Predicció de les tendències de vendes d'articles de moda sobre un dataset multimodal

Casadellà Cors, Guim: 1607484

Abstract—Aquest treball investiga la capacitat de predir les tendències de vendes d'articles de moda utilitzant tècniques tradicionals d'aprenentatge automàtic. S'ha utilitzat el dataset **VISUALLE**, que proporciona informació multimodal sobre diversos articles de roba, incloent dades visuals, textuals i de tendències de vendes. Inicialment, s'ha intentat abordar la tasca com a regressió, però es va evidenciar una dificultat deguda a la distribució desbalancejada de les vendes. Com a alternativa, s'ha optat per una classificació binària per identificar els articles amb vendes d'èxit, aplicant diferents tècniques de balanceig de classes i models de machine learning. Els resultats obtinguts mostren avanços però també evidencien limitacions, suggerint la necessitat d'enfocaments més avançats per millorar la precisió de les prediccions.

Keywords—*fashion, multimodal, machine-learning, forecasting*

1. Introducció

En el sector de la moda, anticipar les vendes futures d'articles és crucial per optimitzar les estratègies de màrqueting i gestionar eficientment l'inventari. Aquest estudi s'ha enfocat en la predicció de les tendències de vendes mitjançant l'ús de dades multimodals proporcionades pel dataset **VISUALLE**. A diferència d'enfocaments més complexos basats en deep learning, s'ha explorat l'eficàcia de metodologies tradicionals d'aprenentatge automàtic per abordar aquesta tasca.

Inicialment, s'ha plantejat la predicció com a problema de regressió per estimar les vendes totals, però els resultats han indicat una tendència dels models a predir valors mitjans a causa de la distribució desigual de les vendes. Per superar aquesta limitació, s'ha reorientat l'objectiu cap a una classificació binària destinada a identificar els articles amb vendes destacades. Aquest canvi ha implicat l'aplicació de diverses tècniques de balanceig de classes i l'ús de models més robustos davant el desbalanceig, com ara el XGBoost Classifier.

A més, s'ha realitzat una enginyeria de característiques enfocada a simplificar les dades de tendències de Google i les descriptors visuals, amb l'objectiu de millorar la rellevància de les variables utilitzades pels models. Tot i els avenços obtinguts, els resultats suggereixen que les tècniques tradicionals presenten dificultats per capturar les complexitats de les dades multimodals, indicant la necessitat d'explorar metodologies més avançades en futurs estudis.

Github del projecte: FashionSalesPrediction

2. Dataset, feina prèvia i objectiu

2.1. VISUALLE

Per a dur a terme el projecte s'ha decidit usar el dataset **VISUALLE**[3], el qual va ser publicat l'any 2021. Aquest, és un dataset multimodal que recull l'informació relacionada a les vendes de certs productes de la companyia real de moda *Nunalie*. Recull vora 45M tendències de venda de 5577 articles nous de roba entre els anys 2016-2019.

Més endavant s'entrarà en més detall, però l'informació multimodal inclou: imatge, metadata textual, informació relacionada amb les tendències de Google de l'article i, finalment, dades de les vendes del producte a partir de la sortida d'aquest.

2.2. Feina prèvia

2.2.1. Article inicial

El dataset va ser introduït inicialment pel paper *paper*. Si bé és un dataset molt ric, amb moltes features diferents, l'article es centrava principalment en fer una aproximació de sèries temporals per explorar l'efecte de les tendències de Google. Aquest primer paper utilitzava una arquitectura de deep learning d'un **transformer**, per a incloure les interaccions entre les diferents modalitats de dades. Per a aconseguir-ho, s'usa una xarxa diferent per a realitzar els encodings de cada modalitat. Seguidament, es permet l'interacció entre elles a través del mecanisme d'atenció al decorder.

