Créer des modèles statistiques plus rapidement et facilement

Paris, 26 Septembre 2012

Sam Gardner

jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW®

Questionnaire

- Comment définiriez vous votre niveau en statistique?
  - 1 = très faible
  - 2 = faible
  - 3 = moyen
  - 4 = bon
  - 5 = très bon
- Quand utilisez-vous les statistiques pour vous aider dans vos décisions?
  - 1 = moins d'une fois par mois
  - 2 = une ou plusieurs fois par mois
  - 3 = une ou plusieurs fois par semaine
  - 4 = une ou plusieurs fois par jour
- Etes-vous familier avec l'utilisation de JMP?
  - Oui / Non

Agenda

- Moyen d'établir de meilleurs modèles
- Méthodes de modélisation statistiques usuelles
  - Arbres de décision
  - Régression
  - Réseau de neurones
- Approches de modélisation statistiques avancées
  - Pas à pas
  - Boosting
  - Modèle moyen, dont les forêts aléatoires
- Cas d'études
  - Ingénierie, R&D, Marketing, Risque

Qu’est ce qu’un modèle statistique?

- Un modèle empirique décrit les données
- Séparer les variations d’une réponse (Y) entre des éléments prédictifs (f(X)) et non prédictifs (erreur résiduelle)

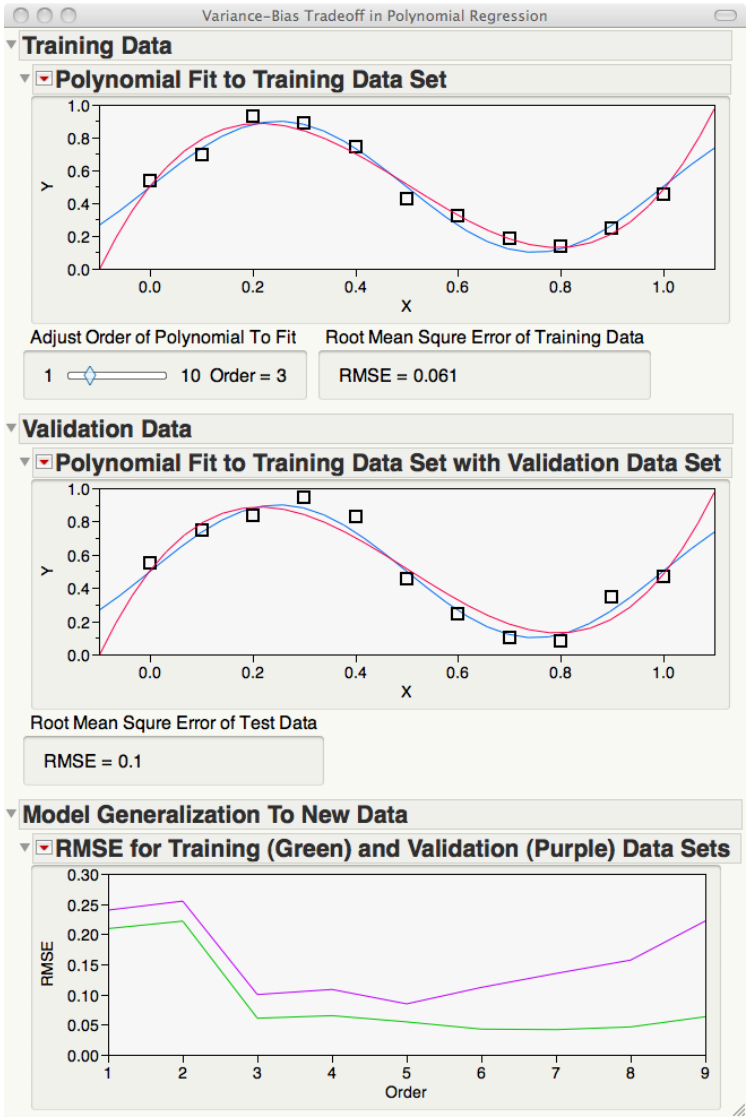
$$Y = f(X) + \text{erreur résiduelle}$$

  - Y est une ou un ensemble de réponses continues ou catégorielles
  - X est une ou un ensemble de prédicteurs continus ou catégoriels
  - *f(X)* décrit la variation prédictive de Y
  - L’erreur décrit la variation non prédictive de Y

Identification d'un modèle statistique exploitable

- "All models are wrong, but some are useful", George Box
- Comment certifier si nos modèles sont utiles?
- Comment éviter de créer des modèles à l'aspect vraiment performant mais qui peuvent induire en erreur?
  - Ou, autrement dit, comment éviter les problématiques de sur-ajustement en attribuant trop de variation de Y à f(X)?

Holdback, outil pour gérer le sur-ajustement



- Holding garde des données non utilisées pour ajuster le modèle.
- Ces données sont utilisées pour sélectionner le modèle (plus petite erreur ou meilleur R2)
- Un troisième sous-ensemble (nommé Test) peut de plus être utilisé afin de vérifier la capacité de généralisation du modèle construit

Options de validation de modèle

- Les grands jeux de données utilisent la méthode de holdback pour en extraire aléatoirement deux ou trois sous-ensemble:
  - Apprentissage: Utilisé pour construire (ajustement ou estimation) le modèle.
  - Validation: utilisé pour sélectionner le meilleur modèles
    - » i.e. construire un modèle f(X) sans sur-ajustement
  - Test: Utilisé pour tester la qualité du modèle.
    - » Donne une idée réaliste de la capacité du modèle à être généralisable sur de nouveaux jeux de données.

Options de validation de modèles (Suite)

- Les petits jeux de données utilisent la méthode k-fold:
  - Divise aléatoirement les données en k groupes séparés
  - Extrait un des échantillon et construit le modèle sur les autres échantillons.
  - L'échantillon extrait est prédit avec le modèle et les mesures sont enregistrées. Ceci est répété pour chaque échantillon.
  - Une erreur moyenne est estimée à travers les différents échantillon et le modèle avec la plus petite erreur est conservé.

Modèles statistiques

- Nous allons utiliser une approche basée sur des cas d'étude pour introduire les différentes techniques usuelles de modélisation statistique en utilisation les approches de validation précédentes:
  - Types de modèles
    - » Arbres de décision
    - » Régression
    - » Réseaux de neurones
  - Approches
    - » Pas à pas
    - » Boosting
    - » Bootstrap et modèles moyens (forêts aléatoires)

Questions?

Etude de cas 1: Arbre de Decision

Exemple R&D

jmp Statistical Discovery.™ From SAS.

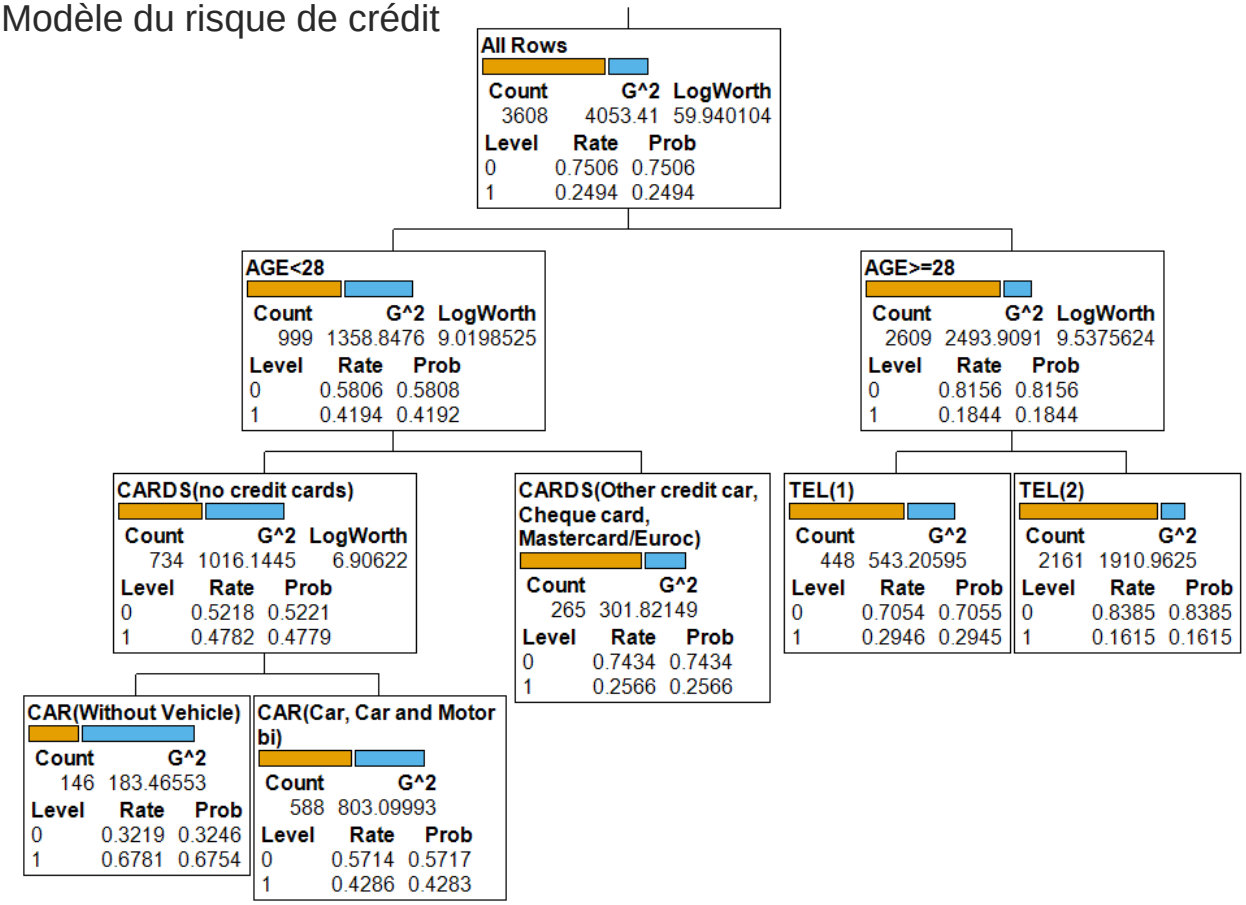
sas THE POWER TO KNOW.

Arbre de Décision

- Aussi connu sous le nom de partition récursif, CHAID, CART
- Les modèles sont une série de déclarations imbriquées IF (), où chaque déclaration IF () peut être considérée comme une branche distincte dans un arbre.
- Les branches sont choisis de telle sorte que la différence dans la réponse moyenne (ou taux de réponse moyenne) entre les branches appariés est maximisée.
  - Faire ainsi, attribue plus de variation de Y en fonction de *X*.
- Algorithme devient plus compliqué et les calculs plus intensives avec la retenue de données (holdback).

Arbre de Décision

Exemple rapide: Modèle du risque de crédit



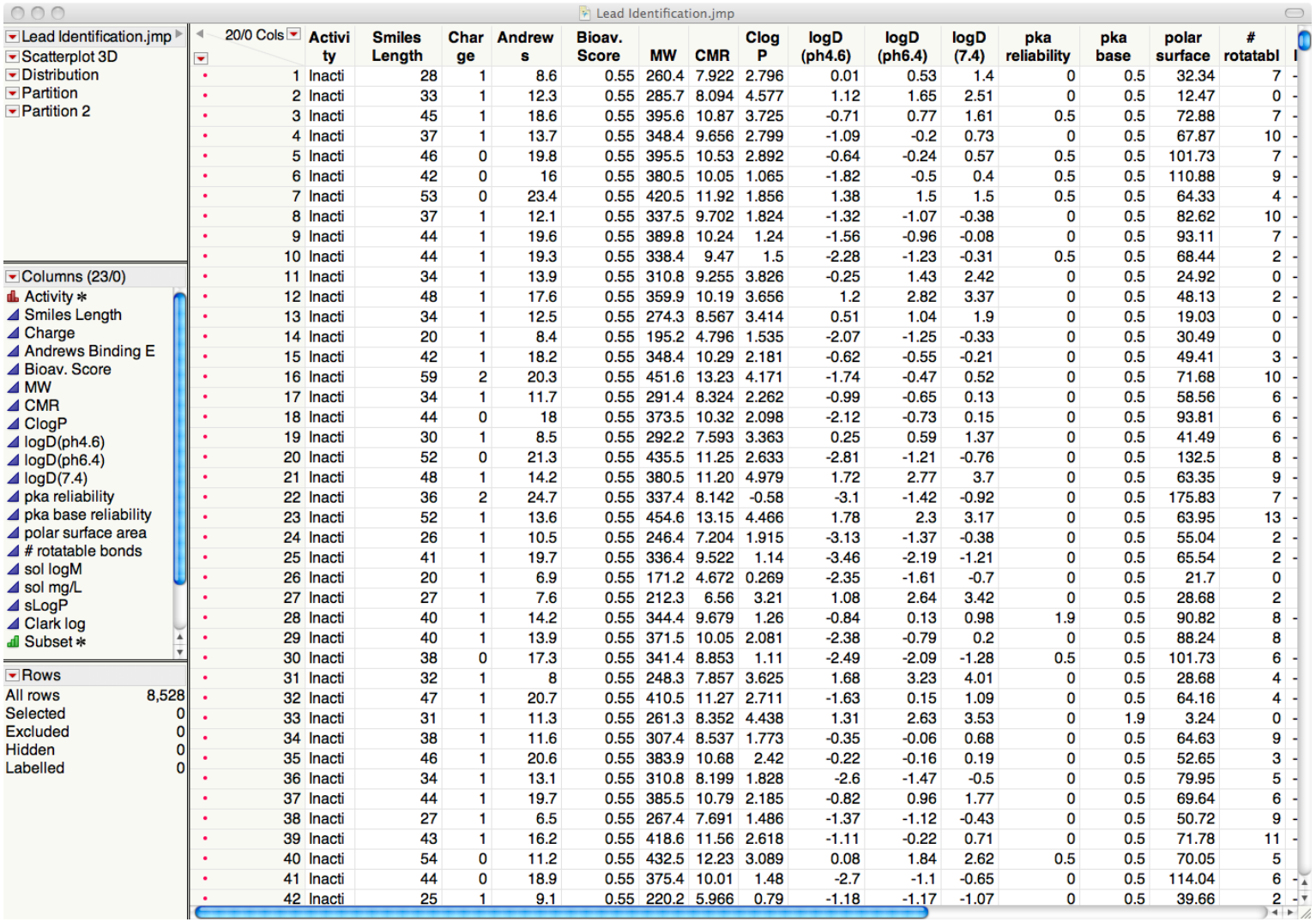
Evaluation du modèle

- Modèles à réponses continues sont évaluées à l'aide de SSE (somme de l'erreur quadratique), des mesures telles que R^2, R^2 ajusté.
  - D'autres alternatives sont des mesures d'information du type AIC et BIC.
- Modèles à réponses catégoriques sont évalués par leur capacité de:
  - trier les données, en utilisant des courbes ROC et les courbes Lift
  - classer une nouvelle observation mesurée a partir des matrices de confusion, ainsi que le taux globale de classification erronée
  - La troisième étude de cas montrera des exemples de cela
- Pour une discussion plus détaillée sur les critères de sélection de modèles, voir *Gardner, S., "Model Selection: Part 1 – Model Selection Criteria", ASQ Statistics Division Newsletter, Volume 29, No. 2, Winter, 2011,* *http://asqstatdiv.org/newsletterarch.php*

