Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

Hodnocení klasifikačních metod na nevyvážených datových souborech

DIPLOMOVÁ PRÁCE

Specializace: Datové inženýrství
Studijní obor: Ekonometrie a operační výzkum

Autor: Bc. Jan Šimůnek
Vedoucí práce: Ing. Petra Tomanová, Ph.D., MSc

Praha, červen 2024

Poděkování

Chtěl bych velice poděkovat Ing. Petře Tomanové, Ph.D., MSc za její cenné rady, odborné konzultace a vedení, které významně přispěly k dokončení této diplomové práce.

Abstrakt

Tato diplomová práce se zaměřuje na analýzu výkonnosti různých klasifikačních metod při práci s nevyváženými datovými soubory. Zkoumá vliv technik vyvážení dat, jako jsou SMOTE, undersampling a oversampling, na výkonnost klasifikačních modelů. Hlavním cílem je zjistit, zda tyto techniky skutečně zlepšují predikční schopnosti klasifikačních modelů. Studie se soustředí na tři konkrétní klasifikační metody: logistickou regresi, Random Forest a Hellinger Distance Decision Tree (HDDT), a sleduje, na jakých datech jsou nejúčinnější. Dále se analyzuje využití Leave-One-Out Cross-Validation (LOOCV) a stanovení prahu pomocí kvantilu jako metod pro hodnocení modelů.

Klíčová slova

Hellinger Distance Decision Tree, Logistická regrese, nevyvážená data, Random Forest

Abstract

This thesis focuses on analyzing the performance of different classification methods when dealing with imbalanced datasets. It investigates the effect of data balancing techniques such as SMOTE, undersampling and oversampling on the performance of classification models. The main goal is to determine whether these techniques actually improve the prediction capabilities of classification models. The study focuses on three specific classification methods: logistic regression, Random Forest and Hellinger Distance Decision Tree (HDDT), and examines on which data they are most effective. It also analyzes the use of Leave-One-Out Cross-Validation (LOOCV) and thresholding using quantile as methods for evaluating models.

Keywords

Hellinger Distance Decision Tree, imbalanced data, Logistic regression, Random Forest

Obsah

Úvod

Ve světě datové analýzy a strojového učení je práce s nevyváženými datasety jedním z nejčastějších a zároveň nejnáročnějších problémů [1]. Nízký počet pozitivních případů značně komplikuje odhad pravděpodobnosti jejich výskytu a následnou predikci. Tento problém je zvláště patrný v oblastech, kde jsou falešně negativní výsledky mnohem kritičtější než falešně pozitivní, například v medicínských datech. Při detekci nemoci může falešně negativní výsledek znamenat, že pacient nedostane potřebnou léčbu včas., což může vést k vážným zdravotním komplikacím.

Hlavní motivací této práce je přispět k řešení výzvy, kterou představují nevyvážené datasety v různých aplikacích. Výzkum efektivních metod pro práci s těmito daty může významně přispět ke zlepšení diagnostických nástrojů, detekci podvodů, credit risku a dalších oblastí.

Práce se zaměřuje na porovnání výkonnosti tří různých metod (Logit, Random Forest a Hellinger Distance Decision Tree) na datasetech s různě nevyváženými vysvětlovanými proměnnými. Současně se zabývá efektivitou různých technik pro vyvážení dat, jako jsou SMOTE, undersampling a oversampling. Výkonnost těchto metod bude hodnocena pomocí metrik Precision, Recall, F1 score, ROC křivka a PR křivka. Hlavním cílem práce je zjistit, zda techniky na vyvážení dat skutečně zlepšují výkonnost modelů, a pokud ano, tak které kombinace algoritmů a technik jsou nejúčinnější. Dalším cílem je zjistit, zda použití Leave-One-Out Cross-Validation (LOOCV) u datasetů s vysoce nevyváženou vysvětlovanou proměnnou zlepšuje výkonnost modelů. Dále bude analyzováno, zda nastavení kvantilu pro seperaci tříd, namísto hledání prahové hodnoty s nejvyšším F1 score, vede k lepší rovnováze mezi precision a recall, tedy zda zvýšení precision způsobuje přiměřené snížení recall.

Hellinger Distance Decision Tree byl jako metoda do této práce zvolen kvůli jeho prokázané robustnosti a odolnosti vůči nevyváženosti dat, jak je detailně popsáno ve studii Cieslaka [2]. Tento přístup je vhodný zejména v kontextech, kde je kritická přesnost a robustnost modelu při práci s nevyváženými daty.

Tímto tématem se zabýval například van den Goorbergh [3], který se se svými kolegy zabýval metodami korekce nevyvážených dat v modelech rizikových predikcí, konkrétně pomocí penalizované (ridge) logistické regrese. V jejich práci zkoumali čtyři metody korekce nevyváženosti tříd: žádnou korekci, náhodný undersampling, náhodný oversampling a SMOTE. Metody korekce nevyváženosti, včetně SMOTE, vedly v této práci k špatně kalibrovaným modelům, kde byla pravděpodobnost příslušnosti k menšinové třídě silně přeceněna.

Práce je strukturována následovně. Nejprve je představena teoretická část, která se věnuje popisu jednotlivých metod a technik pro vyvážení dat. Následně jsou popsány datové sady a způsob jejich úpravy pro účely experimentů. Poté jsou provedeny experimenty s použitím výše zmíněných metod a technik na původních i upravených datech. Nakonec jsou výsledky experimentů analyzovány a porovnány pomocí zvolených metrik.

1. Teorie

1.1 Logistická regrese

Logistická regrese je statistická metoda používaná k modelování binárních nebo dichotomických výsledků, kde výstupní proměnná může nabývat pouze dvou hodnot, typicky označených jako 0 a 1. Dle Kleinbauma a Kleina [4] je tato metoda populární kvůli logistické funkci, která má rozsah mezi 0 a 1. Model je určen k popisu pravděpodobnosti, která je vždy číslo mezi 0 a 1. Tato pravděpodobnost může popisovat například riziko selhání klienta, či riziko, že pacient onemocní určitou chorobou.

1.1.1 Základní principy logistické regrese

1. **Logitová funkce:** Logistická regrese používá logitovou funkci, která transformuje pravděpodobnostní hodnoty do škály od $-\infty$ do $+\infty$. Logitová funkce je definována jako:
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{1.1}$$
kde p je pravděpodobnost, že vysvětlovaná proměnná nabývá hodnoty 1.
2. **Model:** Logistický regresní model vyjadřuje vztah mezi logitovou funkcí a vysvětlujícími proměnnými:
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{1.2}$$
kde β_0 je směrnice a $\beta_1, \beta_2, ..., \beta_k$ jsou regresní koeficienty pro vysvětlující proměnné $X_1, X_2, ..., X_k$ a k je index vysvětlující proměnné.
3. **Odhad parametrů:** Parametry modelu jsou odhadnuty pomocí metody maximální věrohodnosti. Tato metoda poskytuje hodnoty neznámých parametrů, které maximalizují pravděpodobnost získání pozorovaného souboru dat. [5]
4. **Predikce pravděpodobnosti:** Pravděpodobnost, že vysvětlovaná proměnná je 1, je dána inverzní logitovou funkcí:
$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}. \tag{1.3}$$
5. **Hodnocení modelu:** Pro hodnocení přizpůsobení modelu se používají různé statistiky, jako je precision, recall a F1 score. Křivky Receiver Operating Characteristic (ROC) a Precision–Recall (PR) se používají k hodnocení diskriminační schopnosti modelu. Další běžně používaný test je Hosmer-Lemeshow test, který testuje shodu mezi pozorovanými a predikovanými hodnotami [5].
6. **Interpretační prvky:** Interpretačními prvky jsou koeficienty. Hodnoty koeficientů β udávají změnu v logitové hodnotě (log-odds) pro jednotkovou změnu vysvětlující proměnné. Interpretačními prvky může být i poměr šancí. Exponenciace koeficientu e^{β} dává poměr šancí, které udává, jak se mění šance na výsledek 1 s jednotkovou změnou ve vysvětlující proměnné.

1.1.2 Výhody logistické regrese při práci s nevyváženými daty

- **Modelování binárních výsledků:** Logistická regrese je speciálně navržena pro modelování binárních výsledků, kde závislá proměnná nabývá pouze dvou možných hodnot. [6]
- **Jednoduchá interpretace:** Výsledky logistické regrese jsou snadno interpretovatelné, což je výhodné při práci s komplexními daty a modely. Koeficienty lze interpretovat jako logaritmický poměr šancí nebo po exponenciaci jako poměr šancí.

1.2 Random Forest

Random Forest je ensemble learning metoda, která kombinuje výsledky více rozhodovacích stromů pro klasifikaci, regresi nebo jiné úkoly. Je to efektivní metoda pro klasifikační úlohy a dobře zvládá nevyvážené soubory dat. [7] Herrera ho popisuje tak, že je odolný vůči odlehlým hodnotám a šumu, překonává problémy s overfttingem a zvládá mírně nevyvážená data. [8]

1.2.1 Základní principy Random Forest

1. **Bootstrap vzorkování:** Algoritmus začíná bootstrap vzorkováním. Pro každý rozhodovací strom T_b (kde b je index stromu), se náhodně vybírá vzorek s opakováním z tréninkových dat D.

$$D_b = \{(x_i, y_i)\}_{i=1}^{n_b}, \tag{1.4}$$

   kde n_b je velikost bootstrap vzorku (obvykle stejná jako velikost původního tréninkového vzorku n). x_i označuje vektor vysvětlujících proměnných pro i-té pozorování. y_i označuje vysvětlující proměnnou pro i-té pozorování.
2. **Konstrukce rozhodovacích stromů:** Každý strom T_b je trénován na vzorku D_b. Tento proces se opakuje, dokud není dosaženo maximální hloubky stromu nebo jiného zastavovacího kritéria.
3. **Agregace výsledků:** Pro predikci se agregují výsledky všech rozhodovacích stromů. V případě binární klasifikace se obvykle používá většinové hlasování:

$$\hat{y} = \text{mode}\{T_b(x)\}_{b=1}^{B}, \tag{1.5}$$

   kde mode reprezentuje nejčastěji se vyskytující hodnota mezi predikcemi jednotlivých stromů.

1.2.2 Výhody Random Forest při práci s nevyváženými daty

- **Odolnost vůči overfittingu:** Díky použití více stromů a náhodnému výběru vzorků a vlastností je Random Forest odolný vůči overfittingu. [7][8]
- **Efektivní v práci s velkými daty:** Random Forest dobře škáluje s velkými daty a vysokou dimenzionalitou. [8]

- **Vestavěné vyvažování třídy:** Algoritmus může použít různé techniky pro vyvažování třídy, jako je přidělování váh třídám nebo použití oversampling/undersampling technik během bootstrapování.

1.3 Helliner Distance Decision Tree

Hellinger Distance Decision Tree (HDDT) je metoda, která používá Hellingerovu vzdálenost k rozdělování datových uzlů v rozhodovacím stromu. Tento přístup zlepšuje schopnost stromu zvládat nevyvážená data a zvyšuje robustnost modelu.

1.3.1 Základní principy HDDT

1. **Hellinerova vzdálenost:** Hellingerova vzdálenost je metrika používaná k měření podobnosti mezi dvěma pravděpodobnostními rozděleními. Pro dvě pravděpodobnostní rozdělení P a Q je Hellingerova vzdálenost definována jako:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{P(i)} - \sqrt{Q(i)})^2}, \tag{1.6}$$

   kde $P(i)$ a $Q(i)$ jsou pravděpodobnosti pro kategorii i ve dvou různých rozděleních. n značí celkový počet kategorií nebo tříd, přes které se Hellingerova vzdálenost počítá.
2. **Rozhodovací stromy:** Rozhodovací stromy jsou modely založené na pravidlech, které se opakovaně dělí datovou sadu na podmnožiny na základě hodnot vysvětlujících proměnných. HDDT využívá Hellingerovu vzdálenost k výběru nejlepšího dělení v každém uzlu stromu.
3. **Výběr dělení:** V každém uzlu stromu HDDT se spočítají pravděpodobnostní rozdělení P_L a P_R pro levý a pravý poduzel na základě kandidátských dělení. Dělení, které minimalizuje Hellingerovu vzdálenost mezi těmito rozděleními, je vybráno jako nejlepší. Hellingerova vzdálenost v uzlu i se vypočítá jako:

$$H_i = \sqrt{\left(\sqrt{\frac{N_{\text{left}_i}}{N_i}} - \sqrt{\frac{P_{\text{left}_i}}{P_i}}\right)^2 + \left(\sqrt{\frac{N_{\text{right}_i}}{N_i}} - \sqrt{\frac{P_{\text{right}_i}}{P_i}}\right)^2}, \tag{1.7}$$

   kde N_i a P_i jsou počty negativních a pozitivních případů v rodičovském uzlu i, N_{left_i} a P_{left_i} jsou počty případů, které spadají do levého poduzlu, a N_{right_i} a P_{right_i} jsou počty případů, které spadají do pravého poduzlu.
4. **Agregace výsledků:** Pro predikci u binárních úloh se používá většinové hlasování (majority voting) stejně jako u metody Random Forest.

