Analiza și predicția reziliierilor de servicii:
Un studiu utilizând modele de machine learning

ANDREI CHIRICA - MASTER II – IISC

ABSTRACT

Am efectuat analiza unui set de date referitor la rezilierile de servicii, cu accent pe predictivitatea acestor rezilieri. Iată un rezumat al acțiunilor tale:

- numpy
- matplotlib
- pandas
- seaborn

Concepts: • Programming; • Python; • Machine learning • Jupyter Notebook;

Additional Key Words: datasets, programming, compile, aggregation data

Reference Format:
Andrei Chirica - Master II - IISC. 2024. Analiza și Prognozarea Reziliierilor de Servicii: Un Studiu Utilizând Modele de Machine Learning (Ian 2024)

I.Introducere:

Preprocesare și explorare a datelor:

Am încărcat setul de date dintr-un fișier CSV pentru a începe analiza. Prima mea acțiune a fost să examinez primele câteva rânduri ale datelor, astfel încât să am o înțelegere inițială a structurii și conținutului lor.

O decizie importantă a fost să convertesc variabila 'Churn' în valori numerice, asignând 0 pentru clienții care nu au reziliat serviciile și 1 pentru cei care au făcut acest lucru. Acest pas a fost esențial pentru a permite utilizarea acestei variabile ca variabilă țintă în modelul de predicție.

În continuare, am considerat că eliminarea coloanei 'customerID' este benefică pentru analiză, deoarece aceasta nu furniza informații semnificative pentru predicția reziliierilor și putea introduce zgomot în modelele mele.

Pentru a trata valorile lipsă în coloana 'TotalCharges', am optat să aplic o abordare de completare a acestora cu media valorilor existente. Acest lucru a fost necesar pentru a evita pierderea datelor semnificative și pentru a asigura coerența setului de date.

O altă acțiune importantă a fost transformarea variabilelor categorice în valori numerice. Acest proces a inclus maparea categoriilor din variabilele precum 'PhoneService', 'MultipleLines', 'InternetService', etc., într-o reprezentare numerică adecvată pentru a putea fi folosite în modelele mele de machine learning.

Prin aceste etape de preprocesare, am asigurat că datele sunt pregătite pentru a fi utilizate în construirea și evaluarea modelelor de machine learning pentru a prognoza rezilierile de servicii.

Tehnologiile multimedia utilizate în această aplicație de editare audio-video în Python includ:

2.1.cv2 (OpenCV): Este o bibliotecă populară utilizată pentru procesarea și manipularea imaginilor și a fluxurilor video. Cu ajutorul OpenCV, aplicația poate efectua operații de manipulare a imaginilor, cum ar fi redimensionarea, recortarea, filtrarea și extragerea de cadre din fișiere video.

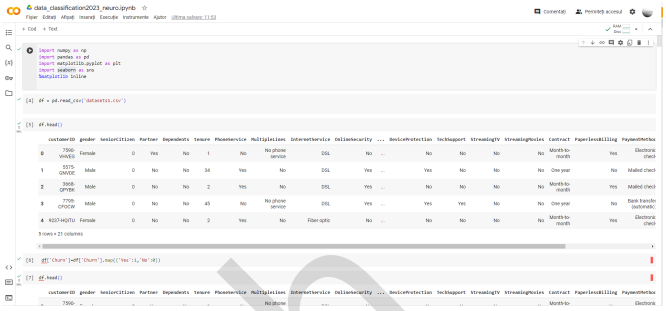

Fig. 1 - Importuri în Google Colab

Author: Andrei Chirica - Master II - IISC.

II.Vizualizare a datelor:

Am ales să utilizez biblioteca Seaborn pentru a crea vizualizări relevante în scopul explorării datelor. Prin diagrame de tip count și hărți de căldură, am căutat să evidențiez relațiile și tendințele în setul de date.

În special, am creat diagrame de tip count pentru variabila 'Churn' pentru a evalua distribuția reziliierilor în setul de date. Aceste diagrame au furnizat o perspectivă clară asupra proporțiilor de clienți care au rămas sau au reziliat serviciile.

Hărțile de căldură au fost utile pentru a explora corelațiile între diferite variabile. Prin colorarea celulelor în funcție de intensitatea corelației, am identificat relațiile puternice sau slabe dintre diverse caracteristici, oferind astfel insights în legătură cu posibile influențe ale acestora asupra variabilei țintă 'Churn'.

Aceste vizualizări au avut rolul de a evidenția modele sau trenduri în datele noastre, furnizând astfel informații utile pentru etapele ulterioare ale analizei și construirii modelelor de machine learning. Am considerat că este important să împărtășesc aceste vizualizări în cadrul proiectului, pentru a facilita înțelegerea și interpretarea datelor de către cei interesați.

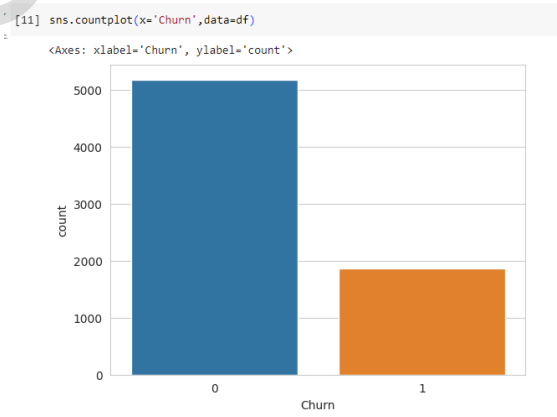

Fig. 2. Diagrame de tip Count

Heat maps:

Pentru a crea o harta de căldură pentru a explora corelațiile dintre variabile, ai folosit codul:
sns.heatmap(df.corr(), annot=True, cmap='viridis')

Această linie de cod a generat o hartă de căldură care a evidențiat corelațiile între diferitele caracteristici ale setului de date.

