No. 1, 2018

RISK INSIGHTS®

Efficacité des modèles de classification

par Louis Rossouw, Gen Re, Le Cap

Les assureurs développent de plus en plus de modèles de prédiction qu'ils utilisent dans leurs processus d'assurance. Ces modèles reposent souvent sur des techniques traditionnelles, mais les techniques d'apprentissage automatique sont de plus en plus appliquées.

Dans la pratique, ces techniques sont appliquées aux cas de sélection des risques, aux demandes de souscription de polices, voire aux déclarations de sinistre. Un exemple peut prendre la forme d'un modèle qui est utilisé pour prédire les cas qui seront évalués comme étant « standard » avant qu'un souscripteur ne s'en aperçoive. C'est ce que l'on appelle un « modèle de classification » utilisé pour classer les données dans des buckets discrets (oui ou non, standard ou non, etc.).

Les techniques de modélisation utilisées dans ces applications peuvent être relativement simples ou bien très complexes. Voici quelques exemples de techniques :

- Régression logique (généralement un modèle linéaire généralisé – MLG)
- Arbres de décision
- Forêts aléatoires
- Machines à vecteurs de support
- Techniques de descente du gradient
- Réseaux de neurones artificiels

Les techniques traditionnelles, telles que des modèles basés sur la régression, produisent des modèles qui sont lisibles par des humains. L'impact de chaque variable que comprend le modèle est clairement visible sur le résultat. La plupart des techniques d'apprentissage automatique produisent des modèles que l'observation directe ne permet pas de comprendre facilement. Ils produisent des résultats, mais les mécanismes internes sont cachés ou trop complexes pour pouvoir parfaitement les comprendre. Ce sont les modèles dits de « boîte noire ».

Avec un tel choix de modèles, comment évaluer la précision et la qualité de ces modèles pour identifier le meilleur ? Comment tenons-nous compte de l'efficacité

Contenu

des modèles faciles à comprendre, par rapport aux différents modèles de boîte noire ? Comment évaluons-nous la valeur relative pour l'assureur de ces modèles ?

Dans la suite de l'article, nous nous concentrerons sur un classifieur binaire simple capable de prédire si une demande de souscription doit être considérée comme standard (sans supplément ni condition de sélection des risques) ou non (refusée, avec une surprime ou clause d'exclusion). Ce modèle peut être utilisé pour passer outre la sélection des risques traditionnelle afin de gagner du temps et faire des économies pour un sous-ensemble de cas.

Données d'entraînement et de tests

Pour créer un modèle classifieur, on a généralement un ensemble de données avec des cas historiques et le résultat enregistré de ces cas. Dans notre exemple de sélection des risques, il peut s'agir de données concernant le demandeur et la décision enregistrée (standard ou non).

Il est préférable de répartir les données passées dans au moins deux catégories :

- Les données d'entraînement qui seront utilisées pour intégrer le(s) modèle(s) en question ; ces données sont utilisées pour « entraîner » le modèle.
- Les données de test qui seront utilisées pour évaluer le(s) modèle(s) afin de déterminer l'efficacité du/des modèle(s)

Les données de validation sont également utilisées très souvent pour affiner les paramètres de modélisation avant de tester le modèle.

La raison principale de la séparation entre les tests et l'entraînement est de s'assurer que le modèle fonctionne bien avec les données sur lesquelles il n'a pas été entraîné. Cela permet notamment d'identifier le problème de surapprentissage ou surajustement, qui se produit lorsqu'un modèle donné semble donner des prédictions trop précises à partir des données d'entraînement. Lorsque ce modèle est ensuite vérifié par rapport aux autres données de test, sa performance se dégrade toutefois considérablement. On constate alors que le modèle « surapprend » les données d'entraînement.

Entre 10 et 30 % des données sont retenues comme données de test. Cela dépend de la disponibilité globale des données et du degré de surapprentissage possible des modèles. Il est possible de sélectionner des données de test comme sous-ensemble aléatoire de vos données ou, par exemple, baser la sélection des données sur le temps (en sélectionnant la dernière année de données comme sous-ensemble de test).

