Projet INGESUP M1 et M2

Probabilité et statistique

Disposer régulièrement d'une vision économique et sociale de leur entreprise est un enjeu majeur pour de nombreux chefs d'entreprise. Cette photographie sociale et économique de l'entreprise permet d'analyser un certain nombre d'indicateurs comme la répartition des effectifs par tranche d'âge, l'encienneté, le sexe, … Dans le domaine du suivi des salariés et des éléments de rémunération, connaître l'historique administratif (emplois occupés, contrats, …) ainsi que l'évolution du salaire depuis l'embauche ou encore la comparaison par rapport à la moyenne de sa catégorie sont des questions de grande importance.

La base de données proposée présente les données observées sur un échantillon de 76 salariés d'une grande entreprise. Pour chaque salarié, on dispose des informations suivantes :

- **id** : identificateur

- **sexe** : variable qualitative à 2 modalités (0 : homme, 1 : femme)

- **nais** : date de naissance. Attention certaines données manquantes sont représentées par 9999

- **educ** : variable quantitative discrète représentant le nombre d'années d'études

- **csp** : variable qualitative à 3 modalités (1 : employé de bureau, 2 : agent de sécurité, 3 : cadre)

- **salemb** : variable quantitative représentant le salaire annuelle d'embauche

- **salac** : variable quantitative représentant la salaire annuelle actuel

- **anc** : variable quantitative représentant l'ancienneté des salariés en nombre de mois dans l'entreprise

- **exp** : variable quantitative représentant l'expérience professionnelle antérieure des salariés, en nombre de mois de travail avant l'entrée dans l'entreprise

- **mino** : variable qualitative codée 1 si l'individu appartient à une minorité ethnique et 0 sinon.

L'objectif dans cette étude est de voir s'il y a des corrélations et des différences significatives en terme de salaire entre les salariés selon par exemple le sexe, l'ancienneté, … D'autres analyses peuvent être menées comme la liaison entre sexe et minorité, CSP – sexe, …, à vous d'être imaginatif.
Au final, peut-on prévoir le salaire actuel en fonction des autres variables ? Si oui quelles sont les variables les plus discriminantes ? Et selon quel modèle ?

Pour évaluer la pertinence de votre modèle, on utilise le plus souvent le RMSLE. En gros, c'est une moyenne des erreurs de prévision au carré.

$$RMSLE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Le RMSLE doit être le plus petit possible.
- $\boldsymbol{n}$ est le nombre d'individus (ici 76)
- $\boldsymbol{p}_i$ est la valeur prévue pour le salaire actuelle pour chaque individu i
- $\boldsymbol{a}_i$ est la valeur actuelle du salaire pour chaque individu i

Restitution de votre travail.
Le travail est à rendre pour le 31 mars sous 3 formes.
- un fichier pdf avec votre analyse, vos graphiques, vos résultats statistiques
- un fichier xls avec les 76 individus et 3 colonnes : l'identifiant, le salaire actuelle et le salaire prévu
- et le fichier R correspondant.

Bon courage

mon email pour toute question et pour la restitution : pascal.bernard@datalone.com