FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Vliv počasí na spotřebu tankového piva
Student:	Marek Jirman
Vedoucí:	Ing. Jan Martínek
Studijní program:	Informatika
Obor / specializace:	Umělá inteligence 2021
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2024/2025

Pokyny pro vypracování

Zadavatel disponuje daty o výtočích tankového piva jednoho z předních výrobců piva v ČR. Tato data jsou uložena v datovém skladu zadavatele. Cílem této práce je zpracovat prediktivní model spotřeby tankového piva zahrnující vhodný zdroj předpovědi počasí v krátkodobém horizontu.

1) Seznamte se s problematikou lokální předpovědi počasí pro týdenní a měsíční horizont v oblastech definovaných zadavatelem. Proveďte rešerši dostupných dat předpovědi počasí – zahrňte otevřená i placená data.
2) Seznamte se s daty dostupnými v datovém skladu zadavatele o výtočích tankového piva.
3) S využitím dat z bodu 1) a 2) navrhněte alespoň dva prediktivní modely predikující spotřebu tankového piva na základě dat o předpovědi počasí. (konkrétní úkoly budou specifikovány po diskuzi s vedoucím).
4) Spolehlivost modelů ověřte na základě historických dat a získané výsledky interpretujte.
5) V případě uspokojivé spolehlivosti prediktivních modelů řešení automatizujte.

Elektronicky schválil/a Ing. Magda Friedjungová, Ph.D. dne 16. listopadu 2023 v Praze.

Bakalářská práce

VLIV POČASÍ NA SPOTŘEBU TANKOVÉHO PIVA

Marek Jirman

Fakulta informačních technologií
Katedra aplikované matematiky
Vedoucí: Ing. Martínek Jan
16. května 2024

Obsah

Seznam obrázků

Seznam tabulek

Chtěl bych poděkovat především vedoucímu panu Ing. Janu Martínkovi za vedení této práce. Dále bych chtěl poděkovat panu Mgr. Petru Šimánkovi za pomoc ohledně dat počasí a při tvorbě modelů. V neposlední řadě bych rád poděkoval rodině a přátelům za podporu při psaní této práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací. Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 citovaného zákona.

V Praze dne 16. května 2024

Abstrakt

Tato bakalářská práce se zaměřuje na nalezení spojitosti mezi počasím a spotřebou tankového piva. Literární část práce se zabývá předpovědí počasí, popsáním podobných prací a představením použitých metod a modelů. Praktická část se zabývá předzpracováním dat, trénováním regresních a klasifikačních modelů nad různými datasety a predikcí výtoče piva.

Při trénování bylo dosaženo nejlepších výsledků u modelů, které jsou trénovány na jednotlivých hospodách, to ukazuje, že výtoč se u každé hospody velice liší. S pomocí nejlepších modelů byla provedena predikce na předpovědi počasí a na historických datech o počasí. U historických dat o počasí bylo navíc provedeno porovnání s reálnými daty. Tyto predikce umožní pivovaru přibližně odhadovat velikost výtoče dle předpovědi počasí a tím optimalizovat dodávku piva do hospod.

Klíčová slova umělá inteligence, strojové učení, datová analýza, počasí, spotřeba piva, CatBoost, Python

Abstract

This bachelor thesis focuses on finding connection between weather and tank beer consumption. The Literary part of the thesis deals with weather forecast, description of similar works and the presentation of used methods and models. The practical part deals with data preprocessing, training of regression and classification models over different datasets and predicting beer outlet.

When training, the best results were achieved with models that are trained on individual pubs, this shows that the outlet varies greatly from pub to pub. Using the best models, predictions were made on weather forecasts and historical weather data. For the historical weather data, there was also made a comparison with real data. These predictions allow the brewery to approximately estimate the size of the outlet according to the weather forecast and optimize the beer supply to the pubs.

Keywords artificial intelligence, machine learning, data analysis, weather, beer consumption, CatBoost, Python

Seznam zkratek

GBM	Gradient Boosting Machine
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
IF	Isolation Forest
IT	Isolation Tree
BST	Binary Search Tree

Úvod

Pivo je jedním z nejoblíbenějších alkoholických nápojů Čechů. Dle statistického úřadu za rok 2022 každý Čech průměrně vypil 142,9 litrů piva. Toto číslo se každoročně zvyšuje. Předpovídání spotřeby v daných regionech na základě počasí by mohlo být velice prospěšné pro výrobce piva. Ti by díky tomu mohli optimalizovat dovoz piva do daných hospod.[1]

Výsledek práce by měl být schopen spotřebu tankového piva predikovat. Nyní spotřebu piva predikují hlavně dle spotřeby z minulých týdnů či měsíců. Často se stává, že buď na dané období dodají piva málo nebo příliš moc.

Práci jsem si zvolil z několika důvodu. Téma mi přijde velice zajímavé, protože rád zajdu s přáteli posedět na pivo a zajímá mě, při jakém počasí chodí lidé nejvíc do hospody. Dále si myslím, že práce s počasím mi přinese užitečné zkušenosti, které se můžou v budoucnu hodit. V neposlední řadě se mi líbí, že má výsledná práce bude mít reálné využití v praxi.

Na začátku práce začnu úvodem do problematiky předpovědi počasí a představením různých webů poskytující data o předpovědi počasí. Provedu rešerši již podobných prací na téma vlivu počasí na spotřebu jak alkoholických nápojů, tak nealkoholických.

Následuje teoretické seznámení s použitými metodami a modely, které při práci použiji. Seznámím se s daty o počasí a o výtoči piva. U dat o počasí vyberu nejlepší možné příznaky, které by mohly mít na spotřebu piva největší vliv, např. déšť, teplotu, sílu větru či oblačnost. Data předzpracuji pro jednodušší práci a budu trénovat a testovat různé modely na historických datech o počasí a datech o výtoči. Vyzkouším různé hyperparametry modelů a různé možnosti rozdělení dat, vše povede k nalezení modelu s největší přesností.

Následuje extrakce dat o předpovědi počasí z vybraného webu a aplikace nejlepšího modelu. Predikce počasí bude maximálně v horizontu příštích 14 dní, protože predikce na více dní dopředu již není tak přesná. Nakonec následuje dodání softwaru do pivovaru a jejich případné využití v praxi.

Cíle

Hlavním cílem je nalezení spojitosti mezi počasím a spotřebou piva pro pivovar poskytující data o výtoči tankového piva, toho bude docíleno pomocí trénování datasetu o výtoči na historických datech o počasí. Dílčím cílem bude ověření spolehlivosti modelů a predikce výtoči tankového piva na reálných datech o předpovědi počasí.

Kapitola 1

Úvod do problematiky předpovědi počasí

V kapitole je představen úvod do problematiky předpovědi počasí.

1.1 Co je to předpověď počasí

„*Meteorologická předpověď je fyzikální úloha, jejíž cílem je vytvořit nejpravděpodobnější scénář nebo scénáře budoucího vývoje atmosféry v časovém horizontu typicky několika hodin, dnů až týdnů, maximálně měsíců.*“[2]

1.2 Rozdělení

Existují 3 základní typy rozdělení předpovědi:

1. podle období, na které je vydána,
   a. velmi krátkodobá – 0-1 den,
   b. krátkodobá – 1-3 dny (1-2 dny),
   c. střednědobá – 3-15 dnů (2-15 dnů),
   d. dlouhodobá – měsíční,
   e. klimatická – většinou desetiletí,
2. podle účelu,
   a. všeobecná – pro veřejnost,
   b. speciální – pro specializované uživatele,
3. podle místa.
   a. oblastní – pro administrativní území,

b. liniová – pro dopravu,

c. místní – pro určitou lokalitu.[2]

1.3 Historie

Předpověď počasí je téma, kterým se lidé zabývají přes tisíce let. Již kolem roku 650 př. n. l. se Babyloňané pomocí mraků a dalších optických fenoménů snažili předpovědět počasí. V roce 340 př. n. l. zkoumal Aristotelés jak by mohl vysvětlit chování počasí, napsal o tom pojednání Meteorologica.[3]

Důležitým milníkem je až polovina 15. století, kdy Nicholas Cusa poprvé popsal hygrometr, který měřil vlhkost vzduchu. Následuje Galileo Galilei s vynálezem teploměru v roce 1592 a Evangelista Torricelli s vynálezem tlakoměru v roce 1643.[3]

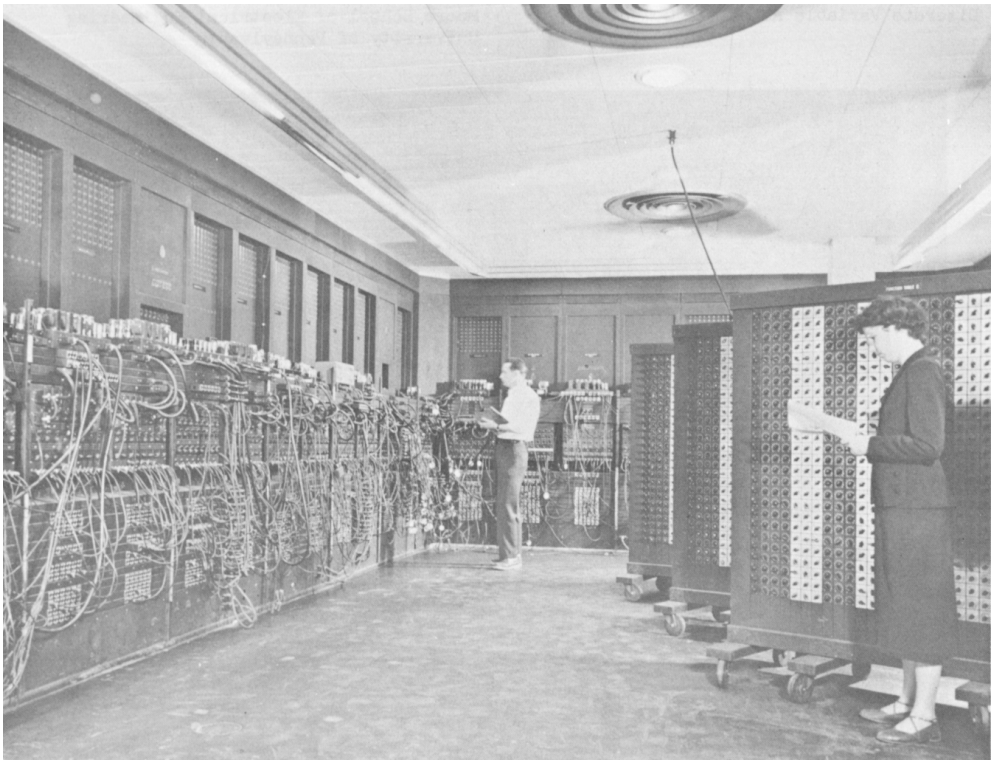
Velkým zlomovým bodem je vynález telegrafu a telegrafních sítí. Tento vynález umožnil rychlý převod pozorování o počasí. Zrodila se synoptická předpověď počasí.[3]

Na začátku 20. století meteorologové Cleveland Abbe z USA a Vilhelm Bjerknes z Norska přišli s nápadem řešení hydrodynamických a termodynamických rovnic a vytvořit tím objektivní předpověď. V roce 1922 Lewis Fry Richardson přišel s metodou předpovídání vývoje atmosféry numericky pomocí metody grafické integrace, která byla kvůli chybě Richardsona neúspěšná. Tyto výpočetní modely spadají do kategorie NWP[1]. Rozvoj těchto modelů přichází až s vývojem počítačů v půlce 20. Na obrázku 1.1 je vidět jeden z těchto počítačů Eniac.[2]

1.4 Předpověď počasí dnes

V dnešní době se předpověď počasí rozlišuje hlavně na krátkodobou a dlouhodobou. V krátkodobé předpovědi se používají nejvíce radary a satelity, které pozorují lokální atmosférické podmínky. V dlouhodobé předpovědi se meteorologové spíše přiklánějí ke klimatologickým metodám.[4]

[1] Numerical Weather Prediction - numerická předpověď počasí



Obrázek 1.1 Počítač pro předpověď počasí Eniac.[5]

1.5 Data o počasí

V této sekci jsou představeny různé weby poskytující předpověď počasí. Každá podsekce se jmenuje podle názvu poskytovatele.

1.5.1 OpenWeather

OpenWeather je firma sídlící v Londýně, poskytující data od roku 2014. Stránka nabízí 3 verze. První základní verze je bezplatná a umožňuje 1000 API volání denně, další API volání jsou zpoplatněné za 0.0014 Eura za jedno volání. Nabízí historická data. Placené verze nabízejí více volání API a další funkce.[6]

Výhody:

- obsahuje data z celého světa,
- poskytuje základní informace o počasí, nabízí současnou předpověď, předpověď na hodinu, na 48 hodin a na 8 dní,
- nabízí stahovat data pomocí API,

- bezplatná a lehce použitelná.

Nevýhody:

- neobsahuje žádné významné nevýhody.[6]

1.5.2 visualcrossing

Visualcrossing je jedním z předních poskytovatelů předpovědi počasí, který funguje již od roku 2003. Stránka nabízí několik verzí. Základní bezplatná verze umožňuje stahovat 1000 záznamů denně a umožňuje stahovat až 50 let stará historická data o předpovědi počasí a hlavně stahovat data o předpovědi počasí na 15 dní. Placené verze umožňují stahovat denně větší počet záznamů, souběžné stahování a další funkce.[7]

Výhody:

- obsahuje data z celého světa,
- poskytuje velký počet informací o počasí denně, či každou hodinu na 15 dní dopředu,
- nabízí různé formáty dat, např. JSON, CSV či excel a jak bylo zmíněno výše, data lze také stahovat pomocí API,
- bezplatná a lehce použitelná.

