No. 7, 2018

RISK INSIGHTS

Décryptage des modèles « boîte noire » : interprétabilité et fiabilité

par Dr. Behrang Jalali, Gen Re, Cologne

Dans la plupart des secteurs, les modèles linéaires généralisés (MLG) et les modèles d'analyse des données spécifiques aux domaines sont habituellement utilisés pour extraire des informations utiles à partir des données. Les calculs mathématiques qui sous-tendent ces analyses sont plutôt simples et les praticiens comme les participants au projet non initiés savent parfaitement interpréter les résultats et ne manquent pas de les appliquer dans le contexte de l'entreprise.

Ces dernières années, des modèles plus pointus, tels que l'apprentissage automatique, sont de plus en plus appliqués et surpassent les modèles conventionnels. Les modèles qui reposent sur l'apprentissage automatique sont utilisés dans un large éventail d'applications, de la classification des images et l'analyse de texte à la segmentation des clients et l'analyse des écarts.

Les modèles d'apprentissage automatique sont généralement qualifiés de modèles « boîte noire ». Cela s'explique non seulement par le haut niveau de détails techniques nécessaires pour les comprendre, mais aussi par les résultats ; par exemple, les schémas reconnus entre la réponse et les variables, qui souvent ne peuvent être formulés en termes de relation concise, contrairement aux modèles linéaires. La capacité à reconnaître des relations non linéaires et non monotones entre les variables et la réponse est toutefois précisément ce qui distingue ces modèles sophistiqués des traditionnels en termes de performances.

Une expérience limitée dans l'application de ces modèles et l'impossibilité d'accéder à des méthodes faciles à comprendre, par exemple, des graphiques informatifs, qui communiquent les résultats aux membres non-techniciens dans une entreprise, empêche de nombreux praticiens et managers d'appliquer ces modèles, notamment dans les secteurs réglementés tels que l'assurance-vie/maladie.

Cet article présente les méthodes destinées à faciliter l'interprétation des résultats (prédictions) des modèles d'apprentissage automatique et de mieux comprendre ces modèles. Ces méthodes aident par ailleurs à évaluer les modèles dans le milieu de l'entreprise, en permettant aux membres du projet d'évaluer la fiabilité d'un modèle et nécessitent une expertise technique limitée. Les méthodes présentées peuvent

Sommaire

La lettre d'information en bref

Risk Insights est une publication technique conçue par Gen Re pour les cadres de l'assurance vie et santé du monde entier. Les articles portent sur des thèmes ayant trait à la gestion des risques, la médecine, les sinistres, la souscription et les questions actuarielles.

Les produits étudiés sont les assurances maladie grave, soins longue durée, pensions d'invalidité, santé et vie.

Tableau 1 – Description des données

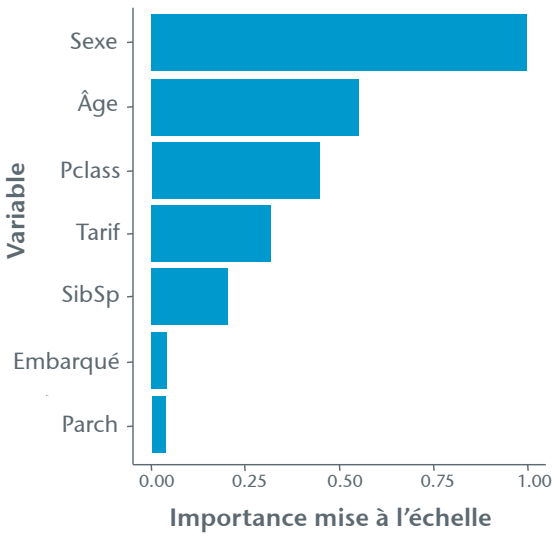
Variables	Description	Valeur
Statut de survie (réponse)	Classe binaire de réponse	Mort (0) ou Survie (1)
Âge	Âge du passager	0 à 80 ans (numérique)
Sexe	Sexe du passager	Masculin ou féminin
Pclass	Classe du passager	1, 2 ou 3
Tarif	Prix du billet	0 à 500 livres sterling (nombre)
SibSp	Nombre de frères/sœurs et conjoint	1,2,3, ..., 8 (nombre discret)
Parch	Nombre de parents et enfants	1,2,3, ..., 6 (nombre discret)
Embarqué	Port d'embarquement	S=Southampton, Q=Queenstown, C=Cherbourg

être appliquées avec toutes sortes d'algorithmes et sont par conséquent compatibles avec tous les modèles et peuvent être utilisées dans les applications de régression et de classification.