1.3.2 Výhody HDDT při práci s nevyváženými daty

- **Odolnost vůči nevyváženým datům:** Díky použití Hellingerovy vzdálenosti je HDDT lépe schopen pracovat s nevyváženými daty, protože tato metrika je méně citlivá na

různé velikosti tříd než jiné metriky, jako je Giniho index nebo entropie. [2]
- **Robustnost a přesnost:** HDDT zvyšuje přesnost a robustnost modelu tím, že vybírá nejlepší dělení na základě Hellingerovy vzdálenosti. [2]
- **Efektivita bez nutnosti samplování:** HDDT mohou dosáhnout vysoké výkonnosti na nevyvážených datech bez nutnosti použití samplovacích metod, což snižuje složitost a potřebu ladění parametrů [2].

1.4 Nevyvážená data

Nevyvážená data představují významný problém v oblasti strojového učení a statistického modelování. Tento problém nastává, když jedna nebo více tříd v datové sadě je výrazně méně častá než ostatní. To může vést k modelům, které mají tendenci ignorovat vzácnější třídy, což může mít závažné důsledky v aplikacích, jako je detekce podvodů, lékařská diagnostika nebo odhalování vad.

1.4.1 Problémy spojené s nevyváženými daty

1. **Zkreslení v modelu:** Tradiční modely trénované na nevyvážených datech mohou být zaujaté ve prospěch většinové třídy, což vede k vysoké přesnosti na úkor nízké citlivosti pro menšinovou třídu. To znamená, že modely mohou mít tendenci přehlížet nebo chybně klasifikovat případy menšinové třídy. [9]
2. **Nevhodné metriky hodnocení:** Klasické metriky jako přesnost (accuracy) mohou být zavádějící, protože mohou maskovat špatnou výkonnost modelu na menšinové třídě. Proto jsou preferovány metriky jako precision, recall, F1 score, ROC křivka a PR křivka, které poskytují lepší přehled o výkonnosti modelu na nevyvážených datech. [10]
3. **Problémy s generalizací:** Modely trénované na nevyvážených datech mohou mít problémy s generalizací a mohou selhat při aplikaci na nové, neviděné vzorky, zvláště pokud se poměr tříd v tréninkové sadě liší od toho ve skutečném světě.

1.4.2 Metody řešení nevyvážených dat

Pro řešení problémů spojených s nevyváženými daty existují následující techniky:

1. **Resampling techniky:** Podrobněji popsány v sekci 1.5
   - **SMOTE Synthetic Minority Over-sampling Technique):** Generování nových syntetických vzorků menšinové třídy.
   - **Oversampling:** Zvýšení počtu vzorků menšinové třídy.
   - **Undersampling:** Snížení počtu vzorků většinové třídy.
2. **Použití vhodných metrik:** Metriky jako precision, recall, F1 score, ROC křivka a PR křivka poskytují lepší přehled o výkonnosti modelu na nevyvážených datech. [10]

3. **Vyvažování tříd:** Přidělování vyšších vah menšinové třídě během trénování modelu, aby model bral v úvahu nerovnoměrné rozložení tříd.
4. **Použití specializovaných algoritmů:** Algoritmy jako Hellinger Distance Decision Tree (HDDT) jsou navrženy tak, aby lépe zvládaly nevyvážená data tím, že používají metriky méně citlivé na nevyváženost tříd. [2]

1.4.3 Leave-One-Out Cross Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) je speciální případ k-fold cross validace, kde k je rovno počtu vzorků v datové sadě. Každý vzorek je jednou použit jako testovací data a zbývající vzorky jako tréninková data. Tento přístup má několik výhod i nevýhod při práci s nevyváženými daty: [11]

1. **Výhody:**
   - Maximální využití dostupných dat pro trénování modelu.
   - Zvětšení absolutního počtu pozorování u menšinové třídy, což zajišťuje, že každý vzorek, včetně těch z menšinové třídy, je použit jako testovací data alespoň jednou, čímž se zvyšuje šance na zachycení výkonu modelu na těchto vzorcích.
2. **Nevýhody:**
   - Výpočetně náročné, zejména pro velké datové sady.
   - Může trpět vysokou variabilitou, pokud jsou data velmi nevyvážená.

LOOCV je užitečný nástroj pro hodnocení výkonnosti modelů, zejména když je k dispozici omezený počet vzorků. Avšak pro velké a nevyvážené datové sady mohou být vhodnější jiné formy cross validace nebo resampling techniky.

1.5 Techniky pro vyvážení dat

V oblasti práce s nevyváženými daty je k dispozici několik technik. Mezi nejčastěji používané patří SMOTE, undersampling a oversampling. Tyto metody upravují poměr tříd v datové sadě, což může vést ke zvýšení výkonnosti modelů strojového učení.

1.5.1 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE je technika oversamplingu, která generuje nové syntetické vzorky menšinové třídy. Místo jednoduchého kopírování existujících vzorků vytváří SMOTE nové vzorky interpolací mezi existujícími vzorky menšinové třídy. Dle Aggarwala [12] se provádí následujícím způsobem:

1. Pro každý vzorek menšinové třídy se vybere k nejbližších sousedů (obvykle $k = 5$).
2. Nový syntetický vzorek je poté vygenerován podél úsečky spojující tento menšinový vzorek s jedním z jeho náhodně vybraných nejbližších sousedů.

3. Tento proces se opakuje, dokud není dosaženo požadovaného počtu nových vzorků s menšinovou třídou.

Nový vzorek x'_i vytvořit jako: [13]

$$x'_i = x_i + \lambda \cdot (x_j - x_i),$$

kde x_i je původní vzorek s menšínovou třídou, x_j je jeden z k nejbližších sousedů s menšinovou třídou a λ je náhodná hodnota v intervalu $[0, 1]$. Náhodná hodnota λ určuje polohu nového syntetického vzorku na úsečce mezi původním vzorkem a jeho sousedem.

1.5.2 Undersampling

Undersampling je technika, která snižuje počet vzorků většinové třídy. Tím se zlepšuje vyváženost mezi třídami a snižuje potenciální zkreslení vůči většinové třídě. Toho lze dosáhnout náhodným odstraněním vzorků z většinové třídy, dokud není dosaženo požadovaného poměru tříd.

Hlavním výhodou undersamplingu je zjednodušení modelu, který nemusí být tolik náchylný k přeučení. Dalšími výhodami jsou snížení doby trénování modelu a může zlepšit generalizaci modelu tím, že odstraní nadbytečná pozorování, která mohou způsobovat zkreslení a nadměrné přizpůsobení specifickým vzorcům v datech.

Naopak významnou nevýhodou je zmenšení velikosti souboru dat, což může negativně ovlivnit prediktivní výkonnost modelu. Odstranění pozorování vede ke ztrátě cenných informací. [14]

1.5.3 Oversampling

Oversampling je metoda, která zahrnuje zvýšení počtu vzorků menšinové třídy. Nejjednodušší forma oversamplingu zahrnuje replikaci existujících vzorků menšinové třídy, aby se zvýšil jejich počet a zlepšila se vyváženost datové sady.

Výhodou je, že nedochází ke ztrátě informací. To je výhodné pro trénování složitých modelů s mnoha parametry. Protože oversampling zahrnuje všechny záznamy z původního souboru dat, nedochází ke ztrátě informací.

Nevýhodou oversamplingu je, že může vést k přeučení, protože zvyšuje přítomnost určitých charakteristik v souboru dat. Ať už prostřednictvím duplicitních nebo syntetických pozorování, může způsobit, že se model příliš zaměří na specifické detaily, které nemusí být široce použitelnými. Také zvyšuje dobu trénování modelu. [14]

1.6 Hodnotící metriky

Při práci s nevyváženými daty je důležité zvolit vhodné metriky pro hodnocení výkonnosti modelů. Mezi klíčové metriky patří precision, recall, F1 score, ROC křivka a PR křivka. Tyto metriky poskytují cenné informace o schopnosti modelu správně klasifikovat vzorky nejen v přítomnosti nevyvážených dat, ale také v různých jiných aplikacích strojového učení.

1.6.1 Precision (Přesnost)

Precision měří, jaký podíl pozorování klasifikovaných jako pozitivní jsou skutečně pozitivní. Je definována jako poměr počtu skutečně pozitivních předpovědí k celkovému počtu pozitivních předpovědí následujícím vzorcem [15]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

kde TP je počet skutečně pozitivních předpovědí (true positives) a FP je počet falešně pozitivní předpovědí (false positives).

Precision je důležitá v kontextech, kde je třeba minimalizovat falešně pozitivní výsledky, například při detekci podvodů nebo spamových zpráv.

1.6.2 Recall (Citlivost)

Recall měří, jaký podíl skutečně pozitivních pozorování je správně identifikováno modelem. Je definován jako poměr počtu skutečně pozitivních předpovědí k celkovému počtu skutečných pozitiv následujícím vzorcem [15]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

kde TP je počet skutečně pozitivních předpovědí (true positives) a FN je počet falešně negativních předpovědí (false negatives).

Recall je klíčový v situacích, kde je důležité minimalizovat falešně negativní výsledky, například v lékařské diagnostice.

1.6.3 F1 score

F1 score je harmonický průměr precision a recall. Tato metrika poskytuje vyvážené měřítko pro modely, které mají různou precision a recall. F1 score lze zapsat vzorcem [15]

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

F1 score je klíčové v situacích, kde je třeba rovnováha mezi precision a recall. Vzhledem k tomu, že harmonický průměr klade větší důraz na nižší hodnoty z obou metrik, F1 score penalizuje extrémně nízké hodnoty precision nebo recall, čímž podporuje modely, které dosahují rovnováhy mezi těmito dvěma aspekty. F1 score je obzvláště vhodné pro hodnocení výkonu modelů na nevyvážených datech. Když jsou data nevyvážená, modely mohou mít tendenci ignorovat menšinovou třídu a soustředit se na většinovou třídu, což může vést k vysoké precision, ale nízkému recall. Maximalizace F1 score může být také použita k nalezení nejvhodnějšího prahu pro klasifikaci predikovaných tříd.

1.6.4 ROC křivka (Receiver Operating Characteristic)

ROC křivka je grafické znázornění výkonu klasifikačního modelu při různých prahových hodnotách. Graf zobrazuje True Positive Rate (TPR) proti False Positive Rate (FPR), které se vypočítají následovně:

- $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$
- $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$,

kde TP je počet skutečně pozitivních předpovědí (true positives), FN je počet falešně negativních předpovědí (false negatives), FP je počet falešně pozitivní předpovědí (false positives) a TN je počet skutečně negativních předpovědí (true negatives).

ROC křivka poskytuje přehled o tom, jak dobře model rozlišuje mezi třídami při různých prahových hodnotách.

AUROC (Area Under ROC curve) napomáhá interpretaci ROC křivky a hodnotí celkovou schopnost modelu rozlišovat mezi třídami. Hodnota AUROC by se měla pohybovat mezi 0.5 a 1, nicméně může nabýt i hodnot pod 0.5. Draelos [16] interpretuje hodnoty AUROC následovně:

- AUROC nižší než 0.5 značí, že model je horší než náhodný klasifikátor.
- AUROC 0.5 odpovídá náhodnému klasifikátoru.
- AUROC mezi 0.5 a 0.7 znamená suboptimální výkonnost.
- AUROC 0.7 - 0.8 značí dobrou výkonnost.
- AUROC vyšší než 0.8 znamená vynikající výkonnost.
- AUROC 1.0 odpovídá dokonalému klasifikátoru.

