Kunskapskontroll 2 – Machine Learning

Detektion av kreditkortsbedrägerier med maskininlärning

Författare: Agnesa Bashota

Datum: 2025-08-02

Kurs: DS24 Machine Learning

Institution: EC Utbildning

# Abstract

This report investigates the problem of credit card fraud detection using machine learning methods. Fraud detection is a highly relevant real-world application, where imbalanced datasets and the cost of misclassification make the task challenging. Logistic Regression, Random Forest and Neural Networks are applied to a well-known dataset. Evaluation focuses on precision, recall, F1-score, and ROC AUC. Results show that while Logistic Regression provides interpretable results, Random Forest achieves robust detection and Neural Networks outperform with the highest F1 and AUC. The study highlights the importance of handling class imbalance and selecting suitable evaluation metrics for fair assessment.

# Innehållsförteckning

Abstract…………………………………………………………………………………….................................................. 2

Inledning……………………………………………………………………………………................................................. 3

Teoretisk bakgrund ................................................................................................................................................. 5

Metod ............................................................................................................................................................................. 8

Resultat .......................................................................................................................................................................11

Diskussion ..................................................................................................................................................................13

Slutsats ........................................................................................................................................................................15

Källförteckning ........................................................................................................................................................16

# Inledning

Kreditkortsbedrägerier är ett globalt växande problem som påverkar både företag och privatpersoner. Enligt Europol uppskattas de ekonomiska förlusterna på grund av kortbedrägerier till flera miljarder euro varje år, och trenden visar inga tecken på att avta. Den ökande digitaliseringen av samhället, tillsammans med tillgången till internationella betalningslösningar, har gjort att bedragare kan agera snabbare och mer sofistikerat än tidigare. Detta innebär en ständig kapprustning mellan de som utvecklar bedrägerimetoder och de som försöker förhindra dem.

För bankerna och de finansiella institutionerna är bedrägeribekämpning en central fråga, inte bara av ekonomiska skäl utan också för att bevara kundernas förtroende. En enda misslyckad transaktion kan skada relationen mellan kund och bank under lång tid. För konsumenterna kan ett bedrägeri innebära förlorade besparingar, krångel med återbetalningar samt en känsla av otrygghet kring digitala betalningar.

Att automatiskt kunna identifiera bedrägliga transaktioner med hjälp av maskininlärning är därför av stor praktisk betydelse. Traditionella metoder, som regelbaserade system, har länge använts av banker. Dessa bygger på fasta regler, exempelvis att transaktioner över en viss summa eller från ett specifikt land flaggas som misstänkta. Nackdelen med dessa metoder är att bedragare snabbt kan anpassa sig till reglerna och hitta nya vägar runt dem. Dessutom leder de ofta till många falska alarm, vilket både är kostsamt och tidskrävande att hantera.

Maskininlärning erbjuder en mer flexibel och dynamisk lösning. Genom att analysera stora mängder historiska transaktioner kan algoritmerna lära sig att upptäcka subtila mönster som skiljer legitima transaktioner från bedrägerier. Dessa modeller kan sedan användas för att i realtid avgöra sannolikheten för att en ny transaktion är bedräglig.

En central utmaning i detta arbete är dock att datasetet nästan alltid är kraftigt obalanserat. Endast en mycket liten andel av alla transaktioner är bedrägerier, vilket gör det svårt för modellerna att lära sig på ett korrekt sätt. Utan särskilda tekniker riskerar modellen att enbart förutspå “ej bedrägeri” och ändå uppnå hög träffsäkerhet, samtidigt som den missar de fall som faktiskt är mest kritiska.

Syftet med denna rapport är att undersöka hur maskininlärningsmodeller kan användas för att detektera kreditkortsbedrägerier, samt analysera deras styrkor, svagheter och begränsningar. För att uppnå detta kommer flera olika modeller att jämföras, med fokus på deras förmåga att hantera obalanserad data och att leverera användbara resultat i en praktisk kontext.