Nevýhody:

- neobsahuje žádné významné nevýhody.[7]

1.5.3 ECMWF

ECMWF[2] je výzkumný institut a také poskytovatel dat o předpovědi počasí založen v roce 1975. ECMWF poskytuje bezplatně velké množství datasetů kolem předpovědi počasí, historické předpovědi počasí a mnoho dalšího.[8]

Výhody:

- obsahuje velký počet informací o počasí,

Nevýhody:

- neintuitivnost stáhnutí dat,
- neumožňuje výběr lokace před stáhnutím,
- neobsahuje hodinové předpovědi.[8]

[2]European Centre for Medium-Range Weather Forecasts

Kapitola 2
Přehled podobných prací

V kapitole jsou představeny práce zabývající se vlivem počasí na konzumaci alkoholických a nealkoholických nápojů. Každá sekce odpovídá jedné práci. Název sekce je vždy shodný s názvem článku.

2.1 Impact of weather changes on consumption of beverages in the hospitality industry

Výzkum proveden v roce 2013 na téma nalezení vzájemné souvislosti počasí a spotřeby alkoholických i nealkoholických nápojů v pohostinství v Chorvatsku. Tato souvislost pomůže manažerům firem v pohostinství předpovídat úroveň poptávky turistů po nápojích. Článek se zaměřuje i na změnu počasí na pobřeží Jaderského moře. Chorvatsko je země, která je hodně závislá na turismu v sezonním období, tedy hlavně v létě, což může hrát v tomto výzkumu velikou roli.[9]

Autoři článku pracují s daty z let 2002–2012 během letní sezony, přesněji od 15. června do 15. září. Data byla zpostředkována firmami v pohostinství v destinacích u Jaderského moře: Poreč, Opatija, Mali Lošinj, Zadar, Šibenik, Split, Hvar a Dubrovnik.[9]

Při hledání závislosti autoři práce použili regresní analýzu. Regresní analýza se snaží najít vztah mezi závislou proměnnou a nezávislými proměnnými. Jejich hypotéza zní takto: „*Množství spotřebovaných nápojů se zvedá při vyšším čísle strávených nocí a teploty vzduchu.*“ Pro ověření hypotézy autoři použili tuto regresní funkci

$$y = 0,264x_1 + 735,7x_2, \tag{2.1}$$

kde y je počet nápojů, x_1 je počet realizovaných přespání a x_2 je průměrná teplota vzduchu.[9, 10]

„*Testy ukázaly, že použité nezávislé proměnné jsou významné na úrovni 0,05 a významnost funkce je testována F-testem.*“ V závěru článku je potvrzeno, že turisté konzumují nápoje, když jsou atmosférické podmínky příznivé, dalším

výsledkem práce bylo, že poptávka nápojů roste při vyšších teplotách a nižší vlhkosti vzduchu.[9]

2.2 Colder weather and fewer sunlight hours increase alcohol consumption and alcoholic cirrhosis worldwide

Článek zaměřen na vliv počasí na spotřebu alkoholu, a jeho vlivu na cirhózu jater. Jedná se o celosvětový výzkum proveden v roce 2018. Práce je rozdělena na část zaměřující se na celý svět a část zaměřující se pouze na Spojené státy americké.[11]

Autoři využili databázi Globální informační systém o alkoholu a zdraví organizace WHO[1] z roku 2014.[11]

Pro získání korelace mezi klimatickými proměnnými a proměnnými o spotřebě alkoholu byly použity korelační testy (Spearmanovo ρ). Spearmanův ρ je neparametrický test měřící sílu a směr asociace mezi dvěma hodnocenými proměnnými. [11, 12]

$$\rho_s = 1 - \frac{6 \sum_i^n d_i^2}{n(n^2 - 1)}, \tag{2.2}$$

kde ρ_s je Spearmanův ρ, d_i jsou rozdíly mezi dvěma úrovněmi každého pozorování a n je počet pozorování.

Hodnota $\rho_s = 1$ znamená dokonalou pozitivní korelaci a hodnota ρ_s = -1 znamená dokonalou negativní korelaci.[11, 12]

V části zaměřující se na celý svět výsledky výpočtů naznačují, že chladnější počasí a méně hodin slunečního svitu nepřímo korelují s vyšší konzumací alkoholu. Dále ukázali souvislost mezi cirhózou jater a spotřebou alkoholu. Část o Spojených státech amerických ukázala stejné výsledky a jenom utvrdila tvrzení ohledně souvislosti mezi cirhózou jater a spotřebou alkoholu. [11]

2.3 The impact of systematic changes in weather on the supply and demand of beverages

Článek o účinku systematických změn v počasí na nabídku a poptávku nápojů. Proveden v Spojených státech amerických pro 52 hlavních trhových zón. Zaměřuje se na krátkodobé i dlouhodobé změny počasí.[13]

Data o počasí byla získána z Národního klimatického datového centra NOAA[2], pokrývají každý měsíc v období 10 let. Data související s kategorií

[1] World Health Organization - Světová zdravotnická organizace

[2] National Oceanic and Atmospheric Administration - Národní úřad pro oceán a atmosféru

LRB[3] byla získána z prodejů těchto nápojů.[13]

Autoři použili při výzkumu dvoukrokový ekonometrický model. Tento model se často používá v makro-ekonomice a ve financích. Výsledkem práce bylo poukázání na fakt, že prodej LRB se meziročně zvyšuje přibližně o 0,21 %, kvůli rostoucího teplotě. Odhady autorů také indikují, že během vln veder se zvyšuje poptávku asi o 2,1 %, za každý stupeň Fahrenheita a během vln chladu klesá poptávka pouze o 0,4 %, za každý stupeň Fahrenheita.[13]

[3]Liquid refreshment beverages - tekuté osvěžující nápoje

Kapitola 3
Teoretická část

V kapitole je představena teoretická část práce. Popisuje metody a modely použité v praktické části.

3.1 CatBoost

CatBoost je algoritmus strojového učení založen na gradient boosting machine (GBM). GMB je technika strojového učení postavená na základě posilování (boosting). Posilování je postupné konstruování nových modelů, kde každý model je závislý na tom minulém. CatBoost při trénování postupně vytváří množinu rozhodovacích stromů. Každý následující strom je konstruován se sníženou ztrátou ve srovnání s předchozími stromy. CatBoost se dá použít pro řešení problémů regrese i klasifikace.[14]

Počáteční parametry určují počet stromů a hloubku stromů. Pro zabránění přeučení se může použít detektor přeučení, který zastaví stavění stromů. Přeučení modelu je nadměrné přizpůsobení se trénovacím datům a nemožnost předpovídat data neznámá.[14]

Před zahájením trénování je provedena kvantizace (quantization) pro každý numerický příznak, což je rozdělení hodnot objektů do disjunktích rozsahů (bucketů) dle hraničních hodnot (splitů). Kvantizace slouží k výběru stromové struktury a také při práci s kategorickými příznaky. CatBoost byl speciálně navržen pro práci s kategoriálními příznaky. Pokud se v datech vyskytují nějaké kategorické příznaky CatBoost je dokáže automaticky převádět na numerické. Trénování je ovlivněno boostrapem, který pomáhá zabránit přeučení a zrychluje konstrukci stromu. Bootstrap je používání výběru s opakováním.[14, 15]

	CatBoost		LightGBM		XGBoost		H2O	
	Tuned	Default	Tuned	Default	Tuned	Default	Tuned	Default
Adult	**0.26974**	0.27298 +1.21%	0.27602 +2.33%	0.28716 +6.46%	0.27542 +2.11%	0.28009 +3.84%	0.27510 +1.99%	0.27607 +2.35%
Amazon	**0.13772**	0.13811 +0.29%	0.16360 +18.80%	0.16716 +21.38%	0.16327 +18.56%	0.16536 +20.07%	0.16264 +18.10%	0.16950 +23.08%
Click prediction	**0.39090**	0.39112 +0.06%	0.39633 +1.39%	0.39749 +1.69%	0.39624 +1.37%	0.39764 +1.73%	0.39759 +1.72%	0.39785 +1.78%
KDD appetency	0.07151	**0.07138** −0.19%	0.07179 +0.40%	0.07482 +4.63%	0.07176 +0.35%	0.07466 +4.41%	0.07246 +1.33%	0.07355 +2.86%
KDD churn	**0.23129**	0.23193 +0.28%	0.23205 +0.33%	0.23565 +1.89%	0.23312 +0.80%	0.23369 +1.04%	0.23275 +0.64%	0.23287 +0.69%
KDD internet	**0.20875**	0.22021 +5.49%	0.22315 +6.90%	0.23627 +13.19%	0.22532 +7.94%	0.23468 +12.43%	0.22209 +6.40%	0.24023 +15.09%
KDD upselling	**0.16613**	0.16674 +0.37%	0.16682 +0.42%	0.17107 +2.98%	0.16632 +0.12%	0.16873 +1.57%	0.16824 +1.28%	0.16981 +2.22%
KDD 98	**0.19467**	0.19479 +0.07%	0.19576 +0.56%	0.19837 +1.91%	0.19568 +0.52%	0.19795 +1.69%	0.19539 +0.37%	0.19607 +0.72%
Kick prediction	**0.28479**	0.28491 +0.05%	0.29566 +3.82%	0.29877 +4.91%	0.29465 +3.47%	0.29816 +4.70%	0.29481 +3.52%	0.29635 +4.06%

Obrázek 3.1 Porovnání výkonnosti algoritmů s CatBoostem.[14]

3.2 Evaluace chyby u regresních modelů

K ověření přesnosti regresních úloh se používají většinou nezáporné funkce, které se nazývají ztrátové funkce $L : \mathbb{R}^2 \to \mathbb{R}$. Pomocí těchto funkcí se ověří chyba modelu, čím je chyba menší tím je model lepší.[16]

3.2.1 Kvadratická ztrátová funkce

Kvadratická ztrátová funkce je jednou ze základních ztrátových funkcí

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2, \tag{3.1}$$

kde Y je vysvětlovaná proměnná a $\hat{Y}$ je její predikce.[16]

3.2.2 Střední kvadratická chyba

Střední kvadratická chyba (MSE) je další ztrátovou funkcí, jedná se o jednu z nejpoužívanějších metrik pro chybovost modelu. Nevýhodou MSE je, že není ve stejných jednotkách jako vysvětlovaná proměnná.

$$\text{MSE} = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2, \tag{3.2}$$

kde n je počet pozorování, y_i je i-té pozorování vysvětlované proměnné a $\hat{y}_i$ je korespondující predikce vysvětlované proměnné.[17]

Podobnou ztrátovou funkcí je odmocnina ze střední kvadratické chyby (RMSE). Výhodou RMSE je, že má stejnou jednotku jako vysvětlovaná proměnná, tím je porovnání jednodušší.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2}, \tag{3.3}$$

kde n je počet pozorování, y_i je i-té pozorování vysvětlované proměnné a $\hat{y}_i$ je korespondující predikce vysvětlované proměnné.[17]

3.2.3 Střední absolutní chyba

Střední absolutní chyba (MAE) je další velice používaná metrika. Výhodou MAE, že je v stejných jednotkách jako vysvětlovaná proměnná, stejně jako u RMSE.

$$\text{MAE} = \frac{1}{n}\sum_{i}^{n}|y_i - \hat{y}_i| \tag{3.4}$$

kde n je počet pozorování, y_i je i-té pozorování vysvětlované proměnné a $\hat{y}_i$ je korespondující predikce vysvětlované proměnné.[17]

3.2.4 Koeficient determinace

Koeficient determinace (R^2) je metrika, která udává jak dobře natrénovaný model odhaduje reálná data. R^2 může mít hodnotu od 0 do 1, kdy 0 je nejhorší přesnost a 1 je nejlepší přesnost.

$$R^2 = 1 - \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{\sum_{i}^{n}(y_i - \overline{y})^2} \tag{3.5}$$

kde y_i je i-té pozorování vysvětlované proměnné, $\hat{y}_i$ je korespondující predikce vysvětlované proměnné a $\overline{y_i}$ je průměr vysvětlované proměnné.[18]

3.3 Evaluace chyby u klasifikačních modelů

Nejčastější volbou pro měření přesnosti u klasifikační úlohy je accuracy (přesnost), která se vypočítá velmi jednoduše a přímočaře

$$accuracy = \frac{\text{počet správně klasifikovaných dat}}{\text{počet všech dat}}.[19] \tag{3.6}$$

3.4 Isolation Forest

Isolation Forest (IF) je algoritmus pro identifikaci a odstranění anomálií. Algoritmus IF funguje jinak než většina existujících metod. Běžné metody se snaží profilovat normální instance a dle nich určit anomálie. Tyto metody jsou optimalizovány na hledání normálních instancí, to může vést k špatnému určení anomálií (můžeme určit správné instance jako anomálie nebo určit příliš málo anomálií), dále jsou omezeny na menší a nízko dimenzionální datasety. Následující definice jsou z [20].

▶ **Definice 3.1.** ***Isolation Tree (IT)***. *Let T be a node of an isolation tree. T is either an external-node with no child, or an internal-node with one test and exactly two daughter nodes (T_l,T_r). A test consists of an attribute q and a split value p such that the test $q < p$ divides data points into T_l and T_r.*

▶ **Definice 3.2.** ***Path Length*** *$h(x)$ of a point x is measured by the number of edges x traverses an IT from the root node until the traversal is terminated at an external node.*

IF využívá toho, že anomálií je málo a jsou odlišné. Vytváří soubor (ensemble) IT a dle průměrné délky cesty v IT určí, které instance jsou anomáliemi. Za anomálie jsou označovány instance s krátkou průměrnou délkou. Toto vede k efektivnímu algoritmu s lineární časovou složitostí a nízkou paměťovou složitostí.[20]

3.4.1 Anomaly score

Skóre anomálie je potřebné pro každou metodu, která se snaží identifikovat anomálie. Toto skóre je výstupem algoritmu IF a dle určené hranice uživatelem určuje, zda je daná instance anomálie. K výpočtu se díky stejné struktuře IT a binárního vyhledávacího stromu (BST), využije průměrné délky neúspěšných vyhledávání v BST

$$c(n) = 2H(n-1) - (2(n-1)/n), \tag{3.7}$$

kde $H(i)$ je harmonické číslo a lze jej odhadnout pomocí $ln(i) + 0,5772156649$ (Eulerovo konstanta) a n je velikost datasetu.[20]

Skóre anomálie je poté definováno jako

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \tag{3.8}$$

kde x je řádek datasetu, n je velikost datasetu a $E(h(x))$ je průměr $h(x)$ v souboru IT.[20]

Výsledek skóre se dá rozdělit na 3 závěry:

- instance vrací s velmi blízko k 1, potom jsou určitě anomáliemi,
- instance vrací s mnohem menší než 0,5, potom mohou bezpečně být označeny za normální,
- všechny instance vrací s kolem 0,5, potom celý vzorek nemá žádné výrazné anomálie.[20]

3.4.2 Algoritmus

Algoritmus IF se dělí na 2 fáze, na trénovací fázi a vyhodnocovací fázi. V trénovací fázi je konstruováno, na základě hyperparametru t, určený počet IT rekurzivním rozdělením trénovací sady, dokud nejsou instance izolovány nebo dokud není dosaženo konkrétní výšky stromu. Trénovací část přímo popisují algoritmy 1 a 2 z [20].

