Hazardous asteroids forecast via Markov random fields

Project for the exam: Probabilistic Modelling (DSE)

Marzio De Corato

Introduction

- **Final Goal** Assessment of forecasts and interpretability for different machine learning algorithms, including the probabilistic models
- **Method** Use a dataset for which the laws that interconnects the different features are known from general principles
- **Dataset** CNEOS asteroids dataset for more than 3500 asteroids
- **Theoretical laws** Celestial mechanics
- **Algorithms involved - probabilistic models** GLASSO, mgm, minforest, mmod
- **Algorithms involved - others** Random forest, Support Vector Machines, Quadratic Discriminant Analysis, Logistic Regression

Celestial mechanics

$$\mathbf{F}_1 = \mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_1 \ddot{\mathbf{r}}_1 \tag{1}$$

$$\mathbf{F}_2 = -\mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_1 \ddot{\mathbf{r}}_2 \tag{2}$$

If we consider the motion of the second item with respect to the first one

$$\ddot{\mathbf{r}} = \ddot{\mathbf{r}}_2 - \ddot{\mathbf{r}}_1 \quad \mu = \mathcal{G}(m_1 + m_2) \tag{3}$$

$$\frac{d^2\mathbf{r}}{dt^2} + \mu\frac{\mathbf{r}}{r^3} = 0 \tag{4}$$

$\mathbf{r} \times \ddot{\mathbf{r}} = 0 \implies \mathbf{r}$ and $\dot{\mathbf{r}}$ lies in the same plane

Celestial mechanics

With polar coordinates $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$

$$\mathbf{r} = r\hat{\mathbf{r}} \tag{5}$$

$$\dot{\mathbf{r}} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\boldsymbol{\theta}} \tag{6}$$

$$\ddot{\mathbf{r}} = \left(\ddot{r} - r\dot{\theta}^2\right)\hat{\mathbf{r}} + \left[\frac{1}{r}\frac{d}{dt}\left(r^2\dot{\theta}\right)\right]\hat{\boldsymbol{\theta}} \tag{7}$$

$$\mathbf{h} = r^2\dot{\theta}\hat{\mathbf{z}} \tag{8}$$

$$h = r^2\dot{\theta} \tag{9}$$

Celestial mechanics [5]: 2^{th} Kepler law

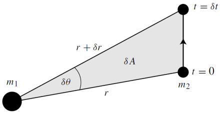


Figure 1: [5]

$$\delta A \approx \frac{1}{2} r(r + dr)\sin(\delta\theta) \approx \frac{1}{2} r^2 \delta\theta \quad (10)$$

$$\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt} = \frac{1}{2} h \quad (11)$$

h is constant $\implies$ 2^{th} Kepler law

Celestial mechanics [5]: 1^{th} Kepler law

Using the substitution $u = \dfrac{1}{r}$ $h = r^2\dot{\theta}$

$$\dot{r} = -\frac{1}{u}\frac{du}{d\theta}\dot{\theta} = -h\frac{du}{d\theta} \tag{12}$$

$$\ddot{r} = -h\frac{d^2u}{d\theta^2}\dot{\theta} = -h^2u^2\frac{d^2u}{d\theta^2} \tag{13}$$

$$\frac{d^2u}{d\theta^2} + u = \frac{\mu}{h^2} \tag{14}$$

$$u = \frac{\mu}{h^2}\left[1 + e\cos(\theta - \phi)\right] \tag{15}$$

Celestial mechanics [5]: 1^{th} Kepler law

$$r = \frac{p}{1 + e\cos(\theta - \phi)} \qquad (16)$$

e is eccentricity

- circle: $e = 0 \qquad p = a$
- ellipse: $0 < e < 1$
  $p = a(1 - e^2)$
- parabola: $e = 1 \qquad p = 2q$
- hyperbola: $e > 1$
  $p = a(e^2 - 1)$

a is the semi-major axis of the conic

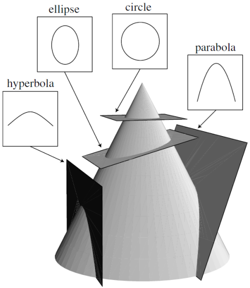


Figure 2: [5]