1.6.5 PR křivka (Precision-Recall)

PR křivka zobrazuje precision proti recall při různých prahových hodnotách. Tato křivka je obzvláště užitečná při práci s nevyváženými daty, protože poskytuje lepší přehled o výkonnosti modelu při identifikaci menšinové třídy. Saito [17] tvrdí, že na základě simulací v jeho práci "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets"je PR křivka více informativním a výkonnějším

grafem pro nevyvážené datasety než ROC křivka. Navzdory tomu, většina studií o binárních klasifikátorech ve spojení s nevyváženými soubory dat používá jako hlavní metodu hodnocení výkonnosti graf ROC.

AUPRC (Area Under PR curve) kvantifikuje PR křivku do jednoho čísla, které shrnuje výkonnost modelu. Vysoká hodnota AUPRC naznačuje, že model má dobrý trade-off mezi precision a recall, což je důležité při práci s nevyváženými daty. Avšak interpretace je složitější než u AUROC, protože narozdíl od AUROC, kde je základní hodnota 0.5 vždy stejná, základní hodnota AUPRC závisí na podílu pozitivních případů v datové sadě, což znamená, že hodnoty AUPRC nejsou absolutní a mění se s různou distribucí tříd. Z toho vyplývá, že při velké nevyváženosti datasetu může i nízká hodnota AUPRC poukazovat na velmi dobrou výkonnost modelu [18].

1.6.6 Význam pro nevyvážená data

- **Precision** pomáhá minimalizovat falešně pozitivní výsledky, což je důležité tam, kde falešné poplachy mohou mít vysoké náklady.
- **Recall** je klíčový pro minimalizaci falešně negativních výsledků, což je důležité tam, kde je důležité zachytit všechny relevantní případy.
- **F1 score** poskytuje vyvážené hodnocení výkonu modelu, když je třeba zohlednit jak precision, tak recall.
- **ROC křivka** je užitečná pro hodnocení celkové schopnosti modelu rozlišovat mezi třídami.
- **PR křivka** je obzvláště vhodná pro nevyvážená data, protože poskytuje lepší přehled o výkonnosti modelu při identifikaci menšinové třídy.

1.7 Stepwise regrese

Stepwise regrese je metoda selekce modelu používaná pro výběr vysvětlujících proměnných, jejímž cílem je identifikovat podmnožinu vstupních vysvětlujících proměnných, které jsou nejvhodnější pro predikci vysvětlované proměnné.

Tato metoda pomáhá vyhnout se nadměrnému přeučení modelu tím, že zahrnuje pouze nejinformativnější vstupní vysvětlující proměnné, čímž zlepšuje zobecnitelnost a interpretovatelnost modelu.

Mezi běžné ukazatele hodnocení výkonu používané v stepwise regresi patří například Akaikeho informační kritérium (AIC), Bayesovo informační kritérium (BIC) a modifikovaný R-squared. [19]

Při použití AIC jako ukazetele pro hodnocení výkonu ve stepwise regresi je cílem nalézt model s nejnižší hodnotou AIC. AIC penalizuje složitost modelu, takže preferuje modely, které vysvětlují data dobře, ale nejsou zbytečně složité.

2. Empirická studie

Tato kapitola se zaměřuje na systematické zpracování a analýzu datasetů, které jsou klíčové pro validaci a testování vybraných modelů. Kapitola je rozdělena do čtyř hlavních sekcí: Popis dat a jejich základní zpracování, Transformace a úprava dat, Rozdělení dat a Specifikace modelu. Každá sekce popisuje jednotlivé kroky a techniky, které byly použity k přípravě a analýze dat.

2.1 Popis dat a jejich základní zpracování

Tato sekce se zaměřuje na podrobný popis použitých datasetů, včetně jejich struktury a typů proměnných. Cílem je poskytnout přehled o použitých datech. Bude zde popsána práce s chybějícími hodnotami a bude zde provedeno základní očištění dat, aby byla zajištěna jejich kvalita a přesnost pro následné analýzy.

2.1.1 Definování algoritmu na základní zpracování dat

Vstupy:

- D: Původní dataset
- x_i: Pozorování v datasetu D, kde i je index pozorování
- x_j: Proměnná v datasetu D, kde j je index proměnné
- x_k: Proměnná v datasetu D, kde k je index proměnné odlišné od j
- T: Prahová hodnota pro detekci vysoké multikolinearity (např. 2)
- M: Maximální povolený věk (např. 95 let)
- P: Prahová hodnota pro procento chybějících hodnot (např. 5 %)

Výstupy:

- D': Očištěný dataset po zpracování.

Kroky:

1. Identifikace a odstranění chybějících hodnot:

- Pokud je podíl chybějících hodnot v proměnné x_j vyšší než P:

$$\forall x_j \in D, \text{ pokud } \frac{\sum_{i=1}^{n} I(x_{ij} = \text{NA})}{n} > P, \text{ bude odstraněna proměnná } x_j$$

- Pokud jsou v datasetu přítomny další chybějící hodnoty, odstranit pozorování, která tyto chybějící hodnoty obsahují:

$$\forall X_i \in D, \text{ pokud } x_{ij} = \text{NA pro jakékoliv } j, \text{ bude odstraněno } X_i$$

2. Odstranění neplatných hodnot:

- Pokud je věk vyšší než M:

$$\forall X_i \in D, \text{ pokud } x_{i,\text{age}} > M, \text{ bude odstraněno } X_i$$

- Pokud je délka zaměstnání o méně než 15 menší než věk, odstranit:

$$\forall X_i \in D, \text{ pokud } x_{i,\text{emp\_length}} > (x_{i,\text{age}} - 15), \text{ bude odstraněno } X_i$$

- (Obecně pro proměnné mající tento význam, nikoliv nutně nazvané *emp_length* nebo *age*)

3. Testování a odstranění aliasovaných (lineárně závislých) proměnných:

- Pro každou dvojici proměnných (x_j, x_k) v datasetu D, kde $j \neq k$:

$$\text{pokud existuje lineární závislost, bude odstraněna proměnná } x_j$$

4. Testování multikolinearity:

- Opakovat, dokud proměnná s maximálním $GVIF^{1/(2Df)}(x_j)$ má hodnotu vyšší než T:

$$\text{Identifikovat proměnnou s maximální hodnotou } GVIF^{1/(2Df)}(x_j).$$

$$\text{Pokud tato hodnota přesahuje } T, \text{ odstranit proměnnou } x_j.$$

5. Kombinování úrovní kategoriálních proměnných:

- Pro každou ordinální proměnnou x_j, pokud má počet hodnot nižší než 5:
  - Pokud se jedná o nejvyšší úroveň, kombinovat ji s nižší úrovní.
  - Pokud se jedná o nejnižší úroveň, kombinovat ji s vyšší úrovní.

6. Výběr relevantních proměnných pomocí stepwise regrese:

- Použít stepwise regresi k výběru nejrelevantnějších proměnných na základě AIC.

  Iterativně přidávat nebo odstraňovat proměnné z modelu, aby se minimalizovala hodnota AIC.

7. Finální dataset:

- $D' = D$ po provedení všech výše uvedených kroků.

2.1.2 Dataset 1: Credit Risk Dataset

Přehled dat

Prvním datasetem je "Credit Risk Dataset"[20] od společnosti Kaggle. Skládá se z různých proměnných souvisejících s žádostmi o půjčku, včetně osobních a finančních informací žadatelů, a používá se k predikci pravděpodobnosti splácení půjčky.

Originální dataset obsahuje 32 581 pozorování a 12 proměnných a obsahuje kvantitativní, kategoriální a binární proměnné. Jako vysvětlovaná proměnná byla v tomto datasetu vybrána proměnná *loan_status*, která popisuje, jestli žadatel řádně splácí úvěr. Tato proměnná je binární a v původním datasetu měla třídy 0 a 1 nevyvážené v poměru 25 474:7108.

Popis proměnných

Seznam proměnných použitých v konečném datasetu je uveden v tabulce 2.1.

Proměnná	Datový typ	Popis
person_income	Kvantitativní	Roční příjem žadatele
person_home_ownership	Kategoriální	Vlastnictký status bydlení žadatele
loan_intent	Kategoriální	Účel půjčky
loan_grade	Kategoriální	Úvěrový stupeň přiřazený k půjčce
loan_amnt	Kvantitativní	Požadovaná částka půjčky
loan_status	Binární	Řádné splácení půjčky žadatelem
loan_percent_income	Kvantitativní	Procento příjmu žádatele určené na splácení půjčky
cb_person_cred_hist_length	Kvantitativní	Délka úvěrové historie žadatele v letech

Tabulka 2.1: Dataset 1 (Credit Risk Dataset) – popis

Chybějící hodnoty

V datasetu bylo 895 chybějících hodnot u proměnné $person\_emp\_length$ a 3116 u proměnné $loan\_int\_rate$. Proměnná $loan\_int\_rate$ byla vzhledem k podílu chybějících hodnot nad 5 % vyřazena. Ostatní chybějící hodnoty byly vzhledem k velikosti datasetu a kvantitativnímu datovému typu proměnné $person\_emp\_length$ vyřazeny.

Odstranění neplatných hodnot

Byla vyřazena pozorování s věkem vyšším než 95 let a také ta, kde délka zaměstnání byla o méně než 15 let nižší než věk. Tyto hodnoty byly považovány za nerealistické.

Testování a odstranění aliasovaných proměnných

Aliasované proměnné nebyly identifikovány.

Testování multikolinearity

Multikolinearita byla měřena pomocí $GVIF^{1/(2Df)}$ skóre. Nejprve bylo nejvyšší $GVIF^{1/(2Df)}$ skóre 2.11 u proměnné $person\_age$. Poněvadž dle Zhanga [21] $GVIF^{1/(2Df)}$ vysokou multikolinearitu indikují hodnoty vyšší než 2, byla proměnná $person\_age$ vyřazena. Po těchto úpravách již žádná proměnná neměla $GVIF^{1/(2Df)}$ skóre vyšší než 2.

Výběr relevantních proměnných pomocí stepwise regrese

Z datasetu byly vyřazeny proměnné $person\_emp\_length$ a $cb\_person\_default\_on\_file$.

Finální dataset

Po všech úpravách obsahuje výsledný dataset 31 679 pozorování a 8 proměnných.

2.1.3 Dataset 2: Myocardial Infarction Complication

Přehled dat

Druhým použitým datasetem je "Myocardial Infarction Complications"[22], který je k dispozici v úložišti strojového učení UCI. Dataset je určen pro klasifikační úlohy, zejména pro předpovídání komplikací po infarktu myokardu. Data zachycují informace o pacientovi v různých fázích: při přijetí a 24, 48 a 72 hodin po přijetí. Dataset obsahuje demografické údaje, anamnézu a klinická měření.

Původní dataset obsahoval 1700 pozorování a 124 proměnných. Vysvětlovanou binární proměnnou v tomto datasetu je proměnná $V114$, která udává, zda pacient trpí supraventrikulární tachykardií. Tato proměnná byla vybrána z důvodu její extrémní nevyváženosti, v původním datasetu byly totiž třídy 0 a 1 u této proměnné nevyvážené v poměru 1680:20.

Popis proměnných

Seznam proměnných použitých v konečném datasetu je uveden v tabulce 2.2.

Proměnná	Datový typ	Popis
V10	Binární	Indikuje, jestli pacient trpí symptomatickou hypertenzí
V27	Binární	Indikuje, jestli pacient trpí diabetem mellitem v anamnéze
V42	Binární	Indikuje, jestli pacient trpěl záchvaty supraventrikulární tachykardie v době přijetí na JIP
V49	Binární	Indikuje, jestli pacient měl infarkt myokardu pravé komory
V100	Kvantativní	Indikuje použití opioidních léků na JIP v prvních hodinách hospitalizace
V107	Binární	Indikuje použití beta-blokátorů na JIP
V113	Binární	Indikuje, jestli pacient trpí fibrilací síní
V114	Binární	Indikuje, jestli pacient trpí supraventrikulární tachykardií
V115	Binární	Indikuje, jestli pacient trpí komorovou tachykardií

Tabulka 2.2: Dataset (Myocardial Infarction Complications) – popis

Chybějící hodnoty

Úprava dat byla zahájena vyřazením proměnné $V1$, která značila index pozorování. Poté byly identifikovány buňky označené jako ? a nahrazeny hodnotou NA. Bylo vypočítáno procento chybějících hodnot pro každou proměnnou a odstraněny byly ty, u nichž chybělo více než 5 % pozorování. Nakonec byla odstraněna pozorování s chybějícími hodnotami u zbývajících proměnných.