Ve vyhodnocovací fázi je z očekávané délky cesty $E(h(x))$ pro každou instanci odvozeno skóre anomálie s. Pro nalezení hlavních m anomálií stačí seřadit data pomocí s v sestupném pořadí.[20]

Algorithm 1 *iForest*(X,t,ψ)

Inputs: X - input data, t - number of trees, ψ - sub-sampling size
Output: a set of t *iTrees*

1: **Initialize** *Forest*
2: set height limit $l = ceiling(\log_2 \psi)$
3: **for** $i = 1$ to t **do**
4: $X' \leftarrow sample(X, \psi)$
5: $Forest \leftarrow Forest \cup iTree\ (X', 0, l)$
6: **end for**
7: **return** *Forest*

Algorithm 2 *iTree*(X, e, l)

Inputs: X - input data, e - current tree height, l - height limit
Output: an $iTree$

if $e \geq l$ or $|X| \leq 1$ **then**
    return $exNode\{Size \leftarrow |X|\}$
else
    let Q be a list of attributes in X
    randomly select an attribute $q \in Q$
    randomly select a split point p from max and min values of attribute q in X
    $X_l \leftarrow filter(X, q < p)$
    $X_r \leftarrow filter(X, q \geq p)$
    return $inNode\{Left \leftarrow iTree(X_l, e+1, l),$
                $Right \leftarrow iTree(X_r, e+1, l),$
                $SplitAtt \leftarrow q,$
                $SplitValue \leftarrow p\}$
end if

Algorithm 3 *PathLength*(x, T, e)

Inputs: x - an instance, T - an iTree, e - current path length; to be initialized to zero when first called
Output: a path length of x

if T is an external node **then**
    return $e + c(T.size)$ {$c(.)$ is defined in Equation 3.7}
end if
$a \leftarrow T.splitAtt$
if $x_a < T.splitValue$ **then**
    return $PathLength(x, T.left, e+1)$
else {$x_a \geq T.splitValue$}
    return $PathLength(x, T.right, e+1)$
end if

3.5 SMOTE

SMOTE je metoda na vyvážení datasetů. Dataset je nevyvážený, pokud klasifikační kategorie nejsou zastoupeny rovnoměrně. Často se totiž stává, že reálné datasety obsahují více „normálních“ záznamů a méně „abnormálních“ záznamů. Metoda SMOTE tedy méně zastoupené kategorie doplní umělými (syntetickými) vzorky, jedná se o takzvaný over–sampling.[21]

SMOTE pracuje s přistupem over–samplingu, při kterém je menšinová třída over–sampled vytvořením syntetických záznamů. Over–sampling menšinové třídy probíhá tak, že pro každý vzorek menšinové třídy se zavedou syntetické příklady podél úseček spojujících libovolných/všech k nejbližších sousedů menšinové třídy. Náhodně se vybere k nejbližších sousedů dle požadované velikosti over–samplingu. Implementace může použít až pět nejbližších sousedů, záleží na velikosti over–samplingu, např. když je požadovaná velikost over–samplingu 200 %, jsou vybráni dva sousedé z pěti nejbližších sousedů a vzorek je generován ve směru každého z nich. Algoritmus je podrobně popsán v 4 a 5.[21]

Algorithm 4 *SMOTE*(T, N, k)

Inputs: Number of minority class samples T, Amount of SMOTE N%, Number of nearest neighbors k

Output: $(N/100) * T$ synthetic minority class samples

(* *If N is less than 100 %, randomize the minority class samples as only a randompercent of them will be SMOTEd..**)
if $N < 100$ **then**
    Randomize the T minority class samples
    $T = (N/100) * T$
    $N = 100$
end if
$N = (int)(N/100)$ (∗ *The amount of SMOTE is assumed to be in integral multiples of 100.* ∗)
k = Number of nearest neighbors
$numattrs$ = Number of attributes
$Sample[][]$: array for original minority class samples
$newindex$: keeps a count of number of synthetic samples generated, initialized to 0
$Synthetic[][]$: array for synthetic samples
(* *Compute k nearest neighbors for each minority class sample only.* *)
for $i = 1$ to T **do**
    Compute k nearest neighbors for i, and save the indices in the $nnarray$
    Populate($N, i, nnaray$)
end for

Algorithm 5 *Populate*(N, i, $nnarray$)

(* *Function to generate the synthetic samples.* *)
while $N \neq 0$ **do**
    Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
    for $attr = 1$ to $numattrs$ **do**
        Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        Compute: $gap =$ random number between 0 and 1
        $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
    end for
    $newindex++$
    $N = N - 1$
end while
return (* *End of Populate.* *)

Kapitola 4
Použité technologie

V kapitole jsou představeny všechny podstatné knihovny použité v implementaci a jejich verze[1]. Celá implementace je naprogramována v jazyce Python, který se pro strojové učení používá nejčastěji.

- `catboost` (1.2.3) je knihovna obsahující algoritmus CatBoost.[22]
- `cfgrib` (0.9.11.0) je prostředí pro načtení souborů grib.[23]
- `imbalanced-learn` (0.12.2) je knihovna pro práci s nevybalancovanými datasety.[24]
- `jupyter` (1.0.0) je knihovna pro práci s programovacími jazyky v interaktivním webovém prostředí.[25]
- `matplotlib` (3.7.5) je knihovna pro vizualizaci dat.[26]
- `numpy` (1.26.4) je knihovna pro vědecké výpočty. Doplňuje Python o práci s vektory, maticemi, vícerozměrnými poli a dalšími matematickými funkcemi.[27]
- `pandas` (2.1.4) je knihovna pro práci s daty a datovou analýzou.[28]
- `seaborn` (0.13.2) je knihovna pro vizualizaci dat.[29]
- `scikit-learn` (1.4.2) je knihovna pro prediktivní analýzu dat.[30]
- `scipy` (1.11.4) je knihovna pro vědecké výpočty. Poskytuje spoustu algoritmů na optimalizaci dat.[31]

[1] Čísla v závorkách odpovídají verzi knihovny

Kapitola 5
Data

V kapitole jsou popsány použité datasety v experimentální části. Dále je popsáno vytvoření finálního datasetu spojením dat o výtoči a dat o počasí a jeho následné předzpracování před použitím modelů.

5.1 Data o výtoči

Data o výtoči obsahují 182172 záznamů, ty jsou rozděleny mezi 41 zákazníků. V datech se také vyskytují nerelevantní záznamy, které je potřeba odstranit.

5.1.1 Popis příznaků

V datasetu se vyskytuje 10 příznaků.

1. `ZAKAZNIK_ID` uvádí ID zákazníka.
2. `ZAKAZNI_JMENO` uvádí jméno zákazníka.
3. `GPS_X` a `GPS_Y` uvádí přibližnou geografickou polohu zákazníka pomocí souřadnic. `GPS_X` odpovídá zeměpisné délce a `GPS_Y` odpovídá zeměpisné šířce. Souřadnice jsou anonymizovány.
4. `DATUM` uvádí datum výtoče. Datum je ve tvaru DD.MM.YYYY.
5. `BRFAM_AG` uvádí typ piva.
6. `HOUR_OF_DAY` uvádí hodinu výtoče.
7. `VYTOC_HL` uvádí hodnotu výtoče v daný den a hodinu v hektolitrech.
8. `DELTA_MIN` uvádí časovou prodlevu od poslední změny počítadla v minutách a tedy čas za který byla výtoč naměřena. Počítadlo se resetuje při odeslání dat od zákazníka na server.
9. `MAPY_CZ_LINK` uvadí odkaz na mapy ukazující odpovídající souřadnice.

5.2 Historická data o počasí

K natrénování modelu je potřeba co nejpřesnější historická předpověď v požadovaných místech a příznaky počasí, které budou mít na výtoč tankového piva největší vliv. Data byla získána z [32]. Tato data byla vybrána z několika důvodů:

1. Data jsou ukládána ve čtvercích o straně 0.1°, což odpovídá přibližně 9 km. Zaměření na požadované hospody bude tedy velice přesné a data budou co nejvíce odpovídat realitě.
2. Data z [32] jsou měřena od roku 1950 do dnes, každý den a každou hodinu, to je potřeba pro správné napárování k datům z hospod.
3. Potřebné příznaky. Pro získání nejlepšího modelu budou třeba příznaky, které by na konzumaci piva měli mít největší vliv. Data z [32] nabízejí všechny potřebné příznaky.
4. Stáhnutí dat je velice jednoduché a přímočaré. Každý stáhnutý dataset obsahuje data v horizontu jednoho měsíce.

5.2.1 Popis příznaků

Pro finální model je vybráno 7 příznaků o počasí.

1. `10m u-component of wind`, „*východní složka větru 10m. Je to horizontální rychlost vzduchu pohybujícího se směrem na východ, ve výšce 10 metrů nad povrchem Země, v metrech za sekundu.*“[32]
2. `10m v-component of wind`, „*severní složka větru 10m. Je to horizontální rychlost vzduchu pohybujícího se směrem na sever, ve výšce 10 metrů nad povrchem Země, v metrech za sekundu.*“[32]
3. `2m temperature`, „*teplota vzduchu ve výšce 2 metrů nad povrchem země, moře nebo vnitrozemských vod.*“ Teplota je měřena v Kelvinech.[32]
4. `Snowfall`, „*nahromaděný celkový sníh, který spadl na zemský povrch. (…) Uvedené jednotky měří hloubku, jakou by měla voda, kdyby sníh roztál a byl rovnoměrně rozprostřen po mřížce.*“[32]
5. `Surface net solar radiation`, „*množství slunečního záření (také známého jako krátkovlnného záření) dopadajícího na povrch Země (jak přímého i rozptýleného) mínus množství odražené od zemského povrchu (které se řídí albedem). (…) Jednotky jsou jouly na metr čtvereční (J* m^{-2}*).*“[32]
6. `Surface solar radiation downwards`, „*množství slunečního záření (také známého jako krátkovlnného záření) dopadajícího na povrch Země. (…) Jednotky jsou jouly na metr čtvereční (J* m^{-2}*).*“[32]

7. `Total precipitation`, „*nahromaděná kapalina a zmrzlá voda, včetně deště a sněhu, která spadne na zemský povrch. (...) Jednotky srážek je hloubka v metrech.*“[32]

5.3 Spojení datasetů

Každý dataset o předpovědi počasí je ve formátu grib. Formát grib je datový formát speciálně určený pro data o počasí. V grib formátu jsou data ukládána do mřížek. Tyto mřížky určuje zeměpisná šířka a zeměpisná délka. Data o výtoči jsou ve formátu xlsb, což je formát binárního souboru Excel.[33]

Prvním krokem je načtení datasetu o výtoči piva, přejmenování jednotlivých příznaků pro lepší přehlednost a přidání prázdných sloupců určené pro příznaky o počasí. Následuje načtení datasetů o předpovědi počasí a vyříznutí požadovaných souřadnic. Do datasetu o výtoči piva se přidají data z datasetu o předpovědi počasí podle souřadnic, data a hodiny.

5.4 Předzpracování datasetu

V sekci je dataset předzpracován pro trénování jednotlivých modelů.

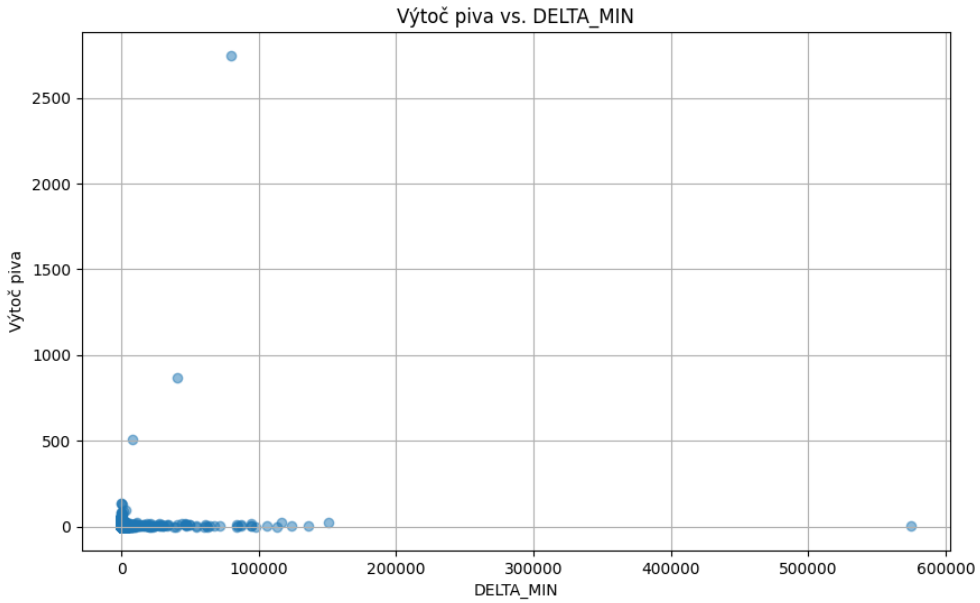
5.4.1 Základní úpravy

Prvními úpravami jsou převody příznaku o výtoči piva na litry a převody příznaku o teplotě na stupně Celsia. Tyto úpravy slouží čistě k lepšímu přehledu v datasetu a přehlednějším vizualizacím.

