Aprobarea creditului: o problemă de clasificare

Duma Amalia Diana

Inginerie Software

Rezumat

Acest raport examinează problema aprobării creditului prin utilizarea unor modele de clasificare binară. Au fost implementate două modele principale: regresia logistică și pădurea aleatorie, ambele optimizate pentru performanță utilizând GridSearchCV. Setul de date utilizat conține caracteristici demografice și financiare ale solicitanților, permițând o analiză detaliată a factorilor influenți. Procesul de preprocesare a inclus transformarea atributelor categorice în numerice și scalarea atributelor relevante. Evaluarea importanței caracteristicilor a evidențiat că istoricul de neplată și anii de angajare sunt factorii cei mai influenți în decizia de aprobare a cardului. Modelul de pădure aleatorie a obținut o acuratețe de 87.92%, demonstrând o performanță superioară față de regresia logistică și modelele similare analizate. Abordarea combinată a acestor modele, alături de tehnici avansate de optimizare, a oferit o soluție robustă și interpretabilă pentru problema aprobării creditului, subliniind importanța unor prelucrări și optimizări riguroase în îmbunătățirea preciziei predicțiilor.

Cuprins

Capitolul 1

Introducere

Problema abordată în acest raport este "Credit card approval" (Aprobarea cardurilor de credit), o problemă de clasificare binară în care scopul este de a determina dacă un solicitant de card de credit ar trebui aprobat sau nu pe baza unor caracteristici demografice și financiare. Având în vedere importanța și complexitatea acestei probleme, au fost propuse și utilizate mai multe abordări în literatura de specialitate pentru a îmbunătăți acuratețea și eficiența predicțiilor.

Pentru acest proiect, au fost implementate două modele de clasificare: regresia logistică și pădurea aleatorie. Regresia logistică a fost aleasă datorită simplității și interpretabilității sale, fiind un model de bază foarte utilizat în probleme de clasificare binară. Pădurea aleatorie a fost selectată datorită capacității sale de a gestiona date complexe și de a oferi informații valoroase despre importanța caracteristicilor. În plus, pentru modelul de pădure aleatorie, a fost utilizat GridSearchCV pentru a explora și identifica cei mai buni hiperparametri, optimizând astfel performanța modelului.

În cadrul studiului, s-a analizat o implementare similară disponibilă pe Kaggle [2], care a utilizat regresia logistică, oferind un punct de referință pentru compararea performanței modelului dezvoltat. De asemenea, s-a studiat un articol științific intitulat "Credit Card Approval Verification Model" [3], care prezintă utilizarea regresiei logistice și a modelului de tip pădure aleatoare pentru problema aprobării cardurilor de credit.

Capitolul 2 descrie setul de date "Credit Card Approvals" de pe Kaggle, cu 690 de înregistrări și 16 coloane. Majoritatea solicitanților au vârsta între 20 și 40 de ani și scoruri de credit sub 10. Cele mai multe cereri provin din industria "Energy", iar majoritatea solicitanților sunt cetățeni prin naștere. Ratele de aprobare variază semnificativ între grupurile etnice. În capitolul 3 sunt detaliate etapele de prelucrare a datelor, incluzând analiza matricei de corelație și evaluarea importanței caracteristicilor utilizând modelul de pădure aleatorie.

Capitolul 4 descrie procesul de creare a modelelor de regresie logistică și pădure aleatorie, incluzând detalii despre optimizarea hiperparametrilor prin "GridSearchCV", evaluarea performanței acestor modele pe baza metricilor de acuratețe și a altor măsuri de performanță relevante. De asemenea, sunt prezentate și comparate rezultatele acurateței modelelor studiate cu cele obținute în cadrul acestui proiect. Ultimul capitol prezintă concluziile acestui referat.