2.2.2. Articles posteriors

A partir de l'article mencionat inicialment, han sortit quasibé una desena de papers amb plantejaments semblants. Fent servir l'ús de models complexos de deep learning per a representar i explicar les dades. [2]

2.3. Objectiu del treball

Un cop presentada la feina prèvia que s'ha fet al voltant del dataset, em va surgir la següent questió, que es pendrà com a objectiu principal i inicial d'aquest treball: **Podem fer predicció de les tendències de vendes d'articles de moda sobre un dataset multimodal amb estratègies tradicionals d'aprenentatge computacional?**

3. Exploració i Enginyeria de Dades

3.1. Variables Catagòriques

Primerament, farem una descripció inicial dels diferents atributs del dataset. Comptem amb tres variables textuals que es poden interpretar com a categòriques. La primera d'elles és la variable **category**. Aquesta descriu quin tipus d'article de roba és, des de `'kimono dress'` (vestit de tipus kimono) a `'short sleeves'` (camisa curta).

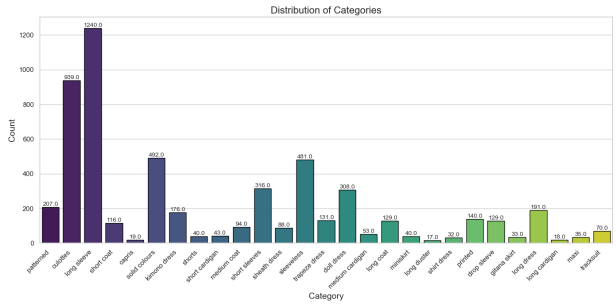


Figure 1. Distribució de les Categories

Com podem notar, hi ha certes categories amb molta més presència que les altres, mentre n'hi ha d'altres amb molt poca representació.

La variable **color** explica el color principal de la peça de roba, amb la següent distribució:

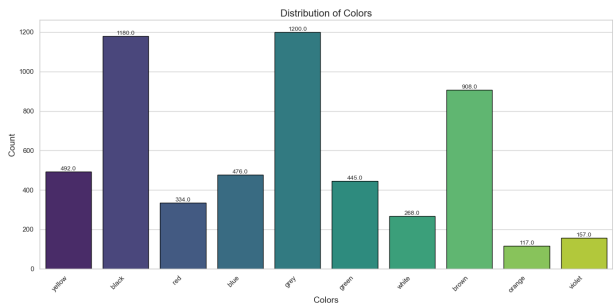


Figure 2. Distribució dels Colors

Finalment, tenim la variable **fabric**, que descriu el tipus de teixit de la peça. Alguns exemples són `'acrylic'` o `'georgette'`.

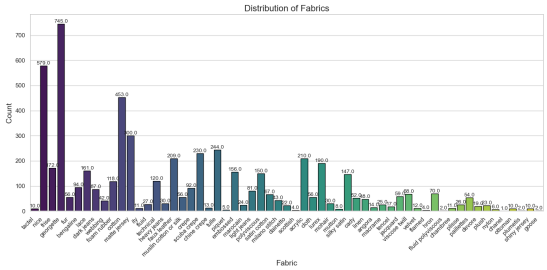


Figure 3. Distribució dels teixits

Aquestes s'han convertit a variables catagòriques mitjançant `LabelEncoders`

3.2. Tendències de Google

A continuació anirem a explorar les variables de les tendències de google. Per a cada una de les peces de roba del dataset, tenim la següent informació: Tant per la **categoria**, el **color** i el **teixit**, un vector de **52** elements numèrics. Aquests són la popularitat de cada un dels descriptors en un rànquing d'analítiques de google, amb valors entre 0 i 1.

És evident que aquests vectors afegeixen molta dimensionalitat al problema, doncs tenim 156 descriptors diferents per a cada article. És per això que s'ha volgut simplificar les features per a extreure'n valor i més interpretabilitat.

