AZZOUZ-THUEROZ

Maxence

Quantmetry

Test technique

1. Statistiques descriptives
2. En vu de préparer les données à l’algorithme d’apprentissage supervisé, certaines données qui n’ont pas un format numérique devront subir une préparation au préalable.

On constate que l’ensemble des variables contient des valeurs manquantes et que ces valeurs manquantes ne sont pas des lignes entières de valeurs manquantes, il convient donc de chercher à conserver ces données en remplaçant les valeurs manquantes par les valeurs adéquates.

Notons que ces valeurs manquantes représentent une faible proportion du set (Une variable contient au maximum 0,5% de valeurs manquantes par rapport à son nombre d’observations).

Intéressons-nous d’abord à la feature liée à la date de réception de la candidature.

Les CVs des candidats ont été récupérés sur 5 années durant. Il semble intéressant de diviser la colonne de temps en trois colonnes ‘années’, ‘mois’, ‘jour’ pour identifier les périodes de temps les plus pertinentes pour expliquer l’embauche. On observe pour chaque variable avec un histogramme que les observations sont uniformément réparties, nous amenant à avoir une médiane égale à la moyenne lorsqu’on arrondit à l’entier, on utilisera donc ces valeurs pour remplacer les valeurs manquantes.

La variable relative à l’âge des candidats représentée à l’aide d’un histogramme et d’une fonction de densité univarié de Parzen-Rosenblatt nous montre que l’âge suit une distribution « en cloche » et impliquant une médiane et une moyenne équivalents et égales à 35 ans. Ceci m’amène donc à utiliser 35 ans comme valeur pour remplacer les valeurs manquantes de la feature ‘âge’.

On observe de même pour les autres variables numériques ‘exp’, ‘salaire’ et ‘note’ des distribution similaires nous amenant à adopter la même stratégie pour remplacer les valeurs manquantes.

Pour les variables catégorielles ‘cheveux’, ‘sexe’, ‘diplôme’, ‘spécialité’ et ‘dispo’, la répartitions des valeurs entre chaque catégorie étant inéquitable, on peut choisir de prendre le mode de la série pour remplacer les valeurs manquantes.

Pour entraîner l’algorithme supervisé visant à prédire si oui ou non un candidat va être embauché, on utilise la variable ‘embauche’ comme cible et les autres variables comme features, excepté ‘cheveux’ dont je considère qu’elle n’est pas pertinente pour expliquer l’embauche. Je conserve aussi la feature ‘sexe’ pour prendre en compte un potentiel biais discriminatoire dans l’embauche du candidat.

1. Pour mesurer tester la dépendance statistique entre deux variables catégorielle que sont la spécialité et le sexe, le test du V de Cramer permet de tester l’intensité de cette association, avec une statistique qui prend une valeur entre -1 et 1, 1 indiquant une liaison très forte. En l’occurrence, le test du V Cramer nous donne la valeur 0.36972941631623524, indiquant une forte liaison statistique entre ces deux variables.
2. Pour mesurer la dépendance statistique entre la couleur de cheveux et le salaire demandé, on peut effectuer un test du Chi 2 dans le cadre d’un régression linéaire. Tout d’abord on transforme la variable cheveux entre quatre variables catégorielles 'blond','brun','chatain' et 'roux'. Ensuite on procède à une régression de la variable salaire sur ces quatres variables. La P-value associé au test du Chi 2 nous laisse penser l’hypothèse nulle de corrélation non significative de ces variables est donc rejeté au profit de l’hypothèse alternative d’un impact significatif de ces quatre variables sur la variable dépendante qu’est le salaire.
3. De même, nous procédons à une régression linéaire de la variable note sur la variable expérience et la P-value basée sur la statistique de Student qui prend comme valeur zéro nous amène à rejeter l’hypothèse de non significativité statistique du coefficient de régression entre ces deux variables.
4. Machine Learning