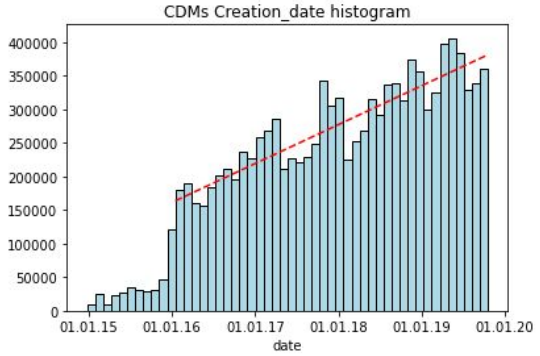
Learning from an imbalanced dataset of conjunction data messages

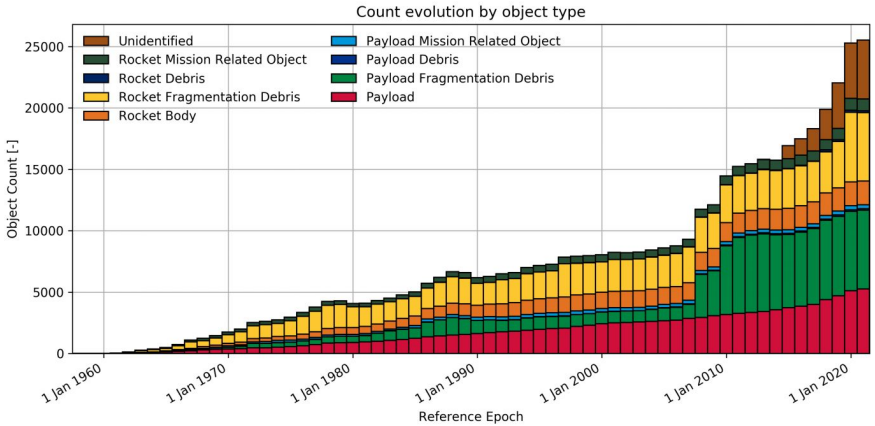
Workshop on Collision Avoidance Challenge Results

Rasit Abay

Introduction

- The number of close approaches between space objects is increasing due to the proliferation of satellites and debris in orbit.
- The number of CDMs issued weekly are a couple of orders of magnitude larger than the actionable ones, and this puts pressure on satellite operators.
- An automated process that can predict collision risk with desired accuracy a couple of days ahead to allow satellite operators to plan for collision avoidance maneuvers is desired.

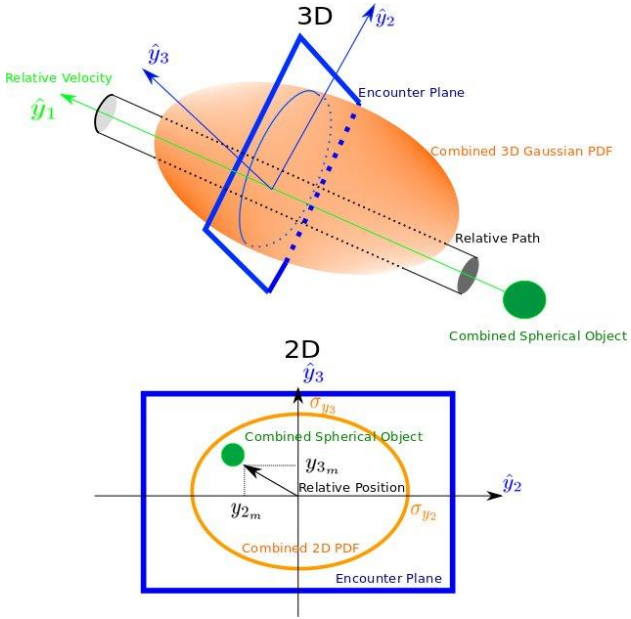




https://www.orbofleet.com/space-debris-space-environment-statistics/

Collision Risk Prediction Preliminaries

- The relative trajectory is assumed to be linear, and this is valid for short encounters.
- The physical properties are not modeled with high-fidelity, and the characterization of debris regarding physical properties is challenging.
- The interaction between space objects and the space environment are challenging to model, and there are unknown unknowns.



$$P_c = \frac{1}{\sqrt{(2\pi)^3 |C_c|}} \int_V e^{\left(-\frac{1}{2}(\Delta r_{ca} + \delta r_{ca})^T C_c^{-1} (\Delta r_{ca} + \delta r_{ca})\right)}$$

https://www.orbofleet.com/space-debris-space-environment-statistics/

Data

Data Source

- Preprocessed CDMs from **2015 to 2019** without full states;
- **13154** events;
- **103** features.

Data Cleaning

- Remove events with **only one CDM**

Imbalanced Dataset

- High-risk: **1.05**%;
- Low-risk: **98.9**%.

Feature Engineering

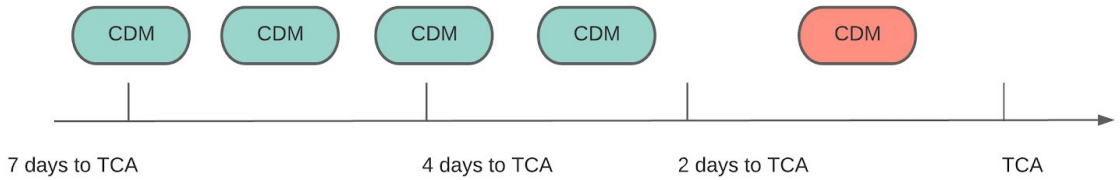
- Number of **CDMs bef. 2 days**;
- **Mean and STD** of PoC values;

Data

- Risk is the most important feature.
- Any attempt to learn from sequence data is very challenging due to sparse and noisy data, summary statistics of time-series of risk values could be used.
- Last available risk is a very strong baseline solution due to the nature of the problem.

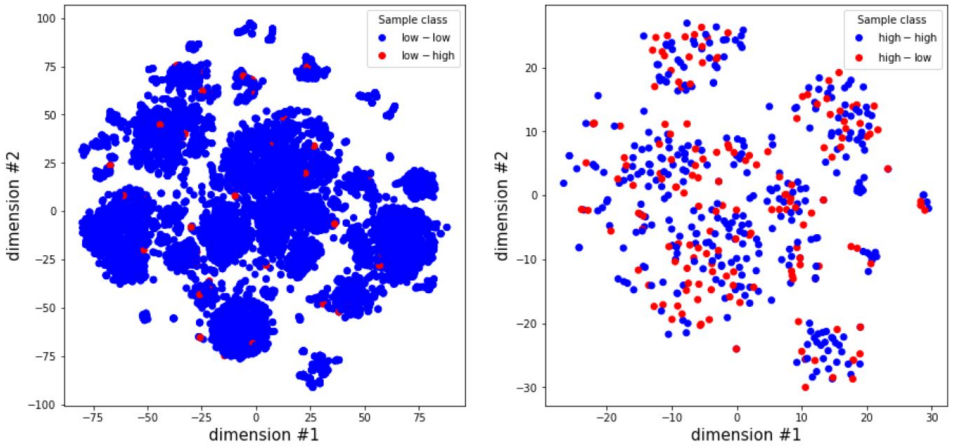
event_id	time_to_tca	mission_id	risk	max_risk_estimate	max_risk_scaling	miss_distance
4	4.966244	19	-7.870632	-6.800245	5.111282	31612.0
4	4.030424	19	-7.968592	-6.807711	5.363402	33272.0
4	3.066467	19	-30.000000	-7.861743	434.669432	33593.0
4	1.797727	19	-30.000000	-8.792368	4334.538505	23709.0
4	1.528458	19	-30.000000	-8.795880	4380.345259	23700.0
4	1.258629	19	-30.000000	-8.763967	4022.819178	23099.0
4	0.973420	19	-30.000000	-8.759451	4008.467679	23059.0
4	0.592587	19	-30.000000	-8.764977	4053.221837	23066.0
4	0.273166	19	-27.650917	-7.819587	50.584357	23080.0

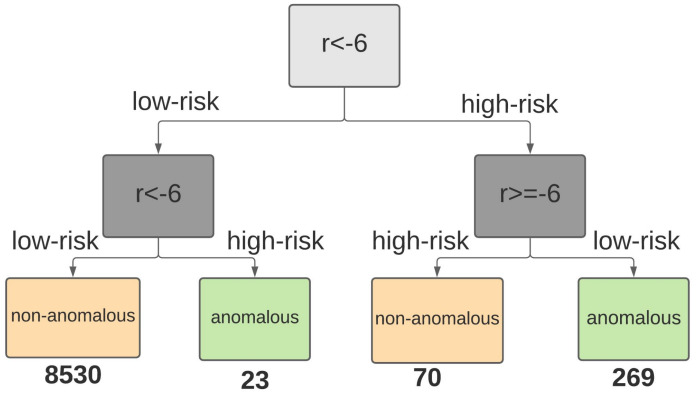
Target label
Input features



Data

- Most features are dependent on each other.
- Data is very imbalanced, and this is related to the problem itself.
- There is a class overlapping.
- There are subgroups.





Metrics

- Loss function needs to match metrics and metrics needs to match the business problem.

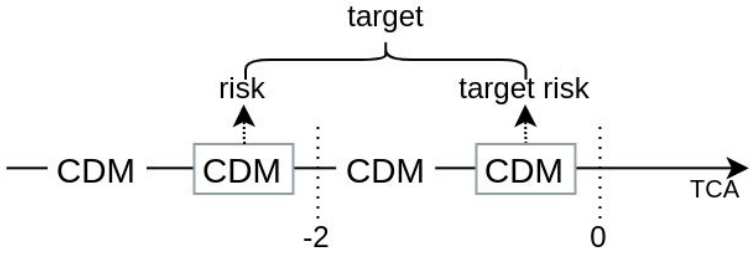
$$L(r, \hat{r}) = \frac{1}{F_2} MSE(r, \hat{r})$$

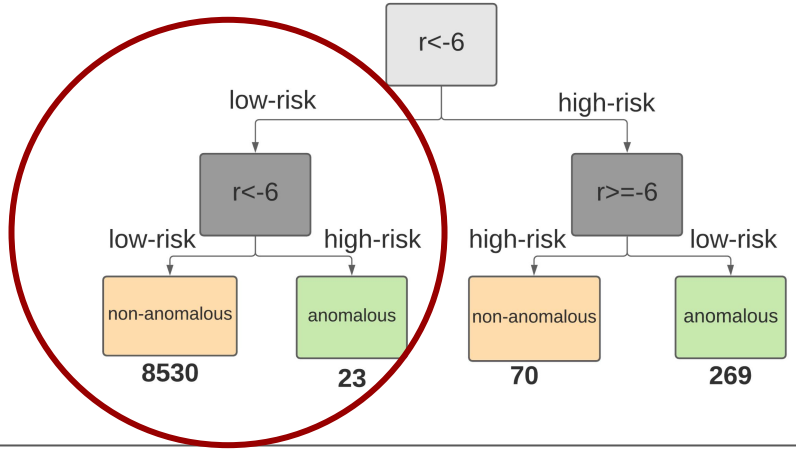
$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

$$MSE(r, \hat{r}) = \frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{r}_i)^2, \{i \mid r_i \geq 10^{-6}\}$$

Model

- Using the latest risk value can be leveraged for predicting target risk values, and it can also be utilised for classification purpose intrinsically by casting the problem as anomaly detection problem.
- Cleaning low-risk to high-risk anomalies (False Negatives) can yield better score due to F2 metric.
- Cleaning high-risk to low-risk anomalies (False Positive) is costly if it fails due to F2 metric.





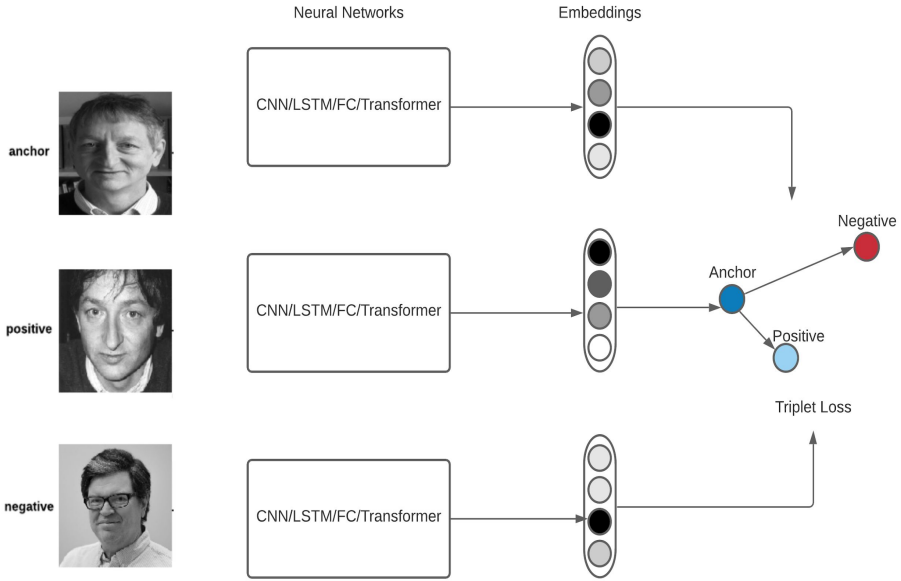
Model

- There is a distribution difference between anomalous and non-anomalous samples.