La matrice de confusion

Pour évaluer la qualité d'un classifieur binaire, il est possible de générer une matrice de confusion avec nos données de test. Il suffirait d'appliquer le modèle sur ces données afin de produire des résultats prédits, et intégrer aussi les résultats réels dans les données. Puis de compiler une matrice avec les nombres de cas pour lesquels le modèle a identifié le résultat comme positif (standard dans notre exemple) et le résultat réel était positif (également standard). C'est le nombre de « vrais positifs ». Nous comptons également le nombre de cas pour lesquels le modèle a correctement prédit le négatif comme des « Vrais négatifs ». Nous comptons également les erreurs lorsque le modèle a prédit un négatif mais le résultat était positif et inversement. La matrice de confusion ci-dessous est ainsi générée :

		Prédit	
Réel		Positif	Négatif
	Positif	Vrais positifs	Faux négatifs
	Négatif	Faux positifs	Vrais négatifs

Dans une matrice de confusion, les faux positifs sont également des erreurs de type I et les faux négatifs sont considérés comme des erreurs de type II.

On peut ensuite utiliser cette matrice de confusion pour classer le modèle à l'aide de divers paramètres :

$$Précision = \frac{Faux\ positifs + vrais\ négatifs}{Nombre\ total\ de\ cas}$$

$$Sensibilité = \frac{Vrais\ positifs}{Vrai\ Positifs + Faux\ négatifs} = \frac{Vrais\ Positifs}{Réllement\ positifs}$$

$$Spécificité = \frac{Vrais\ négatifs}{Vrais\ négatifs + Faux\ positifs} = \frac{Vrais\ négatifs}{Réellement\ négatifs}$$

La mesure de précision est une mesure globale de la précision. La sensibilité indique le nombre de positifs qui sont réellement identifiés comme tels. L'interaction de ces variables indique l'efficacité

d'un modèle. Par exemple, si nous avons deux modèles prédisant si des cas sont standard ou non, nous pouvons tabuler une matrice de confusion pour chaque modèle.

Le modèle A est un modèle très simple (et très imprécis) qui prédit que chaque cas sera standard.

Modèle A		Prédit	
		Standard	Non-standard
Réel	Standard	80	0
	Non-standard	20	0

D'après la matrice de confusion, nous pouvons estimer qu'il y a eu 100 cas. Notre modèle prédit que tous les cas seront standard. Notre sensibilité est alors de 100 % et notre précision semble satisfaisante à 80 %. Cependant, avec une spécificité à 0 %, cela indique que le modèle est peu performant.

Il convient de prendre en compte plusieurs mesures pour connaître l'efficacité d'un modèle. Un modèle plus réaliste ressemblerait à celui reproduit ci-dessous (sur les mêmes données) :

Modèle B		Prédit	
		Standard	Non-standard
Réel	Standard	61	19
	Non-standard	8	12

Dans ce cas, la précision globale est de 73 %, la sensibilité de 76,25 % et la spécificité de 60 %. Ce modèle semble raisonnable.

Scores des modèles et seuil

La plupart des modèles classifieurs ne produisent pas qu'une classification binaire comme résultat. Ils calculent généralement un score pour classer les cas comme positifs ou négatifs. Le score est en général converti en pourcentage. Ce score n'implique pas toujours forcément une vraie probabilité, notamment pour les techniques d'apprentissage automatique, et indique souvent qu'un classement des cas, plutôt qu'une probabilité stricte.

