Pau de Batlle 242171
Lucia Plasencia 240371

Universitat Pompeu Fabra Barcelona

MACHINE LEARNING

Project 2: Application of Classification in Marketing and Sales

2021-2022

INTRODUCCIÓ

En aquest projecte, hem desenvolupat i aplicat diferents tècniques de classificació supervisada. Aquestes s'utilitzen a les empreses per identificar nous clients al mercat, identificar els clients insatisfets, processar correus electrònics, i moltes més utilitats.

L'objectiu final és identificar un set de no-clients per vendre un producte nou a través d'un classificador de Machine Learning. Durant aquest projecte seguirem el següent procés d'aprenentatge automàtic:

1. Comprensió i preparació de dades: exploració del conjunt de dades i enginyeria de característiques.

2. Model Training: formació de l'SVM de referència i els arbres de decisió. Anàlisi de mètriques (record, precisió, mètriques de confusió) i millora de la classificació mitjançant diverses tècniques com el submostreig per equilibrar o conjunt de models.

3. Creació d'una oportunitat de negoci amb Machine Learning: selecció del millor model i identificació de les característiques més importants.

CONTEXT

Treballem com a cap de ciència de dades i IA en una nova empresa d'Internet de les coses. La nostra empresa crea nous productes sense fil. Els nostres companys de màrqueting fan una nova campanya comercial per captar nous clients.

Hem de decidir quines empreses són l'objectiu a ser visitat pels nostres responsables de vendes. Com que el cost d'enviar un responsable de vendes a visitar un client potencial és força elevat, hem de seleccionar aquelles empreses que tinguin més probabilitats de comprar algun dels nostres productes i convertir-se en un nou client.

Per donar suport a la campanya, utilitzarem diverses tècniques de classificació. Treballarem amb termes com recordatori, precisió, falsos positius, etc... per decidir quins són els millors clients potencials.

Pau de Batlle 242171
Lucia Plasencia 240371

COMPRENSIÓ I PREPARACIÓ DE LES DADES

Exercici 1: Let's identify the type of the variables (integer, float, chart...) and the size of the dataset and the file. Which are the variables with more nulls? And with no nulls? Why 'City' variables is considered as object type? Perquè Python considera del tipus objecte les variables que siguin strings, com les variables 'City' .

Les variables amb més nulls son:

Revenue	4746
CNT_CB_DENSITY	3070
CNT_CB_MOB_DENSITY	3070
CNT_CB_FN_DENSITY	3070

Les variables amb més no nulls son:

City	13335
Customer_Flag	13335
CNT_EMPLOYEE	13335
Mobile_potential	13335

Exercici 3: Which are the main differences between customer_dt and noncustomer_dt datasets comparing these variables. Which is the dataset with CNT_EMPLOYEE higher? Which datasets have more outliers in Revenues? Which is the Q1, median and Q3 for Revenuesand Mobile_potential?

Les diferències més importants entre customers and non customers són que els customers tenen uns ingressos més alts, més quantitat de treballadors, més mobile potencial i més companyies a prop que els non customers.

El dataset amb CNT_EMPLOYEE més alt és el de customers.

Veiem els outliers a Revenue dels dos datasets:

Customer Revenue Q1: 1047500.0
Customer Revenue Median: 2200000.0
Customer Revenue Q3: 4195000.0

Non Customer Revenue Q1: 902986.0

Non Customer Revenue Median: 1750000.0
Non Customer Revenue Q3: 3501123.5

Outliers in Customer Revenue: 278
Outliers in Non Customer Revenue: 1469

Com podem veure non customers té 5 vegades més outliers que customers en revenue.