Odstranění neplatných hodnot

Nebyla identifikována žádná zjevná neplatná pozorování, která by bylo třeba odstranit.

Testování a odstranění aliasovaných proměnných

Aliasované proměnné nebyly v tomto datasetu identifikovány.

Testování multikolinearity

Při testování multikolinearity pomocí $GVIF^{1/(2Df)}$ skóre nebyla nalezena žádná proměnná s hodnotou vyšší než 2. Žádná další proměnná tudíž nebyla vyřazena.

Výběr relevantních proměnných pomocí stepwise regrese

Z datasetu bylo vyřazeno 61 proměnných.

Finální dataset

Výsledný dataset obsahuje 1436 pozorování a 9 proměnných.

2.1.4 Dataset 3: Diabetes Health Indicators

Přehled dat

Třetím datasetem je dataset "Diabetes Health Indicators Dataset"[23], který obsahuje indikátory zdraví související s diabetem. Data jsou shromážděna z Behaviorálního rizikového faktoru sledování systému (BRFSS) od amerického Centra pro kontrolu a prevenci nemocí (CDC). Dataset zahrnuje různé zdravotní, demografické a behaviorální faktory.

Použitou vysvětlovanou proměnnou je proměnná *Diabetes_binary*, která byla vybrána z důvodu její poměrně velké nevyváženosti. Třídy 0 a 1 u této binární proměnné jsou nevyvážené v poměru 218 334:35 346.

Popis proměnných

Seznam proměnných použitých v konečném datasetu je uveden v tabulce 2.3.

Chybějící hodnoty

Dataset neobsahuje žádné chybějící hodnoty.

Odstranění neplatných hodnot

Nebyla identifikována žádná zjevná neplatná pozorování, která by bylo třeba odstranit.

Testování a odstranění aliasovaných proměnných

Aliasované proměnné nebyly v tomto datasetu identifikovány.

Testování multikolinearity

Proměnná	Datový typ	Popis
Diabetes_binary	Binární	Indikuje přítomnost diabetu
HighBP	Binární	Indikuje, zda má respondent vysoký krevní tlak
HighChol	Binární	Indikuje, zda má respondent vysokou hladinu cholesterolu
CholCheck	Binární	Indikuje, zda respondent podstoupil kontrolu hladiny cholesterolu
BMI	Kvantitativní	Index tělesné hmotnosti
Smoker	Binární	Indikuje, zda je respondent kuřák
Stroke	Binární	Indikuje, zda měl respondent mrtvici
HeartDiseaseorAttack	Binární	Indikuje, zda měl respondent srdeční onemocnění nebo infarkt
PhysActivity	Binární	Indikuje, zda respondent vykonává fyzickou aktivitu
Veggies	Binární	Indikuje, zda respondent konzumuje zeleninu
HvyAlcoholConsump	Binární	Indikuje, zda respondent konzumuje těžký alkohol
AnyHealthcare	Binární	Indikuje, zda má respondent zdravotní péči
NoDocbcCost	Binární	Indikuje, zda respondent nemohl navštívit lékaře kvůli nákladům
GenHlth	Kategoriální	Obecné hodnocení zdravotního stavu
MentHlth	Kvantitativní	Počet dní s problémy s duševním zdravím za posledních 30 dní
PhysHlth	Kvantitativní	Počet dní s fyzickými zdravotními problémy za posledních 30 dní
DiffWalk	Binární	Indikuje, zda má respondent problémy s chůzí
Sex	Binární	Pohlaví respondenta (0 = muž, 1 = žena)
Age	Kategoriální	Věková skupina respondenta
Education	Kategoriální	Vzdělání respondenta
Income	Kategoriální	Příjmová skupina respondenta

Tabulka 2.3: Dataset 3 (Diabetes Health Indicators) – popis

$GVIF^{1/(2Df)}$ skóre nevycházelo u žádné z proměnných vyšší než 2, a tak nebyla vyřazena žádná proměnná.

Výběr relevantních proměnných pomocí stepwise regrese

Z datasetu byla vyřazena proměnná $Fruits$.

Finální dataset

Výsledný dataset obsahuje 253 680 pozorování a 21 proměnných.

2.1.5 Dataset 4: Framingham

Přehled dat

Čtvrtým datasetem je "Framingham Heart Study"[24]. Tento datový soubor zachycuje zdravotní stav účastníků dlouhodobé a průběžné kohortové kardiovaskulární studie, která byla zahájena v roce 1948. Studie má klíčový význam pro identifikaci rizikových faktorů kardiovaskulárních onemocnění.

Původní dataset zahrnoval 4240 pozorování a 16 proměnných. Jako vysvětlovanou proměnnou

byla použita proměnná *prevalentStroke*, která ukazuje, zda pacient prodělal mrtvici. Tato binární proměnná byla zvolena kvůli své extrémní nevyváženosti, s původním poměrem 4215:25 u tříd 0 a 1.

Popis proměnných

Seznam proměnných použitých v konečném datasetu je uveden v tabulce 2.4.

Proměnná	Datový typ	Popis
education	Ordinální	Úroveň vzdělání účastníka
cigsPerDay	Kvantitativní	Počet cigaret vykouřených za den účastníkem
BPMeds	Binární	Indikuje, zda účastník bere léky na krevní tlak
prevalentStroke	Binární	Indikuje, zda účastník prodělal mrtvici
prevalentHyp	Binární	Indikuje, zda účastník má hypertenzi
TenYearCHD	Binární	Indikuje, jestli se u účastníka vyvinula koronární srdeční choroba během 10 let

Tabulka 2.4: Dataset 4 (Framingham) – popis

Chybějící hodnoty

Proměnná *glucose* byla odstraněna vzhledem k více než 5 % chybějících hodnot. Následně byla odstraněna pozorování s chybějícími hodnotami u zbývajících proměnných.

Odstranění neplatných hodnot

Nebyly identifikovány žádné neplatné hodnoty, které by bylo třeba odstranit.

Testování a odstranění aliasovaných proměnných

Aliasované proměnné nebyly v tomto datasetu identifikovány.

Testování multikolinearity

Testování multikolinearity pomocí $GVIF^{1/(2Df)}$ skóre ukázalo, že žádná proměnná neměla hodnotu vyšší než 2, a proto nebyla vyřazena žádná další proměnná.

Výběr relevantních proměnných pomocí stepwise regrese

Z datasetu bylo vyřazeno 9 proměnných.

Finální dataset

Výsledný dataset se skládá z 3989 pozorování a 6 proměnných.

2.1.6 Dataset 5: SUPPORT2

Přehled dat

Pátým použitým datasetem je SUPPORT2 [25], který je umístěn v úložišti strojového učení UCI, obsahuje 9105 jednotlivých kriticky nemocných pacientů z pěti lékařských center v USA. Data byla shromážděna ve dvou fázích: Fáze I (1989-1991) a fáze II (1992-1994). Data z těchto dvou fází lze modelovat společně. Tento dataset je primárně zaměřen na vývoj a ověření prognostického modelu, který odhaduje přežití v průběhu 180 dnů u vážně nemocných hospitalizovaných dospělých osob.

Původní dataset obsahoval 9105 pozorování a 47 proměnných. Jako vysvětlovaná proměnná byla použita proměnná *diabetes*, která udává, zda má pacient diabetes mellitus. V původním datasetu byla tato binární proměnná, která nabývá hodnot 0 a 1, nevyvážená v poměru 7327:1778.

Popis proměnných

Seznam proměnných použitých v konečném datasetu je uveden v tabulce 2.5.

Proměnná	**Datový typ**	**Popis**
age	Kvantitativní	Věk pacienta
sex	Binární	Pohlaví pacienta
slos	Kvantitativní	Délka od vstupu do studie do propuštění
dzgroup	Kategoriální	Skupina onemocnění
num.co	Ordinální	Počet souběžných onemocnění
charges	Kvantitativní	Náklady na zdravotní péči
race	Kategoriální	Rasa pacienta
sps	Kvantitativní	Fyziologické skóre SUPPORT
aps	Kvantitativní	Fyziologické skóre APACHE III
diabetes	Binární	Indikátor přítomnosti diabetu
dementia	Binární	Indikátor přítomnosti demence
ca	Kategoriální	Status rakoviny
dnr	Kategoriální	Do Not Resuscitate status
hrt	Kvantitativní	Srdeční frekvence
resp	Kvantitativní	Respirační frekvence
temp	Kvantitativní	Tělesná teplota

Tabulka 2.5: Dataset 5 (SUPPORT2) – popis

Chybějící hodnoty

Chybějící hodnoty v proměnné *income* byly převedeny na novou kategorii "no data". Proměnné s více než 5 % chybějících hodnot byly odstraněny a následně byla odstraněna pozorování s chybějícími hodnotami u zbývajících proměnných. Vyřazeny byly především nové proměnné, které byly do datasetu přidány ve fázi II.

Odstranění neplatných hodnot

V datasetu nebyly přítomny žádné zjevné neplatné hodnoty, jež by bylo třeba odstranit.

Testování a odstranění aliasovaných proměnných

Aliasované proměnné, *dzclass* a *adlsc*, byly odstraněny.

Testování multikolinearity

Proměnné $surv2m$, $surv6m$ a *dnrday* byly postupně odstraněny kvůli vysokým hodnotám $GVIF^{1/(2Df)}$. Následně již žádná proměnná nepřekročila hodnotu 2.

Kombinování úrovní kategoriálních proměnných

Úrovně 8 a 9 v proměnné *num.co* byly sloučeny do úrovně 7 kvůli nízkému počtu hodnot.

Výběr relevantních proměnných pomocí stepwise regrese

Z datasetu bylo vyřazeno 11 proměnných.

Finální dataset

Konečný dataset obsahuje 16 proměnných a 8579 pozorování.

2.2 Transformace a úprava dat

V rámci empirické studie bylo provedeno několik datových transformací a byly aplikovány různé techniky pro vyvažování dat, aby bylo možné adekvátně porovnat účinnost klasifikačních modelů v podmínkách silně nevyvážených binárních klasifikačních úloh. Tato sekce popisuje konkrétní kroky a techniky, které byly použity při přípravě dat.

1. SMOTE (Synthetic Minority Over-sampling Technique)

Pro vyvážení tříd ve vysvětlované proměnné byla použita technika SMOTE, která synteticky vytváří nové příklady menšinové třídy na základě již existujících vzorků. Tento krok by měl pomoci zlepšit výkon klasifikačních algoritmů na nevyvážených datech. Tato technika byla implementována pomocí funkce $SMOTE$ z balíčku $smotefamily$. V rámci tohoto procesu bylo nutné provést několik kroků pro správné fungování metody:

1. **Encoding dat:** Kategoriální proměnné byly převedeny na numerické hodnoty pomocí one-hot encodingu. One-hot encoding převádí kategorické proměnné na binární (0 nebo 1) sloupce, což umožňuje správnou aplikaci metody SMOTE a dalších algoritmů strojového učení.
2. **Odstranění jedné kategorie:** Při použití one-hot encodingu byla odstraněna jedna kategorie od každé kategoriální proměnné, aby nedošlo k multikolinearitě. To znamená, že jedna z kategorií je reprezentována implicitně nulovými hodnotami ve všech příslušných binárních sloupcích.
3. **Kombinace SMOTE a undersamplingu:** Po aplikaci SMOTE byla menšinová třída synteticky zvětšena tak, aby její podíl ve vzorku byl větší než 50 %. Následně byl aplikován náhodný undersampling původní menšinové třídy, aby byly obě třídy vyvážené. K tomuto kroku bylo přistoupeno z důvodu, že hodnota dup_size, která u funkce $SMOTE$ z balíčku $smotefamily$ určuje maximální počet syntetických minoritních instancí vytvořených oproti originálním majoritním instancím v datové sadě, musí být u použité funkce diskrétní, což neumožňuje přesné vyvážení tříd u binární vysvětlované proměnné.

Tato kombinace SMOTE a undersamplingu zajišťuje, že výsledná datová sada je vyvážená, což umožňuje lepší trénink klasifikačních algoritmů a zvýšení jejich výkonu při práci s nevyváženými daty. Tento přístup doporučuje například Chawla. [26]

2. Undersampling

Undersampling snižuje počet vzorků většinové třídy náhodným odstraněním části těchto vzorků. V kódu bylo provedeno pomocí funkce *ovun.sample* z balíčku $ROSE$. Undersampling je aplikován tak, že výsledný dataset bude mít stejný počet vzorků pro každou třídu tím, že náhodně odstraní část vzorků většinové třídy, aby jejich počet odpovídal počtu vzorků menšinové třídy.