Spotřeba piva a celková návštěvnost hospod je největší ve večerních hodinách v dny, kdy den poté je volno a lidé tak nemusí do práce či do školy. Do datasetu se tedy přidají 4 další příznaky. První příznak `Day`, určuje den výtoče, pohybuje se v rozmezí od 0-6 (0 označuje pondělí, 6 označuje neděli). Druhý příznak `Holiday`, určuje, zda je daný den svátek (0 není svátek, 1 je svátek). Třetí příznak `Next day is free`, určuje zda je příští den volný (0 další den není volný, 1 další den je volný). Čtvrtý příznak `Month`, určuje měsíc výtoče (1 označuje leden, 12 označuje prosinec).

5.4.2 Odstranění nerelevantních hodnot

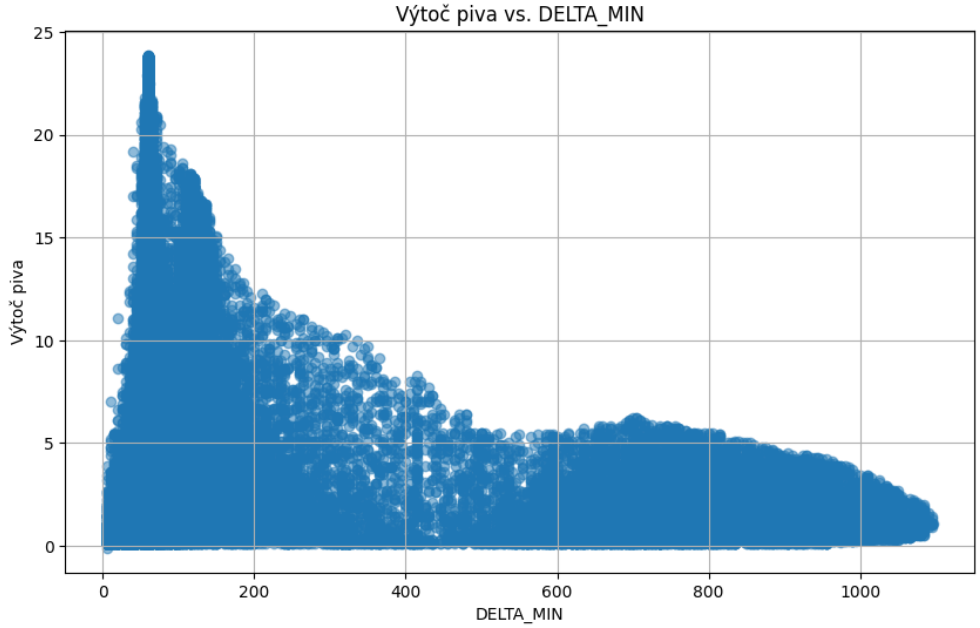
Jak již bylo zmíněno při popisu datasetu o výtoči tankové piva, tak se v tomto datasetu vyskytuje příznak `DELTA_MIN`, který uvadí časovou prodlevu od poslední změny počítadla a tedy dobu za kterou byla výtoč naměřena. V určitých případech dojde k extrémnímu nárůstu stavu počítadla, kdy např. vypadla komunikace mezi hospodou a serverem, kam jsou data odesílána, tyto extrémní nárůsty je potřeba identifikovat a odstranit.



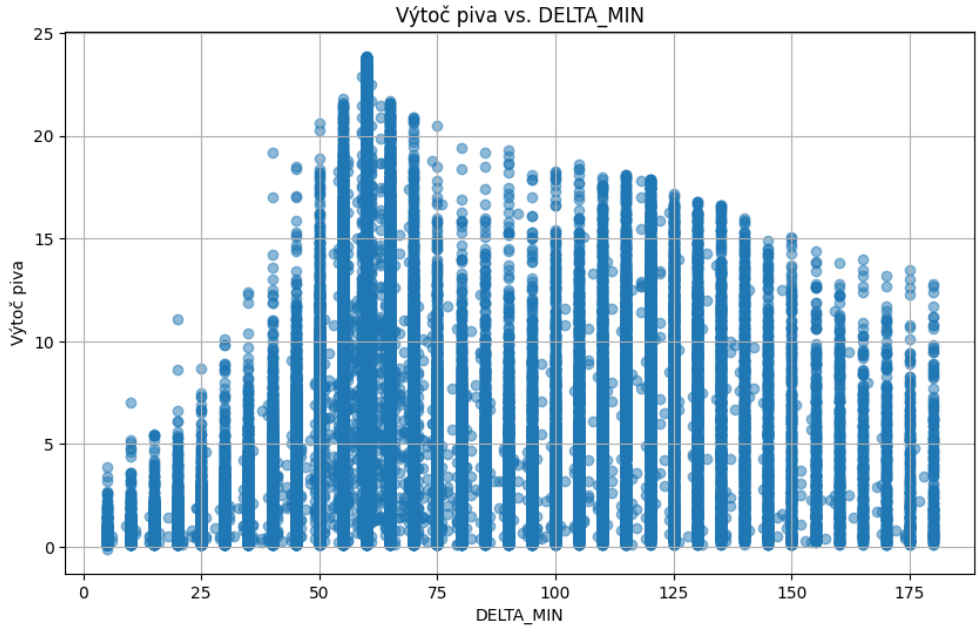
Obrázek 5.1 Data o výtoči před zpracovaním

Na grafu 5.1 je vidět vztah mezi výtočí piva a příznaku `DELTA_MIN` po načtení datasetu. Na ose x je vidět, že hodnoty příznaku `DELTA_MIN` dosahují hodnot přesahující i několik stovek tisíc. Pro odstranění těchto hodnot se použije IF. Pro ladění hyperparametrů IF je použit `RandomizedSearchCV`, který zkouší dle počtu iterací náhodné kombinace v určeném rozmezí. Provede se ladění hyperparametrů `n_estimators` a `contamination`. `n_estimators` určuje počet odhadců IF. `contamination` určuje znečištění IF (podíl anomálií datasetu). Výsledek této úpravy je vidět na 5.2.

Tato úprava slouží hlavně k odstranění výrazných anomálií. Nyní maximální hodnota `DELTA_MIN` přesahuje 1000 minut. Vyšší hodnoty většinou určují změnu stavu počítadla mezi zavřením hospody a otevřením hospody následující den. Dle poskytovatele o výtoči piva se však většina hodnot `DELTA_MIN` větší jak 180 považuje za nerelevantní, proto jsou odstraněny. Výsledek této úpravy je vidět na 5.3.



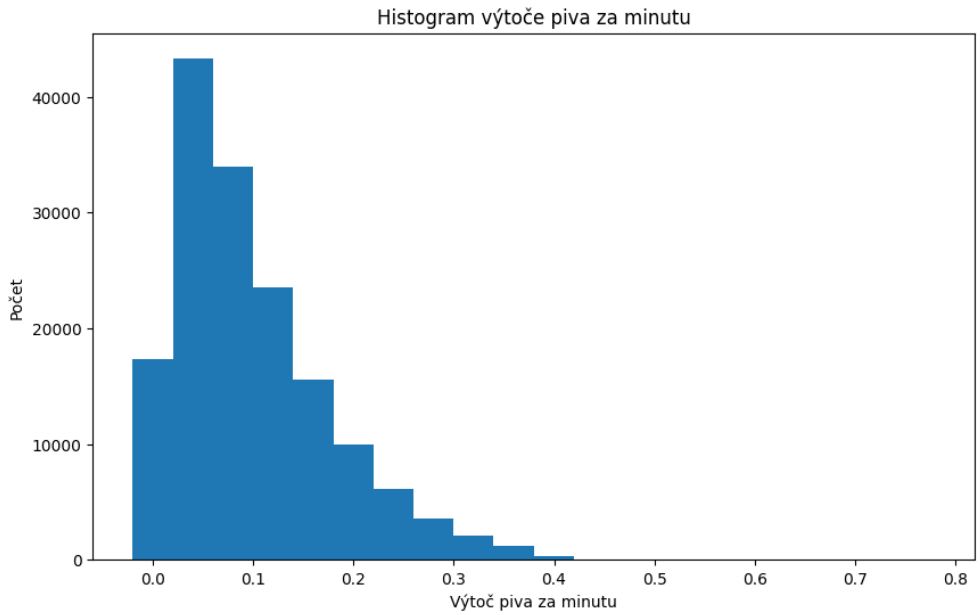
Obrázek 5.2 Data po použití IF



Obrázek 5.3 Data s `DELTA_MIN` menší než 180

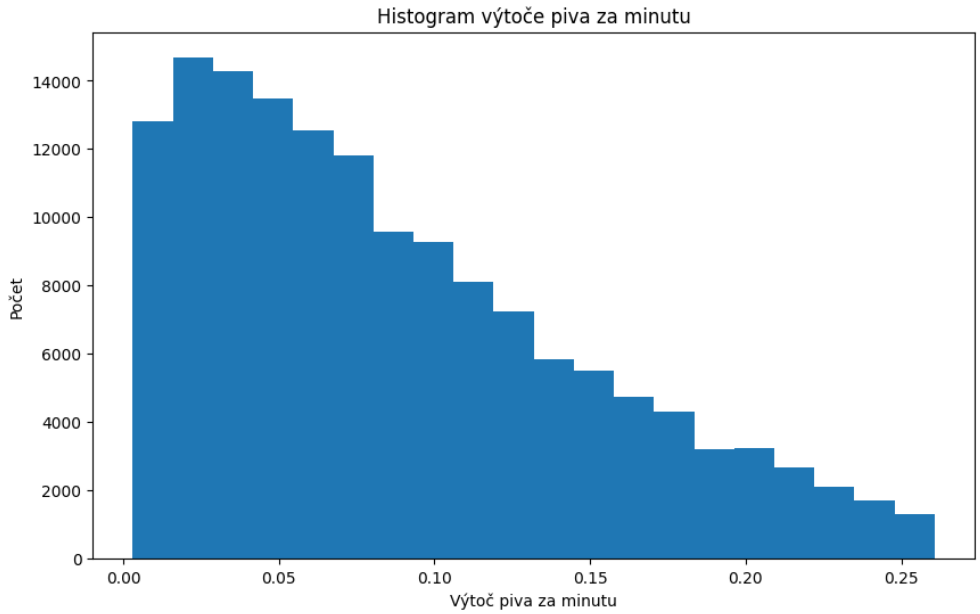
5.4.3 Příprava datasetu pro trénování

Pro trénování datasetu se musí určit vysvětlovaná proměnná. První úvahy navádí k vybrání přiznaku o výtoči v litrech jako vysvětlovanou proměnnou, tato úvaha je však špatná. Jelikož jsou data o výtoči odesílána nerovnoměrně (nejsou odesílána každých 60 minut), tak je výtoč vždy měřena za jiný čas. Vytvoří se tedy nový příznak `Beer outlet per min`, vypočítán jako `Beer outlet/DELTA_MIN`, který určuje výtoč piva za minutu. Odstraní se již nyní nepotřebný příznak `Beer outlet`.



Obrázek 5.4 Histogram příznaku `Beer outlet per min`

Na grafu 5.4 je vidět, že hodnoty příznaku `Beer outlet per min` jsou rozděleny nerovnoměrně a obsahují outliery. Použije se tedy znovu IF na odstranění outlierů. Změna je vidět na grafu 5.5. Nakonec se ještě odstraní nepotřebné příznaky.



Obrázek 5.5 Histogram příznaku `Beer outlet per min` po použití IF

Kapitola 6
Experimenty

V kapitole se natrénují jednotlivé modely. Celkově se vyzkouší 6 různých modelů, 3 regresní modely a 3 klasifikační modely. K trénování všech modelů je použit CatBoost a ladění hyperparametrů probíhá pomocí `RandomizedSearchCV`. Pro zlepšení modelů jsou k ladění vybrány tyto hyperparametry, `iterations`, `learning_rate` a `depth`. `iterations` určuje počet stromů, `learning_rate` určuje rychlost učení, která se použije ke snížení kroku gradientu a `depth` určuje hloubku stromů.[14]

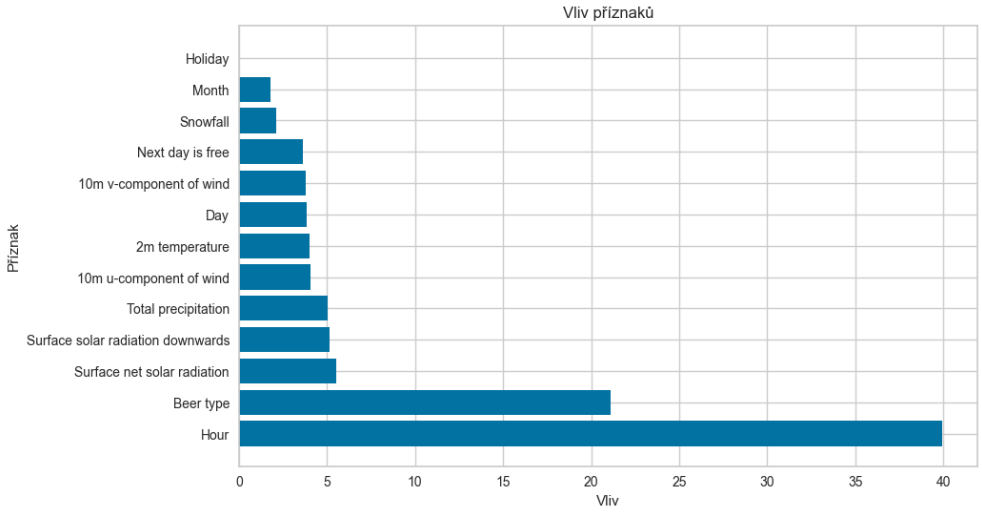
6.1 Regresní úlohy

V sekci se natrénují jednotlivé regresní modely a nakonec se porovnají predikce na validační množině. Kvality modelů jsou hodnoceny pomocí RMSE, MAE a R^2, přičemž nejdůležitější metrikou je R^2, protože udává jak dobře natrénovaný model odhaduje reálná data.