Capitolul 2

Descrierea și analiza setului de date

Setul de date utilizat în acest proiect se numește ”Credit Card Approvals” și se găsește pe kaggle [1]. Este utilizat pentru a analiza și modela factorii care influențează aprobarea cererilor de carduri de credit, ajutând la dezvoltarea algoritmilor de clasificare și la îmbunătățirea proceselor de decizie în instituțiile financiare. Setul de date conține doar 690 de înregistrări și 16 coloane, printre care:

- Gender - 0 = Feminin | 1 = Masculin; Tip: int64
- Age - Vârsta individului în ani; Tip: float64
- Debt - Datorie restantă (scalată); Tip: float64
- Married - 0 = Singur/Divorțat | 1 = Căsătorit; Tip: int64
- BankCustomer - 0 = Nu are un cont bancar | 1 = Are un cont bancar; Tip: int64
- Industry - Sectorul locului de muncă actual sau cel mai recent; Tip: object
- Ethnicity - Etnie; Tip: object
- YearsEmployed - Ani de angajare; Tip: float64
- PriorDefault - ; Tip: int64
- Employed - 0 = Șomer | 1 = Angajat; Tip: int64
- CreditScore - Scorul de credit (scalat); Tip: int64
- DriversLicense - 0 = Nu are permis | 1 = Are permis; Tip: int64
- Citizen - ByBirth | ByOtherMeans | Temporary; Tip: object
- ZipCode - Cod poștal; Tip: int64
- Income - Venit (scalat); Tip: int64
- Approved - 0 = Nu a fost aprobat | 1 = Aprobat; Tip: int64

Mai jos este prezentată distribuția valorilor atributelor din setul de date. Tabelul 2.1 și 2.2 prezintă o analiză detaliată a atributelor numerice din setul de date. Acesta cuprinde informații statistice relevante, cum ar fi media, deviația standard, valoarea minimă și cea maximă pentru fiecare atribut numeric.

	Gender	Age	Debt	Married	BankCustomer	YearsEmployed	PriorDefault
count	690	690	690	690	690	690	690
mean	0.7	31.51	4.76	0.76	0.76	2.22	0.52
std	0.46	11.86	4.98	0.43	0.43	3.35	0.5
min	0	13.75	0	0	0	0	0
25%	0	22.67	1	1	1	0.17	0
50%	1	28.46	2.75	1	1	1	1
75%	1	37.71	7.21	1	1	2.63	1
max	1	80.25	28	1	1	28.5	1

Tabela 2.1: Distribuția valorilor atributelor numerice I

	Employed	CreditScore	DriversLicense	ZipCode	Income	Approved
count	690	690	690	690	690	690
mean	0.43	2.4	0.46	180.55	1017.39	0.44
std	0.5	4.86	0.5	173.97	5210.10	0.5
min	0	0	0	0	0	0
25%	0	0	0	60	0	0
50%	0	0	0	160	5	0
75%	1	3	1	272	395.5	1
max	1	67	1	2000	100000	1

Tabela 2.2: Distribuția valorilor atributelor numerice II

Pe baza histogramelor 2.2 se observă ca majoritatea solicitanților au vârsta între 20 și 40 de ani, au între 0 și 5 ani de experiență, cea mai mare parte a scorurilor de credit sunt sub 10 și veniturile variază semnificativ, dar majoritatea sunt sub 2000.

Trecând la analiza atributelor categorice, Industry, Ethnicity, Citizen, observăm că, deși există o diversitate de industrii, cele mai multe solicitări provin din industria ”Energy”. Majoritatea solicitanților sunt albi, și sunt cetățeni după dreptul de naștere. Aceste observații se pot deduce din figura 2.3.

De asemenea, conform figurii 2.1, analizând procentele de cereri aprobate și neaprobate, se observă variații semnificative între grupurile etnice. De exemplu, grupurile de etnie ”latino” au înregistrat cea mai mică rată de aprobare de doar 14%, în timp ce grupurile de etnie ”black” au cea mai mare rată de aprobare de 63%.