3. Oversampling

Oversampling zvyšuje počet vzorků menšinové třídy duplikováním existujících vzorků. V kódu bylo provedeno pomocí funkce *ovun.sample* z balíčku $ROSE$. Je zajištěno, že výsledný dataset bude mít stejný počet vzorků pro každou třídu tím, že náhodně duplikuje existující vzorky menšinové třídy, dokud jejich počet nedosáhne požadované úrovně.

4. Změny datových typů

Při zpracování dat byla zkontrolována konzistence datových typů všech proměnných. Kategorie byly převedeny na faktory a kvantitativní proměnné na numerické datové typy tam, kde to bylo potřeba. Pokud byl detekován nesoulad, proměnné byly převedeny na vhodné datové typy, což zajistilo správnou aplikaci statistických metod a algoritmů strojového učení. Tento krok byl nezbytný pro odstranění potenciálních chyb při dalším zpracování dat a modelování.

2.3 Rozdělení dat

Data byla rozdělena dat na trénovací a testovací sady, což je klíčové pro zajištění objektivního hodnocení modelu a zabránění overfittingu. Pokud by model byl hodnocen na stejných datech, na kterých byl trénován, mohl by dosahovat výborných výsledků, ale na nových datech by mohl selhávat.

1. Rozdělení 70:30

Všechny datasety byly rozděleny na trénovací a testovací sady s poměrem 70:30. Poměr 70:30 by měl mít dostatečný počet dat pro trénování modelu a zároveň dostatečný počet nezávislých dat pro jeho hodnocení, což poskytuje lepší odhad jejich výkonu na neznámých datech. Avšak u silně nevyvážených dat může tento poměr vést k problémům, kdy trénovací sada obsahuje stále výrazně méně příkladů z menšinové třídy, a proto u těchto datasetů použijeme i LOOCV.

2. Leave-One-Out Cross Validation (LOOCV)

Pro datasety se silně nevyváženými třídami (dataset 2: Myocardial Infarction Complication a dataset 4: Framingham) byla data trénována také pomocí Leave-One-Out Cross Validace (LOOCV). LOOCV je metoda, kde každé jednotlivé pozorování slouží jako testovací data, zatímco všechny ostatní pozorování slouží k trénování modelu. Proces se opakuje pro všechna pozorování. Tento přístup je obzvláště užitečný pro malé nebo extrémně nevyvážené datové sady, protože umožňuje plně využít všechna dostupná data.

2.4 Specifikace modelu

V této sekci jsou popsány modely a metody použité pro analýzu dat. Pro každý model byl specifikován proces trénování, ladění a vyhodnocení.

Trénování a testování modelů:

Každý z vybraných modelů použil všechny proměnné, které prošly základním zpracováním dat. Pro modely Logit, Random Forest a HDDT a pro každou techniku úpravy dat byly provedeny následující kroky:

1. Trénování modelu na trénovací sadě
2. Provedení predikcí na testovací sadě
3. Pro každý práh $threshold \in \{0.01, 0.02, ..., 0.99\}$ byly vypočteny precision, recall a F1 score a byly uloženy nejlepší výsledky (hodnoty prahu, precision, recall, F1 score a matice záměn)
4. Pro každý model a techniku úpravy dat byly vypočteny a zaznamenány následující metriky:
   - Nejlepší F1 score
   - Precision a recall pro nejlepší F1 score
   - Matice záměn pro nejlepší F1 score
   - ROC křivka
   - PR křivka
5. Pro dataset 2: Myocardial Infarction Complication, dataset 3: Diabetes Health Indicators a dataset 4: Framingham byl určen práh pro dělení tříd na základě kvantilu predikcí. K tomuto bylo přistoupeno kvůli výrazné nevyváženosti vysvětlovaných binárních proměnných a medikální povaze těchto datasetů, u kterých bývá primárním cílem minimalizovat false negatives. Predikce byly upraveny na základě určeného kvantilu a byly uloženy precision, recall, F1 score a matice záměn. Kvantily byly vybrány s přihlédnutím k velikosti nevyváženosti vysvětlované proměnné a byly následující:
   - Dataset 2: Myocardial Infarction Complication a 4: Framingham: 95% kvantil
   - Dataset 3: Diabetes Health Indicators: 85% kvantil
6. Na datasetech 2: Myocardial Infarction Complication a 4: Framingham (datasety s nejvíce nevyváženými vysvětlovanými proměnnými) byly modely trénovány pomocí LOOCV.
   - Pro každý záznam i v datasetu byly použity všechny ostatní záznamy jako trénovací sada $1,..., i-1, i+1, ..., n$ a aktuální záznam i jako testovací sada.
   - Modely byly trénovány na trénovací sadě a byly predikovány na testovacím pozorování.
   - Modely byly vyhodnoceny pomocí stejných metrik jako v kroku 4 (nejlepší F1 score, precision, recall, AUROC, AUPRC).

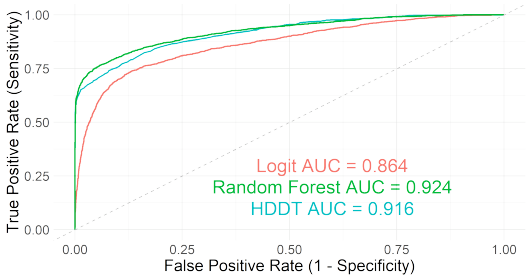
3. Výsledky

Výsledky jsou prezentovány v podobě tabulek a grafů, které ukazují výkonnost modelů na základě různých metrik, včetně přesnosti (Precision), citlivosti (Recall), F1 score, plochy pod ROC křivkou (AUROC) a plochy pod PR křivkou (AUPRC). Tabulky také obsahují informaci "Podíl třídy 0 u y", která značí podíl tříd u vysvětlované proměnné po vyvážení dat a o prahové hodnotě, pro které vychází nejlepší F1 score a z kterého byly výše zmiňované metriky získány.

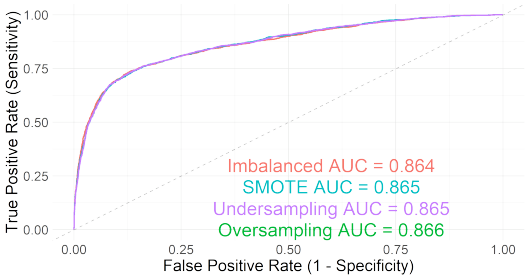
3.1 Dataset 1: Credit Risk Dataset

Metoda	**Data**	**Precision**	**Recall**	**F1**	**AUROC**	**AUPRC**	**Podíl třídy 0 u y**	**Práh**
Logit	Imbalanced	0.893	0.950	0.920	0.864	0.724	0.785	0.46
	SMOTE	0.887	0.952	0.918	0.865	0.716	0.500	0.77
	Undersampling	0.885	0.953	0.918	0.865	0.715	0.500	0.77
	Oversampling	0.876	0.966	0.919	0.866	0.719	0.500	0.81
Random Forest	Imbalanced	0.922	0.996	0.958	0.928	0.879	0.785	0.51
	SMOTE	0.920	0.996	0.956	0.928	0.876	0.500	0.65
	Undersampling	0.918	0.995	0.955	0.924	0.870	0.500	0.77
	Oversampling	0.921	0.995	0.956	0.927	0.875	0.500	0.64
HDDT	Imbalanced	0.916	0.998	0.955	0.921	0.864	0.785	0.89
	SMOTE	0.918	0.993	0.954	0.909	0.848	0.500	0.93
	Undersampling	0.916	0.995	0.954	0.915	0.853	0.500	0.94
	Oversampling	0.914	0.999	0.955	0.915	0.856	0.500	0.97

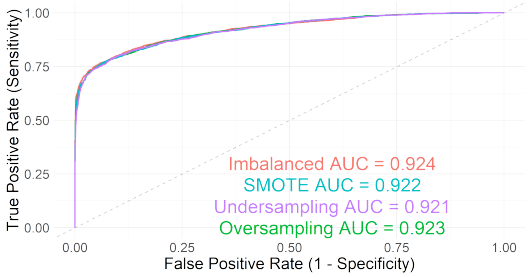
Tabulka 3.1: Dataset 1 (Credit Risk Dataset) – výsledky



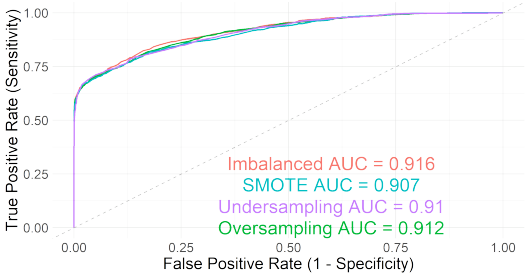
Srovnání modelů na nebalancovaných datech – ROC



Srovnání metod na vyvážení dat u Logitu – ROC



Srovnání metod na vyvážení dat u Random Forest – ROC



Srovnání metod na vyvážení dat u HDDT – ROC

Obrázek 3.1: Srovnání modelů a metod na vyvážení dat – dataset 1 (Credit Risk Dataset)

V tabulce 3.1 a na obrázku 3.1 můžeme vidět, že Logit model vykazuje jen mírné rozdíly ve výkonu napříč různými technikami úpravy dat. Výsledky ukazují, že techniky k vyvážení dat mírně zlepšily AUROC, na druhou stranu se snížily hodnoty AUPRC. K výraznému vylepšení však vyvážením dat nedošlo.

Random Forest model vykazuje vyšší výkon než Logit model napříč všemi metrikami. SMOTE, oversampling ani undersampling nevedou k vylepšení žádné z metrik a jejich výsledky jsou srovnatelné s výsledky z nebalancovaných dat.

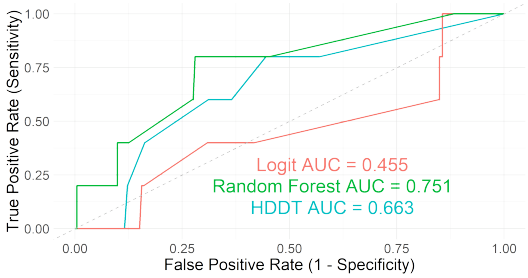
HDDT model dosahuje podobného výkonu jako Random Forest. HDDT model se jeví jako velmi robustní napříč různými technikami úpravy dat. Žádná z technik na balancování dat nevedla k vylepšení predikcí.

Na prvním datasetu, jehož výsledky jsou v tabulce 3.1, metody vyvážení dat nezlepšily výkonnost modelů. Nejlepší hodnoty většinou vycházely pro nevyvážený Random Forest. Výkonnost Logit a HDDT modelů zůstala relativně konzistentní, i když byly aplikovány metody na vyvážení dat. Nejvyššího výkonu dosahoval na tomto datasetu Random Forest.

