Kreditriskmodell med maskininlärning

ECUTBILDNING

Geisol Yissel Urbina

EC utbildning

Fördjupad Python

2025–09

Syftet

Syftet med projektet är att bygga en modell som kan förutsäga om en kund med beviljat lån riskerar att gå i betalningsinställelse (default). Projektet har följ en fullständig maskininlärningspipeline från datagenerering och analys till modellträning, utvärdering och tolkning. Arbetet är utför på syntetiskt genererade data, vilket möjliggör full kontroll och insyn i variabler och beroenden.

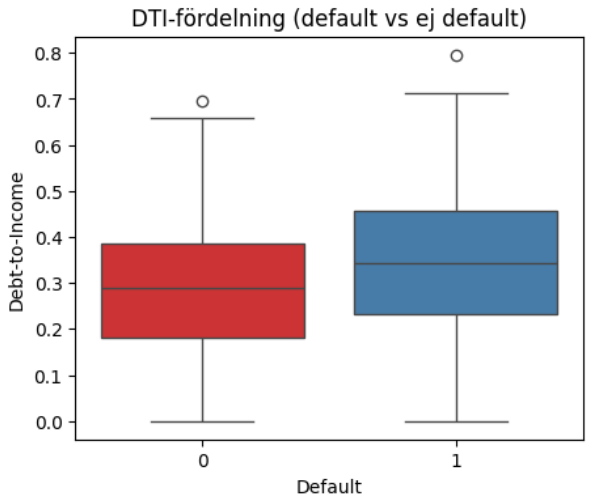
Data och bearbetning

Projektet använder en syntetisk datamängd, skapad specifikt för denna studie med 5000 observationer. Datat innehåller fiktiva men realistiskt konstruerade kundprofiler, inklusive:

- Demografiska faktorer: ålder, inkomst, anställningsstatus och boendeform.
- Finansiella faktorer: kreditpoäng, lånebelopp, DTI (debt-to-income).
- Historik: antal missade betalningar, kundrelationens längd
- Målvariabel: default där 1 = betalningsinställelse, 0= betalar tillbaka.

I datagenereringen implementeras en specifik åldersregel för lånebeviljande, där ingen kund över 74 år tilldelas lån. Denna regel motiveras av att personer i denna åldersgrupp ofta är pensionärer med begränsade ekonomiska resurser. Pensionen räcker i många fall knappt till nödvändiga utgifter som boende, mat och vård, vilket gör risken för betalningsinställelse högre. För att efterlikna realistiska kreditbedömnings principer valdes därför att filtrera bort dessa kunder från modellen. Regeln kontrolleras strikt i koden via en *assert*, som säkerställer att inga över 74 får lån.

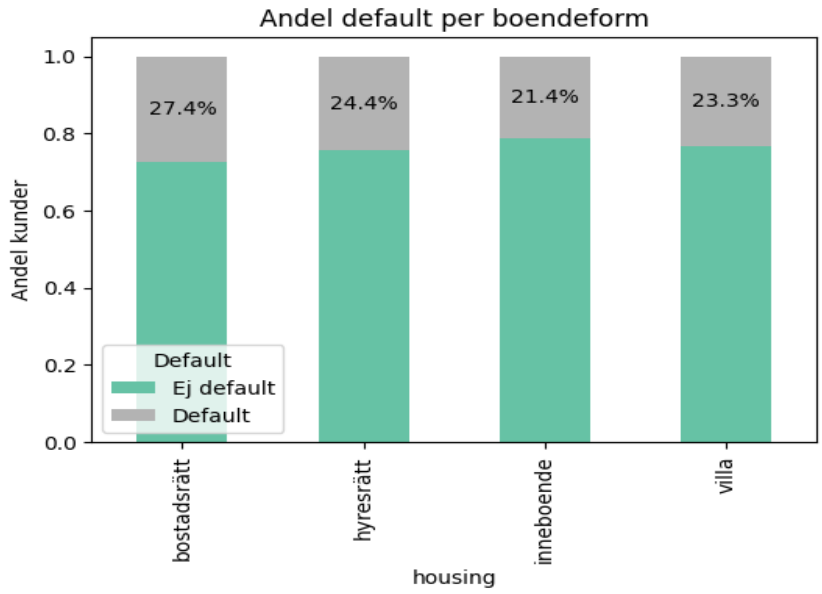
Efter datagenerering utfördes EDA (Exploratory Data Analysis) för att förstå data, identifiera mönster och förbereda inför modellering. Analysis visade hög DTI, många missade betalningar och låg kreditpoäng ökar risken för default. Studenter och pensionärer har högre default-andel, inkomst och kreditpoäng är svagt korrelerade.



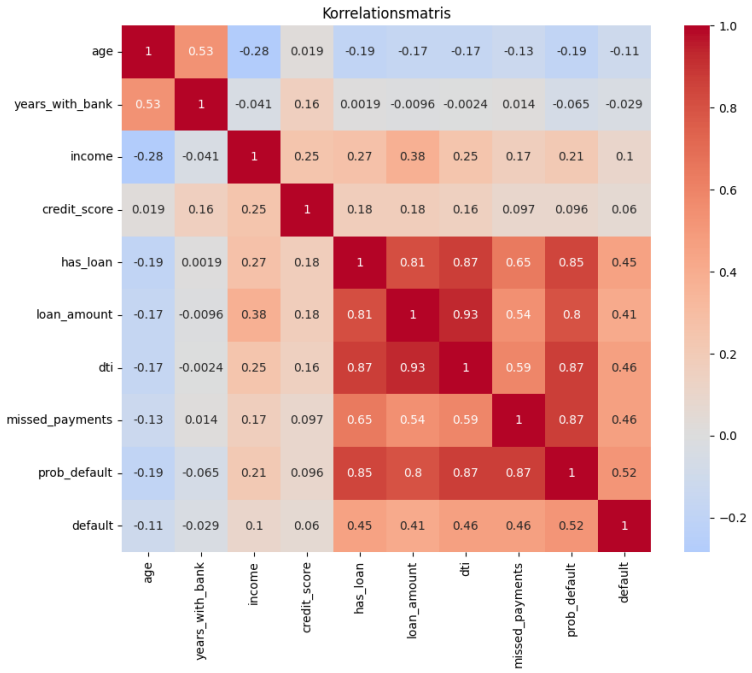
Figur 1. Det visar förhållandet mellan DTI (debt-to-income) och en person som går i default.

Diagrammet visar att det finns en viss skillnad i DTI-fördelning mellan de två grupperna. Gruppen som betalar inte lånet (dvs 1 =default) har en högre genomsnittlig DTI och en större spridning av DTI-värdena. Detta indikerar att ett högre Debt-to-income värde verkar vara associerat med en högre risk att gå i default. Skillnaden är inte stor så det betyder att inte alla personer som går i default har ett högt DTI, och inte heller att alla med ett lågt DTI undviker att gå i default. Däremot är DTI en av de faktorer som kan användas för att bedöma risken.

En analys har även genomförts av varje boendeform. Det visar att bostadsrätt har högsta andelen defaults, med 27.4%. Detta kan bero på att lån för bostadsrätter oftast är stora och långvariga, vilket kan öka risken för default vid ekonomiska svårigheter. Hyresrätt default andelen är 24.4%. Detta är lägre än för bostadsrätter, vilket kan reflektera att lånen som tas av denna grupp kanske är mindre och därmed lättare att hantera. Inneboende-grupp har den lägsta andelen defaults, med 21.4%. Eftersom inneboende ofta har mycket lägre boendekostnader än andra grupper, kan detta innebära att de har mer ekonomiskt utrymme att klara av sina lån. Villa default andelen är 23.3%. Detta är lägre än både hyresrätts- och bostadsrättsgrupperna, men något högre än inneboendegruppen.



Figur 2. Andel default per boendeform.



Figur 3. Korrelationsmatris

Matrisen visar att *age* och *years_with_bank* har en stark positiv korrelation (0.53), vilket är logiskt. Ju äldre du är, desto längre har du sannolikt varit kund hos banken. Det finns också en tydlig positiv korrelation mellan variablerna *has_loan, loan_amount* och *dti.* Exempelvis har *has_loan* och *loan_amount* en korrelation på 0.81, vilket är väntat, om en kund har ett lån är det mycket sannolikt att denne även har ett högt lånebelopp. Vidare ser vi att *has_loan* och *dti* (debt-to-income) har en korrelation på 0.87. Detta tyder på att kunder med lån ofta har en högre skuldsättningsgrad, vilket är logiskt eftersom lånet direkt påverkar skuldsidan. Den starkaste korrelationen i hela matrisen är mellan *loan_amount* och *dti*, med ett värde på 0.93. Detta är förståeligt, eftersom DTI är direkt beroende av lånebeloppet i förhållande till inkomsten, ju större lån, desto högre DTI, givet samma inkomstnivå.

Eftersom så starka korrelationer mellan variabler kan leda till multikollinearitet, är det viktigt att hantera detta innan modellträning. I detta fall valde jag att utesluta *loan_amount* eftersom den är starkt korrelerad med både *has_loan* och *dti,* och *dti* är en mer informativ och sammansatt indikator på skuld i förhållande till inkomst. Detta beslutstöds också av analysen av VIF (Variance Inflation Factor), där *loan_amount* visade hög inflation.

Modellering

Efter att ha analyserat multikollinearitet via både korrelationsmatris och Variance Inflation Factor (VIF), valdes ett urval av prediktorer som visade låg korrelation men hög förklaringskraft i relation till default. De slutliga variablerna som användes i modellen var: numeriska: *age, years_with_bank, income, credit_score, dti, missed_payments*; samt kategoriska: *employment_status* och *housing.*

Tre modeller jämfördes med hjälp av 5-fold stratifierad cross-validation och ROC AUC som utvärderingsmått: Logistic Regression (0,682), Radom Forest (0,620) och XGBoost (0,589). Logistic Regression valdes som slutmodell tack vare både högst prestanda och dess tolkbarhet.

Därefter tränades modellen på hela träningsdatan och förbereddes för kalibrering och tröskeloptimering. När en klassificeringsmodell som logistisk regression tränas för att förutsäga sannolikheter för ett utfall är det avgörande att dessa sannolikheter är välkalibrerade. Det innebär att om modellen bedömer att 100 kunder har 30 % risk att gå i default, så bör ungefär 30 av dem faktiskt göra det. Välkalibrerade sannolikheter gör att modellen kan användas som ett tillförlitligt beslutsstöd i kreditbedömningar. Däremot ger inte standardtröskeln på 0,50 nödvändigtvis den bästa balansen mellan precision och recall. För att hantera detta gjordes en tröskeloptimering, där olika gränsvärden testas och det som maximerar ett valt mått, i det här fallet F1-score, som balanserar precision och recall väljs. Kalibrering förbättrade kvaliteten på modellens sannolikheter, medan tröskeloptimering förbättrade själva klassificeringsbesluten. I kombination leder dessa steg till en mer robust och praktiskt användbar modell, särskilt i situationer där klasserna är obalanserade.

Resultat

Den initiala baslinjemodellen utan kalibrering eller tröskeloptimering gav måttliga resultat. ROC AUC låg på 0,692 vilket indikerar att modellen bara något bättre än slumpen kan skilja mellan kunder som går i default och de som inte gör det. Recall var relativt låg på 26,8 %, vilket innebär att modellen missade en stor andel faktiska defaults, även om precision på 53,6 % innebar att ungefär hälften av de flaggade kunderna faktiskt var riskkunder. Det resulterade i ett F1-score på 0,357. Även sannolikhetskalibreringen var svag, med ett Brier score på 0,173, vilket visar att de predikterade sannolikheterna inte låg nära de faktiska utfallen.

Modell	**ROC AUC**	**Recall**	**Precision**	**F1-score**	**Brier**
Baseline	0.692	26.8%	53.6%	0.357	0.173
Förbättrad LogReg	0.943	56.4%	33.7%	0.422	0.039

Tabell 1. Resultatöversikt

Efter förbättringar i form av feature engineering, ElasticNet-regularisering, kalibrering och tröskeloptimering förbättrades resultaten avsevärt. ROC AUC ökade till 0,943, vilket visar på mycket god förmåga att rangordna kunder efter risknivå. Brier score sjönk kraftigt till 0,039, vilket innebär att de predikterade sannolikheterna låg mycket nära de verkliga utfallen. Recall förbättrades till 56,4 %, vilket betyder att modellen lyckades identifiera en betydligt större andel av riskkunderna, men precisionen sjönk till 33,7 %, vilket återspeglar en ökad andel falska positiva. Detta gav ett något högre F1-score på 0,422, vilket visar att modellen total sett hittade en bättre balans mellan precision och recall.

Sammantaget visar resultaten att förbättringarna kraftigt ökade modellens förmåga att skilja mellan säkra och riskabla kunder samt att ge mer tillförlitliga sannolikheter. Även om precisionen minskade var detta en medveten kompromiss för att höja recall, vilket är rimligt i kreditrisk eftersom det i regel är allvarligare att missa en riskkund än att felaktigt neka en kund som egentligen hade kunnat betala tillbaka.

Slutsats

Analysen visar att även en enkel modell som logistisk regression kan prestera mycket väl när den kombineras med systematiskt arbete kring feature engineering, regularisering, kalibrering och threshold-justering. Mer komplexa modeller som Random Forest och XGBoost kan ge hög prediktiv kraft, men på bekostnad av tolkningsbarhet. En central lärdom är att kalibrering är avgörande när modellen ska användas i faktiska beslut, inte bara för rangordning. Projektets största begränsning är att det bygger på syntetisk data och ännu inte är testat i en produktionsmiljö. För framtiden bör modellerna utvärderas på verkliga eller hybriddata, kompletteras med fler variabler och optimeras mot olika kostnadsstrukturer för felklassificering. Sammantaget visar projektet att en strukturerad metodik kan ge robusta och välkalibrerade kreditriskmodeller som balanserar prediktiv styrka med regulatoriska krav på transparens och tolkbarhet.