UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Pràctica 1

Inteligencia artificial

Xarxes Neuronals I Deep Learning

Autor

Artur Aubach, Altes

Mejia Rota, Cesar Elias

Grup 12

Professors:

Daniel Hinjos García

Luis Antonio Belanche Muñoz

Quadrimestre Primavera 2023/2024

TAULA DE CONTINGUTS

1. IDENTIFICACIÓ DEL PROBLEMA

1.1 Dades

La base de dades seleccionada per a aquest estudi és **smartphone_data.csv**, un conjunt de dades en el qual cada fila correspon a un model de mòbil diferent i cada columna representa una característica específica d'aquest model.

1.2 Objectiu

L'objectiu principal d'aquest projecte és desenvolupar un model de regressió capaç de predir el preu de mòbils basant-se en les seves característiques tècniques. Aquesta capacitat predictiva és fonamental per poder establir estratègies de preus més efectives per a models existents, així com per preveure el preu de dispositius nous abans del seu llançament al mercat. D'aquesta manera, es busca oferir una eina útil per a la presa de decisions en el marc.

1.3 Anàlisis dels models

Per a la tasca de predicció s'han seleccionat dos models diferents: la regressió lineal i el Perceptró Multicapa. Per tant és important considerar les següents passes:

- **Comprovació de Correlació:** Abans de procedir, és vital analitzar la correlació entre les variables independents per evitar problemes de multicolinealitat en la regressió lineal. Aquest pas no és tan crític per al MLP, però pot ajudar a reduir la complexitat del model.
- **Transformació de Característiques Categòriques:** Si les dades inclouen atributs categòrics, hauran de ser convertides a un format numèric mitjançant tècniques com l'encodificació one-hot o l'encodificació de etiquetes, depenent de la naturalesa de la dada i el model a utilitzar.
- **Normalització:** La normalització ajustar les dades a una distribució amb una mitjana de zero i una desviació estàndard d'una cosa que és crucial per al MLP per garantir que el gradient descendent funcioni de manera eficient durant l'entrenament.

2. ANÀLISIS I PREPROCESSAT DE DADES

Per aquest projecte, s'han implementat algunes modificacions respecte al procediment que suggereix el enunciat per evitar possibles fugues de dades (data leaking) durant el preprocessament. A continuació, detallem les etapes realitzades durant aquesta fase:

2.1 Anàlisi Univariant de les Dades

Aquesta anàlisi inicial ens ha permès obtenir una comprensió bàsica del comportament de cada variable independentment, incloent la distribució, la mitjana, la mediana i la variància, ajudant-nos a identificar patrons inicials i anomalies. Tambe ens hem fixat que la variable preu segueix una exponencial.

2.2. Anàlisi Multivariant

L'anàlisi ha revelat la rellevància de la marca del mòbil en la determinació del seu preu, suggerint una associació significativa entre aquesta característica categòrica i la variable dependent.

2.3. Estudi de Balanceig de Classes

Com que l'objectiu del model és una regressió, no s'ha realitzat un estudi de balanceig de classes, ja que aquest és més pertinent per models de classificació.

2.4. Estudi de Missings

S'ha analitzat el percentatge de dades mancants per cada variable. Aquesta informació és crucial per decidir com gestionar aquestes absències en les etapes següents del preprocessament.

2.5. Estudi d'Outliers

Hem investigat sobre la web els outliers relacionats amb el preu i hem comprovat que no es tracta d'errors, sinó que els preus corresponen als que pertoca a cada producte (pàgines linkades a l'annex). No obstant això, hem descobert que aquests outliers presenten propietats característiques que provoquen que el preu sigui excepcionalment alt, com per exemple un mòbil fet d'or. Aquestes característiques no estan reflectides en la base de dades. Atès que el comportament del preu del mòbil no està influït per les característiques comunes sinó per factors externs, hem

decidit excloure aquests individus de l'anàlisi. Continuant amb l'anàlisi dels outliers, hem observat que també existeixen outliers relacionats amb la capacitat de les bateries i mes. Aquests valors, inicialment sospitosos, han estat verificats com a reals mitjançant la consulta de diverses pàgines web

2.6. Creació de Característiques

Basant-nos en variables com "RESOLUTION", "EXTEND MEMORY" i "FAST", s'han creat noves característiques. Les raons específiques i les metodologies aplicades estan detallades en el notebook corresponent.

2.7. Partició de les Dades

Inicialment, les dades van ser barrejades (shuffled) per garantir una distribució aleatòria, posteriorment, es va procedir a dividir el conjunt en dades de formació (train) i de prova (test). La partició de validació es farà després de completar tot el preprocessament.

