Models predictius

Enunciat

Semestre 2023.2

Índice

Introducción

En aquesta activitat usarem el conjunt de dades **Casas** sobre les característiques i preus d'habitatges de l'estat de Washington en USA. S'han recollit dades dels habitatges, com a superfície útil, nombre de banys i lavabos, superfície de la parcel · la, etc,. També es té informació sobre variables de l'entorn, com a vistes i proximitat dels veïns. L'objectiu principal d'aquest estudi és esbrinar quins són els factors que més influeixen a l'hora de determinar el preu d'una casa. Aquest tipus d'anàlisi és molt útil tant per a les agències, com per al mercat immobiliari en general, per a poder predir l'evolució del preu dels habitatges.

L'arxiu conté aproximadament 21600 registres i 16 variables. Les principals variables són:

- date: Data de venda
- price: Preu de venda
- bedrooms: Nombre d'habitacions
- bathrooms: Nombre de banys/lavabos
- sqtf_livingl: Superfície habitable (en peus al quadrat)
- sqft_lot: Superfície de la parcel · la (en peus al quadrat)
- floors: Nombre de plantes
- waterfront: Indica si l'habitatge té accés a un llac
- view: Tipus de vista (variable numèrica)
- condition: Estat de l'habitatge, codificada de l'1 al 5, sent 1 Molt mala i 5 Molt bona
- sqft_basement: Superfície del soterrani (en peus al quadrat)
- yr_built: Any de construcció de l'habitatge
- yr_renovated: Any de renovació de l'habitatge
- sqft_living15: Superfície habitable mitjana dels 15 veïns més pròxims
- sqft_lot15: Superfície de la parcel · la mitjana dels 15 veïns més pròxims

Primer s'estudiaran les possibles relacions lineals entre el preu de l'habitatge i les diferents variables independents. En la segona part de l'activitat es buscaran els possibles factors que intervenen perquè un habitatge tingui un preu més elevat i sigui considerada de luxe. Segons les dades d'aquest estudi el preu mitjà seria de 450000$.

---

1 Regressió lineal

1.1 Preparació de les dades

- Abans de començar amb l'anàlisi es passaran les variables de superfície mesures en peus quadrats a metres quadrats i els dòlars a euros. Podeu usar les conversions següents:

  $1\_\ peusquadrats = 0.0929\_\ metresquadrats$

  $1\_\ dolar = 0.93\_\ euros.$

- Posteriorment reviseu la naturalesa de cadascuna de les variables a estudi i comproveu si és necessari algun canvi. És a dir, si són de tipus character, factor, integer o numeric.

1.2 Estudi de correlació lineal

Una vegada efectuades les conversions, calculeu la correlació lineal entre totes les variables de l'estudi, exceptuant `date` y `zipcode`. Una vegada calculades les correlacions, *mostreu en forma de matriu, només aquelles que tinguin un coeficient de correlació superior a 0.2.* A més es demana ordenar-les de manera decreixent, és a dir del valor més alt de r, al més baix. Interpreteu.

NOTA: Recordeu no duplicar les variables convertides, és a dir no posar, per exemple, la superfície habitable en peus i en metres. Ens quedem amb els metres i euros.

1.3 Generació dels conjunts d'entrenament i de test

Per a poder estimar de forma més objectiva la precisió del model lineal, separarem el conjunt de dades en dues parts: el conjunt d'entrenament (training) i el conjunt de prova (testing). Ajustarem el model de regressió lineal amb el conjunt d'entrenament, i avaluarem la precisió amb el conjunt de prova. Es demana:

Genereu els conjunts de dades per a entrenar el model (training) i per a testar-lo (testing). Es pot fixar la grandària de la mostra d'entrenament a un 80% de l'original.

1.4 Estimació del model de regressió lineal

Estimeu per mínims quadrats ordinaris un model lineal que expliqui el preu dels habitatges en funció de totes les variables utilitzades per a calcular les correlacions en l'apartat anterior

1.4.1 Comprovació de colinealitat

Una vegada ajustat el model es demana comprovar la presència o no de colinealitat. A la vista dels resultats, exclouries alguna variable del model. Raona la teva resposta. Per a la comprovació de la colinealitat, se suggereix comparar els valors del factor de inflacción de la variància (FIV), amb el seu equivalent en el model ajustat, és a dir amb $1/(1-R^2)$, on R^2 és el coeficient de determinació del model. Els valors FIV majors que aquesta quantitat impliquen que la relació entre les variables explicatives és major que la que existeix entre la resposta i els predictors, i per tant donen indicis de multicolinealitat.

Nota: Generalment, valors d'un FIV superiors a 10 donen indicis d'un problema de multicolinealitat, si bé la seva magnitud depèn del model ajustat, però altres autors consideren valors per sobre de 4.

