UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Použití zobecněných lineárních modelů pro predikci prodeje

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Tomáš Fürst**
Vypracoval(a): **Bc. Kristýna Neuwirthová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Kristýna Neuwirthová

Název práce: Použití zobecněných lineárních modelů pro predikci prodeje

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Tomáš Fürst

Rok obhajoby práce: 2021

Abstrakt: Tato práce je zaměřená na prozkoumání teoretického pozadí a využití zobecněných lineárních modelů v predikci prodeje kosmetických výrobků. V první části se zaměřuje na teoretické základy regresní analýzy počínaje jednoduchými lineárními modely, přes kategorické proměnné až po způsoby odhadu a interpretaci regresních parametrů. Následuje klíčová kapitola o zobecněných lineárních modelech blíže zaměřená na modely s Poissonovým a negativně-binomickým rozdělením náhodné veličiny. Druhou část práce tvoří vlastní výzkum s cílem využít popsané metody a zhodnotit možnosti jejich aplikace v problematice predikce prodeje kosmetických výrobků.

Klíčová slova: zobecněné lineární modely, regresní analýza, predikce, Poissonovo rozdělení, negativně-binomické rozdělení, Hooke-Jeevsova metoda

Počet stran: 78

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Kristýna Neuwirthová

Title: Sales prediction by means of gerenalized linear models

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Tomáš Fürst

The year of presentation: 2021

Abstract: This thesis is focused on the use of generalized linear models in sales prediction. In the first part, the theoretical basis of regression analysis is considered from simple linear models, through categorical variables to estimation methods and interpretation of regression parameters. The next chapter focuses on generalized linear models considering closely Poisson models and negative-binomial models. The second part describes my own research aimed at the use and evaluation of the previously described methods in sales prediction area.

Key words: generalized linear models, regression analysis, prediction, Poisson distribution, negative-binomial distribution, Hooke-Jeeves method

Number of pages: 78

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Tomáše Fürsta a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala všem, kteří mě při psaní této práce podporovali. V první řadě panu doktorovi Fürstovi za vedení a zdroj nadšení pro posouvání hranic v odborných dovednostech. Děkuji také pánům Kratochvílovi a Milanovi ze společnosti Oriflame za umožnění spolupráce na praktické části. Velký dík patří mé rodině, nejbližším přátelům a mému partnerovi, bez nichž by dosahování akademických cílů postrádalo smysl.

Úvod

Fenoménem dnešní doby v oblasti podnikání a prodeje se stává využívání velkého množství dat, které je umožněno rozvojem informačních technologií a databází. Firmy mají shromážděné obrovské množství informací o svých klientech i o svých provozních aktivitách. To dává prostor pro rozvoj oblastí Data Science s cílem popsat datové soubory a využít získané informace k optimalizaci a zlepšení procesů. V oblasti výroby a obchodu se může jednat o optimalizaci výroby, distribuce výrobků, skladování či samotného nabízení výrobků klientům. Také je možné z historických trendů vytvářet modely k predikování budoucího vývoje. A právě tímto směrem se bude ubírat tato diplomová práce.

Nástrojů k predikci nabízí Data Science celou řadu. Cílem této práce je prozkoumat a zhodnotit využitelnost zobecněných lineárních modelů, které na rozdíl od obyčejných (nezobecněných) lineárních modelů dovolují popisovat širší škálu reálných situací. Jednou z nich je také modelování počtu prodaných kusů, kdy přirozeně očekáváme, že výstupem modelu budou nezáporná celá čísla. V tomto případě by tedy bylo nesprávné použít obyčejné lineární modely, které předpokládají normální rozdělení vysvětlované proměnné a jejich výstupem jsou tudíž jakékoliv hodnoty oboru reálných čísel. Tato diplomová práce hlouběji analyzuje modely s Poissonovým rozdělením modelující data vyjadřující počet a také modely s negativně-binomickým rozdělením, které se využívají k řešení problému nadměrného rozptylu u Poissonova rozdělení.

Celá práce bude rozdělena do dvou částí – teoretické a praktické. V teoretické části budou popsány základy regresní analýzy opírající se o obyčejné lineární modely, jejichž rozšířením získáme teoretický základ pro zobecněné lineární mo-

dely. Zaměříme se také na způsob odhadů regresních parametrů a na jejich interpretaci, zahrnuty budou také modely s kategorickými proměnnými, které jsou v datech o prodeji výrobků velice časté. U zobecněných lineárních modelů budou rozebrány jejich tři základní složky (náhodná a nenáhodná složka a spojovací funkce) a zaměříme se také na způsob odhadu regresních parametrů. Vyjdeme z obecné metody maximalizace funkce věrohodnosti, která není podmíněna normalitou dat a ukážeme, že výpočetně jednodušší a více používaná iterativní metoda vážených nejmenších čtverců má stejné výstupy, a proto je možné ji využít. Následně bude popsáno teoretické pozadí modelů s Poissonovým rozdělením, problematika nadměrného rozptylu a modely s negativně-binomickým rozdělením. Teoretickou část uzavře krátká kapitola o Hooke-Jeevsově metodě optimalizace funkce více proměnných, která přispěla k výsledkům získaným v praktické části.

Samotná praktická část bude založena na využití popsaných metod na datech společnosti Oriflame, která se zabývá prodejem kosmetiky na několika světových trzích. Hlavním cílem této části bude vytvořit několik modelů predikujících počty prodaných výrobků Oriflamu a dosáhnout co nejlepší predikce. Vyjdeme z popisné statistiky zpracovávaného datového souboru, následně bude vysvětlena myšlenka tvorby modelů, které má tato práce za cíl porovnat. Bude provedena argumentace k výběru srovnávací metriky a představení dvou druhů optimalizace parametrů modelu za účelem dosažení co nejlepšího výsledku vzhledem ke zvolené metrice. Výstupem bude pořadí modelů na základě přesnosti predikce, porovnání vytvořených modelů s modelem Oriflamu a krátká vizualizace predikcí pro ilustraci vzniklých problémů a jejich řešení.

Kapitola 1

Teoretická část

1.1. Regresní analýza

Regresní analýza je statistickou metodou zabývající se vztahem mezi dvěma nebo více proměnnými, kdy vysvětlovaná proměnná je popsána pomocí vysvětlujících proměnných [6]. Mezi typické úlohy regresní analýzy patří modelování vztahů mezi proměnnými v businessu, sociálních i přírodních vědách jako například:

- vztah mezi pracovním výkonem zaměstnance a výsledky testu dovedností,
- vztah mezi slovní zásobou dítěte, jeho věkem a dosaženým vzděláním rodičů,
- vztah mezi počtem prodaných produktů a výší nákladů vynaložených na reklamu.

1.1.1. Základní pojmy a značení

Mezi základní pojmy použité v této práci patří [4]:

vysvětlovaná proměnná taky závislá proměnná, náhodná, značíme Y

vysvětlující proměnná taky nezávislá proměnná, nenáhodná, značíme x

regresní parametry značíme β_i pro $i \in \{0, 1, 2, \dots\}$

střední hodnota náhodné veličiny X je definována jako

$$E(X) = \sum_n x_n \cdot P(X = x_n)$$

pro diskrétní náhodnou veličinu a

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

pro spojitou náhodnou veličinu

rozptyl náhodné veličiny X je definován jako

$$var(X) = E[X - E(X)]^2$$

Počet pozorování je značen n, počet regresních parametrů označujeme p a počet vysvětlujících proměnných k .

1.1.2. Jednoduchá lineární regrese

Model jednoduché lineární regrese s jednou vysvětlující proměnnou má následující formu [6]:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1.1}$$

kde Y_i je i -tá hodnota vysvětlované proměnné, β_0, β_1 jsou regresní parametry, x_i je i-tá hodnota vysvětlující proměnné a ε_i je náhodná odchylka.

Pro náhodnou odchylku předpokládáme splnění následujících podmínek:

- $E(\varepsilon_i) = 0$, $\forall i$
- $var(\varepsilon_i) = \sigma^2$, $\forall i$
- ε_i a ε_j jsou nekorelované pro $i \neq j$, tj. $cov(\varepsilon_i, \varepsilon_j) = 0$.

Regresní model (1.1) je lineární v parametrech β. Pokud jsou splněny předpoklady na náhodnou odchylku, platí

$$E(Y_i) = \beta_0 + \beta_1 x_i, \tag{1.2}$$

tj. regresní funkce vyjadřuje vztah mezi očekávanou hodnotou náhodné veličiny Y a danou hodnotou regresoru x. Z druhé podmínky na náhodnou odchylku (konstantní rozptyl) také vyplývá, že náhodná veličina Y musí mít také konstantní rozptyl, neboť

$$var(Y_i) = var(\beta_0 + \beta_1 x_i + \varepsilon_i) = var(\varepsilon_i) = \sigma^2.$$

Z požadavku na nekorelovanost náhodných odchylek taktéž plyne nekorelovanost hodnot vysvětlované proměnné Y_i a Y_j pro $i \neq j$.

1.1.3. Vícenásobná lineární regrese

Pokud chceme najít vztah mezi více vysvětlujícími proměnnými a vysvětlovanou proměnnou, mluvíme o tzv. vícenásobné lineární regresi. Pakliže máme $p-1$ vysvětlujících proměnných, regresní model má následující tvar:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i. \tag{1.3}$$

Náhodná odchylka musí splňovat stejné podmínky jako v případě jednoduchého lineárního modelu.

Pro maticový zápis tohoto modelu je potřeba definovat vektor vysvětlované proměnné

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},$$

dále matici hodnot vysvětlujících proměnných

$$\mathbf{X}_{n\times p} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{bmatrix},$$

vektor regresních parametrů

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

a vektor náhodných chyb

$$\boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Model (1.3) lze poté zapsat jako

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.4}$$

Pro vektor náhodných odchylek platí

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

a

$$var(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}. \tag{1.5}$$

1.1.4. Odhady regresních parametrů

Cílem regresní analýzy je určit odhady parametrů $\boldsymbol{\beta}$ (ozn. $\hat{\boldsymbol{\beta}}$). Pro nalezení dobrých odhadů $\hat{\boldsymbol{\beta}}$ se používá metoda nejmenších čtverců, která zde bude vysvětlena pro případ jednoduchého lineárního modelu (1.1). Metoda nejmenších čtverců vychází z chyb modelu definovaných jako

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

a hledá parametry $\hat{\boldsymbol{\beta}}$ tak, aby minimalizovala součet čtverců těchto chyb, tj.

$$min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{1.6}$$

Analytické odvození vztahů pro výpočet hodnot $\hat{\beta}_0$ a $\hat{\beta}_1$ vychází z derivace minimalizované funkce podle β_0 a β_1 a položení těchto derivací rovno nule. Lze jej nalézt např. v [6], zde uvedeme pouze výsledné vztahy:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Pro model vícenásobné regrese (1.3) lze metodu nejmenších čtverců zobecnit jako minimalizaci součtu

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{p-1} x_{i,p-1})^2,$$

nebo maticově jako řešení soustavy normálních rovnic

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}, \tag{1.7}$$

z nichž můžeme odhady $\hat{\boldsymbol{\beta}}$ určit jako

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \tag{1.8}$$

Na základě Gauss-Markovovy věty (viz [6]) mají odhady regresních parametrů určené pomocí metody nejmenších čtverců a za dodržení podmínek lineárního modelu následující vlastnosti:

I) jsou nestrannými odhady parametrů $\boldsymbol{\beta}$,

II) jsou lineárními nejlepšími odhady.

Pokud navíc platí, že náhodné odchylky $\boldsymbol{\varepsilon}$ mají normální rozdělení, tj. $\boldsymbol{\varepsilon}$ jsou nezávislé s rozdělením $N(0, \sigma^2)$, odhady regresních parametrů jsou konzistentní a dostačující.

Výše uvedené vlastnosti si nyní definujeme (na základě [4] a [17]).

Definice 1.1.1. *Nechť je dán parametr $\boldsymbol{\theta} \in \mathbb{R}^k, k \geq 1$ a bodový odhad $T = T(X_1, \ldots, X_n)$ reálné funkce $\tau(\boldsymbol{\theta})$. Potom bodový odhad T nazveme*

- *nestranným odhadem parametrické funkce $\tau(\boldsymbol{\theta})$, jestliže platí*

$$E\left[T(\mathbf{X})\right] = \tau(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta,$$

- *nejlepším odhadem parametrické funkce $\tau(\boldsymbol{\theta})$, platí-li*

$$var(T(\mathbf{X})) \leq var(T^*(\mathbf{X})), \quad \forall \boldsymbol{\theta} \in \Theta,$$

  kde $T^*(\mathbf{X})$ je libovolný jiný nestranný odhad $\tau(\boldsymbol{\theta})$,

- *(slabě) konzistentním odhadem parametrické funkce $\tau(\boldsymbol{\theta})$, jestliže platí*

$$\lim_{n\to\infty} P(|T(\mathbf{X}) - \tau(\boldsymbol{\theta})| < \varepsilon) = 1, \quad \forall \varepsilon > 0, \forall \boldsymbol{\theta} \in \Theta,$$

- *je postačujícím (suficientním) odhadem, pokud lze hustotu $f_{\boldsymbol{\theta}}(\mathbf{X})$ zapsat jako*

$$f_{\boldsymbol{\theta}}(\mathbf{X}) = h(\mathbf{X}) g_{\boldsymbol{\theta}}(T(\mathbf{X})),$$

  tj. jako součin funkce h, která nezávisí na $\boldsymbol{\theta}$ a funkce g, která na $\boldsymbol{\theta}$ závisí právě prostřednictvím $T(\mathbf{X})$. Jinými slovy, postačující odhad splňuje podmínku, že jakýkoliv další odhad získaný ze stejného náhodného výběru nepřidá další dodatečnou informaci o hodnotě parametru.

Splnění normality náhodných odchylek taky zajišťuje normalitu vysvětlované proměnné Y. Toto rozdělení pravděpodobnosti náhodných odchylek, resp. vysvětlované proměnné, je také požadováno za účelem testování hypotéz o regresních parametrech nebo určovaní intervalů spolehlivosti.

Očekávané hodnoty potom určíme jako

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

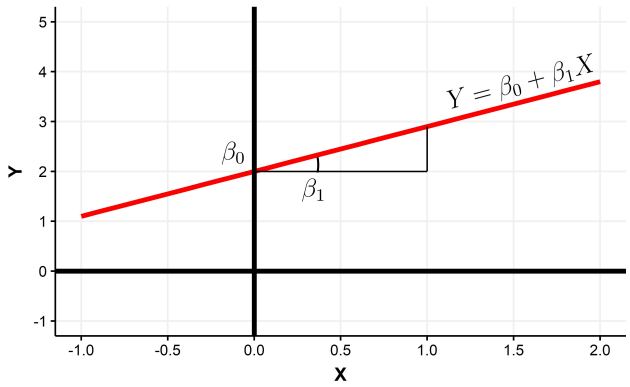
Odchylky, které vypočítáme ze vztahu

$$\mathbf{e_i} = \mathbf{Y} - \hat{\mathbf{Y}}$$

se nazývají rezidua.

1.1.5. Interpretace regresních parametrů

Na obrázku 1.1 je graficky znázorněna interpretace regresních parametrů pro jednoduchý lineární model (1.1). Parametr β_0 lze interpretovat jako očekávanou hodnotu vysvětlované proměnné v případě, že $x = 0$. Parametr β_1 je pak směrnice regresní přímky, respektive změna očekávané hodnoty Y při jednotkové změně vysvětlující proměnné x. Nutno doplnit, že pokud definiční obor náhodné veličiny x neobsahuje nulovou hodnotu, potom parametr β_0 nemá žádnou interpretaci.



Obrázek 1.1: Interpretace parametrů jednoduchého regresního modelu

Interpretace parametrů vícenásobné regrese (1.3) je zobecněním výše uvedeného odstavce. Parametr β_0 je očekávaná hodnota vysvětlované proměnné v případě, že všechny vysvětlující proměnné nabývají hodnoty nula. Parametr

β_i pro $i \in \{1, 2, \ldots, p-1\}$ znamená změnu vysvětlované proměnné při jednotkové změně i-té vysvětlující proměnné za předpokladu, že ostatní vysvětlující proměnné zůstanou konstantní.

1.1.6. Kategorické proměnné v regresních modelech

V regresních modelech nemusí být v roli vysvětlující proměnné pouze spojité náhodné veličiny, ale i náhodné veličiny nabývající několika diskrétních hodnot [6]. Může jít například o pohlaví (muž, žena), barvu (bílá, červená, ...) a další.

Kategorické proměnné, které nabývají pouze dvou hodnot, jsou označovány jako dichotomické proměnné. Pro ilustraci uvažujme proměnnou pohlaví, která nabývá hodnot *muž* a *žena*. Pokud ji chceme využít v regresním modelu, je třeba vytvořit umělou proměnnou x_2 tak, že

$$x_2 = \begin{cases} 0 & \text{proměnná nabývá hodnoty } muž \\ 1 & \text{proměnná nabývá hodnoty } žena. \end{cases}$$

Hodnota *muž* se označuje jako referenční hodnota.

Regresní model by potom mohl mít například následující předpis:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{1.9}$$

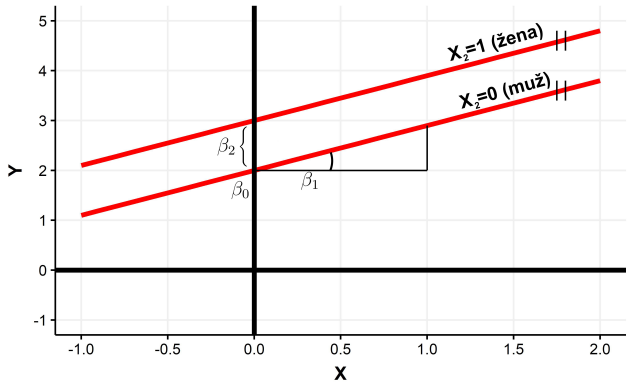
kde Y je vysvětlovaná proměnná (např. hmotnost), x_1 je spojitá vysvětlující proměnná (např. výška) a x_2 je umělá proměnná vyjadřující, zda se jedná o muže či ženu.

Interpretace regresního parametru β_1 je v tomto případě očekávaná změna vysvětlované proměnné při jednotkové změně proměnné x_1 a neměnné hodnotě kategorické proměnné. Význam parametru β_2 lze jednoduše odvodit, když položíme hodnotu x_2 rovnu jedné (tj. jde o ženu). Potom platí

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 1 = (\beta_0 + \beta_2) + \beta_1 x_1,$$

a tedy hodnota β_2 vyjadřuje rozdíl mezi očekávanou hodnotou závislé proměnné pro kategorickou proměnnou nabývající svou referenční hodnotu a pro kategorickou proměnnou nabývající svou druhou hodnotu při x_1 konstantní. V případě

našeho příkladu by tedy šlo o rozdíl očekávané hodnoty hmotnosti mezi muži a ženami při nulové výšce. Interpretace parametrů je zřejmá taky z obrázku 1.2.