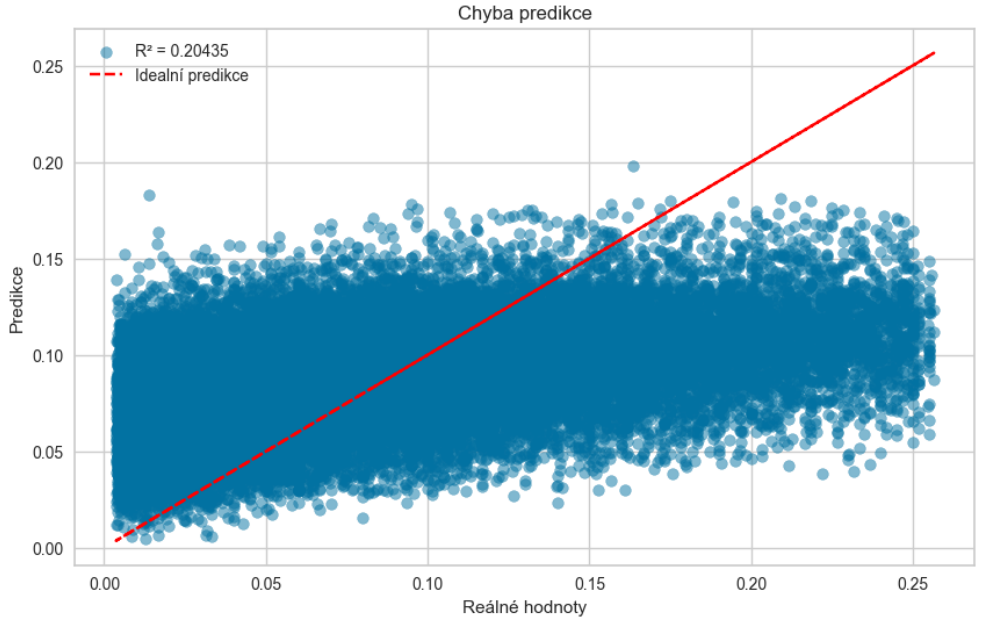
6.1.1 Regresní model

Prvním natrénovaným modelem je regresní model `CatBoostRegressor` na datasetu, který neobsahuje příznak `ID`. Výsledky na validační množině jsou: hodnota RMSE je 0,05450, hodnota MAE je 0,04388 a hodnota R^2 je 0,20435. Tyto výsledky naznačují velmi nekvalitní model.

Na 6.1 je vidět, že největší vliv na predikci vysvětlované proměnné má hodina, kdy je výtoč měřena, dále typ vytočeného piva a poté až následují příznaky o počasí v čele s příznakem o dopadu slunečního záření na zemský povrch. Při pohledu na predikce na grafu 6.2 je vidět, že predikované hodnoty příliš neodpovídají reálným hodnotám. Přímka, která by odhadovala predikované hodnoty, by byla skoro konstantní.



Obrázek 6.1 Vliv jednotlivých příznaků

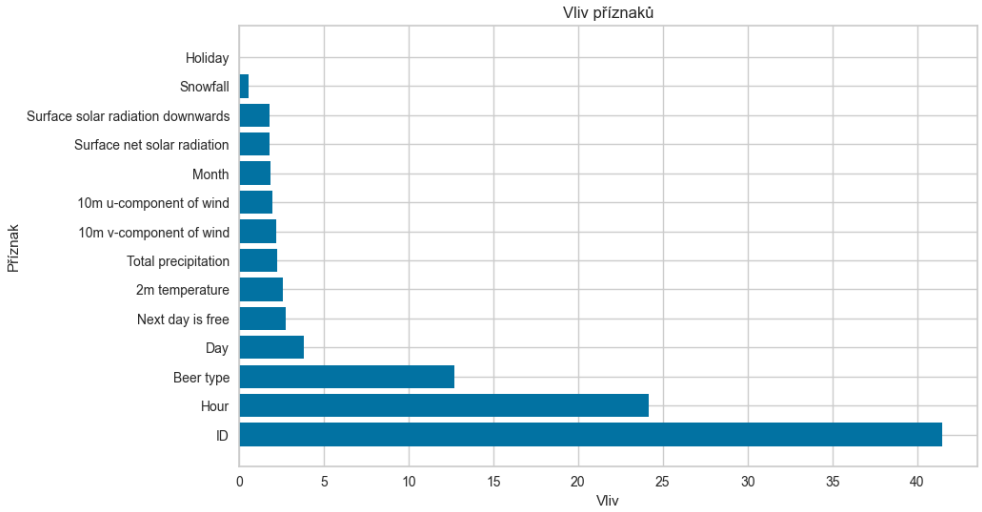


Obrázek 6.2 Chyba predikce

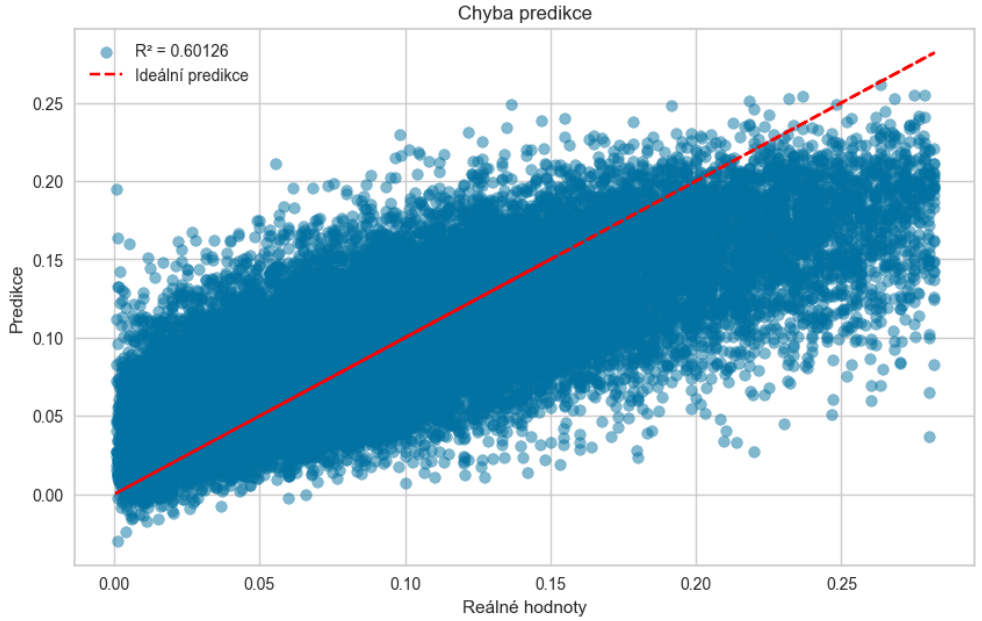
6.1.2 Regresní model s ID

Další model je regresní model `CatBoostRegressor`, nyní ale obsahující příznak `ID`. Výsledky na validační množině jsou: hodnota RMSE je 0,04164, hodnota MAE je 0,03179 a hodnota R^2 je 0,60126. Tyto hodnoty jsou o dost lepší než u předchozího modelu.

Z grafu 6.3 je vidět, že za zlepšením modelu stojí hlavně přidání příznaku `ID`, který má největší vliv na predikci, příznaky o počasí jsou opět v pozadí, nyní v čele s příznakem o teplotě. Zlepšení modelu, díky příznaku `ID`, naznačuje, že hodnoty výtoče se u každé hospody velmi liší. Na 6.4 je opět vidět zlepšení, predikce o dost více odpovídají reálným hodnotám než u minulého modelu.



Obrázek 6.3 Vliv jednotlivých příznaků u regrese s ID



Obrázek 6.4 Predikce vs. reálné hodnoty

6.1.3 Regresní model - jednotlivé hospody

Poslední regresní model je `CatBoostRegressor` natrénovaný na každé hospodé samostatně. Vytvoří se funkce na trénování jednotlivých hospod. V tabulce 6.1 je vidět 5 nejlepších modelů pro jednotlivé hospody s vyšším počtem záznamů a které vytočili pouze jeden druh piva[1]. Jednotlivé modely mají přibližně stejné výsledky jako minulý model obsahující příznak `ID`.

Model	RMSE	MAE	R^2
Hospoda 29	0,03561	0,02808	0,60537
Hospoda 5	0,03197	0,02378	0,60535
Hospoda 12	0,04051	0,03158	0,59906
Hospoda 15	0,03665	0,02779	0,59776
Hospoda 17	0,03791	0,03037	0,59642

Tabulka 6.1 Výsledky metrik pro 5 nejlepších regresních modelů

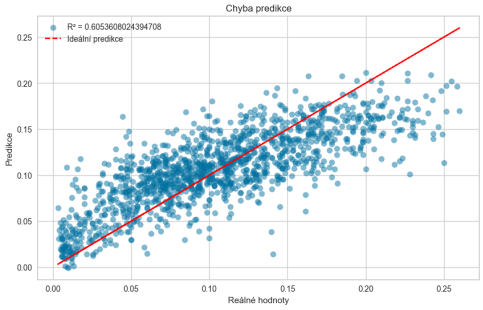
6.1.4 Porovnání výsledků

Nakonec se porovnají výsledky výše popsaných modelů. Jednotlivé modely jsou seřazeny podle R^2 sestupně. V tabulce 6.2 je vidět, že nejlépe predikují vybrané modely trénované na jednotlivých hospodách. Samozřejmě toto je jen 5 nejlepších modelů a ostatní hospody můžou mít predikce horší. Na souboru grafů 6.5 je vidět porovnání predikcí modelů.

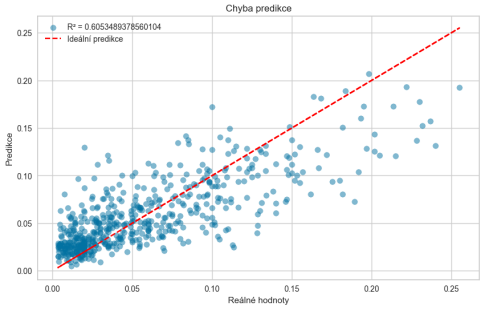
Hospoda	RMSE	MAE	R^2
Hospoda 29	0,03561	0,02808	0,60537
Hospoda 5	0,03197	0,02378	0,60535
Regresní model s ID	0,04164	0,03179	0,60126
Hospoda 12	0,04051	0,03158	0,59906
Hospoda 15	0,03665	0,02779	0,59776
Hospoda 17	0,03791	0,03037	0,59642
Regresní model	0,05450	0,04388	0,20435

Tabulka 6.2 Výsledky regresních modelů

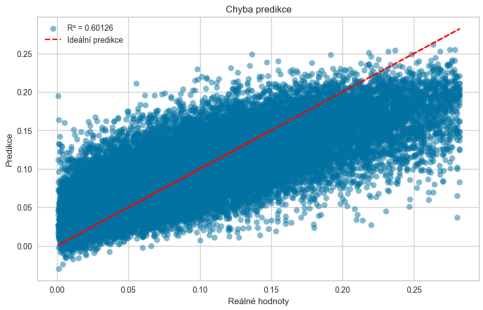
[1] Více druhů piva by mohlo zkreslit výsledky



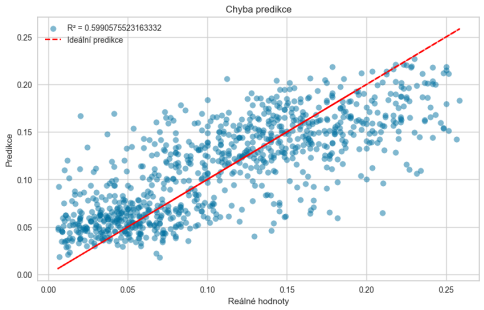
(a) Hospoda 29



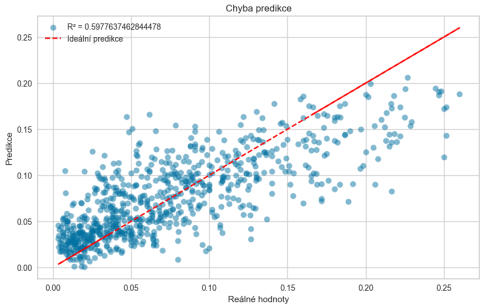
(b) Hospoda 5



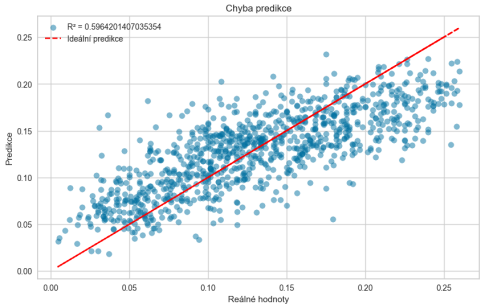
(c) Regresní model s `ID`



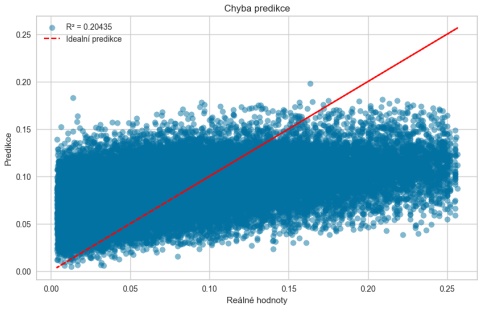
(d) Hospoda 12



(e) Hospoda 15



(f) Hospoda 17



(g) Regresní model

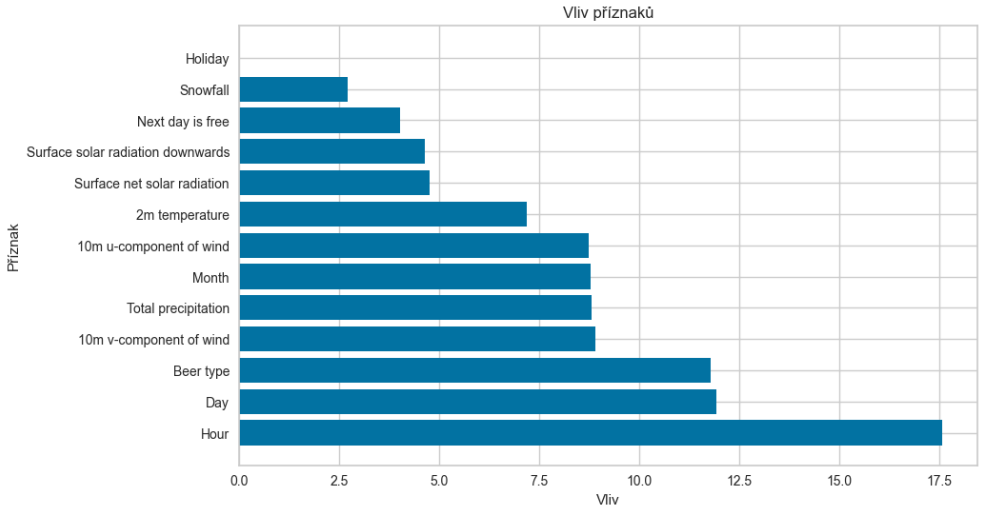
Obrázek 6.5 Porovnání jednotlivých regresních modelů