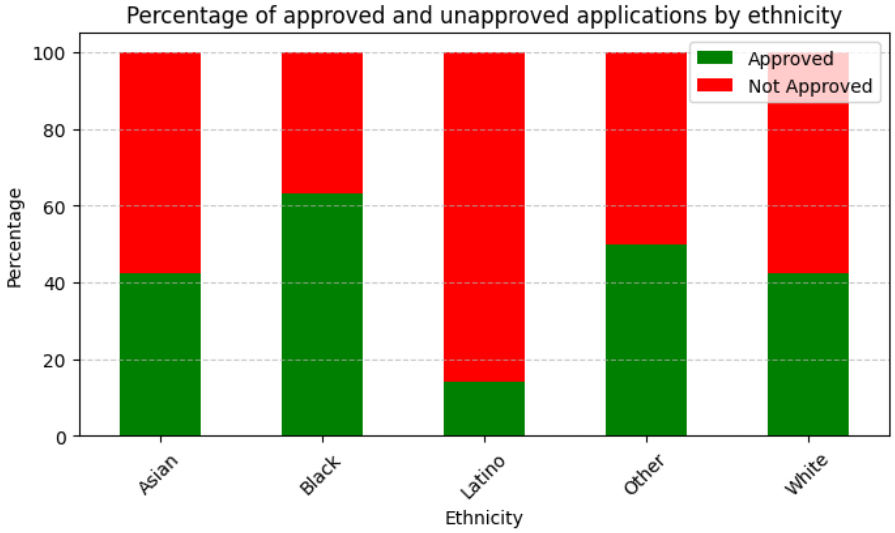

Figura 2.1: Procentul solicitărilor aprobate și ne aprobate per rasă

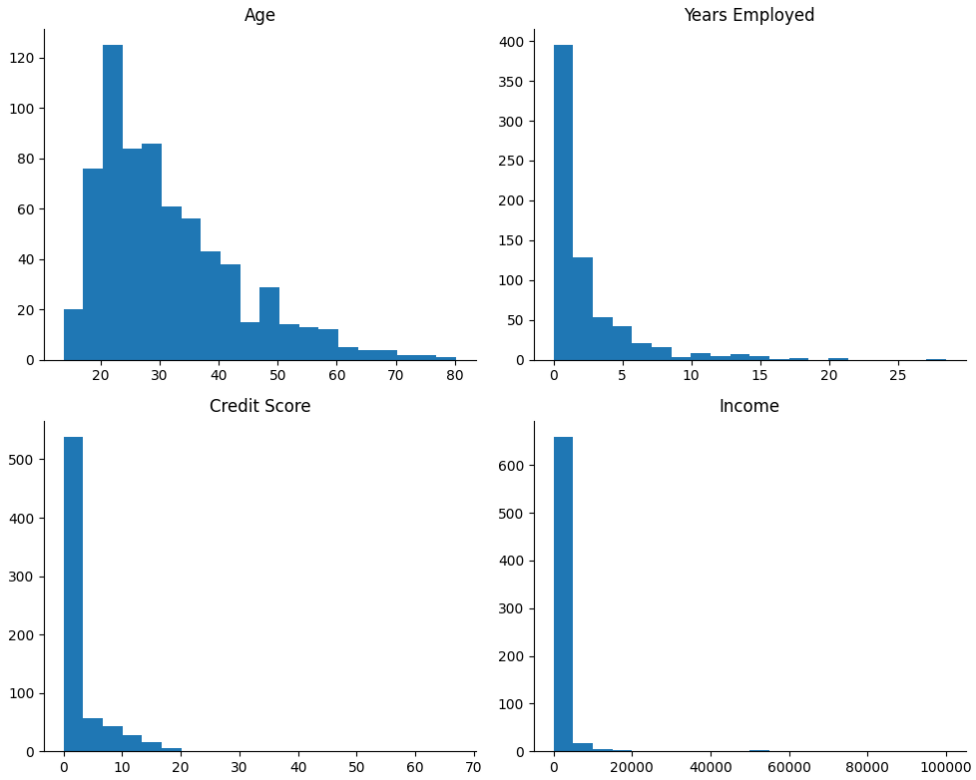

Figura 2.2: Histograme pentru atributele Age, YearsEmployes, CreditScore și Income