Obrázek 1.2: Interpretace parametrů modelu s kategorickou proměnnou

Vytvoření modelu s dichotomickou proměnnou se liší od vytvoření dvou modelů pro jednotlivé hodnoty této proměnné [6]. V případě jednoho modelu se totiž předpokládá, že regresní přímky pro obě hodnoty kategorické proměnné budou mít stejný sklon a navíc hodnoty vysvětlované proměnné pro tyto dvě kategorie mají stejný rozptyl. To v případě dvou modelů neplatí.

Pro vysvětlující proměnné s n hodnotami (kategoriemi) je potřeba vytvořit $n-1$ umělých proměnných. Následující příklad vysvětluje model se spojitou proměnnou, kategorickou proměnnou se 3 hodnotami a dichotomickou kategorickou proměnnou.

Je dán model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22} + \beta_3 x_3, \tag{1.10}$$

kde Y je vysvětlovaná proměnná *prodejní cena šály*,
x_1 je vysvětlující proměnná *délka*,
x_2 je vysvětlující proměnná *země výroby* nabývající hodnot Česko, Čína a Japonsko, tzn.

$$x_{21} = \begin{cases} 1 & \text{vyrobena v Číně} \\ 0 & \text{jinak,} \end{cases}$$

$$x_{22} = \begin{cases} 1 & \text{vyrobena v Japonsku} \\ 0 & \text{jinak,} \end{cases}$$

x_3 je vysvětlující proměnná *materiál* nabývající hodnot bavlna a hedvábí, tzn.

$$x_3 = \begin{cases} 1 & \text{hedvábí} \\ 0 & \text{bavlna.} \end{cases}$$

Referenční kategorie jsou tedy zřejmě země výroby Česko a materiál bavlna. Parametr β_0 reprezentuje očekávanou hodnotu prodejní ceny při nulové délce, pro šálu vyrobenou v Česku z bavlny (referenční kategorie). Hodnota β_1 je očekávaná změna ceny při jednotkové změně délky pro ostatní prediktory neměnné. Parametr β_{21} reprezentuje rozdíl mezi očekávanými hodnotami prodejní ceny pro šálu z Číny a z Česka, parametr β_{22} obdobně pro šálu z Japonska a z Česka (v obou případech pro nulovou délku a materiál v referenční kategorii). A konečně parametr β_3 je rozdíl mezi očekávanou hodnotou ceny pro šálu z bavlny a z hedvábí z Česka s nulovou délkou.

1.2. Zobecněné lineární modely

Zobecněné lineární modely patří taktéž do regresní analýzy, pro komplexnost tohoto tématu jim však v této práci bude vyčleněna samostatná kapitola. Následující odstavce čerpají z [1], [8], [3], [9], [14], [15], [16].

1.2.1. Složky zobecněných lineárních modelů

Podle Agrestiho [1] se zobecněný lineární model skládá ze tří složek:

- náhodná složka,
- nenáhodná (systematická) složka,
- spojovací funkce.

Roli *náhodné složky* hraje vysvětlovaná proměnná (označována Y) s daným rozdělením pravděpodobnosti. Pozorování (hodnoty) vysvětlované proměnné jsou považována za nezávislá. Mohou nabývat binárních hodnot (např. úspěch/neúspěch), celých hodnot (např. počet úspěchů) či jakékoliv reálné hodnoty. V prvním případě mluvíme o náhodné veličině s binomickým rozdělením pravděpodobnosti, ve druhém se náhodná veličina často modeluje pomocí Poissonova rozdělení pravděpodobnosti a nakonec se může jednat o náhodnou veličinu s normálním rozdělením pravděpodobnosti.

Vysvětlující proměnné zastupují *systematickou složku*. V modelu se vyskytují ve formě tzv. lineárního prediktoru, tj.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}.$$

Poslední složkou je *spojovací funkce*, která spojuje střední (očekávanou) hodnotu rozdělení pravděpodobnosti náhodné složky a lineární prediktor. Tento vztah lze zapsat jako

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}, \tag{1.11}$$

případně maticově

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

kde $g(\cdot)$ je spojovací funkce a $\mu = E(Y)$. Nejjednodušší spojovací funkcí je identita, tj. $g(\mu) = \mu$. Z toho plyne souvislost zobecněných lineárních modelů s obyčejnými (nezobecněnými) lineárními modely, tedy že zobecněný lineární model se spojovací funkcí identity je obyčejným lineárním modelem. Dalšími spojovacími funkcemi můžou být například logaritmus ($g(\mu) = ln(\mu)$), potom mluvíme o loglineárních modelech, či logitová spojovací funkce ($g(\mu) = ln(\frac{\mu}{1-\mu})$), která je používána u logistických modelů.

Zobecněné lineární modely zobecňují obyčejné (nezobecněné) lineární modely ve dvou aspektech:

1) umožňují vysvětlované proměnné Y mít jiné rozdělení pravděpodobnosti než normální (jakékoliv z exponenciální třídy rozdělení [8]),

2) umožňují použití jiné funkce spojující očekávanou hodnotu náhodné složky a lineární prediktor než je funkce identity (spojovací funkce musí být monotónní a diferencovatelná [8]).

Metody obyčejných (nezobecněných) lineárních modelů je tedy možné použít pouze pro Y s přibližně normálním rozdělením a s konstantním rozptylem. Toho je možné dosáhnout například různými transformacemi vysvětlované proměnné. Tyto kroky však nejsou v praxi vždy žádoucí, proto je dobré využít aparátu zobecněných lineárních modelů, který pro odhad parametrů využívá metody maximální věrohodnosti a nevyžaduje normalitu Y. Spojovací funkce tedy není volena za účelem dosažení normality ani stabilizace rozptylu.

1.2.2. Exponenciální třída rozdělení

Jak už bylo uvedeno výše, metoda zobecněných lineárních modelů předpokládá, že náhodná složka Y pochází z exponenciální třídy rozdělení [8]. Hustotu jejich rozdělení pravděpodobnosti lze obecně zapsat jako

$$f_Y(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1.12}$$

kde $a(\cdot), b(\cdot)$ a $c(\cdot)$ jsou specifické funkce a θ, ϕ jsou parametry. Pokud je ϕ známý, nazýváme model s parametrem θ modelem v kanonické formě. Tato parametrizace je jen jednou z možných parametrizací předpisu, v různých zdrojích je možné najít i další.

V krátkosti bude uveden princip metody maximální věrohodnosti. Tato metoda vychází z tzv. funkce věrohodnosti [8]. Pokud $f(y;\theta)$ je hustota rozdělení náhodné veličiny y při daných parametrech θ, pak funkce věrohodnosti je definována jako

$$L(\theta, y) = f(y;\theta) = \prod_i f_i(y_i;\theta_i),$$

tedy jde o funkci parametrů při daných datech, kterou je často výhodné zapsat ve tvaru součinu. Cílem metody je maximalizovat funkci $L(\theta, y)$, resp. její logaritmus, tj.

$$ln(L(\theta, y)) = l(\theta, y) = \sum_i f(y_i;\theta_i).$$

Pro distribuce pocházející z exponenciální třídy rozdělení lze při daných datech (podle [15]) zapsat funkci věrohodnosti (1.12) jako

$$L(\theta, \phi; y) = \prod_{i=1}^{n} f_i(y_i;\theta_i,\phi) = exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right\},$$

a tedy logaritmická funkce věrohodnosti je

$$l(\theta, \phi; y) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi). \tag{1.13}$$

V následujících odstavcích bude odvozen obecný předpis pro střední hodnotu a rozptyl náhodných veličin s distribucí z exponenciální třídy rozdělení. Odvození vychází z logaritmické funkce věrohodnosti a platnosti následujících výrazů [16].

Položme $l = \ln(f_Y)$, kde f_Y je hustota náhodné veličiny Y, pak platí

$$l' = \frac{f'}{f} \quad \text{a} \quad l'' = \frac{f''}{f} - \left(\frac{f'}{f}\right)^2. \tag{1.14}$$

Pro hustotu distribuce náhodné veličiny platí

$$\int_{-\infty}^{\infty} f(y;\theta)dy = 1, \tag{1.15}$$

a tedy po derivaci obou stran rovnice platí

$$\int_{-\infty}^{\infty} f'(y;\theta)dy = 0 \quad \text{a} \quad \int_{-\infty}^{\infty} f''(y;\theta)dy = 0. \tag{1.16}$$

Dále za použití definice očekávané hodnoty, platnosti (1.14) a (1.15) lze ukázat, že

$$E(l') = \int_{-\infty}^{\infty} l' \cdot f(y;\theta)dy = \int_{-\infty}^{\infty} \frac{f'}{f} \cdot f(y;\theta)dy = \int_{-\infty}^{\infty} f'dy = 0. \tag{1.17}$$

Ze stejných důvodů také platí

$$E(l'') = \int_{-\infty}^{\infty} l'' \cdot f(y;\theta)dy = \int_{-\infty}^{\infty} \frac{f''}{f} \cdot f(y;\theta)dy - \int_{-\infty}^{\infty} \left(\frac{f'}{f}\right)^2 \cdot f(y;\theta)dy = -E\left[(l')^2\right], \tag{1.18}$$

a tedy platí

$$E\left[l''\right] + E\left[(l')^2\right] = 0. \tag{1.19}$$

Nyní zbývá vyjádřit první a druhou derivaci logaritmické věrohodnostní funkce hustoty (1.13):

$$l' = \frac{\partial l}{\partial \theta} = \frac{Y - b'(\theta)}{a(\phi)} \tag{1.20}$$

a

$$l'' = \frac{\partial^2 l}{\partial^2 \theta} = -\frac{b''(\theta)}{a(\phi)}. \tag{1.21}$$

Z (1.18) a (1.20) plyne

$$E\Big(\frac{Y - b'(\theta)}{a(\phi)}\Big) = 0$$

$$\frac{1}{a(\phi)}\Big(E(Y) - b'(\theta)\Big) = 0,$$

a tedy

$$E(Y) = b'(\theta) = \mu. \tag{1.22}$$

Podobně z (1.19) a (1.21)

$$var(Y) = b''(\theta)a(\phi). \tag{1.23}$$

První člen součinu $b''(\theta)$ závisí pouze na kanonickém parametru, nazývá se funkcí rozptylu a značí se $V(\mu)$. Druhý člen součinu $a(\phi)$ je v modelech pro data vyjadřující počet roven jedné [3].

1.2.3. Odhad parametrů

Jak bylo zmíněno dříve, parametry zobecněných lineárních modelů se odhadují maximalizací logaritmické funkce věrohodnosti. V této části ukážeme, že stejných odhadů lze dosáhnout pomocí iterativní metody vážených nejmenších čtverců (často označována IWLS z anglického iterative weighted least squares), která je právě často používána ve statistických softwarech [8]. Následující odvození vznikla s pomocí [15].

V případě zobecněných lineárních modelů v metodě IWLS použijeme místo proměnné Y proměnnou z , která je definována jako

$$z_i = \eta_i + (y_i - \mu_i)g'(\mu_i), \tag{1.24}$$

kde $\eta_i = \mathbf{X_i}\boldsymbol{\beta}$ je lineární prediktor pro i-té pozorování, $\mu_i = E(Y_i)$ a $g'(\cdot)$ je první derivace spojovací funkce. Algoritmus IWLS probíhá v následujících čtyřech krocích:

I) položíme $\mu_i = y_i$

II) (1) určíme $\mu_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$

(2) $z_i = \mathbf{X}_i\boldsymbol{\beta} + (y_i - \mu_i)g'(\mu_i)$

III) aktualizujeme $\hat{\boldsymbol{\beta}} = (\mathbf{X'WX})^{-1}\mathbf{X'WZ}$, kde $\mathbf{W}$ je diagonální matice vah

$$\mathbf{W} = diag\left\{\left[(g'(\mu_i))^2 a_i(\phi)V(\mu_i)\right]^{-1}\right\}.$$

IV) opakujeme krok II) a III) dokud metoda nezkonverguje.

V první iteraci tohoto algoritmu můžeme vynechat první část druhého kroku, protože hodnotu μ_i jsme položili rovnu y_i. V druhé části druhého kroku pak získáme $z_i = \mathbf{X}_i\boldsymbol{\beta}$, protože $(y_i - \mu_i) = (y_i - y_i) = 0$.

Při odvozování ekvivalence algoritmu IWLS a metody maximalizace funkce věrohodnosti vyjdeme ze soustavy normálních rovnic pro IWLS [15]

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{Z}$$
$$(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}(\mathbf{X}\boldsymbol{\beta} + \mathbf{G}(\mathbf{Y}-\boldsymbol{\mu})),$$

kde $\mathbf{G} = diag\,\{g'(\mu_i)\}$

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} + \mathbf{X}'\mathbf{W}\mathbf{G}(\mathbf{Y} - \boldsymbol{\mu})$$
$$\mathbf{0} = \mathbf{X}'\mathbf{W}\mathbf{G}(\mathbf{Y} - \boldsymbol{\mu}).$$

Pak i-tou složku vektoru $\mathbf{X}'\mathbf{W}\mathbf{G}(\mathbf{Y} - \boldsymbol{\mu})$ můžeme zapsat jako

$$0 = \sum_{j=1}^{n} x_{ij} \frac{g'(\mu_i)}{[g'(\mu_i)]^2\, a_i(\phi) V(\mu_i)} (y_j - \mu_j)$$

$$0 = \sum_{j=1}^{n} x_{ij} \left[a_i(\phi) V(\mu_i)\right]^{-1} (y_j - \mu_j) \tag{1.25}$$

Nyní odvodíme rovnici pro výpočet optimálních parametrů z pohledu funkce věrohodnosti. Za předpokladu, že Y je náhodná veličina s rozdělením pravděpodobnosti z exponenciální třídy s hustotou (1.12), její logaritmická funkce věrohodnosti má tvar

$$l(\theta, \phi; y) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{1.26}$$

Optimální parametry pak získáme určením maxima funkce (1.26), tj. řešením soustavy rovnic parciálních derivací

$$\frac{\partial l}{\partial \beta_i} = 0 \text{ pro } i = 0, 1, \ldots p - 1. \tag{1.27}$$

Všimněme si, že logaritmická funkce věrohodnosti je v rovnici (1.26) zapsána jako součet jednotlivých složek l_i. Pro zjednodušení se nyní zaměříme na i-tou složku logaritmické funkce věrohodnosti. Za využití pravidla pro derivaci složené funkce je tedy potřeba vyřešit

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = 0. \tag{1.28}$$

Jednotlivé složky součinu jsou odvozeny takto:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a_i(\phi)} = \frac{Y_i - \mu_i}{a_i(\phi)} \tag{1.29}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{V(\mu_i)} \tag{1.30}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)} \tag{1.31}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left[\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}\right] = x_{ij} \tag{1.32}$$

Pro výraz (1.30) jsme využili toho, že platí $\mu_i = b'(\theta_i)$ a funkce $b(\cdot)$ je striktně rostoucí, tudíž ji lze invertovat. Potom $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = V(\mu_i)$ (viz (1.22) a (1.23)).

Pro výraz (1.31) jsme využili vztahu $g(\mu_i) = \eta_i$, tj. že lineární prediktor je s očekávanou hodnotou vysvětlované proměnné propojen pomocí spojovací funkce $g(\cdot)$.

Rovnici (1.28) lze tedy za využití výrazů (1.29) - (1.32) zapsat jako

$$\frac{\partial l_i}{\partial \beta_j} = \frac{Y_i - \mu_i}{a_i(\phi)} \cdot \frac{1}{V(\mu_i)} \cdot \frac{1}{g'(\mu_i)} \cdot x_{ij}. \tag{1.33}$$

Pro určení optimálních parametrů je tedy potřeba na základě (1.26) vyřešit soustavu rovnic

$$\sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i)} \cdot \frac{y_i - \mu_i}{a_i(\phi) V(\mu_i)} = 0,$$

což odpovídá rovnici (1.25) pro i-tou složku. Tímto je tedy dokázáno, že výsledky metody IWLS jsou v případě zobecněné lineární regrese ekvivalentní s maximalizací logaritmické funkce věrohodnosti.

1.2.4. Modely s Poissonovým rozdělením

Vysvětlující proměnné často vyjadřují počty, může se jednat například o počet červených lentilek v krabičce či počet prodaných kusů za dané období [1]. Takováto náhodná složka modelu tedy nabývá nezáporných celých hodnot a nejčastěji se modeluje pomocí Poissonova rozdělení pravděpodobnosti.

Poissonovo rozdělení

Poissonovo rozdělení pravděpodobnosti je unimodální rozdělení zešikmené doprava s definičním oborem $\{0, 1, 2, \dots\}$ a s pravděpodobnostní funkcí

$$P(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \text{pro } k \ = 0, 1, 2, \dots .$$

Parametrem Poissonova rozdělení je λ, který zároveň reprezentuje střední hodnotu a rozptyl, tj. pro $Y \sim Po(\lambda)$ platí

$$E(Y) = var(Y) = \lambda.$$

Důkaz tohoto tvrzení je následující [4]:

$$E(Y) = \sum_{k=0}^{\infty} k \cdot P(Y = k) = \sum_{k=0}^{\infty} k\, e^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Byla využita substituce $j = k - 1$ a Taylorův rozvoj exponenciální řady

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}.$$

Rozptyl určíme takto:

$$var(Y) = E(Y^2) - [E(Y)]^2 = E(Y^2) - E(Y) + E(Y) - [E(Y)]^2 = E\left[Y(Y-1)\right] + E(Y) - [E(Y)]^2 .$$

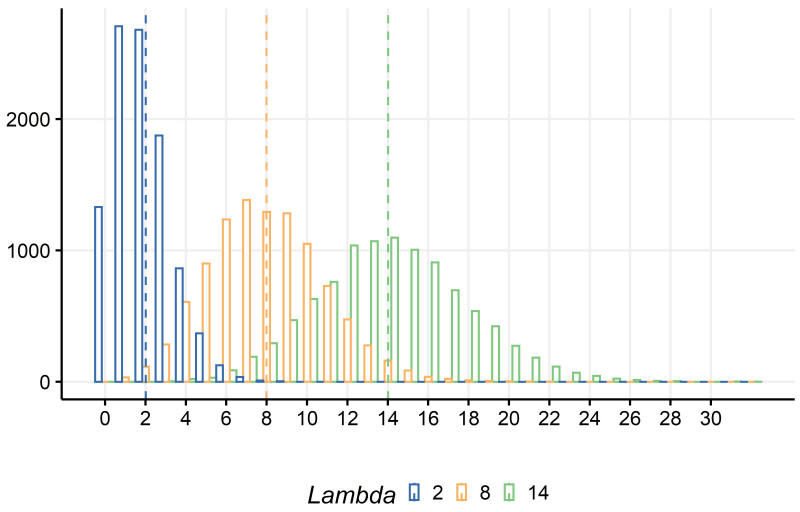
Dále platí (s využitím obdobné substituce a Taylorova rozvoje jako výše)

$$E\left[Y(Y-1)\right] = \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2 e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2,$$

a tedy

$$var(Y) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Ze vztahu mezi střední hodnotou a rozptylem vyplývá, že náhodná veličina s Poissonovým rozdělením s vysokou střední hodnotou má také velký rozptyl. Poissonovo rozdělení s různými hodnotami parametru λ ilustruje obrázek 1.3.