features	mean	min	25%	50%	75%	max
time_to_tca (nonanomalous)	2.32	2.0	2.09	2.18	2.29	6.95
time_to_tca (anomalous)	2.28	2.04	2.13	2.24	2.27	3.78
max_risk_estimate (nonanomalous)	-6.32	-9.81	-7.02	-6.34	-5.71	-2.60
max_risk_estimate (anomalous)	-5.87	-7.59	-6.66	-5.72	-5.08	-3.78
max_risk_scaling (nonanomalous)	2e+4	3.2e-11	19.2	96.0	443.3	1.8e+7
max_risk_scaling (anomalous)	15.2	2.9e-9	1.8	3.65	14.5	172.4
mahalanobis_distance (nonanomalous)	198.5	0.0	31.3	99.8	244.2	6091.5
mahalanobis_distance (anomalous)	42.6	0.0	5.5	12.8	30.8	213.6
miss_distance (nonanomalous)	17869.3	70.0	5359.5	13693.5	27095.0	66175.0
miss_distance (anomalous)	8856.9	232.0	1050.5	3631.0	11535.5	37490.0
number_CDMs (nonanomalous)	10.9	1.0	6.0	14.0	15.0	17.0
number_CDMs (anomalous)	6.8	1.0	2.0	4.0	13.0	15.0
mean_risk_CDMs (nonanomalous)	-17.7	-30.0	-23.8	-17.6	-11.4	-3.9
mean_risk_CDMs (anomalous)	-9.9	-27.8	-11.6	-6.9	-6.2	-3.8
STD_risk_CDMs (nonanomalous)	5.0	0.0	1.1	5.9	7.8	12.8
STD_risk_CDMs (anomalous)	1.6	0.0	0.002	0.28	1.22	9.8
c_position_covariance_det_log (nonanomalous)	18	7	14	16	17	20
c_position_covariance_det_log (anomalous)	45	3	11	13	15	46
c_obs_used (nonanomalous)	67.9	3.0	21.0	30.0	63.0	1502.0
c_obs_used (anomalous)	31.13	9.0	15.0	18.0	29.0	223.0

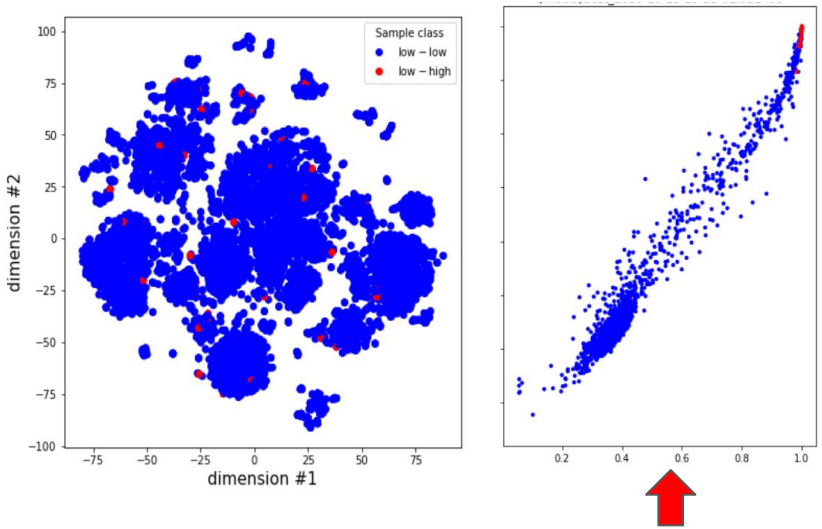
Model

- Simpler models can help with over-fitting when the number of sample is limited.
- Using the most discriminative features can help with over-fitting.
- Ensembling models may increase the accuracy of the predictions.
- Machine learning models should allow incremental learning and defining custom loss functions.
- Standardisation scaling is better when data has outliers.



Results

- The proposed models are ensembled with majority vote (3-fold).
- The distance distribution is computed against all data available.
- Ranked 3rd out of 96 teams.



Teams	Score	MSE_{HR}	F_2
sesc	**0.556**	**0.407**	0.733
dietmarw	0.571	0.437	**0.765**
Magpies	0.585	0.441	0.753
Vidente	0.610	0.436	0.714
DeCRA	0.615	0.457	0.743
Valis	0.628	0.467	0.744
DunderMifflin	0.628	0.451	0.718
madks	0.634	0.476	0.750
vhrique	0.649	0.496	0.764
Spacemeister	0.649	0.479	0.738
LRP baseline	0.694	0.513	0.739

Conclusions

- It is possible to build machine learning models that can perform **better than naive solution**.
- There is **distribution difference between subgroups** of classes, and this can be further leveraged.
- It is beneficial to incorporate **physical properties and behaviours** of resident space objects into the models.
- It is possible that various data sources have been merged and they have different distributions, and **data fusion using machine learning** should be investigated.

Questions