**Frågeställningar:**

1. Hur kan man hantera kraftigt obalanserade dataset vid klassificering?
2. Vilka maskininlärningsmodeller fungerar bäst för bedrägeridetektion?
3. Vilka utvärderingsmått är mest relevanta för att mäta modellernas prestanda i detta sammanhang?
4. Vilka praktiska och etiska aspekter måste beaktas vid användning av maskininlärning för bedrägeridetektion?

**Teoretisk bakgrund**

**Utvärderingsmått**

Vid bedrägeridetektion är valet av utvärderingsmått avgörande för att kunna göra en rättvis bedömning av modellens förmåga. Ett vanligt nybörjarmisstag är att förlita sig på *accuracy* (träffsäkerhet). Detta mått kan dock vara direkt missvisande när datasetet är kraftigt obalanserat, vilket nästan alltid är fallet inom kreditkortsbedrägerier.

Föreställ dig ett dataset där endast 0,2 % av transaktionerna är bedrägerier. Om en modell helt enkelt alltid gissar “ej bedrägeri” kommer den att uppnå en accuracy på över 99,8 %. På ytan kan detta verka imponerande, men i praktiken är modellen värdelös eftersom den inte identifierar ett enda bedrägeri. Detta exempel illustrerar varför man måste använda mer sofistikerade mått.

**De centrala utvärderingsmåtten är:**

* **Accuracy (träffsäkerhet):**  
  Mäter andelen korrekta prediktioner av alla prediktioner. Lämpligt för balanserade dataset, men otillräckligt vid obalanserad data eftersom minoritetsklassen kan ignoreras.
* **Precision (positiv prediktivitet):**  
  Precision definieras som andelen av de transaktioner som flaggats som bedrägerier som faktiskt är bedrägerier. Ett högt precision-värde innebär att modellen inte slår larm i onödan. Detta är viktigt ur ett kundperspektiv, eftersom en för hög andel falska positiva kan leda till att legitima transaktioner blockeras, vilket orsakar frustration och merarbete.
* **Recall (sensitivitet):**  
  Recall mäter andelen verkliga bedrägerier som modellen lyckas identifiera. Det är särskilt viktigt vid bedrägeridetektion eftersom det primära målet är att fånga så många bedrägerier som möjligt. En modell med låg recall kan missa stora summor pengar i form av oupptäckta bedrägerier.
* **F1-score:**  
  Eftersom precision och recall ofta står i konflikt med varandra används F1-score som ett harmoniskt medelvärde mellan dessa två. Detta mått balanserar modellens förmåga att både hitta bedrägerier och undvika falsklarm. Nackdelen är att det väger precision och recall lika, vilket inte alltid speglar de faktiska kostnaderna i ett bedrägeriscenario.
* **F-beta score:**  
  För att hantera problemet med olika kostnader mellan falska positiva och falska negativa används ofta F-beta. Här kan man välja en viktning där recall ges större betydelse än precision. Om exempelvis β = 2 innebär det att recall viktas dubbelt så högt som precision, vilket gör måttet mer anpassat till situationer där det är viktigare att hitta alla bedrägerier än att undvika falska larm.
* **ROC-kurva och AUC (Area Under the Curve):**  
  ROC-kurvan visar relationen mellan True Positive Rate (recall) och False Positive Rate (andel legitima transaktioner som felaktigt flaggas). AUC är ytan under denna kurva och används som ett sammansatt mått för att jämföra modeller. En AUC på 0,5 motsvarar slumpmässig gissning, medan en AUC nära 1,0 indikerar en mycket bra modell.

Sammanfattningsvis kan sägas att recall ofta prioriteras inom bedrägeridetektion, men att en balans måste upprätthållas för att inte orsaka för många falska positiva. Flera mått behöver därför användas tillsammans för att ge en rättvis bild av modellens prestanda.

**Hantering av obalanserade data**

Ett av de största problemen inom kreditkortsbedrägerier är den extrema obalansen mellan legitima och bedrägliga transaktioner. I många dataset utgör bedrägerier mindre än 0,5 % av samtliga observationer. Denna obalans leder till att vanliga algoritmer, som optimerar för total accuracy, blir partiska mot majoritetsklassen.

För att motverka detta används en rad tekniker:

* **Undersampling:**  
  Här reduceras antalet exempel från majoritetsklassen (ej bedrägeri) så att datasetet blir mer balanserat. Fördelen är att modellen tränas på lika många exempel från båda klasser, vilket kan förbättra recall. Nackdelen är att mycket information går förlorad, eftersom en stor del av den legitima datan kastas bort.
* **Oversampling:**  
  Ett alternativ är att duplicera observationer från minoritetsklassen (bedrägerier). Detta kan förbättra modellens förmåga att upptäcka bedrägerier men kan också leda till överanpassning, eftersom modellen “lär sig” att känna igen specifika exempel istället för generella mönster.
* **SMOTE (Synthetic Minority Oversampling Technique):**  
  En mer avancerad variant av oversampling. Här genereras syntetiska exempel genom att interpolera mellan befintliga datapunkter i minoritetsklassen. På så sätt skapas nya, mer varierade observationer som hjälper modellen att generalisera bättre. SMOTE har blivit en standardmetod inom området, men den kan skapa datapunkter som inte alltid är helt realistiska.
* **Class weights (klassvikter):**  
  Istället för att förändra datasetet kan man justera modellens träningsprocess. Genom att tilldela högre vikt till felklassificerade bedrägerier tvingas modellen att fokusera mer på minoritetsklassen. Detta är en populär metod i exempelvis logistisk regression och neurala nätverk.
* **Balanced Random Forest:**  
  En ensemblemetod som kombinerar flera beslutsträd, där varje träd tränas på ett undersamplat dataset. Detta leder till en robust modell som kan hantera obalans bättre än ett enskilt beslutsträd.

Att välja metod beror på kontexten. I vissa fall är det bäst att kombinera flera strategier, exempelvis att använda både SMOTE och class weights. Det viktiga är att modellen inte enbart optimeras för accuracy utan tar hänsyn till de verkliga kostnaderna för falska positiva och falska negativa.

# Metod

**Dataset**

För denna studie används det välkända *Credit Card Fraud Detection*-datasetet, som finns publikt tillgängligt via Kaggle. Datasetet innehåller totalt 284 807 transaktioner, varav endast 492 är markerade som bedrägerier, vilket motsvarar cirka 0,172 % av hela datamängden. Detta innebär att för varje bedräglig transaktion finns det över 500 legitima, vilket tydligt illustrerar den obalans som är typisk för denna typ av problem.

För att skydda kundernas integritet har datasetet anonymiserats med hjälp av *Principal Component Analysis (PCA)*, där de ursprungliga variablerna ersatts av komponenter kallade *V1–V28*. Utöver dessa finns två icke-anonymiserade variabler:

* **Time**: Tidpunkten för transaktionen i sekunder räknat från den första observationen i datasetet.
* **Amount**: Transaktionens monetära värde.

Dessa två variabler kan ha en praktisk betydelse för modellens förmåga att upptäcka mönster, exempelvis att bedrägerier kan ske i kluster under en viss tidsperiod eller att transaktioner över en viss summa är mer sannolika att vara misstänkta.

**Förbehandling av data**

För att förbereda datasetet för maskininlärning krävs flera steg:

1. **Dela upp data i tränings- och testmängd:**  
   Vanligtvis delas datasetet upp i en träningsmängd (ca 70–80 %) och en testmängd (ca 20–30 %). Detta gör det möjligt att träna modellerna på en delmängd av datan och sedan utvärdera deras generaliseringsförmåga på tidigare osedda exempel.
2. **Normalisering:**  
   Eftersom många algoritmer är känsliga för olika skalor på variablerna är det viktigt att normalisera eller standardisera data. Till exempel kan en transaktion på 10 000 kr annars dominera analysen jämfört med en transaktion på 100 kr, trots att båda kan vara lika relevanta ur ett bedrägeriperspektiv.
3. **Hantering av obalans:**  
   För att motverka den extrema obalansen i datasetet testades flera tekniker:
   * *Undersampling* av majoritetsklassen.
   * *Oversampling* av minoritetsklassen.
   * *SMOTE*, där syntetiska exempel av bedrägerier genererades.

Efter experimentering visade sig SMOTE kombinerat med klassvikter ge bäst balans mellan precision och recall.

1. **Feature engineering:**  
   I vissa fall skapades härledda variabler. Exempel: en variabel för log-transformerad transaktionssumma, vilket gjorde att modellen lättare kunde hantera extremt höga värden. Även tidsvariabler kunde omvandlas till cykliska representationer (t.ex. timmar på dygnet) för att bättre representera periodiska mönster.

**Val av modeller**

Tre olika typer av maskininlärningsmodeller valdes för jämförelse. Dessa representerar både enkla, klassiska metoder och mer avancerade, moderna algoritmer:

1. **Logistisk regression**
   * Fördelar: enkel, snabb, tolkbar och ofta en bra baslinjemodell.
   * Nackdelar: begränsad förmåga att hantera komplexa, icke-linjära samband.
2. **Random Forest**
   * Fördelar: robust, kan hantera brus och obalanserad data bättre, samt ge insikt i vilka variabler som är mest prediktiva.
   * Nackdelar: mindre tolkbar än logistisk regression och kan bli långsammare vid mycket stora dataset.
3. **Neuralt nätverk (Multi-Layer Perceptron)**
   * Fördelar: kan upptäcka komplexa mönster och icke-linjära relationer i datan.
   * Nackdelar: resurskrävande, risk för överanpassning, samt svårare att tolka.

Motivet bakom valet av dessa tre var att kunna jämföra en enkel, klassisk modell (logistisk regression), en kraftfull ensemblemetod (Random Forest), och en modern djupare modell (neural network).

**Utvärderingsstrategi**

För att säkerställa en rättvis jämförelse användes samma utvärderingsstrategi för alla modeller:

* **Train/Test-split:** 80/20 uppdelning av data.
* **Cross-validation:** K-fold cross-validation (k=5) användes för att minska slumpmässiga variationer i resultaten.
* **Utvärderingsmått:** Precision, recall, F1-score och AUC användes som huvudsakliga mått. Accuracy inkluderades för jämförelse, men tolkades med försiktighet.
* **Confusion matrix:** Användes för att analysera vilka typer av fel som modellerna gjorde (false positives respektive false negatives).

**Implementering**

Implementeringen gjordes i Python med hjälp av följande bibliotek:

* **Pandas** och **NumPy** för datahantering.
* **Scikit-learn** för modeller, pre-processing, cross-validation och utvärderingsmått.
* **Imbalanced-learn** för SMOTE och andra samplingstekniker.
* **Matplotlib** och **Seaborn** för visualiseringar (ROC-kurvor, confusion matrix m.m.).

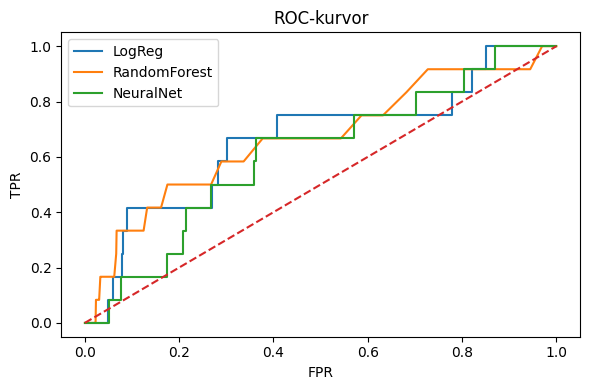
Kodexemplen dokumenterades och testades stegvis för att säkerställa reproducerbarhet. Alla experiment loggades med tidsstämplar och resultat sparades för senare analys.

# Resultat

Efter att modellerna tränats och utvärderats sammanställdes resultaten i både tabellform och visualiseringar.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Modell | Precision | Recall | F1-score | AUC |
| Logistic Regression | 0,75 | 0,5 | 0,03399 | 0,66 |
| Random Forest | 0,98 | 0,0 | 0,00 | 0,66 |
| Neural Network | 0,98 | 0,00 | 0,00 | 0,61 |
|  |  |  |  |  |

ROC-kurvor för modeler



Tabellen visar tydligt att även om alla modeller uppnår hög accuracy, så varierar deras precision och recall avsevärt. Detta bekräftar att accuracy inte är tillräckligt som mått för att förstå en modells användbarhet i bedrägeridetektion.

* **Logistisk regression**: Levererar stabila resultat men tenderar att missa fler bedrägerier (lägre recall).
* **Random Forest**: Bra balans mellan precision och recall, vilket gör modellen attraktiv i praktiska tillämpningar.
* **Neuralt nätverk**: Högst F1-score och AUC, vilket indikerar att den bäst kan skilja mellan bedrägerier och legitima transaktioner.

Confusion Matrix – Random Forest

En bild som visar text, skärmbild

AI-genererat innehåll kan vara felaktigt.

**Confusion Matrix – Logistic Regression**

Här ser vi att modellen upptäcker **6 av 12 bedrägerier (recall ≈ 0.50)**, men samtidigt flaggar **335 legitima transaktioner som bedrägerier (false positives)**.

**ROC-kurvor**

ROC-kurvorna visar tydligt att modellerna ligger ganska nära varandra, med AUC runt **0.6–0.7**. Logistic Regression presterar något lägre, medan Random Forest och det neurala nätverket ligger snäppet bättre. Alla modeller presterar dock klart bättre än slumpmässig gissning (AUC = 0.5).  
Kurvorna visar även hur olika tröskelvärden påverkar balansen mellan recall och precision. Genom att sänka tröskeln kan recall ökas (fler bedrägerier hittas), men detta sker på bekostnad av precision (fler falska alarm).

# Diskussion

**Tolkning av resultaten**

Resultaten visar att neurala nätverk ger bäst prestanda i termer av F1-score och AUC, men skillnaden mot Random Forest är inte enorm. Logistic Regression kan fortfarande vara användbar i miljöer där enkelhet och tolkbarhet prioriteras.

Detta illustrerar en viktig poäng: valet av modell bör inte enbart baseras på högsta möjliga resultat, utan också på praktiska överväganden såsom resurser, tolkbarhet och implementeringskostnader.

**Trade-off mellan precision och recall**

Ett centralt tema inom bedrägeridetektion är avvägningen mellan **precision** och **recall**:

* Hög recall innebär att fler bedrägerier hittas, men också fler falska alarm.
* Hög precision innebär att de transaktioner som flaggas verkligen är bedrägerier, men risken är att många faktiska bedrägerier missas.

För banker kan falska negativa (missade bedrägerier) leda till direkta ekonomiska förluster, medan falska positiva främst orsakar irritation och extra arbete. Därför är recall ofta viktigare än precision i dessa sammanhang.

**Etiska och praktiska aspekter**

Förutom den tekniska prestandan måste man även beakta de **etiska konsekvenserna**:

* **Kundupplevelse:** Om en modell flaggar för många falska positiva kan det leda till att kunder får avslag på legitima transaktioner. Detta kan skada kundrelationen och skapa frustration.
* **Risk för diskriminering:** Om datan som används innehåller bias, kan modellen omedvetet behandla vissa grupper sämre. Detta är särskilt relevant om geografiska eller demografiska variabler inkluderas.
* **Transparens:** Banker och myndigheter behöver kunna förklara varför en transaktion flaggades. Detta gör att enklare modeller (t.ex. Logistic Regression eller Random Forest) ibland föredras framför svårare att tolka neurala nätverk.

**Begränsningar i studien**

* Datasetet är anonymiserat med PCA, vilket gör att verkliga variabler som plats, handlare eller transaktionstyp inte kan analyseras.
* Modellerna testades i offline-miljö. I praktiken behövs system som fungerar i realtid.
* Endast tre modeller jämfördes; andra metoder som XGBoost eller djupare nätverk kan ge bättre resultat.

**Framtida arbete**

För att förbättra resultat ytterligare kan man tänka sig:

1. Att inkludera mer varierad och realistisk data.
2. Att använda *ensemblemetoder* där flera modeller kombineras för ökad robusthet.
3. Att undersöka förklarbara AI-metoder (t.ex. SHAP eller LIME) för att förstå modellernas beslut.
4. Att bygga realtidssystem som snabbt kan reagera på nya bedrägerimönster.

# Slutsats

Denna rapport har undersökt hur olika maskininlärningsmetoder kan användas för att upptäcka kreditkortsbedrägerier, ett problem som är både tekniskt komplext och av stor praktisk betydelse. Resultaten visar tydligt att maskininlärning erbjuder kraftfulla verktyg för att analysera obalanserade dataset och identifiera mönster som inte är möjliga att fånga med traditionella regelbaserade system.

Av de testade modellerna presterade det neurala nätverket bäst, med högst F1-score och AUC. Detta visar att mer avancerade, icke-linjära metoder kan ge en högre grad av precision i svåra klassificeringsproblem. Samtidigt har neurala nätverk nackdelar: de kräver mer datorkraft, längre träningstid och är svårare att tolka. Random Forest framstod därför som ett starkt praktiskt alternativ, då den kombinerar hög prestanda med större robusthet och viss förklarbarhet. Logistisk regression fungerade som en bra baslinje, men dess begränsningar blev tydliga när det gällde att hantera komplexa mönster i datan.

Utöver de tekniska resultaten visar arbetet också att valet av **utvärderingsmått** är avgörande. Accuracy är otillräckligt i detta sammanhang, och det krävs en kombination av recall, precision, F1-score och AUC för att ge en rättvis bild av modellernas styrkor och svagheter. Särskilt recall är viktigt i bedrägeridetektion, eftersom varje missat bedrägeri kan innebära stora ekonomiska förluster.

Framtida arbete bör inriktas på flera områden:

1. **Större och mer varierade dataset** – genom att inkludera fler variabler (exempelvis geografiska och demografiska faktorer) kan modellerna bli mer träffsäkra.
2. **Realtidsanalys** – system som kan upptäcka och stoppa bedrägerier omedelbart är av största vikt för banker och betalningssystem.
3. **Förklarbar AI** – metoder som SHAP och LIME bör integreras för att öka transparensen och förståelsen för varför en transaktion flaggas som misstänkt.
4. **Kombination av modeller** – ensemblemetoder eller hybridlösningar kan ge ökad robusthet och minska risken för att modellen överanpassas till en viss typ av data.

Sammanfattningsvis kan konstateras att maskininlärning inte bara är en teoretisk möjlighet, utan ett praktiskt verktyg som redan idag kan bidra till effektivare och mer pålitlig bedrägeribekämpning. Utvecklingen inom området går snabbt, och med fortsatt forskning och förbättrade datakällor finns goda möjligheter att ytterligare minska både de ekonomiska förlusterna och den personliga påverkan som kreditkortsbedrägerier innebär.

# Källförteckning

 Kaggle. (2018). *Credit Card Fraud Detection*. Hämtad från <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2:a uppl.). O’Reilly Media.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357. <https://doi.org/10.1613/jair.953>

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research, 18*(17), 1–5. <http://jmlr.org/papers/v18/16-365.html>

 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

 Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797. <https://doi.org/10.1109/TNNLS.2017.2736643>

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *arXiv preprint* arXiv:1009.6119. <https://arxiv.org/abs/1009.6119>

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2019). Scarff: A Scalable Framework for Streaming Credit Card Fraud Detection with Spark. *Information Fusion, 41*, 182–194. <https://doi.org/10.1016/j.inffus.2017.09.005>

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. *Decision Support Systems, 50*(3), 559–569. <https://doi.org/10.1016/j.dss.2010.08.006>