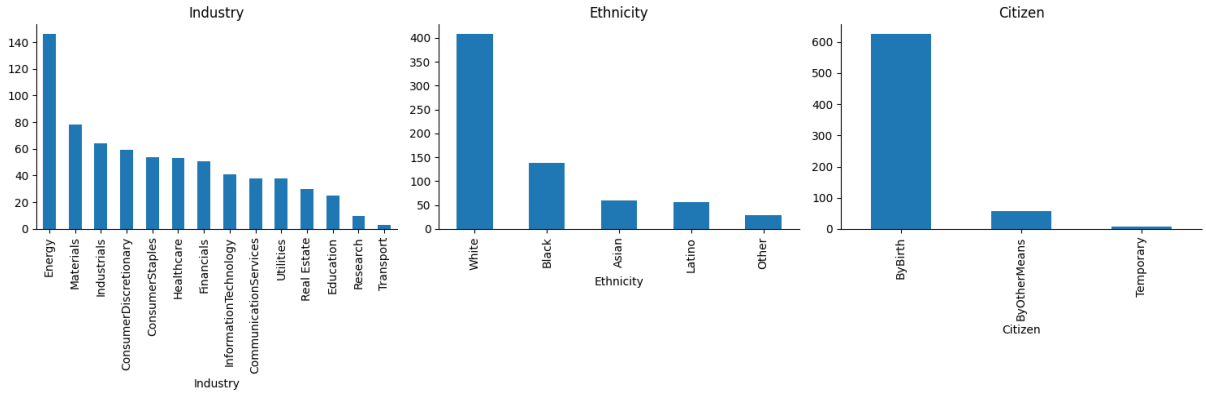

Figura 2.3: Bar plot pentru atributele categorice

Capitolul 3

Prelucrarea datelor

În cadrul procesului de preprocesare a datelor, primul pas a constat în transformarea atributelor categorice în atribute numerice. Acest proces s-a realizat cu ajutorul clasei "LabelEncoder" care atribuie un număr unic fiecărei categorii. Această transformare este esențială pentru analiza matricei de corelație, evaluarea importanței caracteristicilor și pentru construirea modelului de clasificare.

În urma testelor efectuate, s-a decis utilizarea "StandardScaler" pentru scalarea atributului CreditScore, deoarece, deși au fost testate scalările pentru coloanele Age, Debt, YearsEmployed, CreditScore și Income, îmbunătățiri semnificative ale metricilor de performanță au fost observate doar prin scalarea acestui atribut specific. Această scalare transformă valorile astfel încât să aibă o medie de 0 și o deviație standard de 1.

3.1 Analiza matricei de corelație

Analizând matricea de corelație 3.1 putem extrage informații valoroase despre relațiile dintre variabile. Atributul de clasă, Approved, prezintă o corelație puternică cu atributul PriorDefault, având un scor de 0.72, și o corelație moderată cu atributele Employed (0.46) și CreditScore (0.41). Putem deduce că persoanele angajate și cu un scor de credit mai mare sunt mai susceptibile de a fi aprobate. Există o corelație aproape perfectă (0.99) între statutul de căsătorit și faptul de a fi client al băncii, ceea ce sugerează că majoritatea solicitanților căsătoriți sunt și clienți ai băncii. O observație surprinzătoare este că atributul Income are o corelație foarte mică cu celelalte atribute, indicând lipsa unei relații semnificative între venit și acestea.

3.2 Evaluarea importanței caracteristicilor

Pentru evaluarea importanței caracteristicilor s-a ales folosirea unui model de pădure aleatorie, datorită capacității sale de a furniza măsurători precise ale contribuției fiecărei caracteristici la predicțiile făcute. Importanța caracteristicilor este calculată în funcție de cât de mult contribuie fiecare caracteristică la reducerea impurității (de exemplu, indicele Gini sau entropia) pe parcursul creării arborilor de decizie din pădurea aleatorie.

Observând figura 3.2, putem deduce că cea mai mare importanță (aproximativ 29.74%) în decizia de aprobare a creditului o are atributul PriorDefault. Aceasta indică faptul că băncile și instituțiile financiare consideră această informație esențială pentru a determina probabilitatea ca un client să nu-și respecte obligațiile financiare în viitor.

Cu o importanță de 10.30%, numărul de ani de angajare este al doilea cel mai important factor. Aceasta sugerează că stabilitatea în carieră și durata angajării sunt semni-

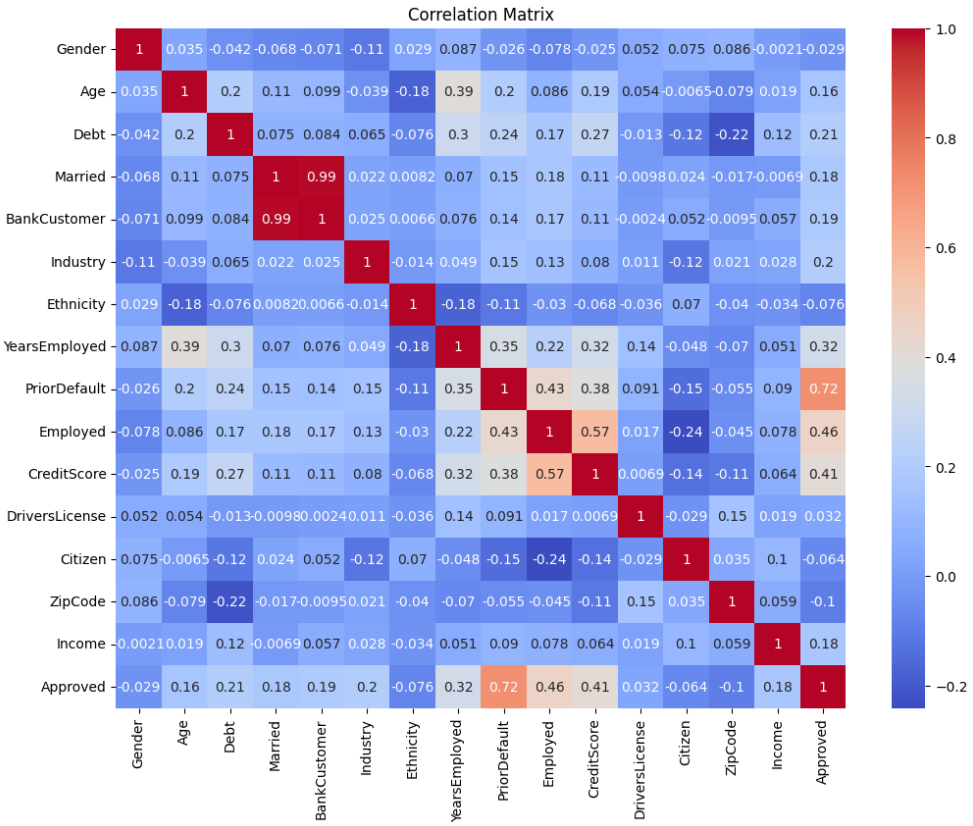

Figura 3.1: Matricea de corelație

ficative în evaluarea capacității unui client de a-și plăti datoriile. O perioadă lungă de angajare indică stabilitate financiară și un venit consistent, factori pozitivi pentru aprobarea unui credit. Scorul de credit, având o importanță de 9.45%, este un alt factor esențial. Un scor de credit mai mare indică un istoric de credit pozitiv și o probabilitate mai mică de neplată. Caracteristici precum ”Citizen”, ”Gender”, ”DriversLicense” , ”Married”, ”BankCustomer” au cele mai mici importanțe de aproximativ 1%.

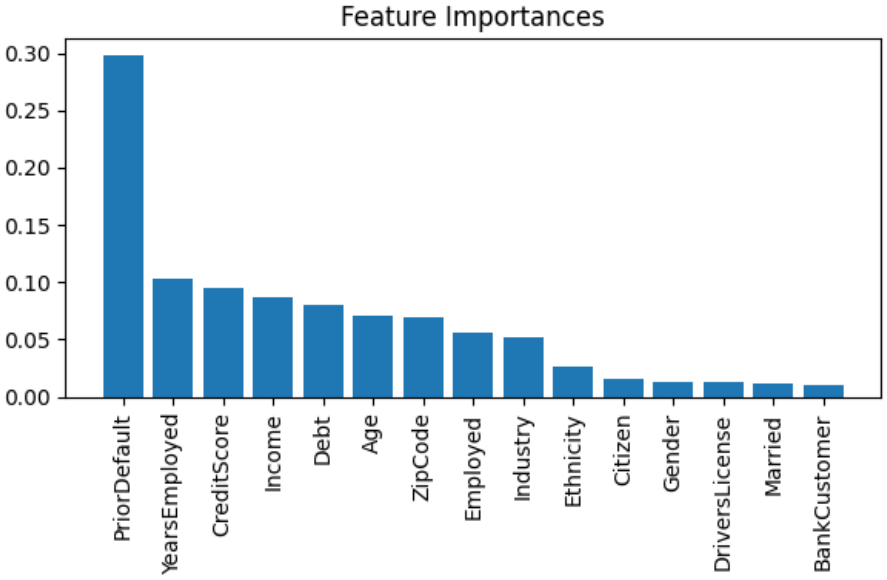

Figura 3.2: Importanța caracteristicilor

Capitolul 4

Crearea și evaluarea modelelor

Pentru această problemă de clasificare, s-a ales testarea unui model de pădure aleatoare. Utilizând parametrii impliciți, s-a obținut o acuratețe de 85.51%. Pentru a optimiza performanța modelului, s-a decis folosirea GridSearchCV, o metodă de căutare exhaustivă a celor mai bune combinații de hiperparametri. GridSearchCV explorează sistematic diferite seturi de parametri și evaluează performanța fiecărei combinații prin validare încrucișată, având scopul de a identifica parametrii care oferă cea mai bună performanță pentru model.