Les données et le modèle

Pour illustrer les différentes méthodes, nous utilisons les données génériques du Titanic,[1] représentant le tragique naufrage du paquebot auquel plus de 60 % des passagers n'ont pas survécu (cf. Tableau 1). Nous appliquons l'algorithme de Gradient Boosting (GBM) comme exemple de modèle d'apprentissage automatique. Le GBM est un modèle arborescent et un ensemble d'apprenants faibles (modèles) qui sont construits selon une séquence pour concevoir un modèle final plus robuste. Le cas échéant, nous utilisons le GBM comme classifieur binaire pour modéliser le statut de survie des passagers.

Figure 1 – Importance des variables d'un modèle GBM classifiant la réponse binaire dans l'ensemble de données du Titanic



Source: Gen Re

Importance des variables

L'importance des variables est la liste de toutes les variables qui ont été incluses dans le modèle, généralement classées dans un ordre décroissant et qui est statistiquement calculée différemment pour les modèles d'apprentissage automatique. Généralement, les variables plus importantes entraînent une baisse plus grande des erreurs dans la description de la variable de réponse (statut de survie modélisé dans ce cas). La figure 1 montre cette liste comme résultat de l'application d'un modèle GBM avec les données du Titanic. Dans les modèles MLG, cette liste peut être similaire en termes de valeurs absolues des coefficients, où les coefficients plus élevés ont un impact plus important (par rapport aux autres variables) dans la description de la réponse.

Dans le périmètre global, il est important d'identifier les variables qui sont plus importantes dans un projet donné, ce qui est particulièrement utile dans les grands ensembles de données dimensionnelles. Il fournit la première mesure dans l'évaluation de la fiabilité d'un modèle, qui est avérée lorsque la liste des variables importantes est conforme aux attentes de domaine et peut aussi rester stable, avec de légères variations des données. (Figure 1)

Diagramme de dépendance partielle

Le diagramme de dépendance partielle (PDP) montre la dépendance d'une réponse sur une variable ou un ensemble de variables. Il montre l'effet marginal de la variable choisie sur la réponse (pour la régression) ou la probabilité de classe (pour la classification). Cet effet est mesuré dans les variations de la réponse moyenne, c'est-à-dire pour la classification de la variation de la probabilité

de classe. L'axe Y n'exprime plus les valeurs de réponse originales. L'idée peut être également comprise en analogie avec l'interprétation des coefficients dans les modèles MLG ; en d'autres termes, la variation de la réponse au regard de la variable choisie, en supposant que les autres variables restent constantes.

Dans la figure 2a, le panneau de gauche montre la dépendance de la réponse (probabilité de classe Survie ou Mort) concernant la variable de l'Âge dans la modélisation des données du Titanic. Nous avons ajouté une fonction supplémentaire aux PDP illustrés qui nous permet de mieux décider du niveau de confiance dans la tranche d'âge et de prendre des décisions prudentes si nécessaire. Le « rug » ajouté (petites coches verticales le long des axes X) montre que ces données sont trop rares après 60 ans, car trop peu de données sont disponibles après 70 ans.

Les PDP améliorent la fiabilité lorsque les résultats sont conformes avec le savoir du domaine, et permettent de comprendre le modèle en visualisant la non-linéarité et les interactions.