1.5 Diagnosi del model.

Per a la diagnosi es tria el model final construído en l'apartat anterior, una vegada decidit si s'elimina o no alguna variable per problemes de colinealitat. Es demana generar un histograma amb els residus (valors observats menys els predits pel model) i posteriorment altres dos gràfics: un amb els valors ajustats enfront dels residus (que ens permetrà veure si la variància és constant) i el gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment(QQ plot). Interpreteu els resultats.

1.6 Predicció del model

Segons el model final, calculeu les prediccions del preu que tindria un habitatge, utilitzant la base de dades de testing. Representeu els valors predits front els valors observats. Avalueu la precisió del model mitjançant l'arrel quadrada de l'error quadràtic mitjà (RMSE). Interpreteu.

2 Regressió logística.

2.1 Preparació de les dades

Es vol estudiar quins són els factors que més influeixen en el preu de l'habitatge en l'estat de Washington.

Per a això, primer es crearà una nova variable dicotòmica anomenada **price_re**. Aquesta nova variable està relacionada amb els valors de la variable `price_eur`. Es codificarà de la manera següent: "preu inferior a 500000 euros" pren el valor 0 i "preu superior o igual a 500000" el valor 1. Per al model de regressió logística, es prendrà com a variable dependent `price_re`.

De manera anàloga als models lineals, separarem el conjunt de dades en dues parts: el conjunt d'entrenament (training2) i el conjunt de prova (testing2). Ajustarem el model de regressió logística amb el conjunt d'entrenament, i avaluarem la precisió amb el conjunt de prova.

Es demana:

Una vegada recodificada la variable dependent, genereu els conjunts de dades per a entrenar el model (training2) i per a testar-lo (testing2). Es pot fixar la grandària de la mostra d'entrenament a un 80% de l'original.

2.2 Estimació del model de regressió logística

Prenent com a base, training2:

Estimeu el model de regressió logística sent la variable dependent `price_re` i prenent com a variables explicatives les triades en el model de regressió lineal final. Tingueu en compte que la variable `price_eur` sense recodificar cal eliminar-la com a variable explicativa. A més la nova variable `price_re`, com les variables `view` i `waterfront` han de ser transformades a factor. A la vista dels resultats, explica si eliminaries o no algun dels factors del model. A aquest model triat, se'n dirà model final.

2.3 Càlcul de les OR (Odss-Ràtio)

Resumir quins de les variables explicatives del model de regressió logística final generat en l'apartat anterior poden considerar-se factors de risc o protecció. Calculeu les OR corresponents i interpreteu.

2.4 Matriu de confusió

A continuació analitzeu la precisió del model final triat, comparant la predicció del model contra el conjunt de prova (testing). S'assumirà que la predicció del model és 1, *habitatges amb preu superior o igual a 500000 euros*, si la probabilitat del model de regressió logística és superior o igual a 0.5 i 0 en cas contrari. Analitzeu la matriu de confusió i les mesures de sensibilitat i especificitat.

2.5 Predicció

Segons el model final, prenent com a base training2, calculeu la probabilitat que un habitatge amb les mateixes característiques que la registrada en la tercera fila de la base de dades, tingués un preu superior o igual a 500000 euros.

2.6 Bondat de l'ajust

a) Avalueu la bondat de l'ajust, mitjançant la *devianza*. Perquè el model final sigui bo la devianza residual ha de ser menor que la devianza nul · la. En aquest cas el model prediu la variable dependent amb major precisió.

b) Avalueu l'eficàcia del model segons el test Chi-quadrat. En aquest cas el valor de l'estadístic Chi-quadrat observat és igual a la diferència de devianzas (nul · la-residual). Calculeu la probabilitat associada a l'estadístic del contrast utilitzant la funció **pchisq**.

2.7 Corba ROC

Dibuixeu la corba ROC i calcular l'àrea sota la corba. Discutiu el resultat.

3 Resum executiu. Conclusions de l'anàlisi

Resumiu les conclusions de l'estudi per a una audiència no tècnica, indicant les respostes a les preguntes de recerca plantejades. El resum no ha d'ocupar més de mitja pàgina.

Nota: aquesta pregunta treballa la competència de comunicació que és molt important en el rol d'analista de dades.

Puntuació de los apartats

- Apartat 1.1 y 1.3 (5%)
- Apartat 1.2 (10%)
- Apartat 1.4 (10%)
- Apartat 1.5 y 1.6 (10%)
- Apartat 2.1 (5%)
- Apartat 2.2 (10%)
- Apartat 2.3 (10%)
- Apartat 2.4 (10%)
- Apartat 2.5 y 2.7 (10%)
- Apartat 2.6 (5%)
- Apartat 3 (10%)
- Qualitat de l'informe dinàmic (5%)