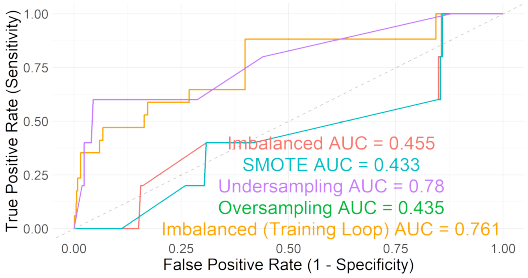
3.2 Dataset 2: Myocardial Infarction Complication

Metoda	Data	Precision	Recall	F1	AUROC	AUPRC	Podíl třídy 0 u y	Práh
Logit	Imbalanced	0.988	1.000	0.994	0.455	0.011	0.988	0.40
	SMOTE	0.988	1.000	0.994	0.433	0.010	0.500	0.99
	Undersampling	0.994	0.760	0.861	0.780	0.080	0.500	0.83
	Oversampling	0.988	1.000	0.994	0.435	0.010	0.500	0.99
	Imbalanced (Quantile 0.95)	0.988	0.951	0.969	0.455	0.011	0.988	0.06
	Imbalanced (Training Loop)	0.988	1.000	0.994	0.761	0.090	0.988	0.45
Random Forest	Imbalanced	0.988	1.000	0.994	0.751	0.060	0.988	0.29
	SMOTE	0.988	1.000	0.994	0.700	0.020	0.500	0.91
	Undersampling	0.991	1.000	0.995	0.715	0.231	0.500	0.80
	Oversampling	0.988	1.000	0.994	0.707	0.019	0.500	0.91
	Imbalanced (Quantile 0.95)	0.990	0.953	0.971	0.751	0.060	0.988	0.05
	Imbalanced (Training Loop)	0.988	1.000	0.994	0.736	0.050	0.988	0.34
HDDT	Imbalanced	0.988	1.000	0.994	0.663	0.018	0.988	0.67
	SMOTE	0.988	1.000	0.994	0.593	0.009	0.500	0.99
	Undersampling	0.988	1.000	0.994	0.604	0.016	0.500	0.86
	Oversampling	0.988	1.000	0.994	0.575	0.009	0.500	0.99
	Imbalanced (Quantile 0.95)	0.988	0.948	0.968	0.663	0.018	0.988	<0.01
	Imbalanced (Training Loop)	0.989	1.000	0.994	0.718	0.087	0.988	0.50

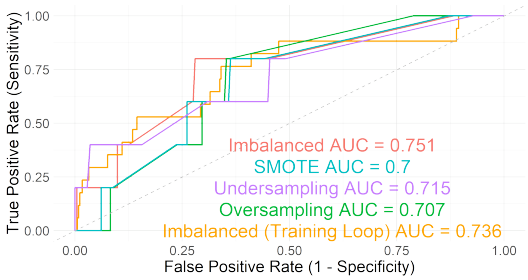
Tabulka 3.2: Dataset 2 (Myocardial Infarction Complication) – výsledky



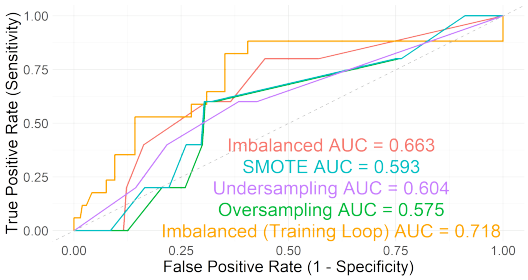
Srovnání modelů na nebalancovaných datech – ROC



Srovnání metod na vyvážení dat u Logitu – ROC



Srovnání metod na vyvážení dat u Random Forest – ROC



Srovnání metod na vyvážení dat u HDDT – ROC

Obrázek 3.2: Srovnání modelů a metod na vyvážení dat – dataset 2 (Myocardial Infarction Complication)

V tabulce 3.2 můžeme vidět, že Logit model vykazuje špatné výsledky a významné rozdíly ve výkonu napříč různými technikami úpravy dat. Dle hodnot AUROC můžeme vidět, že většina modelů klasifikuje hůře než náhodný klasifikátor. Jedinou technikou na vyvažování dat, která se zdá být účinnou, je undersampling, který v AUROC dosahuje dokonce nejlepší hodnoty napříč všemi metodami a modely u tohoto datasetu. Podobných výsledků u AUROC a AUPRC dosahuje model trénovaný a testovaný pomocí LOOCV, což můžeme vidět i na obrázku 3.2. Nízká výkonnost Logitu může být způsobena lineární povahou modelu nebo overfittingem.

Random Forest model vykazuje vyšší výkon než Logit model napříč většinou metrik a všechny modely tentokrát klasifikují výrazně lépe než náhodný klasifikátor. Můžeme vidět, že model používající undersampling dosahoval nejlepší hodnoty AUPRC. To může být způsobeno tím, že Random Forest může efektivněji využít náhodnost v podmnožinách dat k vytvoření rozmanitějších stromů nebo snížením rizika overfittingu na minoritní třídu. Použití LOOCV hodnotu AUROC ani AUPRC vůči nevyváženému datasetu nevylepšilo. Ostatní metody nepřinesly zlepšení.

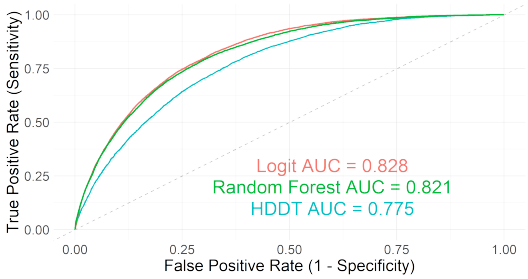
HDDT model dosahuje špatných výkonů napříč většinou technik na úpravu dat. Nicméně narozdíl od Logitu všechny modely klasifikují lépe než náhodný klasifikátor. SMOTE, undersampling ani oversampling nevedl u HDDT k zlepšení. Nejlepších výkonů dosahoval HDDT při použití LOOCV. Nízká výkonnost HDDT u SMOTE a oversamplingu by mohla být způsobena velkým šumem, který narušují schopnost HDDT správně rozpoznat vzory v datech. Nízká výkonnost u undersamplingu by mohla být pravděpodobně způsobena nedostatkem dat.

Na datasetu 2 metody vyvážení dat přinesly různé výsledky pro různé modely. Jedinou metodou na balancování dat, která se zdála být účinnou byl u Logitu i Random Forestu undersampling. Žádná z ostatních technik na balancování dat výsledky výrazně nevylepšila. LOOCV vedlo u Logitu a HDDT k zlepšení, což by mohlo nastiňovat to, že by na silně nevyvážených datasetech mohl být za určitých podmínek účinný. Logit klasifikoval podobně jako náhodný klasifikátor a nejlepším modelem na tomto datasetu je Random Forest.

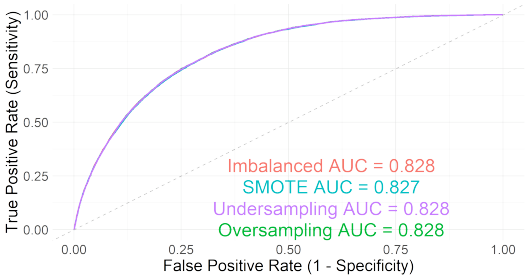
3.3 Dataset 3: Diabetes Health Indicators

Metoda	Data	Precision	Recall	F1	AUROC	AUPRC	Podíl třídy 0 u y	Práh
Logit	Imbalanced	0.873	0.987	0.927	0.828	0.420	0.861	0.54
	SMOTE	0.875	0.985	0.927	0.827	0.417	0.500	0.88
	Undersampling	0.875	0.985	0.927	0.828	0.418	0.500	0.88
	Oversampling	0.876	0.984	0.926	0.828	0.418	0.500	0.88
	Imbalanced (Quantile 0.85)	0.913	0.902	0.907	0.828	0.420	0.861	0.30
Random Forest	Imbalanced	0.873	0.988	0.927	0.821	0.422	0.861	0.53
	SMOTE	0.864	0.997	0.926	0.813	0.390	0.500	0.83
	Undersampling	0.872	0.989	0.926	0.824	0.415	0.500	0.87
	Oversampling	0.864	0.997	0.926	0.814	0.393	0.500	0.82
	Imbalanced (Quantile 0.85)	0.912	0.901	0.906	0.821	0.422	0.861	0.31
HDDT	Imbalanced	0.864	0.997	0.926	0.775	0.345	0.861	0.81
	SMOTE	0.861	1.000	0.925	0.753	0.308	0.500	0.99
	Undersampling	0.863	0.997	0.925	0.783	0.349	0.500	0.97
	Oversampling	0.861	1.000	0.925	0.748	0.309	0.500	0.99
	Imbalanced (Quantile 0.85)	0.900	0.889	0.895	0.775	0.345	0.861	0.35

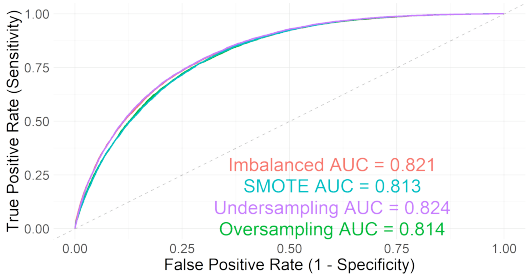
Tabulka 3.3: Dataset 3 (Diabetes Health Indicators) – výsledky



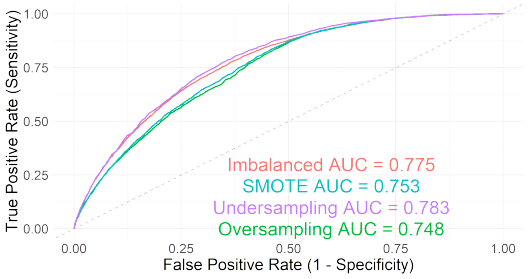
Srovnání modelů na nebalancovaných datech – ROC



Srovnání metod na vyvážení dat u Logitu – ROC



Srovnání metod na vyvážení dat u Random Forest – ROC



Srovnání metod na vyvážení dat u HDDT – ROC

Obrázek 3.3: Srovnání modelů a metod na vyvážení dat – dataset 3 (Diabetes Health Indicators)

V tabulce 3.3 můžeme vidět, že Logit model vykazuje jen mírné rozdíly ve výkonu napříč různými technikami úpravy dat. Na obrázku 3.3 můžeme vidět, že ROC křivky napříč technikami na vyvážení dat dokonce splývají. Výsledky ukazují, že metody k vyvážení dat (SMOTE, undersampling a oversampling) nevedou k výrazným změnám ve výkonnosti modelu. Hodnoty precision, recall a F1 score zůstávají relativně konstantní. Hodnota AUPRC se může zdát nízká, ale vzhledem k nevyváženosti datasetu ji lze interpretovat jako vysokou. Použití kvantilu 0.85 pro stanovení prahu zlepšilo precision, ale zároveň výrazně snížilo recall a o něco méně výrazně snížilo F1 score. Zvýšení precision však v určitých situacích může být důležitější než to, že se výrazně sníží recall.

Random Forest model vykazuje podobný výkon jako Logit model. Výsledky naznačují, že metody SMOTE a oversampling nevedou k výrazným zlepšením. Undersampling mírně zvýšil hodnotu AUROC. Použití kvantilu 0.85 pro stanovení prahu vedlo k podobným výsledkům jako model Logit.

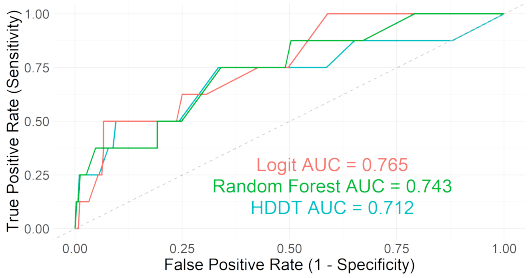
HDDT model dosahuje nižšího výkonu než Random Forest a Logit. AUROC a AUPRC hodnoty jsou výrazně nižší napříč všemi technikami na balancování dat, což naznačuje větší problémy s odlišením tříd. Žádná z technik na balancování dat nepomohla k lepším výsledkům. Použití kvantilu 0.85 pro stanovení prahu vedlo k podobným výsledkům jako u dvou předchozích modelů. Horší výkonnost HDDT může být způsobena tím, že kvůli velikosti tohoto datasetu byl u HDDT z důvodu výpočetní náročnosti nastaven vyšší minimální počet vzorků potřebných pro rozdělení uzlu než u ostatních datasetů. To může vést k tomu, že některé komplení vztahy v datech mohou být přehlíženy, což vede k nižší přesnosti modelu.

Na datasetu 3 metody vyvážení dat nepřinesly výrazné zlepšení pro žádný z modelů. Stanovení prahu pomocí kvantilu vedlo k určitému zlepšení precision, ale zároveň k výraznému snížení recall. Při specifických úlohách však může být tento postup užitečný. Nejlepších výsledků na tomto datasetu dosahoval Logit.