Obrázek 1.3: Histogramy Poissonových distribucí s různými hodnotami λ, svislé čáry značí očekávané hodnoty

Podle [8] platí, že pro $\lambda \to \infty$ má náhodná veličina $\frac{Y-\lambda}{\sqrt{\lambda}}$ přibližně rozdělení $N(0,1)$.

Skutečnost, že Poissonovo rozdělení je z exponenciální třídy rozdělení lze ukázat úpravami funkce pravděpodobnosti ([16]):

$$P(Y=k) = e^{-\lambda}\frac{\lambda^k}{k!} = \exp\left\{\ln\left(e^{-\lambda}\frac{\lambda^k}{k!}\right)\right\} = \exp\left\{k \cdot \ln(\lambda) - \lambda - \ln(k!)\right\},$$

z čehož při použití substituce $\eta = \ln(\lambda)$ lze vyjádřit kanonickou formu

$$P(Y = k) = \exp\{k \cdot \eta - e^{\eta} - \ln(k!)\}.$$

Při porovnání s obecným předpisem pravděpodobnostní funkce distribucí z exponenciální třídy rozdělení (1.12) je zřejmé, že $a(\phi) = 1$ (což platí pro všechny distribuce popisující počet, viz kapitola 1.2.1) a $b(\theta) = e^{\eta}$. Potom můžeme ze vztahů (1.22) a (1.23) obecně platných pro distribuce z exponenciální třídy rozdělení odvodit očekávanou hodnotu a rozptyl náhodné veličiny mající Poissonovo rozdělení jako

$$E(Y) = b'(\theta) = e^{\eta} = e^{\ln(\lambda)} = \lambda,$$
$$var(Y) = b''(\theta) = e^{\eta} = e^{\ln(\lambda)} = \lambda.$$

Model s Poissonovým rozdělením

Obecný zápis zobecněných lineárních modelů (1.11) lze v případě modelů s Poissonovým rozdělením zapsat jako

$$ln(\mu) = ln(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1},$$

a tedy spojovací funkcí v tomto případě je přirozený logaritmus. Proto se tyto modely také nazývají loglineární [8].

Interpretace regresních koeficientů je díky logaritmické spojovací funkci jiná než u lineárních modelů uvedených v kapitole 1.1.5 a bude po vzoru [1] vysvětlena na Poissonovském modelu s jednou vysvětlující proměnnou ve formě

$$ln(\mu) = \beta_0 + \beta_1 x. \tag{1.34}$$

Je zřejmé, že

$$\mu = e^{\beta_0 + \beta_1 x}$$

a také platí

$$\frac{\mu(x+1)}{\mu(x)} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = \frac{e^{\beta_0} e^{\beta_1 x} e^{\beta_1}}{e^{\beta_0} e^{\beta_1 x}} = e^{\beta_1}.$$

Tedy pro očekávanou hodnotu vysvětlované proměnné v bodě $x + 1$ platí

$$\mu(x + 1) = e^{\beta_1}\mu(x),$$

což vyjádřeno slovy znamená, že jednotková změna vysvětlující proměnné x má multiplikativní efekt na vysvětlovanou proměnnou a tedy pro $\beta_1 > 0$ nabývá e^{β_1}-krát vyšší hodnotu, případně pro $\beta_1 < 0$ nabývá e^{β_1}-krát nižší hodnotu.

1.2.5. Nadměrný rozptyl

Jak bylo zmíněno v předchozí kapitole, modely s Poissonovým rozdělením předpokládají, že očekávaná hodnota náhodné složky je rovna jejímu rozptylu [1]. V praxi je však tato situace spíše výjimečná, mnohem častěji u dat pozorujeme tzv. nadměrný rozptyl [9] (anglicky *overdispersion*), tj. situaci, kdy

$$var(Y) > E(Y).$$

Častým důvodem nadměrného rozptylu je pozitivní korelace mezi pozorovanými hodnotami vysvětlované proměnné, případně určité shlukování dat, kdy každá skupina má jiný rozptyl [3]. Nadměrný rozptyl může způsobit podhodnocení rozptylu odhadnutých parametrů. Analýza nadměrného rozptylu může probíhat pomocí Pearsonovy χ^2 statistiky [14]

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

která má $n - k$ stupňů volnosti a potom se odhad disperzního parametru

$$\sigma^2 = \frac{\chi^2}{n - k}$$

porovnává s hodnotou jedna. Mohou nastat tři možnosti

- $\sigma^2 = 1$, potom je podmínka na očekávanou hodnotu a rozptyl splněna,
- $\sigma^2 > 1$, potom model trpí nadměrným rozptylem (overdispersion),

- $\sigma^2 < 1$, potom model trpí malým rozptylem (underdispersion, vyskytuje se velmi zřídka).

Nadměrný rozptyl může být takzvaně zjevný [3], to je v případech, kdy například do modelu nezahrneme důležitou vysvětlující proměnnou, data obsahují odlehlé hodnoty, v modelu je potřeba doplnit interakce či je nutné prediktor transformovat. Řešení tohoto druhu nadměrného rozptylu je zřejmé, jde o doplnění chybějícího prediktoru do modelu, provedení transformace a podobně. Pokud problém nadměrného rozptylu stále přetrvává, je potřeba jej řešit jinými metodami, které jsou blíže vysvětleny v [3]. Jako příklady uveďme:

- škálování čtvercových chyb,
- využití robustních odhadů rozptylu,
- kvazi Poissonovské metody,
- negativně-binomické rozdělení.

Nadměrný rozptyl je také často způsoben velkým množstvím nulových hodnot v datech. Tyto situace se nejčastěji modelují pomocí tzv. zero-inflated modelů. Více o těchto modelech se lze dočíst například v [3].

1.2.6. Modely s negativně-binomickým rozdělením

Pokud je v modelu s Poissonovým rozdělením porušen předpoklad pro rovnost očekávané hodnoty a rozptylu, využívá se k modelování dat nejčastěji negativně-binomické rozdělení [3]. Existuje několik forem tohoto rozdělení např.

- NB2 ($V = \mu + \alpha\mu^2$)
- NB1 ($V = \mu + \alpha\mu$)
- NB-C (kanonický)
- omezený NB2, NB1, NB-C

- ZINB (zero-inflated)

Tyto a další formy negativně-binomického rozdělení jsou více rozebrány v [3]. V následujících odstavcích bude popsáno nejčastěji používané rozdělení označované NB2. Funkce rozptylu je v tomto případě ve formě

$$V(\mu) = \mu + \alpha\mu^2$$

a tedy rozptyl je funkcí očekávané hodnoty.

Negativně-binomický model (dále jen NB2) lze odvodit ze smíšeného Poisson-gamma modelu nebo z vlastností exponenciální třídy rozdělení. V prvním případě je základem předpis Poissonova rozdělení s nehomogenním rozptylem, který má gamma rozdělení se střední hodnotou jedna. V této práci bude více rozebrán druhý postup.

Odvození negativně-binomického rozdělení

Exponenciální třída rozdělení je charakterizována zejména společnou formou zápisu hustoty (1.12). Mezi její členy patří taktéž binomické rozdělení s obecně známou pravděpodobnostní funkcí

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y},$$

kde p je pravděpodobnost úspěchu v n nezávislých pokusech [3]. Geometrické rozdělení, které je taky členem exponenciální třídy, modeluje počet neúspěchů před prvním úspěchem a NB2 rozdělení modeluje počet neúspěchů před r-tým úspěchem. Název NB2 rozdělení tedy plyne z toho, že binomické rozdělení popisuje počet úspěchů, zatímco NB2 počet neúspěchů. Spojitost mezi NB2 a Poissonovým rozdělením plyne taktéž ze série nezávislých Bernoulliho pokusů. Poissonovo rozdělení totiž modeluje pravděpodobnost nastání y úspěchů v sérii n pokusů, kdy n je nekonečně veliké a pravděpodobnost nastání úspěchu p je malá. Je tedy potřeba znát pouze očekávaný počet úspěchů, který označujeme jako λ.

Pravděpodobnostní funkci NB2 rozdělení lze odvodit z následující úvahy [3]: uvažujme y jako počet neúspěchů před r-tým úspěchem a situaci, kdy r-tý úspěch

nastal v x-tém pokusu, přičemž pravděpodobnost nastání úspěchu značíme p. Předchozích $r-1$ úspěchů tedy mohlo nastat v jakémkoliv z $x-1$ pokusů, a tedy r-tý úspěch v x-tém pokusu můžeme získat $\binom{x-1}{r-1}$ způsoby. V každém způsobu může dojít k $x-r$ neúspěchům s pravděpodobností $p^r(1-p)^{x-r}$ a platí, že celkový počet pokusů x je roven součtu počtu neúspěchů y a počtu úspěchů r, tj. $x = y + r$. S využitím těchto úvah můžeme zapsat pravděpodobnostní funkci NB2 jako

$$P(Y = y_i) = \binom{y_i + r - 1}{r - 1} p^r (1-p)^{y_i}, \text{pro } y_i = 0, 1, \ldots. \tag{1.35}$$

Tuto pravděpodobnostní funkci je možné přepsat do formy odpovídající exponenciální třídě rozdělení (viz (1.12)):

$$f(y; p, r) = \exp\left\{ y_i \ln(1-p) + r \ln(p) + \ln\binom{y_i + r - 1}{r - 1} \right\}.$$

V tomto případě platí

$$\theta = \ln(1-p), \tag{1.36}$$

a tedy

$$p = 1 - \exp(\theta).$$

Funkce $b(\theta)$ má pro NB2 tvar

$$b(\theta) = -r \ln(p) = -r \ln(1 - \exp(\theta))$$

a funkce $a(\phi)$ je rovna 1 (jako pro všechny modely popisující počty, viz kapitola 1.2.1). Z vlastností exponenciální třídy rozdělení odvozených v kapitole 1.2.1 lze tedy pro NB2 určit očekávanou hodnotu jako

$$b'(\theta) = \frac{\partial b}{\partial \theta} = -r \frac{1}{1 - \exp(\theta)} (-\exp(\theta)),$$

což lze s využitím vztahu (1.36) zapsat jako

$$b'(\theta) = r \frac{1-p}{1-(1-p)} = r \frac{1-p}{p} = E(Y).$$

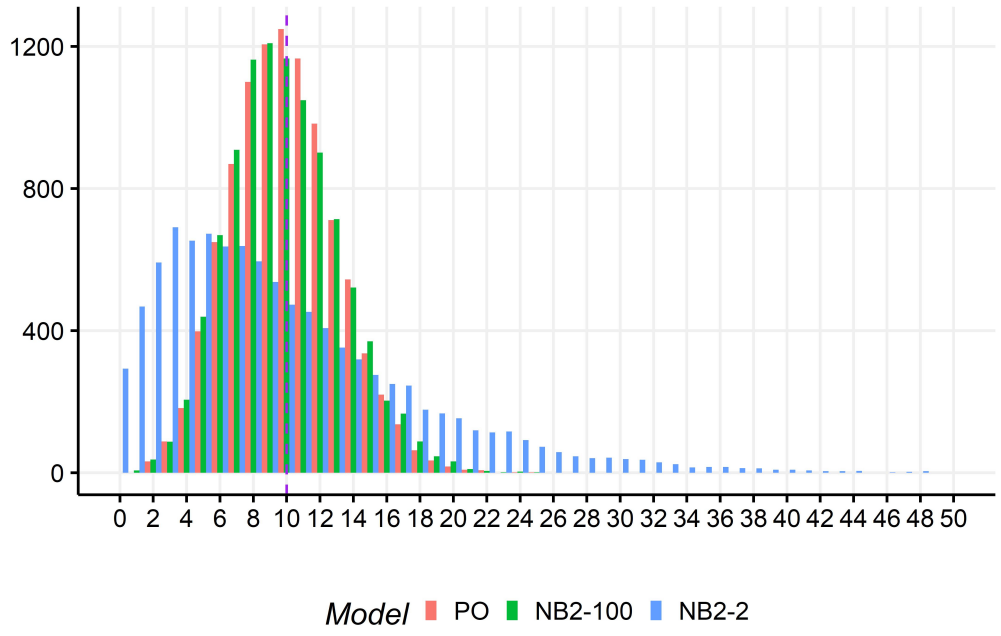
Pro rozptyl potom platí

$$b''(\theta) = r\frac{\exp(\theta)(1-\exp(\theta)) - \exp(\theta)(-\exp(\theta))}{(1-\exp(\theta))^2} = r\frac{\exp(\theta) - \exp(\theta)^2 + \exp(\theta)^2}{(1-\exp(\theta))^2}$$

a pokud opět využijeme vztah (1.36), platí

$$b''(\theta) = r\frac{1-p}{p^2} = var(Y).$$

Histogramy na obrázku 1.4 porovnávají náhodnou veličinu s Poissonovým rozdělením se střední hodnotou 10 a dvě náhodné veličiny s NB2 rozdělením s tou samou střední hodnotou. První náhodná veličina označená NB2-100 má hodnotu parametru r rovnu 100 a druhá označená NB2-2 má $r = 2$. I z histogramů je zřejmé, že platí pro $r \to \infty$ NB2 rozdělení konverguje k Poissonovu rozdělení [3].



Obrázek 1.4: Histogramy Poissonovy distribuce v porovnání s NB2 s $r = 100$ a $r = 2$, očekávaná hodnota u všech distribucí je 10

1.3. Hooke-Jeevesova metoda optimalizace

Ačkoliv je tato práce zaměřená na aplikaci zobecněných lineárních modelů v praxi, v této kapitole bude popsána numerická metoda optimalizace funkce více proměnných, jelikož vedla k citelnému zlepšení výsledků predikce. Tato kapitola je pouze letmým přehledem a čerpá z [7].

Základní úlohou nepodmíněné optimalizace je nalézt $\mathbf{x}^* \in \mathbb{R}^n$ takové, že $f(\mathbf{x}^*)$ je lokální či globální minimum funkce $f(\mathbf{x})$. Protože v praktické části půjde o minimalizaci nediferencovatelné funkce (tj. funkce, která nemá derivaci), není potřeba vysvětlovat pojmy jako jsou gradient či stacionární bod. Pro úplnost však uvedeme definici lokálního minima úlohy nepodmíněné optimalizace.

Definice 1.3.1. *Nechť $f(\mathbf{x})$ je daná funkce. Bod $\mathbf{x}^* \in \mathbb{R}^n$ se nazývá bodem lokálního minima funkce f, jestliže existuje $\varepsilon > 0$ takové, že*

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \qquad \forall \mathbf{x} \in B(\mathbf{x}^*, \varepsilon),$$

kde $B(\mathbf{x}^*, \varepsilon) = \{\mathbf{y} \in \mathbb{R}^n : 0 \leq ||\mathbf{x}^* - \mathbf{y}|| < \varepsilon\}$, přičemž $||.||$ značí euklidovskou normu vektoru v $\mathbb{R}^n$.

Pokud platí

$$f(\mathbf{x}^*) < f(\mathbf{x}) \qquad \forall \mathbf{x} \in B(\mathbf{x}^*, \varepsilon), \mathbf{x} \neq \mathbf{x}^*,$$

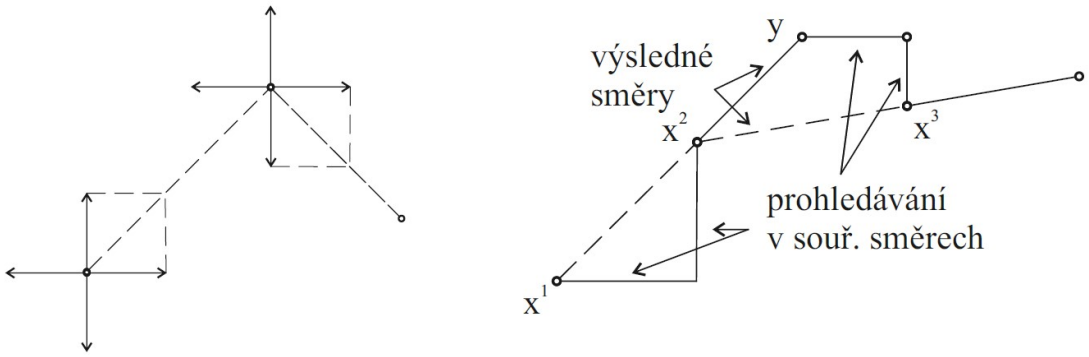
pak hovoříme o ostrém lokálním minimu.

V praktické části této práce bude využita jedna z metod optimalizace nediferencovatelné funkce více proměnných, a to konkrétně Hooke-Jeevesova metoda. Dalšími možnostmi nalezení optima jsou například Nelder-Meadova, Powelova či Rosenbrockova metoda. Hooke-Jeevesova metoda (dále jen H-J) má dvě fáze:

1. prohledávání funkce ve směrech daných souřadnicovými osami,

2. stanovení výsledného směru pro minimalizaci na základě výsledků 1. fáze.

Schéma 1.5 graficky znázorňuje H-J metodu v dvojrozměrném prostoru.



Obrázek 1.5: Dvě fáze H-J metody (převzato z [7])

Následující algoritmus H-J metody je převzat z [7].