6.2 Klasifikační úlohy

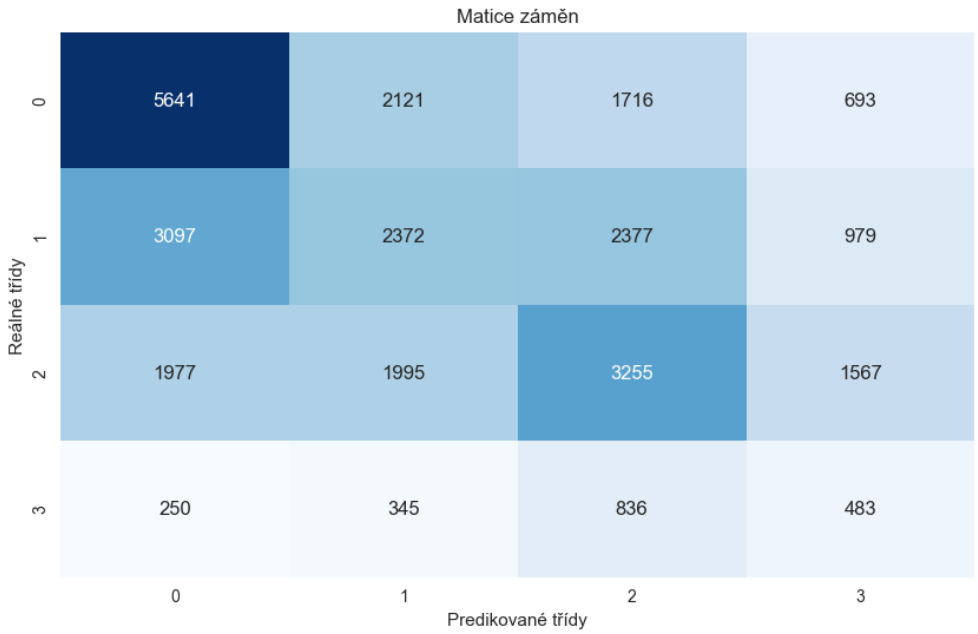
Na grafu 5.5 je vidět nerovnoměrnost dat, tato nerovnoměrnost může způsobit špatné predikce méně zastoupených dat. Z tohoto důvodu se převede úloha na klasifikační a použije se metoda SMOTE. Úloha se převede na klasifikační vytvořením několika rozmezí výtoče piva za minutu. Vytvoří se 4 rozmezí výtoče za minutu, první rozmezí je od 0 do 0,05, druhé rozmezí je od 0,05 do 0,1, třetí rozmezí je od 0,1 do 0,2 a čtvrté rozmezí je od 0,2 do 0,3. Po převedení úlohy na klasifikační je možné použit SMOTE. Kvalita modelu je hodnocena pomocí accuracy.

6.2.1 Klasifikační model

Jako u regresní úlohy se nejdřív pracuje s datasetem, který neobsahuje příznak `ID`, tentokrát se však jedná o model `CatBoostClassifier`. Hodnota accuracy na validační množině je 0,39560. Tento výsledek opět jako u regresního modelu bez příznaku `ID` naznačuje nekvalitní model, protože model určí klasifikační třídu správně pouze v necelých 40 %. Vliv příznaků je vidět na grafu 6.6, opět v čele s hodinou a dnem výtoče. Rozhodování modelu je vidět na matici záměn6.7.



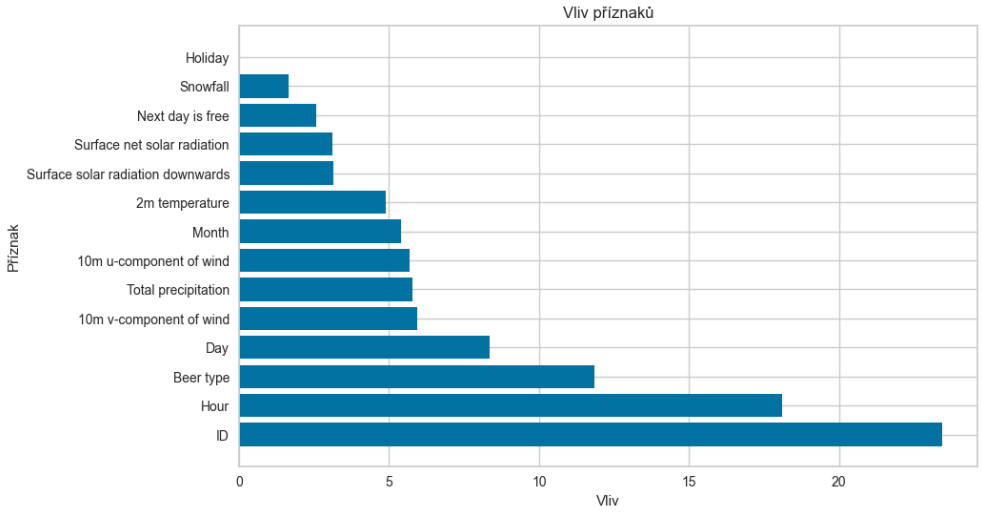
Obrázek 6.6 Vliv jednotlivých příznaků u klasifikace



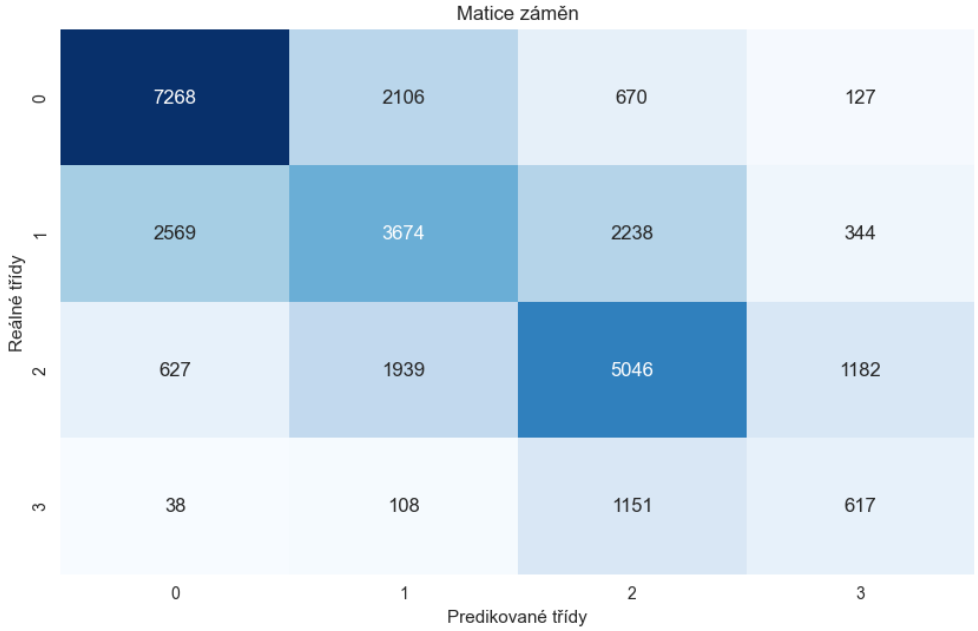
Obrázek 6.7 Matice záměn

6.2.2 Klasifikační model s ID

Další model je opět klasifikační model `CatBoostClassifier` s příznakem `ID`. Hodnota accuracy na validační množině je 0,55902. Tyto hodnoty jsou lepší než bez příznaku `ID`, ale zlepšení není tak markantní jako u regresního modelu. Na grafu 6.8 je vidět, že nyní má největší vliv příznak `ID`, jako u regresního modelu. Zlepšení se dá také pozorovat na matici záměn 6.9.



Obrázek 6.8 Vliv jednotlivých příznaků u klasifikace s `ID`



Obrázek 6.9 Matice záměn

6.2.3 Klasifikační model - jednotlivé hospody

Poslední klasifikační model je `CatBoostRegressor` natrénovaný na každé hospodě samostatně. V tabulce 6.1 je vidět 5 nejlepších modelů pro hospody, které opět vytočili pouze jeden druh piva.

Hospoda	accuracy
Hospoda 25	0,67986
Hospoda 24	0,66059
Hospoda 5	0,65693
Hospoda 36	0,64173
Hospoda 20	0,60976

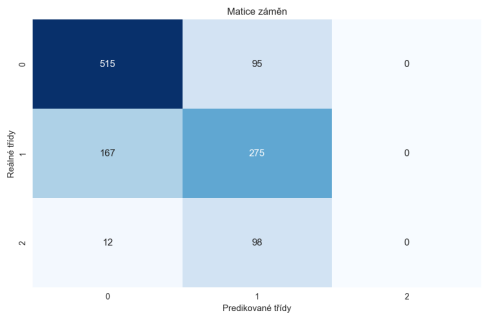
Tabulka 6.3 Přesnost 5 nejlepších klasifikačních modelů

6.2.4 Porovnání výsledků

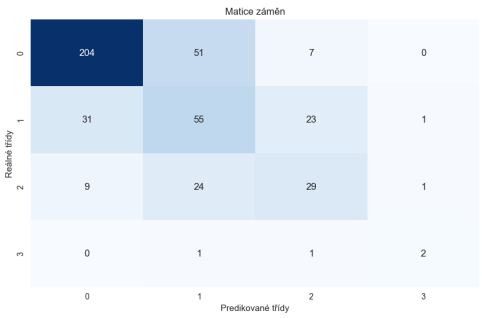
Porovnání výsledků v tabulce 6.2 výše popsaných modelů ukazuje jako u regresních modelů, že modely trénované na jednotlivých hospodách mají lepší přesnost. Modely jsou seřazeny podle accuracy sestupně. Porovnání přesností jednotlivých modelů je také možné pozorovat na maticích záměn 6.10.

Hospoda	accuracy
Hospoda 25	0.67986
Hospoda 24	0.66059
Hospoda 5	0.65693
Hospoda 36	0.64173
Hospoda 20	0.60976
Klasifikační model s ID	0,55902
Klasifikační model	0,39560

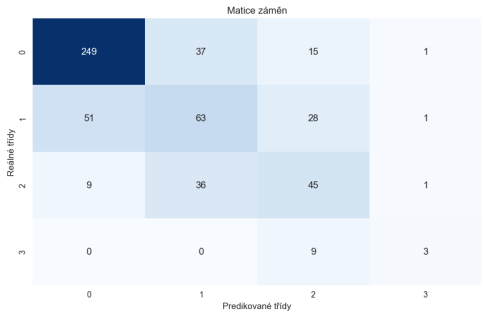
Tabulka 6.4 Přesnost klasifikačních modelů



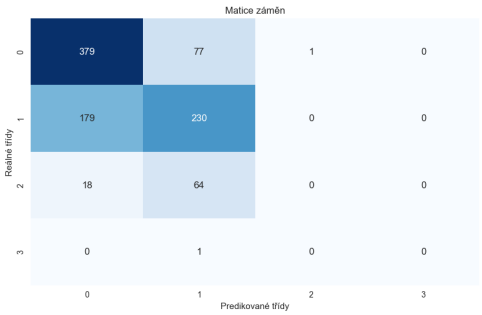
(a) Hospoda 25



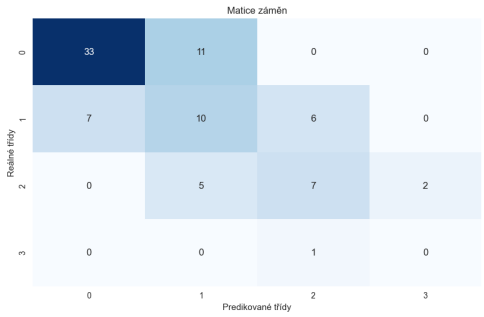
(b) Hospoda 24



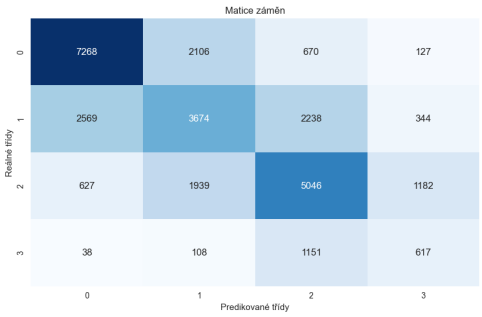
(c) Hospoda 5



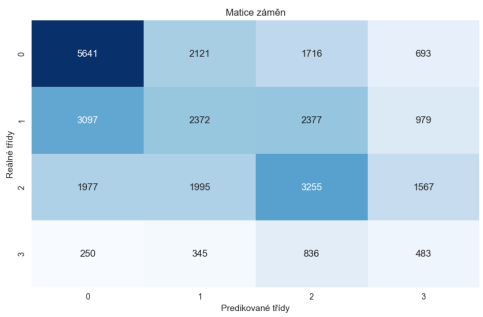
(d) Hospoda 36



(e) Hospoda 20



(f) Klasifikační model s `ID`



(g) Klasifikační model

Obrázek 6.10 Porovnání jednotlivých klasifikačních modelů pomocí matic záměn

Kapitola 7
Predikce

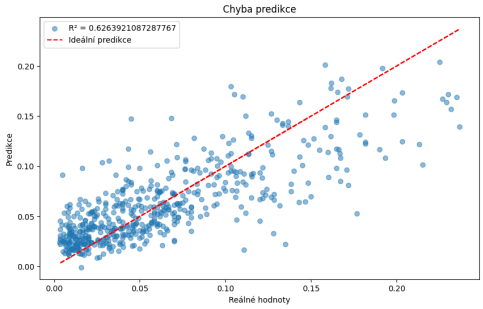
V kapitole jsou provedeny predikce na nejlepších modelech z předchozí kapitoly. Nejlepšími modely dle minulé kapitoly jsou modely trénované na jednotlivých hospodách. Použije se tedy 5 těchto modelů z každé úlohy a vyzkouší se predikce, jak na testovací množině, na reálné předpovědi počasí a také na krátkém časovém úseku hospod, které měly nejlepší výsledky. Data o reálné předpovědi počasí jsou z intervalu od 17.5.2024 do 30.5.2024 a jsou stažena z [7], protože z popsaných poskytovatelů dat počasí poskytuje nejlepší příznaky a nejlíp se s ním pracuje.

