ANALYSIS OF CREDIT APPROVAL DATA PROJECT

By Susana Navarro

Descripció del Projecte Acadèmic

Tasques del projecte acadèmic:

1. Presentació del conjunt de dades escollit: explicació i observacions generals del conjunt de dades a utilitzar per a fer el projecte final.
2. Característiques generals: sense necessitat d'entrar molt en detall, explicar les principals característiques que defineixen aquest conjunt de dades. Tipologia, sector, tipus de dades, font, context, etc.
3. Definició de les variables: explicació teòrica de les principals variables que conté el conjunt de dades.
4. Presentació dels objectius: detallar els objectius inicials marcats de cara a extreure informació rellevant del conjunt de dades.

El propòsit d'aquest projecte és aplicar les tècniques analítiques i de predicció apreses durant el curs/bootcamp de IT_Academy en l'Itinerari de Data Science. Per aquest propòsit utilitzarem el conjunt de dades que correspón a sol.licituds de targetes de crèdit. Les dades estàn disponibles en el repositori de UCI Machine Learning:
https://archive.ics.uci.edu/ml/datasets/credit+approval

Descripció del Projecte Tècnic

1. Abstracte

El risc de crèdit en els consells d'administració dels bancs gira bàsicament al voltant de la determinació de la probabilitat d'incompliment o la solvència d'un client i el seu cost, en el cas que succeeixi. Actualment, amb el creixent del nombre de noves sol·licituds de targetes de crèdit i l'augment d'incompliments de pagament de factures durant la recent pandemia, els bancs estàn exigint un anàlisi del risc més acurat i una millora en l'eficiència de la seva gestió per tal de mitigar el risc creditici. És important tenir en compte els factors clau i anticipar la probabilitat d'incompliment del consumidor ya que pot afectar al flux de caixa i conduir a l'acumulació de retards en el balanç la qual cosa tindria un efecte exponencial en la pèrdua d'actius si el banc és una organització que cotitza en borsa.

Aquí, és on entren en joc els models d'aprenentatge automàtic, ja que permet als bancs i les grans institucions financeres a predir si els seus clients pagarán o no els seus préstecs, per tal mantenir un sistema financer estable i saludable.

L'anàlisi de crèdit ha d'identificar i avaluar els factors que poden afectar als clients en la devolució del seu crèdit. És a dir, seleccionar les variables significatives que expliquin millor el mòdel predictiu que finalment, es traslladarà a la resta de la població.

2. Antecedents

En l'actualitat, els bancs utilitzen classificadors construïts a través de xarxes de Bayes i KNN en el pronòstic de riscos. Les targetes de calificació FICO són una estratègia típica de control de riscos en la indústria monetària que utilitzen dades personals dels clients e informacions històriques per a preveure la probabilitat de futurs incompliments i préstecs.

En definitiva, existeix una gran diversitat de tècniques existents que incorporen anàlisis estadístics, eines de mineria de dades o intel·ligència artificial amb Machine Learning. Una de les tècnica més clàssiques en tècniques de credit scoring és la regressió logística, que ofereix bons resultats estadístics. Un altre enfocament clàssic és sintetitzar la informació de la base de dades de clients a través de regles i d'arbres de decisió, i finalment, altres aproximacions més noves en els models que es basen en l'aplicació de xarxes neuronals, implementant algorismes evolutius, splines de regressió adaptativa, màquines de vectors de suport (SVM) o lògica borrosa.

Per obtenir una xarxa bayesiana s'ha d'especificar una estructura gràfica i una funció de probabilitat conjunta que ve especificada pel producte de la probabilitats de cada node donats els seus pares, cosa que implica que en la majoria de les ocasions no es coneixen ni l'estructura ni les probabilitats. Aquesta és la raó per la qual s'han desenvolupat diferents mètodes d'aprenentatge per obtenir la xarxa bayesiana quan les dades són conegudes.