Celestial mechanics [5]: 3^{th} Kepler law

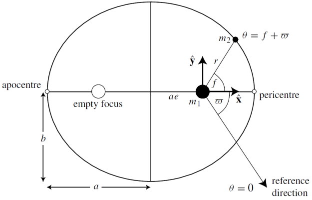


Figure 3: [5]

$$b^2 = a^2(1 - e^2) \tag{17}$$

$$r = \frac{a(1 - e^2)}{1 + e \cdot cos(\theta - \phi)} \tag{18}$$

Area swept in one orbital period T

$$A = \pi ab$$

We know that: $hT/2 \quad h^2 = \mu a(1 - e^2)$

Therefore

$$T^2 = \frac{4\pi^2}{\mu} a^3 \tag{19}$$

3^{th} Kepler law

$$\frac{m_c + m}{m_c + m'} = \left(\frac{a}{a'}\right)\left(\frac{T'}{T}\right)^2 \tag{20}$$

But since $m, m' << m_c$

$$(\acute{a})^3 \approx (T')^2 \tag{21}$$

And therefore

$$T' \approx a'^{3/2} \tag{22}$$

The mass of the asteroid is not involve

d

Orbital parameters

Mean motion

$$n = \frac{2\pi}{T} \tag{23}$$

$$v_{perihelion} = na\sqrt{\frac{1+e}{1-e}} \tag{24}$$

$$v_{aphelion} = na\sqrt{\frac{1-e}{1+e}} \tag{25}$$

Orbital parameters

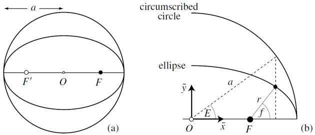


Figure 4: [5]

Mean anomaly

$$M = n(t - \tau) \tag{26}$$

- $M = f = 0 \quad t = \tau$ Perihelion
- $M = f = \pi \quad t = \tau + T/2$ Aphelion

$$M = E - e \sin E \tag{27}$$

Jupiter Tisserard invariant

$$T_P = \frac{a_p}{a} + 2 \cos I \sqrt{\frac{a}{a_p}(1 - e^2)} \tag{28}$$

Orbital parameters

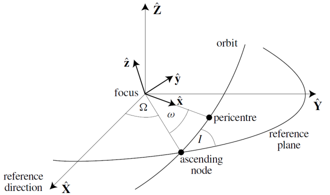


Figure 5: [5]

I: inclination of the orbit

Ω: longitude of the ascending node

Magnitude

$$\Phi = \frac{L}{4\pi r^2} \tag{29}$$

$$m = -2.5 \log_{10} \Phi + C \tag{30}$$

$$m_1 - m_2 = -2.5 \log_{10} \frac{\Phi_1}{\Phi_2} \tag{31}$$

$$M - m = -2.5 \log_{10} \frac{\Phi \cdot d^2}{\Phi \cdot 10^2} \tag{32}$$

$$M = m + 5 - 5 log_{10} d \tag{33}$$

Where Φ is the flux for a sphere of radius r, m the relative magnitude and M the Absolute magnitude

Magnitude

$$\Phi = \frac{L}{4\pi r^2} \tag{34}$$

$$m = -2.5 \log_{10} \Phi + C \tag{35}$$

$$m_1 - m_2 = -2.5 \log_{10} \frac{\Phi_1}{\Phi_2} \tag{36}$$

$$M - m = -2.5 \log_{10} \frac{\Phi \cdot d^2}{\Phi \cdot 10^2} \tag{37}$$

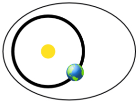
$$M = m + 5 - 5 log_{10} d \tag{38}$$

Where Φ is the flux for a sphere of radius r, m the relative magnitude and M the Absolute magnitude

Classification

Amors

Earth-approaching NEAs with orbits exterior to Earth's but interior to Mars' (named after asteroid (1221) Amor)



$$a > 1.0 \text{ AU}$$
$$1.017 \text{ AU} < q < 1.3 \text{ AU}$$

Apollos

Earth-crossing NEAs with semi-major axes larger than Earth's (named after asteroid (1862) Apollo)

$$a > 1.0 \text{ AU}$$
$$q < 1.017 \text{ AU}$$

Atens

Earth-crossing NEAs with semi-major axes smaller than Earth's (named after asteroid (2062) Aten)

$$a < 1.0 \text{ AU}$$
$$Q > 0.983 \text{ AU}$$

Atiras

NEAs whose orbits are contained entirely within the orbit of the Earth (named after asteroid (163693) Atira)

$$a < 1.0 \text{ AU}$$
$$Q < 0.983 \text{ AU}$$

(q = perihelion distance, Q = aphelion distance, a = semi-major axis)

Classification

- **Potentially Hazardous Asteroids**: MOID ≤ 0.05 au
  $M \leq 22.0$ NEAs whose Minimum Orbit Intersection Distance (MOID) with the Earth is 0.05 au or less and whose absolute magnitude (M) is 22.0 or brighter