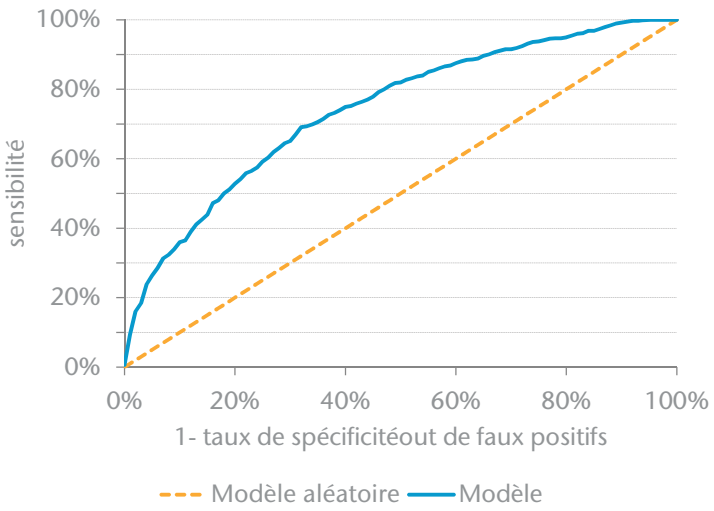
Sachant que chaque cas aurait un score, il conviendrait d'attribuer un seuil (ou une limite) au-delà duquel le résultat du modèle pourrait être considéré comme positif. Par exemple, un modèle pourrait donner différents scores pour différents cas, et avec un seuil de 80 %, traiterait un cas comme un cas standard uniquement lorsque le score est supérieur à 80 %.

Chacun de ces seuils suppose une matrice de confusion spécifique. Avec un seuil de 0 %, nous finirions avec un modèle A représenté ci-dessus, nous prédirons chaque cas comme standard, avec la sensibilité de 100 % mais une spécificité de 0 %. De la même manière, avec un seuil de 100 %, tous les cas seront classés comme non-standard, avec une sensibilité de 0 % et une spécificité de 100 %.

Fonction d'efficacité du récepteur ou courbe ROC

La modification du seuil d'un modèle donné sur toutes les valeurs entre 0 et 100 % permet de tracer une courbe des différentes valeurs de spécificité et de sensibilité. La courbe ROC trace la sensibilité (axe Y) par rapport à la spécificité (axe X) pour chaque valeur du seuil. Notez que la spécificité (1) est également le taux de faux positifs. La figure 1 est un exemple de cette courbe ROC.

Figure 1 – Courbe caractéristique du récepteur



En bas à gauche, on retrouve le cas où le seuil est de 100 %. C'est l'équivalent de notre modèle pour prédire tous les cas comme étant non-standard (sensibilité de 0 % mais spécificité de 100 % ou taux de faux positifs est de 0 %). En haut à droite, nous avons le cas de figure dans lequel tous les cas sont prédits comme standard (sensibilité de 100 % mais spécificité de 0 % ou taux de faux positifs est de 100 %).

La ligne diagonale en pointillés représente le résultat attendu pour un modèle attribuant des scores à des cas de façon aléatoire. La ligne bleue représente les résultats pour un modèle donné qui est plus fiable qu'une sélection aléatoire. La qualité globale du modèle pourrait être déterminée en évaluant la zone sous la courbe bleue. Dans ce cas, elle est estimée à 74 %, ce qui représente un modèle plutôt fiable.

Nous pouvons choisir deux cas aléatoires, un parmi les cas standard réels et un autre parmi des cas non-standard réels. Il a été démontré que la zone située sous la courbe (74 % dans cet exemple) équivaut à la probabilité que le cas standard obtienne un score supérieur à celui du cas non-standard. Cette zone fournit donc une mesure globale de l'efficacité du modèle pour opérer un tri entre les cas standard et les cas non-standard.

Le modèle aléatoire présente une zone sous la courbe de 50 %, ce qui est l'exigence minimale sur laquelle il convient de s'améliorer. Un modèle parfait aurait une zone sous la courbe équivalente à un, et le modèle produirait un point sur le coin supérieur gauche de la figure 1.

Généralement, les modèles avec une zone sous la courbe comprise entre 50 et 60 % sont considérés comme erronés, et les modèles dépassant 90 % comme très bons. Un résultat compris entre 70 et 80 % est considéré comme moyen à bon. Notez qu'un modèle avec une zone sous la courbe de moins de 50 % doit en fait être inversé, car il prévoit le résultat inverse avec une plus grande efficacité ; généralement la zone sous la courbe varie entre 50 % et 100 %.