Algoritmus Hooke-Jeevesovy metody

- Inicializace

  zvolíme startovací bod $\mathbf{x}^1$ a $\varepsilon > 0$ pro ukončovací kritérium

  položíme $\mathbf{y}^1 = \mathbf{x}^1, k \ = j = 1$

- Fáze 1

  určíme α_j tak, že $f(\mathbf{y}^j + \alpha_j \mathbf{e}_j) = \min_{\alpha \in \mathbb{R}} f(\mathbf{y}^j + \alpha \mathbf{e}_j)$

  kde $\mathbf{e}_j, j = 1, \ldots, n$, jsou souřadnicové směry

  položíme $\mathbf{y}^{j+1} = \mathbf{y}^j + \alpha_j \mathbf{e}_j$

  IF $j < n$ THEN

  položíme $j = j + 1$ a opakujeme fázi 1

  ELSE $(j = n)$

  položíme $\mathbf{x}^{k+1} = \mathbf{y}^{n+1}$

  IF $\left|\left|\mathbf{x}^{k+1} - \mathbf{x}^k\right|\right| < \varepsilon$ THEN

  výsledek je $\bar{\mathbf{x}} = \mathbf{x}^{k+1}$ STOP

  ELSE přejdeme na fázi 2

  END IF

  END IF

- Fáze 2

  položíme $\mathbf{d} = \mathbf{x}^{k+1} - \mathbf{x}^k$

  určíme $\hat{\lambda}$ tak, že $f(\mathbf{x}^{k+1} + \hat{\lambda}\mathbf{d}) = \min_{\lambda \in \mathbb{R}} f(\mathbf{x}^{k+1} + \lambda \mathbf{d})$

  položíme $\mathbf{y}^1 = \mathbf{x}^{k+1} + \hat{\lambda}\mathbf{d}$

  položíme $j = 1, k \ = k + 1$

  přejdeme na fázi 1

V případě praktické části této práce jde o minimalizaci v p-rozměrném prostoru (p je počet regresních parametrů), vektor $\mathbf{x}$ z algoritmu je tedy vektorem parametrů $\boldsymbol{\beta}$ a f je funkce MAPE (více v kapitole 2.9.3).

Kapitola 2

Praktická část

2.1. Úvod

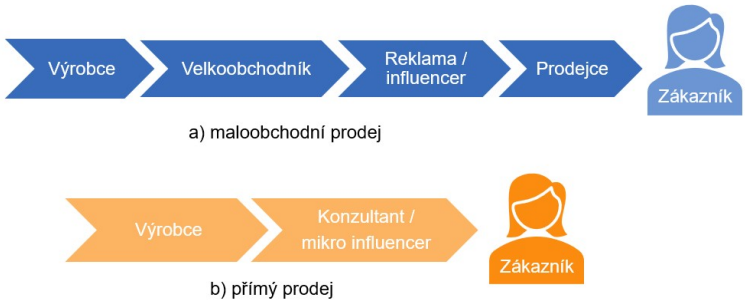
Téma této diplomové práce vzešlo z projektu Oriflame IT academy, kterou pořádala v roce 2018 firma ORIFLAME SOFTWARE s. r. o. Cílem původního projektu bylo prozkoumat potenciál využití machine learning (strojového učení) v prostředí Microsoft Azure Machine Learning Studia pro zlepšení predikčního modelu firmy ORIFLAME SOFTWARE s. r. o. Tento záměr se následně přetransformoval v dohodu o napsání této diplomové práce, která je ze strany ORIFLAME SOFTWARE s. r. o. zaštiťována panem Jaroslavem Kratochvílem (Senior manager) a panem Janem Milanem (Team leader).

2.2. Firma Oriflame

Oriflame je kosmetická společnost založená v roce 1967 ve Švédsku bratry Jonasem a Robertem af Jochnickovými [10]. V současné době působí na více než 60 světových trzích, na ten český vstoupila v roce 1990. Velmi si zakládá na udržitelnosti a environmentální zodpovědnosti. Své portfolio tedy staví na produktech vyvinutých v souladu s vědou i přírodou a zdůrazňují jejich bezpečnost, spolehlivost a efektivitu [13].

Prodejní strategie společnosti Oriflame je založena na tzv. multilevel marketingu (také síťový marketing). Ten vychází z principu přímého prodeje, kdy nezávislí prodejci nabízejí zboží firmy přímo zákazníkům, získávají provize z pro-

deje a zároveň jsou motivováni rozšiřovat prodejní síť dané firmy tím, že profitují také z aktivity prodejců, které získali pro firmu [18]. Rozdíl mezi maloobchodním a přímým prodejem zobrazuje diagram na obrázku 2.1.



Obrázek 2.1: Rozdíl mezi maloobchodním a přímým prodejem [2]

V prodejní síti společnosti Oriflame se pro prodejce používá název Brand Partners (v této práci budou označováni dříve používaným výrazem konzultanti). V roce 2019 pro Oriflame pracovalo kolem 3 milionů nezávislých konzultantů, z toho třetina v Asii a Turecku, třetina ve státech tzv. Společenství nezávislých národů (SNS, anglicky CIS), čtvrtina v Evropě a Africe a zbývajících cca 10 % v Latinské Americe [2]. Největší část prodejů v roce 2019 proběhla na území Asie a Turecka (34 %). V současné době se Oriflame zaměřuje na přesunutí prodeje na internetovou platformu. V roce 2019 bylo prostřednictvím internetu provedeno 96 % objednávek (v rámci celosvětového prodeje) [2]. To ovšem neznamená, že by se tento způsob prodeje obešel bez konzultantů. Zákazníci si objednávají jejich prostřednictvím přes internetové rozhraní. Zároveň konzultanti používají mobilní aplikaci, která obsahuje informace o produktech, aktuální nabídku v katalogu nebo také přehled o stavu objednaného zboží [11].

Konzultanti Oriflamu nabízejí výrobky na základě nabídek v katalogu. Každý rok se dělí do 17 kampaní a každá kampaň odpovídá jednomu katalogu. Katalog se skládá z jednotlivých nabídek, které jsou charakterizovány vlastnostmi blíže popsanými v kapitole 2.4. Hodnoty těchto vlastností jsou použity v této práci pro tvorbu predikčního modelu. Konzultanti jsou k prodeji motivováni provizí,

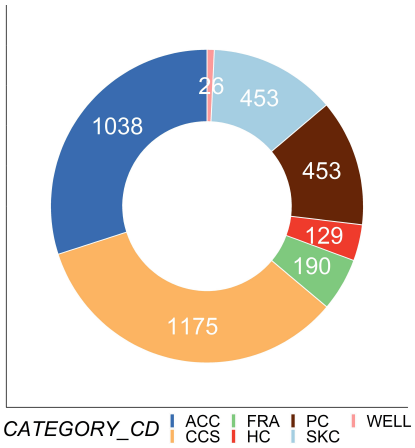
která se primárně odvíjí od rozdílu mezi katalogovou cenou a cenou pro konzultanta (30% zisk) [12]. Každý konzultant má navíc možnost budovat svou kariéru a postupovat na kariérním žebříčku tím, že získává nové konzultanty a vede je k prodeji a budování další sítě konzultantů. Tím se může vypracovat z pozice kosmetického poradce až na pozici direktora spojenou s dalšími finančními odměnami.

2.3. Výrobky

Výrobky společnosti Oriflame jsou rozděleny do 7 kategorií:

- péče o pleť (SKC) zahrnuje např. krémy, čistící produkty, pleťové masky či opalovací péči,
- dekorativní kosmetika (CCS) je sortiment od rtěnek, make-upů, přes oční stíny až po laky na nehty,
- vůně (FRA) nabízí toaletní a parfémované vody, deodoranty i vůně pro muže,
- péče o tělo (PC) pokrývá hydratační péči o tělo, ruce i nohy, sprchové gely, tělové peelingy či péči o zuby,
- péče o vlasy (HC) zahrnuje šampony, kondicionéry i stylingovou péči,
- doplňky (ACC) pokrývají sortiment od náušnic, náramků, přes hodinky až po peněženky a šátky,
- wellness (WELL) nabízí převážně doplňky stravy.

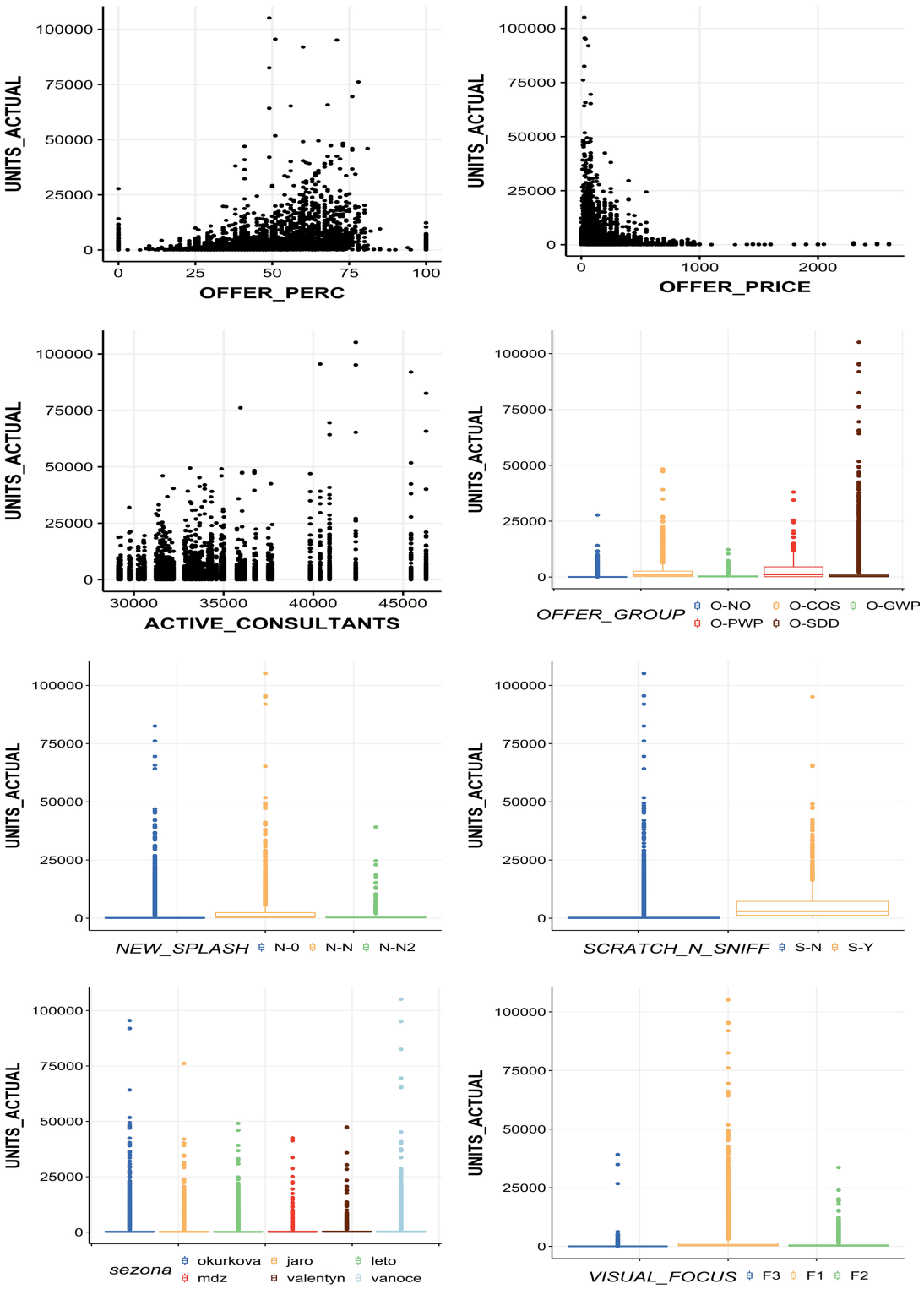
Ve zpracovávané datové sadě je 3 464 výrobků. Graf 2.2 zobrazuje počty výrobků v jednotlivých kategoriích. Kategorie se dále dělí na sektory a segmenty. Toto jemnější dělení již však není v této práci uvažováno.



Obrázek 2.2: Počty výrobků v jednotlivých kategoriích

2.4. Data

Firma ORIFLAME SOFTWARE s. r. o. poskytla datový soubor obsahující informace o prodejích na jednom z trhů v letech 2015-2018. Data jsou citlivým majetkem ORIFLAME SOFTWARE s. r. o., a proto není možné je přiložit (byla podepsána dohoda o mlčenlivosti). Použity budou jen souhrnné hodnoty a výstupní modely. Celkem šlo o 50 677 záznamů, každý záznam je informace o prodeji jedné položky v jedné kampani. Na obrázku 2.3 můžeme vidět přehled závislosti počtu prodaných kusů na jednotlivých použitých proměnných. Dále jsou v této kapitole popsány všechny proměnné podrobněji.

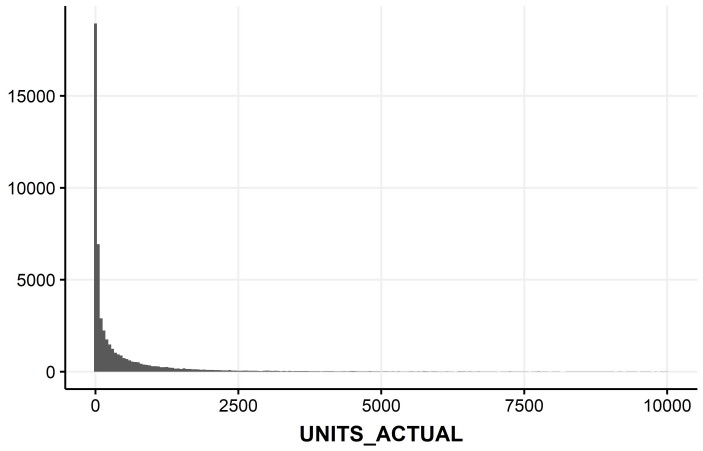


Obrázek 2.3: Přehled vztahů mezi počtem prodaných výrobků a jednotlivými proměnnými

Vysvětlovaná proměnná Model vytvořený v této práci má za úkol predikovat prodej výrobků firmy Oriflame. Skutečné hodnoty prodejů jsou v datech popsány ve dvou proměnných: UNITS_ACTUAL a UPA_ACTUAL. První zmíněná proměnná udává počet prodaných kusů daného výrobku v dané kampani. Souhrnné hodnoty UNITS_ACTUAL obsahuje tabulka 2.1. Rozdělení této proměnné je silně zešikmené (viz histogram na grafu 2.4). Jelikož se jedná o počty, budeme tuto proměnnou považovat za proměnnou s Poissonovým rozdělením.

Tabulka 2.1: Rozdělení hodnot UNITS_ACTUAL

Min.	$Q_{0.25}$	Medián	Průměr	$Q_{0.75}$	Max.
0	10	69	645.454	456	105 152



Obrázek 2.4: Histogram UNITS_ACTUAL

Oproti tomu UPA_ACTUAL v sobě zahrnuje také informaci o počtu prodejců v dané kampani. UPA je zkratkou slov Units Per Active, v překladu Jednotky na aktivního (prodejce), a lze ji vypočítat jako podíl počtu prodaných kusů a počtu aktivních konzultantů. Tato proměnná již tedy není celočíselná, a proto ji nelze považovat za proměnnou s Poissonovým rozdělením. Souhrnné hodnoty UPA_ACTUAL obsahuje tabulka 2.2.

Tabulka 2.2: Rozdělení hodnot UPA_ACTUAL

Min.	$Q_{0.25}$	Medián	Průměr	$Q_{0.75}$	Max.
0	0.0003	0.002	0.019	0.013	2.482

Zatímco ve firmě ORIFLAME SOFTWARE s. r. o. se pomocí modelů predikuje hodnota UPA_ACTUAL, v této práci jsou všechny modely zaměřené na predikci UNITS_ACTUAL. Navíc byla každá hodnota zvýšena o jedničku, abychom mohli vypočítat metriku Odchylka (viz kapitola 2.6, problém s nulami ve jmenovateli). Tato úprava byla použita i pro data v testovací sadě (viz kapitola 2.9), a tudíž nemá žádný vliv na vypočtené metriky.

OFFER_PRICE Proměnná OFFER_PRICE obsahuje cenu výrobku v katalogu. Do modelu je tato proměnná zařazená po škálování, tj. v tomto případě jsou všechny hodnoty poděleny průměrnou cenou výrobků v letech 2015-2017 (testovací sada, viz dále). Souhrnné hodnoty OFFER_PRICE obsahuje tabulka 2.3.

Tabulka 2.3: Rozdělení hodnot OFFER_PRICE (bez škálování)

Min.	$Q_{0.25}$	Medián	Průměr	$Q_{0.75}$	Max.
0	99	169	235.297	299	2 599

OFFER_PERC Hodnota slevy v procentech je zaznamenána v proměnné OFFER_PERC. Pro použití v modelu byly hodnoty vyděleny číslem 100 tak, aby se pohybovaly v intervalu od nuly do jedné. Mezi hodnotami OFFER_PERC a UNITS_ACTUAL je možné pozorovat kvadratickou závislost (viz graf na obrázku 2.3), a proto byla tato proměnná do modelu zařazená v druhé mocnině. Souhrnné hodnoty OFFER_PERC obsahuje tabulka 2.4.

Tabulka 2.4: Rozdělení hodnot OFFER_PERC (bez škálování)

Min.	$Q_{0.25}$	Medián	Průměr	$Q_{0.75}$	Max.
0	0	25	22.408	41	100

ACTIVE_CONSULTANTS Proměnná ACTIVE_CONSULTANTS udává počet aktivních konzultantů v daném období. Aktivní konzultant je ten, který se aktivně zapojuje do prodeje. Do modelů byla tato proměnná zahrnutá po vydělení číslem 10 000. Souhrnné hodnoty ACTIVE_CONSULTANTS obsahuje tabulka 2.5.

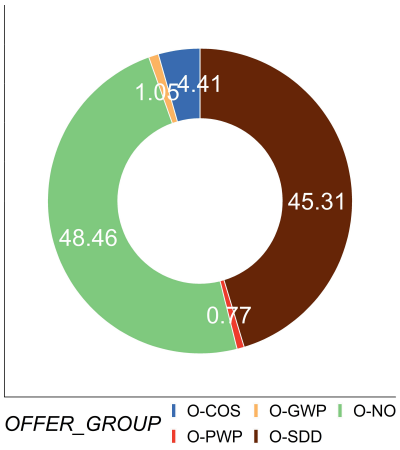
Tabulka 2.5: Rozdělení hodnot ACTIVE_CONSULTANTS (bez škálování)

Min.	$Q_{0.25}$	Medián	Průměr	$Q_{0.75}$	Max.
29 140	31 649	33 599	34 267.900	35 740	46 281

OFFER_GROUP Nabídky produktů jsou rozděleny do několika skupin, které zachycuje proměnná OFFER_GROUP. Jde o kategorickou proměnnou s pěti úrovněmi:

- No offer (bez nabídky, NO) - výrobek se v dané kampani nenabízí se slevou,
- Straight Discount (přímá sleva, SDD) - výrobek se v dané kampani nabízí se slevou (procento z původní ceny),
- Combine offer/set (kombinovaná nabídka/sada, COS) - koupí daného výrobku v kampani získává kupující možnost zakoupení jiného výrobku se slevou,
- Gift with purchase (GWP), Purchase with purchase (PWP) - nabídky omezené nákupem jiného výrobků či jinou podmínkou.