Classification

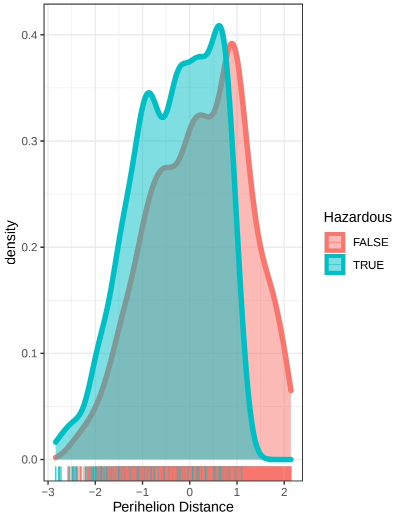
- The asteroid dataset was retrieved from Kaggle [2], which reports into a more machine readable form the dataset of The Center for Near-Earth Object Studies (CNEOS) [3], a NASA research centre.
- 3552 Asteroids
- Among the 40 the features, the ones connected only to the other name of the asteroid, or connected only to the name of the orbit and the one connected with the orbiting planet (since for all it was the Earth) were discarted
- The proportion hazardous/not hazardous was set 1:5
- The continuous measures were standardised and demeaned

Features

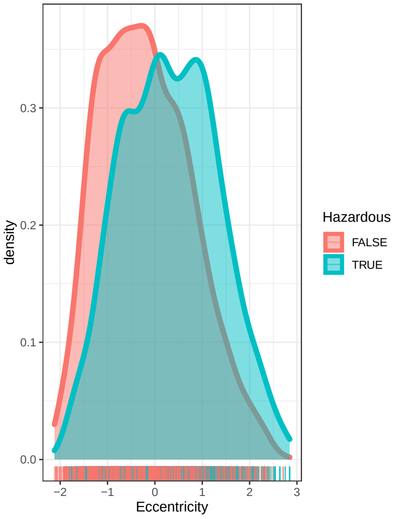
Features	Type
Neo Reference ID	not used
Absolute Magnitude	Continuous
Est Dia in KM (min)	Continuous
Est Dia in KM (max)	Continuous
Close Approach Date	Continuous
Epoch Date Close Approach	Continuous
Relative_Velocity	Continuous
Miss_Dist	Continuous
Min_Orbit_Intersection	Continuous
Jupiter_Tisserand_Invariant	Continuous
Epoch_Osculation	Continuous
Eccentricity	Continuous

Features

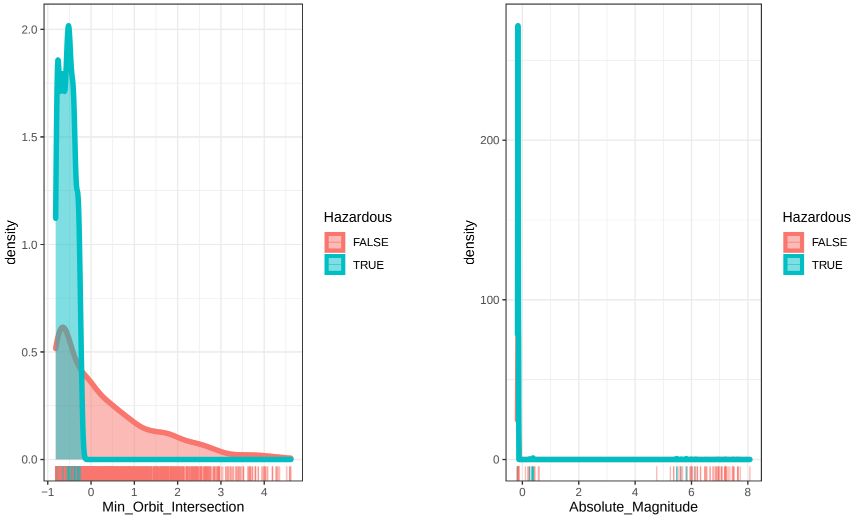
Features	Type
Semi Major Axis	Continuous
Inclination	Continuous
Asc Node Longitude	Continuous
Orbital Period	Continuous
Perihelion Distance	Continuous
Perihelion Arg	Continuous
Perihelion Time	Continuous
Mean_Anomaly	Continuous
Mean_Motion	Continuous
Hazardous	Categorical (Binary)