En aquest procés d'enginyeria de característiques, per a cada mostra i i cada tendència j, es calculen la mitjana mòbil i la suma mòbil sobre finestres de longitud 4 i 12 períodes. A continuació podem veure, per a una mostra concreta, l'efecte que té aquesta creació de variables:

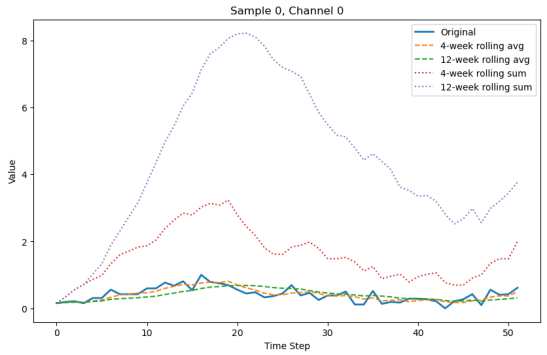


Figure 4. Exemple de tendències de Google

3.3. Variables temporals

De cara al tractament temporal de les mostres, tenim la data de sortida del model. Hem discretitzat la data per a extreure varies features:

- **Week**: Setmana de l'any
- **Month**: Mes de l'any
- **Year**: Any
- **Is_Weekend**: Binària que descriu si ens trobem al cap de setmana o no
- **Quarter**: Quatrimestre
- **Season:**: Estació de l'any

3.4. Descriptors visuals

3.4.1. Bag of Visual Words

A continuació s'explicarà el procés que s'ha dut a terme per a extreure valor descriptiu de les imatges. Aquest procés s'ha dut a terme amb l'algorisme **Bag of Visual Words**. Com que últimament s'ha fet molta literatura a classe al voltant d'aquest algorisme, no s'entraran en detalls del funcionament i només en decisions d'implementació. S'ha usat el feature extractor **SIFT**[1], doncs és el State of the Art. En un primer moment es deixava que l'algorisme extragués també els keypoints, com es veu a la següent figura. *Nota*: Els exemples estan en blanc i negre, però per l'implementació final s'usen els 3 canals de color.

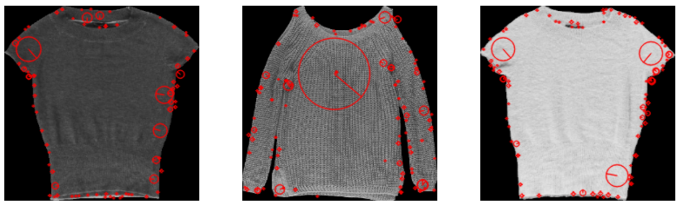


Figure 5. Exemples inicials dels keypoints

El problema amb aquesta aproximació és, tal com veiem, que al tenir unes imatges molt planes, doncs les textures acostumen a ser bastant regulars, treu molts pocs descriptors a aquestes zones. És per això que es va canviar per a usar un **Dense_SIFT** amb un mapa de keypoints creat cada 10 píxels. Amb una imatge d'entrada de 256x256, obtenim uns 625 descriptors per imatge.

Un cop processades totes les imatges, obtenim un total de 3434080 descriptors. Aquests llavors els passem per un calgorisme de clustering, en el nostre cas un **KMeans** amb 200 descriptors. Llavors, obtenim per a cada imatge un historgrama d'aquests que ens servirà com a les features visuals de cada una de les imatges

3.4.2. Embeddings CLIP

Com a afegidor, s'ha estat provat també l'efecte d'usar el model **Fashion CLIP** com a feature extractor visual, per a determinar si aconseguia capturar millors descriptors de les imatges. Aquest usa com a model base la feina d'*OpenAI* i s'ha fine-tunnejat en productes de moda FashionCLIP-GitHub. Aquest projecta en un espai d'embeddings de mida 512 les imatges. Un cop feta la inferència per a totes, hem guardat els vectors resultants. Cal mencionar que, tot i que s'han provat en els diferents apartats d'investigació, sempre han donat resultats iguals o pitjors que l'anterior. Així que s'han **acabat descartant**.

3.5. Vendes

A continuació farem exploració de la variable que ens interessa predir. Veiem que el dataset ens dona 12 valors relacionats amb les vendes dels articles de moda. Aquests corresponen a les vendes de l'article en les 12 setmanes que segueixen la seva sortida. Notem que els valors estan compresos entre 0 i 1, essent la majoria molt pròxims a 0.