7.1 Testovací množina

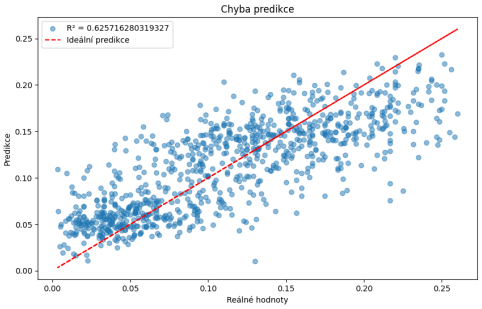
Jelikož přesnost na datech o reálné předpovědi počasí není zatím možná vyhodnotit, použije se k určení přesnosti testovací množina. Výsledky na testovací množině jsou vidět v tabulkách 7.1 a 7.2. Vizualizaci chyb je také možné pozorovat na souborech grafů 7.1 a 7.2.

Model	**RMSE**	**MAE**	**R^2**
Hospoda 5	0,03127	0,02332	0,62639
Hospoda 12	0,03947	0,03129	0,62572
Hospoda 17	0,03680	0,02948	0,61362
Hospoda 15	0,03860	0,02975	0,57228
Hospoda 29	0,03659	0,02885	0,57129

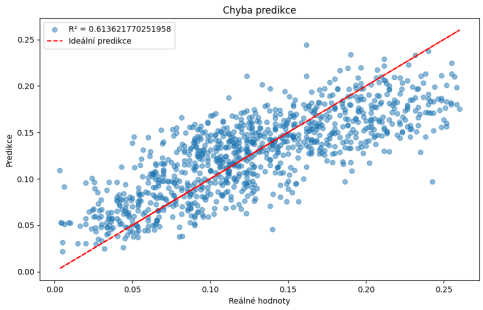
Tabulka 7.1 Výsledky metrik pro 5 nejlepších regresních modelů na testovací množině



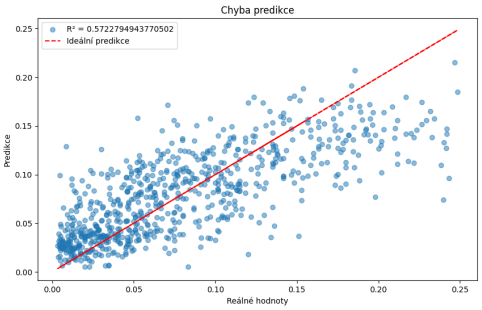
(a) Hospoda 5



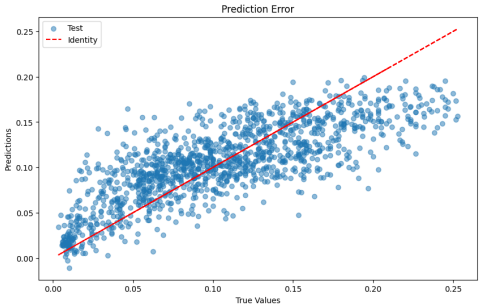
(b) Hospoda 12



(c) Hospoda 17



(d) Hospoda 15

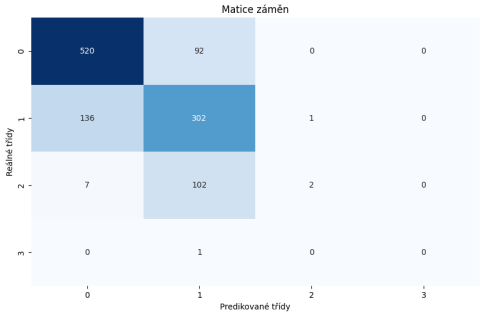


(e) Hospoda 29

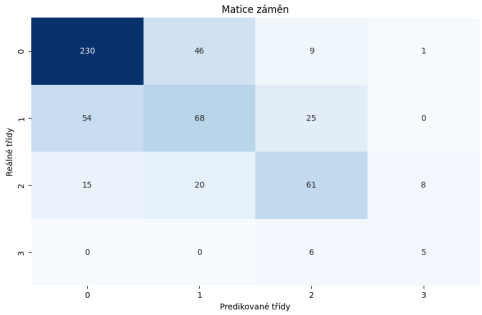
Obrázek 7.1 Predikce pro regresní modely

Hospoda	**accuracy**
Hospoda 25	0,70851
Hospoda 5	0,66423
Hospoda 24	0,66515
Hospoda 36	0,63368
Hospoda 20	0,56098

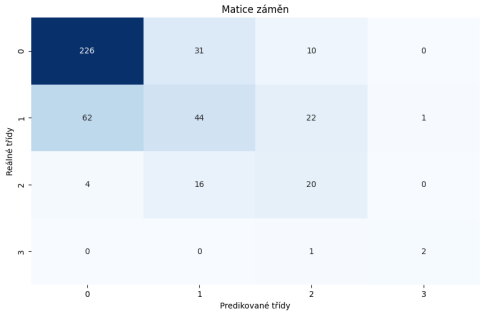
Tabulka 7.2 Výsledky metrik pro 5 nejlepších klasifikačních modelů na testovací množině



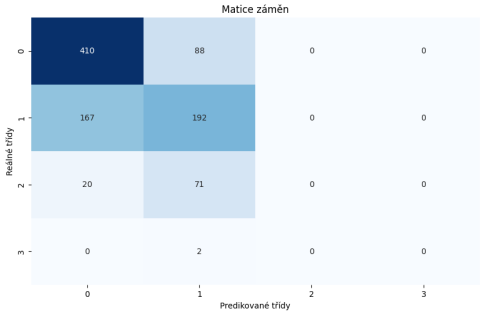
(a) Hospoda 25



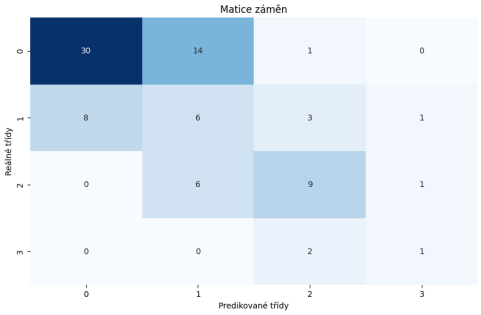
(b) Hospoda 5



(c) Hospoda 24



(d) Hospoda 36



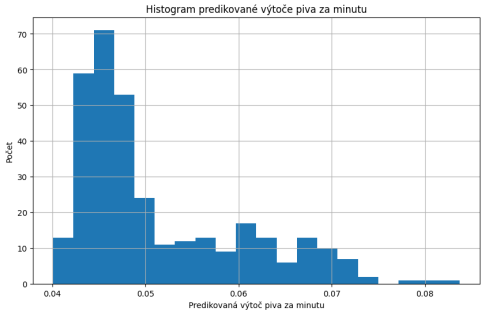
(e) Hospoda 20

Obrázek 7.2 Matice záměn pro klasifikační modely

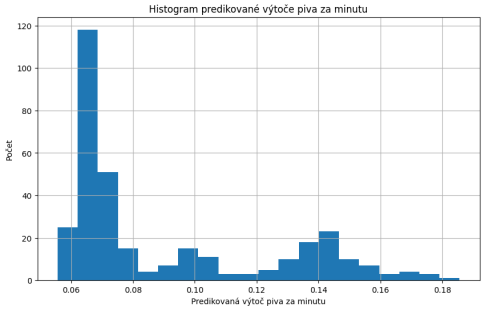
7.2 Reálná předpověď počasí

První částí je stáhnutí dat a předzpracováním těchto dat, jelikož stažená data nemají stejné příznaky jako data, na kterých jsou modely natrénované. Přidají se časové příznaky `Day`, `Holiday` a `Next day is free`, `Month`. Přidá se typ vytáčeného piva a přejmenují se sloupce. Jelikož předpověď počasí obsahuje pouze 1 příznak o slunečním záření, nahradí tento příznak oba příznaky z trénovacího datasetu ve správných jednotkách, toto nahrazení nezpůsobí žádné výraznější chyby. Stejný postup se zvolí i u příznaků popisující rychlost větru. Odstraní se nepotřebné příznaky a dataset se rozdělí na 5 datasetů pro hospody u regresních a klasifikačních modelů.

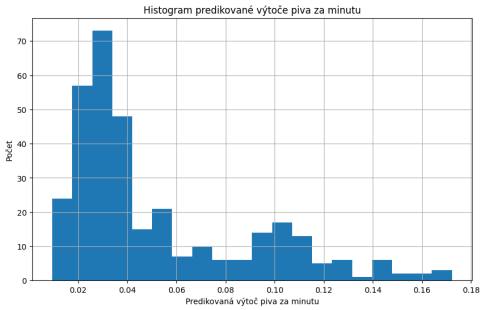
Po upravení dat se načtou uložené nejlepší regresní a klasifikační modely a provede se predikce na počasí v jednotlivých hospodách. Tyto výsledky není možné vyhodnotit, protože není možné vědět, jaká hodnota výtoče bude. Toto ověření už závisí na poskytovateli dat o výtoči piva. Na souboru grafů v 7.3 jsou vidět histogramy predikcí pro regresní modely a na souboru grafů 7.4 jsou vidět sloupcové grafy predikcí jednotlivých tříd.



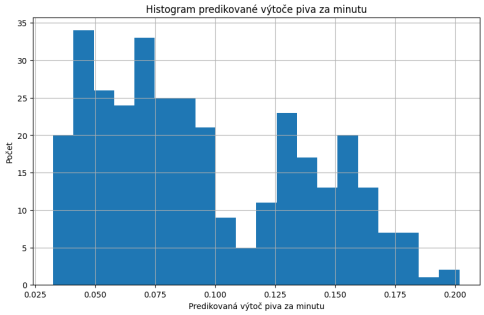
(a) Hospoda 5



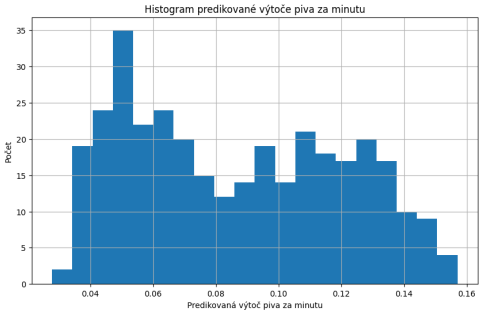
(b) Hospoda 12



(c) Hospoda 15

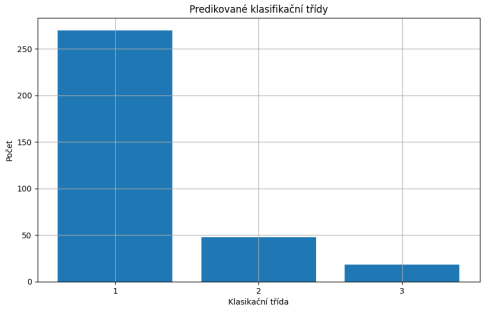


(d) Hospoda 17

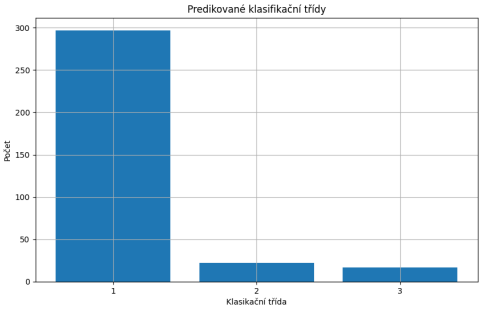


(e) Hospoda 29

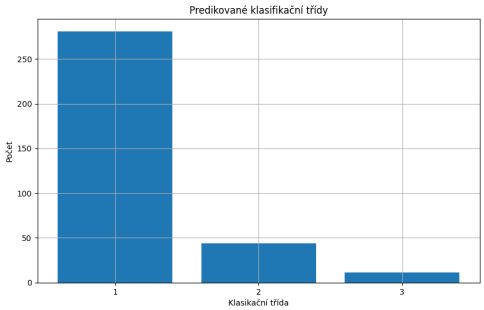
Obrázek 7.3 Histogramy predikce výtoče regresními modely v jednotlivých hospodách