a) Perihelion Distance



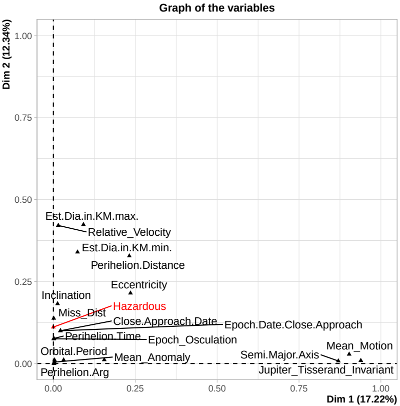
b) Eccentricity

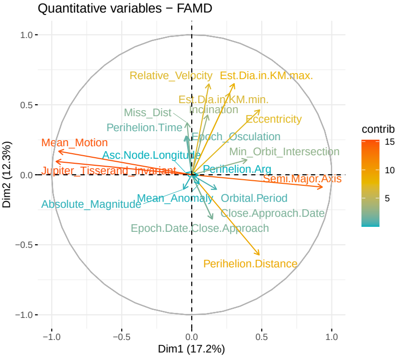


c) Min orbit intersection

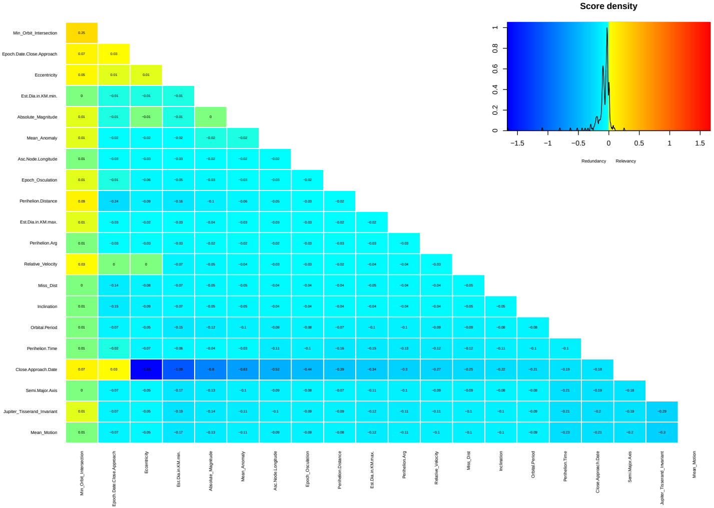
d) Absolute magnitude

FAMD

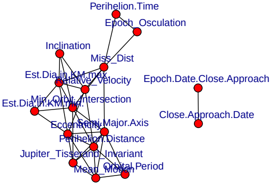

Graph of the variables
Dim 2 (12.34%)
Dim 1 (17.22%)
Est.Dia.in.KM.max.
Relative_Velocity
Est.Dia.in.KM.min.
Perihelion.Distance
Eccentricity
Inclination
Hazardous
Miss_Dist
Close.Approach.Date
Epoch.Date.Close.Approach
Perihelion.Time
Epoch_Osculation
Orbital.Period
Mean_Anomaly
Mean_Motion
Semi.Major.Axis
Perihelion.Arg
Jupiter_Tisserand_Invariant

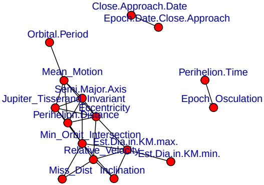

Quantitative variables − FAMD
Dim2 (12.3%)
Dim1 (17.2%)
contrib
Relative_Velocity
Est.Dia.in.KM.max.
Est.Dia.in.KM.min.
Inclination
Miss_Dist
Eccentricity
Perihelion.Time
Epoch_Osculation
Mean_Motion
Min_Orbit_Intersection
Asc.Node.Longitude
Perihelion.Arg
Jupiter_Tisserand_Invariant
Semi.Major.Axis
Absolute_Magnitude
Mean_Anomaly
Orbital.Period
Close.Approach.Date
Epoch.Date.Close.Approach
Perihelion.Distance