Etude de cas R&D: Conception et découverte de composant chimique actif

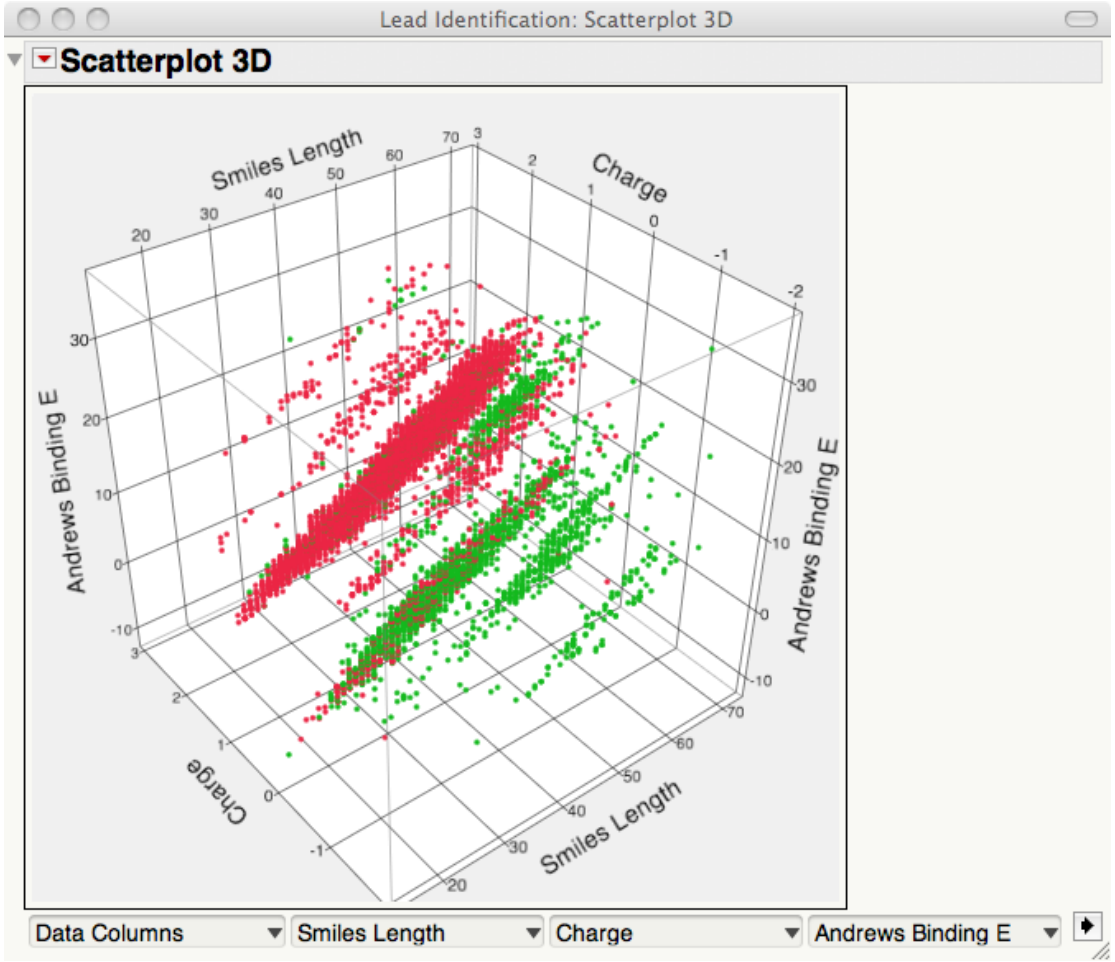
- Découverte de composant chimique actif.
- Echantillon de 8528 composant chimiques d'une banque de données a plus de 2M de composants.
- 18 descripteurs chimiques (ou propriétés).
- Exemple de composants chimiques testés contre une protéine nouvellement identifiée associée à une maladie, afin de déterminer lequel des composants chimiques présentent une activité biologique potentielle contre cette protéine (et potentiellement contre la maladie).
- Résultat obtenu est appelé « activité » et prend des valeurs de « inactifs » ou « actifs ».
- L'objectif est de construire un modèle permettant de prédire les composants chimiques actifs et ensuite utiliser le modèle prédictif pour sélectionner les composants chimiques les plus susceptibles d'être actif à partir de la base de données.

Les données sont simulées



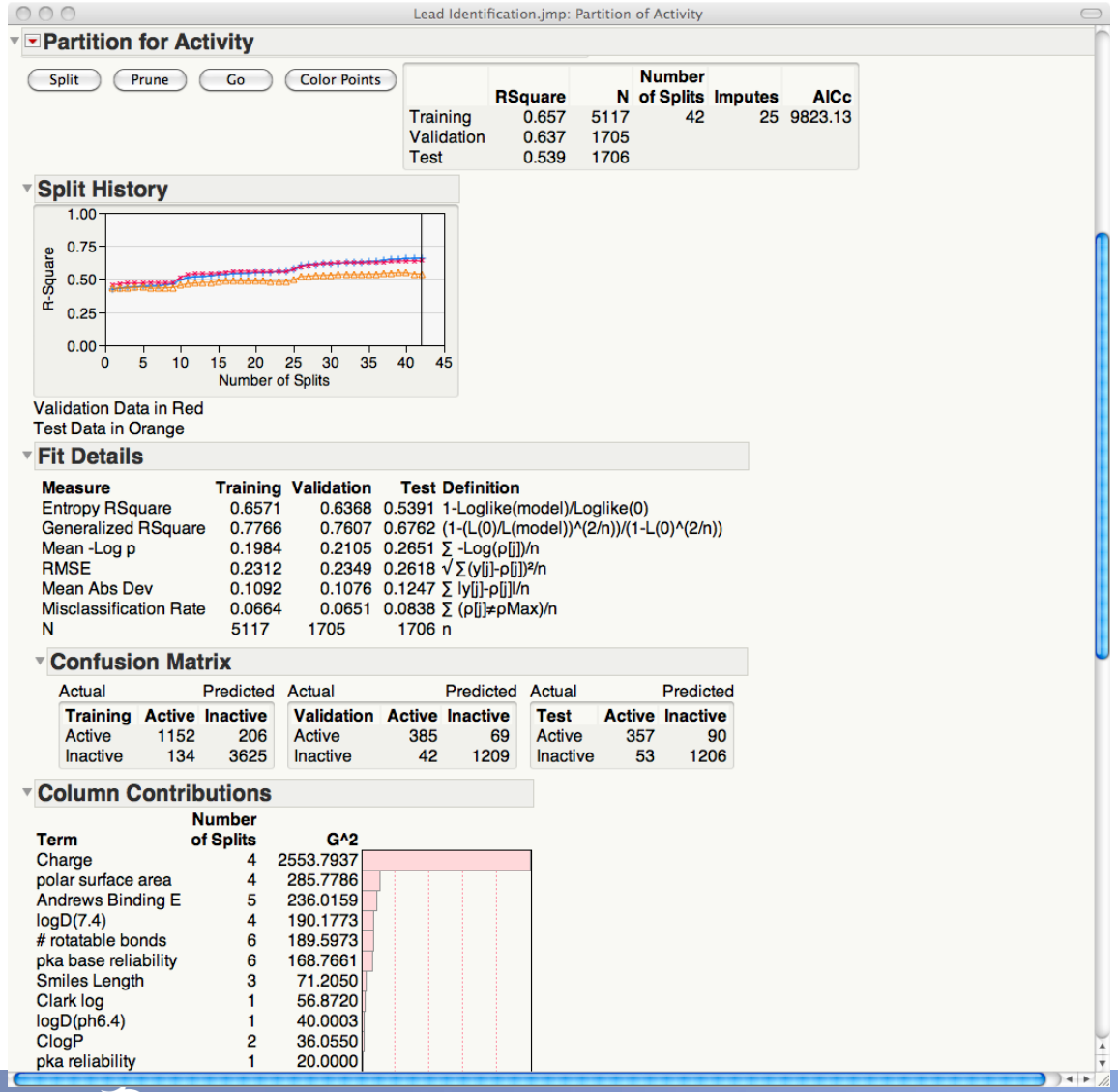
	Activity	Smiles Length	Charge	Andrews	Bioav. Score	MW	CMR	ClogP	logD (ph4.6)	logD (ph6.4)	logD (7.4)	pka reliability	pka base	polar surface	# rotatabl
1	Inacti	28	1	8.6	0.55	260.4	7.922	2.796	0.01	0.53	1.4	0	0.5	32.34	7
2	Inacti	33	1	12.3	0.55	285.7	8.094	4.577	1.12	1.65	2.51	0	0.5	12.47	0
3	Inacti	45	1	18.6	0.55	395.6	10.87	3.725	-0.71	0.77	1.61	0.5	0.5	72.88	7
4	Inacti	37	1	13.7	0.55	348.4	9.656	2.799	-1.09	-0.2	0.73	0	0.5	67.87	10
5	Inacti	46	0	19.8	0.55	395.5	10.53	2.892	-0.64	-0.24	0.57	0.5	0.5	101.73	7
6	Inacti	42	0	16	0.55	380.5	10.05	1.065	-1.82	-0.5	0.4	0.5	0.5	110.88	9
7	Inacti	53	0	23.4	0.55	420.5	11.92	1.856	1.38	1.5	1.5	0.5	0.5	64.33	4
8	Inacti	37	1	12.1	0.55	337.5	9.702	1.824	-1.32	-1.07	-0.38	0	0.5	82.62	10
9	Inacti	44	1	19.6	0.55	389.8	10.24	1.24	-1.56	-0.96	-0.08	0	0.5	93.11	7
10	Inacti	44	1	19.3	0.55	338.4	9.47	1.5	-2.28	-1.23	-0.31	0.5	0.5	68.44	2
11	Inacti	34	1	13.9	0.55	310.8	9.255	3.826	-0.25	1.43	2.42	0	0.5	24.92	0
12	Inacti	48	1	17.6	0.55	359.9	10.19	3.656	1.2	2.82	3.37	0	0.5	48.13	2
13	Inacti	34	1	12.5	0.55	274.3	8.567	3.414	0.51	1.04	1.9	0	0.5	19.03	0
14	Inacti	20	1	8.4	0.55	195.2	4.796	1.535	-2.07	-1.25	-0.33	0	0.5	30.49	0
15	Inacti	42	1	18.2	0.55	348.4	10.29	2.181	-0.62	-0.55	-0.21	0	0.5	49.41	3
16	Inacti	59	2	20.3	0.55	451.6	13.23	4.171	-1.74	-0.47	0.52	0	0.5	71.68	10
17	Inacti	34	1	11.7	0.55	291.4	8.324	2.262	-0.99	-0.65	0.13	0	0.5	58.56	6
18	Inacti	44	0	18	0.55	373.5	10.32	2.098	-2.12	-0.73	0.15	0	0.5	93.81	6
19	Inacti	30	1	8.5	0.55	292.2	7.593	3.363	0.25	0.59	1.37	0	0.5	41.49	6
20	Inacti	52	0	21.3	0.55	435.5	11.25	2.633	-2.81	-1.21	-0.76	0	0.5	132.5	8
21	Inacti	48	1	14.2	0.55	380.5	11.20	4.979	1.72	2.77	3.7	0	0.5	63.35	9
22	Inacti	36	2	24.7	0.55	337.4	8.142	-0.58	-3.1	-1.42	-0.92	0	0.5	175.83	7
23	Inacti	52	1	13.6	0.55	454.6	13.15	4.466	1.78	2.3	3.17	0	0.5	63.95	13
24	Inacti	26	1	10.5	0.55	246.4	7.204	1.915	-3.13	-1.37	-0.38	0	0.5	55.04	2
25	Inacti	41	1	19.7	0.55	336.4	9.522	1.14	-3.46	-2.19	-1.21	0	0.5	65.54	2
26	Inacti	20	1	6.9	0.55	171.2	4.672	0.269	-2.35	-1.61	-0.7	0	0.5	21.7	0
27	Inacti	27	1	7.6	0.55	212.3	6.56	3.21	1.08	2.64	3.42	0	0.5	28.68	2
28	Inacti	40	1	14.2	0.55	344.4	9.679	1.26	-0.84	0.13	0.98	1.9	0.5	90.82	8
29	Inacti	40	1	13.9	0.55	371.5	10.05	2.081	-2.38	-0.79	0.2	0	0.5	88.24	8
30	Inacti	38	0	17.3	0.55	341.4	8.853	1.11	-2.49	-2.09	-1.28	0.5	0.5	101.73	6
31	Inacti	32	1	8	0.55	248.3	7.857	3.625	1.68	3.23	4.01	0	0.5	28.68	4
32	Inacti	47	1	20.7	0.55	410.5	11.27	2.711	-1.63	0.15	1.09	0	0.5	64.16	4
33	Inacti	31	1	11.3	0.55	261.3	8.352	4.438	1.31	2.63	3.53	0	1.9	3.24	0
34	Inacti	38	1	11.6	0.55	307.4	8.537	1.773	-0.35	-0.06	0.68	0	0.5	64.63	9
35	Inacti	46	1	20.6	0.55	383.9	10.68	2.42	-0.22	-0.16	0.19	0	0.5	52.65	3
36	Inacti	34	1	13.1	0.55	310.8	8.199	1.828	-2.6	-1.47	-0.5	0	0.5	79.95	5
37	Inacti	44	1	19.7	0.55	385.5	10.79	2.185	-0.82	0.96	1.77	0	0.5	69.64	6
38	Inacti	27	1	6.5	0.55	267.4	7.691	1.486	-1.37	-1.12	-0.43	0	0.5	50.72	9
39	Inacti	43	1	16.2	0.55	418.6	11.56	2.618	-1.11	-0.22	0.71	0	0.5	71.78	11
40	Inacti	54	0	11.2	0.55	432.5	12.23	3.089	0.08	1.84	2.62	0.5	0.5	70.05	5
41	Inacti	44	0	18.9	0.55	375.4	10.01	1.48	-2.7	-1.1	-0.65	0	0.5	114.04	6
42	Inacti	25	1	9.1	0.55	220.2	5.966	0.79	-1.18	-1.17	-1.07	0	0.5	39.66	2

Exploration graphique des relations



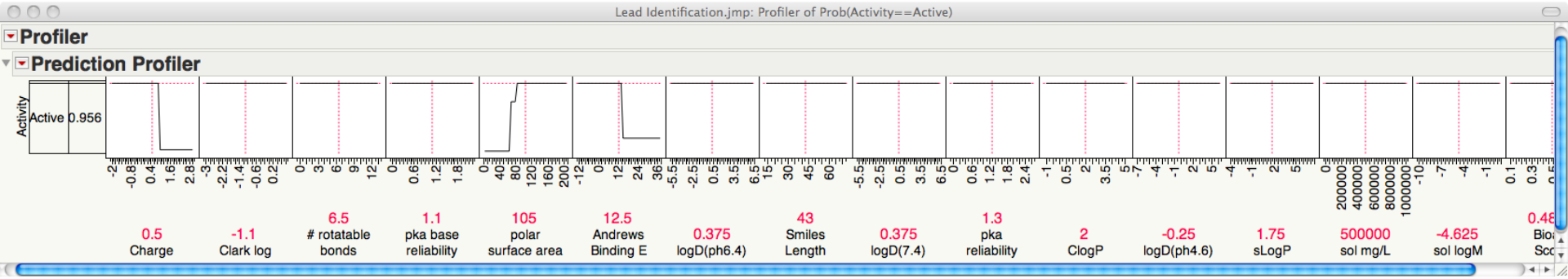
- Nous aimerions identifier les facteurs prédictifs les plus susceptibles de séparer ou distinguer les points rouges (inactif) des points verts (actif).
- JMP nous permet de changer les prédicteurs sur les trois axes en utilisant les listes de sélection au bas du graphique.
- Cependant, le nombre de combinaisons de 3 à partir de 18 prédicteurs est trop grand pour nous permettre d'identifier rapidement les principales variables prédictives fiables, à l'aide seulement d'outils de visualisation.