Jelikož jsou úrovně GWP a PWP zastoupeny minimálně (viz graf 2.5), v modelu byly použity pouze první tři úrovně.

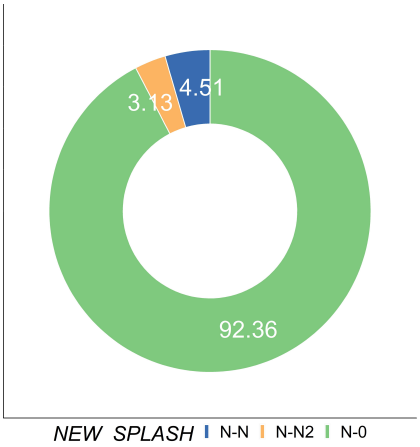


Obrázek 2.5: Zastoupení úrovní OFFER_GROUP v datech (%)

NEW_SPLASH U nabídek je taky sledováno, jestli jsou na trh uvedeny jako novinky, případně jsou nabízeny podruhé či opakovaně. Tuto skutečnost zachycuje kategorická proměnná NEW_SPLASH, která má 3 úrovně:

- novinka (N-N): výrobek je v dané kampani nabízen poprvé,
- druhý prodej (N-N2): výrobek je v dané kampani nabízen podruhé,
- opakovaně (N-O): výrobek byl nabízen již minimálně dvakrát.

Zastoupení jednotlivých úrovní v datech zobrazuje graf 2.6.

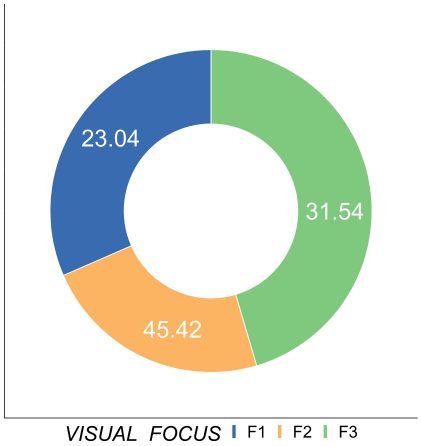


Obrázek 2.6: Zastoupení úrovní NEW_SPLASH v datech (%)

VISUAL_FOCUS Proměnná VISUAL_FOCUS obsahuje údaje o velikosti obrázku výrobku v katalogu dané kampaně. Opět jde o kategorickou proměnnou, která má 3 úrovně:

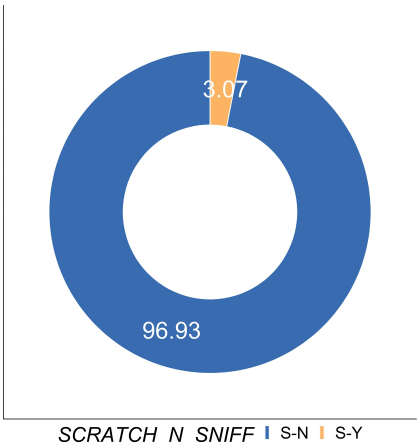
- přes celou stranu (F1),
- přes půlku strany (F2),
- malý (F3).

Zastoupení jednotlivých úrovní v datech zobrazuje graf 2.7.



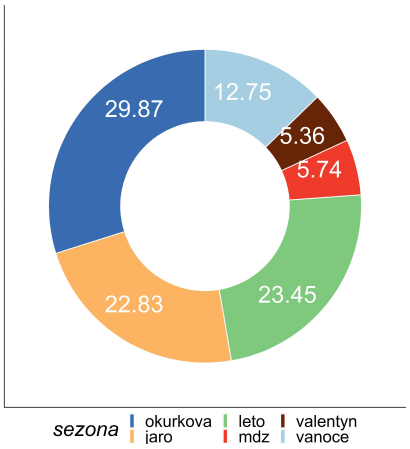
Obrázek 2.7: Zastoupení jednotlivých úrovní VISUAL_FOCUS v datech (%)

SCRATCH_N_SNIFF Některé výrobky jsou v katalogu doprovázeny tzv. scratch&sniff oblastí, která slouží pro otestování vůně daného výrobku. Jedná se pouze o výrobky kategorií vůně, péče o vlasy, péče o tělo a péče o pleť. Kategorická proměnná SCRATCH_N_SNIFF tedy nabývá pouze dvou hodnot - ano (S-Y) a ne (S-N) a jejich zastoupení v datech zobrazuje graf 2.8.



Obrázek 2.8: Zastoupení jednotlivých úrovní SCRATCH_N_SNIFF v datech (%, pouze data pro kategorie FRA, HC, PC, SKC)

sezona Proměnná sezona byla vytvořená za účelem rozdělit 17 kampaní v každém roce do delších časových celků. Má 6 úrovní: jaro, Valentýn, MDŽ, léto, Vánoce a okurková sezóna (období mezi ostatními sezónami). Zastoupení jednotlivých sezón v datech zobrazuje graf 2.9.

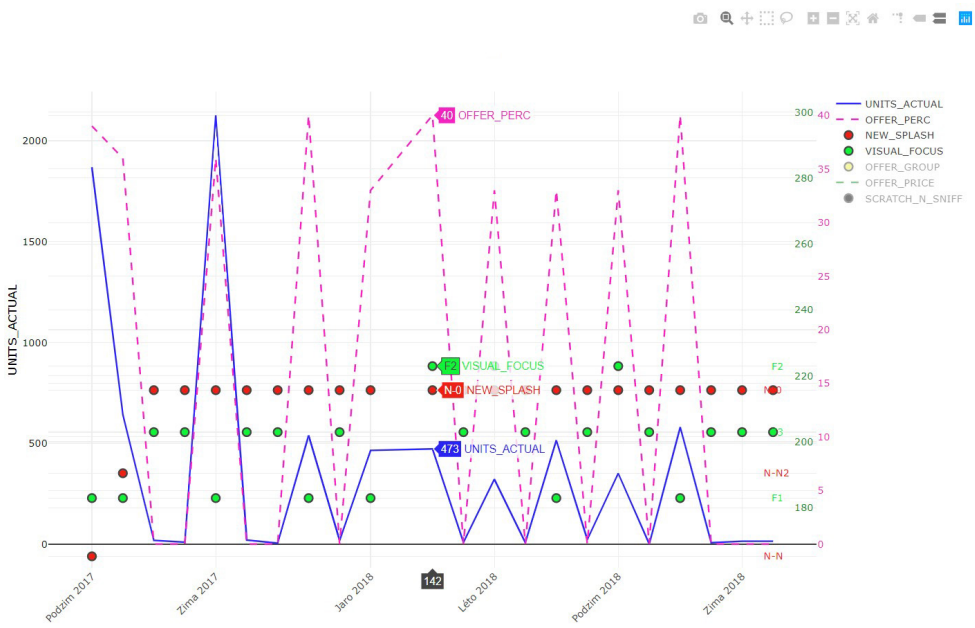


Obrázek 2.9: Zastoupení jednotlivých sezón v datech (%)

PROD_ID Každý výrobek má vlastní identifikátor, tzv. PROD_ID. Tato proměnná je v modelu zahrnuta jako kategorická, a tím je zajištěno, aby základní úroveň modelu (bez dalších prediktorů) byla pro každý výrobek jiná.

2.5. Vizualizace dat

Prvním krokem před prací na samotném modelu byla tvorba nástroje pro vizualizaci dat. K tomu byl použit programovací jazyk Python. Příklad jedné vizualizace je na obrázku 2.10. Tento graf zobrazuje prodej daného výrobku společně s ostatními proměnnými. Jde o interaktivní graf, ve kterém je možné přibližovat vybrané oblasti, zapínat či vypínat zobrazení jednotlivých proměnných (rozkliknutím v legendě vpravo nahoře) či přejížděním kurzoru po ploše grafu zobrazovat hodnoty vybraných prediktorů pro zvolený časový okamžik.



Obrázek 2.10: Interaktivní vizualizace dat

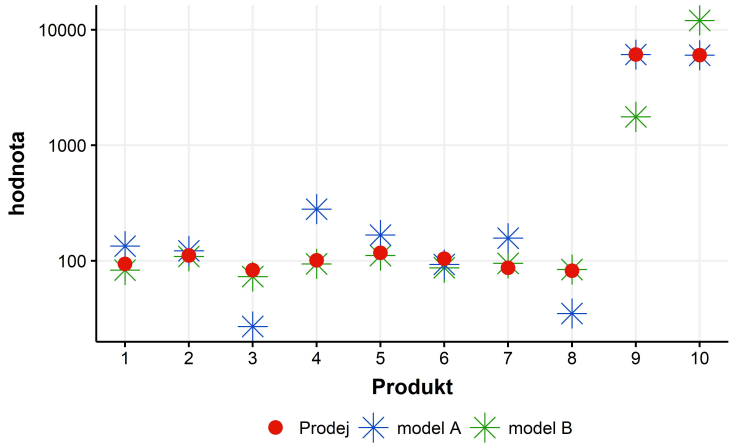
2.6. Metriky

Vytvořené modely je potřeba porovnávat pomocí nějaké metriky. Ve firmě ORIFLAME SOFTWARE s. r. o. používají vlastní metriku zvanou accuracy (česky přesnost, acc). Její hodnota se určí následovně:

$$accuracy = 1 - \frac{\sum_{i=1}^{n} |ACT_i - FIT_i|}{\sum_{i=1}^{n} ACT_i}, \tag{2.1}$$

kde ACT_i jsou skutečné hodnoty prodeje, FIT_i jsou predikované hodnoty prodeje a n je počet výrobků ve skupině. Hodnoty této metriky jsou shora ohraničené hodnotou jedna a zespoda neohraničené. Platí, že čím větší accuracy, tím lepší model. Interpretace této metriky je ze statistického pohledu problematická. Jednak proto, že může nabývat záporných hodnot, a také proto, že chyby na výrobcích s vysokými prodeji mají na hodnotu accuracy tak velký vliv, že se chyby na výrobcích s malými prodeji téměř vůbec neprojeví.

Toto chování je zřejmé z grafu 2.11. Ten ukazuje smyšlené prodeje (červené puntíky) 10 výrobků v jedné kampani a predikci dvou různých modelů: modelu A (modré hvězdy) a modelu B (zelené hvězdy). Je důležité si uvědomit, že graf je v logaritmickém měřítku, protože prvních 8 výrobků se prodává výrazně méně, než poslední dva. Model A predikuje s velkými chybami na prvních 8 výrobcích a s malými chybami na posledních dvou výrobcích. Hodnota accuracy pro model A je 0.9628. Model B predikuje s malými chybami na prvních 8 výrobcích, avšak s velkými chybami na posledních dvou výrobcích. Hodnota accuraccy pro tento model je 0.1940, i když v 80 % případů predikoval docela přesně. Proto by modely založené na této metrice byly v podstatě ovlivněny pouze několika výrobky s nejvyššími prodeji a nebyly by validní pro ostatní výrobky, kterých je většina.



Obrázek 2.11: Demonstrace chování accuracy při různých predikcích

V této práci je za účelem porovnání modelů použita střední absolutní relativní chyba predikce (MAPE, anglicky Mean absolute percentage error). Její předpis je následující:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|ACT_i - FIT_i|}{ACT_i}. \tag{2.2}$$

Metriku MAPE lze interpretovat jako průměrné procentuální absolutní odchýlení predikcí od skutečné hodnoty a nabývá hodnot od nuly do nekonečna. V tomto případě je lepším modelem ten s nižší hodnotou MAPE.

Vzhledem k rozumnějšímu výpočtu a interpretovatelnosti metriky MAPE je tvorba modelů v této práci založena právě na ní, pro porovnání výsledků predikcí firmy ORIFLAME SOFTWARE s. r. o. s našimi výsledky jsou však uváděny i hodnoty accuracy.

2.7. Predikce v prostředí Python

V programovacím jazyku Python vznikl taktéž predikční algoritmus, tzv. naivní model, aby bylo možné posoudit, zda modely vytvořené na základě statistické teorie jsou lepší než naivní metody. Tento algoritmus je založen na vyhledávání nejpodobnější nabídky v historii každého výrobku a funguje pouze pro výrobky mající v letech 2015-2017 více než 2 prodeje a v roce 2018 alespoň jeden prodej. Podobnost nabídek byla posuzována následovně: u každé z proměnných uvedených v tabulce 2.6 se vyhodnotí, zda jejich hodnota v minulých výskytech spadá do intervalu specifikovaného v téže tabulce, případně zda se u kategorických prediktorů shoduje kategorie. Hodnota podobnosti dané historické nabídky s nabídkou, pro kterou je počítána predikce, se pak určí jako vážený průměr přes všechny sledované prediktory (váhy uvedeny v tabulce 2.6). Pokud jsou pro nabídku, pro jejíž prodej je počítána predikce, nalezeny dvě nabídky v historii se stejnou mírou podobnosti, predikcí prodeje se stává průměrná hodnota těchto shodných nabídek.

Tabulka 2.7 uvádí smyšlený příklad výpočtu predikce prodeje pro nabídku danou hodnotami ve sloupci Pred. nabídka. Sloupce N1 a N2 uvádějí hodnoty

Tabulka 2.6: Predikce hledáním podobné nabídky

Proměnná	Interval	Váha
OFFER_PERC	±5%	0.9
CAMPAIGN_MONTH	±3 měsíce	0.6
OFFER_GROUP	shoda	0.7
OFFER_CODE	shoda	0.5
OFFER_PRICE	±100 Kč	0.8
SCRATCH_N_SNIFF	shoda	0.7
VISUAL_FOCUS	shoda	0.6

nabídek v historii a jejich prodaný počet kusů, sloupce V1 a V2 obsahují hodnotu 1, pokud dané prediktory spadají do intervalu, a hodnotu 0, pokud nespadají. Z těchto dvou nabídek by jako predikce byla použita nabídka číslo 1, protože má vyšší hodnotu shody (1*0.9+1*0.6+1*0.7+1*0.5+1*0.8+1*0.7=4.2). Predikovaný počet prodaných kusů je tedy v tomto případě roven 380.

Tabulka 2.7: Příklad fungování predikčního algoritmu

Proměnná	Váha	Pred. nabídka	N1	V1	N2	V2
OFFER_PERC	0.9	40	45	1	10	0
CAMPAIGN_MONTH	0.6	duben	květen	1	listopad	0
OFFER_GROUP	0.7	SDD	SDD	1	GWP	0
OFFER_CODE	0.5	SD	SD	1	GWP	0
OFFER_PRICE	0.8	350	280	1	400	1
SCRATCH_N_SNIFF	0.7	Y	Y	1	N	0
VISUAL_FOCUS	0.6	F1	F2	0	F3	0
UNITS_ACTUAL		?	380		200	
Hodnota shody				4.2		0.8

Tabulka 2.8 udává hodnoty MAPE vypočítané na nabídkách v roce 2018 pro jednotlivé kategorie.

Tabulka 2.8: MAPE pro jednotlivé kategorie

ACC	CCS	FRA	HC	PC	SKC	WELL
1.969	1.414	11.629	1.264	8.828	7.191	0.545

2.8. Software

Modely v této diplomové práci byly vytvořeny v programu R verze 3.6.1 v prostředí Rstudio verze 1.2.5001. Dále byl v již menším rozsahu použit programovací jazyk Python v prostředí JupyterLab. Tabulka 2.9 uvádí přehled využívaných balíčků v programu R, stručný popis jejich zaměření a vybrané funkce. Grafy byly tvořeny pomocí šablony z volně dostupného internetového fóra RPubs (viz [5]).

Tabulka 2.9: Přehled použitých balíčků a vybraných funkcí v R

Balíček	Popis	Vybrané využívané funkce
dfoptim	Optimalizační algoritmy nediferencovatelných funkcí	`hjk()`: Hooke-Jeevesův algoritmus optimalizace nediferencovatelné funkce
stats	Statistické funkce (defaultní balíček)	`glm()`: tvorba zobecněných lineárních modelů `model.matrix()`: tvorba matice designu modelu
MASS	Další statistické funkce a datasety	`glm.nb()`: tvorba modelu pro negativně-binomické rozdělení
Stargazer	Formátování tabulek pro LaTeX a další	`stargazer()`
tidyverse	Soubor balíčků pro jednotnou manipulaci s daty	
ggplot2	Systém pro tvorbu grafiky	`ggplot()`

Data, na kterých jsou postaveny modely v následujících kapitolách, nemohou být s touto prací publikována, proto také není přiložen kód, který z větší části obsahuje přípravu a manipulaci se samotnými daty. Modely byly tvořeny pomocí zabudovaných funkcí zmíněných v tabulce výše (`glm()` s parametrem `family=poisson(link = "log")`, `glm.nb()` a `hjk()`).

2.9. Tvorba modelu

Tvorba modelu je založena na principu křížové validace (cross-validation) jakožto způsobu zajišťování efektivity vytvořeného modelu [19]. Účelem je vyhnout se tzv. overfittingu či underfittingu. Overfitting je situace, kdy model příliš dobře funguje na datech, na kterých byl vytvořen, avšak predikce pro nová data nejsou příliš přesné. Tento model není schopen dostatečně zobecňovat. Na druhou stranu v případě underfittingu je model až příliš obecný.

Nejjednodušším případem křížové validace je rozdělení dat do dvou sad - na trénovací a testovací sadu [19]. Obvykle se celá datová sada rozdělí na dvě části náhodně v poměru 70:30 či 80:20. Následně je na trénovací sadě vytvořen model, který se použije pro predikci hodnot testovací sady. Data z testovací sady nesmí být použita pro tvorbu modelu. Posledním krokem je výpočet zvolené metriky pomocí predikcí na testovací sadě a skutečných hodnot výstupní proměnné.

Druhou variantou křížové validace je K-fold validace [19]. Spočívá v opakovaném dělení celé datové sady na trénovací a testovací část a výpočtu dané metriky. Výstupem je průměrná hodnota zvolené metriky. Tato metoda je méně vychýlená než předchozí, ale má větší výpočetní náročnost.

V případě predikce prodeje byla nasimulována skutečná situace, ve které by byl predikční model použit, tj. v určitém časovém okamžiku je známa historie prodeje za nějakou dobu a je vytvořen model pro predikci do budoucna. Nedává tedy smysl data rozdělovat náhodně. Z důvodu omezené výpočetní síly byla zvolena první metoda validace. Za historická data byly považovány údaje z let 2015-2017 (trénovací sada) a za budoucnost data z roku 2018 (testovací sada). V tomto případě byl poměr trénovací ku testovací sadě zhruba 70:30.