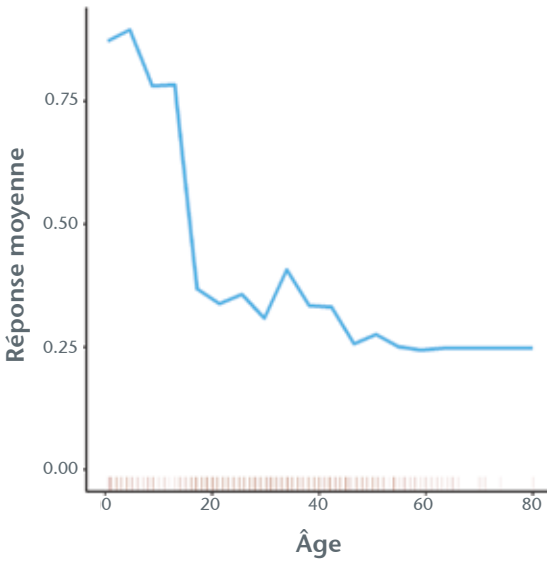
En utilisant deux variables, présentées dans la fenêtre de droite de la figure 2b, le PDP nous permet d'étudier le lien possible entre deux variables dans la description du comportement de la réponse. Dans cet exemple, l'existence d'une interaction entre les deux variables est importante. Comme on peut le constater, la probabilité de survie moyenne diminue avec l'âge (ce qui indique que les passagers plus âgés ont moins de chances de survivre) et augmente pour les femmes, quel que soit l'Âge. Pour les moins de 15 ans (enfants dans ces données), l'Âge est un facteur plus important que le Sexe. L'interprétation ci-dessus est conforme aux hypothèses raisonnables qui supposent que les enfants et les femmes voyageant à bord du Titanic sont montés les premiers à bord des canots de sauvetage, ce qui indique que ce modèle est fiable.

Les PDP fournissent des informations globales lorsque les données sont prises dans leur ensemble, et sont localement informatifs en termes de variables individuelles.

Modèles de substitution

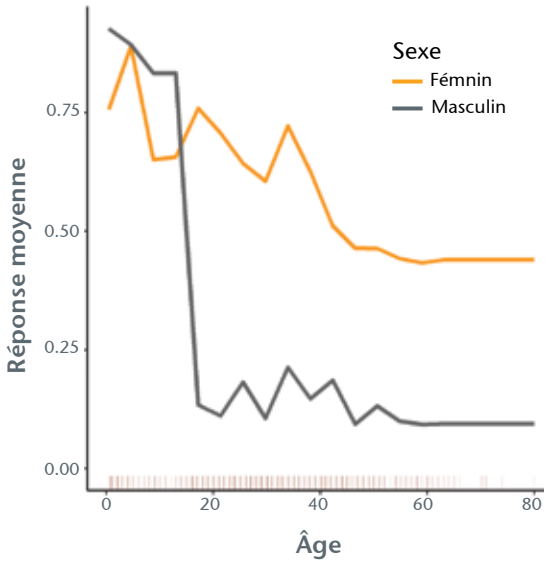
Il existe une autre méthode pour interpréter les résultats des modèles d'apprentissage automatique qui consiste à approcher les prédictions complexes avec un modèle de substitution plus simple. L'idée consiste à entraîner un modèle simple, comme un arbre de décision ou de régression linéaire, avec les données originales, mais en utilisant les prédictions du modèle complexe comme réponse. Dans la pratique, le modèle simple, qui est moins précis que le modèle complexe réel, est utilisé pour visualiser et expliquer les schémas reconnus.

Figure 2a – L'effet de la variable de l'âge dans la variation de la réponse moyenne (probabilité de classe survie ou mort)