Arbre de Décision avec segmentation (holdback) pour la sélection et test du modèle



- 42 divisions sont requises pour maximiser le R carre de validation a 0.636
- Modèle qui en résulte prédit 53,9% de la variation pour les valeurs de l'activité de données de test (Test Entropie R Carre de 0.539).
- Matrice de confusion pour les données de test montre que le modèle prédit correctement 357 de 447 substances chimiques actives et 1206 de 1259 produits chimiques inactifs.
- Contributions des colonnes donne un classement des prédicteurs.

Conclusions: Interprétation du modèle



- Le profileur permet de comprendre les relations et établir des critères de recherche pour identifier d'autres produits chimiques potentiellement actifs au sein de la base de données de 2M de produits chimiques, de mieux cibler les produits chimiques dans des essais complémentaires, et plus spécifiquement:
  - Charge faible
  - Grande surface polaire
  - Faible Andrews Binding E

Cas d'étude 2: Régression pas à pas (K-fold)

Exemple R&D avec un petit jeu de données

jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW.

Régression (réponse continue)

$$Y = f(X_1, X_2, \ldots, X_k)$$

- Exemples

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_k X_k$$

$$Y = a_0 + \sum_i a_i X_i + \sum\sum_{i<j} a_{ij} X_i X_j$$

Régression (réponse catégorielle)

$$P[Y = \text{target}] = f(X_1, X_2, ..., X_k)$$

- Exemple –Régression Logistique

$$P[Y = \text{target}] = \frac{1}{1 + e^{-f(X_1, X_2, ..., X_k)}}$$

$$f(X_1, X_2, ..., X_k) = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_k X_k$$

Pas à pas

- Ajoute ou élimine automatiquement des termes de modélisation à *f(X).*
- Appelé “pas à pas” car cela fonctionne de manière séquentielle par ajout ou suppression d’un petit nombre de terme par étape.
- La procédure pas à pas prendra de nombreuses étapes pour construire un modèle final.
- Voir Gardner, S. “Model Selection: Part 2 – Model Selection Procedures “, ASQ Statistics Division Newsletter, Volume 29, No. 3, Spring, 2011, http://asqstatdiv.org/newsletterarch.php, pour avoir plus de détails sur la régression logistique sur la modélisation d’une réponse continue.

Cas d'étude R&D : Développement de produit

- Première année d'essais de développement pour un nouveau supplément alimentaire destiné à augmenter la prise de poids dans un temps de plus en plus court afin d'améliorer la production de volaille (dinde, poulet, etc).
- Le gain de poids a été contrôlé précédemment dans des conditions de laboratoire fortement contrôlées.
- Un but des essais de développement est de voir les performances du produit dans des conditions de vie réelles dans différentes régions géographiques et avec différentes variété de dindonneau afin de valider la robustesse du produit selon les variétés.

Cas d'étude R&D : Développement de produit

- Incertitude concernant la dose optimale ainsi que la formulation de ce complément alimentaire
- L'Objectif des essais de développement de la première année est de déterminer:
  - La meilleure dose de ce complément alimentaire
  - La meilleure formulation de ce complément alimentaire
  - L'application des même performances selon les variétés (afin d'avoir une vision de la taille du marché potentiel)

Données masquées

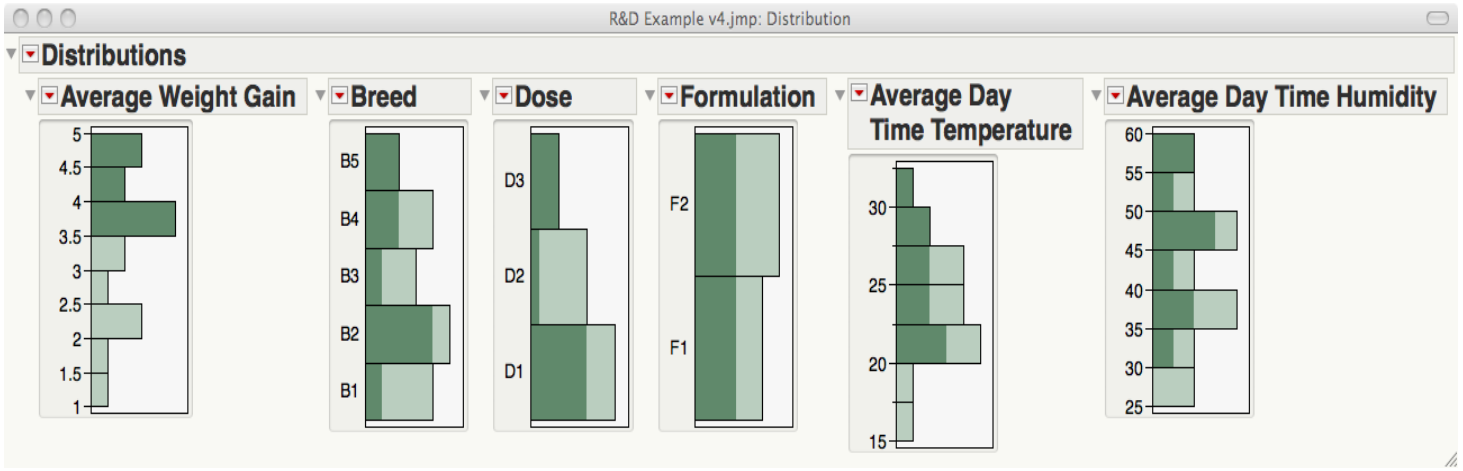
R&D Example v4.jmp

6/0 Cols
18/0

	Breed	Dose	Formulation	Average Day Time Temperature	Average Day Time Humidity	Average Weight Gain
1	B1	D2	F1	20.7	36.2	1.85
2	B2	D1	F2	25.2	46.8	3.72
3	B3	D2	F2	15.9	26.4	2.18
4	B4	D2	F2	17.9	27.2	2.26
5	B3	D2	F1	20	33	3.52
6	B2	D1	F2	21	34.8	3.4
7	B2	D1	F1	25.2	47	4.38
8	B5	D3	F1	20.8	38.6	4.04
9	B1	D1	F1	23.5	44	2.09
10	B5	D1	F1	20.3	36.2	4.57
11	B4	D1	F2	23.8	35.8	2.9
12	B4	D3	F2	29.1	53.6	3.81
13	B4	D3	F1	30.3	57.6	3.69
14	B2	D1	F2	24.5	47.2	3.95
15	B1	D2	F1	25	50.4	1.11
16	B3	D2	F2	26.2	47.8	3.12
17	B1	D1	F2	23.3	41.6	4.77
18	B2	D1	F2	28.2	57.2	4.64

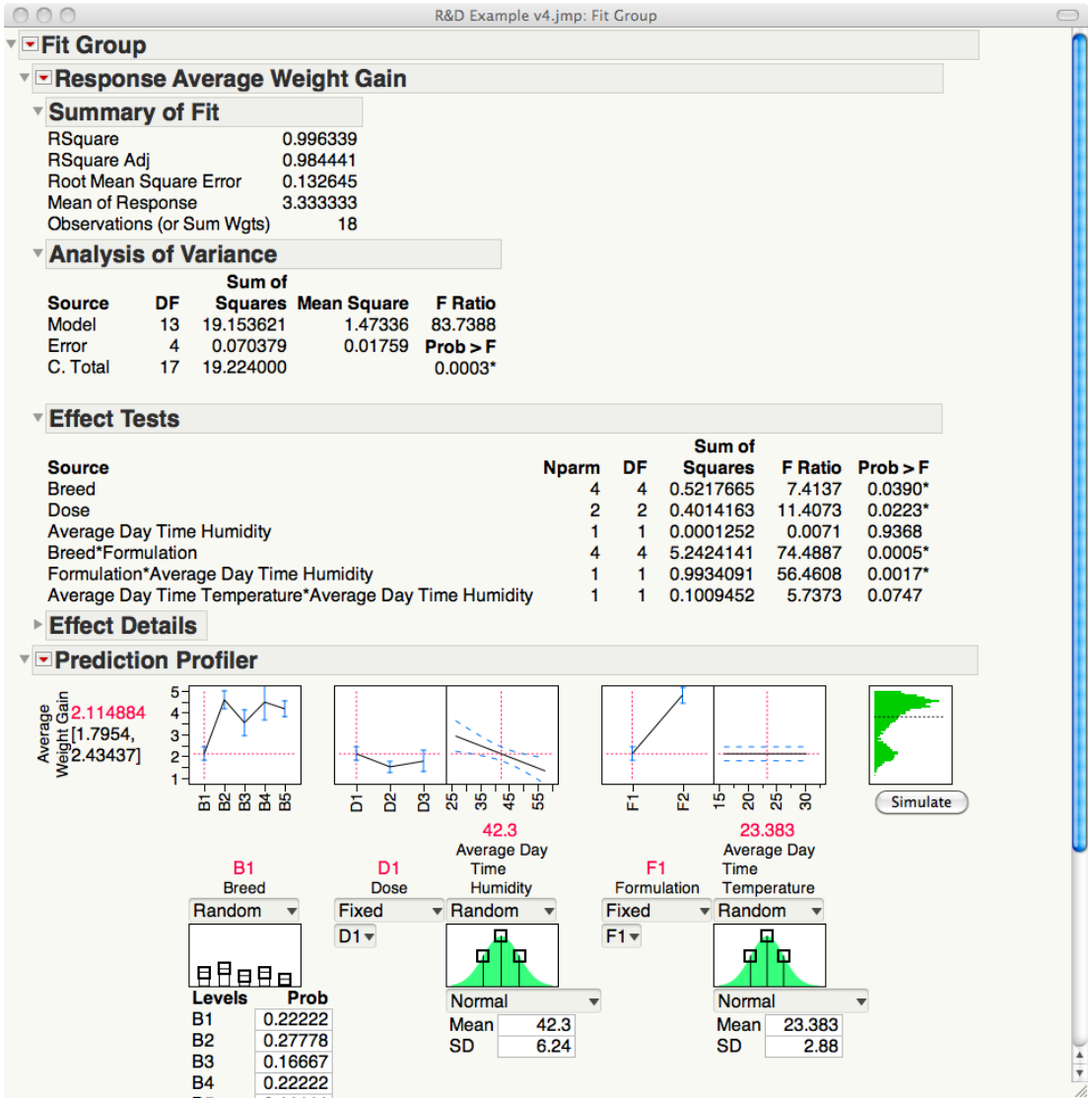
- Chaque ligne représente un résumé d'un essai de complément alimentaire dans une zone géographique (18 zones mondiales).
- Chacun des essais a été fait sur une des trois doses, sur une des deux formulation et sur une des cinq races. Le gain de poids moyen a été enregistré à la fin des essais, ainsi que la température de jour moyenne et l'humidité