2.8. Tractament de Missings

Després de l'anàlisi inicial i la partició de les dades, es va decidir imputar els valors mancants. S'ha dut a terme una estratègia on es generaven intencionadament dades mancants, seguit d'un anàlisi comparatiu entre diferents models d'imputació, resultant en la selecció de MICE com el mètode més efectiu per les dades numèriques i Hot-Deck per a dades categòriques,

2.9. Codificació de Variables

Com es va esmentar en la secció 1.3, per a que els nostres models funcionin correctament, totes les variables han de ser numèriques. Per a les variables categòriques que no presenten un ordre inherent, s'ha aplicat l'estratègia de codificació OneHotEncoder.

2.10. Reavaluació Post-Preprocessat

Un cop completat el preprocessat, s'ha realitzat una nova anàlisi univariant per observar com han evolucionat les dades a conseqüència de les intervencions realitzades, assegurant-se que les transformacions aplicades mantinguin la integritat de la informació original al mateix temps que optimitzin la seva utilitat per a la modelització.

3. PREPARACIÓ DE VARIABLES

Per garantir la qualitat i eficàcia dels models de predicció que s'utilitzaran en aquest estudi, la preparació adequada de les variables és fonamental. Aquesta etapa inclou diverses tècniques clau per optimitzar les dades.

3.1 Separació de la Variables objectiu

La separació de la variable objectiu de les variables predictores és un dels primers passos en la preparació de dades per a l'aprenentatge automàtic. Aquest procés és crucial per assegura que el model no té accés a la variable objectiu durant l'entrenament, evitant així el risc de sobreajustament.

3.2 Transformació variable preu

Tal com hem observat en l'apartat "2.1. Anàlisi Univariant de les Dades", la variable objectiu "preu" presenta una distribució exponencial. Per aquest motiu, hem decidit aplicar una transformació logarítmica a aquesta variable. Aquesta transformació busca normalitzar la distribució dels preus i millorar l'eficàcia del model a l'hora de captar i interpretar les tendències subtils en les dades.

3.2 Normalització de les Variables

Conforme es va mencionar en l'apartat "1.3 Anàlisi dels models utilitzats", la normalització de les dades és crucial, especialment quan s'utilitzen models com el Perceptró Multicapa. La normalització implica ajustar les dades perquè tinguin una mitjana de 0 i una desviació estàndard 1, facilitant així l'aprenentatge dels models en minimitzar les problemàtiques associades amb les diferents escales de les variables.

3.3 Anàlisi de Correlacions entre Variables Numèriques

Una part essencial de la preparació de les dades és l'anàlisi de correlacions entre les variables numèriques. Aquest procés ajuda a identificar i eliminar la multicolinealitat, que pot causar overfitting i inestabilitats en els models de regressió. A més, després de l'aplicació de la codificació OneHot, el nombre de variables s'ha incrementat significativament, augmentant així el risc de correlacions altes. Per això, s'ha decidit eliminar aquelles variables que tenen una correlació

superior al 0.85, assegurant-se que les dades mantinguin la màxima independència possible entre elles ('brand_name_apple', 'processor_brand_bionic', 'os_android').

3.4 Estudi de Dimensionalitat amb PCA

Tot i que la reducció de la dimensionalitat pot comprometre la interpretabilitat de les dades, aquest risc es minimitza quan es tracta de models com xarxes neuronals, on la comprensió detallada de cada variable individual no és tan crítica com en altres contextos analítics. Per aquest motiu, s'ha aplicat l'anàlisi de components principals per reduir el nombre de variables i concentrar-se en les més informatives. S'ha escollit mantenir aquelles components que expliquen fins al 97% de la variància total, ja que l'anàlisi gràfica ha mostrat que més enllà d'aquest punt, la pendència de la curva de variància acumulada es reduïa dràsticament.

4. MODELITZACIÓ

4.1. Model lineal base

El model explica aproximadament el 91.28% de la variabilitat en els preus dels telèfons mòbils, basant-se en les seves característiques. El valor de R^2 és raonablement alt per ser un model inicial, la qual cosa suggereix que s'ha realitzat un bon preprocessament de les dades. Tot i que comparat amb el test, que només dóna un 69,12%, pot donar indicis de sobre ajustament o alta variància.

La majoria dels residus semblen estar concentrats prop de la línia zero, especialment per a valors predits entre -1 i 1. Això generalment és un bon indicador i mostra que el model és prou precís per a la majoria de les prediccions.

Visualment, el gràfic suggereix una possible heteroscedasticitat ja que la dispersió dels residus sembla augmentar amb el valor de les prediccions. Això es nota per la manera en què els residus es dispersen més àmpliament a mesura que augmenten els valors predits, en lloc de mantenir un ample constant al voltant de la línia vermellosa puntejada que representa un residu de zero.

