Institut Régional de Formation des Allocations Familiales
67 avenue Jean Jaurès - 75019 PARIS CEDEX 19 - Tél. : 01 71 13 36 18
Siret : 381 050 996 00127 – APE 8559 A – N° déclaration d'activité : 11 75 48596 75

MODLOG : Modélisation LOGIT et typologies avancées avec Python dans Databricks

Description :

Le métier de chargés d'études consiste à manipuler des données brutes dans le but de réaliser des analyses sociologiques, statistiques et démographiques afin notamment d'apporter les éléments pour aider à la décision. Ces analyses requièrent des compétences à la fois théoriques, méthodologiques et pratiques en programmation.
En outre, les fonctionnalités de Databricks offrent de nouvelles perspectives d'analyse de données plus avancées (régression, modèles de prévision, machine Learning, analyse géospatiale, ...). Les chargés d'études doivent donc être formés sur les méthodes statistiques théoriques et pratiques afin de maitriser la construction de modèles LOGIT mais aussi d'approfondir l'analyse des données à travers des techniques de réduction de dimensionnalité et de segmentation.
Ils devront être capable de préparer et encoder des variables explicatives, construire et ajuster un modèle LOGIT et, enfin, segmenter une population via ACP et CAH.

Objectifs opérationnels

1.Préparer et encoder des variables explicatives

- Sélectionner les variables pertinentes et traiter les données (valeurs manquantes, doublons, normalisation),
- Encoder les variables qualitatives et créer des variables dérivées si nécessaire,
- Détecter et corriger la colinéarité entre variables explicatives.

2. Construire et ajuster un modèle LOGIT

- Estimer un modèle LOGIT et interpréter ses coefficients,
- Évaluer la performance du modèle (AUC, courbe ROC, tests de significativité),
- Comparer et ajuster les modèles à l'aide de critères tels que AIC et BIC.

3. Segmenter une population via ACP et CAH,

- Réaliser une ACP pour réduire la dimensionnalité et interpréter les axes principaux,
- Appliquer une CAH sur les composantes principales pour créer des groupes homogènes,
- Analyser et interpréter les segments obtenus à l'aide de visualisations adaptée.

Objectifs :

- Interpréter un odds ratio et analyser la courbe ROC pour évaluer la performance d'un modèle LOGIT,
- Mesurer la qualité d'un modèle à l'aide de critères tels que AIC, BIC et AUC,
- Réaliser une Analyse en Composantes Principales (ACP) pour réduire la dimensionnalité et faciliter l'interprétation des donnée,

- Construire des typologies en combinant ACP et Classification Ascendante Hiérarchique (CAH) pour segmenter des populations.

Programme :

1. Préparation des données et construction du LOGIT

- Rappel des concepts / Modèle de régression linéaire - théorie + LOGIT
- Sélection des variables explicatives, encodage des qualitatives (dummies) et standardisation des quantitatives
- Formulation du modèle log□odds et estimation via statsmodels et scikit learn
- Isolation de l'effet "toutes choses égales" dans le Logit

2. Évaluation et interprétation du modèle LOGIT

- Mesures de qualité : pseudo□R^2, AIC, BIC
- Courbe ROC, calcul de l'AUC, matrice de confusion et détermination du seuil optimal
- Hétéroscédasticité : détection (test de Breusch□Pagan) et correction (erreurs□types robustes)
- Tests de colinéarité (matrices de corrélation / VIF) pas de lien avec le modèle ridge alors que présent dans le support

3. Typologies avancées et diversité

- Rappel théorique des concepts
- Analyse en composantes principales (ACP) pour synthèse des données quantitatives
- Classification ascendante hiérarchique (CAH) pour segmentation (distances euclidienne, Manhattan (à valider) ; liaisons Ward, complete) ou autre méthode de segmentation (K-means) : plus rapide, plus imagée, peu coûteux en calculs.
- Mesures de diversité : variance, entropie de Shannon

Méthode pédagogique :

- Apports théoriques
- Exemples concrets en Caf
- Exercices d'application

Modalités d'évaluation et de validation :

- **Evaluation de positionnement** : sous forme d'un questionnaire ou d'un tour de table avec le formateur pour valider les prérequis, pour évaluer les besoins individuels et pour déterminer le niveau de connaissances
- **Evaluation des acquis** : validation de la compréhension et de l'acquisition des connaissances sous forme de mises en situations, de réflexions collectives et d'utilisation d'outils de diagnostic
- **Evaluation à chaud :** à la fin de la formation, un bilan oral est effectué par le formateur et une évaluation écrite adressée aux stagiaires permettent d'apprécier la qualité de la prestation et de mesurer l'efficacité de l'action au regard des objectifs globaux
- **Evaluation à froid** : réalisée avec un outil interne Caf
- **Attestation de suivi** : Feuille de présence
- **Certificat de réalisation** mentionnant la nature, la durée de l'action est remis aux stagiaires à l'issue de la formation

Accessibilité :

Nous mettons tout en œuvre afin d'offrir aux personnes en situation de handicap des conditions optimales d'accès et d'apprentissage. N'hésitez pas à contacter Sandrine Decottignies référent handicap sandrine.decottignies@caf92.caf.fr - 01 87 02 85 05 directement pour lui signaler vos besoins spécifiques.

Public cible :
Chargés d'études, Data Analystes, Data Scientiste, Contrôleurs de gestion ayant des connaissances initiales en statistiques

Pré-requis :

- Maitriser le langage python dans Databricks
- Maitriser les concepts de statistique descriptive (variable quali/quanti, moyenne, quantiles, variance, écart-type...),
- Avoir une approche théorique de la régression linéaire simple (Moindre carrés ordinaires MCO)
- Utiliser les données du Sid de la branche Famille

Programmation :
Nous contacter Planification sur le site irfaf.fr

Catégorie :
Formation PERSPICAF - Métiers de l'observation socio-économique et de l'appui au pilotage

Lieu :
IRFAF - 67 avenue Jean Jaurès 75019 Paris

Coût forfait / stagiaire : 1459.00 €
Coût par forfait / groupe : 7295.00 €
Modalités animation :

- Présentiel

Formateur :

Effectif :
5 à 8 participants

Durée en jours :
2,5

Durée en heures :
17h30