Per més detalls consultar aquí

3. Introducció

Les tècniques de ML s'utilitzen per a avaluar el risc de crèdit i automatitzar la qualificació creditícia en predir correctament l'elegibilitat del client utilitzant dades demogràfiques del client i dades transaccionals històriques. A més, ML ajuda els bancs a prendre decisions automatitzades amb menys posibilitats d'error, basades en dades dels solicitants. A més, d'utilitzar les dades bancàries d'una manera més productiva i eficient, agilitza la interacció amb el client mitjançant l'eliminació de processos manuals i feixucs.

4. Objetiu

Predir l'aprovació de targetes de crèdit utilitzant el classificador Logístic, XGboost i Random Forest per augmentar la precisió mitjançant la millora en la qualitat de les dades ja que és una de les principals raons dels resultats del model.

5. Problemàtica

Investigadors han realitzat aplicacions d'aprenentatge automàtic sobre qualificació creditícia i prediccions d'incompliment de clients. Els investigadors han conclòs que el mètode SVM (màquina de vectors de suport) i ANN (Xarxa Neuronal Artificial) van funcionar millor que altres classificadors. No obstant això, és important estudiar altres algorismes i comprobar com es comporten amb el conjunt de dades donat.

En termes generals, les targetes de puntuació de crèdit es basen en dades històriques. Una vegada que es troba amb grans fluctuacions econòmiques. Els models anteriors poden perdre el seu poder predictiu original. El model logístic és un mètode comú per a la qualificació creditícia. Perquè el model Logistic és adequat per a tasques de classificació binària i pot calcular els coeficients de cada característica.

Examinarem tres mètodes e identificarem el millor algorisme de classificació per a predir l'elegibilitat del client per la solicitud de la targeta de crèdit i minimitzar la possible pèrdua del crèdit.

6. Metodología

Primer, començarem carregant i veient el conjunt de dades. Veurem que el conjunt de dades té una mescla de característiques numèriques i no numèriques, que conté valors de diferents rangs, a més de que conté una sèrie d'entrades que manca.Tindrem que pre-processar el conjunt de dades per a garantir que el model d'aprenentatge automàtic que triem pugui fer bones prediccions. Després que les nostres dades estiguin en bona forma, farem una anàlisi exploratòria de dades per a construir les nostres intuïcions. Finalment, construirem un model d'aprenentatge automàtic que pugui predir si s'acceptarà la sol·licitud d'una targeta de crèdit d'un individu.

Les tècniques inclouen visualització de dades, regles d'associació i classificació, i predicció amb tècniques de regressió logística, arbres de decisió i XGBoots.

Aquest projecte d'anàlisi està organitzat de la següent manera:

- Generi diverses visualitzacions de dades per a comprendre les dades subjacents
- Realitzar transformacions de dades segons sigui necessari
- Desenvolupar preguntes de recerca sobre les dades
- Generar i aplicar el model per a respondre a les preguntes de recerca
- Evaluació y comparativa dels models de predicció

6.1. Base de dades

El conjunt de dades d'aprovació de crèdit consta de 690 files, que representen a 690 persones que sol·liciten una targeta de crèdit i 16 variables en total. Les primeres 15 variables representen diversos atributs de l'individu com el gènere, edat, estat civil, anys d'ocupació, etc. La variable 16 és la d'interès o la dependent: crèdit aprovat (o no). Conté el resultat de la sol·licitud, ja sigui positiu (representat per "+") que significa aprovat o negatiu (representat per "-") que significa rebutjat. Aquest conjunt de dades és un conjunt de dades de variables múltiples, que conté dades contínues, nominals i categòrics juntament amb missing values.

A tenir en compte, la mida de la mostra, ja que juga un paper determinant en la bondat dels models de classificació. Els mètodes de classificació afavoreixen en general la classe majoritària excepte en el cas del classificador bayesià Naïves Bayes, que classifica millor la classe minoritària.

6.2. Definició de les variables i processament

Aquest conjunt de dades és interessant perquè hi ha un cunjunt d'atributs de tipus continu, nominal i categòrics.
Característiques del conjunt de dades: multivariant
Tipus de data dels atributs: Categòric, Enter, Real.

Attribute Information:

A1: b, a.

A2: continuous.

A3: continuous.

A4: u, y, l, t.

A5: g, p, gg.

A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.

A7: v, h, bb, j, n, z, dd, ff, o.