Compréhension graphique des relations



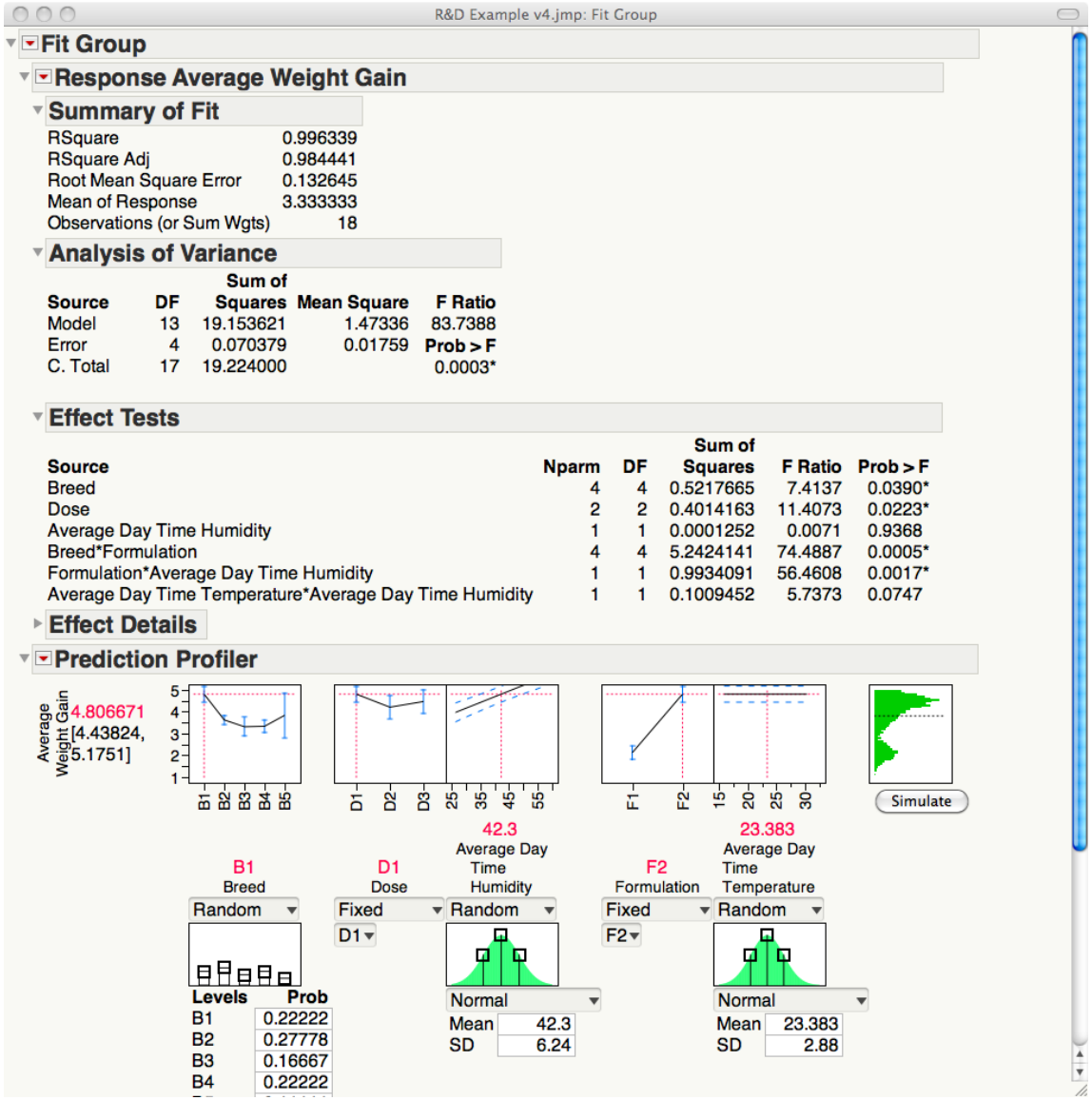
- Les fortes prises de poids ont été sélectionnées et sont apparent plus foncés
- Un prédicteur potentiel aura lui aussi sa zone foncée localisée dans une ou plusieurs région de son graphique.
- La dose D2 et les variétés B1 et B3, peu de température et d'humidité semblent associés avec une prise faible de poids (zone plus claires des graphiques).

Régression pas à pas avec validation K-Fold



- Le modèle exprime 98% de la variation de la prise de poids.
- Les variétés, doses, formulations, humidité et températures impactent tous la prises de poids, directement ou en interaction avec d'autres facteurs.
- Le profileur permet de comprendre et d'utiliser le modèle pour prendre des décisions.

Conclusions



- Changer la formulation de F1 par F2 montre une forte interaction entre la variété et l'humidité..
- La variété B1 a une meilleure prise de poids avec la formulation F2 que F1
- Des discussions avec le groupe de formulation a suggéré une problématique de stabilité de la formulation F1 dans des milieux fortement humides. Et une étude de cette stabilité a été envisagée.
- Les résultats de cette étude déterminera la viabilité de la formulation F1.
- Par contre, basé sur ces résultats, la formulation de F2 semble viable, quelque soit la dose. Et de nouveaux essais comparant la dose économique de F2 avec des formulation concurrentes sont envisagées.

Questions?

Etude de Cas 3: Forêt aléatoire et boosted trees

Exemple marketing: la rétention de clients

jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW.

Foret Aléatoire

- Une autre approche consiste à construire de nombreux modèles et de calculer la moyenne de l'ensemble pour obtenir un modèle global qui a un meilleur choix que tout autre modèles individuels.
- Une approche à la construction de «nombreux» modelés est d'utiliser des méthodes de type bootstrap agrégation (parfois appelée «bagging»).
- Un échantillon bootstrap (un ré-échantillonnage de 100% des données avec remplacement) est généré et un modèle est construit sur cet échantillon bootstrap.
  - Ceci est répété beaucoup de fois.
  - Le modèle final est la moyenne de tous les modèles bootstrap générés.
- JMP utilise une technique appelée Bootstrap Forest (BSF)
  - Ce type de méthode est également connue en tant que foret aléatoire.
  - Dans un BSF, des sous-ensembles de variables sont échantillonnés à chaque étape de la construction d'un arbre
  - Cela permet a des prédicteurs potentiellement faiblement corrélées à jouer un rôle dans le modèle, plutôt que les prédicteurs fortement corrélés.

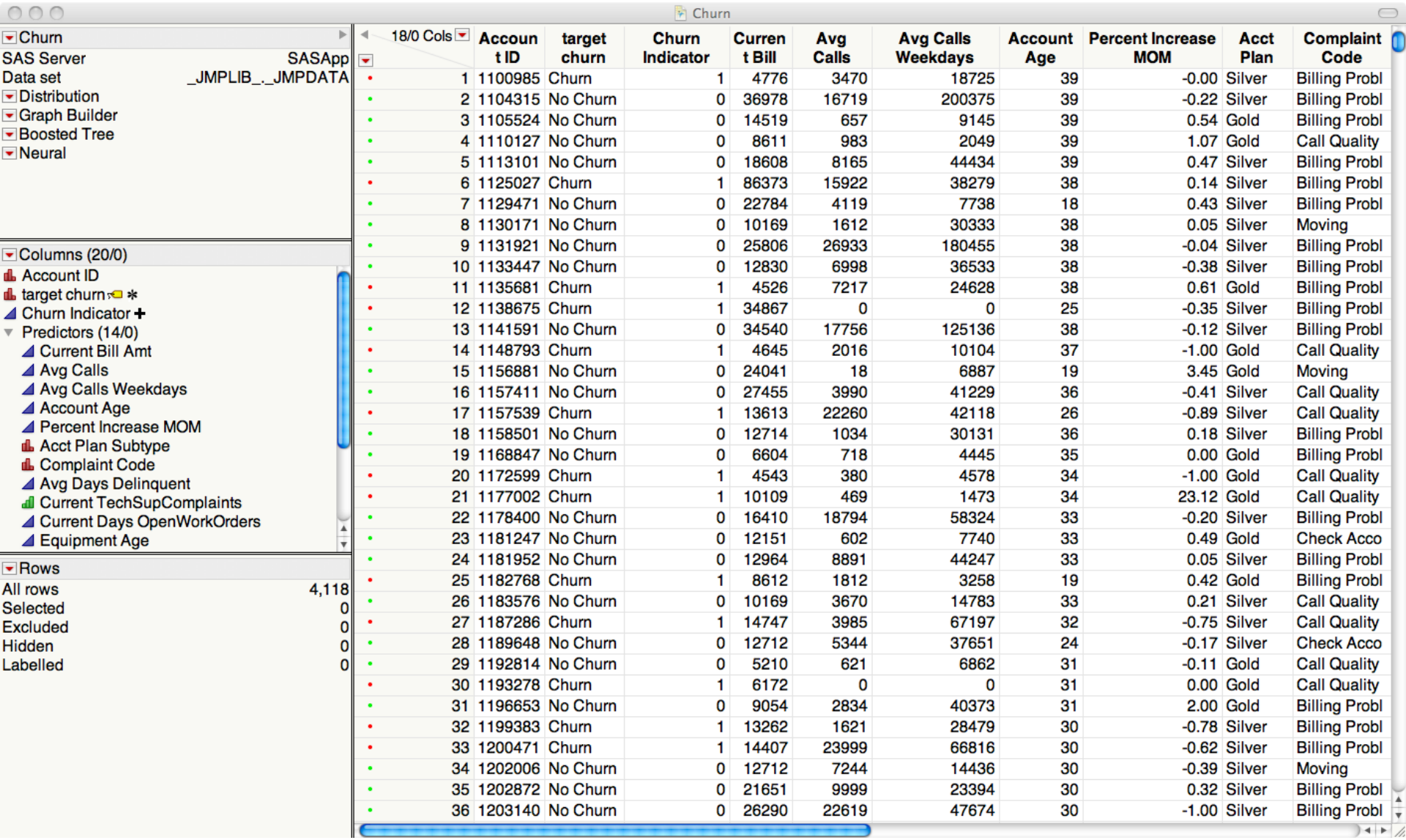
Boosting

- Le Boosting (ou Gradient Boosting) est une nouvelle conception en data mining, où les modèles sont construits en couches.
- Chaque couche du modèle «apprend de manière faible»
  - En d'autre mot, ces modèles prédisent la réponse plutôt mal
  - Et sont souvent de simples et petits modèles
- En commençant par la première couche, le modèle « faible » est ajusté, et les résidus sont calculés à partir de ce modèle.
- La couche suivante du modèle est ajustée aux résidus de la couche précédente, et ces nouveaux résidus sont enregistrés à partir de ce nouveau modèle ajusté.
- Cela continue jusqu'à ce qu'un certain nombre de couches ont été ajustées, et ou une décision a été faite tel que l'ajout de couches successives n'améliore plus le modèle.
- Le modèle final est l'accumulation de toutes les couches du modèle.

Exemple marketing: la rétention de clients

- L’exemple se base de données venant d’une société de télécommunication (téléphone portable) composée de 4,118 dossiers clients.
- La variable d'intérêt est la colonne appelée « churn » ou le taux de désabonnement, qui prend deux valeurs:
  - “Churn” pour indiquer qu’un client a changé de fournisseur et
  - “No Churn” pour indiquer qu’un client utilise encore "notre" service.
- Objectif: construire un modèle permettant de prédire la proportion de désabonnement et utiliser le modèle afin de conserver davantage de clients à l'avenir.
- Les facteurs prédictifs potentiels sont « Montant de la facture » jusque « Etat ». Indicateur Churn est une variable alternative numérique qui cible le taux de désabonnement avec, par exemple une valeur de 1 qui équivaut à la valeur désabonnement ou « churn ».

Les données sont simulées

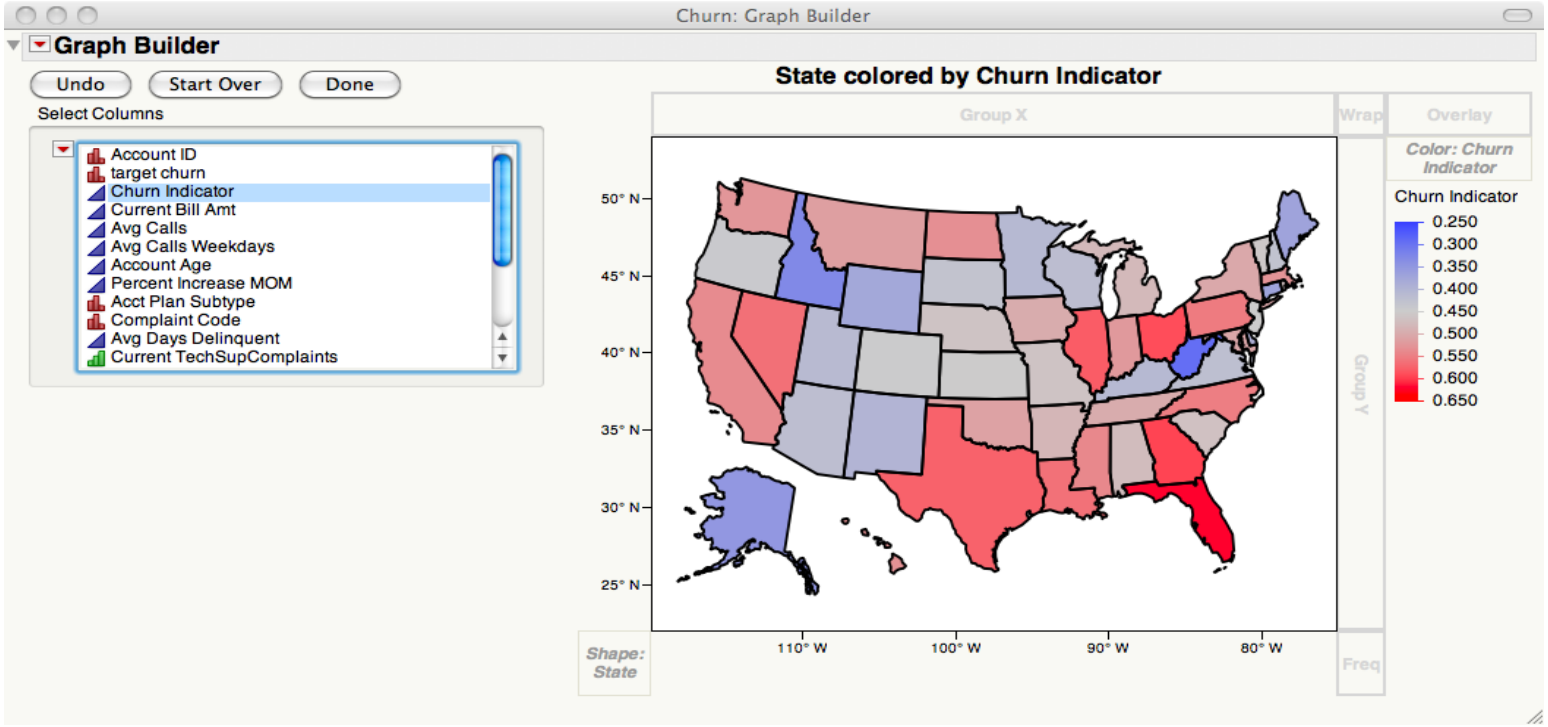


	Account ID	target churn	Churn Indicator	Current Bill	Avg Calls	Avg Calls Weekdays	Account Age	Percent Increase MOM	Acct Plan	Complaint Code
1	1100985	Churn	1	4776	3470	18725	39	-0.00	Silver	Billing Probl
2	1104315	No Churn	0	36978	16719	200375	39	-0.22	Silver	Billing Probl
3	1105524	No Churn	0	14519	657	9145	39	0.54	Gold	Billing Probl
4	1110127	No Churn	0	8611	983	2049	39	1.07	Gold	Call Quality
5	1113101	No Churn	0	18608	8165	44434	39	0.47	Silver	Billing Probl
6	1125027	Churn	1	86373	15922	38279	38	0.14	Silver	Billing Probl
7	1129471	No Churn	0	22784	4119	7738	18	0.43	Silver	Billing Probl
8	1130171	No Churn	0	10169	1612	30333	38	0.05	Silver	Moving
9	1131921	No Churn	0	25806	26933	180455	38	-0.04	Silver	Billing Probl
10	1133447	No Churn	0	12830	6998	36533	38	-0.38	Silver	Billing Probl
11	1135681	Churn	1	4526	7217	24628	38	0.61	Gold	Billing Probl
12	1138675	Churn	1	34867	0	0	25	-0.35	Silver	Billing Probl
13	1141591	No Churn	0	34540	17756	125136	38	-0.12	Silver	Billing Probl
14	1148793	Churn	1	4645	2016	10104	37	-1.00	Gold	Call Quality
15	1156881	No Churn	0	24041	18	6887	19	3.45	Gold	Moving
16	1157411	No Churn	0	27455	3990	41229	36	-0.41	Silver	Call Quality
17	1157539	Churn	1	13613	22260	42118	26	-0.89	Silver	Call Quality
18	1158501	No Churn	0	12714	1034	30131	36	0.18	Silver	Billing Probl
19	1168847	No Churn	0	6604	718	4445	35	0.00	Gold	Billing Probl
20	1172599	Churn	1	4543	380	4578	34	-1.00	Gold	Call Quality
21	1177002	Churn	1	10109	469	1473	34	23.12	Gold	Call Quality
22	1178400	No Churn	0	16410	18794	58324	33	-0.20	Silver	Billing Probl
23	1181247	No Churn	0	12151	602	7740	33	0.49	Gold	Check Acco
24	1181952	No Churn	0	12964	8891	44247	33	0.05	Silver	Billing Probl
25	1182768	Churn	1	8612	1812	3258	19	0.42	Gold	Billing Probl
26	1183576	No Churn	0	10169	3670	14783	33	0.21	Silver	Call Quality
27	1187286	Churn	1	14747	3985	67197	32	-0.75	Silver	Call Quality
28	1189648	No Churn	0	12712	5344	37651	24	-0.17	Silver	Check Acco
29	1192814	No Churn	0	5210	621	6862	31	-0.11	Gold	Call Quality
30	1193278	Churn	1	6172	0	0	31	0.00	Gold	Call Quality
31	1196653	No Churn	0	9054	2834	40373	31	2.00	Gold	Billing Probl
32	1199383	Churn	1	13262	1621	28479	30	-0.78	Silver	Billing Probl
33	1200471	Churn	1	14407	23999	66816	30	-0.62	Silver	Billing Probl
34	1202006	No Churn	0	12712	7244	14436	30	-0.39	Silver	Moving
35	1202872	No Churn	0	21651	9999	23394	30	0.32	Silver	Billing Probl
36	1203140	No Churn	0	26290	22619	47674	30	-1.00	Silver	Billing Probl

Exploration graphique des relations

- Les clients qui ont changés de fournisseur ont été sélectionnés et sont mis en surbrillance.
- Variables prédictives potentiellement bonnes sont celles où la surbrillance se regroupe en une ou plusieurs régions des graphiques de distribution. Par exemple:
  - Moyenne des jours impayés (Avg Days Delinquent) semble être un prédicteur utile ciblant potentiellement le taux de désabonnement puisque la majorité des désabonnés prend des valeurs plus élevées.
  - Code des plaintes (Complaint Code), en particulier une forte proportion de clients ayant une qualité pauvre des appels ont tendance à se désabonner.

Exploration graphique des relations



- Visuellement, nous pouvons constater une tendance de désabonnement pour chaque états des Etats-Unis où les Etats avec plus de taux de désabonnement sont de couleur rouge et le taux de désabonnement plus faible sont les états en couleur bleue.
- Bien que n'étant pas possible de montrer ici, en cliquant sur un état, une fenêtre apparaitra indiquant les valeurs « Churn » et « No Churn » pour cet état particulier, par exemple si nous cliquions sur la Floride et de garder la souris positionnée sur la Floride, nous verrions que notre table de données contient 53 clients en Floride dont 32 désabonnés.

Boosted tree avec segmentation (holdback) pour la sélection et test du modèle

Churn: Partition of target churn

Boosted Tree for target churn

Specifications

Target Column:	target churn	Number of training rows:	2455
Validation Column:	Subset	Number of validation rows:	824
Number of Layers:	50	Number of test rows:	839
Splits Per Tree:	3		
Learning Rate:	0.1		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4085	0.3340	0.3651	1-Loglike(model)/Loglike(0)
Generalized R-Square	0.5762	0.4942	0.5295	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4090	0.4615	0.4399	∑ -Log(ρ[j])/n
RMSE	0.3585	0.3879	0.3755	√∑ (y[j]-ρ[j])²/n
Mean Abs Dev	0.2919	0.3131	0.3064	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1800	0.2184	0.1931	∑ (ρ[j]≠ρMax)/n
N	2455	824	839	n

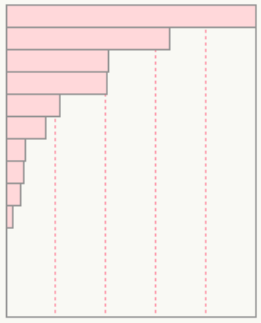
Confusion Matrix

Actual	Predicted		Actual	Predicted		Actual	Predicted	
Training	**No Churn**	**Churn**	**Validation**	**No Churn**	**Churn**	**Test**	**No Churn**	**Churn**
No Churn	1043	254	No Churn	331	89	No Churn	348	82
Churn	188	970	Churn	91	313	Churn	80	329

Cumulative Validation

Column Contributions

Term	Number of Splits	G^2
Complaint Code	30	89675.965
Avg Days Delinquent	26	58808.254
Percent Increase MOM	36	36872.065
Account Age	13	36167.861
Condition of Current Handset	5	19376.157
Equipment Age	6	14517.184
Acct Plan Subtype	20	7059.632
Avg Calls Weekdays	7	6608.625
Current Days OpenWorkOrders	1	5188.086
State	2	2448.805
Current Bill Amt	1	247.475
Current TechSupComplaints	3	141.269
Avg Calls	0	0.000
Avg Hours WorkOrderOpenned	0	0.000



- Le R Carre Généralisé de 0.5295 pour le sous-ensemble de test nous montre que le modèle prédit 52.9% de la variation explique par la variable “churn”.
- Taux de classification erronée est de 0,1931 pour le sous-ensemble de test, nous dit que le modèle devrait classer de façon erronée de 19% les intentions de client.
- Troisième table dans la matrice de confusion fournit la liste des erreurs de classification susceptibles de se produire si ce modèle est utilisé pour prédire le comportement des autres clients dans la base de données.
- Sur les 839 clients dans le sous-ensemble de test, des 409 désabonnés , le modèle prédit 329 désabonnés correctement soit 80% des clients qui ont désabonnés sont correctement prédit.
- La contribution de colonnes classe les prédicteurs dans l'ordre de leur importance dans le modèle.

Foret aléatoire avec segmentation (holdback) pour la sélection et test du modèle

Churn: Partition of target churn

Bootstrap Forest for target churn

Specifications

Target Column:	target churn	Training rows:	2455
Validation Column:	Subset	Validation rows:	824
		Test Rows	839
Number of trees in the forest::	100	Number of terms:	14
Number of terms sampled per split:	3	Bootstrap samples:	2455
		Minimum Splits Per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4771	0.3190	0.3423	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6448	0.4765	0.5037	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3616	0.4719	0.4556	∑ -Log(ρ[j])/n
RMSE	0.3229	0.3903	0.3811	√∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2819	0.3350	0.3286	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1063	0.2100	0.1967	∑ (ρ[j]≠ρMax)/n
N	2455	824	839	n

Confusion Matrix

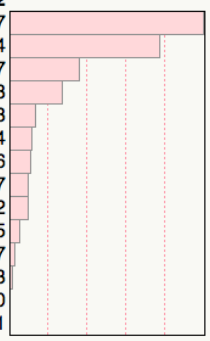
Actual	Predicted		Actual	Predicted		Actual	Predicted	
Training	**No Churn**	**Churn**	**Validation**	**No Churn**	**Churn**	**Test**	**No Churn**	**Churn**
No Churn	1158	139	No Churn	336	84	No Churn	346	84
Churn	122	1036	Churn	89	315	Churn	81	328

Cumulative Validation

Per-Tree Summaries

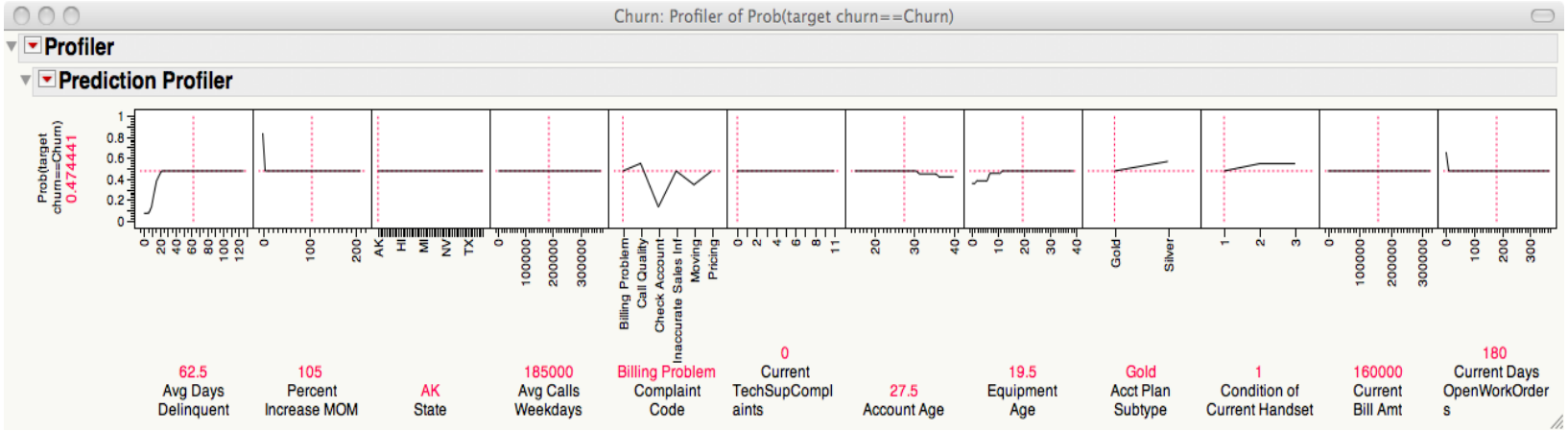
Column Contributions

Term	Number of Splits	G^2
Avg Days Delinquent	990	32953.517
Percent Increase MOM	962	25498.924
Complaint Code	638	11859.677
State	526	8974.888
Avg Calls Weekdays	686	4382.748
Account Age	774	3751.154
Avg Calls	653	3567.896
Current Bill Amt	661	3186.187
Equipment Age	728	3146.152
Current TechSupComplaints	457	1846.725
Acct Plan Subtype	346	973.997
Current Days OpenWorkOrders	148	462.613
Avg Hours WorkOrderOpenned	59	165.530
Condition of Current Handset	73	128.361



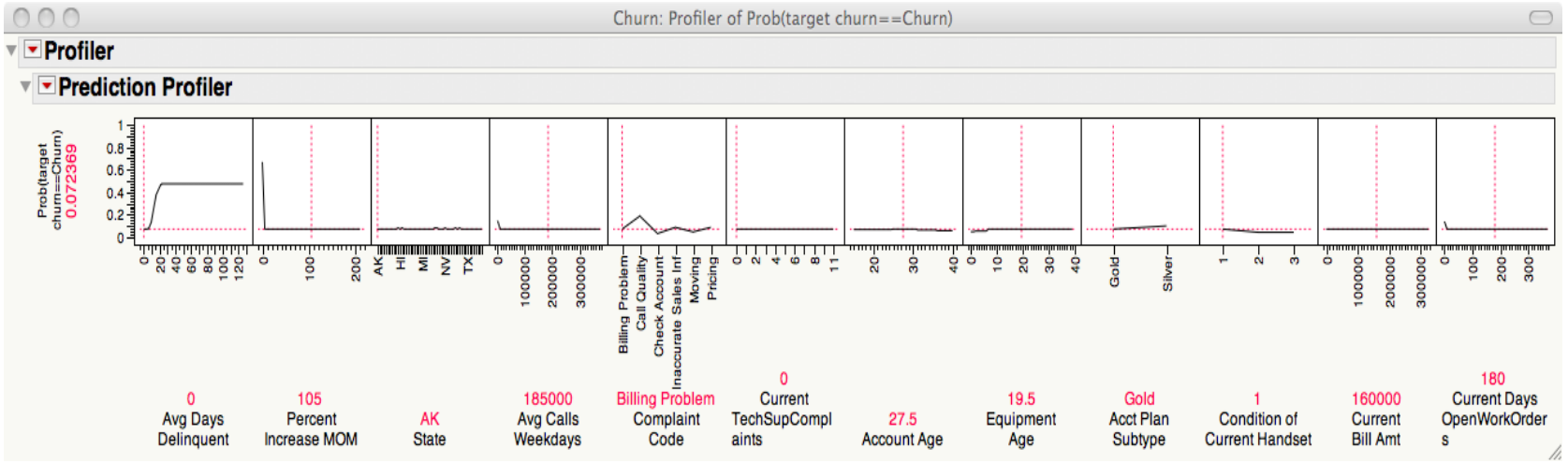
- Le R Carre Généralisé de 0.5037 pour le sous-ensemble de test nous montre que le modèle prédit 50.4% de la variation explique par la variable “churn”.
- Taux de classification erronée est de 0,1967 pour sous-ensemble de test, nous dit que le modèle devrait classer de façon erronée de 19% les intentions de client.
- Troisième table dans la matrice de confusion fournit la liste des erreurs de classification susceptibles de se produire si ce modèle est utilisé pour prédire le comportement des autres clients dans la base de données.
- Sur les 839 clients dans le sous-ensemble de test, des 409 désabonnés, le modèle prédit 328 désabonnés correctement soit 80% des clients qui ont désabonnés sont correctement prédit.
- La contribution de colonnes classe les prédicteurs dans l'ordre de leur importance dans le modèle.

Conclusions: Interpretation du (Boosted) modèle



- Le profileur permet de comprendre les relations et établir des critères de recherche pour identifier d'autres clients avec une plus forte intention de se désabonner. En particulier les clients cibles sont:
  - Avec des plaintes relatives à la qualité des appels, des problèmes de facturation, des informations inexactes de prix de ventes.
  - Qui payent leurs factures avec plus de 20 jours de retard.
  - Avec des augmentations d'utilisation mensuellement plus faible.

Conclusions: Interpretation du (Boosted) modèle



- Interroger le modèle en faisant glisser les lignes verticales pointillées rouges pour voir l'impact et l'évolution de la valeur d'intérêt d'un ou plusieurs prédicteurs sur la probabilité de désabonnement.
- Par exemple, le graphique ci-dessus diffère du graphe précédent par le fait que les jours impayés ont été changé à 0 au lieu de la valeur précédente de 62,5.
- La probabilité de taux de désabonnement est tombé à 0,07 de 0,47.
- Cette fonctionnalité nous permet de déterminer les moyens de réduire le taux de désabonnement.
- Le modèle nous propose quelques stratégies pour augmenter la fidélisation des clients, par exemple:
  - encourager les clients à payer leurs factures rapidement,
  - encourager une utilisation croissante du téléphone mobile et
  - augmenter la qualité des appels

Questions?

Cas d'étude 4: Réseaux de neurones

Exemple marketing: La rétention de clients

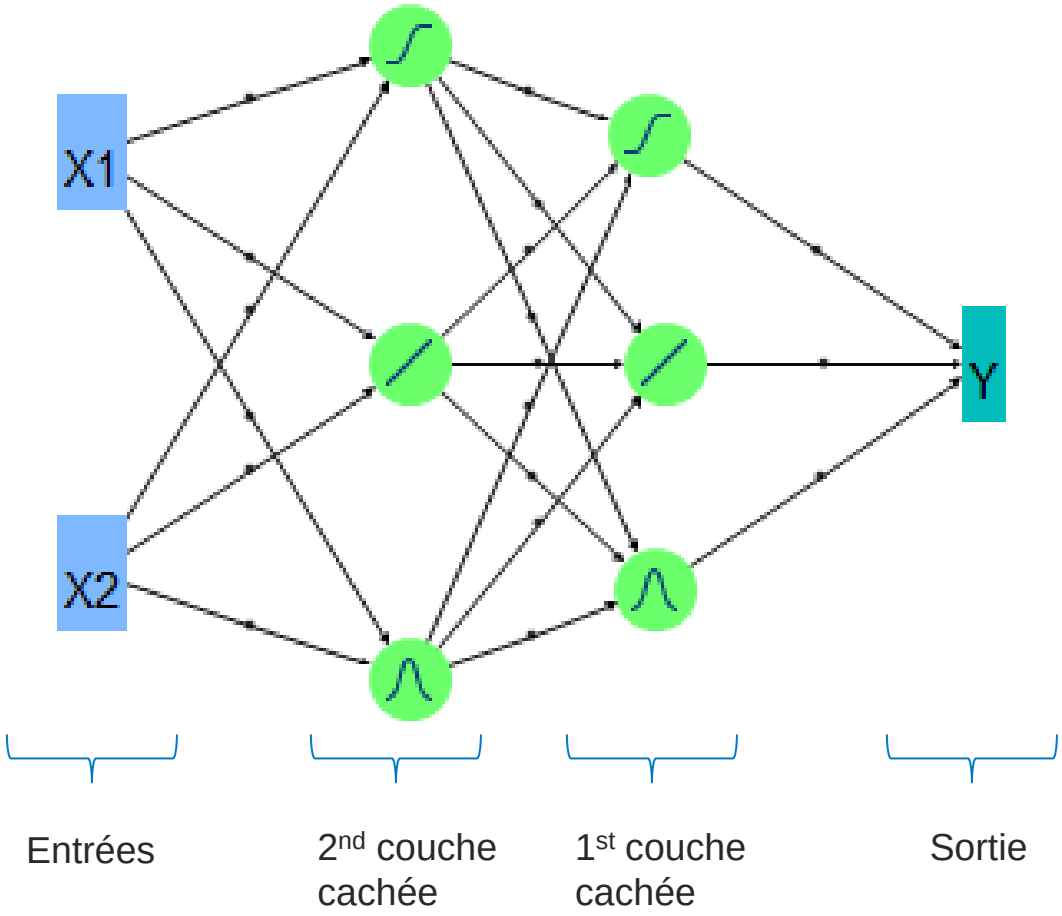
jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW.

Réseaux de neurones

- Les réseaux de neurones sont des modèles très flexibles et potentiellement non-linéaires.
- Un réseau de neurones peut être vu comme un ensemble de blocs et nœuds connectés.
- Soit un exemple général avec une réponse Y et deux prédicteurs X1 et X2. Un exemple classique de réseau de neurone ajustant ces données est donné dans le diagramme suivant

Réseau de neurones: exemple de diagramme



Equation d'un modèle de réseau de neurone

- 2nd couche cachée:
  - Soit une combinaison linéaire des entrées

$$w_0^{2j} + w_1^{2j} X_1 + w_2^{2j} X_2$$

  - Transformée telle que:

$$h_{2j} = f_{2j}(w_0^{2j} + w_1^{2j} X_1 + w_2^{2j} X_2)$$

- 1st couche cachée
  - Soit une combinaison linéaire de la seconde couche

$$w_0^{1j} + \sum_k w_k^{1j} h_{2k}$$

  - Transformée telle que:

$$h_{1j} = f_{1j}\left(w_0^{1j} + \sum_k w_k^{1j} h_{2k} \right)$$

Equation d'un modèle de réseau de neurone

- Modèle final
  - » Soit une combinaison linéaire de la première couche

$$y = w_0 + \sum_j w_j h_{1,j}$$

  - » Note: si la réponse est catégorielle, on applique une tranformation de la sortie avec la fonction logit.

$$y = \text{logit}\left(w_0 + \sum_j w_j h_{1,j} \right)$$

$$\text{logit}(u) = \frac{1}{1+e^{-u}}$$

Réseau de neurones

- Points forts
  - Ce sont des modèles mathématiquement complexes mais facilement représentés comme un diagramme de réseau.
  - Peut être hautement calculatoire et consommateur de temps à construire
  - Peut modéliser une variété impressionnantes de relations
    - » La flexibilité est la principale force des réseaux de neurones.
  - Les modèles NN sont enclins au sur-ajustement
    - » JMP a plusieurs techniques pour éviter le sur-ajustement
      - » Utilisation de technique de validation
      - » Techniques d'arrêt d'ajustement, résultant en moins de sur-ajustement

- Voir Gotwalt, C., "JMP® 9 Neural Platform Numerics", Feb 2011, http://www.jmp.com/blind/whitepapers/wp_jmp9_neural_104886.pdf

Fonctionnalités de la plateforme JMP pour les réseaux de neurones

- *Plateforme Réseaux de neurones de JMP*
  - **Validation croisée**
  - Gestion des valeurs manquantes
  - Sélection automatique de la taille du réseau avec le **boosting**
  - **Transformation automatique des variables d'entrées**, sauvegarde possible de ces données transformées
  - **Fonction loss résistant aux valeurs extrêmes**
  - Capacité à ajuster **des réseaux à une ou deux couches**
  - Capacité à choisir parmi **trois fonctions d'activation** (tanh, linear, gaussian)

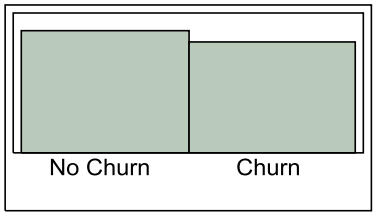
Exemple – Désabonnement téléphonique

- Echantillon de données de la base d'une compagnie téléphonique concernant 4118 données clients.
- La variable à étudier est la colonne nommée "target churn", qui prennent deux valeurs:
  - ***Churn*** pour indiquer qu'un client à changer d'opérateur
  - ***No churn*** pour indiquer que le client est toujours consommateur.
- **But:** Construire un modèle pour prédire les prédispositions ses clients à se désabonner ou à changer d'opérateur et utiliser ce modèle pour conserver davantage de clients dans le futur.
- Les prédicteurs potentiels sont les montants de factures au travers des états. La variable « Churn indicator » est une alternative numérique à « target churn » avec une valeur de 1 pour le cas de désabonnement.

Réponse et jeux de validation croisés

- La réponse est répartie
  - Usuellement, un taux de désabonnement de 47% est TRES mauvais
  - Ce jeu de données a équilibré le sous groupe de désabonné afin de créer un modèle adapté pour la prévision du groupe churn.

target churn



Frequencies

Level	Count	Prob
No Chur	2147	0.52137
Churn	1971	0.47863
Total	4118	1.00000
N Missin	0	

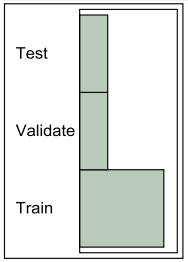
2 Level

- Validation Role
  - Assignation aléatoire
    - » Apprentissage, Validation Test
  - Répartition équilibrée sur les différents niveaux de réponse.

Report

Distributions

Subset



Frequencies

Level	Count	Prob
Train	2455	0.59616
Validat	824	0.20010
Test	839	0.20374
Total	4118	1.00000
N Missin	0	

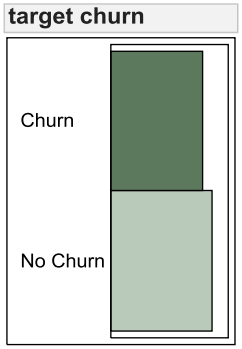
3 Level

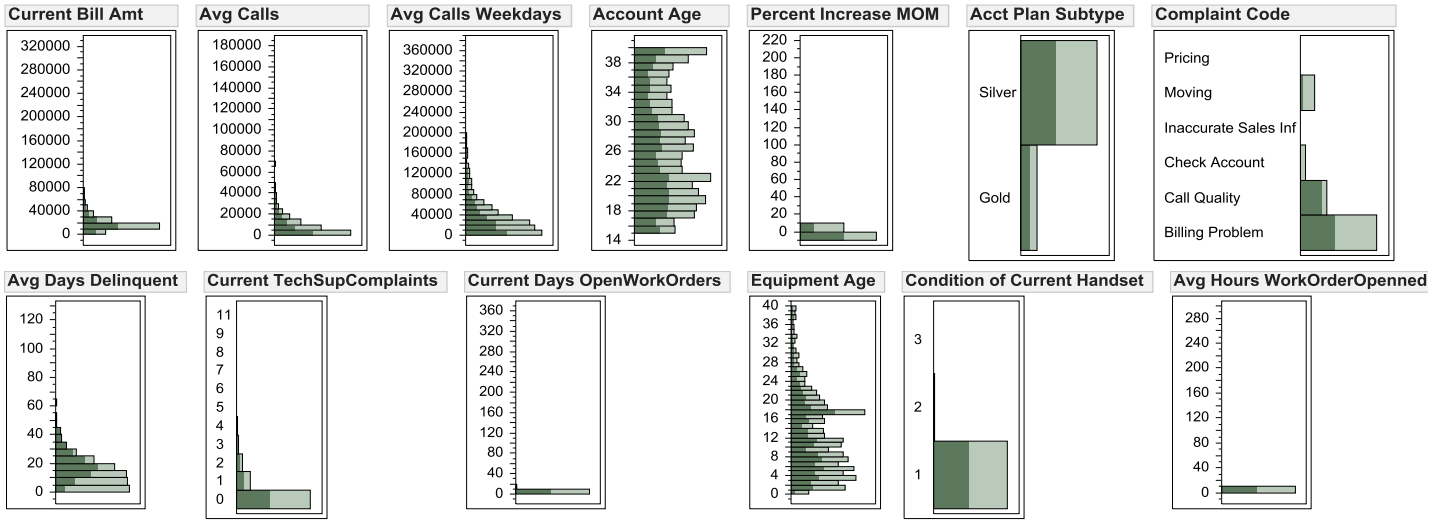
Tabulate

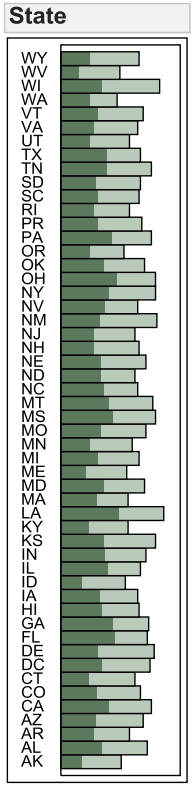
	Subset		
	Train	Validate	Test
target churn	Column %	Column %	Column %
No Churn	52.8%	51.0%	51.3%
Churn	47.2%	49.0%	48.7%

Predicteurs

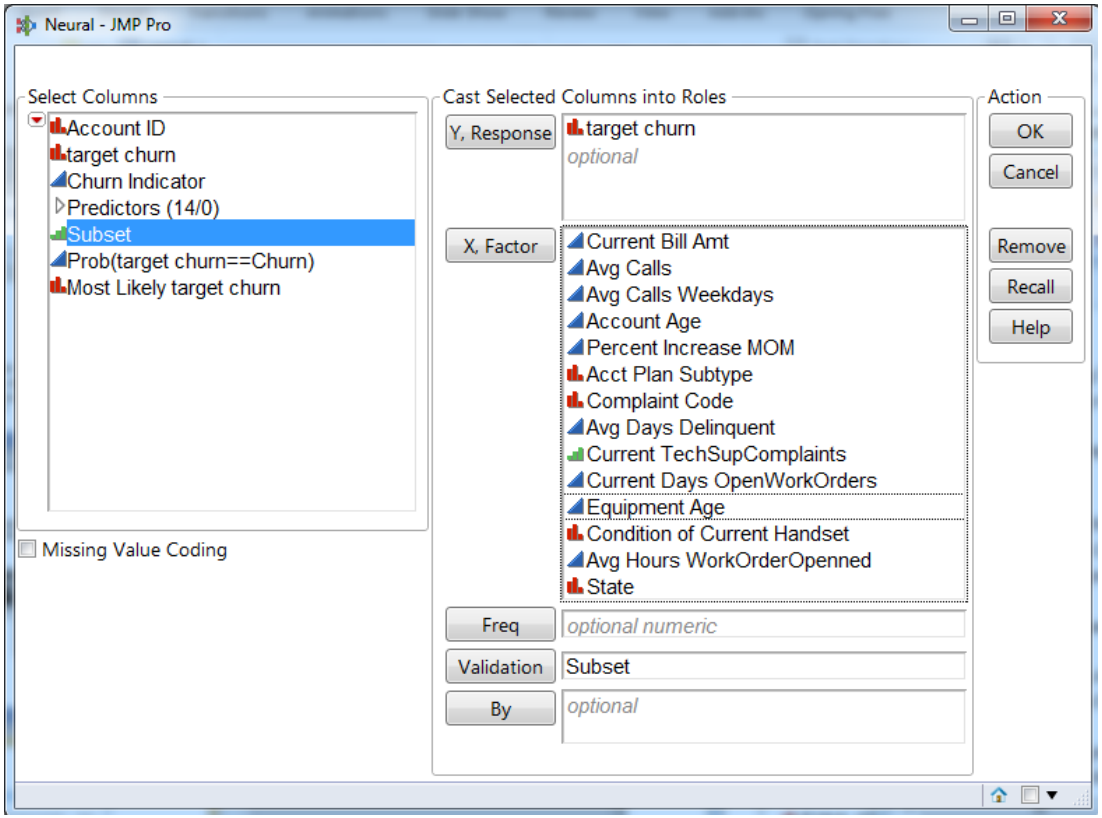
- Il n'est pas aisé de trouver des prédicteurs importants avec les outils d'exploration de données