V problematice predikce prodejů je potřeba počítat také s uváděním nových výrobků na trh. Jelikož je proměnná PROD_ID použita v modelu jako kategorická a tedy každý výrobek má svůj parametr (blíže vysvětleno v kapitole 1.1.6), je potřeba přiřadit nějakou hodnotu regresního parametru těm výrobkům, které se nevyskytují v trénovací sadě. Tento problém je v této práci vyřešen následovně: výrobkům, pro které neexistuje natrénovaný regresní parametr, je při predikci

přiřazena průměrná hodnota parametrů příslušících k proměnné PROD_ID výrobků z odpovídající kategorie.

V této práci je v rámci jedné formy modelu vytvořeno 7 oddělených modelů - pro každou kategorii jeden (viz kapitola 2.3). Předpokladem je, že se prodej produktů v každé kategorii chová odlišně.

2.9.1. Předpis regresního modelu

Obecný předpis základního modelu je následující:

$$ln(E(UNITS\_ACTUAL+1)) = konst.+OFFER\_GROUP+OFFER\_PRICE+ \\ +OFFER\_PERC^2 + NEW\_SPLASH + VISUAL\_FOCUS+ \\ +SCRATCH\_N\_SNIFF + sezona + PROD\_ID. \quad (2.3)$$

V této práci jsou prezentovány tři typy modelů:

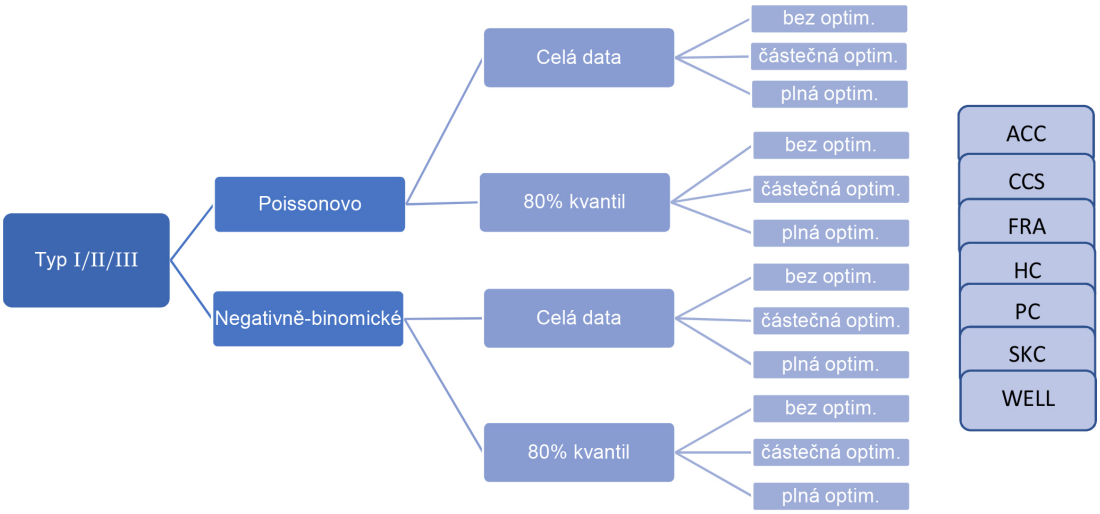
I) základní ve formě podle (2.3) (ve druhé mocnině pouze OFFER_PERC),

II) základní obohacený o proměnnou ACTIVE_CONSULTANTS,

III) s proměnnou ACTIVE_CONSULTANTS, ve druhé mocnině OFFER_PERC a OFFER_PRICE.

Příslušné spojité proměnné jsou použity ve škálované formě tak, jak je zmíněno v kapitole 2.4. Nejedná se o lineární modely, ale o regresní modely s Poissonovým případně negativně-binomickým rozdělením (viz kapitoly 1.2.4 a 1.2.6).

Na obrázku 2.12 je schéma všech vytvořených modelů. Pro každý typ (tj. I, II a III) vznikl jak model s Poissonovým tak s negativně-binomickým rozdělením, ty se dále dělí na modely na všech datech a na podmnožině dat (konkrétně na 80% kvantilu prodejů v jednotlivých kategoriích, více v kapitole 2.9.2), a následně se ještě dělí na modely bez optimalizace, s optimalizací částečnou a s optimalizací na všech parametrech (více v kapitole 2.9.3). Je také třeba zdůraznit, že v rámci každého modelu byl vytvořen zvlášť model pro každou kategorii výrobků. Celkem tedy vzniklo 252 vektorů regresních parametrů (3*2*2*3*7) a k nim přísluší 252 predikcí na testovací sadě.

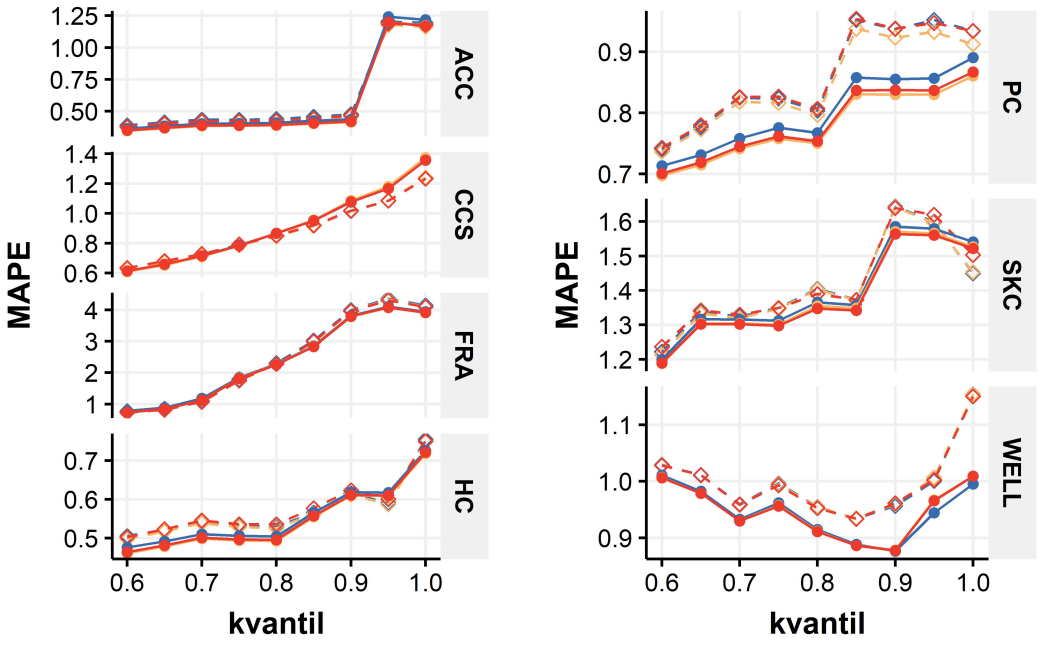


Obrázek 2.12: Schéma vytvořených modelů

2.9.2. Modelování na podmnožinách dat

V praxi se predikční modely nepoužívají pro predikci všech nabídek, ale pouze těch, které nejsou nikterak extrémní (prodeje nejsou příliš vysoké). Tyto extrémní hodnoty poté predikuje expert na základě svých zkušeností a znalostí. Tento postup je v práci zaveden tzv. modelováním na podmnožinách dat. Z testovací sady dat je pro vytvoření modelu použita podmnožina mající hodnoty prodeje menší či rovny určité hodnotě kvantilu prodejů v dané kategorii. Jinými slovy, byl vytvořen model na i % dat ($i \in \langle 60, 100 \rangle$) s nejnižší hodnotou UNITS_ACTUAL. Tento postup byl zopakován celkem 9krát pro každou kategorii výrobků (pro $i = \{60, 65, \ldots, 95, 100\}$). Následně byla zvolena hodnota i tak, aby metrika MAPE dosahovala ve všech kategoriích optimální výše. Na základě grafu 2.13 byla zvolena úroveň 80% kvantilu. Jde o úroveň, kdy se MAPE signifikantně zmenšilo u všech kategorií a zároveň nedošlo k přílišné ztrátě informací (čím více dat je ořezáno, tím méně informací je využito k tvorbě modelu). V grafu je možné pozorovat, že se odchylky pro jednotlivé typy modelů příliš neliší (v některých případech se křivky překrývají tak, že není možné rozeznat všechny 3 barvy odpovídající 3 typům modelů). Pokud chceme porovnávat vzájemně modely PO a NB, hodnoty MAPE vycházejí odlišně, nicméně křivky mají podobné ten-

dence. Tímto postupem byla z trénovací sady vyřazena extrémní data, a to dává potenciál k lepším predikcím použitého modelu (za předpokladu predikce extrémních hodnot odborníkem).



Obrázek 2.13: Hodnoty MAPE při různých úrovních kvantilů (modrá = typ I, žlutá = typ II, červená = typ III, přerušovaná čára= PO model, plná čára= NB model)

2.9.3. Optimalizace

Použité funkce v softwaru R fungují na principu hledání minima pomocí iterativní metody vážených čtverců (IWLS, iterative weighted least squares). Účelem této práce je najít optimální model vzhledem k metrice MAPE (viz 2.6). Metoda IWLS má však jinou účelovou funkci. Byly vytvořeny tzv. optimalizované modely, kdy se regresní parametry, které byly výstupem funkcí `glm()` a `glm.nb()`, optimalizovaly na trénovací sadě vzhledem k funkci

$$f(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|ACT_i - FIT_i|}{ACT_i}, \tag{2.4}$$

kde hodnoty FIT_i jsou závislé na volbě vektoru regresních parametrů $\boldsymbol{\beta}$.

K této optimalizaci byla zvolena Hooke-Jeevesova metoda optimalizace nediferencovatelné funkce, protože funkce 2.4 nemá triviální derivaci.

Optimalizace byla prováděna ve dvou variantách:

1. na všech parametrech,

2. na parametrech nepříslušících proměnné PROD_ID.

V případě první varianty šlo o časově velmi náročné výpočty (kvůli vysoké dimenzi optimalizovaných parametrů), které nebylo možné s dostupným hardwarovým vybavením dokončit. Proto bylo ukončovací kritérium omezeno na délku kroku menší než 0.1. Proto tato optimalizace nezkonvergovala k minimu, avšak i tak přinesla zlepšení výsledků (viz kapitola 2.10.1).

Ve druhém případě již šlo o optimalizaci menšího množství parametrů, a tak bylo možné nechat algoritmus zkonvergovat.

Tím, že byly regresní parametry optimalizovány uvedenou metodou, je model omezen pouze na bodové predikce, zanedbáváme tedy všechny ostatní charakteristiky odhadu parametrů a vyrovnaných hodnot.

2.10. Výsledky

V této kapitole budou popsány predikční schopnosti vytvořených modelů pomocí metriky MAPE. Modely budou porovnány s modelem používaným pro predikci v ORIFLAME SOFTWARE s. r. o. Nutno podotknout, že se nejedná o predikci pomocí regrese, ale pomocí jedné z metod strojového učení. Taky budou prezentovány hodnoty regresních parametrů jednotlivých modelů a bude popsán jejich význam.

Klíč pro názvy jednotlivých modelů je následující:

- typ modelu
  - I)
  - II)
  - III)
- druh regrese
  - PO = Poissonovo rozdělení
  - NB = negativně-binomické rozdělení
- trénovací datová sada
  - full = trénováno na všech datech z let 2015-2017
  - Q8 = trénováno na 80% kvantilu dat z let 2015-2017
- optimalizace
  - bez = žádná optimalizace
  - polo = částečná optimalizace
  - opt = úplná optimalizace

Názvy modelů jsou potom ve tvaru:
TypModelu_DruhRegrese_TrénovacíDatováSada_Optimalizace.

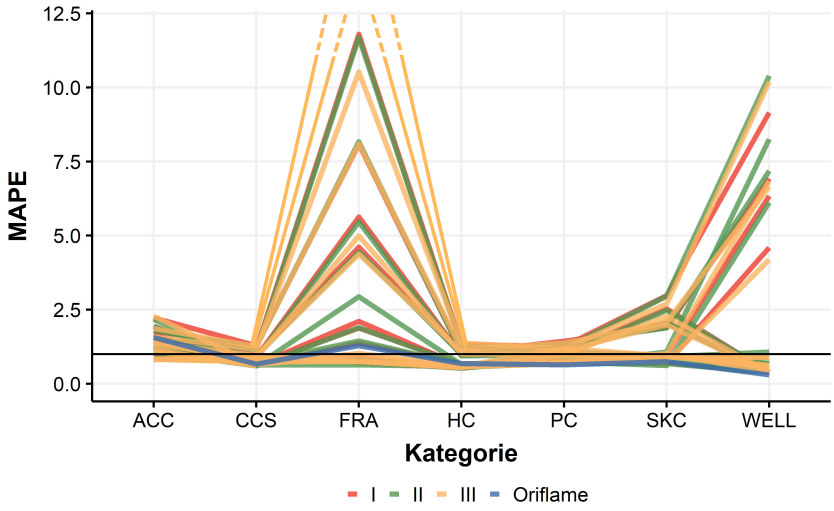
2.10.1. Porovnání modelů dle metriky MAPE

Jak již bylo vysvětleno v kapitole 2.9.1, bylo vytvořeno 252 regresních modelů. V tabulce 2.10 je pořadí nejlepších modelů dle MAPE vypočítaného na testovací sadě. Kromě čísla pořadí tabulka obsahuje název modelu a průměrné MAPE přes všechny výrobky.

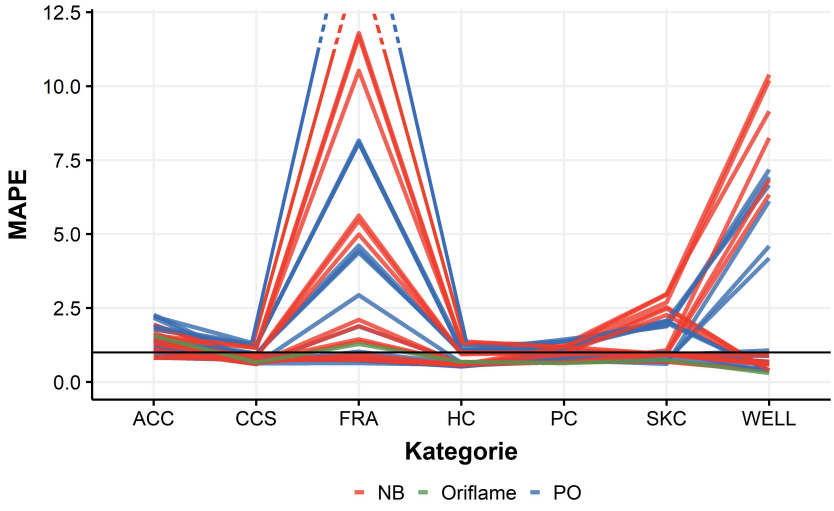
Tabulka 2.10: Pořadí nejlepších modelů dle MAPE na testovací sadě

Pořadí	Model	MAPE
1	II_PO_Q8_polo	0.7756
2	III_PO_Q8_polo	0.7763
3	I_PO_Q8_polo	0.7770
4	II_NB_Q8_opt	0.7792
5	ORIFLAME	0.7955
6	III_NB_Q8_polo	0.8044
7	II_NB_Q8_polo	0.8047
8	I_NB_Q8_polo	0.8070
9	I_PO_Q8_opt	0.8226
10	I_NB_Q8_opt	0.8285
...	...	...
36	Predikce v Pythonu	4.6914
...	...	...

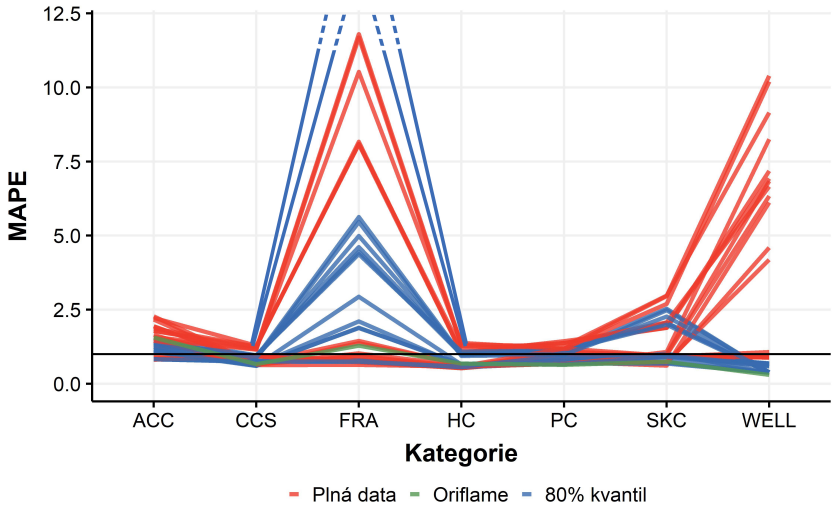
V grafech 2.15 - 2.17 jsou zobrazeny hodnoty MAPE pro všechny modely vypočítané podle vztahu (2.2) jako průměr pro jednotlivé kategorie. Tyto hodnoty jsou počítané pouze na testovací sadě a navíc je graf shora oříznut (pro lepší přehlednost). V prvním grafu jsou modely pomocí barev rozděleny na tři typy (I, II, III), ve druhém podle druhu regrese (NB a PO), ve třetím podle toho, jestli byl model tvořen na celých datech nebo jen na podmnožině dat (Plná data, 80% kvantil) a v posledním grafu podle druhu optimalizace (bez optim., polo-optim., optim.). Jasný trend v kvalitě modelu je možné pozorovat pouze na základě rozdělení modelů podle použitého druhu optimalizace. Je zřejmé, že nejlepší modely vznikly za pomocí polo-optimalizace.