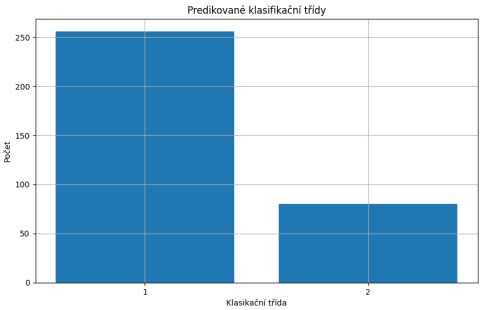
(a) Hospoda 5



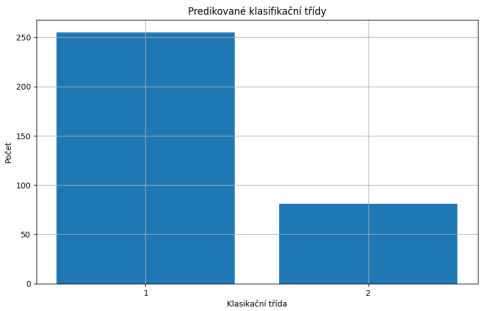
(b) Hospoda 20



(c) Hospoda 24



(d) Hospoda 25

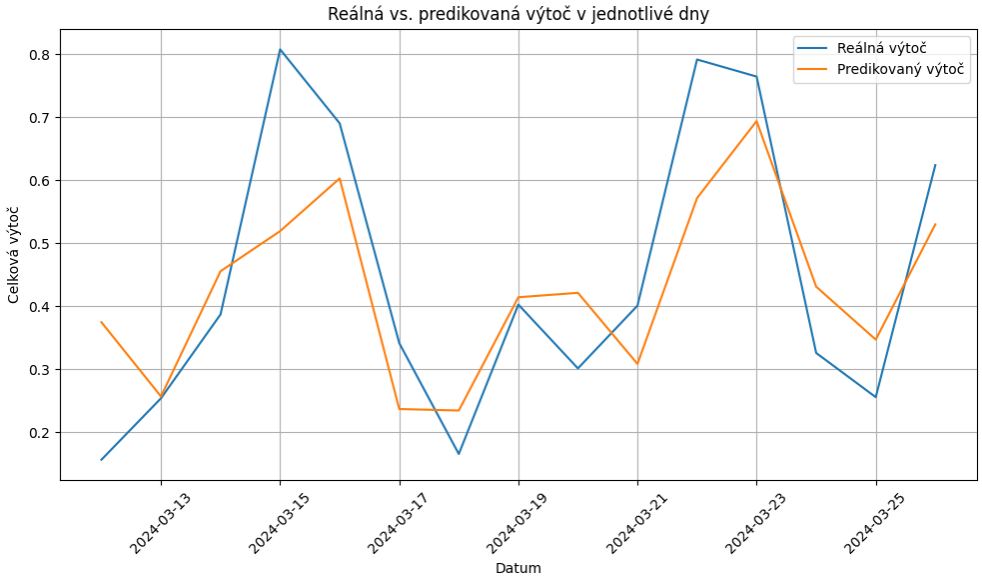


(e) Hospoda 36

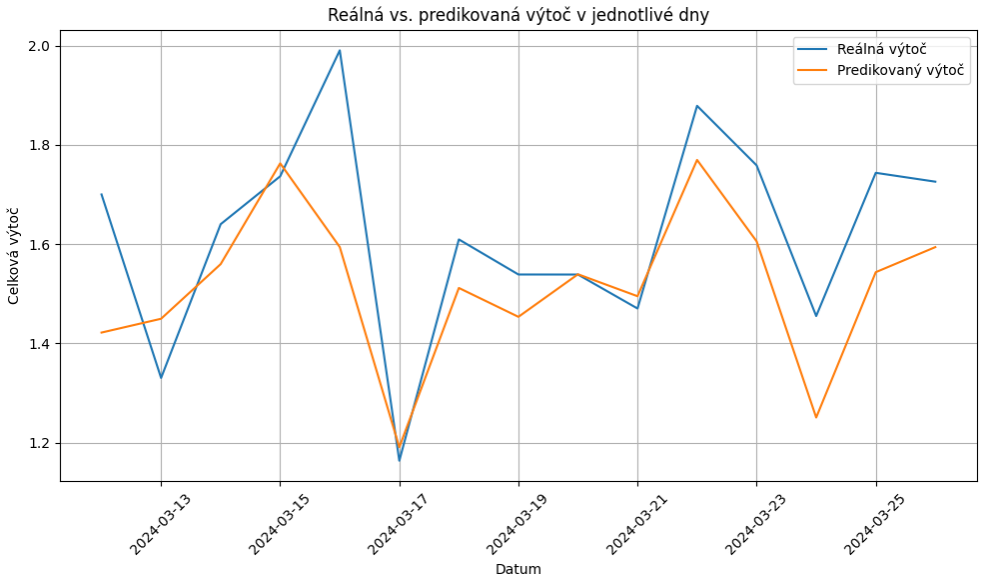
Obrázek 7.4 Sloupcové grafy predikce výtoče klasifikačními modely v jednotlivých hospodách

7.3 Ověření predikce v horizontu 14 dní

Jako poslední ověření predikcí modelů se použijí 2 nejlepší regresní modely na predikci výtoče za posledních 14 dní, kdy má daná hospoda ještě informace o výtoči. Tento časový úsek je vyříznut a na zbytku je model natrénován. Predikce modelu posledních 14 dní se porovná s reálnými hodnotami. Tato porovnání jsou vidět na grafech 7.5 a 7.6.



Obrázek 7.5 Predikce vs. reálné hodnoty během 14 dnů v hospodě 5



Obrázek 7.6 Predikce vs. reálné hodnoty během 14 dnů v hospodě 29

Závěr

Hlavním cílem práce bylo najít souvislost mezi počasím a výtočí tankového piva. Bylo natrénováno několik regresních a klasifikačních modelů. Nejlepšími modely se ukázaly býti modely natrénované na jednotlivých hospodách, to poukazuje na fakt, že u každé hospody se výtoč markantně liší. Při pozorování vlivů jednotlivých příznaků na modely, zkoumané příznaky počasí byly až v pozadí. Největší vliv na výtoč měla daná hospoda, hodina, kdy byla výtoč natočena, typ piva, den a až poté příznaky o počasí. To neznamená, že počasí nemá žádný vliv na spotřebu piva, ale tento vliv není tak významný oproti např. hodině.

Dílčím cílem bylo ověření spolehlivosti modelů a predikce výtoče tankového piva dle předpovědi počasí. U predikcí na testovací množině u regresních modelů se hodnota R^2 pohybují kolem 0,6 a RMSE kolem 0,035, což znamená, že se model splete v průměru o 0,035 litru za minutu. U klasifikačních modelů se přesnost pohybuje kolem 0,65, což znamená, že model v 63 % určí správnou třídu klasifikace. Predikce na reálné předpovědi počasí byla provedena a vizualizována a je na poskytovateli výtoče piva zhodnotit dané modely a případně je nasadit. Výsledkem práce je několik funkčních modelů, které dokáží krátkodobě predikovat výtoč tankové piva, která je z větší části ovlivněna místem a časovými údaji výtoče a z části také počasím.

Dané téma by rozhodně mělo být v budoucnosti znovu prozkoumáno. Při novém prozkoumání bych rozhodně doporučil vyzkoušet jiné modely, např. neuronové sítě. Za pokus také stojí vyzkoušení jiných příznaků o počasí, např. vlhkost, viditelnost, tlak a další. Modelům by také pomohlo vyzkoušení více kombinací při ladění modelů. Dané modely budou také lepší při větším množství dat, protože z porovnání predikcí a reálných hodnot bylo vidět, že dané modely hůře předpovídali výtoč během víkendu a lépe předpovídali výtoč během všedních dní.

Bibliografie

1. CIESLAR, Jan. Několikaletý růst spotřeby potravin se vloni zastavil [online]. 2023. Dostupné také z: https://www.czso.cz/csu/czso/nekolikalety-rust-spotreby-potravin-se-vloni-zastavil.
2. ŠÁLEK, Milan. *Monitoring a předpověď počasí* [online]. MUNI, Informační systém Masarykovy univerzity, 2013. Dostupné také z: https://is.muni.cz/el/sci/podzim2013/Z0076/um/Monitoring_a_predpoved_pocasi__dr._Salek.pdf.
3. GRAHAM, Steve; PARKINSON, Claire; CHAHINE, Mous. Weather forecasting through the ages [online]. 2002. Dostupné také z: https://earthobservatory.nasa.gov/features/WxForecasting/wx2.php.
4. CAHIR, John J. Weather forecasting [online]. 2024. Dostupné také z: https://www.britannica.com/science/weather-forecasting.
5. U.S. ARMY PHOTO. *The ENIAC main control panel at the Moore School of Electrical Engineering*. 2023. Dostupné také z: https://ftp.arl.army.mil/ftp/historic-computers/png/eniac1.png.
6. OPENWEATHER. *One Call API 3.0* [online]. 2022. Dostupné také z: https://openweathermap.org/api/one-call-3#current.
7. VISUALCROSSING. *Weather Query Builder* [online]. 2020. Dostupné také z: https://www.visualcrossing.com/weather/weather-data-services.
8. ECMWF. *Open data* [online]. 2021. Dostupné také z: https://www.ecmwf.int/en/forecasts/datasets/open-data.
9. CEROVIĆ, Zdenko; HORVAT, Đuro. Impact of weather changes on consumption of beverages in the hospitality industry [online]. 2013. Dostupné z DOI: 10.20867/thm.19.2.3.

10. DOBROVOLNÝ, Petr. *Regresní počet* [online]. MUNI, Informační systém Masarykovy univerzity, 2014. Dostupné také z: https://is.muni.cz/el/sci/podzim2016/Z1069/um/50668892/Statistika_7_regresni_pocet.pdf.

11. AL., Meritxell Ventura-Cots et. Colder Weather and Fewer Sunlight Hours Increase Alcohol Consumption and Alcoholic Cirrhosis Worldwide [online]. 2018. Dostupné z DOI: 10.1002/hep.30315.

12. GUPTA, Aryan. *Spearman's Rank Correlation: The Definitive Guide To Understand* [online]. 2023. Dostupné také z: https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation.

13. KELEŞ, Büşra; P.E., Patricia Gómez-Acevedo; SHAIKH, Nazrul I. The impact of systematic changes in weather on the supply and demand of beverages [online]. 2017. Dostupné z DOI: 10.1016/j.ijpe.2017.08.002.

14. YANDEX. *CatBoost* [online]. 2018. Dostupné také z: https://catboost.ai/en/docs/concepts/algorithm-main-stages.

15. VAŠATA, Daniel. *Ensemble metody - náhodné lesy, AdaBoost* [online]. 2024. Dostupné také z: https://courses.fit.cvut.cz/BI-ML1/lectures/files/BI-ML1-09-cs-handout.pdf.

16. VAŠATA, Daniel. *Lineární regrese - metoda nejmenších čtverců* [online]. 2021. Dostupné také z: https://courses.fit.cvut.cz/BI-ML1/lectures/files/BI-ML1-05-cs-handout.pdf.

17. MATALONGA, Hugo. Choosing between MAE, MSE and RMSE [online]. 2023. Dostupné také z: https://hmatalonga.com/blog/choosing-between-mae-mse-and-rmse/.

18. NEWCASTLE UNIVERSITY. Coefficient of Determination, R-squared [online]. 2022. Dostupné také z: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html.

19. VAŠATA, Daniel. *Supervizované učení, klasifikační úloha, rozhodovací stromy* [online]. 2021. Dostupné také z: https://courses.fit.cvut.cz/BI-ML1/lectures/files/BI-ML1-02-cs-handout.pdf.

20. LIU, Fei Tony; VICTORIA, Kai Ming Ting; ZHOU, Zhi-Hua. Isolation Forest [online]. 2009. Dostupné z DOI: 10.1109/ICDM.2008.17.

21. CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique [online]. 2002. Dostupné z DOI: 10.48550/arXiv.1106.1813.

22. YANDEX. *CatBoost* [online]. 2018. Dostupné také z: https://catboost.ai/.

23. RUSSELL, Iain. *GitHub* [online]. 2018. Dostupné také z: https://github.com/ecmwf/cfgrib.

24. NOGUEIRA, Fernando. *imbalanced-learn documentation* [online]. 2014. Dostupné také z: https://imbalanced-learn.org/stable/.

25. AL., Afshin Darian et. *Jupyter* [online]. 2014. Dostupné také z: https://jupyter.org/.

26. HUNTER, John. *Matplotlib* [online]. 2009. Dostupné také z: https://matplotlib.org/.

27. AL., Sebastian Berg et. *Numpy* [online]. 2005. Dostupné také z: https://numpy.org/.

28. BOSSCHE ET. AL., Joris Van den. *Pandas* [online]. 2008. Dostupné také z: https://pandas.pydata.org/.

29. WASKOM, Michael L. *seaborn: statistical data visualization*. Sv. 6. The Open Journal, 2021. Č. 60. Dostupné z DOI: 10.21105/joss.03021.

30. BOSSCHE ET. AL., Joris Van den. *scikit-learn* [online]. 2007. Dostupné také z: https://scikit-learn.org/.

31. AL., Andrew Nelson et. *Scipy* [online]. 2001. Dostupné také z: https://scipy.org/.

32. J., Muñoz Sabater. *ERA5-Land hourly data from 1950 to present* [online]. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2019. Dostupné z DOI: 10.24381/cds.e2161bac.

33. CAMPMANY, Elies. WHAT IS A GRIB FILE? [Online]. 2023. Dostupné také z: https://vortexfdc.com/blog/what-is-a-grib-file/.

Obsah příloh

- `readme.txt` ..... stručný popis obsahu média
- `src`
  - `impl` ..... celá implemtance
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF