Rapport TP

Python for data analysis

Github: https://github.com/Stephane-Bcd/TP-Python-for-data-analysis-Regression

31-01-2020 BOUCAUD Stéphane

ESILV IBO A5

SOMMAIRE

INTRODUCTION

CONFIGURATION DE L'ENVIONNEMENT

J'ai rencontré pas mal de soucis lors de la configuration de mon environnement de développement notamment sur jupyter notebook.

Après pas mal de temps perdu là dessus (problèmes de ancement de jupyter, des executions des commandes de linux, etc etc), je suis donc passé sur une install toute fraiche de Ubuntu sur laquelle je serai libre de coder en python sans plus être embêté.

DATASET

Le dataset qui m'a été fourni provient d'ici:

https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log

Ce dataset contient des données anonymisées d'incidents provenant de la plateforme ServiceNowTM utilisée par une entreprise IT. Elles proviennent d'une base de données relationnelle et est sous format CSV.

ATTRIBUTS DU DATASET

1. number: incident identifier (24,918 different values);
2. incident state: eight levels controlling the incident management process transitions from opening until closing the case;
3. active: boolean attribute that shows whether the record is active or closed/canceled;
4. reassignment_count: number of times the incident has the group or the support analysts changed;
5. reopen_count: number of times the incident resolution was rejected by the caller;
6. sys_mod_count: number of incident updates until that moment;
7. made_sla: boolean attribute that shows whether the incident exceeded the target SLA;
8. caller_id: identifier of the user affected;
9. opened_by: identifier of the user who reported the incident;
10. opened_at: incident user opening date and time;
11. sys_created_by: identifier of the user who registered the incident;
12. sys_created_at: incident system creation date and time;
13. sys_updated_by: identifier of the user who updated the incident and generated the current log record;
14. sys_updated_at: incident system update date and time;
15. contact_type: categorical attribute that shows by what means the incident was reported;
16. location: identifier of the location of the place affected;
17. category: first-level description of the affected service;
18. subcategory: second-level description of the affected service (related to the first level description, i.e., to category);
19. u_symptom: description of the user perception about service availability;
20. cmdb_ci: (confirmation item) identifier used to report the affected item (not mandatory);
21. impact: description of the impact caused by the incident (values: 1â€“High; 2â€“Medium; 3â€–Low);
22. urgency: description of the urgency informed by the user for the incident resolution (values: 1â€“High; 2â€–Medium; 3â€–Low);
23. priority: calculated by the system based on 'impact' and 'urgency';
24. assignment_group: identifier of the support group in charge of the incident;

25. assigned_to: identifier of the user in charge of the incident;
26. knowledge: boolean attribute that shows whether a knowledge base document was used to resolve the incident;
27. u_priority_confirmation: boolean attribute that shows whether the priority field has been double-checked;
28. notify: categorical attribute that shows whether notifications were generated for the incident;
29. problem_id: identifier of the problem associated with the incident;
30. rfc: (request for change) identifier of the change request associated with the incident;
31. vendor: identifier of the vendor in charge of the incident;
32. caused_by: identifier of the RFC responsible by the incident;
33. close_code: identifier of the resolution of the incident;
34. resolved_by: identifier of the user who resolved the incident;
35. resolved_at: incident user resolution date and time (dependent variable);
36. closed_at: incident user close date and time (dependent variable).

ARCHITECTURE DU PROJET

A la racine du projet git, vous trouverez deux fichiers en particulier:

requirements.txt et install_required_packages.sh.

Le premier sert à installer les librairies requises pour les scripts python.

On les installe en utilisant la commande sudo pip3 install -r requirements.txt.

Le deuxième sert à installer les packages linux requis, comme par exemple python ou jupyter notebook.

Le fichier incident_event_log.csv est le dataset que je dois analyser dans le cadre de ce projet (voir sections précédentes pour plus d'infos).

Le dossier documents contient Le rapport actuel, mon powerpoint ainsi que le sujet du devoir.

Le dossier code, comme son nom l'indique contient tout le code du déroulé de ce devoir.

DOSSIER CODE

import_data.py est le script python qui va importer, filtrer et formatter les données afin de faire plus tard du modelling des données, notamment de la regression.

ANALYSE DES DONNÉES

IMPORT DES DONNÉES

C'est dans le ficher de code import_data.py que cela se passe.

La fonction import_data s'en charge et prend en paramètre le chemin vers le ficher des données en CSV.

Elle utilise la librairie pandas et retourne un dataframe

FILTRAGE ET FORMATTAGE DES DONNÉES

C'est dans le ficher de code import_data.py que cela se passe.

La fonction transform_data se charge de filtrer et transformer les données.