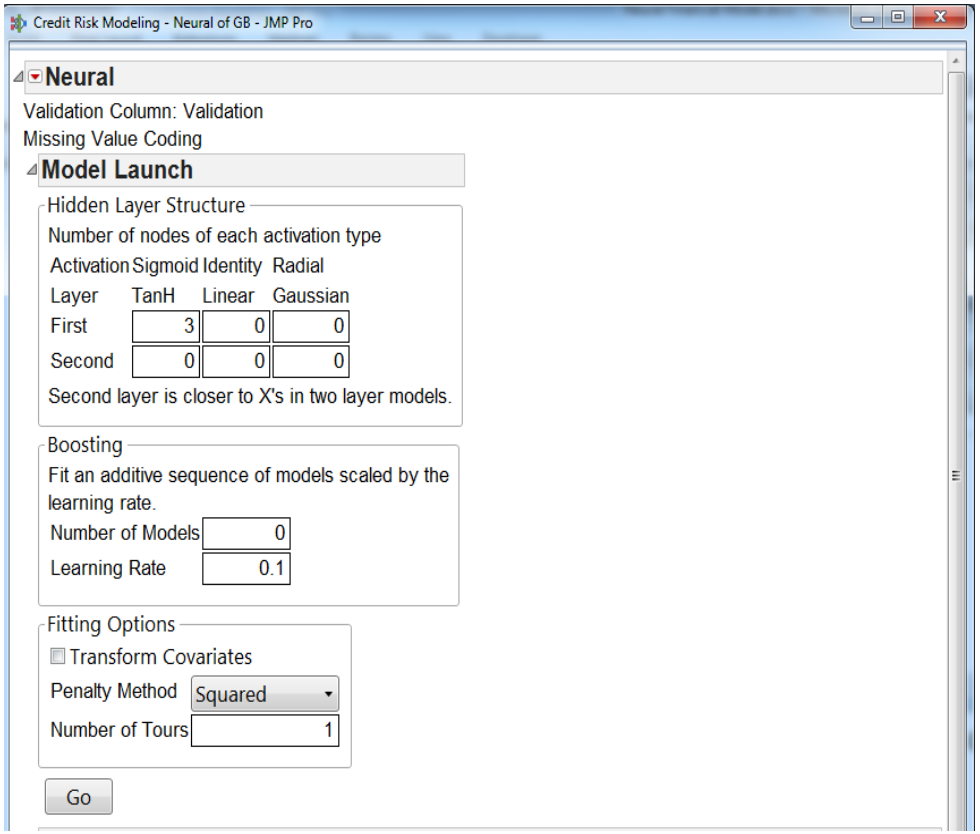




Lancement de la plateforme

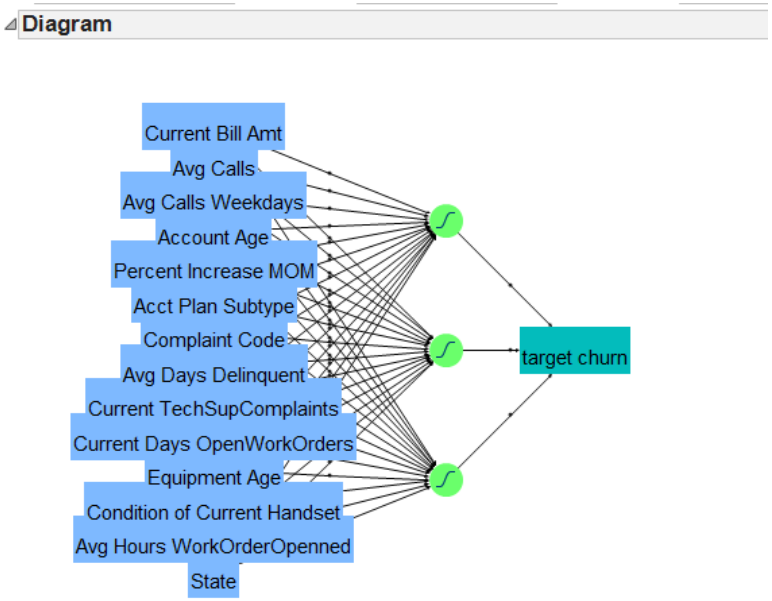


Fenêtre initiale



- Vous pouvez ajuster différents modèles
- Il faut spécifier la structure du réseau de neurones
- Possibilité d'utiliser les techniques de boosting.
- Techniques avancées d'ajustement
  - Transformation des covariables (X's)
  - Nombre de tours
  - Si Y est continu, il est possible d'utiliser la méthode résistant aux valeurs extrêmes

NN – Une couche cachée avec 3 noeuds Tanh



Training

target churn	Measures
Generalized RSquar	0.3995386
Entropy RSquare	0.2571947
RMSE	0.4095411
Mean Abs Dev	0.342912
Misclassification Rat	0.2338086
-LogLikelihood	1261.0896
Sum Freq	2455

Confusion Matrix

Actu / Predicte

target churn	No Churn	Churn
No Chur	1022	275
Churn	299	859

Confusion Rates

Actu / Predicte

target churn	No Churn	Churn
No Chur	0.78797	0.21203
Churn	0.25820	0.74180

Validation

target churn	Measures
Generalized RSquar	0.3387773
Entropy RSquare	0.2114831
RMSE	0.4256592
Mean Abs Dev	0.3566694
Misclassification Rat	0.2512136
-LogLikelihood	450.24149
Sum Freq	824

Confusion Matrix

Actu / Predicte

target churn	No Churn	Churn
No Chur	339	81
Churn	126	278

Confusion Rates

Actu / Predicte

target churn	No Churn	Churn
No Chur	0.80714	0.19286
Churn	0.31188	0.68812

Test

target churn	Measures
Generalized RSquar	0.3585006
Entropy RSquare	0.2259521
RMSE	0.4209513
Mean Abs Dev	0.3542463
Misclassification Rat	0.2455304
-LogLikelihood	449.94447
Sum Freq	839

Confusion Matrix

Actu / Predicte

target churn	No Churn	Churn
No Chur	345	85
Churn	121	288

Confusion Rates

Actu / Predicte

target churn	No Churn	Churn
No Chur	0.80233	0.19767
Churn	0.29584	0.70416

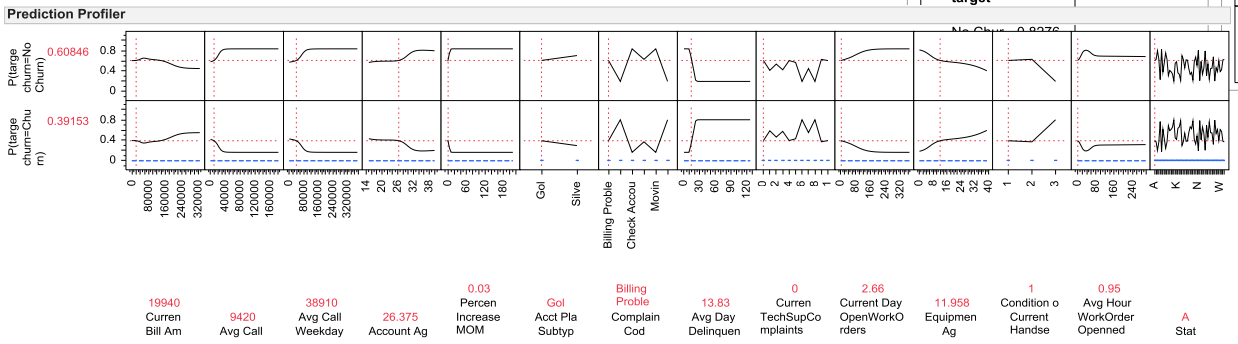
Training

Validation

target churn	Area
No Chur	0.8040
Churn	0.8042

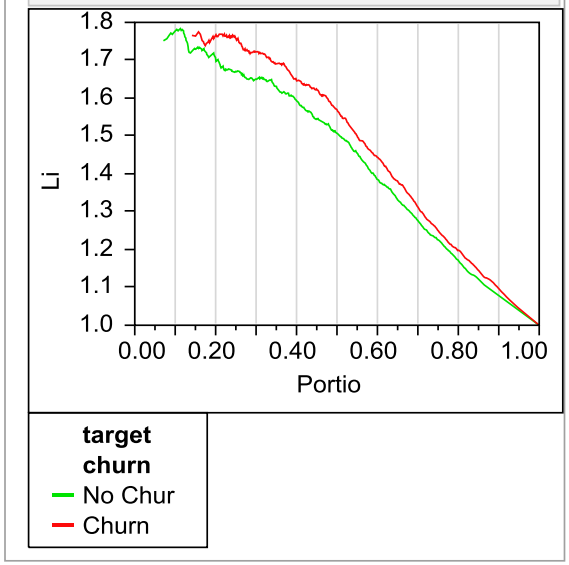
Test

target churn	Area
No Chur	0.8146
Churn	0.8144

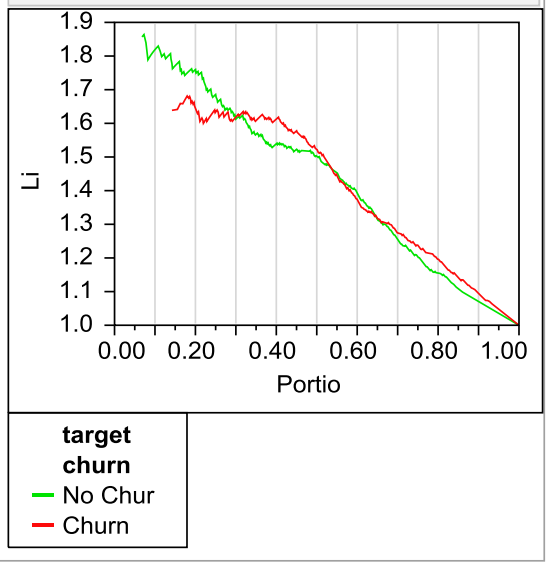


NN – Une couche cachée avec 3 noeuds Tanh

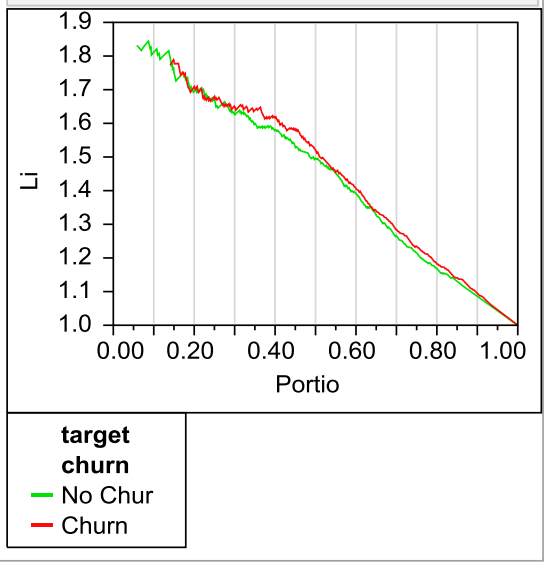
Training



Validation



Test



NN – 2 couches cachées

Model Launch

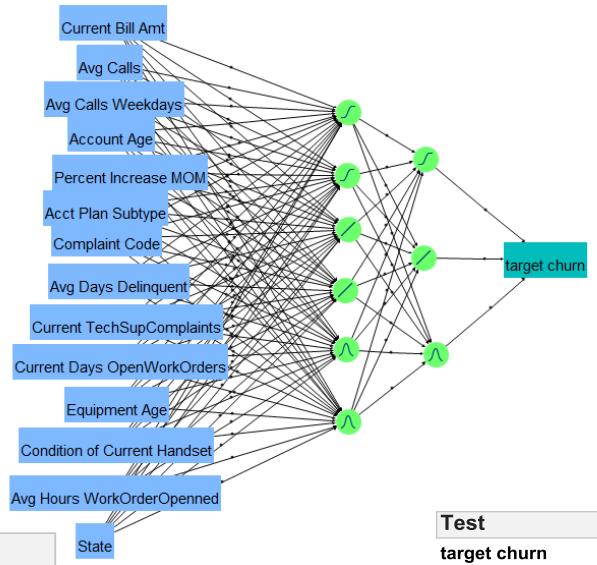
Hidden Layer Structure

Number of nodes of each activation type

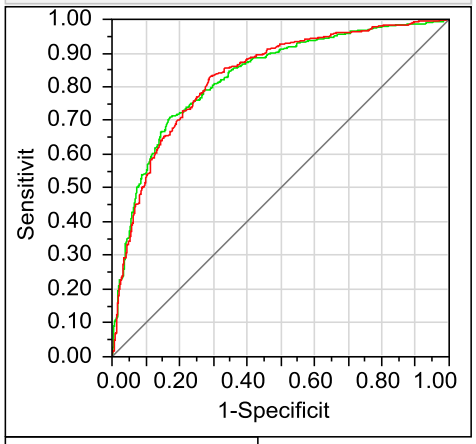
Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	1	1	1
Second	2	2	2

Second layer is closer to X's in two layer models.

Diagram

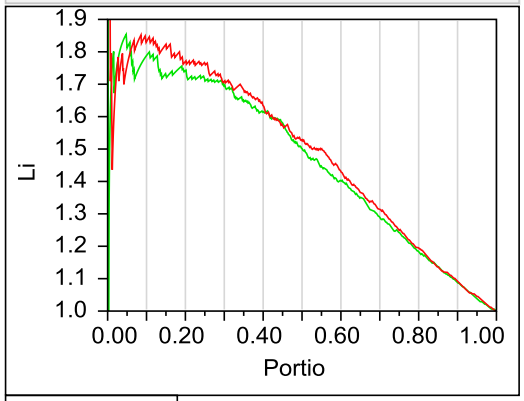


Receiver Operating Characteristic on Test Data



target churn	Area
No Chur	0.8307
Churn	0.8307

Lift Curve on Test Data



target churn
No Chur
Churn

Test

target churn	Measures
Generalized RSquar	0.4009732
Entropy RSquare	0.258091
RMSE	0.4107082
Mean Abs Dev	0.3238983
Misclassification Rat	0.2467223
-LogLikelihood	431.26254
Sum Freq	839

Confusion Matrix

Actu	Predicte	
target churn	No Churn	Churn
No Chur	342	88
Churn	119	290

Confusion Rates

Actu	Predicte	
target churn	No Churn	Churn
No Chur	0.79535	0.20465
Churn	0.29095	0.70905

Courbes de ROC

- Le plus la courbe de ROC est au dessus de la ligne à 45 degrés, meilleur est le modèle comparé à un modèle aléatoire.
- La courbe de ROC est construite sur la table ordonnées (e.g. ordonnancement des données de la plus haute Prob[GB==1] à la plus faible).
  - Pour chaque ligne, si la valeur réelle pour GB==1, alors la courbe augmente (verticale), sinon, elle reste stable (horizontale).
- Une bonne mesure de prédiction est le calcul de l'aire sous la courbe (AUC) qui est une valeur calculée à partir de l'aire sous la courbe de ROC et qui est dans l'intervalle [0,1].
  - Une valeur supérieur à 0.5 indique que le modèle est meilleur qu'un tirage aléatoire.