Mutual information



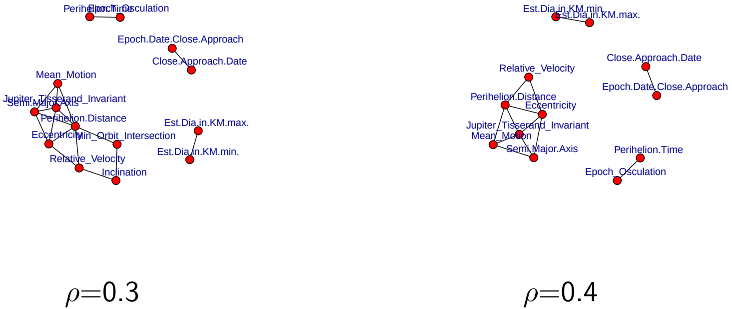
GLASSO





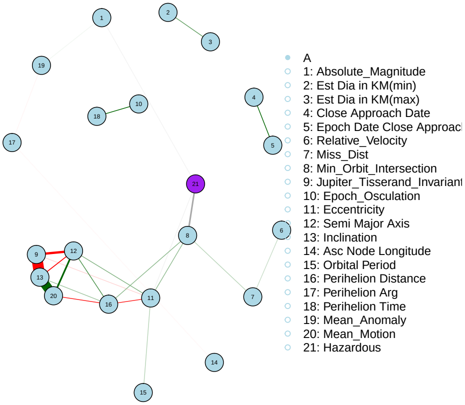
ρ=0.2 ρ=0.2

GLASSO

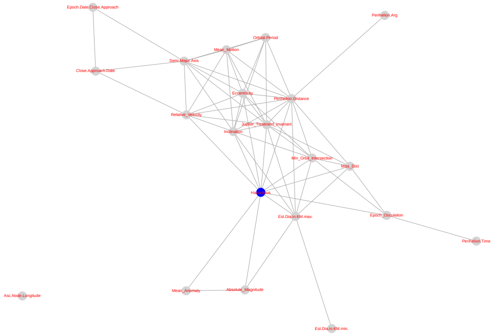


Perihelion.Time
Epoch_Osculation
Epoch.Date.Close.Approach
Close.Approach.Date
Mean_Motion
Jupiter_Tisserand_Invariant
Semi.Major.Axis
Perihelion.Distance
Est.Dia.in.KM.max.
Eccentricity
Min_Orbit_Intersection
Est.Dia.in.KM.min.
Relative_Velocity
Inclination
ρ=0.3
Est.Dia.in.KM.min.
Est.Dia.in.KM.max.
Close.Approach.Date
Relative_Velocity
Epoch.Date.Close.Approach
Perihelion.Distance
Eccentricity
Jupiter_Tisserand_Invariant
Mean_Motion
Semi.Major.Axis
Perihelion.Time
Epoch_Osculation
ρ=0.4

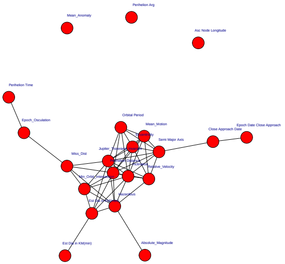
Mixed interactions: mgm



Mixed interactions: minforest

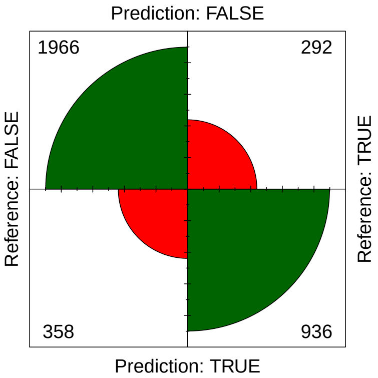


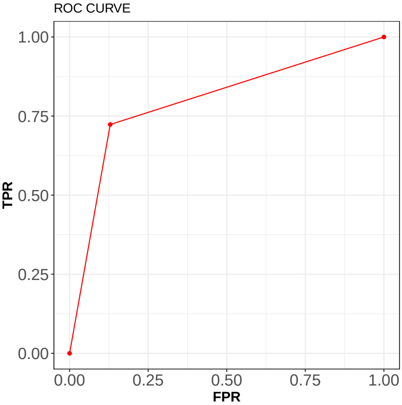
Mixed interactions: mmod



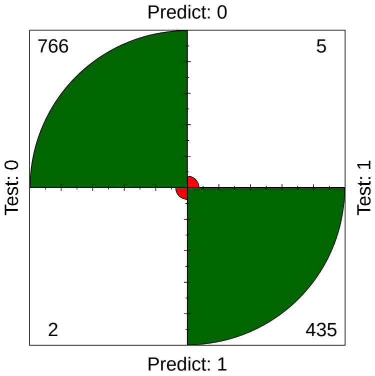
Mixed interactions

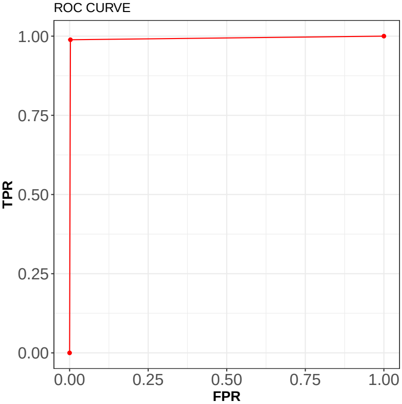
The mgm model is the one that has the list of connection more coherent with the celestial mechanics laws.