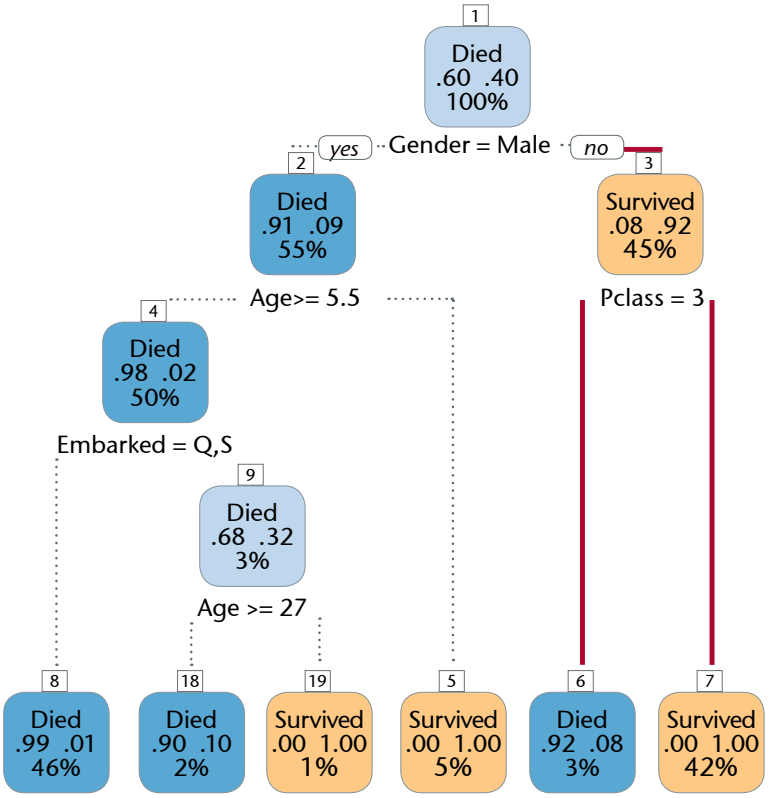
Source: Gen Re

Figure 2b – Variation de la réponse moyenne en termes d'âge et de sexe



Source: Gen Re

Figure 3 – Approcher un modèle GBM complexe avec un arbre de décision simple, visualisant les schémas reconnus, c.-à-d. les règles métiers



Source: Gen Re

Arbre de décision simple

La figure 3 montre un arbre de décision entraîné, en utilisant les probabilités prévues d'un modèle GBM comme réponse pour visualiser les relations reconnues entre les variables impliquées et la réponse. Nous vous rappelons que le modèle GBM est en soi un ensemble de 50 arbres et certains de ces arbres peuvent comprendre jusqu'à six niveaux.

Ces chemins faciles à comprendre peuvent être interprétés et utilisés comme règles métiers.

Pour lire cet arbre, concentrez-vous sur les chemins situés à droite (marqués en rouge). Le premier nœud en haut, la racine, contient toutes les données utilisées dans cet arbre (100 %) et montre les deux probabilités, à cette profondeur, pour les labels Mort et Survie, 60 % et 40 % respectivement. Passant à la seconde étape, suivant la suggestion du graphique pour les femmes (la branche « Non » si le Sexe n'est pas masculin), nous passons sur le nœud orange de droite. Nous avons ici 45 % des données, soit les observations qui ont été prédites comme Survie par 92 % de probabilité et seulement 8 % de probabilité pour le label Mort. Jusqu'à ce niveau, en dehors de l'interprétation, le modèle est cohérent avec les informations historiques selon lesquelles les femmes ont plus de chances de survivre, et ce modèle semble par conséquent fiable (pour cette branche). (voir Figure 3)

En continuant la branche ci-dessus, pour les passagers de la troisième classe (partie gauche) le chemin se termine sur le dernier nœud bleu (le dernier nœud est appelé feuille). 92 % de chances s'appliquent au label Mort. Pour la partie droite qui se termine sur la feuille orange, il est suggéré que les observations représentant des passagers de sexe féminin dans la première ou la deuxième classe auraient toutes les chances de survivre (100 % survécu et 0 % mort) et qui contient 42 % des données.

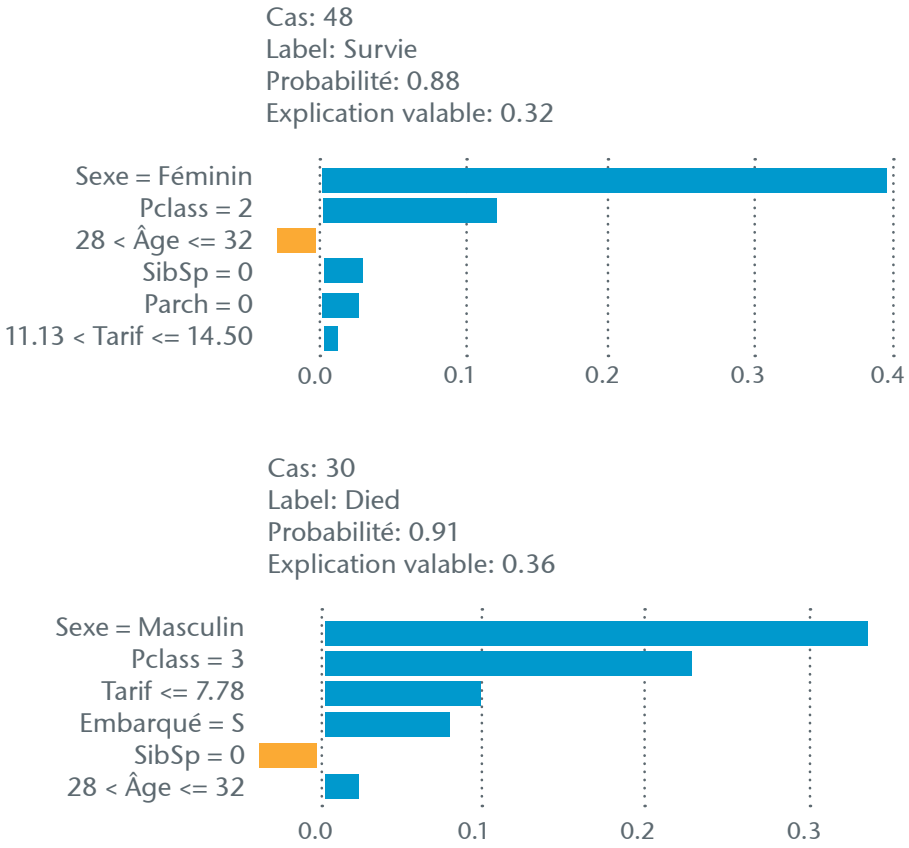
Un modèle de substitution sert également de mesure de la fiabilité lorsque les règles suggérées sont compréhensibles selon le contexte de l'entreprise.