Coefficient de Gini

Le coefficient de Gini sert également à évaluer les modèles classifieurs. Il est directement associé à la zone sous courbe ROC évoquée ci-dessus. Le coefficient de Gini est calculé à partir de la zone sous la courbe (AUC) sous la forme 2AUC–1. Une zone de 74 % sous la courbe donne un coefficient de Gini de 48 %, ce qui est honorable. Le coefficient de Gini varie entre 0 et 100 %, bien qu'il puisse être négatif, auquel cas le modèle devrait être inversé comme dans le cas d'une zone sous la courbe de moins de 50 %.

Comparaison des modèles

Différents modèles et techniques de modélisation aboutiraient à des résultats différents sur différentes données et différents scénarios de problème. Pour comparer deux modèles différents et sélectionner le meilleur, nous pouvons comparer chaque zone sous la courbe ROC (ou coefficients de Gini). On constate qu'un modèle A est meilleur si sa zone sous la courbe est plus grande que B.

Il existe toutefois certains critères. Si nous traçons les modèles et que la courbe ROC de A traverse la courbe de B, nous pouvons affirmer que A est toujours meilleur. À certains niveaux de sensibilité et de spécificité, B peut être meilleur. Toutefois, si la courbe ROC de A ne croise pas celle de B et que sa zone sous la courbe est plus grande, cela signifie que A est toujours meilleur que B.

Pour les modèles dont les courbes ROC et les zones sous les courbes sont proches, des considérations d'ordre pratique, comme les détails de l'implémentation, peuvent entrer en ligne de compte.

Modèles d'ensemble

Il existe différentes méthodes pour améliorer un modèle selon la technique utilisée. Une technique intéressante consiste à construire des modèles d'ensemble. Une fois que nous avons plusieurs modèles qui produisent un score pour un résultat donné, nous pouvons commencer à les combiner dans des configurations intéressantes pour produire des scores d'ensemble. Ceux-ci peuvent être utilisés pour améliorer la zone sous la courbe pour ces modèles.

Prenons par exemple un classifieur par forêts aléatoires et un modèle de régression logique, qui prédisent les deux des risques standard. Un nouveau score peut être calculé comme la moyenne de ces deux classifieurs et l'évaluer ensuite comme un autre modèle. La zone sous la courbe s'améliore généralement pour ces modèles d'ensemble.

Nous pourrions également construire un modèle de modèles en intégrant une combinaison de modèle de régression logique aux différents modèles sous-jacents modélisant le résultat final, en utilisant principalement les données pour suggérer une pondération des différents modèles.

Un nouvel échantillon de données de validation (non utilisé pour l'entraînement ou des tests) serait généralement retenu pour produire ce type de modèle d'ensemble.

Optimisation des affaires

Pour un modèle avec une courbe ROC donnée, nous pouvons maintenant décider de la manière d'appliquer ce modèle en pratique. Sachant que chaque type d'erreur (faux positifs et faux négatifs) aurait des coûts et des avantages pour l'entreprise, nous pouvons ensuite estimer le point auquel le coût est réduit ou auquel l'avantage est maximisé, fixant ainsi le meilleur seuil pour chaque modèle. Ces résultats peuvent nous permettre de comparer les meilleures performances des différents modèles pour prendre une décision finale sur un modèle en particulier.

Dans l'exemple de la modélisation des décisions de tarification standard, nous pouvons prendre en compte les implications de chaque catégorie de résultat sur la valeur actualisée des profits par demande :

- Les vrais positifs (cas correctement classés par le modèle comme standard) pourraient voir la valeur par police augmenter, car nous constaterions des placements plus élevés de cette catégorie en raison de frais de vente par police plus faibles (taux de conversion plus élevé du fait d'un processus rationalisé pour le client) et des frais de sélection des risques et médicaux plus bas.
- Les faux positifs (cas classés à tort comme standard) peuvent être confrontés à des problèmes d'augmentation des sinistres par rapport aux primes. Il existe également un risque de sélection adverse si les demandeurs savent comment influencer leurs scores.
- Les faux négatifs (cas qui sont classés à tort comme non-standard) peuvent être très proches du processus actuel. Il se peut donc que nous devions utiliser un taux de conversion actuel et des taux d'utilisation pour évaluer la valeur actualisée des profits par demande (en supposant que l'underwriting suive notre approche actuelle).
- Les vrais négatifs (cas correctement classés comme négatifs) peuvent encore une fois être cohérents avec notre traitement actuel des cas non-standard.

