeiras linhas do dataset original são apresentadas na Tabela I, como ela possui muitas colunas, apenas uma parte delas é mostrada como exemplo. por conta disso, não foi possível tirar insights relevantes a partir dessa visualização preliminar.

Tabela I
Linhas iniciais do dataset original.

Coluna	Registro 1	Registro 2	Registo 3	Registro 4	Registro 5
event_id	0	0	0	0	0
time_to_tca	1.566798	1.207494	0.952193	0.579669	0.257806
mission_id	5	5	5	5	5
risk	-10.204955	-10.355758	-10.345631	-10.337809	-10.391260
max_risk_estimate	-7.834756	-7.848937	-7.847406	-7.845880	-7.852942
max_risk_scaling	8.602101	8.956374	8.932195	8.913444	9.036838
miss_distance	14923.000000	14544.000000	14475.000000	14579.000000	14510.000000
relative_speed	13792.000000	13792.000000	13792.000000	13792.000000	13792.000000
relative_position_r	453.800000	474.300000	474.600000	472.700000	478.700000
relative_position_t	5976.600000	5821.200000	5796.200000	5838.900000	5811.100000

II. Preparação e Limpeza de Dados

O objetivo da preparação e limpeza de dados foi reduzir a dimensionalidade do dataset, para começar a análise com um conjunto de dados mais manejável e compreensível.

O primeiro passo foi tentar identificar colunas que pudessem ser descartadas logo no início, como identificadores ou colunas com valores constantes ou pouco relevantes para o problema. Contudo, a o dataset praticamente não tem colunas identificadoras, e nenhuma coluna possuia um baixo desvio padrão quando comparado ao valor médio dela. Também foi feita uma rápida pesquisa sobre o significado de algumas das coluna, com a qual foi possível remover algumas que pareciam ter menos relevância inicial para o problema, como os valores já calculados de correlação e aqueles usados pelo algoritmo oficial de detecção de risco. Mesmo assim, não possível eliminar muitas colunas e o dataset permaneceu com uma alta dimensionalidade.

Durante as primeiras iterações, foi escolhido a remoção de colunas não numéricas, a única característica identificada foi a que representa o tipo do objeto espacial (satélite, lixo espacial, etc). Inicialmente, a coluna foi enumerada, mas não foi identificado grande relevância com o risco de colisão e foi optado por removê-la, o que também simplificou o processo de análise.

O desafio da alta dimensionalidade do dataset possui duas causas iniciais: a duplicidade de colunas entre o chaser e o target, são *80* colunas que representam a mesma característica para o chaser e para o target; e a presença de colunas que representam a mesma relação, tanto da perspectiva do target quanto do chaser. Por exemplo, as colunas do apogeu (14 e 15) e perigeu (16 e 17), que existem tanto para o chaser quanto para o target; e as colunas de correlação que representam a mesma relação entre velocidade e posição, mas de perspectivas distintas.

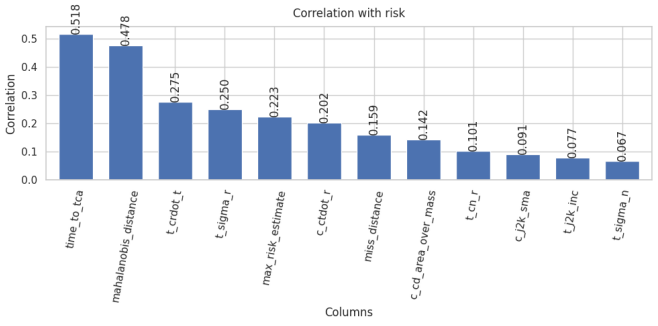
A escolha por remover as colunas de correlação foi feita para reduzir a dimensionalidade, dessas *80* colunas, *32* são de correlação entre métricas como velocidade relativa, posição, etc. A remoção dessas colunas não teve um impacto significativo na análise, já que essas informações podem ser inferidas se necessárias.

A. Filtragem Inicial

A filtragem inicial foi alterada iterativamente e experimentalmente ao longo do EDA. Como a reinclusão de colunas já descartadas, alterar o conjunto inicial de colunas descartadas e alteração nos valores de thresholds. A principal mudança entre as iterações foi a quantidade de colunas descartadas, devido a aos valores de thresholds. A mudança nas colunas descartadas inicialmente, no entanto, teve um impacto limitado, indicando que o processo de seleção foi robusto no que tange à importância e relevância percebida das colunas. Apesar de ter reduzido o dataset para *41* colunas, ainda não possível identificar padrões claros e relações relevantes entre as variáveis.

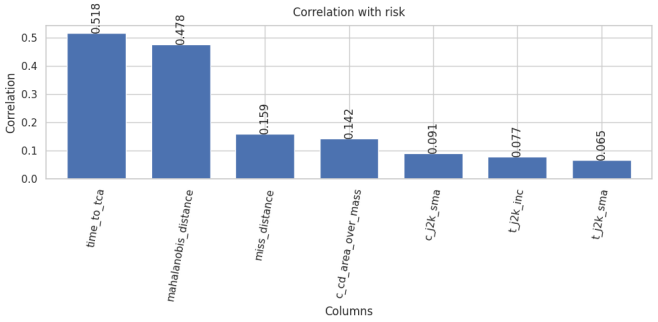
A Figura 1 mostra a correlação das colunas selecionadas ao final de todas as etapas com a variável alvo, usando o dataset original.

Figura 1. Correlação das colunas selecionadas com a variável alvo no dataset original.



A Figura 2 mostra a correlação com a variável alvo, mas realizando a filtragem inicial.

Figura 2. Correlação das colunas selecionadas com a variável alvo após a filtragem inicial.



É possível observar que a filtragem inicial removeu sete colunas, adicionou uma, e manteve quatro colunas.

As sete colunas removidas foram:

- *t_crdot_t* (1),
- *t_sigma_r* (2),
- *max_risk_estimate* (3),
- *c_ctdot_r* (4),
- *t_cn_r* (5),
- *t_j2k_inc* (12),
- *t_sigma_n* (6)

A coluna incluida foi:

- *t_j2k_sma* (13)

E as quatro colunas mantidas foram:

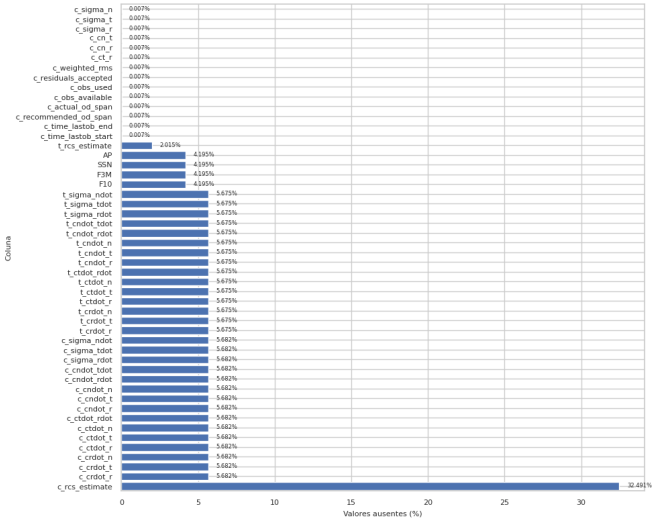
- *time_to_tca* (7),
- *mahalanobis_distance* (8),
- *miss_distance* (9) e
- *c_cd_area_over_mass* (10).

B. Tratamento de Valores Ausentes e Infinitos

O passo seguinte foi tratar os valores ausentes (NaN) e infinitos (+inf e -inf) no dataset original, para que eles não atrapalhassem a análise subsequente.

A Figura 3 mostra a porcentagem de valores ausentes em cada coluna numérica do dataset original.

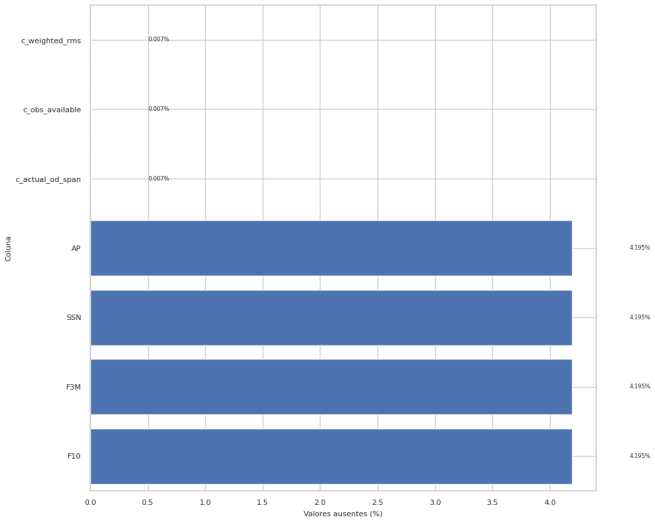
Figura 3. Porcentagem de valores ausentes por coluna no dataset original.



A Figura 4 mostra a porcentagem de valores ausentes em cada coluna numérica do dataset filtrado. É possível observar que a porcentagem de valores se divide em três grupos principais:

- *Trinta e quatro* colunas sem valores ausentes (não mostradas na figura);
- *Três* colunas com 0.007% de valores ausentes;
- *Quatro* colunas com 4.195% de valores ausentes;

Figura 4. Porcentagem de valores ausentes por coluna no dataset filtrado.



A escolha da estratégia para remover os valores ausentes considerou que: para o grupo com 0.007%, a quantidade é muito baixa; e, para o grupo com 4.195%, as colunas são valores que descrevem caracteristicas solares no momento da observação, e que provavelmente não seriam usados na série temporal de qualquer forma. Com isso, a escolha foi por uma abordagem simples, sem maiores análises, substituindo os valores ausentes pela média da coluna.

Não foram encontrados valores infinitos para o positivo ou para o negativo no dataset.

C. Tratamento de Outliers

III. Redução de Dimensionalidade

A. Remoção por Baixa Quantidade de Valores Únicos

B. Remoção por Baixa Dispersão

C. Remoção por Multicolinearidade (Alta Correlação)

D. Remoção por Baixa Correlação com o Alvo (Risk)

E. Remoção por Correlação Similar ao Alvo (Risk)

IV. Estatísticas Descritivas do Dataset

V. Insights

A. Análise das Colunas Selecionadas

Insights sobre as colunas removidas: (2, 3)

- *t_crdot_t* (1): grande quantidade de valores faltantes e pode ser inferida a partir de outras colunas mais confiáveis;
- *t_sigma_r* (2): representa a incerteza na posição radial do target, que é uma métrica derivada e pode não ser tão relevante quanto as características primárias do chaser e do target;
- *max_risk_estimate* (3): deveria ser removida de qualquer forma, pois ela representa o valor calculado pelo algoritmo oficial de detecção de risco, que é o que estamos tentando prever. Incluir essa coluna no modelo poderia levar a um vazamento de dados, onde o modelo aprende a prever o risco com base em uma variável que já contém essa informação;
- *c_ctdot_r* (4): também possui possui uma alta quantidade de valores faltantes e possui uma correlação com o risco, de 0.202 (2), próxima da correlação da coluna *miss_distance* (9), de 0.159, que é mais confiável por não possuir valores ausentes.
- *t_cn_r* (5): representa a componente normal da taxa de variação da distância entre o chaser e o target, que também pode ser derivada de outras colunas mais fundamentais;
- *t_sigma_n*: foi mantida a coluna *mahalanobis_distance* (8), que também representa uma medida de incerteza na órbita, e possui uma correlação muito maior com o risco. E a coluna *t_j2lk_sma* (13) foi inserida, ela possui uma correlação com o risco apenas um pouco menor, de 0.065 ao invés de 0.067, e carrega uma informação a mais sobre a órbita do target, que pode ser relevante para a predição da colisão;

Pela Figura 3, é possível observar que as colunas *t_crdot_t* (1) e *c_ctdot_r* (4) possuem mais de 5.6% de valores ausentes, o que representa uma alta quantidade de valores ausentes em relação aos outros dados do dataset, ficando atrás apenas da coluna *c_rcs_extimate* (18), com mais de 30% de valores faltantes.