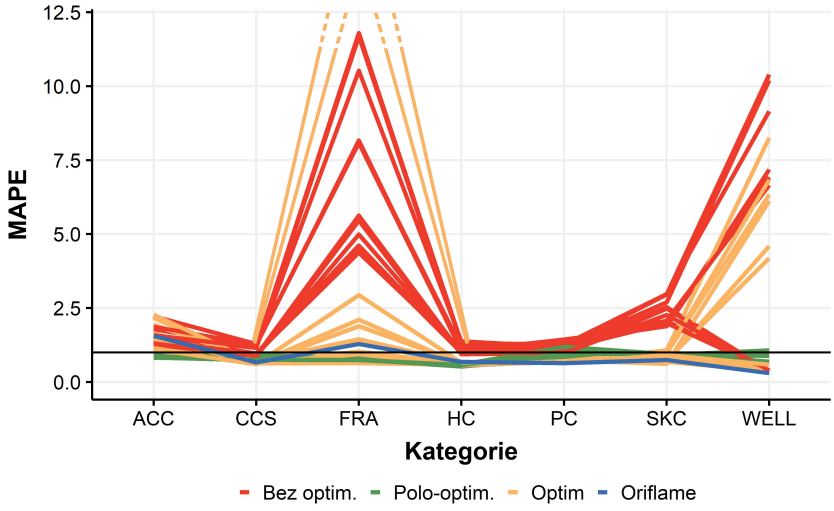
Obrázek 2.14: Hodnoty MAPE na testovací sadě pro všechny modely rozdělené dle typu modelu



Obrázek 2.15: Hodnoty MAPE na testovací sadě pro všechny modely rozdělené dle druhu regrese



Obrázek 2.16: Hodnoty MAPE na testovací sadě pro všechny modely rozdělené dle použitých dat



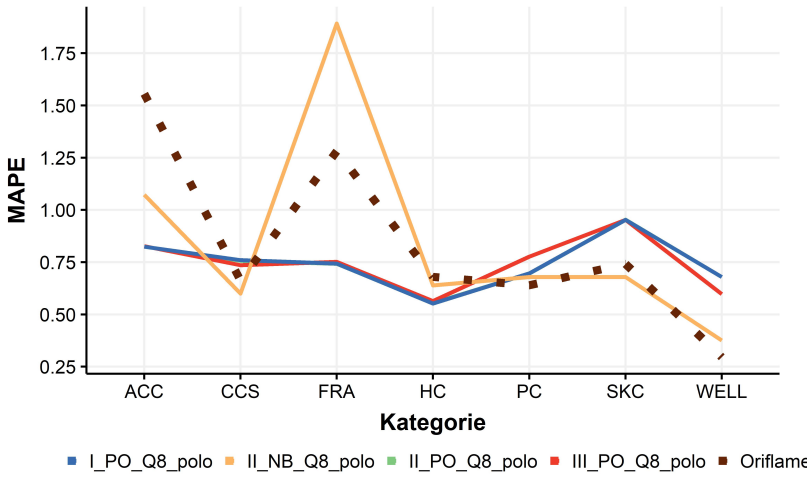
Obrázek 2.17: Hodnoty MAPE na testovací sadě pro všechny modely rozdělené dle optimalizace

V tabulce 2.11 jsou vypsány hodnoty MAPE pro jednotlivé kategorie na 4 nejlepších modelech (dle průměrných MAPE) a v posledním sloupci jsou hodnoty

pro model ORIFLAME. Jde o modely typu II, III a I s Poissonovým rozdělením (PO) na 80 % dat (Q8) s částečnou optimalizací (polo). Tyto hodnoty jsou pro lepší přehlednost vizualizovány na obrázku 2.18. Z průměrných hodnot pro jednotlivé modely vyplývá, že mezi jednotlivými typy modelů není příliš velký rozdíl. A proto by bylo možné místo složitějšího modelu typu III použít základní model typu I bez větší újmy na kvalitě. Taky je zřejmé, že částečná optimalizace na podmnožině vede k nejlepším modelům. Dále v této práci budou prezentovány hodnoty týkající se modelu *I_PO_Q8_polo*, který bude označován jako vítězný model.

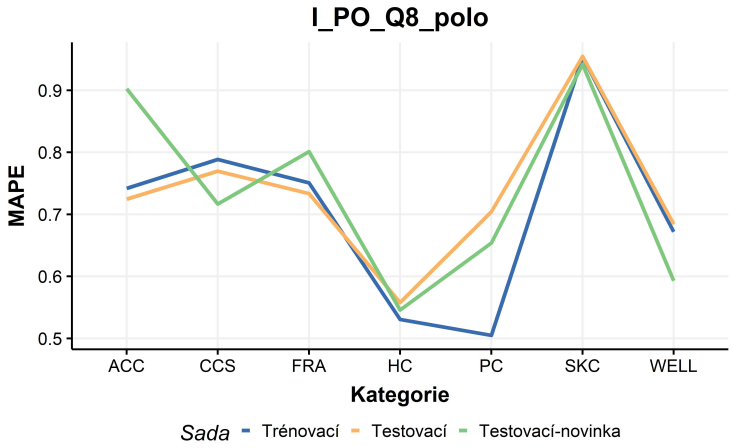
Tabulka 2.11: Hodnoty MAPE pro 3 nejlepší modely na testovací sadě

	II_PO_Q8_polo	III_PO_Q8_polo	I_PO_Q8_polo	II_NB_Q8_opt	ORIFLAME
ACC	0.8241	0.8269	0.8241	1.0726	1.5563
CCS	0.7569	0.7357	0.7598	0.5998	0.6616
FRA	0.7423	0.7517	0.7438	1.8921	1.2813
HC	0.5548	0.5642	0.5527	0.6394	0.6804
PC	0.6962	0.7764	0.6971	0.6794	0.6396
SKC	0.9534	0.9531	0.9531	0.6792	0.7422
WELL	0.6792	0.5972	0.6791	0.3757	0.2958
Průměr	0.7756	0.7763	0.7770	0.7792	0.7955



Obrázek 2.18: Hodnoty MAPE pro 4 nejlepší modely na testovací sadě

Obrázek 2.19 zobrazuje hodnoty MAPE pro model I_PO_Q8_polo na trénovací a testovací sadě. Testovací sada je poté ještě rozdělena na stávající výrobky, které se vyskytovaly i v sadě trénovací a poté na novinky, které do modelu vstoupily poprvé. Je zřejmé, že nejde o případ overfittingu (viz kapitola 2.9), protože MAPE na testovací sadě není výrazně vyšší než na trénovací sadě. V tomto případě nebyl overfitting ani očekáván, jelikož počet odhadovaných parametrů je výrazně menší než počet pozorování (nabídek).



Obrázek 2.19: Hodnoty MAPE na trénovací a testovací sadě (rozdělená na novinky a stávající výrobky)

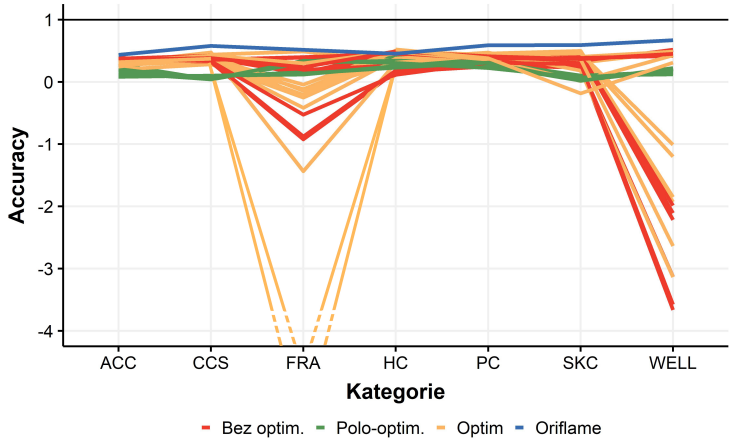
2.10.2. Porovnání modelů dle accuracy

Jak již bylo uvedeno v kapitole 2.6, za rozhodující metriku je v této práci považována MAPE. Pro porovnání s procesy probíhajícími v ORIFLAME SOFTWARE s. r. o. zde budou vytvořené modely zběžně porovnány i na základě metriky accuracy. Tabulka 2.12 obsahuje 10 nejlepších modelů právě na základě accuracy vypočítaných jako průměr na testovací sadě.

Tabulka 2.12: Pořadí nejlepších modelů dle accuracy na testovací sadě

	Model	Accuracy
1	ORIFLAME	0.5666
2	II_PO_full_opt	0.4294
3	I_PO_Q8_bez	0.3812
4	II_PO_Q8_bez	0.3800
5	III_PO_Q8_bez	0.3777
6	III_NB_Q8_bez	0.3770
7	II_PO_full_bez	0.3666
8	III_NB_full_opt	0.3656
9	III_PO_full_bez	0.3618
10	II_NB_Q8_bez	0.3593
...	...	...

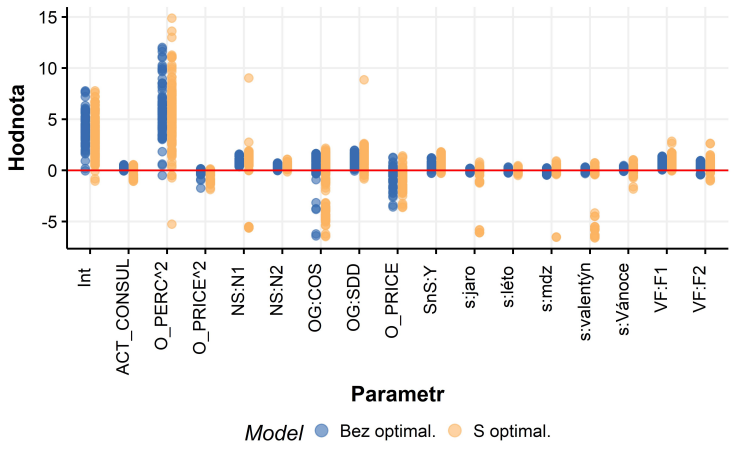
Pro jednotlivé kategorie lze hodnoty accuracy porovnat na grafu 2.20 (omezený zdola pro lepší přehlednost). Je zřejmé, že v tomto případě se jednotlivé skupiny modelů (bez optimalizace, s polo-optimalizací a s optimalizací) vzájemně promíchávají a nelze sledovat jednotné chování. Je to dáno faktem, že optimalizace probíhala vzhledem k funkci MAPE. Modrá linie opět zobrazuje accuracy modelu ORIFLAME SOFTWARE s. r. o.



Obrázek 2.20: Hodnoty accuracy pro všechny modely na testovací sadě

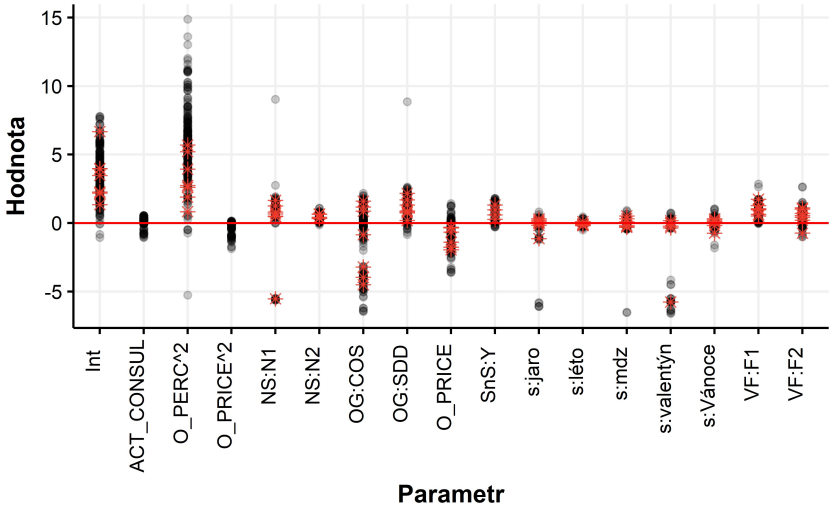
2.10.3. Velikost efektů (regresní parametry modelů)

Celkem bylo vypočítáno 3822 parametrů nepříslušejících proměnné PROD_ID. Jejich hodnoty zobrazuje graf 2.21, kde je možné rozlišit parametry modelů, které byly optimalizovány od těch, které optimalizovány nebyly. Je zřejmé, že optimalizace vedla spíše k extremizaci parametrů.



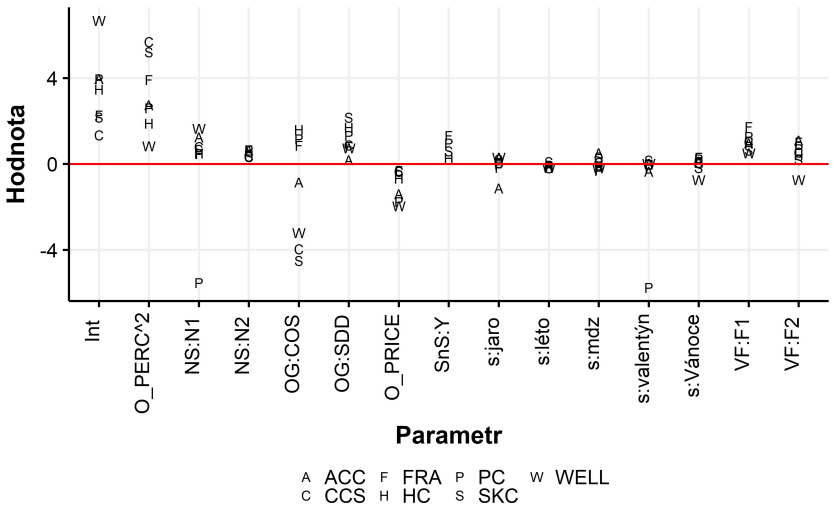
Obrázek 2.21: Hodnoty parametrů ve všech modelech

Graf 2.22 zobrazuje opět parametry všech modelů nenáležící proměnné PROD_ID se zvýrazněnými parametry vítězného modelu. Ten je blíže specifikován v kapitole 2.10.1. Obecně lze soudit, že parametry vítězného modelu nenabývají nikterak extrémních hodnot.



Obrázek 2.22: Parametry vítězného modelu (červené hvězdičky) ve srovnání se všemi ostatními

Obrázek 2.23 zobrazuje parametry pouze vítězného modelu rozdělené po jednotlivých kategoriích. Pro lepší pochopení obrázku je dobré připomenout, že model se dělí na sedm podmodelů odpovídajících právě sedmi kategoriím výrobků. U každého názvu parametru tedy nalezneme sedm vyznačených bodů.



Obrázek 2.23: Parametry vítězného modelu

V následujících odstavcích budou okomentovány velikosti efektů podle jednotlivých kategorií, které jsou odlišeny na obrázku 2.24. Nutno poznamenat, že jsou zobrazeny parametry všech modelů, tj. i těch, které neměly dobré predikce. Proto některé hodnoty nejsou příliš vypovídající. Přesnou interpretaci jednotlivých velikostí efektů lze odvodit na základě kapitoly 1.2.4.

Hodnoty absolutního členu (Int.) jsou ve všech kategoriích kladné, až na kategorii FRA. Tyto hodnoty jsou očekávané počty prodaných výrobků při základních úrovních všech kategorických proměnných a nulových hodnotách spojitých proměnných. Záporná hodnota absolutního členu díky logaritmické spojovací funkci (respektive její inverzi - exponenciální hodnotě) značí kladné prodeje. V realitě však nemůže nastat například situace, kdy je prediktor OFFER_PRICE (cena) roven nule a zároveň sleva na své základní úrovni (bez slevy). Proto je tento parametr neinterpretovatelný.

Počet aktivních konzultantů (ACT_CONSUL) má čistě kladné hodnoty parametrů pouze v případě HC a PC. Nejčastěji nabývají malých kladných hodnot. Nutno připomenout, že tento parametr se nevyskytuje v modelech typu I.

Sleva (O_PERC^2) má dle očekávání vysoce kladný vliv na prodej výrobků. Výjimku tvoří kategorie FRA, kde v některých případech koeficienty ukazují záporný vliv. Mohlo by jít o fakt, že zákazníci mají u vůní tendenci kupovat si své oblíbené bez ohledu na slevu.

Stejně tak u ceny (O_PRICE případně O_PRICE^2), kde je očekáván záporný vliv na prodej, se nejspíš z podobných důvodů u kategorie FRA ukazuje vliv opačný. Kladný efekt, i když ne již tak výrazně, měla cena i v několika případech u kategorie SKC a WELL. Opět je nutné zdůraznit, že cena v druhé mocnině figurovala pouze v modelech typu III.

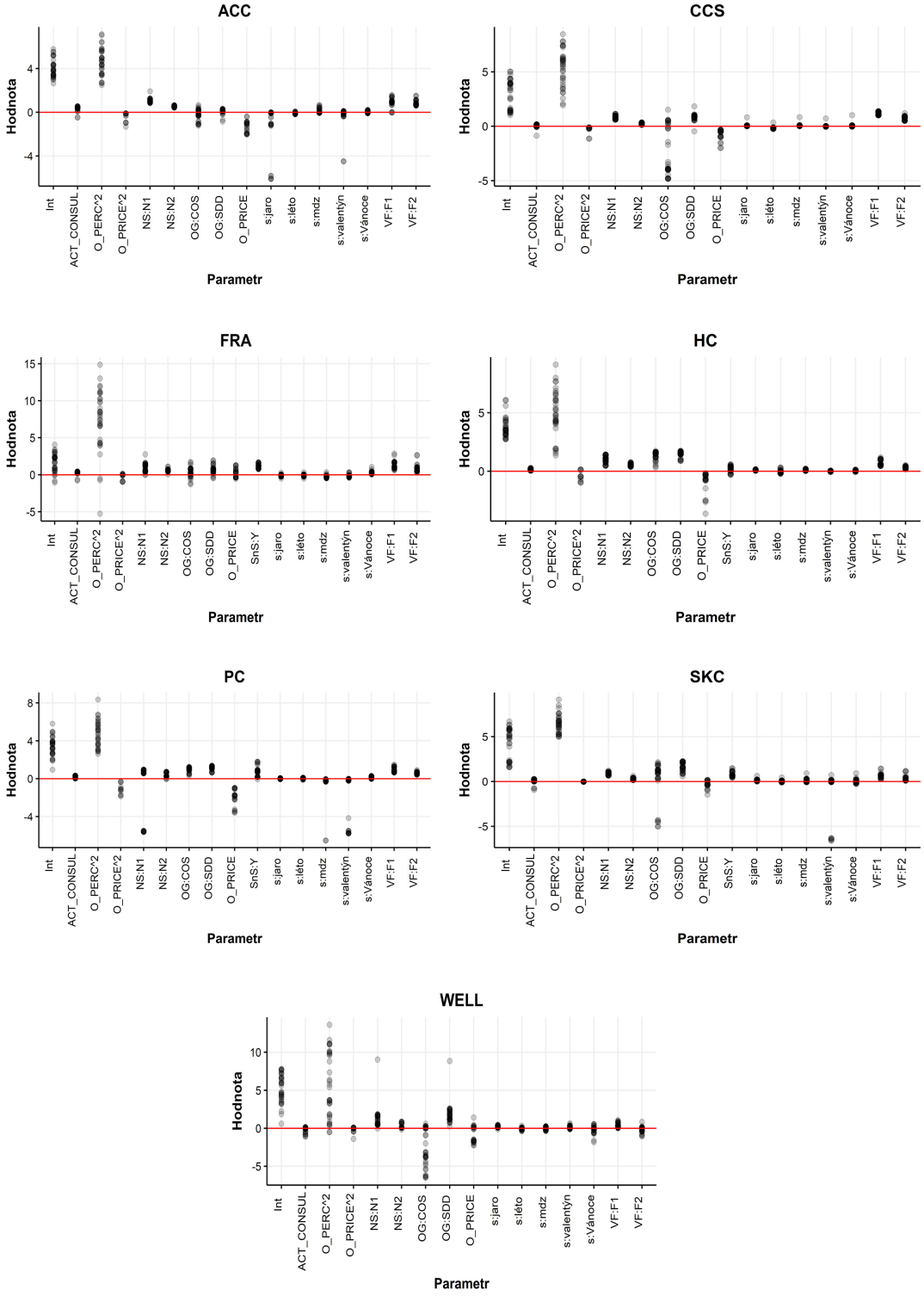
Proměnná NEW_SPLASH (NS) má z drtivé většiny kladný efekt na prodej. To naznačuje, že zákazníci si rádi objednávají novinky, případně výrobky uvedené na trh podruhé.

