Examen final

Professeur Olaf Kouamo
SDS - Spark Pour La Data Science
Année académique 2019-2020

December 22, 2019

Exercice 1. Téléchargez le dataset qui se situe dans *data/HousingData.csv*. Ce dataset contient les informations sur la valeur mediane des logements occupés par leur propriétaires en millier de dollars. L'idée est de prédire cette valeur en fonction des informations décrivant les zones. Informations décrite ci dessous

- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
- INDUS: Proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: Nitric oxide concentration (parts per 10 million)
- RM: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distances to five Boston employment centers
- RAD: Index of accessibility to radial highways
- TAX: Full-value property tax rate per $10,000
- PTRATIO: Pupil-teacher ratio by town
- B: $1000(B_k{}^{-}0.63)^2$, where Bk is the proportion of [people of African American descent] by town
- LSTAT: Percentage of lower status of the population

Notre variable ciblé étant :
MEDV: Median value of owner-occupied homes in $1000s

1. faire l'inventaire du nombre de lignes et de colonnes du data set.

2. tracer la distribution de la variable cible.

3. tracer la matrice de corrélation entre les variables.

4. Quelles sont les variables qui contiennent des valeurs manquantes et quelle est la proportion des valeurs manquantes pour chacune de ces variables? Imputer ces valeurs manquantes par une méthode de votre choix. Vous prendrez soin d'expliquer brievement la méthode utilisée

5. Calculer ainsi la matrice X constituée des variables explicatives et la cible Y. Ensuite découper les observations en X_{train}, Y_{train}, X_{test}, Y_{test}, ou la première partie servira de base d'apprentissage et la seconde de base de test.

6. Mettre en place un modele de regression linéaire pour prédire la variable cible.

   - On sélectionnera pour une première modelisation les variables les plus correlées à la target.
   - on verifiera la cohérence et la significativité de chacun des coefficients estimés avec un test statistique.
   - Ensuite, on rajoutera progressivement les autres variables et on regardera l'impact de ces dernières sur la cible.

7. tester plusieurs modèles non linéaaires et plusieurs combinaisons de paramètres afin de fournir le modèle avec la meilleure prédiction. On définira de facon claire et précise les méthodes d'évaluation de modèles mis en place. Pour ce faire, on comparera les erreurs obtenus sur l'ensemble test.

Exercice 2. Télecharger le fichier $data/african\_crises.csv$ représentant les crises économiques systémiques apparues dans certains pays africains de 1860 à 2014. L'objet de l'etude est de prédire en fonction de certains aggregats macro-économique décrits ci dessous si le pays va être en crise ou pas pendant une année donnée. Les variables dépendantes et la target sont décrites ci dessous:

- country: The name of the country
- year:The year of the observation
- systemic_crisis: "0" means that no systemic crisis occurred in the year and "1" means that a systemic crisis occurred in the year.
- exch_usd: The exchange rate of the country vis-a-vis the USD
- domestic_debt_in_default: "0" means that no sovereign domestic debt default occurred in the year and "1" means that a sovereign domestic debt default occurred in the year

- sovereign_external_debt_default: "0" means that no sovereign external debt default occurred in the year and "1" means that a sovereign external debt default occurred in the year
- gdp_weighted_default: The total debt in default vis-a-vis the GDP
- inflation_annual_cpi: The annual CPI Inflation rate
- independence: "0" means "no independence" and "1" means "independence"
- currency_crises: "0" means that no currency crisis occurred in the year and "1" means that a currency crisis occurred in the year
- inflation_crises"0" means that no inflation crisis occurred in the year and "1" means that an inflation crisis occurred in the year
- banking_crisis: "no_crisis" means that no banking crisis occurred in the year and "crisis" means that a banking crisis occurred in the year

1. Faire l'inventaire des variables explicatives dans les données sachant que la variable cible est *banking_crisis*.
2. Citer 3 différents modèles qui peuvent être utilisés pour répondre à la question posée.
3. Remplacer la colonne pays par les indicatifs de ces derniers: Pour simplifier, on utilisera les deux premières lettres de l'orthographe de chaque pays.
4. Pour chaque pays, donner la proportion des années de crises
    - Existe t-il des variables catégorielles parmi les variables explicatives? Si oui les dummifier
    - Quelle est la proportion de crises dans les observations (pays, année)?
    - Calculer la matrice X constituée des variables explicatives et la cible Y. Ensuite découper les observations en $X_{train}, Y_{train}, X_{test}, Y_{test}$, ou la première partie servira de base d'apprentissage et la seconde de base de test.
    - Modéliser la probabilité de défaut de payement (crise économique) d'un pays pour une année donnée par plusieurs modèles de votre choix. Pour choisir le meilleur modèle, on comparera alors les différents scores, précisions et rappels obtenus.