III. Construirea modelelor de machine learning:

Am început prin a antrena un model de regresie logistică pe setul de date complet. Scopul acestui pas a fost să învățăm relațiile dintre variabilele independente și variabila țintă 'Churn'. Linia de cod relevantă a fost:

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X, y)
```

Acest cod a creat și antrenat modelul de regresie logistică, permitându-ne să evaluăm performanța acestuia pe setul de date.

Apoi, am avansat spre construirea unui clasificator K-nearest neighbors (KNN) cu diferite valori ale parametrului k. Am explorat variantele de k pentru a identifica cea mai potrivită valoare pentru predicții precise.
Codul relevant a fost:

```
From sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X, y)
```

Am utilizat, de asemenea, o buclă pentru a explora mai multe valori ale lui k și a evalua performanța fiecărui model KNN pe setul de date.

Ultima etapă a implicat evaluarea modelului pe un set de testare și compararea performanței dintre regresia logistică și KNN. Acest lucru a fost realizat cu linii de cod care au prevăzut rezultatele și au calculat apoi acuratețea modelelor:

```
y_pred = logreg.predict(X_test)
accuracy_logreg = metrics.accuracy_score(y_test, y_pred)
y_pred_knn = knn.predict(X_test)
accuracy_knn = metrics.accuracy_score(y_test, y_pred_knn)
```

IV. Validare încrucișată și Optimizare a Modelului KNN:

Am dorit să evaluăm performanța modelului K-nearest neighbors (KNN) într-un mod robust, astfel că am ales să efectuăm validare încrucișată. Scopul acestei etape a fost să evaluăm modul în care modelul se comportă pe date diferite și să identificăm cel mai potrivit parametru k.

Am inițiat un proces de validare încrucișată utilizând un interval de valori pentru k, în vederea identificării celei mai bune configurații a modelului. Codul relevant a fost:

```
from sklearn.model_selection import cross_val_score
k_range = list(range(1, 31))
scores = []
For k in k_range:
        knn = KNeighborsClassifier(n_neighbors=k)
        scores.append(np.mean(cross_val_score(knn, X, y, cv=10, scoring='accuracy')))
```

Acest cod a parcurs diferite valori ale lui k și a evaluat performanța modelului KNN pentru fiecare dintre acestea prin intermediul unei validări încrucișate cu 10 fold-uri. Rezultatele au fost înregistrate pentru a putea identifica valoarea optimă a lui k. Următorul pas a constat în identificarea valorii optime a lui k pentru cea mai bună performanță. Aceasta a fost realizată prin găsirea valorii maximale a scorurilor obținute în procesul de validare încrucișată și asocierea acesteia cu valoarea corespunzătoare a lui k:

```
optimal_k = k_range[np.argmax(scores)]
```

Prin această abordare, am obținut valoarea optimă a lui k, care ne-a permis să optimizăm performanța modelului KNN pe setul de date. Acest pas a fost crucial pentru a asigura o predictivitate cât mai precisă a rezultatelor.

Fig.3 Performanța modelului KNN

V. Concluzii si perspective viitoare:

Analiza datelor și modelele de machine learning, cum ar fi regresia logistică și K-nearest neighbors, au relevat o precizie semnificativă în anticiparea rezilierilor de servicii. Pentru a consolida eficacitatea acestor modele, perspectivele viitoare ar putea implica explorarea unor algoritmi avansați suplimentari, integrarea de noi caracteristici în analiza datelor și fine-tuning-ul parametrilor pentru a optimiza performanța și a evalua impactul lor practic în contextul real al industriei serviciilor și al retenției clienților.

Bibliografie

1. https://ieeexplore.ieee.org/document/4160265
2. https://joss.theoj.org/papers/10.21105/joss.00205
3.https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
4.https://www.researchgate.net/publication/372302037_Machine_Learning_with_Python_Theory_and_Implementation?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6ImhvbWUiLCJwYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19
5.https://www.researchgate.net/publication/309045280_Toward_a_new_Advanced_Hydrologic_Prediction_Service_AHPS?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6ImhvbWUiLCJwYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19
6.https://diva-portal.org/smash/resultList.jsf?aq=%5B%5B%7B%22organisationId%22%3A%22879223%22%7D%5D%5D&aq2=%5B%5B%5D%5D&aqe=%5B%5D&af=%5B%5D&language=en&dswid=-7716
7.https://github.com/blondeincode/Telco_customer_churn_modelling/blob/main/Telco_Customer_Churn_Modelling.ipynb
8.https://github.com/viandwip/Employee-Attrition-Prediction-by-Using-Machine-Learning-main
9.https://github.com/ha-mou-ahmed/Prediction-of-the-number-of-resignations-in-a-company-with-supervised-learning-algorithms-in-Python-
10.https://www.kaggle.com/code/kutlukatalay/telco-churn-prediction