Les modèles de substitution sont utiles pour expliquer les relations graphiquement non linéaires et non monotones entre les variables et la réponse.

Approcher un modèle complexe avec un arbre de décision simple peut être considéré comme applicable comme interprétation globale, car il décrit les contributions moyennes (ou globales) des variables à la réponse prédite. Cela peut être aussi considéré comme local lorsqu'il est utilisé pour expliquer un chemin de décision d'une (ou d'un groupe d') observation spécifique.

Notez que rien ne garantit que ce substitut représenterait le modèle complexe fidèlement. Dans la pratique, il convient d'expérimenter avec les paramètres de cette méthode et d'examiner la distribution de la réponse prédite (probabilités dans l'exemple ci-dessus). Dans cet exemple, nous avons sélectionné les observations qui sont modélisées avec une grande confiance, c.-à-d. les probabilités prévues proches de 0 et 1, visant à faire ressortir des schémas plus importants.

Figure 4 – Résultat LIME pour la 48e et la 30e observations, un cas pour chaque label prédit.
L'axe X est le poids des principales variables basé sur un modèle de régression (substitut local). Les variables bleues corroborent la prédiction, tandis que les oranges la contredisent.



Source: Gen Re

Explications agnostiques en termes de modèle interprétables locales

Les explications agnostiques en termes de modèle interprétables locales (LIME) fournissent une explication pour des prédictions en déterminant les variables les plus importantes (de renforcement ou contradictoires) des observations données. Cette méthode avait pour vocation de servir d'approche d'interprétation agnostique en termes de modèle pour la classification des textes et des images. Ses capacités puissantes ont été vite propagées à d'autres applications, comme l'analyse de la segmentation des clients.

Voici comment fonctionne la technique LIME : chaque observation (ligne de données) est simulée en créant des milliers de permutations. Le modèle complexe entraîné établit des prédictions à partir de ces permutations. Un modèle plus simple, comme une régression linéaire, modélise ensuite ces données générées pour identifier les variables plus importantes. Les variables pour chaque observation seront pondérées en fonction d'un modèle de régression, comme illustré sur la figure 4 par exemple pour deux observations sélectionnées au hasard.

Dans la figure 4, le résultat de LIME est montré pour deux observations sélectionnées au hasard (cas dans la sortie LIME) en utilisant les données du Titanic illustrées. Les deux cas ont été sélectionnés à partir d'un échantillon d'observations avec des prédictions correctes, ce qui signifie que la réponse observée a été correctement modélisée. Cet exercice peut être également réalisé pour des prédictions fausses (cas de faux positifs ou faux négatifs) pour un examen plus approfondi de ces observations.

Le panneau en haut montre que le label prédit est Survie et possède 88 % de probabilités, soit une prédiction fiable. Nous constatons que le Sexe (féminin dans cette observation) et la classe du passager (Pclass) ont un poids important dans la détermination de cette probabilité. Ces deux variables, qui confirment le résultat prédit, sont parfaitement alignées avec le contexte de

cet accident. Le fait que l'Âge contredise cette prédiction a peu d'importance, compte tenu de son importance limitée dans cette observation. Si l'on observe les autres variables de la liste, le fait de n'avoir aucun membre de sa famille augmente les probabilités de monter à bord du bateau, et hormis le bruit ou les fluctuations pour le faible poids négatif de l'Âge, cette recommandation est raisonnable.

