STT7335 - Travail de session

Véronique Tremblay

Hiver 2020

Données

Le jeu de données «entrain_valid.csv» contient les résultats d'une enquête menée auprès de personnes ayant obtenu un diplôme en 2010. Voici une brève description des variables.

Variables de nature sociodémographique:

- age: Âge au moment d'obtenir le diplôme en 2010.
- genre: Homme ou femme.
- langue : Langue parlée à la maison.
- minorite_visible : Est-ce que le répondant s'identifie à une minorité visible? (Oui ou non).
- scol_mere : Niveau de scolarité de la mère.
- scol_pere : Niveau de scolarité du père.
- scolarite_avant : Niveau de scolarité avant de commencer le diplôme obtenu en 2010.
- poids: poids d'échantillonnage.

Variables concernant le diplôme obtenu en 2010 et les emprunts qui y sont liés:

- diplome_secteur_2010: Secteur dans lequel le diplôme a été obtenu en 2010.
  - 1 Éducation
  - 2 Arts visuels et d'interprétation, et technologie des communications
  - 3 Sciences humaines
  - 4 Sciences sociales et de comportements, et droit
  - 5 Commerce, gestion et administration publique
  - 6 Sciences physiques et de la vie, et technologie
  - 7 Mathématiques, informatique et sciences de l'information
  - 8 Architecture, génie et services connexes
  - 9 Agriculture, ressources naturelles et conservation
  - 10 Santé, parcs, récréation et conditionnement physique
  - 96 Ne s'applique pas
  - 99 Non déclaré
- scolarite_obtenue_2010: Niveau de scolarité obtenu en 2010.
- emprunt: Montant total de l'emprunt au moment de l'obtention du diplôme en 2010.
- emprunt_gouv: Emprunt au gouvernement pour études (oui ou non).

Variables concernant le statut du diplômé en 2013:

- etude_temps_plein: Est-ce que la personne est toujours étudiante à plein temps en 2013?
- secteur_emploi_2013: Secteur dans lequel la personne travaille en 2013

  - 1 Secteur de la production de biens
  - 2 Commerce, transport et entreposage
  - 3 Finance, assurances, services immobiliers, administration publique
  - 4 Services professionnels, scientifiques et techniques
  - 5 Services d'enseignement
  - 6 Soins de santé et assistance sociale
  - 7 Autres services
  - 96 Ne s'applique pas
  - 99 Non déclaré

- satisfaction_emploi: Est-ce que la personne est satisfaite de son emploi en 2013? (Oui ou non)
- salaire_2013: Revenus d'emploi en 2013.

- sur_sous_qualification: Est-ce que la personne est sur-qualifiée ou sous-qualifiée par rapport à l'emploi occupé en 2013.

  - 0 Exigence en éducation non précisée
  - 1 Repondant a plus d'éducation que requis pour l'emploi
  - 2 Repondant a même éducation que requis pour l'emploi
  - 3 Repondant a moins d'éducation que requis pour l'emploi
  - 6 Ne s'applique pas
  - 9 Non déclaré

- lien_emploi_etude: Nature du lien entre l'emploi occupé en 2013 et le diplôme obtenu en 2010.

  - 1 Un lien étroit
  - 2 Lien faible
  - 3 Aucun lien
  - 6 Ne s'applique pas
  - 7 Ne sait pas
  - 9 Non déclaré

- dette_2013: Montant de la dette en 2013

Questions

Vous devrez:

1. Préparer le jeu de données et en faire une analyse exploratoire complète, incluant la création de regroupements (segmentation). Votre rapport devra expliquer les étapes réalisées et présenter les tableaux et les graphiques **les plus intéressants**. (3 pages maximum)
2. Créer un modèle permettant de prédire la satisfaction par rapport à l'emploi et **discuter** des facteurs qui affectent ce niveau de satisfaction. Votre rapport doit décrire brièvement votre méthodologie. (1 page)
3. Créer un modèle permettant de prédire le revenu annuel d'un diplômé à partir des données **connues à l'obtention du diplôme en 2010**. Votre rapport devra comprendre une explication de votre méthodologie et une comparaison de la performance de vos modèles. (1 page)
4. Créer un modèle permettant de prédire (au moment de l'obtention du diplôme) si une personne qui a fait un emprunt aura ou non remboursé plus de la moitié de son emprunt trois ans après l'obtention du diplôme (excluez les personnes qui sont encore à l'école à temps plein en 2013). Votre rapport doit présenter votre méthodologie et une discussion des difficultés rencontrées. Vous devez aussi analyser

la performance de votre modèle selon au moins trois critères. Discutez des résultats. (1 page). Analysez la performance de votre modèle en fonction des variables sensibles et **discutez** des résultats en considérant, par exemple, le contexte où l'on souhaiterait utiliser ce modèle pour accorder ou non un prêt. (1 page)

Pour les points 3 et 4, vous devrez faire une prédiction sur un échantillon test.