Customer Mobile_potential Q1: 1621.0556861314792
Customer Mobile_potential Median: 1948.4376609395997
Customer Mobile_potential Q3: 2116.47407406108

Non Customer Mobile_potential Q1: 1513.3835966627996
Non Customer Mobile_potential Median: 1797.0542780039395
Non Customer Mobile_potential Q3: 2035.08284022468

Exercici 5: Calculate the ratio of the values of City for customer_dt and noncustomer_dt datasets. Compare the ratio of each category of each dataset.

ratio customer city: 0.6119221411192214
ratio non customer city: 0.4365818332298779

Aquest ratio ens indica que hi ha més ciutats per persona en customer que en non customer

Exercici 6: Calculate the length of X_train and X_test datasets. Is it aligned with the test_size value selected in the split?

X_train size: 34092
X_test size: 16803
test_size: 0.33

Total_size = X_train + X_test = 50895 → Total_size * 0.33 = 16795 ≈ X_test (aligned)

Exercici 7: Draw the histograms of y_trainand y_test. Is the dataset balanced (similar number of rows for each class or Target) or imbalanced ? How do you think it could affect to quality of the classifier?

Com podem veure en els histogrames hi ha molts més non customers que customers, per tant l'impacte que pot causar és que podrà predir millor quins no seran customers ja que té molts més inputs, és a dir, més dades per entrenar.

Pau de Batlle 242171
Lucia Plasencia 240371

MODEL TRAINING

Exercici 12: Train a SVM and Decision Tree algorithm with the new final_dataset. Evaluate the recall, precision and confusion matrix of all 2 models. Which has better accuracy? Which is the model with better recall? Which model do you recommend to classify both classes? Justify your answer.

L'algoritme que té més accuracy i recall a les dues parts (train i test) és el decision tree, per tant és el que triariem per classificar les dues classes, ja que ens dóna millors resultats que l'SVM.

Exercici 13: Build a voting ensemble formed by a SVM and Decision Tree and train it with the balanced training dataset. Calculate the precision, recall and confusion matrix of the new classifier. Is it better than any of the previous baseline models? Justify your answer.

No és millor que cap dels altres models, ja que la precisió i el recall tenen valors més baixos, i també l'accuracy, per tant no ens fa una classificació tan bona.

Exercici 14: Build a Bagging ensemble based on Random Forest. Random Forest is considered a bagging ensemble formed by Decision Trees algorithms. Train the Random Forest with the balanced training dataset, i.e. X_train and y_train. Calculate the precision, recall and confusion matrix of the new classifier. Is it better than any of the previous baseline models? Justify your answer

.

És millor que tots els altres models, ja que la precisió i el recall tenen valors molt bons a les dues parts (train i test) , i la seva accuracy és la més alta fins ara, per tant ens fa una classificació molt bona.

Exercici 16: Build a Boosting ensemble based on Gradient Tree Boosting (GBT). There are several boosting algorithms as Adaboost, etc. Train the GBT with the balanced training dataset, i.e. X_train and y_train. Calculate the precision, recall and confusion matrix of the new classifier. Is it better than any of the previous baseline models? Justify your answer.

No és millor que cap dels altres models, ja que la precisió i el recall tenen valors baixos, sobretot a la part de train. Encara i així, el valor de l'accuracy al test és força bo.

Exercici 17: Plot the histograms of the probabilities resulting of the prediction of the GBT model for class 0 and class 1. Compare it with histogram of Random Forest. Which one classifies better from your point of view? Why?

Podem concloure que GBT ens predeix millor ja que hi ha més distinció entre les dues classes. Si mirem el random forest veiem que hi ha molt solapament entre les dues variables i en canvi en el GBT estan més desplaçades cap a l'esquerra i a la dreta.

Pau de Batlle 242171
Lucia Plasencia 240371

OPORTUNITAT DE NEGOCI AMB MACHINE LEARNING

Exercici 18: Execute the prediction for the selected model. Adjust the cutoff value to optimize the classifier if you consider necessary. How many non customers are you going to send to the sales managers to sell our products to them?

En el nostre class el que volem és que el model ens classifiqui bé la majoria de customers per poder vendre els productes. Això també ens afavoreix ja que clasificarem non customers com a customers ja que segurament tindran moltes característiques en comú i els hi podrà interessar comprar.