Au vu des valeurs pour chaque élément ci-dessus, il est possible d'estimer un seuil optimal pour notre modèle, seuil qui apportera une valeur maximale par demande à l'entreprise. Cela correspondrait à un point unique sur la courbe ROC. La sensibilité du résultat sélectionné doit être testée par rapport aux hypothèses car la plupart de ces dernières, établies pour déterminer la valeur des différents scénarios, pourraient être subjectives.

Si nous disposons de plusieurs modèles (où les courbes ROC se croisent comme précédemment), ou un modèle pour lequel nous ne pouvons pas calculer la courbe ROC, nous pouvons alors comparer la meilleure valeur produite à partir des différents modèles pour déterminer le meilleur d'entre eux. Pour comparer ces valeurs, il convient de s'assurer que le coût de fonctionnement du modèle est inclus. Certains modèles peuvent avoir des implications en matière de coûts, soit du fait des données utilisées, soit en raison de problèmes d'application techniques.

Conclusion

Cet article a présenté certaines approches permettant d'évaluer la performance des modèles utilisés dans le cadre de problèmes de classification.

Nous avons apporté les éclaircissements suivants :

- Il est nécessaire d'avoir des échantillons de données pour les tests.
- Il est possible de comparer la performance de ces modèles en utilisant des mesures dans une matrice de confusion.
- Il est possible d'estimer certaines données, comme la spécificité et la sensibilité à partir de ces matrices de confusion.
- On peut également envisager d'utiliser la courbe ROC et la zone sous la courbe pour se faire une idée de la qualité du modèle.
- Des techniques d'ensemble peuvent être utilisées pour combiner des scores de différents modèles pour produire de meilleurs modèles.
- Une fois que nous appliquons des valeurs commerciales attendues pour différents résultats du modèle, il peut être plus simple de déterminer la valeur qui représenterait le meilleur compromis entre la sensibilité et la spécificité pour l'entreprise.

- Il peut y avoir des considérations pratiques à prendre en compte, notamment des hypothèses subjectives dans l'évaluation de divers résultats et des détails techniques de mise en œuvre qui peuvent donner lieu à des résultats différents.
- La modification du processus peut entraîner des changements de comportement, ce qui invalide la modélisation. Dans ce contexte, il convient d'évaluer les comportements particulièrement anti-sélectifs.

Cette vue d'ensemble relativement simple donne une idée de la manière d'évaluer ces modèles objectivement et de la façon d'évaluer la valeur du modèle pour l'entreprise en tenant compte de sa performance.

Références :

James, G., Witten, D., Hasties, T.& Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Matloff, N. (2017). Statistical Regression and Classification: From Linear Models to Machine Learning. CRC Press.

Fawcett, T. (2006) An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

L'auteur

Louis Rossouw *occupe les postes d'actuaire analyste et de chargé d'études au sein du bureau de Gen Re au Cap pour lequel il est chargé des marchés britannique et sud-africain. Louis a rejoint la société en 2001, après avoir travaillé sur la mise au point des barèmes individuels et sur le développement de produits, la tarification de groupe et les provisions. Il a également passé deux ans au sein de la filiale de Gen Re à Singapour en tant qu'actuaire régional en chef. Il peut être contacté par tél. au +27 21 412 7712 ou par courriel lrossouw@genre.com.*

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

General Reinsurance AG–Succursale Paris
21, rue Balzac
75008 Paris
Tel. +33 1 5367 7676
Fax +33 1 5367 46464

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

Photos: © getty images – underworld111, aedkais, Fredex8