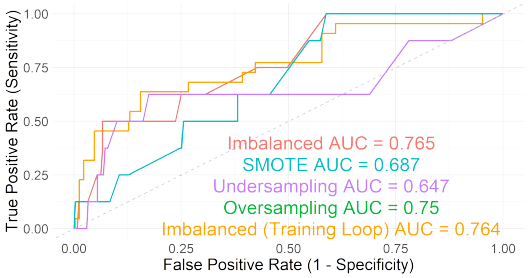
3.4 Dataset 4: Framingham

Metoda	Data	Precision	Recall	F1	AUROC	AUPRC	Podíl třídy 0 u y	Práh
Logit	Imbalanced	0.993	1.000	0.997	0.765	0.024	0.995	0.16
	SMOTE	0.993	1.000	0.997	0.687	0.035	0.500	0.94
	Undersampling	0.994	0.967	0.980	0.647	0.017	0.500	0.92
	Oversampling	0.993	1.000	0.997	0.750	0.021	0.500	0.99
	Imbalanced (Quantile 0.95)	0.994	0.968	0.981	0.765	0.024	0.995	0.02
	Imbalanced (Training Loop)	0.994	1.000	0.997	0.764	0.031	0.994	0.15
Random Forest	Imbalanced	0.993	1.000	0.997	0.743	0.054	0.995	0.13
	SMOTE	0.993	1.000	0.997	0.718	0.042	0.500	0.95
	Undersampling	0.993	1.000	0.997	0.664	0.044	0.500	0.93
	Oversampling	0.993	1.000	0.997	0.767	0.059	0.500	0.96
	Imbalanced (Quantile 0.95)	0.996	0.952	0.973	0.743	0.054	0.995	0.03
	Imbalanced (Training Loop)	0.994	1.000	0.997	0.735	0.035	0.994	0.18
HDDT	Imbalanced	0.993	1.000	0.997	0.712	0.050	0.995	0.50
	SMOTE	0.993	1.000	0.997	0.708	0.052	0.500	0.95
	Undersampling	0.993	1.000	0.997	0.626	0.011	0.500	0.88
	Oversampling	0.994	0.996	0.995	0.731	0.062	0.500	0.98
	Imbalanced (Quantile 0.95)	0.995	0.957	0.976	0.712	0.050	0.995	<0.01
	Imbalanced (Training Loop)	0.994	1.000	0.997	0.708	0.030	0.994	0.67

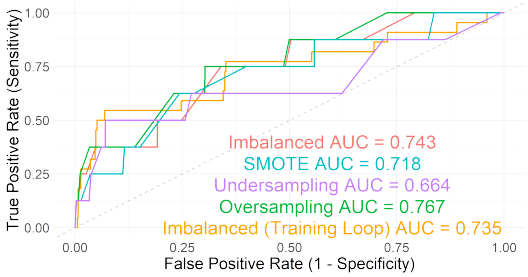
Tabulka 3.4: Dataset 4 (Framingham) – výsledky



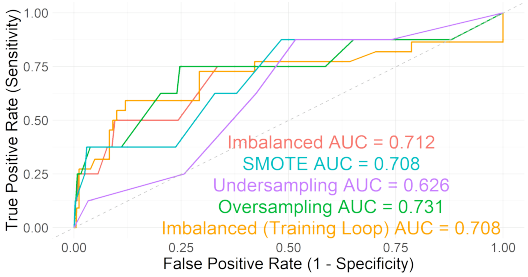
Srovnání modelů na nebalancovaných datech – ROC



Srovnání metod na vyvážení dat u Logitu – ROC



Srovnání metod na vyvážení dat u Random Forest – ROC



Srovnání metod na vyvážení dat u HDDT – ROC

Obrázek 3.4: Srovnání modelů a metod na vyvážení dat – dataset 4 (Framingham)

V tabulce 3.4 můžeme vidět, že Logit model vykazuje významné rozdíly ve výkonu napříč různými technikami úpravy dat a dosahuje nižšího výkonu než Random Forest a HDDT. Při použití technik na vyvážení dat došlo většinou ke snížení výkonnosti modelu.

Random Forest model vykazuje podobnou výkonnost jako Logit. Stejně jako u Logitu techniky na vyvážení dat většinou snížily výkonnost modelu, avšak oversampling vedl k mírnému zlepšení. Stanovení prahu pomocí kvantilu 0.95 zlepšilo precision jen nepatrně. Použití LOOCV nevedlo ke zlepšení.

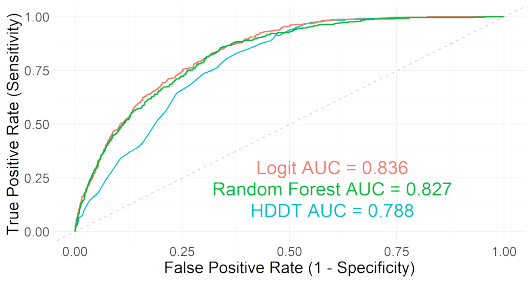
HDDT model dosahuje o něco slabších výkonů než ostatní modely, což můžeme vidět například na ROC křivce pro nebalancovaná data na obrázku 3.4. Použití technik k balancování dat nevedlo k výraznému zlepšení výkonnosti modelu, ale oversampling podobně jako u Random Forestu mírně předčil nebalancovaná data. Na HDDT byla jako u jediného modelu nepříliš efektivní LOOCV, což by mohlo být způsobeno nastavením vyššího minimálního počtu vzorků potřebných pro rozdělení uzlu z důvodu výpočetní náročnosti.

Na datasetu 4 metody vyvážení dat dosáhly různých výsledků. Zatímco undersampling dosahoval vždy neuspokojivých výsledků, tak naopak oversampling vedl u Random Forestu a HDDT k nepatrnému zlepšení. Stanovení prahu pomocí kvantilu vedlo jen k mírnému zlepšení precision. U HDDT nebylo efektivní použití LOOCV. Nejvýkonnějším na tomto datasetu byl těsně před Logitem Random Forest.

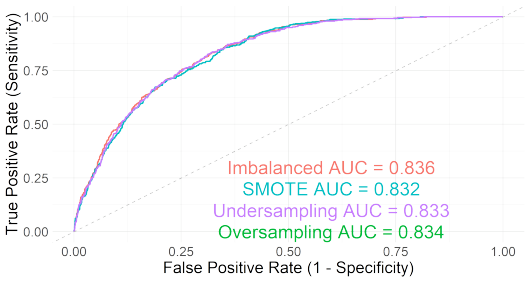
3.5 Dataset 5: SUPPORT2

Metoda	**Data**	**Precision**	**Recall**	**F1**	**AUROC**	**AUPRC**	**Podíl třídy 0 u y**	**Práh**
Logit	Imbalanced	0.828	0.984	0.899	0.836	0.527	0.805	0.65
	SMOTE	0.828	0.978	0.897	0.832	0.513	0.500	0.90
	Undersampling	0.828	0.981	0.898	0.833	0.514	0.500	0.89
	Oversampling	0.828	0.981	0.898	0.834	0.517	0.500	0.89
Random Forest	Imbalanced	0.838	0.968	0.898	0.827	0.516	0.805	0.49
	SMOTE	0.821	0.986	0.896	0.822	0.488	0.500	0.78
	Undersampling	0.833	0.976	0.899	0.834	0.514	0.500	0.78
	Oversampling	0.829	0.979	0.898	0.834	0.520	0.500	0.64
HDDT	Imbalanced	0.814	0.991	0.894	0.788	0.414	0.805	0.92
	SMOTE	0.806	0.999	0.892	0.780	0.410	0.500	0.99
	Undersampling	0.805	1.000	0.892	0.772	0.401	0.500	0.98
	Oversampling	0.811	0.992	0.892	0.790	0.420	0.500	0.98

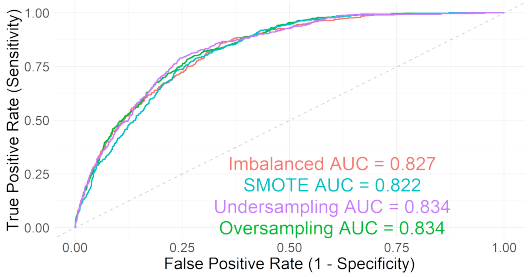
Tabulka 3.5: Dataset 5 (SUPPORT2) – výsledky



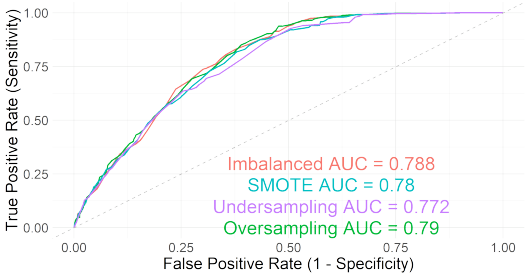
Srovnání modelů na nebalancovaných datech – ROC



Srovnání metod na vyvážení dat u Logitu – ROC



Srovnání metod na vyvážení dat u Random Forest – ROC



Srovnání metod na vyvážení dat u HDDT – ROC

Obrázek 3.5: Srovnání modelů a metod na vyvážení dat – dataset 5 (SUPPORT2)

V tabulce 3.5 můžeme vidět, že Logit model vykazuje jen mírné rozdíly ve výkonu napříč různými technikami úpravy dat. Výsledky ukazují, že metody k vyvážení dat nevedou k výrazným změnám ve výkonnosti modelu. Vyvážení dat nepřineslo zlepšení oproti nevyváženým datům.

Random Forest model má podobný výkon jako Logit. Použití technik (konkrétně undersamplingu a oversamplingu) na balancování dat vedlo tentokrát k mírnému zlepšení výkonnosti modelu. Na nebalancovaných datech vycházel nejvyšší precision, na druhou stranu vycházel nižší recall, což vedlo k nižším hodnotám AUROC a AUPRC.

HDDT model dosahuje nižšího výkonu než Logit a Random Forest. Horší výkonnost HDDT modelu můžeme vidět i na obrázku 3.5 na příkladu pro srovnání modelů na nebalancovaných datech. Použití techniky SMOTE a undersamplingu vedlo k horším výsledkům u většiny metrik. Oversampling vedl k nepatrnému zlepšení AUROC a AUPRC vůči nevyváženým datům.

Na datasetu 5 metody vyvážení dat přinesly mírné zlepšení jen pro Random Forest. U Logitu vycházely nejlepší výsledky pro nevyvážená data. Nejvýkonnějším modelem na tomto datasetu byl Logit.

4. Diskuze

Výsledky této práce poskytly přehled o výkonnosti jednotlivých modelů (Logit, Random Forest a Hellinger Distance Decision Tree) na několika různých datasetech s nevyváženou binární vysvětlovanou proměnnou. Tyto výsledky byly hodnoceny pomocí různých metrik, včetně Precision, Recall, F1 score, AUROC a AUPRC. Analyzovány byly také různé techniky vyvážení dat (SMOTE, undersampling, oversampling), a dále bylo testováno, zda použití Leave-One-Out Cross-Validation (LOOCV) vedlo ke zlepšení výkonnosti modelů a zda stanovení prahu pomocí kvantilu vedlo k rozumnému zvýšení precision.

Výsledky naznačují, že různé modely a techniky vyvážení dat mají odlišný dopad na výkonnost v závislosti na charakteristikách datasetu. Z tabulky 4.1 je zřejmé, že model Random Forest se ukázal jako nejrobustnější napříč většinou datasetů, zejména při použití nevyvážených dat. HDDT, který byl navržen pro práci s nevyváženými daty, dosáhl smíšených výsledků, což naznačuje, že jeho efektivita může být omezena specifickými vlastnostmi dat a nemusí být vhodný pro použití na určitých typech datasetů.

Dataset	**Model**	**Data**	**Precision**	**Recall**	**F1**	**AUROC**	**AUPRC**	**Podíl třídy 0 u y**
1: Credit Risk	Random Forest	Imbalanced	0.922	0.996	0.958	0.928	0.879	0.785
2: Myocardial Infarction	Random Forest	Undersampling	0.991	1.000	0.995	0.715	0.231	0.500
3: Diabetes Health	Logit	Imbalanced	0.873	0.987	0.927	0.828	0.420	0.861
4: Framingham	Random Forest	Oversampling	0.993	1.000	0.997	0.767	0.059	0.500
5: SUPPORT2	Logit	Imbalanced	0.828	0.984	0.899	0.836	0.527	0.805

Tabulka 4.1: Shrnutí nejlepších výsledků napříč datasety

Výsledky této práce potvrzují význam volby správného modelu a jeho optimalizace při práci s nevyváženými datasety. Zatímco techniky vyvážení dat (SMOTE, undersampling, oversampling) mohou v některých případech přinést zlepšení, často se ukazuje, že správně zvolený a optimalizovaný model dokáže efektivně pracovat s nevyváženými daty bez potřeby těchto technik. U silně nevyvážených datasetů se použití Leave-One-Out Cross-Validation (LOOCV) na trénování a testování dat většinou ukázalo jako užitečné.

Logit vykazoval konzistentní výsledky napříč většinou datasetů. Nejlepších výsledků však bylo většinou dosaženo na méně nevyvážených datech, což naznačuje, že jednoduchost a robustnost logistické regrese mohou být výhodou při práci s mírně nevyváženými daty.