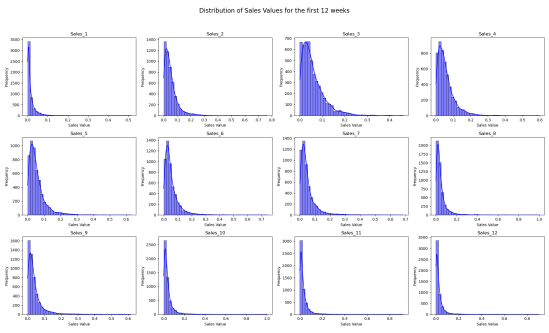


Figure 6. Distribució 12 setmanes de vendes

Cal mencionar que ha resultat bastant complicada la interpretació d'aquests resultats. Els creadors del dataset mencionen que els valors han estat normalitzats amb certa funció que els permet mantenir privacitat.

Si bé fer predicció dels 12 resultats en paral·lel sembla complex, s'ha fet una primera modificació agregant-los tots sota una mateixa

variable `total_sales`, que és la suma de les 12 anteriors. Presenta la següent distribució, amb molts valors centrats entre 0 i 1 i una cua per la dreta llarga però amb pocs valors.

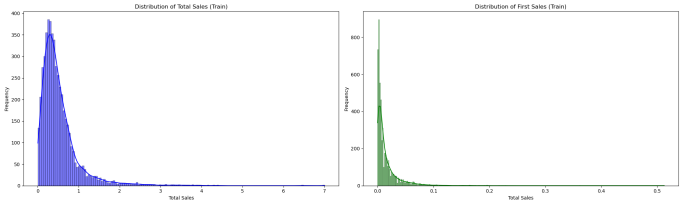


Figure 7. Enter Caption

També s'ha considerat el cas de realitzar la predicció només per les vendes de la primera setmana, amb la distribució de la figura de la dreta.

4. Primers models

Com s'ha comentat a la introducció, l'objectiu inicial del projecte era realitzar una predicció de les vendes dels articles de moda. En un primer instant, aquesta es va modelitzar com una tasca de regressió.

4.1. Regressió

Les primeres proves s'han realitzat amb un model de regressió lineal `LinearRegressor` amb mínimes tècniques de regularització, usant totes les features mencionades i predint sobre `total_sales`. En el següent gràfic, que compara els valors reals amb els predits, podem veure el comportament del model:

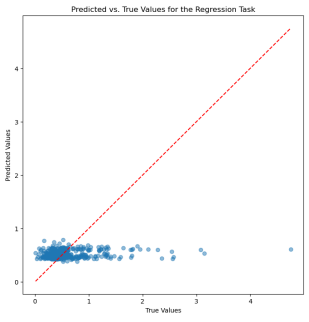


Figure 8. Resultats inicials regressió

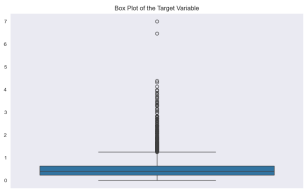


Figure 9. Diagrama Caixes i Bigots

Es pot veure clarament que el model tendeix a predir la mitjana de les dades, al voltant de 0.5, per a tot el dataset. Això és degut al fet que, com veiem al gràfic de la dreta, la densitat de dades en aquest punt és enorme, i això dificulta la regressió. Primerament, s'han provat varis mètodes de regularització, com ara la `Lasso`, `Ridge` o l'`ElasticNet` que les combina, però donaven resultats molt semblants. En aquests casos s'han normalitzat les dades. Posteriorment, s'han fet proves amb models més complexos, com ara l'`XGBRegressor`, un `RandomForest` o `AdaBoostRegressor`. En aquests últims s'han fet diverses proves amb els hiperparàmetres. Per a mirar d'eliminar aquesta tendència cap a la mitjana, s'han fet tant proves amb molts estimadors com amb molt pocs, per a mirar de lluitar contra l'overfitting.P erò no han resultat massa útil. A continuació s'inclouen els resultats obtinguts provant de fer les prediccions no per `total_sales`, sinó pels resultats de vendes de la primera setmana, a l'espera que com que hi ha menys outliers, millorés. També s'inclouen els resultats que s'han obtingut després de fer una transformació logarítmica sobre les dades. Es buscava una transformació que ajudés a explicar millor les dades a predir, però no ha estat possible mitigar el problema:

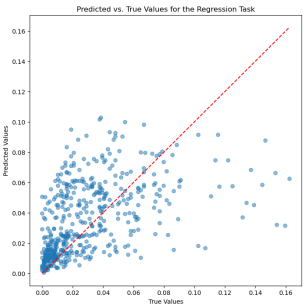


Figure 10. Vendes primera setmana

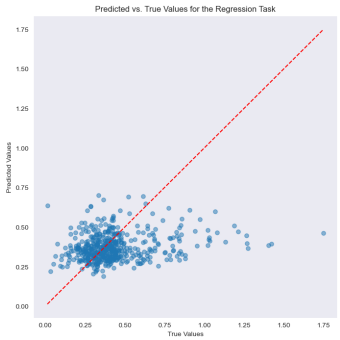


Figure 11. Transofrmació logarítmica

Cal mencionar que aquests han estat alguns dels resultats als quals s'ha arribat, deixant-ne molts d'altres fora. A banda, s'han fet altres proves, per exemple amb subconjunts de les features, arribant a resultats molt semblants. En aquest punt, la conclusió ha estat que, donada la complexitat de la tasca i les features que tenim, resulta molt difícil realitzar regressió per aquestes dades. Això ho confirma el següent plot fet amb la tècnica `UMAP`:

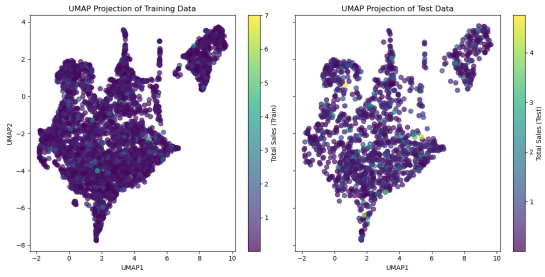


Figure 12. UMAP total sales

Ressalta com no estem explicant les dades amb les features que tenim.

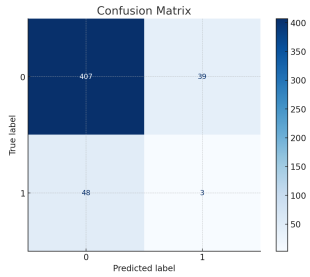
4.2. Classificació binària

Donat que sembla ser massa complicada pel model, s'han buscat maneres de simplificar la tasca. Així, s'ha optat per la **binarització** de les prediccions. S'ha determinat un valor de **treshold** que marcarà, per tots els valors y que estiguin per sobre, com a **èxit**. Llavors, l'objectiu del projecte ha virat a intentar ser capaços de predir quines vendes seràn un èxit respecte les altres. Inicialment, es va determinar aquest valor a 1 per a determinar la viabilitat de la tècnica.

Com que l'objectiu és detectar aquells casos d'èxit, tenim un **desbalanceig** important de classes, amb només un 16% de presència de la classe positiva. Aquesta realitat s'ha tingut en compte a l'hora d'aplicar les diferents estratègies mencionades a continuació:

- Us inicial d'un `RandomForestClassifier` amb el paràmetre `class_weight='balanced'`.
- Model `XGBClassifier` usant el paràmetre `scale_pos_weight`
- Aplicació de vàries tècniques de ressampling per a balancejar les classes, tant per **models lineals** com per **models boosting**:
  - **over_sampling** amb la tècnica `SMOTE`.
  - **under_sampling** amb el mòdul `RandomUnderSampler`.
  - Balanceig **combinat** amb la tècnica `SMOTEENN`

La següent és una matriu de confussió representativa dels resultats obtinguts amb les diferents tècniques. Concreatement, amb la tècnica `SMOTE` seguit d'un `LogisticRegressor`



Notem com els models són incapaços de diferenciar els casos d'èxit, tendint a predir la mateixa proporció en ambdós casos. Fet que, comprovat amb plots `UMAP` semblants a l'anterior, ens porten a cercar altres formes de millora.