Obě úrovně druhu nabídky (OG) mají čistě kladný efekt pouze v případě kategorie HC a PC. Úroveň COS nabývá překvapivě často záporných hodnot.

Možnost vyzkoušet si vůni výrobku (SnS) má v případě kategorie FRA kladný efekt na prodej, což odpovídá očekávání. Tento prediktor se také s kladnými koeficienty v drtivé většině vyskytuje u kategorií HC, PC a SKC.

Vliv jednotlivých sezón (s) je pro každou kategorii jiný. V případě ACC na prodej kladně působí sezóna MDŽ a Vánoc. U CCS jsou kladné efekty všech úrovní sezón. Kategorie FRA se z pohledu sezóny nejvíce prodává o Vánocích. U HC je nejvýraznější efekt léta. Výrazný záporný efekt má sezóna Valentýn v případě kategorie PC. Podobný jev nastal u kategorie SKC, kde mají však ostatní sezóny vyšší kladné hodnoty. V případě kategorie WELL jsou efekty jak kladné tak záporné pro všechny úrovně sezón.

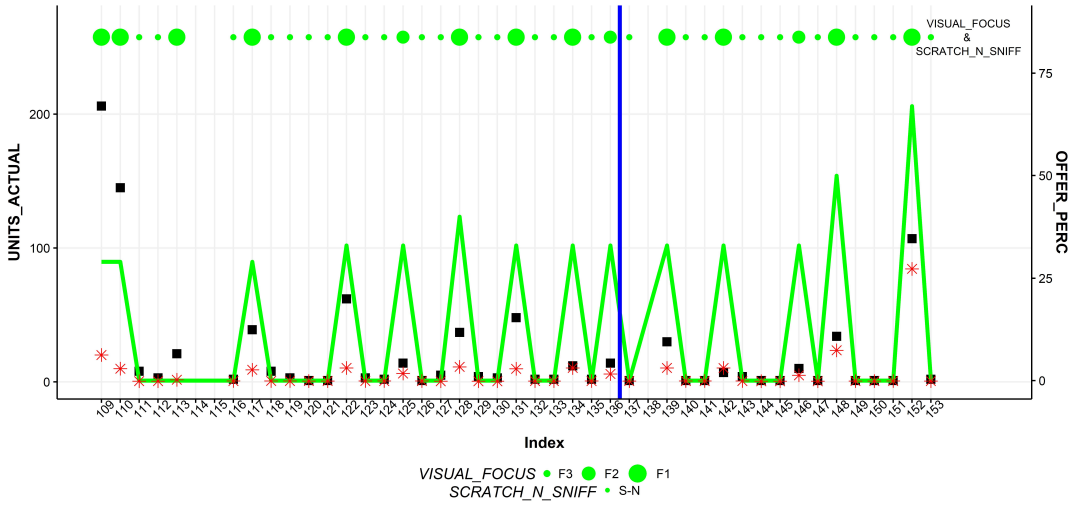
Efekt největšího rozměru obrázku (VF) je ve všech kategoriích větší než při střední velikosti. Pouze u kategorie ACC a SKC se zdají býti docela vyrovnané.



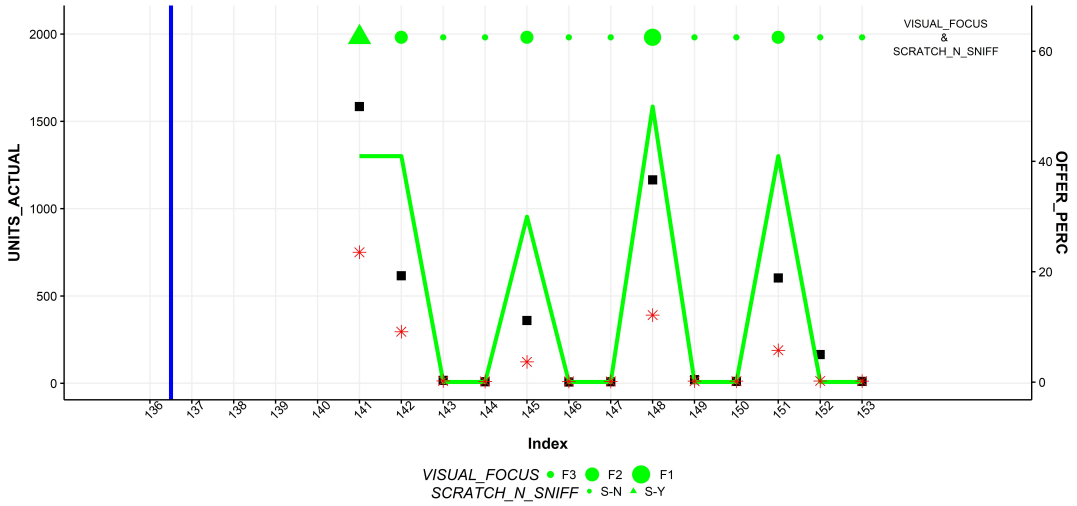
Obrázek 2.24: Parametry modelů pro jednotlivé kategorie

2.10.4. Vizualizace predikcí

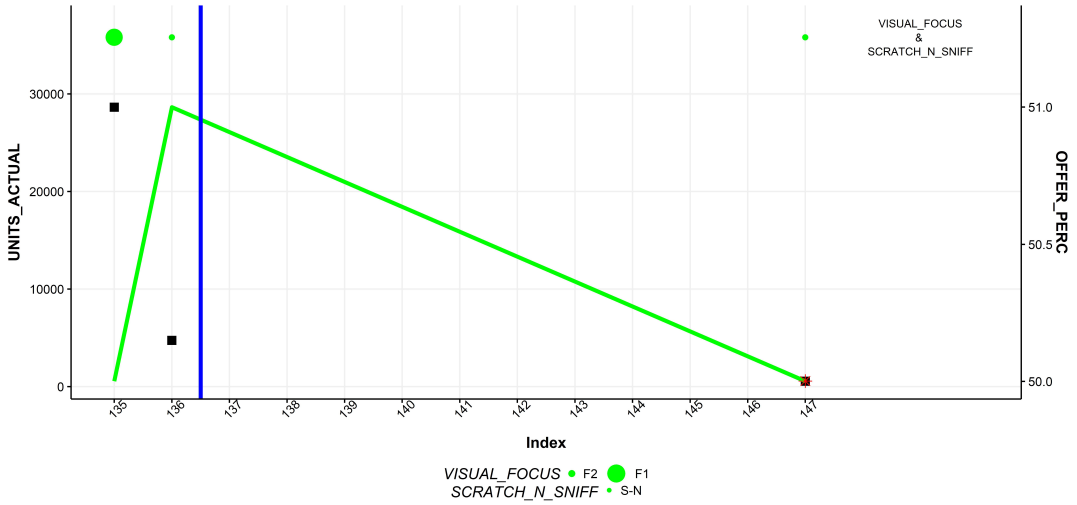
Na následujících stránkách jsou graficky znázorněny příklady predikcí prodejů vybraných výrobků za použití vítězného modelu. Černé čtverce značí skutečnou hodnotu prodeje, červené hvězdy jsou predikce. Svislá modrá čára odděluje období zahrnuté do trénovací sady (do kampaně číslo 136) od období v testovací sadě (od kampaně číslo 137). Na horizontální ose je číslo kampaně, na hlavní vertikální ose je počet prodaných kusů a na vedlejší vertikální ose je sleva. Zelená čára znázorňuje hodnotu slevy dané nabídky a přísluší k vedlejší vertikální ose. Symboly nad grafem znázorňují, jestli nabídka disponuje scratch&sniff oblastí nebo ne a jejich velikost odpovídá velikosti obrázků v katalogu (F1 je největší, F3 je nejmenší).



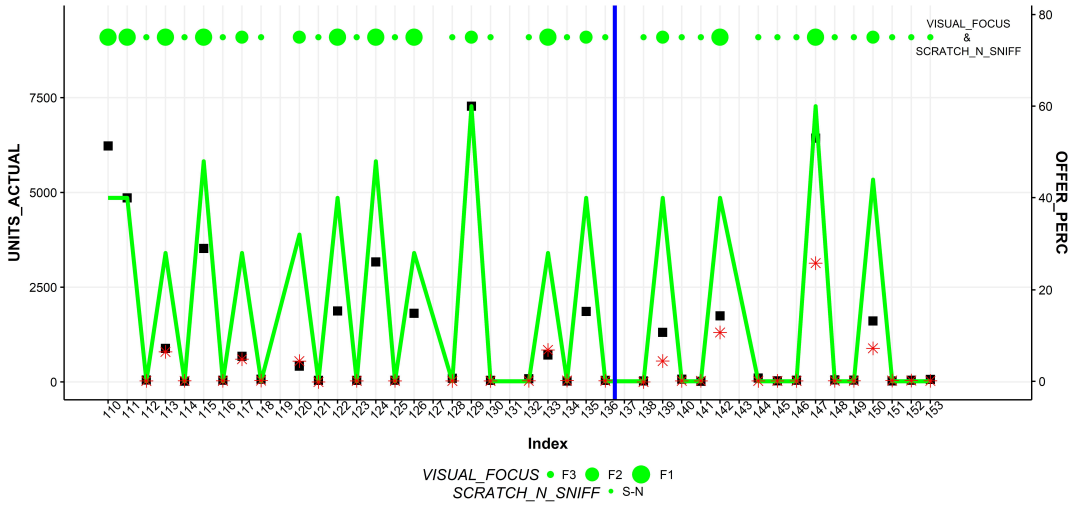
Obrázek 2.25: Predikce na výrobku z kategorie Dekorativní kosmetika, MAPE na trénovací sadě 0.698, na testovací sadě 0.432



Obrázek 2.26: Predikce na výrobku z kategorie Péče o vlasy, nový výrobek (bez dat v trénovací sadě), MAPE na testovací sadě 0.430



Obrázek 2.27: Predikce na výrobku z kategorie Péče o tělo, prodeje v trénovací sadě nad 80% kvantilem, proto data nebyla pro trénink použita, MAPE na testovací sadě 0.022



Obrázek 2.28: Predikce na výrobku z kategorie Péče o tělo, některé prodeje v trénovací sadě nad 80% kvantilem, proto model neumí zcela dobře predikovat odlehlé hodnoty v kampani 147 (predikuje expert), MAPE na trénovací sadě 0.429, na testovací sadě 0.471

Závěr

Cílem této práce bylo prozkoumat teoretické pozadí a využití zobecněných lineárních modelů v predikci prodeje. V teoretické části byly vysvětleny základy obyčejných (nezobecněných) lineárních modelů jakožto nejčastěji používané metody regresní analýzy. Důraz byl kladen na odhady parametrů, jejich interpretaci a také využití kategorických proměnných, které jsou v datech o výrobcích velmi časté. Následující kapitola se již zaměřovala na zobecněné lineární modely, které jsou rozšířením obyčejných lineárních modelů o případy, kdy není zajištěna normalita rozdělení vysvětlované proměnné (tj. vysvětlovaná proměnná má jakoukoliv distribuci z exponenciální třídy distribucí) či je potřeba využít jinou spojovací funkci než je funkce identity. Byly popsány tři základní složky zobecněných lineárních modelů, odvozeny základní charakteristiky obecně pro distribuce z exponenciální třídy rozdělení a taky byly popsány dva způsoby odhadů parametrů zobecněných lineárních modelů. Dále se teoretická část zaměřila na distribuce popisující počet, tj. na modely s Poissonovým rozdělením a negativně-binomickým rozdělením.

Praktické využití popsaných metod bylo testováno ve druhé části diplomové práce. Byla popsána základní myšlenka praktické části, na jejímž základě bylo vytvořeno 252 modelů jako kombinace různých přístupů – tři typy předpisů regresního modelu, Poissonovo či negativně-binomické rozdělení, modelování na celých datech či na 80 % dat s nejmenšími hodnotami, bez optimalizace a s částečnou či úplnou optimalizací. Modely byly srovnávány na základě dvou metrik, jejichž klady a zápory byly diskutovány v samostatné sekci. Výsledkem pak bylo pořadí modelů na základě zvolené metriky MAPE a výběr nejvýhodnějšího mo-

delu. Ten byl také porovnán s výstupy predikčního nástroje společnosti Oriflame, na jejíž datech se praktická práce zakládala. V závěru práce byly prezentovány vizualizace predikcí prodejů několika výrobků, na nichž byly demonstrovány vzniklé komplikace a jejich řešení.

Výsledky praktické části lze shrnout následovně: na přesnost predikcí měla největší vliv následná optimalizace parametrů modelu. Ta byla provedena pomocí Hooke-Jeevesovy metody optimalizace nediferencovatelné funkce více proměnných a hledala takové regresní parametry, ve kterých funkce MAPE nabývá svého minima. Podstatné také bylo budovat model pouze na 80 % dat s nejnižším prodejem, jelikož zbývající data velmi zhoršují kvalitu predikcí a v praxi jsou vysoké prodeje predikovány experty. Optimalizace na všech parametrech se nejevila jako příliš efektivní hlavně vzhledem k časové náročnosti. Zobecněné lineární modely tedy jsou použitelné v oblasti predikce prodeje, což ovšem nevylučuje možnost, že existují lepší metody k dosažení tohoto cíle (např. pokročilé metody Machine Learningu).

Dalšími kroky pro zlepšení predikčních schopností modelu by mohlo být:

- optimalizace všech parametrů bez omezení ukončovacího kritéria,
- více iterací trénování modelu (využití lepší metody křížové validace),
- tvorba jednoho modelu na všech kategoriích dohromady,
- zahrnutí informace o variabilitě koeficientů do hodnocení modelu.

Tyto kroky jsou podmíněny větší výpočetní silou, tj. využitím jiného počítače či výpočtem na cloudu.

Literatura

[1] AGRESTI, Alan. An Introduction to Categorical Data Analysis [online]. 2nd Edition. New Jersey: JohnWiley & Sons, Inc., Hoboken, New Jersey, 2007 [cit. 2020-06-15]. ISBN 978-0-471-22618-5.
Dostupné z: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf.

[2] Annual Report 2019 [online]. [cit. 2020-04-17].
Dostupné z: https://vp233.alertir.com/sites/default/files/report/oriflame_annual_report_2019_final_compressed_0.pdf?v2assets.

[3] HILBE, Joseph M. Negative Binomial Regression [online]. 2nd Edition. New York: Cambridge University Press, 2011 [cit. 2020-07-05]. ISBN 978-0-521-19815-8.
Dostupné z: https://www.cambridge.org/core/books/negative-binomial-regression/12D6281A46B9A980DC6021080C9419E7.

[4] HRON, Karel. Základy počtu pravděpodobnosti a metod matematické statistiky. 2. dopl. vydání. Olomouc : Univerzita Palackého v Olomouci, 2015. ISBN 978-80-244-4774-2 (váz.).

[5] KOUNDIYA DESIRAJU. ggplot theme for publication ready Plots. RPubs [online]. 2015 [cit. 2020-07-05].
Dostupné z: https://rpubs.com/Koundy/71792.

[6] KUTNER, Michael H., Christopher J. NACHTSHEIM, John NETER a William LI. Applied Linear Statistical Models [online]. 5th Edition. New York: McGraw-Hill Irwin, 2005 [cit. 2020-06-15]. ISBN 0-07-238688-6.
Dostupné z: http://users.stat.ufl.edu/ rohitpatra/4210/KNNL.pdf.

[7] MACHALOVÁ, Jitka a Horymír NETUKA. Numerické metody nepodmíněné optimalizace. Olomouc: Univerzita Palackého v Olomouci, 2013, 142 s. Skripta. ISBN 978-80-244-3403-2.

[8] MCCULLAGH, P. a J. A. NELDER. Generalized Linear Models [online]. 2nd Edition. London, Harpenden: Chapman and Hall/CRC, 1983 [cit. 2020-07-05]. ISBN 978-0412317606.
Dostupné z: http://www.utstat.toronto.edu/ brunner/oldclass/2201s11/readings/glmbook.pdf.

[9] Nadměrný rozptyl - overdispersion. Matematická biologie [online]. Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity [cit. 2020-07-15].
Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat–regresni-modelovani–logisticky-regresni-model-a-jine-zobecnene-linearni-modely–nadmerny-rozptyl-overdispersion.

[10] Oriflame [online]. [cit. 2020-04-10].
Dostupné z: https://cs.wikipedia.org/wiki/Oriflame.

[11] Oriflame app [online]. [cit. 2020-04-10].
Dostupné z: https://play.google.com/store/apps/details?id=com.oriflame.oriflame&hl=en.

[12] ORIFLAME CESTA K ÚSPĚCHU, vydání pro lídry, vnitřní brožura pro konzultanty společnosti Oriflame

[13] Oriflame - Sweden -Who we are [online]. [cit. 2020-04-10].
Dostupné z: https://corporate.oriflame.com/About-Oriflame/.

[14] PAYNE, Elizabeth H., Mulugeta GEBREGZIABHER, James W. HARDIN, Viswanathan RAMAKRISHAN a Leonard E. EGEDE. An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. Cummun Stat Simul Comput. 2018, 2018(47), 1722–1738. DOI: 10.1080/03610918.2017.1323223.

[15] STATISTICSMATT. Generalized Linear Models. YouTube [online]. 2020 [cit. 2020-07-15].
Dostupné z: https://www.youtube.com/playlist?list=PLmM_3MA2HWpYoG_q69ZzbCSwvriLuimoO.

[16] STATISTICSMATT. Exponential Family. YouTube [online]. 2020 [cit. 2020-07-27].
Dostupné z: https://www.youtube.com/playlist?list=PLmM_3MA2HWpZGS954BEBe0eu6kc2qw7oW.

[17] Sufficient statistics. Wikipedia [online]. 18.5.2020 [cit. 2020-08-10].
Dostupné z: https://en.wikipedia.org/wiki/Sufficient_statistic#cite_note-Fisher1922-1.

[18] What is Direct Selling, and Is it a Good Home Business Option?[online]. [cit. 2020-04-10].
Dostupné z: https://www.thebalancesmb.com/what-is-direct-selling-1794391.

[19] Why and how to Cross Validate a Model? [online]. [cit. 2020-04-14].
Dostupné z: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f.