Random Forest se ukázal jako nejvýkonnější napříč většinou datasetů, zejména na nevyvážených datech. Jeho flexibilita a schopnost adaptace na různá data často eliminují potřebu speciálních technik vyvážení dat.

HDDT dosáhl smíšených výsledků. Přestože je navržen pro práci s nevyváženými daty, jeho výkon byl mnohdy horší než u Logitu a Random Forestu, což může být způsobeno specifickými omezeními tohoto modelu při práci s různými typy dat.

Techniky vyvážení dat měly různorodý dopad v závislosti na modelu a datasetu. Celkově však nebyly mnohdy účinné nebo nevedly k výraznému zlepšení výkonu. Největší zlepšení bylo pozorováno u Random Forest modelu při použití undersamplingu na datasetu 2: Myocardial Infarction Complication, kdy jeho použití výrazně zvýšilo AUPRC, která je u silně nevyvážených dat důležitá. Tento výsledek naznačuje, že správná technika může v určitých případech výrazně zlepšit výkon modelu, avšak volba správného modelu a jeho nastavení je klíčová.

Použití Leave-One-Out Cross-Validation (LOOCV) na trénování a testování se ukázalo být jako užitečné pro práci se silně nevyváženými daty. Tento přístup umožňuje maximální využití dostupných dat pro trénink, což je zvláště důležité v případě nevyvážených datasetů, kde každé pozorování může být klíčové. LOOCV přinesla zlepšení zejména u modelů Logit a HDDT.

Stanovení prahu pomocí kvantilu vedlo k mírnému zlepšení precision, ale často za cenu výrazného snížení recall. Zvýšení precision může být ale při specifických případech užitečné a důležitější i přes výrazné snížení recall.

Některé modely, jako například HDDT, jsou citlivé na nastavení parametrů, což může ovlivnit jejich výkonnost. Tato citlivost znamená, že pečlivé ladění parametrů je nezbytné pro dosažení optimálních výsledků. Výsledky také závisí na specifikách datasetů, což znamená, že dosažené výsledky nemusí být generalizovatelné na všechny typy nevyvážených dat. Dalším významným omezením je výpočetní náročnost některých metod, jako je LOOCV, což omezuje jejich praktické použití u velkých datasetů. Tato metoda, ačkoli může přinést významná zlepšení, vyžaduje značné výpočetní zdroje, což může být překážkou při práci s rozsáhlými daty.

Další výzkum by se mohl zaměřit na optimalizaci parametrů modelů, zejména u HDDT, aby se zlepšila jejich výkonnost na různých datasetech. Testování dalších typů datasetů by mohlo pomoci ověřit generalizovatelnost výsledků a identifikovat specifické situace, kde jsou určité techniky vyvážení dat nejúčinnější. Vhodné na práci s nevyváženými daty by mohly být neuronové sítě, což zmiňoval například Priya [27].

Závěr

Výsledky této diplomové práce poskytly přehled o výkonnostních charakteristikách různých klasifikačních modelů (Logit, Random Forest a Hellinger Distance Decision Tree) na několika datasetech s nevyváženou binární vysvětlovanou proměnnou. Výkonnost byla hodnocena pomocí metrik jako Precision, Recall, F1 score, AUROC a AUPRC. Byly analyzovány různé techniky vyvážení dat (SMOTE, undersampling a oversampling) a bylo testováno použití Leave-One-Out Cross-Validation (LOOCV) a stanovení prahu pomocí kvantilu.

Hlavní zjištění ukazují, že model Random Forest se ukázal jako nejrobustnější napříč většinou datasetů, zejména při použití nevyvážených dat. HDDT dosáhl dobré výkonnosti na vysoce nevyvážených datasetech, ale na méně nevyvážených datasetech měl smíšené výsledky. Techniky vyvážení dat měly různorodý dopad na výkonnost modelů a často nevedly k významnému zlepšení. LOOCV se ukázala jako užitečný nástroj při práci na nevyvážených datech, zatímco stanovení prahu pomocí kvantilu vedlo k mírnému zlepšení precision, ale za cenu výrazného snížení recall, což však ve specifických sitaucích, například ve zdravotnictví, nemusí příliš vadit.

Další výzkum by se měl zaměřit na optimalizaci parametrů modelů, zejména u HDDT, aby se zlepšila jejich výkonnost na různých datasetech. Testování dalších typů datasetů by mohlo pomoci ověřit přenositelnost výsledků a identifikovat specifické situace, kde jsou určité techniky vyvážení dat nejúčinnější. Zkoumání kombinace různých technik vyvážení dat a modelů by mohlo vést k nalezení optimálních přístupů pro různé typy nevyvážených dat. Navíc by mohl být výzkum rozšířen o pokročilé metody, jako jsou neuronové sítě, které se ukazují jako velmi účinné pro práci s nevyváženými daty díky své schopnosti automaticky extrahovat relevantní rysy z dat.

Tato diplomová práce potvrzuje význam správného výběru a optimalizace modelů při práci s nevyváženými daty. Výsledky ukázaly, že techniky vyvážení dat mohou být v určitých situacích užitečné, ale často není nutné je aplikovat, pokud je zvolen správný model a jeho nastavení. Random Forest se ukázal jako nejvýkonnější model napříč většinou datasetů. Logit a především HDDT vykazovaly smíšené výsledky, což zdůrazňuje důležitost správného nastavení parametrů. Dále se potvrdila užitečnost metod jako LOOCV. Celkově tato práce přispívá k lepšímu pochopení problematiky práce s nevyváženými daty a nabízí směry pro další výzkum v této oblasti.

Bibliografie

1. ARAF, Imane, IDRI, Ali a CHAIRI, Ikram.
   Cost-sensitive learning for imbalanced medical data: a review.
   Artificial Intelligence Review. 2024, roč. 57, s. 1–72. Dostupné také z:
   https://link.springer.com/article/10.1007/s10462-023-10652-8.
   Accessed: 2024-06-15.
2. CIESLAK, David A., HOENS, Tom R., CHAWLA, Nitesh V. a
   KEGELMEYER, William P.
   Hellinger distance decision trees are robust and skew-insensitive. 2012.
   Dostupné také z: https://www.researchgate.net/publication/220451886_
   Hellinger_distance_decision_trees_are_robust_and_skew-insensitive.
3. GOORBERGH, Ruben van den, SMEDEN, Maarten van, TIMMERMAN, Dirk a
   VAN CALSTER, Ben. The harm of class imbalance corrections for risk prediction
   models: illustration and simulation using logistic regression.
   Journal of the American Medical Informatics Association. 2022, roč. 29, č. 11.
   Dostupné z DOI: 10.1093/jamia/ocac093.
4. KLEINBAUM, David G. a KLEIN, Mitchel. *Logistic Regression: A Self-Learning Text*.
   3rd edition. Springer, 2010. ISBN 978-1-4419-1741-6.
   Dostupné také z: https://dmrocke.ucdavis.edu/Class/EPI204-Spring-
   2021/2010_Book_LogisticRegression.pdf.
5. HOSMER, David W., LEMESHOW, Stanley a STURDIVANT, Rodney X.
   Applied Logistic Regression. Third. Hoboken, New Jersey: John Wiley & Sons, 2013.
   ISBN 978-0-470-58247-3. Dostupné z DOI: 10.1002/9781118548387.
6. ALSUBAYHIN, Abdulrahman, RAMZAN, Muhammad Sher a ALZAHRANI, Bander.
   Crime Prediction Model using Three Classification Techniques: Random Forest, Logistic
   Regression, and LightGBM.
   International Journal of Advanced Computer Science and Applications.
   2024, roč. 15, č. 1, s. 240–251. Dostupné z DOI: 10.14569/IJACSA.2024.0150123.
7. BREIMAN, Leo. Random Forests. *Machine Learning*. 2001, roč. 45, č. 1, s. 5–32.
   Dostupné z DOI: 10.1023/A:1010933404324.
8. HERRERA, Victor M., KHOSHGOFTAAR, Taghi M., VILLANUSTRE, Flavio a
   FURHT, Borko. Random forest implementation and optimization for Big Data analytics
   on LexisNexis's high performance computing cluster platform. *Journal of Big Data*.
   2019, roč. 6, č. 68. Dostupné z DOI: 10.1186/s40537-019-0232-1.
9. CHEN, Wuxing, YANG, Kaixiang, YU, Zhiwen, SHI, Yifan a CHEN, C. L. Philip.
   A survey on imbalanced learning: latest research, applications and future directions.
   Artificial Intelligence Review. 2024, roč. 57.
   Dostupné z DOI: 10.1007/s10462-024-10759-6.

10. MAYO, Matthew. *Tips for Handling Imbalanced Data in Machine Learning*. 2024. Dostupné také z: https://machinelearningmastery.com/tips-handling-imbalanced-data-machine-learning/. Accessed: 2024-06-25.

11. BROWNLEE, Jason. *LOOCV for Evaluating Machine Learning Algorithms*. 2020. Dostupné také z: https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/. Accessed: 2024-06-25.

12. AGGARWAL, Charu C. *Data Mining: The Textbook*. Springer, 2015. Dostupné také z: https://github.com/gawainxu/books/blob/master/Data%20Mining_%20The%20Textbook%20%5BAggarwal%202015-04-14%5D.pdf.

13. CHEN, Jinying, LALOR, John, LIU, Weisong et al. Detecting Hypoglycemia Incidents Reported in Patients' Secure Messages: Using Cost-Sensitive Learning and Oversampling to Reduce Data Imbalance (Preprint). *Journal of Medical Internet Research*. 2018, roč. 21, č. 3. Dostupné z DOI: 10.2196/11990. License: CC BY 4.0.

14. MOHAMMED, Roweida, RAWASHDEH, Jumanah a ABDULLAH, Malak. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. Jordan University of Science a Technology, 2020. Dostupné z DOI: 10.1109/ICICS49469.2020.239556.

15. GOUTTE, Cyril a GAUSSIER, Eric. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: *Advances in Information Retrieval*. 2005, s. 345–359. Dostupné z DOI: 10.1007/978-3-540-31865-1_25.

16. DRAELOS, Rachel. *Measuring Performance: AUC & AUROC*. 2019. Dostupné také z: https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/. Accessed: 2024-06-23.

17. SAITO, Taku a REHMSMEIER, Marc. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. 2015, roč. 10, č. 3, e0118432. Dostupné z DOI: 10.1371/journal.pone.0118432.

18. DRAELOS, Rachel. *Measuring Performance: AUPRC*. 2019. Dostupné také z: https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/. Accessed: 2024-06-23.

19. KHADKA, Nirajan. Stepwise Regression: How It Works and When to Use It. *Data Aspirant*. 2023. Dostupné také z: https://dataaspirant.com/stepwise-regression/. Accessed: 2024-06-26.

20. LAOTSE. *Credit Risk Dataset*. 2023. Dostupné také z: https://www.kaggle.com/datasets/laotse/credit-risk-dataset. Accessed: 2024-06-16.

21. ZHANG, Tianqi, TIAN, Yu, LIU, Yan, LI, Jia, HUANG, Xiaohong a YU, Xiaofei. A comprehensive review of deep learning-based image segmentation techniques. *Computational Intelligence and Neuroscience*. 2023. Dostupné z DOI: 10.1155/2023/10064664. Accessed: 2024-06-17.

22. REPOSITORY, UCI Machine Learning. *Myocardial Infarction Complications Data Set*. 2021. Dostupné také z: https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications. Accessed: 2024-06-17.

23. TEBOUL, Alex. *Diabetes Health Indicators Dataset*. 2021. Dostupné také z: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Accessed: 2024-06-17.

24. BHARDWAJ, Ashish. *Framingham Heart Study Dataset*. 2021. Dostupné také z: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset. Accessed: 2024-06-17.

25. HARRELL, Frank. *SUPPORT2* [UCI Machine Learning Repository]. 2023. Accessed: 2024-06-19.

26. CHAWLA, Nitesh V., BOWYER, Kevin W., HALL, Lawrence O. a KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002, roč. 16, s. 321–357. Dostupné z DOI: 10.1613/jair.953.

27. PRIYA, S. a UTHRA, R. Annie. Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data. *Complex & Intelligent Systems*. 2021, roč. 7, č. 5, s. 3499–3510. Dostupné z DOI: 10.1007/s40747-021-00456-0.