Figure 13. Matriu Confussió

5. Models especialitzats

5.1. Intuïció

Fins aquí els problemes han estat bastant complicats. És ben clar que la correlació entre les variables que volem explicar, i la complexitat del que volem predir resulta advers. Així i tot, el procés de **discretitzar** la sortida sembla donar un camí més prometedor. És per això que es seguirà explorant. Finalment, s'ha decidit realitzar la següent simplificació per **classificació binària** de cara a la variable de predicció, ara anomenada `exit_sales_month`. Es farà predicció sobre les vendes **durant el primer mes**. A nivell conceptual, resulta adient voler predir si un article serà un èxit durant el primer mes. Això permet a la companyia destinar de forma adient els esforços de màrqueting. A part, resulta amb una distribució més estable que les anteriors, que facilita les prediccions. S'ha decidit marcar el **treshold** d'èxit com al percentil `80`. És a dir, es vol identificar el top 20% de les vendes. Seguidament veiem com queda aquesta distribució:

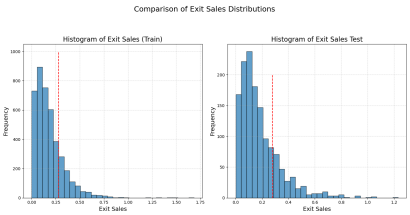


Figure 14. Distribució vendes primer mes

Si ve de moment no hi ha massa canvi respecte l'apartat anterior, ha estat el següent gràfic que ha donat l'intuïció per al segon plantejament del projecte:

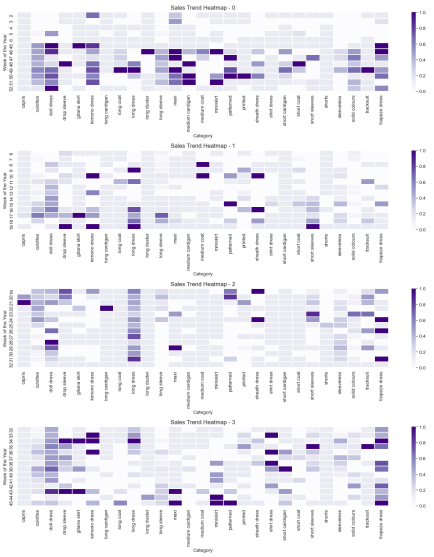
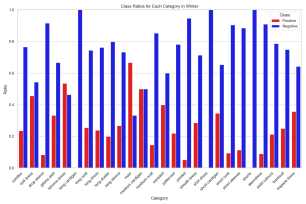


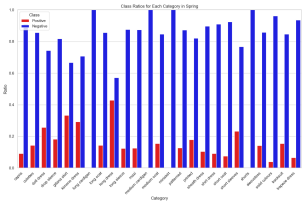
Figure 15. Dades per estació i categoria

Aquest plot mostra la següent informació: Diferenciant per setmanes classificades per estacions, i agrupant els articles per categoria, mostra la mitjana de la variable binària `exit_sales_month`. Podem apreciar que algunes categories com ara `kimono dress` estan molt presents durant l'hivern, són quasi nul·les durant l'estiu.

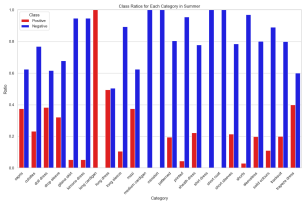
Podem veure les tendències de classes per categoria i estació als següents gràfics:



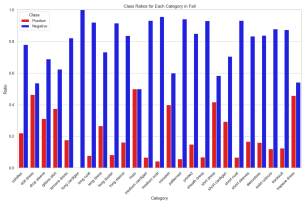
(a) Hivern



(b) Primavera



(c) Estiu



(d) Tardor

Figure 16. Ratios entre classes per estació i categoria

A continuació s'usarà aquestes idees per a millorar el desenvolupament del model.