Dans le panneau du bas, nous observons un label Mort prédit. Un passager qui est un homme, qui appartient à la troisième classe (et possédant un billet très peu cher) qui n'est pas trop jeune, sont des arguments raisonnables pour labelliser correctement et en toute confiance cette observation comme Mort.

En vérifiant une fraction significative de cas, le résultat LIME doit permettre de :

1) Évaluer la fiabilité du modèle
2) Rechercher les principales variables importantes qui corroborent les prédictions correctes
3) Mieux examiner les cas individuels au besoin pour d'autres exigences de l'entreprise, comme la segmentation des clients, les études des déclarations de sinistre et la sélection des risques

Comme on peut s'y attendre dans l'exemple ci-dessus, cette méthode est une approche d'interprétation locale, car il existe des modèles de substitution (linéaires) locaux pour chaque observation.

Résumé et conclusions

Les modèles d'apprentissage automatique sont plus performants et se prêtent à un éventail plus large d'applications comparé aux analyses traditionnelles. L'application de ces modèles avancés dans le cycle de vie des projets d'analyse, en collaboration avec des experts du domaine, peut certainement ajouter de la valeur à l'entreprise.

Dans cet article, nous présentons des techniques agnostiques qui garantissent davantage de transparence dans l'application de modèles avancés en permettant aux utilisateurs d'interpréter les résultats dans le contexte de l'entreprise, et améliorer la compréhension et la fiabilité des modèles.

Ces techniques peuvent être appliquées à la modélisation de différentes applications dans l'assurance, comme la tarification, l'analyse des portefeuilles, la sélection des risques simplifiée, la segmentation des clients et l'étude des sinistres. Veuillez trouver ci-dessous des descriptions courtes des méthodes proposées et de quelle manière elles permettent d'interpréter les résultats :

- L'importance des variables est une liste de résultats qui décrit la contribution globale des variables incluses dans la modélisation de la réponse.
- Le diagramme de dépendance partielle (PDP) montre comment la réponse modélisée moyenne change en termes d'effet marginal d'une variable spécifique. Les PDP constituent un outil puissant pour visualiser les interactions possibles et peuvent être utiles pour communiquer les informations d'un modèle complexe à un public non initié.
- Les modèles de substitution sont des modèles simples qui approchent les modèles complexes pour expliquer les schémas reconnus, par exemple :
  - Un arbre de décision simple peut illustrer visuellement le chemin ou les règles métiers entre la réponse et les variables.
  - La technique LIME permet d'analyser le raisonnement qui sous-tend les prédictions du modèle en ce qu'il fournit l'importance des variables au niveau local.

L'auteur

Dr. Behrang Jalali *est responsable de la mise en place et du développement de capacités d'analyse avancées pour Life-Health International. Il gère des projets analytiques en collaboration avec les unités commerciales, y compris la conception des expériences et l'application de techniques modernes pour créer de la valeur ajoutée pour les clients. Avant de rejoindre Gen Re, il travaillait comme chercheur en astrophysique computationnelle. Il peut être contacté par tél. au +49 221 9738 799 ou par e-mail à l'adresse behrang.jalali@genre.com.*

Sources

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2015). An Introduction to Statistical Learning with applications in R. Springer.

Hall, P., Phan, W. and SriSatish, A. (2017). Ideas on intrepreting machine learning; O'Reilly Media.

Riberio, M. T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

Note

1 Data from "titanic" package: https://cran.r-project.org/web/packages/titanic/titanic.pdf

The difference is...the quality of the promise®

Gen Re®

genre.com | genre.com/perspective | Twitter: @Gen_Re

General Reinsurance AG
Theodor-Heuss-Ring 11
50668 Cologne, Germany
Tel. +49 221 9738 0
Fax +49 221 9738 494

General Reinsurance AG–Succursale Paris
21, rue Balzac
750008 Paris
Tel. +33 1 5367 7676
Fax +33 1 5367 4646

Editors:
Ulrich Pasdika, ulrich.pasdika@genre.com
Ross Campbell, ross_campbell@genre.com

Photos: © getty images – undrey, simoningate, TonyBaggett