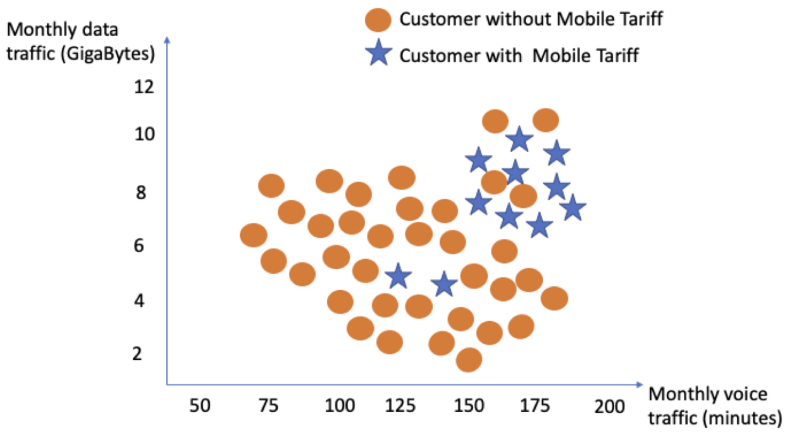
Exercici 19: Order the features by importance. Which are the top 3 features to discriminate between non customers and customers?

Fent servir el model de GBT les classes que tenen més importància al classificar entre customer i non customer son les següents:

1. ['Mobile_potential', 0.15363875540742541]
2. ['CNT_CB_MOB_DENSITY', 0.15098567904433854]
3. ['City_coded', 0.145344428042884]

Exercici 20: In this project, we have used classification techniques to identify potential customers. We have adjusted the main classification parameters as cutoff, recall and precision according to the final purpose: in out case, identify non customers that could be interested to buy our products. Consider a new campaign focused on accelerating the sales of an existing mobile tariff to our customers. Answer the following questions:

In this case, which is the target variable? Which are target=0 samples? And target=1?
As the marketing campaing is oriented to our customers we will have further information about them in our internal systems. In particular, we could add to the information of the previous section 3 new variables: data and voice consumption and mobile expense. Adding more data to the dataset may imply more computational data and cost. Would you add these 3 new variables to dataset? Justify your answer.
Today the mobile tariff is not very popular among our customers. Will the training dataset be balanced or unbalanced? Justify your answer.



En aquest cas el que volem vendre és una tarifa de mòbil als nostres clients i el que ens interessa és trobar tots aquells clients més probables a comprar-la. Això es pot fer mirant quins clients que no tenen la tarifa (0) i que s'assemblen més als clients que si la tenen (1), per tant un altre cop volem classificar tots els clients amb tarifa correctes i de pas identificar tots aquells clients sense tarifa que entren dins aquesta classificació ja que seran els més probables a comprar-la.

Sí que estaria bé afegir aquestes 3 noves columnes ja que ens aporten informació directa que pot augmentar la precisió del model encara que augmenti una mica més el temps de computació i complicació del model.

El dataset clarament no serà balancejat ja que hi ha molts més clients que no tenen la tarifa que clients que si que la tenen.

Describe in terms of monthly data traffic and monthly voice traffic the pattern of target 1 customers
Draw a plane to separate both classes
According to the previous plane, which are the customers to be phoned to sell the mobile tariff?
Could you estimate the precision and recall of the classification?

Aquells clients que sí que tenen la tarifa o target 1, són els que més minuts i dades fan servir a excepció de dues persones.

Si ho fem per Decision Tree tots aquells clients amb més de 125 minuts de veu i més de 4 GB de dades són possibles clients a qui hem de trucar. El Hiperplà seria una recta que talla en els 150 minuts i 12GB i els clients a trucar son tots aquells per sobre de la recta.

Per calcular la precisió i el recall hauriem de fer la taula de confusió i mirar quins s'han classificat bé i quins no.

We hereby declare that, except for the code provided by the course instructors, all of our code, report, and figures were produced by ourselves.