5.2. Partició de dades i models

Donada la realitat de les dades que s'acaba de comentar, es podrien aprofitar aquestes particions per a fer un model per a cada una d'elles. Així, el model no hauria de centrar-se en aprendre les estacionalitats, i podrà posar èmfasi a extreure les característiques importants.

5.2.1. Pipeline

1. Calculem el `treshold 80%` sobre el dataset de **train**.
2. Particionem el dataset per **estacions**.
3. Per cada estació, separem les dades en cada una de les categories.
4. Realitzem tasques de **balanceig** i **regularització**.
5. Decidim si podem **construir** el model per la partició.
6. Fem el **fit** del **model**.
7. Calculem les diferents **mètriques** sobre el conjunt de test de la partició.
8. Les guardem per a futura anàlisis.

5.2.2. Models

Cal mencionar que moltes d'aquestes particions **no contenen dades**, o són molt minoritàries. Aquestes s'han tractat posteriorment.

Les tècniques de regularització han estat semblants a les descrites a l'apartat anterior, essent el `RandomUnderSampler` la que millors resultats donava, en els casos aplicables.

El model que ha resultat desenvolupar-se millor ha estat el `XGBClassifier`, amb uns `1500` estimadors, segurament òptim degut al gran nombre d'atributs provinent de les features visuals.

5.3. Mètriques

5.3.1. Generals

Donat que estem usant molts models, les mètriques s'han d'agregar d'alguna manera perquè siguin més interpretables. De cada un dels models, i per cada classe, recollim la `Precisió`, `Recall` i `F1-score`. També les mètriques ponderades pels pesos de classe. Després de llargues iteracions d'hiperparàmetres, models i tècniques, s'ha arribat a aquests resultats:

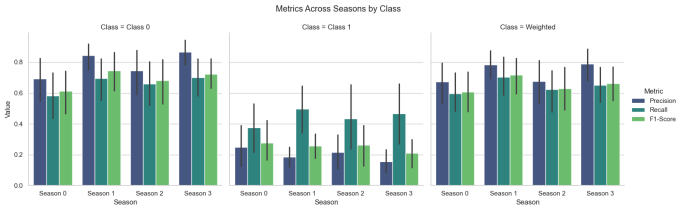


Figure 17. Mètriques models múltiples

Com és d'esperar, les mètriques per la `Classe_0` són millors que les de la `Classe_1`, doncs és la més present en el dataset i el model tendeix a predir-la més. S'han fet moltes, moltes proves, però per aquesta direcció no he pogut dur-lo més enllà.

Era inevitable comprovar l'efecte dels trets visuals en aquesta última fase del projecte. Si bé s'han anat incloent en els apartats anteriors perquè donaven certa ajuda als models, en aquesta última aproximació podem comprovar que la seva aportació és nul·la. Fet que ens porta a pensar que, com s'ha anat veient, la correlació entre les features visuals i les vendes dels articles no l'estem podent capturar, si és que hi és.

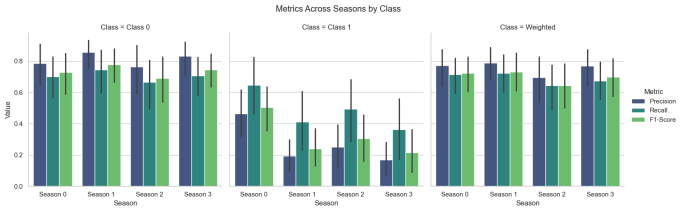


Figure 18. Mètriques sense imatges

5.3.2. End to End

Analitzant els resultats, s'ha vist la següent direcció de millora. Com s'ha comentat anteriorment, certes particions només presenten les dades en una classe a train i/o test. Fins ara estàvem obviant-les en computar les mètriques. Aquesta aproximació no és realista en un context **end2end**, doncs tindríem un desenvolupament perfecte en aquests casos. A part, estem desaprofitant la raó principal per la qual hem fet aquestes particions, que és aprofitar les distribucions de cada una de les particions. Tenint-ho en compte, les mètriques milloren significativament, com es pot veure:

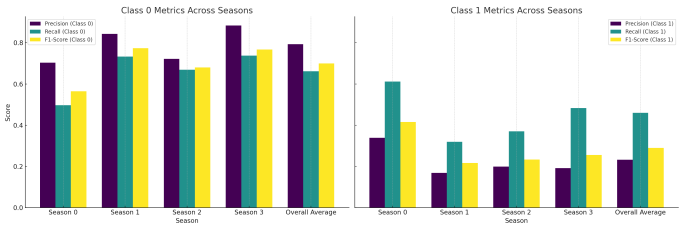


Figure 19. Mètriques end to end

5.4. Models Bagging

Com a últim intent, s'ha provat l'efecte del **Bagging** sobre els models anteriors de la següent manera:

1. Guardem tots els N models de les **particions anteriors**.
2. Obtenim N prediccions de **tot** el dataset de test.
3. Calculem la mitjana entre els N predictors.
4. Binaritzem la sortida, per cert `Treshold`.

D'aquesta manera, hem obtingut els següents resultats:

- Precisió|Recall classe 0: 0.81|0.83
- Precisió|Recall classe 1: 0.23|0.21

Que no milloren massa una predicció de simplement la proporció de classes inicial 80/20.

6. Conclusions

L'objectiu inicial del projecte era realitzar prediccions sobre les tendències de vendes d'articles de moda amb estratègies tradicionals d'aprenentatge computacional. Amb una aproximació inicial per regressió, doncs presentaria una major precisió en les prediccions. Ha resultat una primera tasca impossible, doncs donada la distribució condensada de les dades no s'ha pogut regularitzar.

L'aproximació de **binarització** de la sortida ha donat resultats més prometedors. Cosa que és molt comprensible, doncs s'ha simplificat el model significativament. Així i tot, seguia donant problemes tot i els esforços. És per això que s'ha continuat utilitzant.

Seguint en aquesta línea de simplificació, la partició del problema en estacions i classes ha estat un pas endavant. No només ha millorat lleugerament les mètriques, sino que ha aportat valuosos "insights" sobre la tasca en qüestió. Així i tot, és impossible negar que els resultats **no han estat els esperats**. Especialment centrant-nos en l'efecte, o millor dit en el contraefecte d'afegir les **features visuals** al model.

Amb els extensos resultats del treball, s'ha justificat que **NO** és possible abordar les prediccions de vendes d'articles de roba amb **tècniques tradicionals d'aprenentatge computacional**. Anant més enllà, la correlació entre els trets visuals dels articles de moda i les seves vendes sembla no modelitzable amb aquestes tècniques.

Tot i que amb aquesta aproximació no s'ha pogut arribar més lluny, sembla important recalcar possibles **millores futures**. La natural continuació del treball és, un cop esgotades les eines de **machine learning tradicionals**, mirar per la via del **deep learning**. Sembla adient que, seguint els articles mencionats a l'inici, una estratègia d'**embeddings** multimodals i **transofrmers** podria funcionar. La intuïció ve donada que, gràcies al mecanisme d'**atenció**, aquests models poden aprendre a centrar-se en aquelles features de la imatge més rellevants donades unes característiques categòriques, donant així resultats més esperançadors.

A nivell **personal**, aquest treball ha estat un gran repte. El fet que cap de les proves donés resultats favorables, m'ha obligat a fer experiments de forma exhaustiva per a rebutjar les hipòtesis de forma consistent i no "conformar-me" amb un resultat. Ha estat també molt enriquidor per aquest mateix sentit, doncs, tot i que al final no ha acabat de sortir, m'ha obligat a realitzar moltes hipòtesis per contrastar, modificant els objectius a mesura que anava aprenent del problema i esprémer totes les eines que tinc.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] S. I. Papadopoulos, C. Koutlis, S. Papadopoulos, and I. Kompatsiaris, "Multimodal quasi-autoregression: Forecasting the visual popularity of new fashion products", *arXiv preprint arXiv:2204.04014*, 2022.

[3] G. Skenderi, C. Joppi, M. Denitto, and M. Cristani, "Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends", *arXiv preprint arXiv:2109.09824*, 2024.