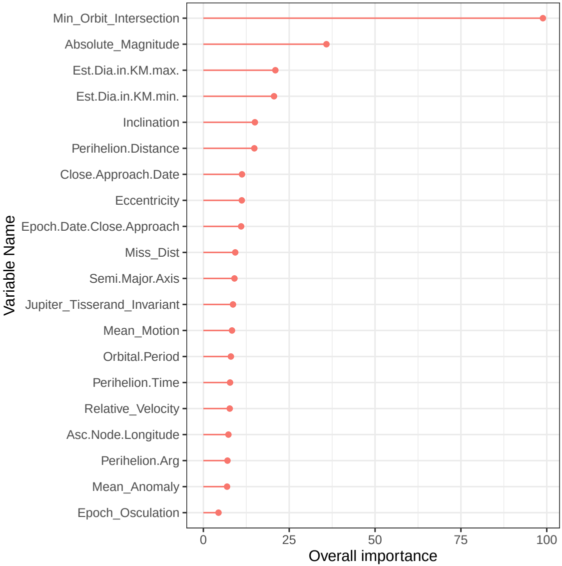


Random Forest

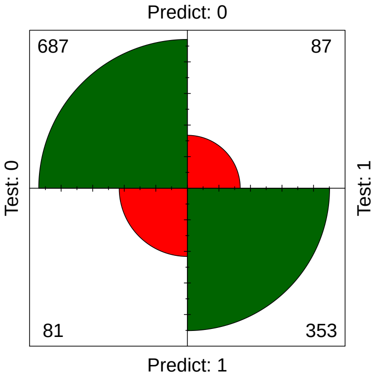


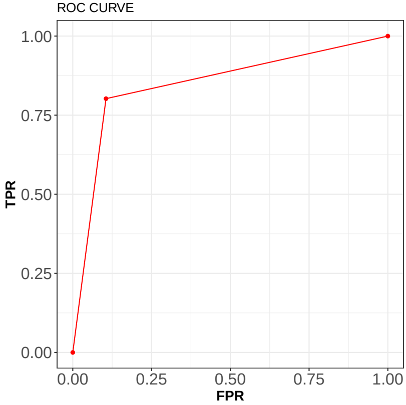


Random Forest

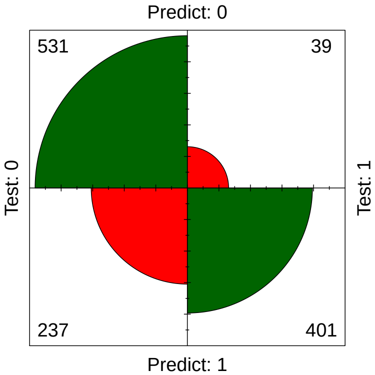


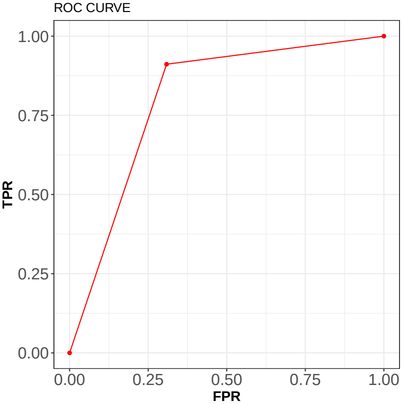
Support Vector Machines



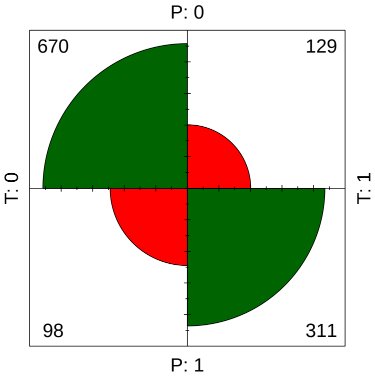


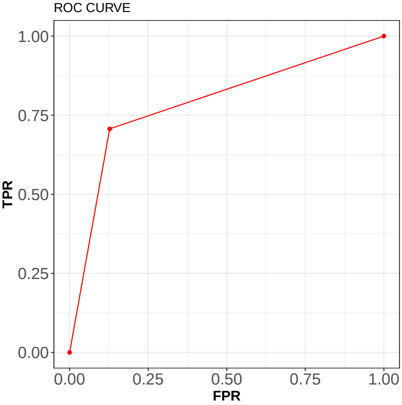
Quadratic Discriminant Analysis (QDA)





Logistic regression





ϕ coefficient

Table 1: ϕ coefficient (also known as Matthews correlation coefficient)