A8: continuous.

A9: t, f.

A10: t, f.

A11: continuous.

A12: t, f.

A13: g, p, s.

A14: continuous.

A15: continuous.

A16: +,- (class attribute)

Definim el nom dels atributs:

'A1':'Gender', 'A2':'Age', 'A3':'Debt', 'A4':'Married', 'A5':'BankCustomer', 'A6':'EducationLevel',

'A7':'Ethnicity', 'A8':'YearsEmployed', 'A9':'PriorDefault', 'A10':'Employed', 'A11':'CreditScore',

'A12':'DriversLicense', 'A13':'Citizen', 'A14':'ZipCode', 'A15':'Income', 'class':'Approval'

El preprocessament de les dades inclou neteja de les dades, integració de dades, transformació de dades, reducció de dades, imputació de missing values, entre altres tasques. A continuació es mostren algunes de les transformacions de dades que es van realitzar en el conjunt de dades d'Aprovació de crèdit abans d'aplicar qualsevol tècnica de EDA.

El conjunt de dades d'aprovació de crèdit conté valors categòrics que es transformen en valors binaris o factors d'1 i 0. Per exemple, la varable 'Aprovat' que té valors de + i – es canvia a 0 i 1 respectivament. 1, per la targeta aprovada. De manera similar, l'ayribut Gènere té valors 'a' canviats a 0 que representen a homes i 'b' canviats a 1. I 'PriorDefault' tots dos tenen valors categòrics 't' i 'f' que es transformen en 1 i 0. 1 com a valor binari es considera sí previamente el sol.licitant ha incorregut en impagaments.

Missing Values constitueixen més del 5% de tot el conjunt de dades. I els valors que falten estan representats per 'nan'. Els convertim en valor numèrics o en object depenent si correspon a una variable categórica o numérica. Imputem el el promig (mean) i el mode, és a dir, el valor més repetit.

Noms de variables: Inicialment, els camps es van nomenar d'A1-A16, però amb l'ajuda d'una certa documentació disponible, se'ls va canviar el nom de manera adequada.

Posteriorment, transformem les variables categòriques en númeriques mitjancant el mètode LabelEncoder(), i estandaritzem les variables en un rang de valors compresos entre 0 i 1.

6.3 Llibreries d'anàlisis

Pandas, numpy para el càlcul matemàtic, análisis i processamient de les dades

Per visualització: matplotlib, seaborn

Scipy, per a optimització, àlgebra lineal, integració, interpolació, funcions especials.

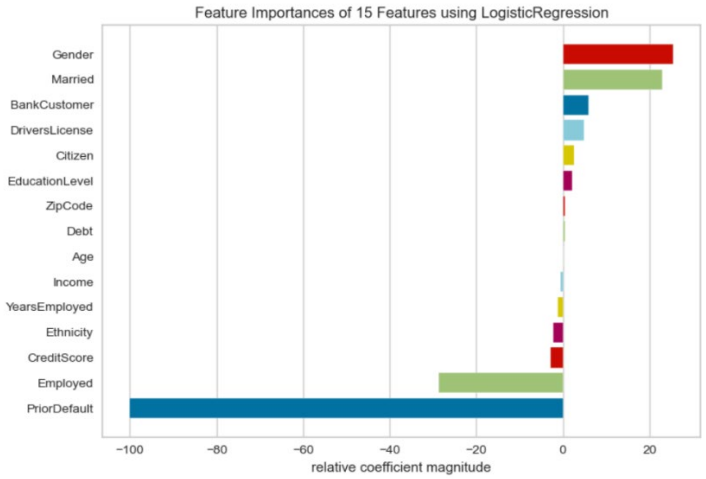
Sklearn, conté mòduls especialitzats per tècniques de ML