Insights sobre as colunas mantidas: (2, 3)

- *time_to_tca* (7): essêncial pois forma a série temporal dos eventos;
- *mahalanobis_distance* (8): é a coluna com a maior correlação com o risco, de 0.478. Permite medir a incerteza da órbita, a confiabilidade dos dados. E evita redundância entre as caracterísitcas, pois possibilita a remoção de outras colunas de incerteza;
- *miss_distance* (9): é uma métrica direta da proximidade entre os dois objetos espaciais na aproximação máxima. Uma menor distância de aproximação está diretamente relacionada a um maior risco de colisão, tornando essa coluna essencial para a análise. Também possui uma correlação significativa com o risco, de 0.159 e auxilia a observar a órbita dos objetos.
- *c_cd_area_over_mass* (10): o coeficiente balístico do chaser influencia sua resistência ao arrasto atmosférico, afetando sua trajetória orbital. Apesar de sua correlação moderada com o risco, de 0.122 e de não variar ao longo do tempo (eventos da série temporal), fornece insights sobre como as características físicas do chaser impactam o risco de colisão.
- *c_j2k_sma* (11) e *t_j2k_inc* (12): representam elementos da órbita do chaser e do target, respectivamente. Essas colunas possuem correlações menores com o risco, de 0.065 e -0.049, mas podem ser importantes para entender a dinâmica orbital dos objetos.

As colunas mantidas encapsulam aspectos relevantes para a predição do risco de colisão. A coluna *mahalanobis_distance* (8) captura a incerteza sobre a órbita, caracterizada pelas colunas *c_j2k_sma* (11), *t_j2k_inc* (12), e c_cd_area_over_mass (10), enquanto a coluna *miss_distance* (9) fornece caracteriza a geometria da aproximação. A série temporal, representada pela coluna *time_to_tca* (7), une essas características, permitindo a análise da evolução do risco ao longo do tempo sob três aspectos principais: a órbita dos objetos envolvidos, a incerteza dela, e a geometria da aproximação.

B. Análise da Correlação das Variáveis com o Risco

C. Análise dos Padrões Temporais (Time-Series)

VI. Divisão do Dataset em Treino, Validação e Teste

APÊNDICE A
SIGNIFICADO DAS COLUNAS MENCIONADAS

1) *t_crdot_t*: Correlação entre a velocidade radial e a posição along-track do target.
2) *t_sigma_r*: Incerteza (desvio padrão) na posição radial do target.
3) *max_risk_estimate*: Estimativa máxima de risco calculada pelo algoritmo oficial de detecção de risco.
4) *c_ctdot_r*: Correlação entre a velocidade along-track e a posição radial do chaser.
5) *t_cn_r*: Componente normal da taxa de variação da distância entre o chaser e o target.
6) *t_sigma_n*: Incerteza (desvio padrão) na posição normal do target.
7) *time_to_tca*: Tempo até a aproximação máxima entre o chaser e o target.
8) *mahalanobis_distance*: Medida de incerteza da órbita, indicando a confiabilidade dos dados.
9) *miss_distance*: Distância mínima prevista entre o chaser e o target na aproximação máxima.
10) *c_cd_area_over_mass*: Coeficiênte balístico do chaser, relacionado à resistência ao arrasto atmosférico.
11) *c_j2k_sma*: Semi-eixo maior da órbita do chaser.
12) *t_j2k_inc*: Inclinação da órbita do target.
13) *t_j2k_sma*: Semi-eixo maior da órbita do target.
14) *t_h_apo*: Ponto mais distante da órbita do target em relação à Terra.
15) *c_h_apo*: Ponto mais próximo da órbita do chaser em relação à Terra.
16) *t_h_peri*: Ponto mais próximo da órbita do target em relação à Terra.
17) *c_h_peri*: Ponto mais próximo da órbita do chaser em relação à Terra.
18) *c_rcs_extimate*: Estimativa da seção transversal do radar do chaser.