Algorithm	ϕ
RF	0.9876
SVM	0.7111
logistic	0.6173
mgm	0.5997
QDA	0.5562

Conclusions: forecast performances vs intepretability

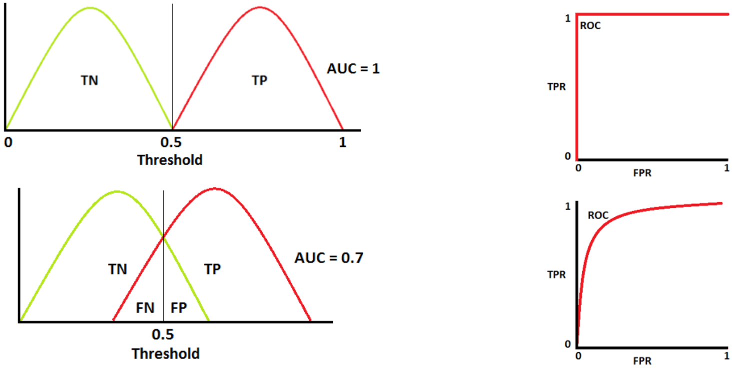
- The mgm algorithm is not the best one in term of performances, but it provides the connections between the features. On the other side, except for the variable importance in RF, the other are black box one
- The mgm model, as the other graphical model is open to a true scientific validation, the other not.
- The probabilistic models lack in the forecast is definitely compensated by their interetability
- This is meaningful since this two features are in conflict
- The probabilistic models provide a good trade-off between intepretability and forecast performances, as long as one is interest to produce a really scientific result (e.g if the only aim is the forecast the RF is definitely better. However how long one can trust to the RF result ?)

ϕ coefficent

	Actual - N	**Actual - P**
Predicted - N	#TP	#FN
Predicted - P	#FP	#TP

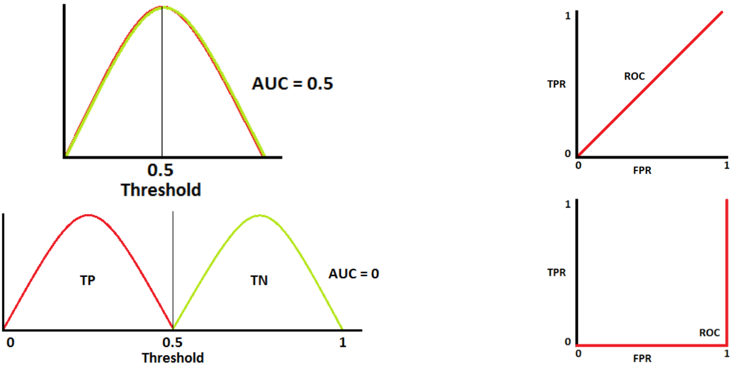
$$\phi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}$$

Receiver operating characteristic



Images taken from [4]

Receiver operating characteristic



Images taken from [4]

Factor analysis of mixed data - FAMD

r(z,k) correlation coefficient (z and k quantitative)

η^2(z,q) correlation ratio (z quantitative and q qualitative)

$$\text{PCA} \rightarrow max \sum_k r^2(z,k)$$

$$\text{MCA} \rightarrow max \sum_q \eta^2(z,q)$$

$$\text{FAMD} \rightarrow max \sum_k r^2(z,k) + max \sum_q \eta^2(z,q)$$

mgm algorithm

$$P(X_s|X_{\setminus s}) = \exp\left\{E_s(X_{\setminus s})\phi_s\left(X_s\right) + B_s(X_s) - \Phi\left(X_{\setminus s}\right)\right\} \tag{39}$$

ϕ_s function of sufficient statistics B_s base measure

$$P(X) = \exp\left\{\sum_{s \in V}\theta_s\phi_s(X_s) + \sum_{s \in V}\sum_{r \in N(s)}\theta_{s,r}\phi_s(X_s)\phi_r(X_r) + \ldots + \sum_{r_1,\ldots,r_k \in C}\theta_{r1,\ldots,rk}\prod_{j=1}^{k}\phi_{rj}\left(X_{rj}\right) + \sum_{s \in V}B_s(X_S) - \Phi(\theta)\right. \tag{40}$$

$$\hat{\theta} = arg\min_{theta}\{-\mathcal{L}(\theta, X) + \lambda||\theta||_1\} \quad ||\theta||_1 = \sum_{j=1}^{J}|\theta_j| \tag{41}$$

GLASSO

$$L_{pen}(K, \hat{\mu}) = log\, \det(K) - tr(K\ S) - \rho||K|| \tag{42}$$

$$K = \Sigma^{-1}$$

S: empirical covariance matrix

Parameters tuning

$$p\left(\mathbf{y}|\boldsymbol{\theta}\right) = \frac{1}{Z(\theta)} \exp\left(\sum_c \boldsymbol{\theta}_c^T \phi_c\left(\mathbf{y}\right)\right) \tag{43}$$

$$\mathcal{L}\left(\boldsymbol{\theta}\right) := \frac{1}{N} \sum_i \log p\left(\mathbf{y}_i|\boldsymbol{\theta}\right) = \frac{1}{N} \sum_i \left[\sum_c \boldsymbol{\theta}_c^T \phi_c(y_i) - \log Z\left(\boldsymbol{\theta}\right)\right] \tag{44}$$