Xgboost biblioteca per aplicar el model de clasificació XGBoost

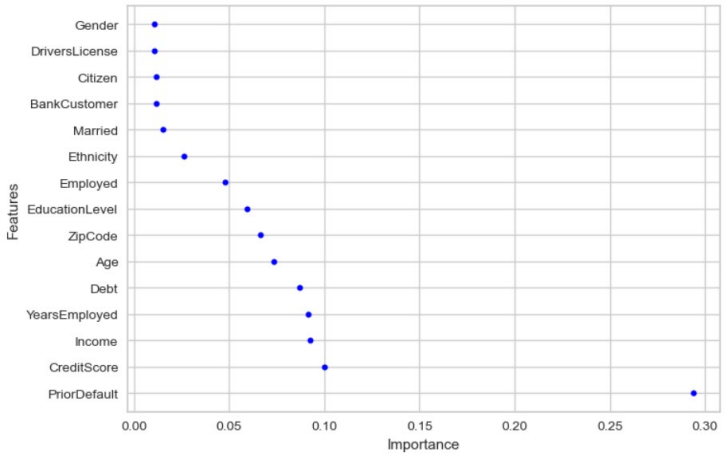
7. Resultats

7.1. Relevancia de les variables socioeconòmiques

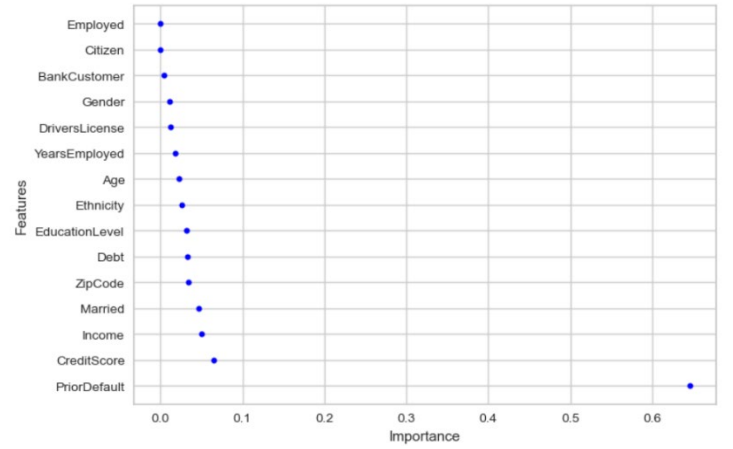
Mitjançant la Informació del valors (IV) + WOE + feature_importances() arrivem a la conclusión que les variables més relevants o significatives per obtenir un resultat més precís en la predicció segóns el model aplica tés el següent:




```
plot_importance(randFC, X, 20)
```




```
plot_importance(XGB, X, 20)
```



En els tres casos, coincideis en l'atribut 'PriorDefault', amb diferencia el més significatiu, li següeix en importancia, CreditScore, Income, Debt, Employed, YearsEmployed, EdicationLevel pero en menor o major grau depenent del model de predicció, com podem veure en els gràfics anteriors.

7.2. Predicció

Amb el mètode LazyPredict ens permet llançar diferents models de scikit-Learn amb els paràmetres per defecte per a comprovar el rendiment de cadascun d'ells en el nostre dataset i enfocar-nos a optimitzar aquells que millor s'ajusten.

El resultat és el seguent:

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RandomForestClassifier	0.86	0.86	0.86	0.86	0.21
NuSVC	0.85	0.85	0.85	0.85	0.03
ExtraTreesClassifier	0.85	0.84	0.84	0.85	0.13
XGBClassifier	0.84	0.84	0.84	0.84	0.15
LogisticRegression	0.84	0.84	0.84	0.84	0.02
LGBMClassifier	0.84	0.84	0.84	0.84	0.09
Perceptron	0.83	0.84	0.84	0.83	0.02
LinearSVC	0.83	0.84	0.84	0.83	0.03
SVC	0.84	0.84	0.84	0.84	0.04

Després d'observar el resultat, procedim a construir els models de predicció que millor s'ajusten:

Model	**Accurancy**
LogísticRegression	0.826 | 0.83
RandomForestClassifier	0.859
XGBClassifier	0.840

8. Conclusió

Podem afirmar que els models de predicció són acertats que podem millorar el nivel de precisió si optimitzem els paràmetres de cada model, com em comprovat amb el model Logistic Regression. El que millor s'ajusta és el model de Random Forest Classifier.

Referent als atributs, hem observant que depen del model predictiu a aplicar podem variar el conjunt d'atributs amb major relevancia.