4.2. Perceptron multicapa base

Per el primer model de perceptró, hem utilitzat una una capa del tamany del dataset com a única capa. Els model no ha arribat a convergir, i per tant hem ideat com a solucions augmentar el número màxim de epochs i pujar el learning rate. Inicialitzarem tots els paràmetres de manera arbitrària i els anirem modificant en busca del millor model possible.

4.3. Perceptron multicapa segona iteració

El segon model si que ha convergit amb els canvis mencionats anteriorment, amb uns resultats per a la pèrdua (loss) i el R^2 score de 0.04 i 0.96 al train i 0.31 i 0.48 al test. Això ens fa veure que el model pateix overfitting. Abans de solucionar això, però, volem evitar que els models que fem pròximament facin més epochs de les que realment necessiten, per això afegirem un early stopper al següent model.

4.4. Perceptron multicapa tercera iteració

Amb el canvi pertinent veiem com el número de epoch realitzades a passat de ser 800 a unes 200. El següent canvi ara anirà orientat a poder entendre millor les gràfiques, ja que loss i R^2 comencen amb valors molt extrems. Per intentar canviar això, intentarem inicialitzar les xarxes amb certs valors i no a 0.

4.5. Perceptron multicapa quarta iteració

El canvi dut a terme no ha millorat les gràfiques com volíem, però ens ha servit per millorar el rendiment de la xarxa i per tant el mantindrem. Ara si que tractarem l'overfitting, i per fer-ho farem servir regularització. Utilitzarem regularització l1.

4.6. Perceptron multicapa cinquena iteració

La regularització ha sigut efectiva, obtenint un 0.9 de R^2 score al train i un 0.85 al test. Intentarem millorar aquests resultats augmentant el número de capes i de neurones per capa de la xarxa.

4.7. Perceptron multicapa sisena i setena iteració

Provant a afegir capes hem afegit primer una nova capa i, al veure la millora en els resultats, hem afegit dues capes més. Aquestes capes noves no han millorat gairebé res l'R^2 i, en canvi, ha fet que augmenti la pèrdua (loss), per tant hem decidit quedar-nos amb la versió de dues capes. Com a últim intent de millora, intentarem disminuir el batch size per veure si obtenim millors resultats.

4.8. Perceptron multicapa iteració

Després de reduir el batch size, els resultats no han millorat i el model ha passat de estar sobre ajustat a estar sota ajustat. Tenint en compte que un batch size més gran implica menys cost computacional, hem decidit quedar-nos amb el model de la sisena iteració. A continuació comentarem els resultats d'aquest model, tot i que es poden trobar tots els resultats i les gràfiques de cada model al notebook adjunt.

4.9. Perceptron multicapa: model final

El model final consta de 2 capes, amb 59 i 118 neurones respectivament, i una capa de output amb 1 sola neurona. Les capes ocultes tenen una funció d'activació ReLU mentre que la de output, al ser una regressió, té la funció identitat. El model ha obtingut una pèrdua i un R^2 score de 0.14 i 0.92 al train i de 0.16 i 0.89 al test.

5. CONCLUSIONS

El projecte ha demostrat una capacitat excepcional per optimitzar i aplicar metodologies de preprocessament i modelització en la predicció de preus de mòbils. L'anàlisi detallada de les dades ha proporcionat una base sòlida per a la transformació de variables i l'eliminació d'inconsistències, com ara outliers i valors mancants, millorant significativament la qualitat de les dades per a l'entrenament dels models, com s'ha pogut observar en els bons resultats del model lineal. La implementació de diverses iteracions del Perceptró Multicapa ha evidenciat la importància de l'ajustament dels paràmetres per millorar la generalització del model. El model final, amb un R^2 de 0.89 en el conjunt de prova, reflecteix una adaptació adequada, mostrant un equilibri entre complexitat i rendiment, apte per a la presa de decisions en estratègies de preus.

6. ANNEX

OUTLIERS

price

- https://vertumobile.in/collections/new-signature-touch
- https://www.smartprix.com/mobiles/xiaomi-redmi-k20-pro-signature-edition-ppd12yins6c1
- https://www.smartprix.com/mobiles/huawei-mate-50-rs-porsche-design-ppd17mlgrez6
- https://www.smartprix.com/mobiles/huawei-mate-30-rs-porsche-design-ppd13522ip1y
- https://www.91mobiles.com/xiaomi-mi-mix-alpha-price-in-india

battery_capacity

- https://shorturl.at/wJO08
- https://shorturl.at/ntxFH