Parameters tuning

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \frac{1}{N} \sum_i \left[\phi_c(y_i) - \frac{\partial}{\partial \boldsymbol{\theta}_c} \log Z(\boldsymbol{\theta}) \right] \tag{45}$$

$$\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}\left[\phi_c(\mathbf{y}) | \, \theta\right] = \sum_{\mathbf{y}} \phi_c(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\theta}) \tag{46}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \left[\frac{1}{N} \sum_i \phi_c(y_i) \right] - \mathbb{E}\left[\phi_c(\mathbf{y})\right] \tag{47}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] - \mathbb{E}_{p(\cdot|\boldsymbol{\theta})}\left[\phi_c(\mathbf{y})\right] \tag{48}$$

$$\mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] = \mathbb{E}_{p(\cdot|\boldsymbol{\theta})}\left[\phi_c(\mathbf{y})\right] \tag{49}$$

mmod

$$f(i,y) = p(i)(2\pi)^{-q/2} det(\Sigma)^{-1/2} \\ exp\left[-\frac{1}{2}(y-\mu(i))^T \Sigma^{-1}(y-\mu(i))\right] \tag{50}$$

$$f(i,y) = \exp\left\{g(i) + \sum_u h^u(i) y_u - \frac{1}{2}\sum_{uv} y_u y_v k_{uv}\right\} \\ = \exp\left\{g(i) + h(i)^T y - \frac{1}{2} y^T K y\right\} \tag{51}$$

where $g(i)$, $h(i)$ and K are the canonical parameters

mmod

$$
\begin{aligned}
K &= \Sigma^{-1} \\
h(i) &= \Sigma^{-1}\mu(i) \\
g(i) &= \log p(i) - \frac{1}{2}\log det(\Sigma) \\
&\quad - \frac{1}{2}\mu(i)^T\Sigma^{-1}\mu(i) - \frac{q}{2}\log 2\pi
\end{aligned}
\tag{52}
$$

Graphical models

$$p(x_1, x_2, ..., x_n) \tag{53}$$

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(x_V|x_{1:V-1}) \tag{54}$$

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{55}$$

$$p(\mathbf{x}_{1:V}) = p(x_1)\prod_{t=1}^{V} p(x_t|x_{t-1}) \tag{56}$$

Graphical models

Theorem (Hammersley-Clifford)

*A positive distribution p(**y**)¿0 satisfies the CI properties of an indirect graph G iif p can be represented as a product of factor, one per maximal clique, i.e.*

$$p(\mathbf{y}|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \tag{57}$$

where C is the set of all the (maximal) cliques of G, and Z(θ) is the partition function given by

$$Z(\theta) := \sum_{y} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \tag{58}$$

Note that this partition function is what ensures the overall distribution sums to 1

Graphical models

$$p(y|\theta) = \frac{1}{Z(\theta)} exp\left(-\sum_c E(y_c|\theta_c)\right) \tag{59}$$

$$\psi_c(y_c|\theta_c) = exp\left(-E(y_c|\theta_c)\right) \tag{60}$$

Information theory

$$H(X) = -\sum_{x \in X} p(x) logp(x) \tag{61}$$

$$H(X, Y) = -\sum_{x \in \chi} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{62}$$

$$\begin{aligned} H(X|Y) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X) \end{aligned} \tag{63}$$

$$H(X, Y) = H(X) + H(Y|X) \tag{64}$$

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \quad D(p||q) \geq 0 \tag{65}$$

Information theory

$$I(X;Y) = \sum(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \\ = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{66}$$

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1) \tag{67}$$

$$g(\alpha, \mathbf{C}, \mathbf{S}, f_i) = MI(f_i; \mathbf{C}) - \sum_{f_s \in S} \alpha(f_i, f_s, \mathbf{C}, \mathbf{S}) MI(f_i; f_s) \tag{68}$$

Ref

[1] https://cneos.jpl.nasa.gov/about/neo_groups.html.

[2] https://www.kaggle.com/shrutimehta/nasa-asteroids-classification.

[3] https://cneos.jpl.nasa.gov/.

[4] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[5] Carl D Murray and Stanley F Dermott. *Solar system dynamics*. Cambridge university press, 1999.