În urma testului, parametrii returnați de GridSearchCV nu au îmbunătățit semnificativ metricile de performanță. Cu toate acestea, bazându-se pe rezultatele obținute și încercând diferite combinații suplimentare de parametri, s-a reușit ajustarea modelului și obținerea unei acurateți de 87.92%. Aceste rezultate, impreună cu parametrii optimizați se pot observa în tabelul 4.1.

	n_estimators	min_samples_split	Accuracy	Precision	Recall	F1 Score
m1	100	2	85.51%	83.17%	86.60%	84.85%
m2	40	5	87.92%	86.00%	88.66%	87.31%

Tabela 4.1: Performanța modelelor tip pădure aleatoare

S-a implementat un model de regresie logistică și s-a obținut o acuratețe de 83.57%, în timp ce un alt model studiat pe Kaggle a obținut 74.20% [2] și un altul găsit într-un articol științific a avut 83.7% [3]. Regresia logistică este un model de clasificare folosit pentru a prezice probabilitatea apariției unui eveniment prin ajustarea datelor la o funcție logistică. Statistica acestor modele se poate observa în tabelul 4.2. M1, M2, M3 reprezintă modelele create, având și coloana "Sursă" goală.

Nume	Tip	Sursa	Accuracy
M1	Pădure aleatoare		85.51%
M2	Pădure aleatoare		87.92%
PaperRF	Pădure aleatoare	Google scholar [3]	86.9%
M3	Regresie logistică		83.57%
KaggleLR	Regresie logistică	Kaggle [2]	74.20%
PaperLR	Regresie logistică	Google scholar [3]	83.7%

Tabela 4.2: Performanța modelelor create și studiate

Capitolul 5

Concluzii

Implementarea modelului de tip pădure aleatoare a permis nu doar realizarea de predicții, ci și evaluarea importanței caracteristicilor, oferind o înțelegere mai profundă a factorilor determinanți în aprobarea cardurilor de credit. Optimizarea hiperparametrilor folosind diferite teste și "GridSearchCV" a dus la îmbunătățirea performanței modelului, deși nu toate combinațiile testate au adus beneficii semnificative. Cu toate acestea, ajustarea fină a parametrilor a permis atingerea unei acurateți de 87.92%, demonstrând importanța procesului iterativ de optimizare.

Printre avantajele abordării propuse se numără: simplitatea și interpretabilitatea regresiei logistice, care facilitează înțelegerea factorilor influenți în deciziile de aprobare, capacitatea modelului pădure aleatoare de a gestiona date complexe și de a evalua importanța caracteristicilor și pptimizarea performanței prin utilizarea GridSearchCV, care explorează diverse seturi de hiperparametri pentru a găsi combinația optimă.

Cu toate acestea, există și câteva dezavantaje și limitări ale abordării propuse. Regresia logistică poate fi limitată de relațiile liniare dintre variabile, fiind mai puțin eficientă în captarea relațiilor complexe. Pădurea aleatorie, deși puternică și flexibilă, poate deveni dificil de interpretat în cazul unui număr mare de arbori și caracteristici. Procesul de optimizare prin GridSearchCV durează și consumă multe resurse, necesitând multiple runde de testare pentru a identifica cei mai buni hiperparametri.

În concluzie, abordarea combinată a regresiei logistice și a modelului Random Forest, alături de tehnici avansate de prelucrare a datelor și optimizare a hiperparametrilor, a demonstrat o performanță robustă în problema de aprobare a cardurilor de credit. Totuși, este esențială continuarea explorării și rafinării acestor metode pentru a îmbunătăți și mai mult precizia și eficiența predicțiilor.

Bibliografie

[1] Samuel Cortinhas. Credit card approvals (clean data). https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data/data, 2022.

[2] Anna Leontjeva. Python for data science). https://www.kaggle.com/code/annitrolla/python-for-data-science#Modelling, 2022.

[3] Umabhanu Tanikella. *Credit Card Approval Verification Model*. PhD thesis, California State University San Marcos, 2020.