Courbes de Lift

- Les courbes de Lift utilisent les données ordonnées, comme les courbes de ROC
- L'axe horizontal est la proportion rangée des données
  - 0.10 == les 10% premières données ordonnées
- L'axe vertical est la qualité du sous jeu
- Le ratio du nombre de la cible désirée sur le nombre total de point est calculé
- La courbe de Lift montre combien un modèle amène la cible souhaitée dans chaque portion, comparé à un modèle aléatoire

Réseaux de neurones - Boosting

- Pour les réseaux de neurones, le boosting, par un arrangement mathématique va grossir le réseau de neurone.
  - En fait, un réseau de neurone boosté (un modèle fin, avec pleins de couches boostées) est mathématiquement équivalent à un large réseau de neurones (c'est-à-dire, avec le même nombre de couches, avec plus de nœuds à chaque couche)
- Pour cet exemple, nous allons booster 30 fois le réseau de neurones (1,1,1)(1,1,1) NN, ce qui résultera à un réseau de neurones de deux couches, et 90 nœuds chaque couche.
  - Note: Ce modèle final prend quelques minutes à se construire. Les réseaux de neurones restent les modèles les plus gourmand en calcul comparés aux autres considérés.

Réseaux de neurones - Boosting

Model Launch

Hidden Layer Structure

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	1	1	1
Second	1	1	1

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models	50
Learning Rate	0.1

Model NTanH(1)NLinear(1)NGaussian(1)NTanH2(1)NLinear2(1)NGaussian2(1)NBoost(14)

Training

target churn	Measures
Generalized RSquar	0.5010825
Entropy RSquare	0.340296
RMSE	0.381727
Mean Abs Dev	0.3155884
Misclassification Rat	0.2032587
-LogLikelihood	1120.0053
Sum Freq	2455

Confusion Matrix

Actu / Predicte

target churn	No Churn	Churn
No Chur	1050	247
Churn	252	906

Confusion Rates

Actu / Predicte

target churn	No Churn	Churn
No Chur	0.80956	0.19044
Churn	0.21762	0.78238

Validation

target churn	Measures
Generalized RSquar	0.4100705
Entropy RSquare	0.2651439
RMSE	0.4090436
Mean Abs Dev	0.3369502
Misclassification Rat	0.243932
-LogLikelihood	419.6013
Sum Freq	824

Confusion Matrix

Actu / Predicte

target churn	No Churn	Churn
No Chur	334	86
Churn	115	289

Confusion Rates

Actu / Predicte

target churn	No Churn	Churn
No Chur	0.79524	0.20476
Churn	0.28465	0.71535

Test

target churn	Measures
Generalized RSquar	0.4313059
Entropy RSquare	0.2819526
RMSE	0.4030799
Mean Abs Dev	0.3338644
Misclassification Rat	0.2348033
-LogLikelihood	417.39211
Sum Freq	839

Confusion Matrix

Actu / Predicte

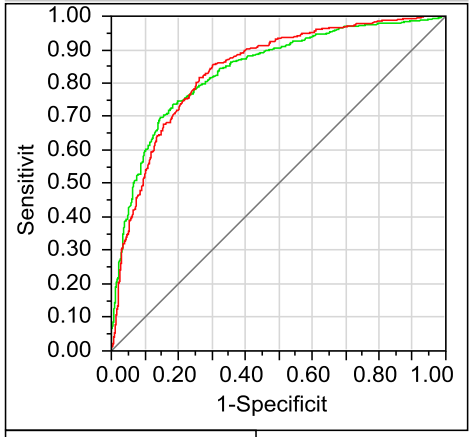
target churn	No Churn	Churn
No Chur	341	89
Churn	108	301

Confusion Rates

Actu / Predicte

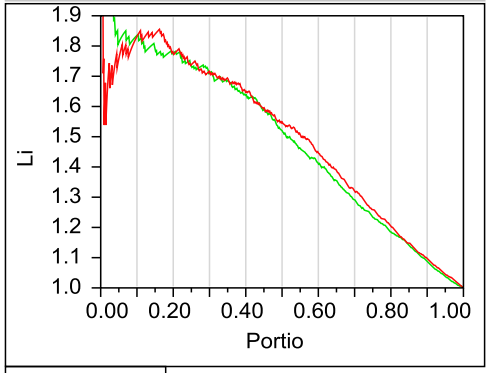
target churn	No Churn	Churn
No Chur	0.79302	0.20698
Churn	0.26406	0.73594

Receiver Operating Characteristic on Test Data



target churn	Area
No Chur	0.8404
Churn	0.8404

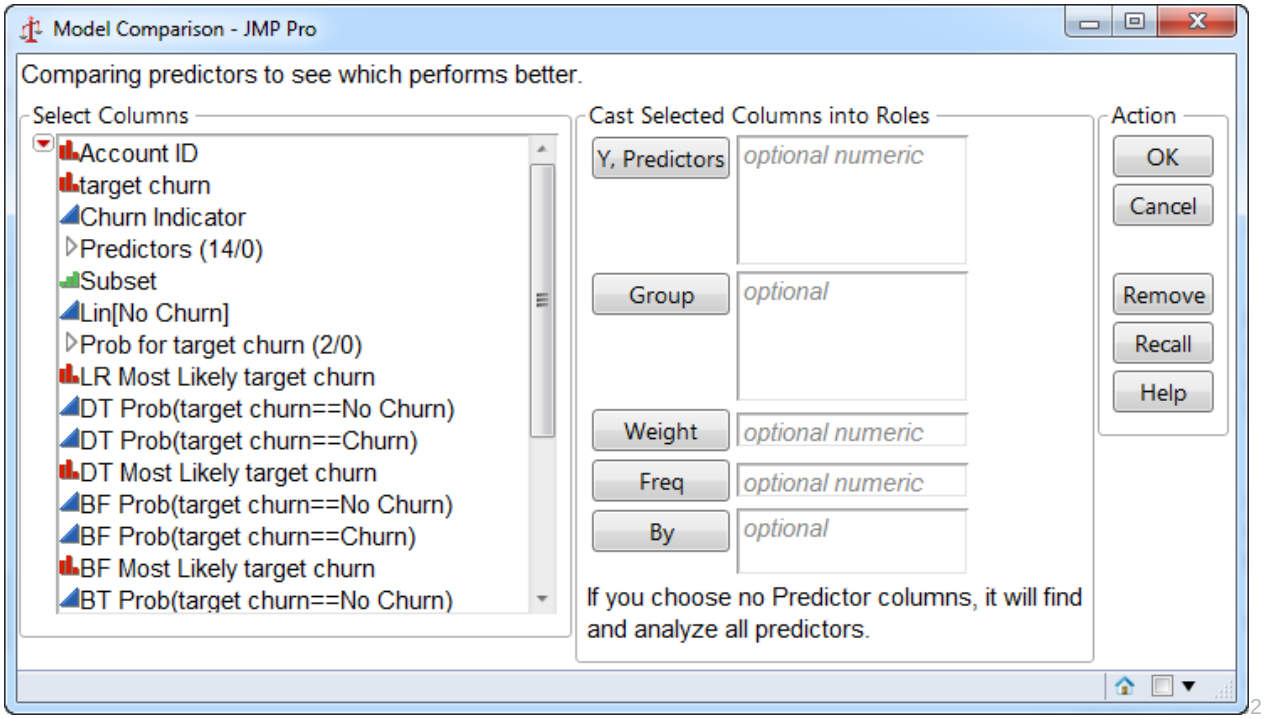
Lift Curve on Test Data



target churn
- No Chur
- Churn

Comparaison de modèles

- Sauver les formules de prédiction des différents modèles dans la table de données
- Analyse > Modelisation > Comparaison de modèles
- Cliquer OK



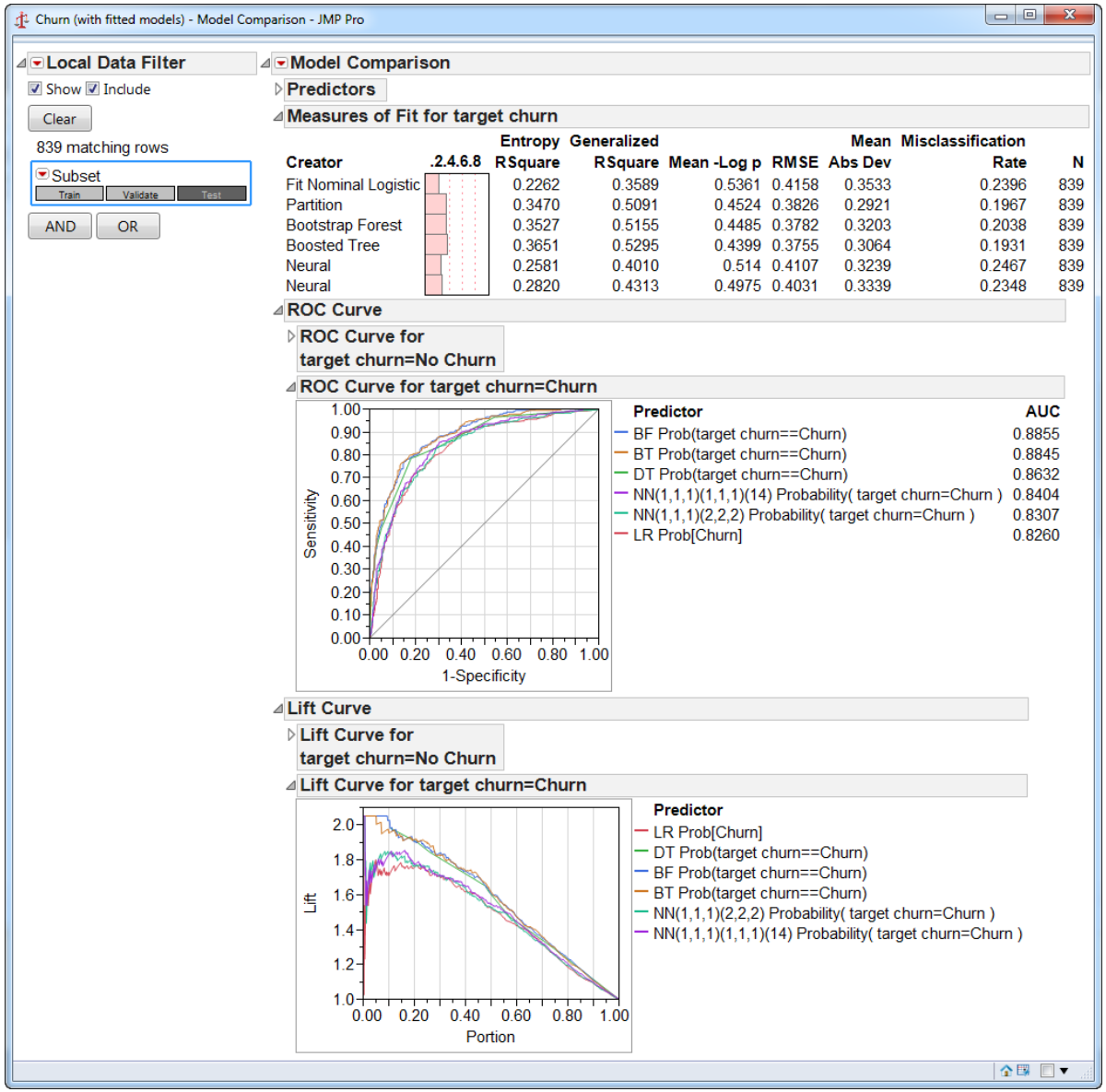
Comparaison de modèles

Predictors

Target Colum	Predictors		
target churn	**Categor**	**Probability Colum**	
	No Chur	LR Prob[No Churn]	Fit Nominal Logisti
	Churn	LR Prob[Churn]	
	Categor	**Probability Column**	
	No Chur	DT Prob(target churn==No Churn	Partition
	Churn	DT Prob(target churn==Churn)	
	Categor	**Probability Column**	
	No Chur	BF Prob(target churn==No Churn	Bootstrap Forest
	Churn	BF Prob(target churn==Churn)	
	Categor	**Probability Column**	
	No Chur	BT Prob(target churn==No Churn	Boosted Tree
	Churn	BT Prob(target churn==Churn)	
	Categor	**Probability Column**	
	No Chur	NN(1,1,1)(2,2,2) Probability(target churn=No Churn	Neural
	Churn	NN(1,1,1)(2,2,2) Probability(target churn=Churn)	
	Categor	**Probability Column**	
	No Chur	NN(1,1,1)(1,1,1)(14) Probability(target churn=No Churn)	Neural
	Churn	NN(1,1,1)(1,1,1)(14) Probability(target churn=Churn	

Comparaison de modèles

- Filtre de données locales pour se focaluser sur le jeu de test.
- Les deux premiers modèles sont les forêts aléatoires et les Boosted trees



Questions?

Résumé

- Les modèles statistiques peuvent séparer la variation dans des réponses continues ou catégorielles en composants prédictifs ou non prédictifs
- Les stratégies de Holdback ou de K-Fold aide à la généralisation du modèle.
- Introduction des modèles classiques de modélisation statistiques
  - Arbres de décision, régression, et réseaux de neurones
  - variation
- Quelques approches utiles de modélisation statistiques
  - Pas à pas, Boosting, modèle moyen, e.g. forêts aléatoires

Résumé

- JMP apporte des stratégie de validation ainsi que des modèles statistiques et approches
- Permet aux utilisateurs avec des niveaux en statistiques différents d'utiliser des modèles plus facilement et rapidement
- Optimise les délais de décision
- Dépendant de la situation
  - Résolution d'avantage de problèmes dans un délais fixé.
  - Meilleure décision
  - Moins de risques de mauvaise décision
  - Evite les conséquences de mauvaise décision
  - Gain de temps
  - Moins de cycles d'apprentissage.